Wireless Communications and Mobile Computing

# Metaheuristic Algorithms for Big Data Analytics within the Internet of Things

Lead Guest Editor: Rajesh Kaluri
Guest Editors: Thippa Reddy Gadekallu, Celestine Iwendi, Lalit Garg, and Mamoun Alazab

# Metaheuristic Algorithms for Big Data Analytics within the Internet of Things

Wireless Communications and Mobile Computing

# Metaheuristic Algorithms for Big Data Analytics within the Internet of Things

Lead Guest Editor: Rajesh Kaluri
Guest Editors: Thippa Reddy Gadekallu, Celestine Iwendi, Lalit Garg, and Mamoun Alazab

# Chief Editor



## Associate Editors



## Academic Editors

# Contents

# Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9820259, 1 page
https://doi.org/10.1155/2023/9820259

*Retraction*

# Retracted: The Importance of Traditional Sports into College Physical Education Based on Big Data Dynamic Programming Algorithm

**Wireless Communications and Mobile Computing**

Received 3 October 2023; Accepted 3 October 2023; Published 4 October 2023



This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

In addition, our investigation has also shown that one or more of the following human-subject reporting requirements has not been met in this article: ethical approval by an Institutional Review Board (IRB) committee or equivalent, patient/participant consent to participate, and/or agreement to publish patient/participant details (where relevant).

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

## References

[1] Z. Zheng, "The Importance of Traditional Sports into College Physical Education Based on Big Data Dynamic Programming Algorithm," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 2996940, 13 pages, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 1149842, 18 pages
https://doi.org/10.1155/2022/1149842

*Research Article*

# IoT-Based Response Time Analysis of Messages for Smart Autonomous Collision Avoidance System Using Controller Area Network

**Anil Kumar Biswal,[1] Debabrata Singh,[2] Binod Kumar Pattanayak,[1] Debabrata Samanta,[3] Amit Banerjee,[4] A. Y. Seteikin,[5,6] and I. G. Samusev[5]**

[1]*Department of CSE, ITER, SOA Deemed to Be University, India*
[2]*Department of CA, ITER, SOA Deemed to Be University, India*
[3]*Department of Computer Science, CHRIST University, India*
[4]*Physics Department, Bidhan Chandra College, Asansol 713 303, India*
[5]*Immanuel Kant Baltic Federal, University, Kaliningrad, 236000, Russian Federation and Amur State University, Blagoveshchensk 675027, Russia*
[6]*Amur State University, Blagoveshchensk 675027, Russia*

Correspondence should be addressed to Amit Banerjee; amitbanerjee.nus@gmail.com and I. G. Samusev; is.cranz@gmail.com

Received 12 August 2021; Accepted 3 March 2022; Published 8 April 2022

Academic Editor: Rajesh Kaluri



Many accidents and serious problems occur on the road due to the rapid increase in traffic congestion in all sections of the country. Autonomous vehicles provide a solution to successfully and cost-effectively avoid this problem while minimizing user disruption. Currently, more engaging electromechanical elements with an analog interface are used to develop affordable automobiles for efficient and cost-effective operation for a smart driving platform with a semiautonomous automobile, strengthening the vehicle involvement of the driver while increasing safety. As a result, it takes longer for various car elements to respond, which causes more problems during message transmission. This project aims to create a Controller Area Network (CAN) for analyzing message response times by incorporating a few application nodes on the IoT platform, such as an antilock braking system, flexible cruise control, and seat belt section, for some real-time control system applications. These application nodes are car analytical parts that are linked to IoT modules to prevent collisions. An autonomous device for collision avoidance and obstacle detection in a vehicle can impact road accidents if the CAN protocol is implemented.

## 1. Introduction

In recent years, traffic congestion, driver drowsiness and reckless driving represent a big problem in the different areas around the world that are seriously affected by road accidents in the transport system [1]. Unconditional circumstances are controlled by intelligent autonomous vehicles because of the above. Therefore, through the implementation of collision avoidance mechanisms in the vehicle, the automation domain offers a forum for monitoring reckless driving as well as driver fatigue [2]. The "Automotive Serial Controller Area Network" protocol is used for designing an intelligent control car with a huge range of serial bus communication control system [3].

Large types of embedded systems require high-speed communication platforms for providing automotive industrial control. But various industries are not supported with automation that needs to the operated with Controller Area

Network (CAN) protocol [4]. A serial communication bus like CAN was considered in the International Standardization Organization (ISO), which has replaced the complex wiring control with a two-wire bus, thereby adding a multi-master communication serial bus that can transmit messages to various parts of the network system [5].

Currently, the Automobiles are being constructed with microcontrollers and more electrical parts as we know that is the central part of the controlling unit and various types of devices or circuitries connected to it. This process is very complex to interpret and improve performance by using several connections and electrical lines linked to a microcontroller [6]. The communication area for the project is an implanted CAN networking system which provides effective data transfer, allowing multiple microcontrollers and devices to be connected with a popular CAN bus using the CAN protocol, then the connection of all items with the consideration of optimum priority and speed [7].

This protocol also offers a high-speed serial data frame communication interface, low-cost physical medium, short message frame length, and, at the same time, it adds a high-level detection or correction mechanism for errors in different communication network nodes [8]. The evolution of embedded systems and software has been used in modern times to build smart autonomous vehicles over 40-50 percent globally and this percentage of progress is only expected to increase with improvising road safety and security features [9] [10]. Due to the process of digitization for constructing smart vehicles with the use of IoT modules that can create a huge number of datasets [11]. So that, the dimensionality reduction and security of datasets are required to be managed and also yield good results through the blockchain-big data technique [12]. In general, the protocols between the network (sensor) nodes for physical communication with the IoT data link layers, the sensor nodes, are described by the CAN protocol [13]. Here, the ultrasonic sensor node application process is used to measure the distance between the vehicle and the road barrier [14] [15].

This system determines the distance of the obstacle by an ultrasonic sensor to control the motor speed that has been designed using Arduino UNO, IoT modules, and CAN base serial communications protocol. When the sensor node is connected to the CAN bus, which provides a rapid response to measure the distance of an obstacle, the message is automatically forwarded to the Arduino Uno module to track the vehicle's engine movement and steering. This proposed system provides an environment to enhance the driver-vehicle platform to make a semi-autonomous vehicle system with the help of developing and implementing a digital driving system [16] [17]. The optimal response time calculation is impossible in the existing system, which takes maximum time to communicate with each part of the vehicle due to massive datasets. These datasets shorten the response time of an existing system that causes vehicle accidents while driving.

*1.1. Motivation of This Work.* Designing a process of response time analysis of a smart autonomous collision avoidance system messages based on IoT modules and CAN serial communications protocol to prevent any road accidents by taking an optimized range of message length and message ID for providing timing response in conversion. In recent years, the autonomous operation has been extensively applied to vehicles for road safety issues.

*1.2. Contribution to This Work*

(i) The smart vehicle is developed by the Controller Area Network (CAN) that is accessed in various real-time suits to link internal-level communication facilities to shared units of car control systems, e.g., industrial and home automation and medical equipment, which is a "broadcast" type of bus. In other words, there is no address part of sending or receiving nodes. The network can accept to receive or transmit the messages sent by all nodes, where the acceptance test is performed after receiving the message from each node

(ii) The messages are checked by each node whether it is irrelevant to that particular node or not. When a message is pertinent, then it is received by that node. Otherwise, it is not accepted. The priority node can send the first message for transmission, which depends on an 11-bit identifier. Here, an identifier is uniquely identified all over the network and is used to tag the content of the message. A numeric value is added to each message, which controls its priority on the bus, thereby recognizing the contents of the message

(iii) When the bus is not loaded by any task, then some nodes can be ready to communicate with each other. But during this period, where the CAN bus attempts to forward messages from more than two nodes concurrently, then the identifier field is uniquely defining the priority of the message through the network. The messages are securely transmitted in the sequence of priority without missing anyone, which is possible with this technique. If a numerical value of the identifier is lower, then it is treated as a higher priority. This means that the message with more prevalent ID bits (i.e., Bit 0) will overwrite all nodes so that only the predominant message will finally be acknowledged by each node after arbitration of the ID

(iv) Using the Arduino controller and serial communication protocol on different device nodes, collision avoidance, and obstacle detection techniques are implemented on the smart vehicle via IoT modules. To detect any obstacle through a different ultrasonic sensor on the road, we can produce a message frame to relay the node to a vehicle's engine

The rest of the paper is structured as follows: the literature survey pertaining to this field is included in Section 2. Section 3 describes the network protocol model and the proposed system, and Section 4 represents the proposed framework and operating theory, along with its implementation processes and components. In Section 5, we describe the simulation setup and result from analysis, and in Section 6, we conclude our paper with some references.

## 2. Literature Survey

Control Area Network (CAN) is detailed in [18] that provides a communication network between control units in automotive industries. CAN provides vast advantages and then it is widely used in distinct industries including military, aviation, electronics, factories and many more. Here, the microcontrollers and devices communicate among themselves using CAN in the absence of a host computer and there is no need to follow heavy access of the main controller. In [19], The author describes CAN-BUS to be an essential network technology for communication which is implemented in the automobile network communication sector with some characteristics like real-time implementation, reliability, and flexibility.

In [20], the current wireless home automation requires a greater amount of RF recipients and thus the frequency range varies. Electromagnetic waves may lose messages, and the cost and complexity of multi-home automation will be significant. This process is crucial for detecting problem areas. Due to that, it allows controlling only limited devices. The author talks about a Controller Area Network (CAN) bus used to send and receive messages between automotive devices [21]. There are possibilities of errors when transmitting messages via the nodes. To detect those errors, a Controller Area Network Adaptive Fault Diagnostic Algorithm detects all of the CAN's defective nodes.

In [22], the authors discuss the current parameters of vital signs for patients in critical care units; patient's bedside are equipped with devices that keep intensivists and other medical staff informed. This information allows paramedical personnel to take the necessary measures for disease prevention and cure. Extracting CAN messages from automotive ECUs can be made successfully as detailed in [23]. It gives the details of the construction of software and hardware, which interfaces directly into the car with the CAN network.

It includes CAN bus transceiver behavioural models. Thermal behaviour can be allowed for different types of simulations for verification in reasonable CPU time from core verifications for detailed analysis of the integrity of signals [24]. This review examines the research done on the Controller Area Network (CAN) reliability analysis. In recent decades, schedulability analysis has been extended to an advanced technique that can determine whether or not the time limits of several jobs performed by a single CPU or a distributed system for nontrivial systems. [25]. This is a description and illustration of a reliability analysis method that focuses on auto systems based on CAN, that also considers the impact of the error on schedulability analysis [26].

In [27] proposed an automotive CAN cluster for processing messages by using a gateway mechanism. So, this is used for worst-case response analysis (WCRT) for finding lower and upper bound on the response time of the CAN cluster of automobiles. It is efficient to monitor a large-scale CAN cluster, and then its performance will be improved by reducing unnecessary conservation in process of designing. The WCRT analysis of CAN with sporadic message execution in a multicore automotive gateway protocol has been designed by Xie et al. [28]. This process is constructed on global and portioned scheduling to evaluate real message sets and guides the design optimization. This gateway technique can remove the bottleneck of the message execution with the use of a small message execution delay, but its real-time process can be improved through multiple execution units. Alaei et al. [29] proposed a method for improving message response time by using a statistical based algorithm. In this paper, the stuff bits were reduced through the Statistical Mask Calculation (SMC) which provided better performance than the existing process. But the validation will require to improve in the reliability of the CAN network by minimizing bit stuffing.

The CAN is generally applied in various sectors like industrial, home automation, transportation, medical sector, and thermal plant, etc. that shows worst-case response time (WCRT) at the time of execution. This occurs due to delay in the periodic frame of the message, desynchronization of the message frames, improper scheduling of frames, etc., which is the main reason for WCRT. But in this current or proposed system is improvising the process response time of message transmission in various units of the system that makes the whole process full of automation. This paper provides a technique for analyzing the response time of CAN through an enhanced method of bit stuffing, message format, and error handling mechanism to optimal way for the handling of huge datasets. The existing process of the CAN bus is designed to measure lower and upper bound on the response time by using a cluster gateway algorithm which is performed their activity on a large-scale cluster of the message. But it sometimes does not give perfect responses due to unnecessary conversion in the process. That is why this proposed approach algorithm provides a perfect observation to finding the response time of optimized or prioritized message conversion in CAN bus that helps to avoid the cause of the collision.

## 3. Proposed System

An automation domain that is a versatile way to monitor the motor movement from a collision due to any obstacle is the objective of the proposed method. The primary goal of this phase is to ensure protection from unconditional road accidents due to traffic congestion and reckless driving.

*3.1. Serial Communication Protocols.* The growing demand for transmission of message has been controlled through the different protocols for communication, which is based on applications to build in a networked and internet-connected environment [30] [31] [32]. But these protocols vary from one another at the time of communication. So the upper and lower end protocols form the transmission procedures covering the various communication nodes of automobiles [33]. The device's CAN control bus and address bus are referred to as the higher end. Together with the increased focus on the distributed systems and networking, the cost benefit and advanced capabilities of silicon technology have led to the need for new highly organized communication methods in the area of field bus application. The automobile expects scalable control systems with a high

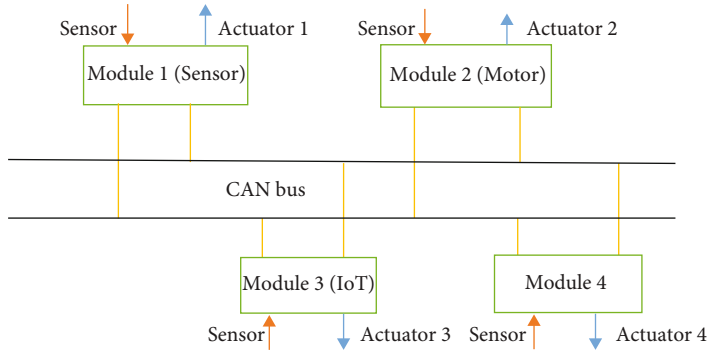



Figure 1: Illustration of a distributed CAN bus network.

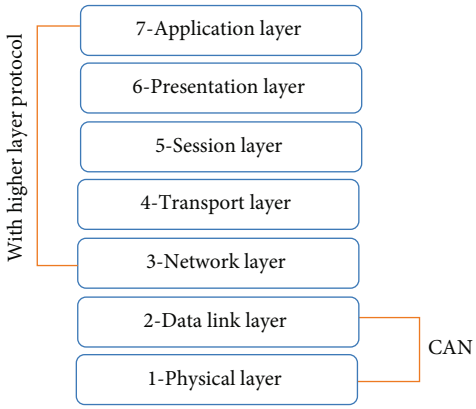



Figure 2: Controller Area Network (CAN) protocol defines in OSI model

standardization degree. The high level of standardization in the hardware and software modules leads to reusable systems that are ready to adapt in any single application setting to the various requirements and solutions [34]. The automotive vehicle has been developed by adding many electronic components with IoT modules for providing safety and to improve collision avoidance systems. Consequently, they need more and more hardwired, dedicated signal lines because of the complexity of the sharing data transmission control architecture of the system. This prompted the replacement by network architecture of the current wiring mode where the network system communicates via a common bus to all of the nodes. So this communication is ideally performed through CAN protocol which was developed by R. Bosch [35]. The CAN is used to communicate together in real time at speeding up to 1 Mbps, sensor and actuator via a two-wire serial data bus [36]. Focused on the concept of the "Shared Variables," the Virtual Levelled Systems Architecture (VLSA) model forms the generic interface architecture that is central to the CAN protocol [37]. Individual tasks are handled by distributed controllers in the VLSA architecture, with each one responsible for a portion of the total control programme. Through their sensors and actuators, nodes in a distributed system interact with the real process. The nodes use a dynamic, priority-based arbitration system to send messages on the bus. Figure 1 shows illustration of a distributed CAN bus network.

The nodes use a dynamic arbitration method based on priority to pass messages on the bus [38]. The nodes filter out the corresponding messages by filtering the message algorithm. Any message sent on the bus is delivered to every node in your network. Based on the message received, the application will send control signals to the device via the actuators. Jitter occurs when data packets are sent over your network connection with a temporal delay. Congestion on the network, as well as route changes, are common causes.

*3.1.1. CAN Protocol.* The CAN is described as two protocol standards such as ISO 11898 and ISO 11519 [39]. The ISO 11898 standard monitors high speed communication up to 1Mbps in physical layer of OSI model. The upper limit for ISO 11519 is 125 kbps that is consisting of a sub-layer of Logical Link Control (LLC) and a sub-layer of Media Access Control (MAC) in data link layer [40] [41]. Controller Area Network (CAN) Protocol defines in OSI model represents in Figure 2.

To keep data and monitor information, the Data Link Layer constructs data frames. Generally, some additional services such as detecting frames with bit stuffing and also used to re-transmit faulty data frames at the time of communication.

The CAN Physical Layer in one given network transmits data between different nodes; it decides the mode of transmission of signals and thus addresses issues such as the encryption, timing and synchronization of the communicated data signal [42]. With the implementation of the CAN protocol, the receivers of the sensor nodes data set will be transmitted to the control unit of the device, which normalizes the physical and data link layers of the OSI communication model for the automation domain, while the higher-layer protocols such as CAL/CAN Open and the CAN Kingdom, System Net, define the application layer [41]. The upper layer of ISO/OSI model's highest level, the application layer, communicates with an application program. The OSI layer closest to the end-user is the application layer.

*3.2. CAN Message Transfer.* The maximum load for utility is 94 bits and restricted format communications of varying but limited lengths are used by CAN. There is no particular address in the messages. Instead, it can be thought that the messages are addressed by four separate frame types for

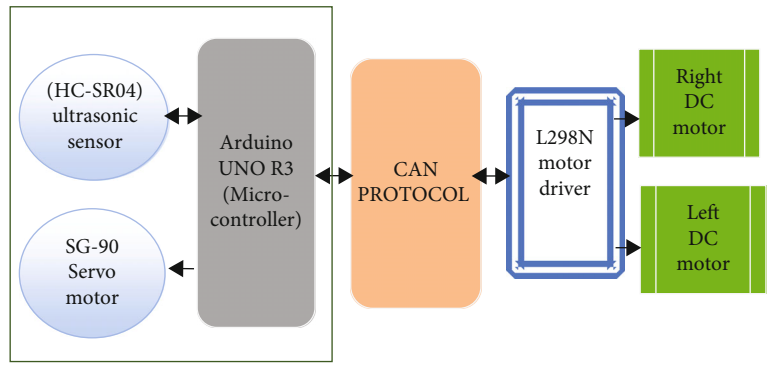



Figure 3: Proposed block diagram of smart autonomous collision avoidance system using the CAN protocol.

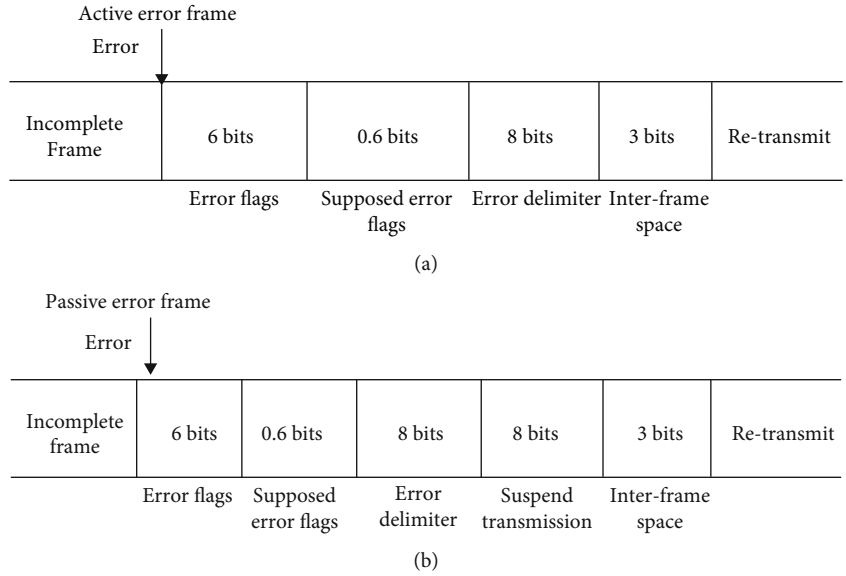



Figure 4: CAN error frame: (a) active error frame and (b) passive error frame.

1 : for$i = jdownto1$
2 : $\pi_{i,2\cdots n_i} = \pi_{i,1}$
3 : $RT_i = WCRT(\pi_{i,2\cdots n_i})$
4 : $while$(RT_i > DL_i)$do$
5 : $decrement\pi_{i,2\cdots n_i}$ (maximum priority)
6 : if$(\pi_{i,2\cdots n_i} < \pi_{i,1})then$
7 : $returnFail$
8 : $endif$
9 : $RT_i = WCRT(\pi_{i,2\cdots n_i})$
10 : $endwhile$
11 : $endfor$
12: Return success

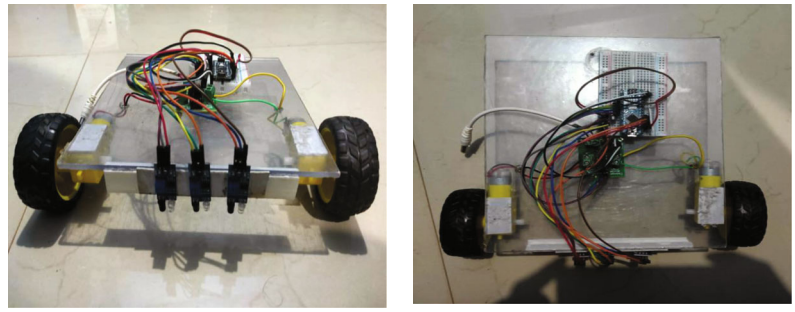
Algorithm 1: Message priority task algorithm.

communications such as a data frame, a remote frame, and an error frame for sending and reporting a detected data.

*3.2.1. Data Frame.* The CAN systems are used to transfer eight bytes of data frames with fixed data lengths through the network. Eight separate bit fields are composed of a message frame: frame start bit, data arbitration, control, data, CRC frame, acknowledgement, frame end field and Inter-Data space. So this protocol is defined by two frameworks base and extended format [43] [44].

The CAN 2.0A specifies base format CAN systems with standard 11-bit frame identifiers. But the CAN 2.0B identifies extended a format CAN system that has 29-bit frame identification. Where the CAN 2.0B supports both 11 bit and 29 bit identifiers, but the CAN 2.0A only supports 11-bit frame. The extended format is used on complex heavy traffic networks where the number of messages generated by network transmitters is greater than the number of possible CAN ID codes that may be given to them. The Standard CAN 11-bit ID provides the Extended CAN 29bit for 2, or 2047 separate message ID, whereas the CAN 29bit ID is stretched to provide 2 or 538 million identifiers [45].

So the vehicular conflicts may arise due to cross-wind, unbalanced friction coefficients, also a flat tire, the driver's behaviour is taken into account. The CAN bus is used by the Arduino UNO R3 module system as it relies on many IoT control units residing in the vehicle's Engine Control Unit (ECU- L298N Motor Driver) which depends significantly on the selection of the braking mechanisms (such as



(a) (b)

Figure 5: (a) Front and (b) top view of Smart Autonomous Collision Avoidance Vehicle.

Table 1: Details of Network Parameters.

| Network parameters | Value in numbers (nos) (variable message ID) | Value in numbers (nos) (variable message length) |
|---|---|---|
| Message length | 5-8 | 1-5000 |
| Message ID | 1-5000 | 1-500 |

hydraulic, pneumatic systems, electro-hydraulics, or even the electro-mechanics) and the usability of the Electronic Control Unit (ECU) as depicted in Figure 3 [46].

## 4. Working Principle and Methodology

*4.1. Measuring Response Time of CAN.* Measurement of the CAN message worst-case latencies for real-time analysis can be conducted on a fixed priority response time analysis scheduling standard [47]. The response time can be calculated using a worst-case message queuing configuration. The typical way to express the worst case action is to assume a collection of streams of traffic and also producing a fixed priority of queue messages on a periodic basis [48].

In the proposed model with message streams (MS) is processed in CPU scheduling with three elements of the messages $<J_i, Q_i, C_i>$, where $J_i$ is the queuing jitter, $Q_i$ is the queuing delay and $C_i$ is the communication delay of message i. When lower priority messages are forwarded it take a long time to be delayed in the queue $(Q_i)$. Then the real-time need to send the message by bus, due to communication delay $(C_i)$. The response time of worst-case error messages of the CAN bus is generally calculated, which shows the overhead error $EO_i$ in terms of E(t) denotes the maximum time required to signal and retrieve errors during the interval t. The response time analysis of worst-case (WR) can be determined by:

$$WR_i = J_i + Q_i + C_i,$$

$$Q_i = B_i + \sum_{j \epsilon hp(i)} \left\lceil \frac{(Q_i + J_j + \tau_{bit})}{T_j} \right\rceil C_j + E(Q_i + C_i),$$

$$WR_i = J_i + B_i + \sum_{j \epsilon hp(i)} \left\lceil \frac{(Q_i + J_j + \tau_{bit})}{T_j} \right\rceil C_j + P_k,$$

$$P_k = C_i + E(Q_i + C_i), B_i = \max_{\forall k \epsilon lm(i)} (C_k),$$

$$C_i = \left( mh + 8P_i + 15 + \left\lceil \frac{(mh + 8P_i - 1)}{4} \right\rceil \right) T_{bit},$$

$$EO_i = 15_t au_{bit} + \max_{K \epsilon hp(i) \cup \{i\}} (C_k + 31_t au_{bit}), \quad (1)$$

(i) Captures the effect of external interference as an error in many frames, rather than allowing the interference pattern to be defined and that explanation gives the consequence of message transmissions.

(ii) The period of interference does not reflect the potential delay, e.g. assuming that interference with duration $15_t au_{bit}$ will in the worst case give an error overhead.

(iii) Only allows relatively simple interference patterns with an initial burst and a residual error rate.

(iv) Does not conveniently capture interference from multiple sources.

*4.1.1. Features of CAN Error Handling.* Error Active is the default mode for a node. When any of the two Error Counters rises above 127, the node goes into Error Passive mode, and when the Transmit Error Counter rises beyond 255, the node goes into Bus Off mode. When an Error Active node identifies errors, it transmits Active Error Flags. The signal was scattered throughout propagation due to several nodes exchanging sample thresholds. The CAN bus can cause errors which are also used for error finding and auto-checking tools to attain resources of source-based controlling, bit stuffing, CRC bit checks, as well as testing format of the message frame [49]. So, it is depicted in Figures 4(a) and 4(b).

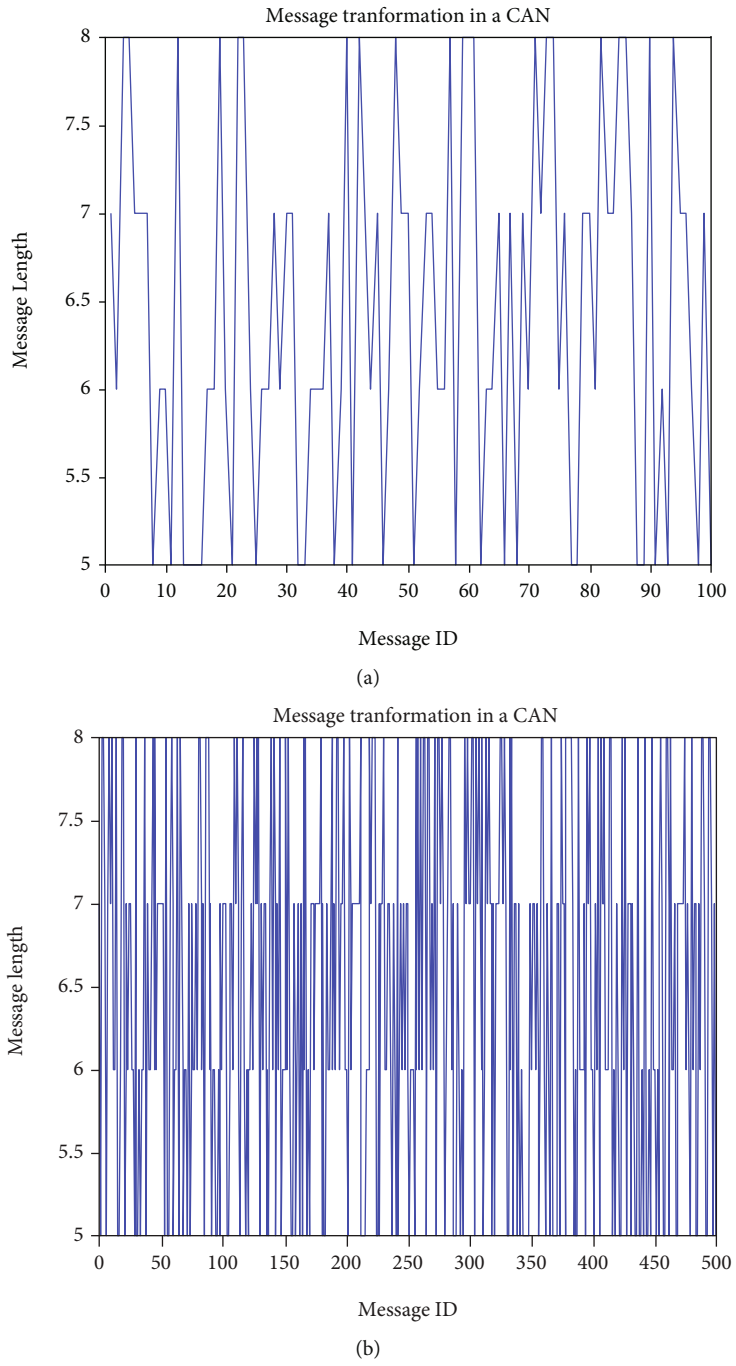



Figure 6: (a) Message length = 5 – 8 and ID = 100 nos and (b) message length = 5 – 8 and ID = 500 nos.

*4.2. Bit-Stuffing Effect and Retransmission of CAN Message Frame.* When a CAN node detects an error in a transmission, it sends an error flag consisting of six bits with the same polarity. The bit stuffing method prevents six consecutive bits from having the same polarity by adding a bit of opposite polarity after the fifth bit. If the number of bit-stuffing increases then the re-transmission of CAN messages can be increased. But the bit-stuffing decreases then the re-transmission of CAN messages can be decreased at the time of communication. The message bit pattern is a set of stuffed bits that requires probability distribution of each bit frame format [50]. So, the distribution of communication time can be collected from the number of stuff bits that is defined as

$$CD_m(t) = CD_m(t) + \phi(b)\tau_{bit} \tag{2}$$

At the time when the message communication is not successfully transmitted to the destination, it is due to the delays

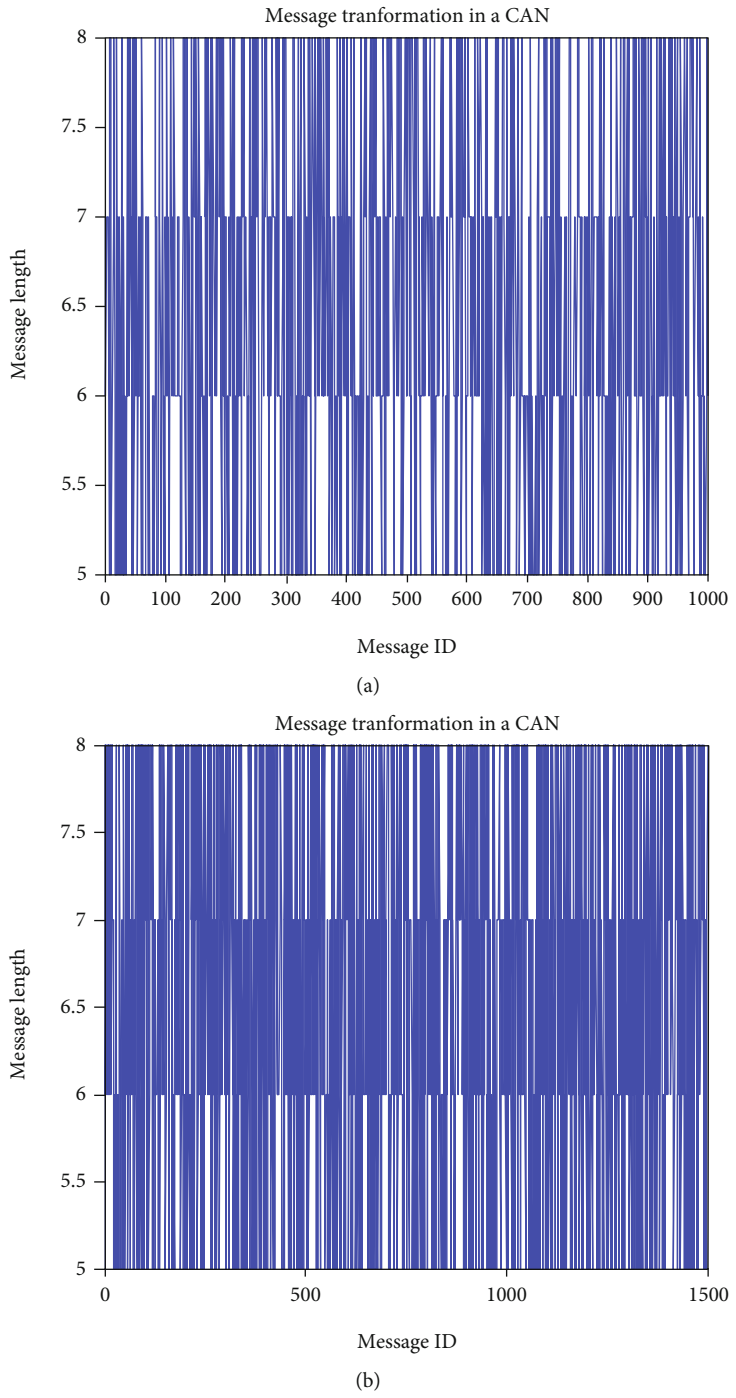



FIGURE 7: (a) Message length = 5 – 8 and ID = 1000 nos and (b) message length = 5 – 8 and ID = 1500 nos.

or occurrence of any noise. Due to that reason, the message may need to be re-transmitted which is denoted by $RT_i$ for message i. So, It can be expressed as:

$$RT_i = \lceil S_i * PR_i \rceil \tag{3}$$

Where the total frames set defines as $S_i$ of a message i, and the percentage of need for message re-transmission size can be defined by $PR_i$. In the case of non-complex data, $PR_i$=0 and for other types of data, it is expressed as $PR_i$>0. Then the worst-case communication time $(WC_i)$ can be calculated for message i without any error situation and can be expressed as

$$WC_i = S_i * p_k * \tau_{bit} \tag{4}$$

Here, when the bits are stuffed, then the worst case data packet size is denoted as $p_k$.

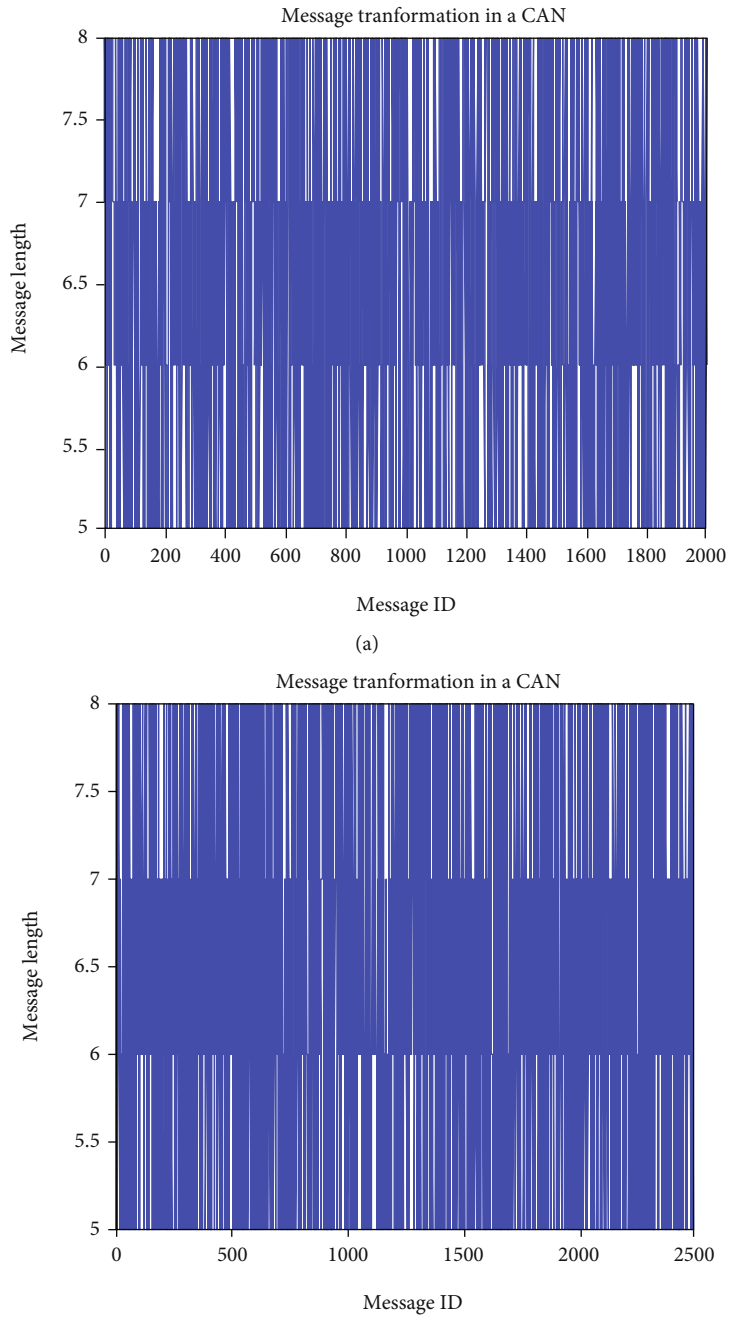



Figure 8: (a) Message length = 5 – 8 and ID = 2000 nos and (b) message length = 5 – 8 and ID = 2500 nos.

*4.3. Phase Communication Time and Optimal Action of Message Frame.* The instant jitter function can be expressed as $j_{i,m} = P_{i,m} - (m * TS_i + \phi_i)$, and according to that, system the jitter sum can be given as

$$J_S = \sum_{n=1}^{t} \sum_{n=1}^{t} j_{i,m} \tag{5}$$

where $j_{i,m}$ denotes the instant jitter of the $k^t h$ data frame of the sensor node m, the beginning time of the communication is denoted as $P_{i,m}$, $\phi_i$ is the start-up phasing of the system and the communication time interval of sensor node i is expressed as $TS_i$. So the alteration between the predicted starting and actual time of communication is expressed by the expression $P_{i,m} - (m * TS_i + \phi_i)$. The fitness can be calculated for optimal action of the system as

$$F(t) = J_S. \tag{6}$$

When a number of messages is used for transmission, then the crossover condition occurs and due to that,

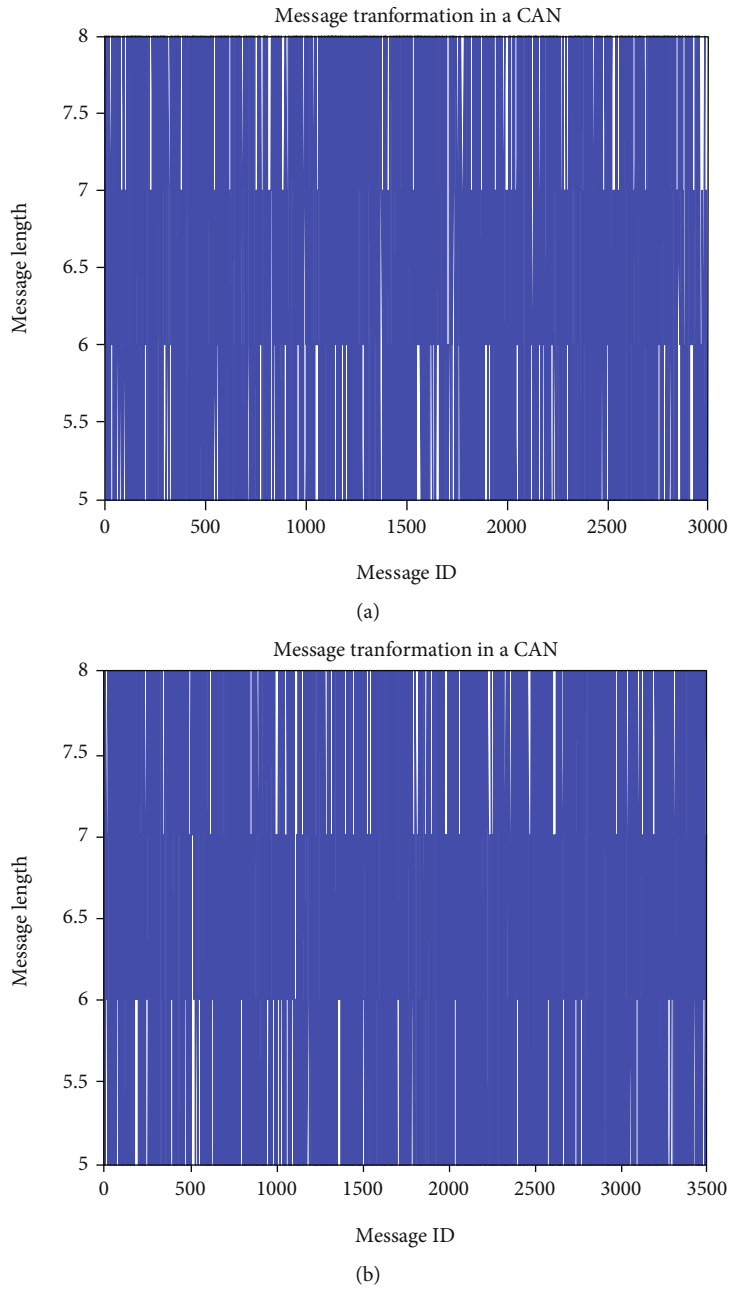



Figure 9: (a) Message length = 5 – 8 and ID = 3000 nos and (b) message length = 5 – 8 and ID = 3500 nos.

messages are queued according to their priorities. The successors are formed during processing by crossover action. It requires optimal scheduling; thus, we can choose a crossover probability of 1.0. Again, this process has improved the optimization by using efficient transformation probability as follows:

$$P_{trans} = \begin{cases} \dfrac{0.1(f_m - f)}{f_m - f_a}, & f \geq f_a, \\ 0.2, & f \leq f_a. \end{cases} \tag{7}$$

*4.4. Message Analysis in CAN Protocol.* The various level of ECUs is accessed in automotive applications to transfer signals as a form of message for steering of wheel speeds, gear selection and position of all controlled nodes of vehicles, measured through the CAN [47]. There can be more than 2500 separate signals in a high-end car, each essentially substituting an isolated connection in a conventional point-to-point connection unit [51].

These signals are used to read the location of a footbrake; when it is pressed, the back-light section can be finding changes in signal to on brake light to avoid collision by ECU of IoT section of the CAN bus. When the messages

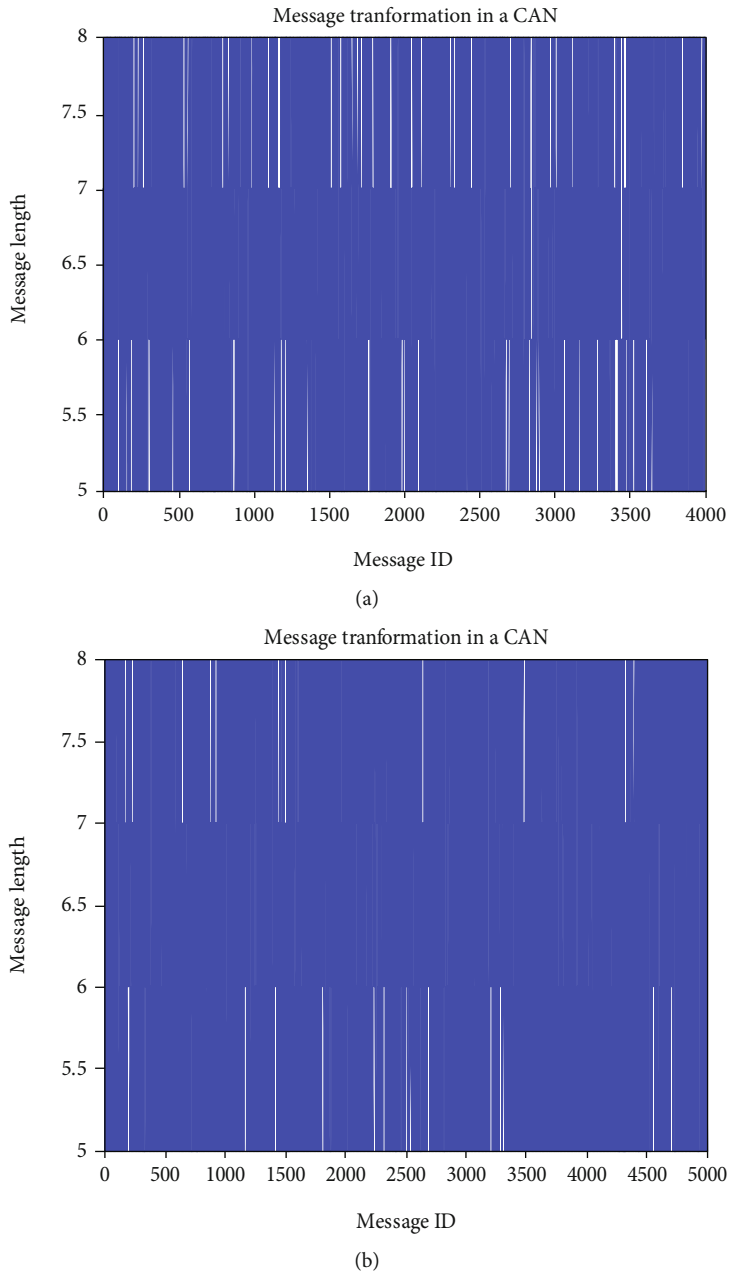



FIGURE 10: (a) Message length =5-8 and ID =4000 nos and (b) message length =5-8 and ID =5000 nos.

of CAN are linked with the deadline, then it responds within their time constraints as often as once every five milliseconds due to the stability and engine control system of an automobile.

*4.4.1. Message Formats.* The restricted data context is discussed in four message formats such as message frame, isolated node, error bit, and excess load bit [52]. A data frame starts with the begin-of-frame bit (BOF), 11-bit ID, and the distant transfer request (DTR) bit [53] [54]. The area of arbitration forms the ID and the DTR bit.

The control field contains six bits that also specify the length of bytes in the data field, which can range from 0 to 8 bytes. Whereas a CRC bit is used to verify whether the bit sequence has been modified or not in the data field. The transmitter uses the 2-bit acknowledgment field (ACK) to obtain correct frame recognition from of receiver. The end of a message frame signal is denoted in the 7-bit end of frame (EOF) which is expanded to a twenty-nine-bit ID recipient. A 21-bit extended database framework is also available.

(i) Response time calculations under normal case

(ii) Graph of maximum achievable utilization versus $T_f$

(iii) Graphs of response time of any message versus $T_f$

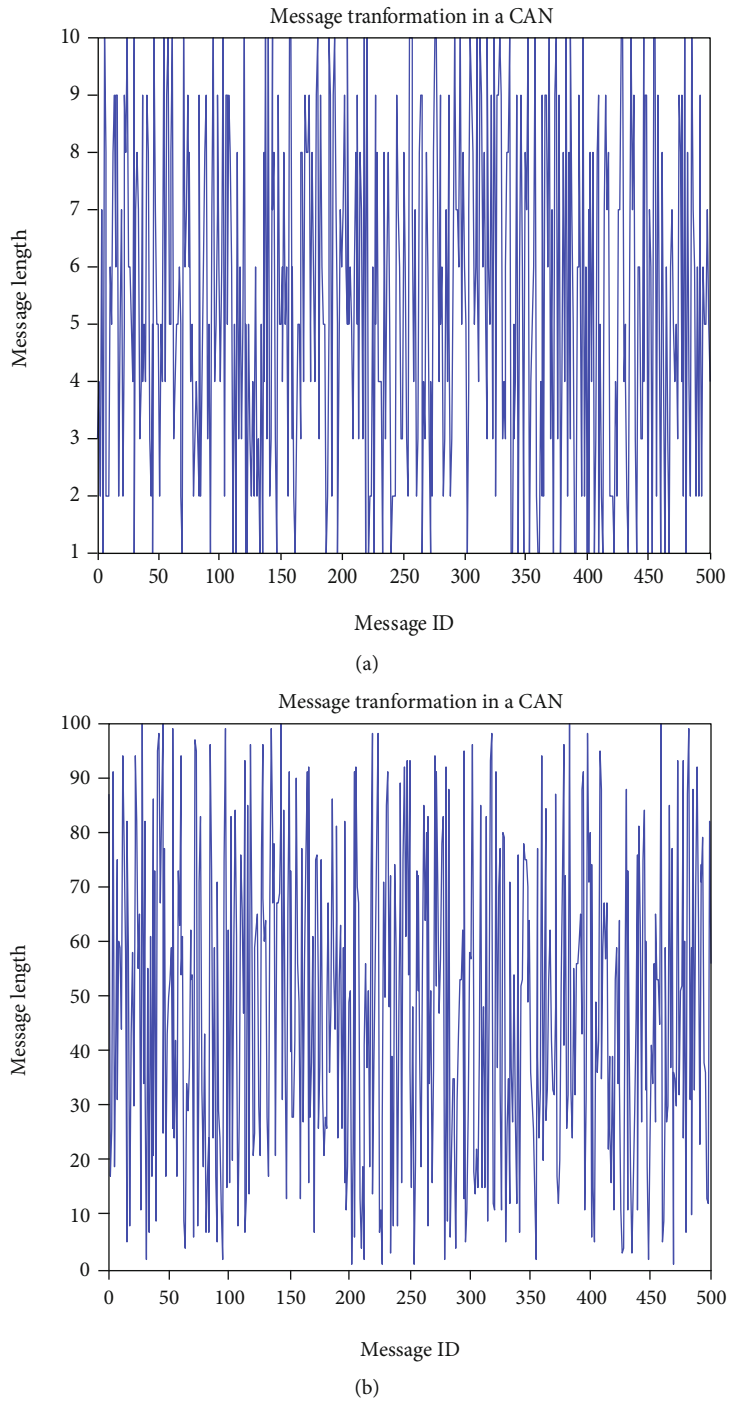



Figure 11: (a) Message length = 1-10 and ID =500 nos and (b) message length = 1-100 and ID =500 nos.

(iv) Worst case tolerable value for $T_f$

Suppose for every first frame, message priorities are already assigned; that is $\Pi_i$, 1 is assigned for each $1 \le i \le m$ to first frame. In the absence of general losses, it can be assumed that messages are organized accordingly: that is $i < j$ implies that $\pi_i, 1 < \pi_j, 1$. If all subsequent frames are assigned priority $\pi_i$,2 $n_i$, assume that WCRT(i, $\pi_i$,2$n_i$) will use the methods shown above to find the worst-case response time $RT_i$ designed for message i. In order to determine whether a feasible priority assignment exists, one can then use the algorithm given below in Algorithm 1.

This algorithm is optimal in that it always identifies one of the priorities for the two levels allocated. This algorithm starts with the lowest priority message frame, and then the worst time complexity is O(n2). The algorithm is proposed for analyzing the response time of the priority of messages from message ID and length in the CAN network. However, this algorithm evaluates response time based on message ID and length.

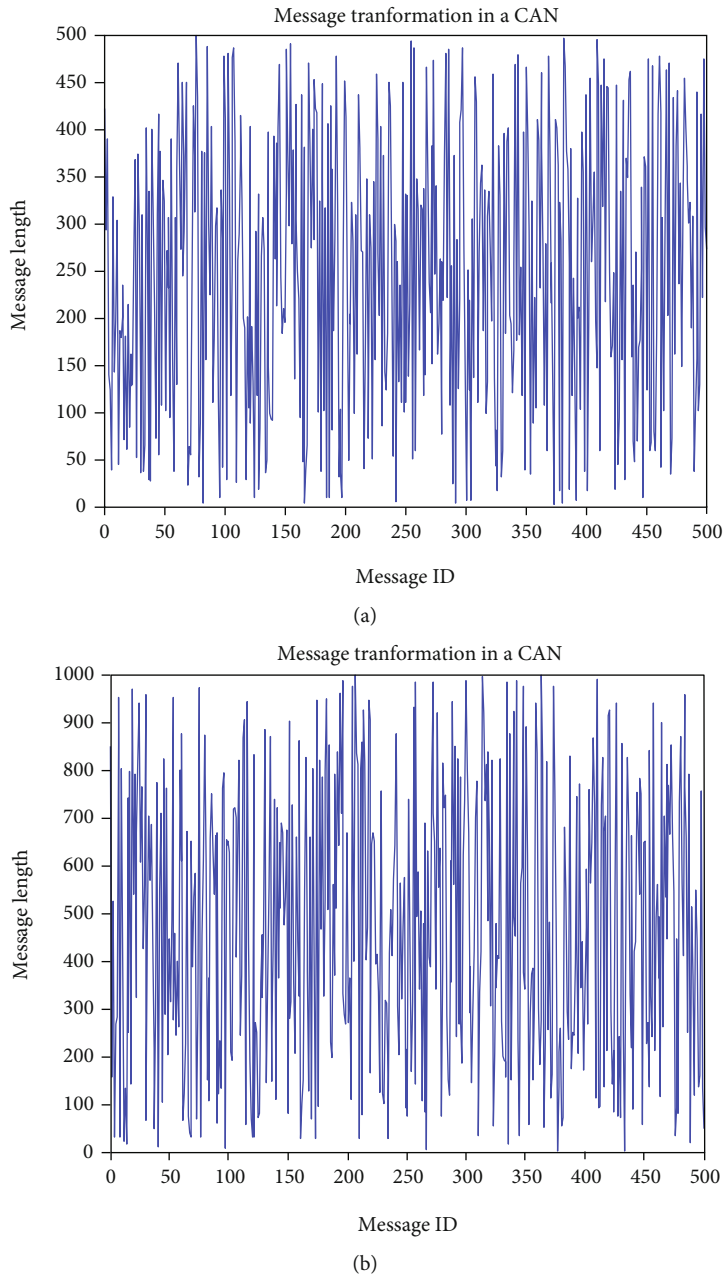



FIGURE 12: (a) Message length = 1-500 and ID =500 nos and (b) message length = 1-1000 and ID =500 nos.

*4.5. Components Description.* Different types of hardware components are required to design this proposed system:

*4.5.1. Arduino UNO R3 Controller.* It is an open-source and IDE microcontroller that controls every movement of sensor nodes and other system network nodes [55] [56]. The C or C ++ language are simply used for programming.

*4.5.2. HC-SR04 Ultrasonic Ranging Module.* The sensor module is typically used in the 2 cm-400 cm range to measure the distance of the obstacle [57]. Thus, the angle of 15 degrees with a voltage of 5 V dc is made.

*4.5.3. L298N Dual Bridge Motor Driver Module.* L298N is a driver circuit with two inputs that makes the system to be independently enabled or disabled and the motor movement can also be controlled [58] [59]. In this case, the pulse of PWM is used to set the service period for signalling.

*4.5.4. DC Motors.* DC motor is accompanied by the two 150 rpm DC motors, which needs 12 V of voltage and 1-2 Amp current to start moving from right to left.

*4.5.5. SG 90 Microservo Motor.* It is a very lightweight server motor with high strength, which can rotate easily about 180

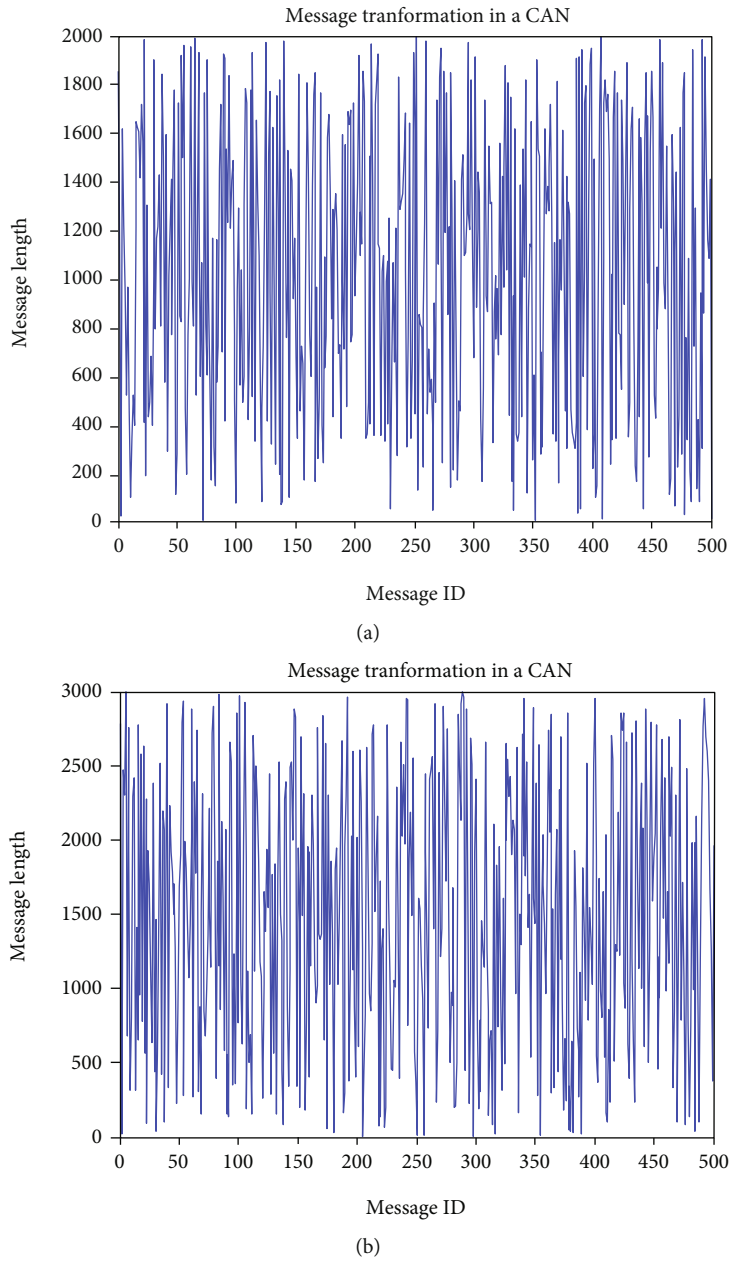



Figure 13: (a) Message length = 1-2000 and ID =500 nos and (b) message length = 1-3000 and ID =500 nos.

degrees (90 per path) [60] [61]. However, the movement is regulated through servo code, hardware and library.

## 5. Simulation Setup and Results Discussion

The intelligent self-employed vehicle is moving forward, which calculates the distance of an obstacle automatically. When an obstacle is detected within 20 cm via an ultrasonic sensor, the message frames obtained are transmitted to the controller. The CAN protocol code is received by the controller, which instructs the command to regulate motor movement from left to right and back. The collision avoidance algorithm was successfully implemented in order to reduce the problem. The front and top view of the Smart Autonomous Collision Avoidance Vehicle is shown in Figures 5(a) and 5(b).

*5.1. Response Time Analysis.* One of the types of field bus control devices used in networking is CAN. It is a protocol system based on a packet. Communication can be accomplished using the CAN protocol between different devices. The CAN bus is used to control the unit of transmission and receiving unit, which is mainly implemented due to low costs. The CAN multi-master node cannot simultaneously be transmitting and accepting messages that consist of a message ID as well as the message frame is communicated consecutively to the bus.

The CAN carriers detect multiple access protocols with collision detection and message priority arbitration, and

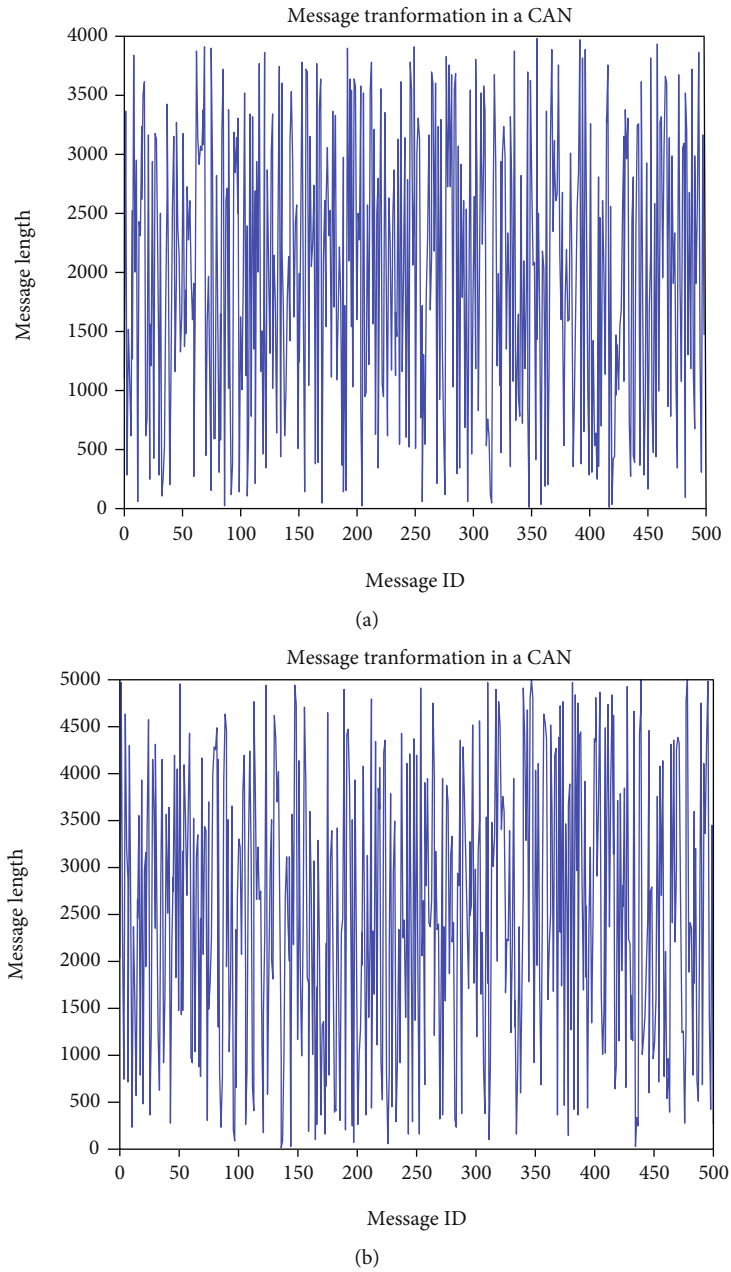



Figure 14: (a) Message length = 1-4000 and ID =500 nos and (b) message length = 1-5000 and ID =500 nos.

there are two types of protocols used in it. For flow control, the technique is used to confirm sensor data integrity, the cyclic redundancy checks (CRC) for an error control mechanism, which also manages the remote frames and the overload frames. CAN-based Database file (.dbc) is taken from the website CAN RT.dbc. It includes numerous attributes such as Name, Statement, ID, Duration, Signals, and Extended. We just take value for message length and message ID in this document which is shown in Table 1. So, these parameters observe the response of messages in the field of communication within the CAN network. After checking various data numbers, we get the following results, which are shown below. Figure 6 shows (a) message length = 5 – 8 and ID = 100 nos and (b) message length = 5 – 8 and ID = 500 nos.

*5.1.1. Result for Constant Message Length and Variable Message ID.* Figure 7 shows (a) message length = 5 – 8 and ID = 1000 nos and (b) message length = 5 – 8 and ID = 1500 nos. Figure 8 shows (a) message length = 5 – 8 and ID = 2000 nos and (b) message length = 5 – 8 and ID = 2500 nos. Figure 9 shows (a) message length = 5 – 8 and ID = 3000 nos and (b) message length = 5 – 8 and ID = 3500 nos. Figure 10 shows (a) message length = 5 – 8 and ID = 4000 nos and (b) message length = 5 – 8 and ID = 5000 nos. In the simulation result, we got some figures which represent

different outcome from the use of CAN RT.dbc. Figures 6–10 represent that the message transformation is performed in the CAN network of two classes of inputs generated by taking fixed values for message length 5-8 and random values for message ID like 100 nos, 500 nos, 1000 nos, 1500 nos, 2000 nos, 2500 nos, 3000 nos, 3500 nos, 4000 nos, 5000 nos, respectively. When the simulation of the above network is performed by changing of message ID parameter, which shows response time (located through the white line) that varies through the process. The above analysis shows that the response time is more delayed by increasing message ID numbers. Similarly, the message is highly responded to at fewer messages ID which is shown in the above figures.

*5.1.2. Result for Constant Message ID and Variable Message Length.* We have some figures in the simulation outcome that show different outcomes from the use of CAN RT.dbc. In the CAN network, figures reflect the message conversion of two groups of inputs created by taking fixed values for message ID 500 numbers and random values for message length such as 1-10, 1-100, 1-500, 1-1000, 1-2000, 1-3000, 1-4000, 1-5000. When this network simulation is performed by changing the message length parameter with fixed message length and it shows no more changes in the response time of the process in the above figures. Figure 11 shows (a) message length = 1 − 10 and ID = 500 nos and (b) message length = 1 − 100 and ID = 500 nos. Figure 12 shows (a) message length = 1 − 500 and ID = 500 nos and (b) message length = 1 − 1000 and ID = 500 nos. Figure 13 shows (a) message length = 1 − 2000 and ID = 500 nos and (b) message length = 1 − 3000 and ID = 500 nos. Figure 14 shows (a) message length = 1 − 4000 and ID = 500 nos and (b) message length = 1 − 5000 and ID = 500 nos.

*5.2. Discussion.* In a practical scenario, several attributes such as Name, Statement, ID, Extended, Length, Signals have to be considered from the collected database. If all these attributes are combined, we can get an unschedulable framework in certain situations. Therefore, some of the attributes already in the database, such as Name, Statement, Expanded, and Signals, need to be removed. For communication, some powerful attributes such as ID and length have to be considered for a particular message.

We might easily get a very negative analysis to find errors if there are many such data and we compose them. We have shown from the simulation outcome that there was no difference in the simulation outcome for different message length values by holding the message ID unchanged. In other words, we can assume that message length variance has less influence than message ID. We take the message ID from 1 to 5000 in this paper and check the network conjunction result. We also shift the message's duration from 1 to 5000 and examine the transformation impact of this change in values.

The existing approaches tested in a limited range of message length and message ID values which is not provided a clear idea about the performance of the CAN network. But this paper is taking 1-5000 numbers values for message length and message ID. According to the proposed experiment is evaluating performance in two ways like (i) constant message length with variance number of values (1-5000) for message ID and (ii) constant message ID with variance number of values (1-5000) for message length. Thus the performance is varied on the case of constant message length with variance number of values (1-5000) for message ID, but the consistent performance is evaluated on the case of constant message ID with variance number of values (1-5000) for message length.

## 6. Conclusion and Future Scope

While several solution algorithms and concepts have been developed over several years to solve conflict issues using CAN communication on the vehicle network, there have been few attempts to develop a solution to the handling of errors. This is a comparison-based analysis of the variable message ID and constant message length figures that we concluded that if we send no more messages at a time, then there would be more conflict on the network. The recipient will not get the real message due to conflict, and there will also be a risk of receiving more than one message at a single node. The CAN protocol is used to provide a secure and robust serial communication bus from sensor nodes to the control unit of an automated system. When the sensor node of the IoT module is received, a message frame can be transmitted to the destination node that can be responded to in time. The phase of communication time and the optimal action of the message frame are utilised to build a flexible format for transmission of a frame from sender end to receiver end, which implies that a system node can receive a message frame and respond to it via an acknowledged frame bit. The proposed scheme achieves high precision, determining the location of an obstacle and then monitoring the impact of collision time.

More number of experiments could be carried out and future directions are:

(a) To assess the efficiency of the algorithms, a large number of experiments with more tasks have to be tested

(b) Secondly, it is important to evaluate large-size problem cases using periodic preemptive tasks

A new automatically moving algorithm between the EDF algorithm and the ACO scheduling algorithm should be developed in the future to work with overloaded conditions.

## Data Availability

The IoT data that support the findings of this study are available on request from the corresponding author.

## Conflicts of Interest

The authors of this manuscript declared that they do not have any conflict of interest.

## Acknowledgments

This paper was partially supported by the Ministry of Science and Higher Education of the Russian Federation (Assignment No. 075-02-2021-1748) and Device Development Programme (DST/TDT/DDP-38/2021), by the Department of Science Technology, Ministry of Science and Technology, Government of India.

## References

[1] G. Li, W. Lai, X. Sui et al., "Influence of traffic congestion on driver behavior in post-congestion driving," *Accident Analysis & Prevention*, vol. 141, p. 105508, 2020.

[2] P. A. Hancock, T. Kajaks, J. K. Caird et al., "Challenges to human drivers in increasingly automated vehicles," *Human Factors*, vol. 62, no. 2, pp. 310–328, 2020.

[3] U. Kiencke, S. Dais, and M. Litschel, "Automotive serial controller area network," *SAE Transactions*, vol. 95, pp. 823–828, 1986.

[4] Y.-J. Kim, H.-Y. Lee, and J.-G. Chung, "4-bit data arrangement algorithm for can compression," in *2018 International SoC Design Conference (ISOCC)*, pp. 216-217, Daegu, Korea (South), 2018.

[5] S. Dekanic, R. Grbic, T. Maruna, and I. Kolak, "Integration of can bus drivers and uds on aurix platform," in *2018 Zooming Innovation in Consumer Technologies Conference (ZINC)*, pp. 39–42, Novi Sad, Serbia, 2018.

[6] I. Gonzalez and A. J. Calderon, "Integration of open source hardware arduino platform in automation systems applied to smart grids/micro-grids," *Sustainable Energy Technologies and Assessments*, vol. 36, article 100557, 2019.

[7] G. Kornaros, O. Tomoutzoglou, D. Mbakoyiannis et al., "Towards holistic secure networking in connected vehicles through securing CAN- bus communication and firmware-over-the-air updating," *Journal of Systems Architecture*, vol. 109, p. 101761, 2020.

[8] D.-S. Kim and H. Tran-Dang, "Industrial sensors and controls in communication networks," in *Computer Communications and Networks*, Springer International Publishing, Cham, 2019.

[9] H. Lu, Q. Liu, D. Tian, Y. Li, H. Kim, and S. Serikawa, "The cognitive internet of vehicles for autonomous driving," *IEEE Network*, vol. 33, no. 3, pp. 65–73, 2019.

[10] L. Chelouah, F. Semchedine, and L. Bouallouche-Medjkoune, "Localization protocols for mobile wireless sensor networks: a survey," *Computers & Electrical Engineering*, vol. 71, pp. 733–751, 2018.

[11] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[12] N. Deepa, Q.-V. Pham, D. C. Nguyen et al., "A survey on blockchain for big data: Approaches, opportunities, and future directions," 2020, https://arxiv.org/abs/2009.00858.

[13] A. L. Kun, S. Boll, and A. Schmidt, "Shifting gears: user interfaces in the age of autonomous driving," *IEEE Pervasive Computing*, vol. 15, no. 1, pp. 32–38, 2016.

[14] J. Sanchez-Garcıa, J. Garcıa-Campos, M. Arzamendia, D. G. Reina, S. Toral, and D. Gregor, "A survey on unmanned aerial and aquatic vehicle multi-hop networks: Wireless communications, evaluation tools and applications," *Computer Communications*, vol. 119, pp. 43–65, 2018.

[15] A. K. Biswal, D. Singh, B. K. Pattanayak, D. Samanta, and M.-H. Yang, "Iot-based smart alert system for drowsy driver detection," *Wireless Communications and Mobile Computing*, vol. 2021, 13 pages, 2021.

[16] IndustryJournalPro, "Forward collision avoidance radar market research report 2019-2025," 2019, 2021, https://industryjournalpro.com/forward-collision-avoidance-radar-market-research-report-2019-2025/.

[17] Hyundai, "12 Hyundai safety features that make driving today safer than ever," 2020, 2021, https://www.hyundai.com/au/en/why-hyundai/myhyundai-news/issues/2020/issue-02/12-hyundai-safetyfeatures-that-make-driving-today-safer-than-ever.

[18] T. Nolte, M. Nolin, and H. A. Hansson, "Real-time server-based communication with can," *IEEE Transactions on Industrial Informatics*, vol. 1, no. 3, pp. 192–201, 2005.

[19] W. L. Ng, C. K. Ng, B. M. Ali, N. K. Noordin, and F. Z. Rokhani, "Review of researches in controller area networks evolution and applications," *Proceedings of the Asia-Pacific Advanced Network*, vol. 30, pp. 14–21, 2013.

[20] A. S. Shinde and V. B. Dharmadhikari, *Controller Area Network for Vehicle Automation*, CiteSeer, 2012.

[21] G. Yu, C. Zhou, and S. Huang, "A protocol for automatic node-id binding in canopen networks," *JCM*, vol. 7, no. 10, pp. 765–773, 2012.

[22] T. Vijayan, *Controller Area Network in Modern Home Automation*, CiteSeer, 2012.

[23] C.-L. Wey, C.-H. Hsu, K.-C. Chang, P.-C. Jui, and M.-T. Shiue, "Emi prevention of can-bus-based communication in battery management systems," *International Journal of Engineering & Computer Science IJECS-IJENS*, vol. 13, no. 5, pp. 6–12, 2013.

[24] K. C. Emani, K. Kam, M. Zawodniok, Y. R. Zheng, and J. Sarangapani, "Improvement of can bus performance by using error-correction codes," in *2007 IEEE Region 5 Technical Conference*, pp. 205–210, Fayetteville, AR, USA, 2007.

[25] H. A. Hansson, T. Nolte, C. Norstrom, and S. Punnekkat, "Integrating reliability and timing analysis of can-based systems," *IEEE Transactions on Industrial Electronics*, vol. 49, no. 6, pp. 1240–1250, 2002.

[26] R. I. Davis, A. Burns, R. J. Bril, and J. J. Lukkien, "Controller area network (can) schedulability analysis: refuted, revisited and revised," *Real-Time Systems*, vol. 35, no. 3, pp. 239–272, 2007.

[27] G. Xie, G. Zeng, R. Kurachi, H. Takada, R. Li, and K. Li, "Exploiting Sparsity to Accelerate Fully Connected Layers of CNN-Based Applications on Mobile SoCs," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 17, no. 2, pp. 1–25, 2018.

[28] G. Xie, G. Zeng, R. Kurachi et al., "Wcrt analysis and evaluation for sporadic message-processing tasks in multicore automotive gateways," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 38, no. 2, pp. 281–294, 2019.

[29] R. Alaei, P. Moallem, and A. Bohlooli, "Statistical based algorithm for reducing bit stuffing in the controller area networks," *Microelectronics Journal*, vol. 101, p. 104794, 2020.

[30] K. Liu, E. Fridman, and L. Hetel, "Networked control systems in the presence of scheduling protocols and communication delays," *SIAM Journal on Control and Optimization*, vol. 53, no. 4, pp. 1768–1788, 2015.

[31] B. H. C. Orak, F. Y. Okay, M. Guzel, S. Murt, and S. Ozdemir, "Comparative analysis of iot communication protocols," in *2018 International symposium on networks, computers and communications (ISNCC)*, pp. 1–6, Rome, Italy, 2018.

[32] K. T. Nguyen, M. Laurent, and N. Oualha, "Survey on secure communication protocols for the internet of things," *Ad Hoc Networks*, vol. 32, pp. 17–31, 2015.

[33] S. Krishnamoorthy, *Design of an ASIC chip for a Controller Area Network (CAN) protocol controller, [Ph.D. thesis]*, Texas Tech University, 2006.

[34] J. Long, "Automobile electronic control network design based on can bus," in *2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pp. 9–12, Xiamen, China, 2018.

[35] K. Pazul, *Controller area network (can) basics*, Microchip Technology Inc, 1999.

[36] K. H. Johansson, M. Torngren, and L. Nielsen, "Vehicle applications of controller area network," in *Handbook of networked and embedded control systems*, Springer, 2005.

[37] W. E. Lawrenz, *System design tools for networked systems in cars*, SAE Technical Paper, 1990.

[38] Q. Zhu, Z. Dongmei, and S. Xunwen, "Distributed remote temperature monitoring and acquisition system based on can bus," in *2010 Prognostics and System Health Management Conference*, pp. 1–4, Macao, China, 2010.

[39] SC HPL, "Introduction to the controller area network (can)," in *Application Report SLOA101*, Texas Instruments, 2002.

[40] P. Santi, "Topology control in wireless ad hoc and sensor networks," *ACM Computing Surveys (CSUR)*, vol. 37, no. 2, pp. 164–194, 2005.

[41] M. A. C. Aung and K. P. Thant, "Detection and mitigation of wireless link layer attacks," in *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, pp. 173–178, London, UK, 2017.

[42] Y. Liu, Q. Zhang, and L. Ni, "Opportunity-based topology control in wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 3, pp. 405–416, 2010.

[43] S. Hasnaoui, O. Kallel, R. Kbaier, and S. B. Ahmed, "An implementation of a proposed modification of can protocol on can fieldbus controller component for supporting a dynamic priority policy," in *38th IAS Annual Meeting on Conference Record of the Industry Applications Conference, 2003*, pp. 23–31, Salt Lake City, UT, USA, 2003.

[44] S. Hong and W.-H. Kim, "Bandwidth allocation scheme in can protocol," *IEE Proceedings-Control Theory and Applications*, vol. 147, no. 1, pp. 37–44, 2000.

[45] X. Wang and W. Guo, "The design of rs232 and can protocol converter based on pic mcu," *Computer and Information Science*, vol. 2, no. 3, pp. 176–181, 2009.

[46] Q. Lin, Y. Zhang, A. Van Mieghem et al., "Design and experiment of a sun-powered smart building envelope with automatic control," *Energy and Buildings*, vol. 223, article 110173, 2020.

[47] G. I. Mary, Z. C. Alex, and L. Jenkins, "Response time analysis of messages in controller area network: a review," *Journal of Computer Networks and Communications*, vol. 2013, Article ID 148015, 11 pages, 2013.

[48] Y. Wang and M. Saksena, "Scheduling fixed-priority tasks with preemption threshold," in *Proceedings Sixth International Conference on Real-Time Computing Systems and Applications. RTCSA'99 (Cat. No. PR00306)*, pp. 328–335, Hongkong, 1999.

[49] J. Van Waes, J. Lannoo, A. Degraeve, D. Vanoost, D. Pissoort, and J. Boydens, "Effectiveness of cyclic redundancy checks under harsh electromagnetic disturbances," in *2017 International Symposium on Electromagnetic Compatibility-EMC EUROPE*, pp. 1–6, Angers, France, 2017.

[50] G. Leen and D. Heffernan, "Ttcan: a new time-triggered controller area network," *Microprocessors and Microsystems*, vol. 26, no. 2, pp. 77–94, 2002.

[51] K. Tindell, A. Burns, and A. J. Wellings, "Calculating controller area network (can) message response times," *Control Engineering Practice*, vol. 3, no. 8, pp. 1163–1169, 1995.

[52] R. Sato and S. Fukumoto, "Response-time analysis for controller area networks with randomly occurring messages," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 3893–3902, 2020.

[53] S. Mubeen, J. Maki-Turja, and M. Sjodin, "Extending Worst Case Response-Time Analysis for Mixed Messages in Controller Area Network With Priority and FIFO Queues," *IEEE Access*, vol. 2, pp. 365–380, 2014.

[54] G. Cena, I. C. Bertolotti, T. Hu, and A. Valenzano, "Can with extensible in-frame reply: protocol definition and prototype implementation," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2436–2446, 2017.

[55] N. S. Kumar, B. Vuayalakshmi, R. J. Prarthana, and A. Shankar, "Iot based smart garbage alert system using arduino uno," in *2016 IEEE Region 10 Conference (TENCON)*, pp. 1028–1034, Singapore, 2016.

[56] A. K. Biswal, D. Singh, and B. K. Pattanayak, "Iot-based voice-controlled energy-efficient intelligent traffic and street light monitoring system," in *Green Technology for Smart City and Society*, Springer, 2021.

[57] V. Zhmud, N. Kondratiev, K. Kuznetsov, V. Trubin, and L. Dimitrov, "Application of ultrasonic sensor for measuring distances in robotics," *Journal of Physics: Conference Series*, vol. 1015, article 032189, 2018.

[58] M. H. Y. Yumei, "High-precision servo controller of dc micromotor based on dspic30f4011," *Electronic Measurement Technology*, vol. 10, 2010.

[59] M. M. Maung, M. M. Latt, and C. M. Nwe, "Dc motor angular position control using pid controller with friction compensation," *International Journal of Scientific and Research Publications*, vol. 8, no. 11, p. 149, 2018.

[60] S. K. Mostaque and B. Karmakar, "Low cost arduino based voice controlled pick and drop service with movable robotic arm," *European Journal of Engineering and Technology Research*, vol. 1, no. 5, pp. 29–33, 2018.

[61] K. Neeraja, P. R. C. Rao, D. S. Maloji, and D. M. A. Hussain, "Implementation of security system for bank using open cv and rfid," *International Journal of Engineering & Technology*, vol. 7, no. 2-7, p. 187, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3123615, 14 pages
https://doi.org/10.1155/2022/3123615

*Research Article*

# MR-Pareto: A Multiattribute Opportunistic Routing Method Based on Pareto Optimal Solution for Mobile Crowdsensing

**Xiao Han, Huiqiang Wang, Jing Tan, Hongwu Lv, and Chengbo Wang**

*College of Computer Science and Technology, Harbin Engineering University, 115001 Harbin, China*

Correspondence should be addressed to Huiqiang Wang; wanghuiqiang@hrbeu.edu.cn

Received 12 August 2021; Revised 5 January 2022; Accepted 25 January 2022; Published 15 March 2022

Academic Editor: Rajesh Kaluri



Mobile crowdsensing (MCS) is a new perception mode for solving large-scale mobile sensing tasks. Traditional data transmission methods are inapplicable, as the MCS is affected by coverage, user preference, and network access cost. Opportunistic network data transmission schemes in MCS have recently witnessed a surge of research efforts due to their ability of high delivery and low consumption. However, existing works only focus on the impact of the geographical location of nodes on user needs or the interaction between social information and data, which do not take into account the temporal and spatial characteristics of nodes. To address these issues, this paper proposes a multiattribute routing method based on Pareto optimal (MR-Pareto) solution to construct a balance between the energy consumption and resource constraints of nodes in transmission protocols. First, the attribute relationship between nodes is analyzed, which was aimed at selecting the nodes within a contact time threshold. Then, based on a nondominated sorting approach, we achieve a Pareto optimal set of candidate nodes. Finally, the relay nodes for forwarding messages are determined by comparing the cache size and the remaining energy. The experimental results demonstrate that our proposed method has low network overhead, low packet loss, and high message delivery rate, compared to epidemic and prophet routing strategies.

## 1. Introduction

With the development of pervasive computing, mobile crowdsensing technology, and intelligent terminal devices, intelligent systems with integrated sensing and computing communication capabilities have been widely deployed. MCS migrates perception tasks from a centralized platform to a distributed computing terminal across the space-time dimension, which can not only achieve data analysis and understanding but also make large-scale and high-precision environment perception possible [1, 2]. Nowadays, MCS has attracted great attentions from various research institutions [3].

In traditional data transmission methods, it is difficult to satisfy the actual needs of data acquisition by only using predeployed dense sensor nodes in an uncertain and large-scale sensing environment, such as limited network resources and heterogeneous sensing terminals. MCS can use ubiquitous intelligent terminals and networks (i.e., WiFi, 4G, and 5G) cooperatively to improve the sensing quality, which is especially suitable for the environments with low network coverage or expensive access network (e.g., remote areas, large cities, and disaster recovery). The performance of MCS routing method plays an important role in sensing the task transmission quality. MCS has the characteristics of sparse node distribution, intermittent connection, and dynamic topology change, which is similar to the opportunistic routing. Therefore, data forwarding through opportunistic routing transmission mode can leverage the advantages of MCS [4]. The advantages are as follows: (1) opportunistic routing method can reduce the cost of network deployment, make full use of millions of mobile devices to build large-scale sensing networks, and ensure the privacy of user data. (2) It does not need any centralized servers or infrastructures for communication and management. Through the opportunity contact between mobile users, the mode of "store-carry-

forward" [5] is adopted to transmit the sensing data, which can reduce the workload of cellular network in dense areas and make the maintenance of network easier.

Up to now, the research on MCS opportunistic routing methods has mainly focused on the following aspects: (1) node location, which influences the impact of user demand; (2) data interaction, which determines how to select a suitable set of users; (3) balance between node energy and cache. Although the aforementioned aspects will affect the routing transmission performance of the MCS, the routing algorithm that comprehensively considers them is rare.

There are aspects that have an impact on the routing transmission performance of MCS. However, it is rare to consider all three aspects together.

Under the environment of weak network and limited resources, we propose a multiattribute opportunistic routing method for MCS that can achieve efficient node transmission by selecting the appropriate relay nodes and node local resource information (i.e., node motion state, node history connection, node energy, and capacity) and reducing network energy consumption and overhead. We considered energy consumption, cache level, and internode relationship as combined parameters in the routing process, instead of considering them separately as in literature [6, 7]. In addition, we added spatiotemporal characteristics to further constrain the nodes' forwarding decisions.

Our major contributions can be summarized as follows.

(1) We proposed a multiattribute routing method based on Pareto optimal (MR-Pareto) and evaluated the stability of node connection based on the user motion state. The energy and cache attributes of user terminals are introduced to improve the data delivery rate and reduce the communication load

(2) By analyzing the impact of different indices on candidate nodes in MCS, we noticed that the optimal candidate nodes are related to node centrality, correlation, and similarity. Through these three indices, we achieved the Pareto optimal result by using nondominated sorting

(3) Acceptance rates have historically run at slightly over 50%. There is no sufficient room within the technical program to accept all submissions

The rest of the paper is organized as follows: Section 2 discusses the related work. Section 3 introduces the detection of node location in the network. The opportunistic routing method based on Pareto optimal solution is presented in Section 4. In Section 5, we evaluate our proposed method and present evaluation results. Finally, we conclude this paper in Section 6.

## 2. Related Work

In the last decade, many opportunistic routing schemes have been designed to facilitate process of data routing. Generally, opportunistic routing schemes can be classified into geographic schemes, link state aware schemes, and probability schemes. The first category covers schemes based on locations. The second category covers schemes that take link states and nodes' energy into account. In the third category, probability is used to tackle the problems of node mobility.

Geographic opportunistic routing can overcome the issue of lack of infrastructure due to the dynamic property of MCS. A number of geographic metrics have been studied [8]. For instance, LIPS is a geographic scheme where the complication caused by the multiturning point structure can be overcame by the virtual plane mirror algorithm [9]. To handle asymmetric links, FQ-AGO [10] utilizes a fuzzy logic approach and employs the Q-learning algorithm to select the most stable link. MGOR uses a multiple channel to improve routing efficiency and take opportunistic effective one-hop throughput as a new local metric to solve the impacts of multiple rates, as well as candidate selection, prioritization, and coordination. To reduce energy cost of multipath routing approach, EQGOR [11] selects and prioritizes the forwarding candidate set in an efficient manner, which is suitable for WSNs in respect of energy efficiency, latency, and time complexity.

Opportunistic routing improves the reliability of packet delivery by broadcasting of the wireless link. Choosing the next relay at each hop based on the link state, opportunistic routing guarantees the packet delivery ratio. In [12], Li et al. propose the probability prediction-based reliable and efficient opportunistic routing (PRO) algorithm. Based on PRO, the variation of signal-to-interference-plus-noise ratio (SINR) and packet queue length (PQL) of the receiver can be predicted. In [13], Zhang et al. propose a link availability probability prediction model and a new concept called the link correlation which is used to represent the influence of different link combinations. Based on these conclusions, a street-centric opportunistic routing protocol which is based on the expected transmission cost over a multihop path is proposed. Wu and Ma [14] formulate a rate distortion model to represent the fact that the immoderate utilization of wireless fading channels could incur high distortion due to high probabilities of video package loss and damage. Based on the model, the authors propose the routing algorithm to seek a balance between the distortion and delay.

Since it is hard to decide whether an encountered node is a good relay at the moment of encounter, choosing the relay nodes based on the probability is a feasible solution. In [13], Zhang et al. propose a link model by Wiener process to represent the influence of different link combinations in network topology. Based on the model, the authors design an opportunistic routing metric called the expected transmission cost over a multihop path. To dynamically determine relay candidate set and take into account the effects of uncertainty in node wake-up times, Zhang et al. [15] propose an opportunistic routing which constructs candidate set based on the relatively stable topology and duty-cycle length information.

Most of the existing studies only forces on one attribute, such as the geographic location or the probability. In order to improve network performance in mobile crowdsensing networks, it is necessary to comprehensively consider the influence of spatiotemporal correlation between perception

nodes on network performance and evaluate the importance of each attribute by objective weight when judging each indicator. Therefore, this paper takes the relationship between nodes as the main evaluation index, fuses the nodes' motion state in space, and proposes a method named multiattribute routing based on Pareto optimal (MR-Pareto), which uses the nondominant ordering of objective weighted evaluation method based on focusing on the energy of nodes and cache.

## 3. Node Position Detection

Since the nodes are mobile, the network links are prone to be interrupted. Therefore, based on the sharing of geographical location information, the nodes that can successfully forward messages within the time threshold are screened out in this paper. Each node periodically broadcasts Hello information within a hop range. When the node $i$ receives the Hello packet sent by the node $j$ or the location information that the node broadcasts, the node $i$ can determine the encounter with the node $j$. This information contains the geographical position, movement speed, and direction at the current time. With the help of the publishing mechanism of geographical location information, the region of the neighbor node within the range of the forwarding node can be determined, and the node can scan the updated neighbor list within a certain period to detect the future geographic location of the neighbor nodes combined with the saved location coordinate information. $(x_i, y_i)$ is defined as the position coordinates of node $i$, and $(x_i, y_i)$ is defined as the position coordinates of node $j$, and then, the distance between two communication nodes $\mathrm{Dist}(i, j)$ can be calculated by formula (1):

$$\mathrm{Dist}(i,j) = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}. \tag{1}$$

Figure 1 is the schematic diagram of the neighbor nodes in the detection communication range waiting for the message to be forwarded. Suppose $\mathrm{Range}(i)$ is the communication range of node $i$, $\mathrm{Dist}_{\mathrm{now}}(i, j)$ is the current distance between node $i$ and node $j$, and $\mathrm{Dist}_{\mathrm{pro}}(i, j)$ is the sine component of the velocity vector difference between node $i$ and node $j$ when the node $j$ reaches the boundary of the communication range of node $i$. $\mathrm{Dist}_{\mathrm{rest}}(i, j)$ means the remaining distance of sustainable connection between the two nodes within the communication range, which can be calculated as follows:

$$\mathrm{Dist}_{\mathrm{rest}}(i,j) = \sqrt{\mathrm{Range}^2(i) - \mathrm{Dist}^2_{\mathrm{pro}}(i,j)} - \mathrm{Dist}_{\mathrm{now}}(i,j) \cdot \cos\theta_0. \tag{2}$$

To prevent link interruption caused by node movement, it is necessary to detect whether the location of neighboring nodes around the forwarding node is still within the reachable communication range after exceeding the transmission time threshold $T_{\mathrm{threshold}}(i, j)$. If the node can ensure that the link is always connected during the process of forwarding all the different messages to the neighbor nodes, it can be confirmed that this node meets the requirements of location transmission.

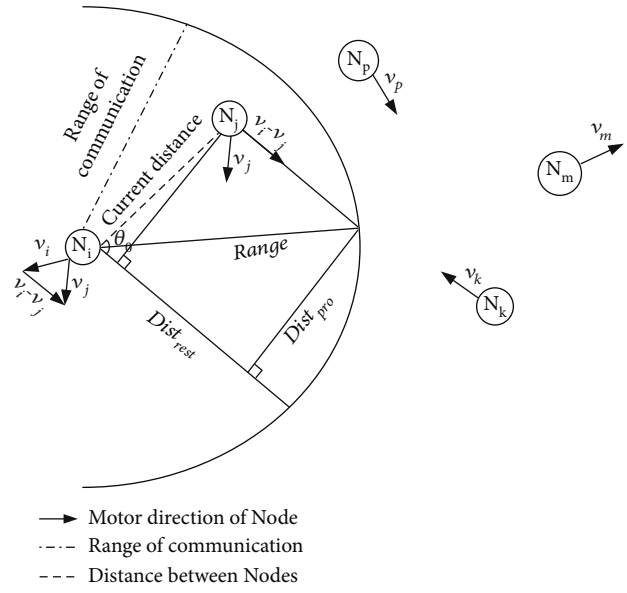



Figure 1: The node probes the neighbor nodes.

By comparing the relationship between the transmission time threshold of link continuous connection and the remaining connection time during message forwarding between two nodes, we can determine whether the forwarding requirements are met. $T_{\mathrm{threshold}}(i, j)$ is determined by the length of node exchange message, and the remaining connection time $T_{\mathrm{rest}}(i, j)$ between node $i$ and node $j$ can be calculated in equation (3):

$$T_{\mathrm{rest}}(i,j) = \frac{\mathrm{Dist}_{\mathrm{rest}}(i,j)}{\left|v_i - v_j\right|}. \tag{3}$$

## 4. A Multiattribute Routing Method Based on Pareto Optimality

*4.1. Pareto Candidate Node Set.* Due to the lack of fixed infrastructure and the constantly changing network topology in mobile group-intelligence perceptions, each node needs to meet its neighbor nodes to carry out data transmission. Since the communications between nodes are influenced by trajectory and behavior factors, it is typical to choose a reliable and utility node transmission task when choosing relay nodes. Two node contact should be relatively stable and frequent, so that the transmission of sensory data can be accomplished in order to better use collaboration between the nodes to adapt to the change in the network structure. Therefore, the centrality, similarity, and relevance of nodes are taken as the basis for the selection of relay nodes in this paper.

*Definition 1.* Node centrality.

The number of current neighboring nodes is commonly taken as the criterion to measure the centrality of nodes. [16]

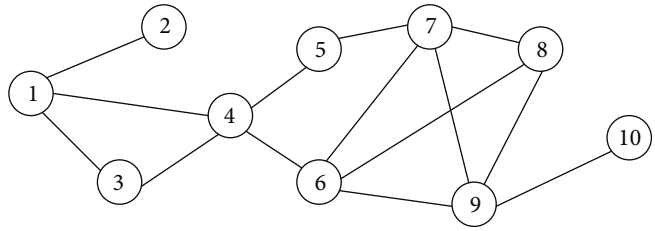



Figure 2: Small-scale mobile crowdsensing network.

showed that according to the data generated by multiarea mobile devices, after comparison and analysis, nodes form certain connection relationships after multiple contacts, which facilitates the understanding of social behaviors of potential users in wireless networks and leads to information diffusion. The more centrality a node has in the MCS network, the more it can promote network communication. Centrality is usually represented by the degree of a node in an undirected graph. In an undirected graph with nodes, the centrality of the node is defined as

$$\mathrm{CEN}(i) = \sum_{j=1}^{n} A_{ij} (i \neq j), \tag{4}$$

where $\mathrm{CEN}(i)$ represents the center degree of the node $i$ and is also the sum of nodes that can be contacted by the node $i$ and other $n-1$ nodes in the network and $A$ represents the adjacency matrix of node $i$ and node $j$ contact. It is assumed that the contact of nodes is bidirectional. If there is a path from node to node, there is also a path from node $i$ to node $j$. $A$ is the symmetric matrix of $n$ by $n$. If there are edges between node $i$ and node $j$, we have $A_{ij} = 1$; otherwise, $A_{ij} = 1$. Therefore, the node center degree can be calculated by formula (4). It can not only reflect the relationship between the node and other nodes but also reflect the key degree of this node according to the network scale. As shown in Figure 2, a mobile crowdsensing network with a number of 10 nodes can be obtained. According to the above definition, the centrality of node 4 is 4.

*Definition 2.* Node correlation.

Node correlation is used to represent the probability of meeting nodes that can transmit messages through collaboration in mobile crowdsensing network. When the probability of confrontation between the node and the destination node is higher, the ability of the node to forward to the destination node is stronger, and the delivery rate of messages in MCS network is higher. The duration of connection establishment and the time interval from the last mutual contact represent the time and frequency of each node's contact with other nodes.

Assuming that the contact time follows an exponential distribution, it represents the probability of the duration of two nodes in a certain period of time [17]. The node correlation is shown in equation (5):

$$\mathrm{COR}(i,j) = 1 - \exp\left(-\frac{U(i,j)}{I(i,j)}\right). \tag{5}$$

where $\mathrm{COR}(i,j)$ represents the probability of the node $i$ and the node $j$ meeting within a period of time, $U(i,j)$ is the length of the last connection between the node $i$ and the node $j$, and $I(i,j)$ is the time interval between the node $i$ and the node $j$ from last connection. The longer the connect time of duration (the connection time) between nodes, the higher the value of $COR(i,j)$; the shorter the interval from the last connection, the higher the value of $COR(i,j)$. If the node $i$ and the node $j$ experience a long period of time since the last encounter, the probability of meeting is updated by the decay factor $\gamma$ within the time $t$ from the last encounter, as shown in equation (6). If two nodes can meet again after a short contact, and the span of the two encounters is shorter, the possibility of the two nodes meeting again in the future is greater.

$$\mathrm{COR}(i,j) = 1 - \mathrm{COR}(i,j)_{\mathrm{old}} \times \gamma^{t}. \tag{6}$$

*Definition 3.* Node similarity.

Two nodes are more similar if they have several common neighboring nodes and they often meet. $N(i)$ and $N(j)$ are defined as the set of neighbor nodes of nodes $i$ and $j$, respectively. $S(i,j)$ can be measured by

$$S(i,j) = N(i) \cap N(j). \tag{7}$$

Since each node $i$ cannot obtain global information, the similarity of all common neighbors is estimated by exchanging $N(i)$ with $N(j)$ and when $N(i)$ satisfied $N(j)$. Newman [18] calculated and verified the relationship between the number of common neighbors at time $t$ and the probability of their future cooperation. The similarity of the node $i$ and the node $j$ in MCS network is represented by the ratio of the intersection set of all common neighbor nodes existing between other nodes encountered by these two nodes and the union set of their respective neighbor nodes, and the similarity $SU(i,j)$ is shown in equation (8).

$$\mathrm{SU}(i,j) = \left|\frac{N(i) \cap N(j)}{N(i) \cup N(j)}\right|. \tag{8}$$

If the intersection number of neighbor nodes set of the two nodes is larger, it can be seen that the two nodes are more likely to forward messages through the common relay node. Therefore, the calculation of similarity can reflect the

efficiency of message transmission to a certain extent. When calculating the similarity process between two nodes, the symmetric matrix can be expressed as the similarity between nodes in MCS network by referring to the graph theory of node centrality. In Figure 2, according to the above definition, it can be determined that the similarity of node 4 and node 7 is 1/3, and their common neighbor nodes are node 5 and node 6.

*4.2. Nondominated Sorting.* Since the centrality, relevance, and similarity of nodes are taken as comprehensive evaluation indexes in this paper, in order to make these three attributes reflect the overall characteristics of the relationship between nodes, it is necessary to select the comprehensive evaluation method of multiple factors to calculate the weight of each attribute element, so as to maximize the characteristic values of the relationship between nodes. According to the comprehensive evaluation of multiple factors, it can be classified as subjective weighting evaluation method and objective weighting evaluation method [19]. The subjective weighting evaluation method takes the attention level of the strategic decision-maker in each evaluation indicator as the standard and determines the weight distribution of multiple indicators by subjective assumption. Because the raw data in this method follow the subjective judgment of experience and human factors interfere greatly, it has obvious limitations in the comprehensive evaluation of the three attributes representing the relationship between nodes in MCS network. Therefore, this paper adopts the nondominant ordering method in the objective weighted evaluation method, which can solve the problem of how to obtain the optimal solution by combining the objective functions constructed by the three attributes when the three objective functions of node centrality, relevance, and similarity are in conflict. However, this kind of problem can only seek the noninferior solution; that is, Pareto is optimal. Therefore, the quantization problem of the relationship between the attributes of three nodes is transformed into the problem of comparing the overall characteristics between nodes.

*Definition 4.* Pareto dominance.

For multiple objective function minimization problems, assumption of a vector is$\bar{f}(\bar{X}) = (f_1(\bar{X}), f_2(\bar{X}), \cdots, f_n(\bar{X}))$, which is made up of$n$target components$f_i(i = 1, 2, \cdots, n)$, given any two decision variables$\bar{X}_u, \bar{X}_v \in U$: If and only if, for$\forall i \in \{1, 2, \cdots, n\}$, there is$f_i(\bar{X}_u) \le f_i(\bar{X}_v)$and there is at least one$j \in \{1, 2, \cdots, n\}$. Make$f_j(\bar{X}_u) \le f_j(\bar{X}_v)$true; thus,$\bar{X}_u$ has weak-dominated of$\bar{X}_v$. If and only if,$\exists i \in \{1, 2, \cdots, n\}$, then$f_i(\bar{X}_u) < f_i(\bar{X}_v)$; at the same time,$\exists j \in \{1, 2, \cdots, n\}$. Make$f_j(\bar{X}_u) > f_j(\bar{X}_v)$true; thus,$\bar{X}_u$non-dominates$\bar{X}_v$.

The proposal of nondominated sorting method is to solve the Pareto optimal solution set [20]; the method is based on the Pareto solutions of individuals to hierarchical groups, aiming the algorithm used the cycle to adapt to the grading at the mercy of the form, the search direction to the Pareto optimal solution set to calculate the final result.

If there is only one objective function, the maximum or minimum solution is in the limit position in the global, so it is better than other solutions in the solution set. If the problem consists of multiple objective function, multiple function cannot be achieved in the process of solving the absolute equilibrium state. It is difficult to find a solution to make the multiple objective function to achieve the best effect; in other words, even though a solution can ensure that the result of the objective function is optimal, the rest of the function is not necessarily the best result. Therefore, for multiobjective optimization problems, there is often a set, which cannot be compared between all objective functions and their characteristics; that is, the utility of an objective function cannot be reduced without increasing the utility of any objective function. This solution called nondominated solution or Pareto optimal solution is defined as follows:

For multiobjective minimization problems, assumption of a vector is $\bar{f}(\bar{X}) = (f_1(\bar{X}), f_2(\bar{X}), \cdots, f_n(\bar{X}))$, which is made up of $n$ target components $f_1(i = 1, 2, \cdots, n)$. Set $\bar{X}_u \in U$ is the decision variable; if $\bar{X}_u$ is Pareto's optimal solution and then if and only if, when the decision variable $v = f(\bar{X}_v) = (v_1, \cdots v_n)$ cannot dominate $u = f(\bar{X}_u) = (u_1, \cdots u_n)$, $\bar{X}_v \in U$, do not make formula (9):

$$\forall i \in \{1, 2, \cdots, n\}, v_i \le u_i \cap \exists i \in \{1, 2, \cdots, n\} | v_i \le u_i. \quad (9)$$

To find the Pareto optimal set means to find the Pareto optimal front. As shown in Figure 3, the points on the line represent viable choices, where smaller values are better than larger ones. Point $C$ is not at the Pareto boundary because it is controlled by points $A$ and $B$. Points $A$ and $B$ are not strictly occupied by any other points, so they do lie on the boundary.

When a node performs a fast nondominant ordering of its neighbor nodes, the population size is assumed to be $P$, and two parameters $n_p$ and $S_p$ of each individual $p$ in $P$ need to be calculated in this algorithm. $n_p$ in the population represents the number of individuals of dominant individual $p$, which is determined by the first nondominant layer of the number of individuals in the feasible solution set that can be dominated and is expressed as the set of individuals that are dominated by individual $p$, and its quantitative determination is based on the result that all individuals in the feasible solution can be dominated by individual $p$ constituting the set. After traversing the entire population, the complexity of the algorithm is $O(mN^2)$ [21]. The main steps of the algorithm are as follows:

(1) Identify each $n_p = 0$ individual in the population and store it in the current nondominant set $F_1$

(2) Consider of all the individual $i$ in the present constituted set of nondominating and then look at the set of individuals $S_i$ which they can dominate; if there exists an individual $l$ in $S_i$, then reduce the dominated set of individual $L$ by 1, namely, $n_l = n_l - 1$ (due to the fact that the elements which the individual $l$ can control in the governing set $S_i$ are already

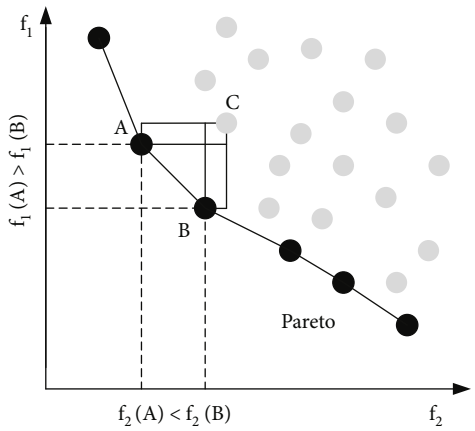



Figure 3: Pareto frontier.

stored in $S_i$); if $n_l$ reduces to zero, namely, when $n_l = 0$, store the individual $l$ into other set $H$

(3) Consider the first nondominant solution set as the entire solution set which represents the optimal set of individuals, since only the disposable individuals can be constrained by any other individuals in the set, for each individual in the set based on a similar nondominant order. Then, the hierarchy of the set is defined by the above steps, and the same nondominant order is given until each individual in the set is graded

The selection of Pareto candidate node set refers to the Pareto optimal solution set obtained by the social context information results between nodes, taking the three attributes of centrality, relevance, and similarity as objective functions and computing the objective weighting evaluation method of nondominant ordering. Not only does this algorithm take into account the local connections between nodes but also avoids the interference of subjective factors on the determination of weights, as shown in Algorithm 1.

*4.3. Relay Node Determination.* In mobile crowdsensing system, users usually use small smart mobile terminals as perception devices. Although these devices have considerable computing power and perception capacity, they often have certain limitations in energy or cache reserve. If the power or cache of the device cannot meet the work needs, the perception task undertaken by the device will not be completed, and the performance of the whole network will be greatly reduced [22].

Due to node energy consumption, which is mainly caused by the data transmission task processing, signal processing and hardware operation [23], data transmission, and the task process, the energy consumption is larger. Therefore, the model is established based on the energy which mainly considers two aspects of data transmission and task processing and data transmission energy consumption including node scanning, sending messages, and receiving messages, three parts. The energy consumption of node scanning is defined as the energy consumed when the node

```
1: Input: Node Set P
2: Output: Candidate Node SetPareto
3: Procedure ParetoCandidate
4: Calculate the centrality of node
   CEN(i) = ∑_{j=1}^{n} A_ij (i ≠ j)
5: Calculate the correlation of node
   COR_ij = 1 − COR(i, j)_old × γ^t
6: Calculate the similarity utility of node
   SU_ij = |N(i) ∩ N(j)/N(i) ∪ N(j)|
7: LET F be the first dominating set
   % Non-dominated Sorting
8: for p in P do
9:   let Sp be the dominating set
10:  for q in Pdo
11:      add q to the Sp list when p dominates q
           n_p add 1 when p is dominated by q
           take n_p of the individual is 0 be F1
12:  end for
13: end for
14: F.append(F_1)
15: SET index = 0
16: while F [index] do
17:   let Q be the dominated individual set
        Sort Sp
        F.append(Q)
        index ⟵ index + 1
18: end while
```

Algorithm 1: Select nodes from candidate node set Pareto.

carries out periodic scanning of the surrounding environment, including the energy consumption when scanning the node channel preparing for communication and sensing data. Let $e_s$ be the energy lost in a single scan of the node, $t$ be the working time of the node and be the scan cycle, and then, the energy consumption of node scan $E_s$ can be defined as formula (10):

$$E_s = e_s \times \frac{t}{T}. \tag{10}$$

When a node transmits data, the data length is usually used to measure the amount of energy consumed. Therefore, the energy consumption during message sending and receiving is set as a positive correlation function with the corresponding message size. When the message size is larger, the energy consumption during transmission is larger. Let $e_t$, $e_r$, and $M_{size}$ denote the energy required to send a unit message, the energy required to receive a unit message, and the message size, respectively. Energy consumption for sending messages is given by

$$E_t = e_t \times M_{\text{size}}, \tag{11}$$

and energy consumption for receiving messages is given by

$$E_r = e_r \times M_{\text{size}}. \tag{12}$$

Since task processing represents mostly the energy consumed by tasks running in the background of user nodes, the use of cache space is taken as the measurement factor of energy consumption required by node tasks. If the cache space used by nodes is large, it indicates that there are more tasks running in the background, thus consuming more energy. Let $e_d$ denote the energy consumption per unit message processing and $B$ denote the use size of cache space. We can compute the task processing energy consumption by

$$E_d = e_d \times B. \tag{13}$$

Let $E$, $E_{\text{consume}}$, and $E_{\text{current}}$ denote the total energy value of the node, the energy consumed by the node, and the residual energy of the node, respectively. $E_{\text{consume}}$ and $E_{\text{current}}$ can be computed by

$$\begin{aligned} E_{\text{consume}} &= E_s + E_t + E_r + E_d, \\ E_{\text{current}} &= E - E_{\text{consume}}. \end{aligned} \tag{14}$$

In order to deal with the problem of excessive depletion of node resources in mobile crowdsensing network, this algorithm takes full account of the dynamic changes of residual energy of nodes and cache space when selecting relay nodes and defines the measurement value of relay nodes as shown in formula (15)

$$EB = \alpha E_{\text{current}} + \beta B, \tag{15}$$

where $\alpha$ and $\beta$ denote the tuning parameter. When the measure value of the node $EB$ is greater than the average value of the measure value of all nodes in the Pareto candidate node set, this node is regarded as a relay node.

*4.4. Message Forwarding Strategy.* In the message forwarding phase, nodes detect all neighbor nodes, according to the update location information table and the neighbor table. Then, the nodes use the geographical position information table to search the neighbor node that meets the requirements for the forward. If a node finds other that meets the communication requirements and set it as the destination node, the node will forward message directly. If not, the node needs to establish candidate node set through the node relationship. Otherwise, the nodes are filtered by the relationship between nodes to obtain the candidate node set. Then, choose those nodes of which the measure value $EB$ is higher than the average as the relay node from the candidate node set for multicopy forwarding.

*4.5. Algorithm Implementation of MR-Pareto*

(1) Node $N$ initiates the forward request and broadcasts its location information

(2) Evaluate whether any node responds to the forwarding request of the node $n$. If other nodes respond to node requests, execute (3); Otherwise, abandon this forwarding

(3) Node $n$ and the responding neighbor nodes swap location information; calculate the remaining connection time between nodes $T_{test}$, and then, execute (4)

(4) Determine the remaining connection threshold between nodes $T_{\text{threshold}}$ according to the forwarding message length

(5) Compare the values of the remaining connection time between nodes $T_{\text{rest}}$ and the remaining connection threshold between nodes $T_{\text{threshold}}$. If $T_{\text{rest}} > T_{\text{threshold}}$, thus, this node is considered to meet the forwarding condition

(6) Get the relationship between nodes, calculate the center degree, correlation degree, and similarity degree of nodes in the detection set conforming to node position, and then, execute (7).

(7) The Pareto solution set of the objective function composed of three attributes of the nodes is calculated as the candidate nodes set through nondominant ordering

(8) Get the nodes' energy $E_{\text{current}}$ and the cache size $B$ among the candidate nodes set

(9) Compare the measurements of node's energy and cache size $EB$ to its mean value. If $EB$ is greater than its mean value, the message is forwarded. Otherwise, ignore this node

## 5. Performance Evaluation

*5.1. Performance Index.* In this paper, the simulation software ONE is used as the experimental platform. Moreover, the MR-Pareto routing method is compared with many traditional opportunistic routing methods.

(1) The residual energy is used to evaluate the lifetime of the mobile crowdsensing network, and the average residual energy of the network nodes $\text{energy}_{\text{avg}}$ is shown in formula (16), where $\text{energy}(n_i)$ represents the residual energy of network node $n_i$

$$\text{energy}_{\text{avg}} = \frac{\sum_{i=1}^{n} \text{energy}(n_i)}{n} \tag{16}$$

(2) The network overhead is expressed as the total number of all message forwarding, as shown in formula (17), where $\text{transmission}(M_i)$ is the number of forwards of message $M_i$

$$\text{overhead} = \sum_{i=1}^{m} \text{transmission}(M_i) \tag{17}$$

(3) Message delivery rate reflects the situation where messages can be delivered to the destination nodes successfully through collaboration, as shown in

formula (18), where deliver$_c$ount is the number of successfully delivered messages and create$_c$ount indicates the quantity of messages in the network

$$\text{delay}_{\text{pro}} = \frac{\text{deliver}_c\text{ount}}{\text{create}_c\text{ount}} \tag{18}$$

(4) Packet loss quantity is used to measure the total packet loss quantity of nodes under different routing strategies and congestion control strategies, as shown in formula (19), where drop$(n_i)$ represent the packet loss quantity of node $n_i$

$$\text{drop} = \sum_{i=1}^{n} \text{drop}(n_i) \tag{19}$$

*5.2. Simulation Environment.* Three groups of nodes with different movement speeds were set up in the experiment to compare the network efficiency of the MCS routing strategy. For the first group, the number of nodes was set as 80, and the movement speed was 0.5-1.5 km/h. In the second group, 40 nodes were set, and the movement speed was 2.7-13.9 km/h. The number of nodes in the third group is set to 6, the movement speed is set to 7-10 km/h, and two of them set their transmission range size to 1000 m. The specific parameters are shown in Table 1.

Table 1: User matches with drone.

| Parameters name | Parameters values |
|---|---|
| Scene size | 4500 m × 3400 m |
| First group of nodes | 80 × (0.5 – 1.5 km/h) |
| Second group of nodes | 40 × (2.7 – 13.9 km/h) |
| Third group of nodes | 6 × (7 – 10 km/h) |
| Mobile model of the first group | Random walk movement |
| Mobile model of the second group | Map route movement |
| Mobile model of the third group | Map route movement |
| Transmission range | 900 m and 900 m |
| Bandwidth | 250 KBps |
| Range of message size | 500 k – 1 M |
| Message generation interval | 5-15 s, 15-25 s, 25-35 s, 35-45 s, and 45-55 s |
| TTL | 300 min |
| Node cache | 10M, 15M, 20M, 25M, and 30M |
| Routing strategy | MR-Pareto, direct, epidemic, and prophet |
| Cache strategy | FIFO |
| $\alpha$ | 100 |
| $\beta$ | 1 |
| Initial energy $E$ | 100 J |
| Scanning energy cost $e_s$ | 0.02 J/time |
| Receiving energy cost $e_r$ | $2.4 \times 10^{-7}$ J/bit |
| Sending energy cost $e_t$ | $3.3 \times 10^{-7}$ J/bit |
| Simulation time | 12 h |

*5.3. Simulation Results.* The energy of the energy network, network overhead, packet loss quantity, and message delivery rate of the MR-Pareto routing algorithm was further compared under different simulation times. The results of this comparison are reported in Figures 4–7. Other parameters in the simulation parameter setting table remain unchanged, and the performance of each routing strategy is evaluated by changing the simulation time. Gradually increase the simulation time from 6 hours to 18 hours. It is proved that the performance of MR-Pareto routing strategy is better than that of epidemic and prophet traditional routing strategy at different simulation time.

As can be seen from Figure 8, compared with epidemic and prophet routing strategies, the average residual energy of the network significantly improves, especially with the increase of simulation time. In the whole simulation process, the energy consumption of epidemic and prophet routing strategies is approximately equal. When the simulation time was 6 hours, the MR-Pareto routing strategy and the epidemic and prophet routing strategy improved by 2.3% on average, and when the simulation time increased to 18 hours, the MR-Pareto routing strategy and the epidemic and prophet routing strategy improved by 14.2% on average. Since most mobile nodes work continuously and cannot provide power in time, the longer the working time, the more energy can be saved.

Figure 4 illustrates the network overhead of the MR-Pareto, epidemic, and prophet routing strategies over simulation time. The epidemic and prophet routing strategies are consistently above 61.5% in the simulation time, while the MR-Pareto routing strategy remains at a relatively low level in terms of network overhead, which remains below 3% in the experiment. As can be seen from the figure, compared with the epidemic and prophet routing strategies, the network overhead of the MR-Pareto routing strategy has decreased by 48.3%. The reason is that the MR-Pareto routing strategy optimizes the selection of relay nodes and reduces the number of messages forwarding and the network resources occupied by message transmission.

In Figure 5, the three routing strategies increase with the simulation time. In the whole simulation process, the epidemic routing strategy has an average 7.2% higher packet loss than the prophet routing strategy. The simulation time of the MR-Pareto routing strategy increased from 6 hours to 12 hours, and the number of lost packets increased from 7,954 to 20,543. MR-Pareto is 37.5% lower than epidemic routing strategy. Compared with prophet, the reduction was 35.8%. The MPOP routing strategy performs well in controlling the number of lost packets.

Figure 6 analyzes the trend of MR-Pareto routing strategy and epidemic and prophet routing strategy over simulation time. When the simulation time is 6 hours, the advantage of MR-Pareto routing strategy is small, but with

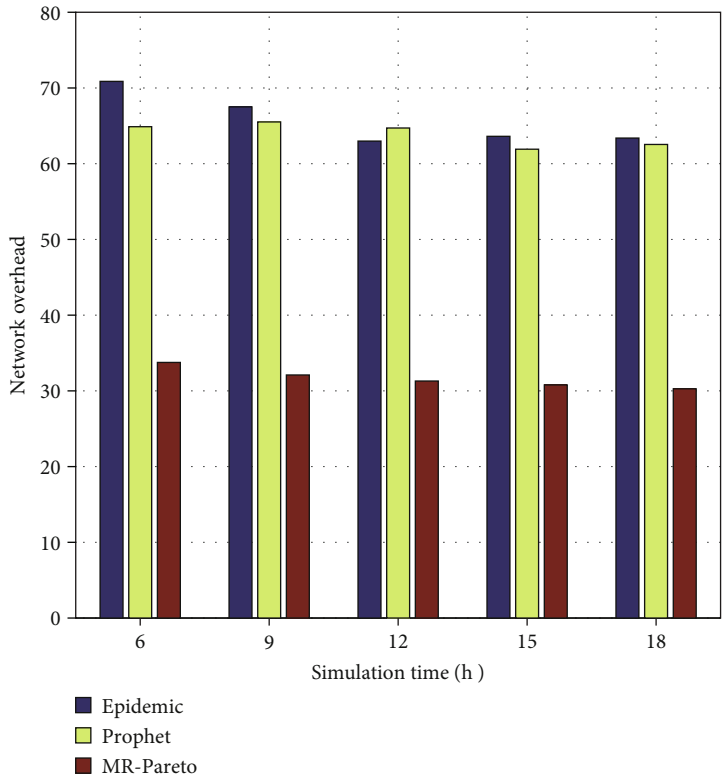



Figure 4: Network overhead changes with simulation time.

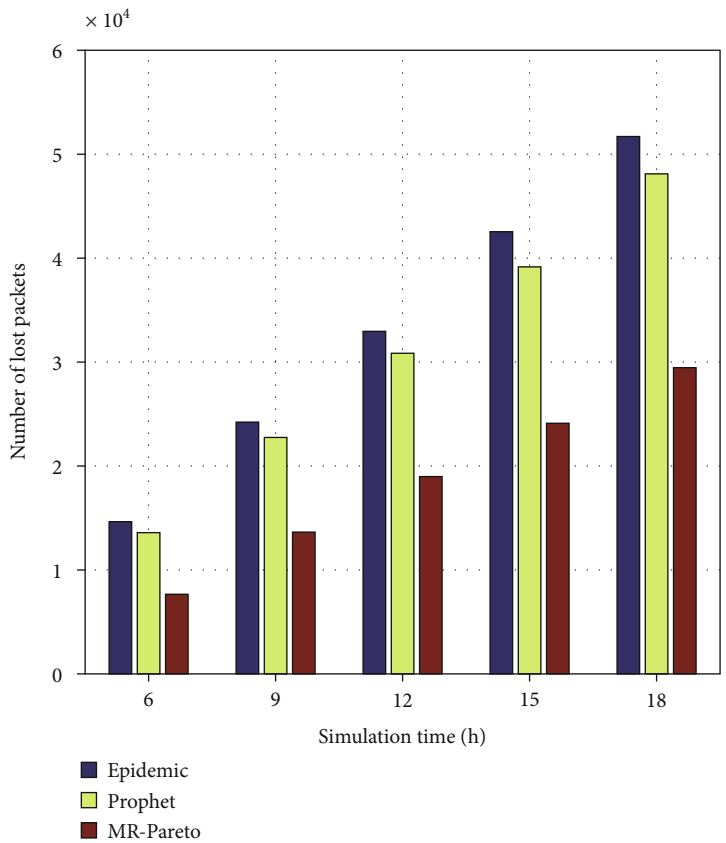



Figure 5: Number of lost packages change with simulation time.

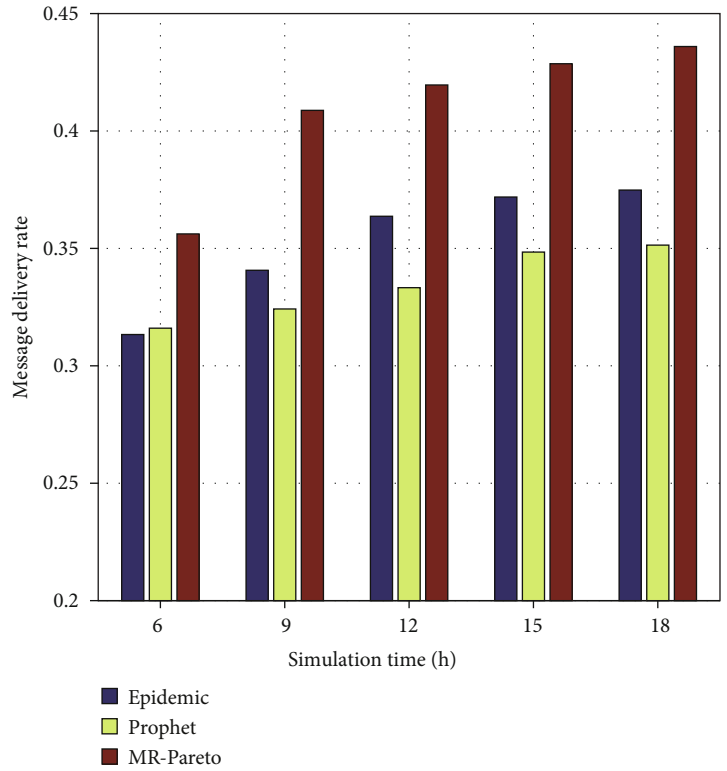



Figure 6: Message delivery rate changes with simulation time.

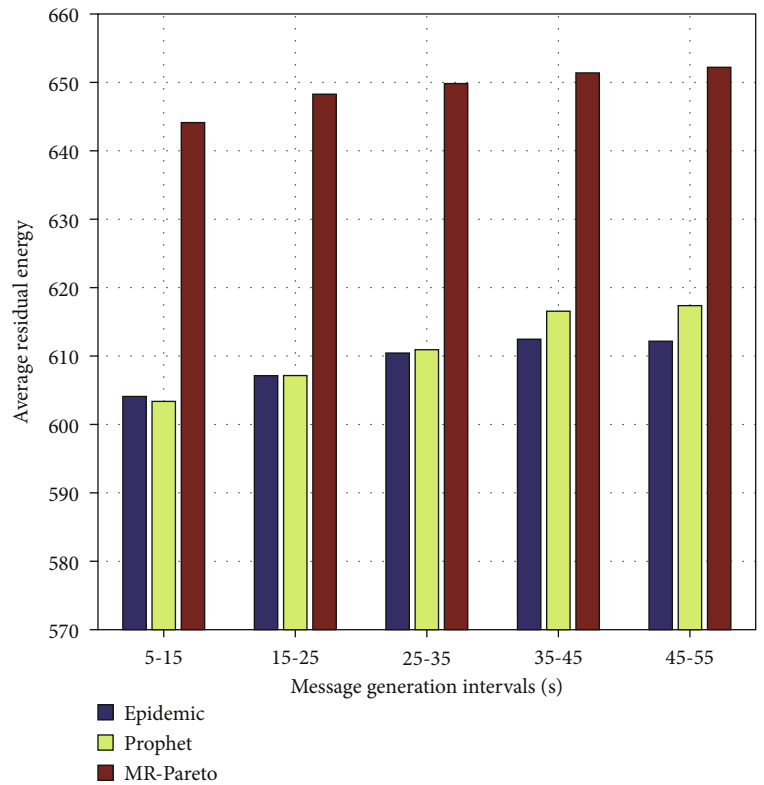



Figure 7: Average residual energy changes with the message generation intervals.

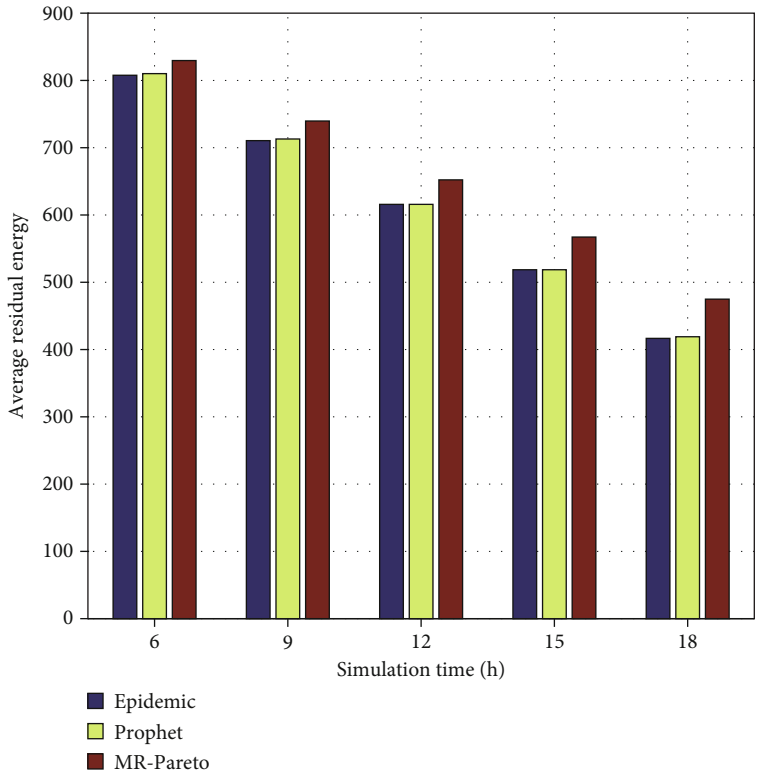



Figure 8: Average residual energy changes with simulation time.

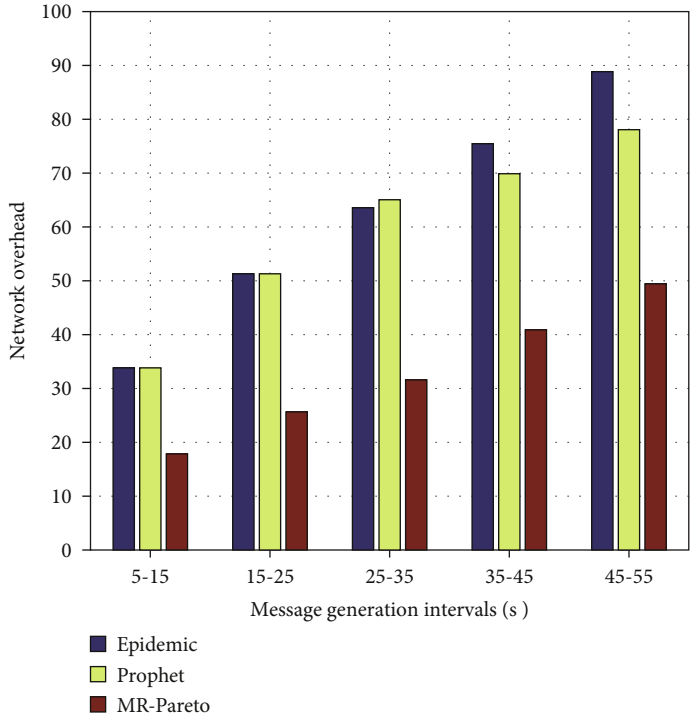



Figure 9: Network overhead changes with the message generation intervals.

the increase of time, the delivery rate of MR-Pareto message keeps higher and higher, staying above 0.4, increasing by 14.1% and 19.2%, respectively, compared with the other two routing strategies. Compared with MR-Pareto, due to the spread and forwarding in the network, epidemic makes all encounter nodes to carry a copy of the message, which

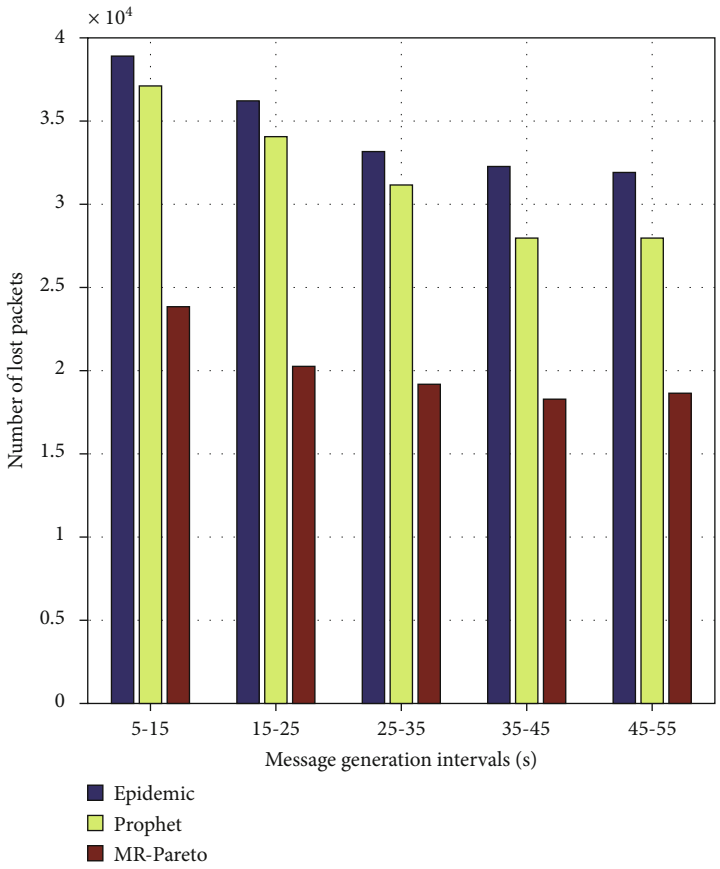



Figure 10: Number of lost packets changes with the message generation interval.

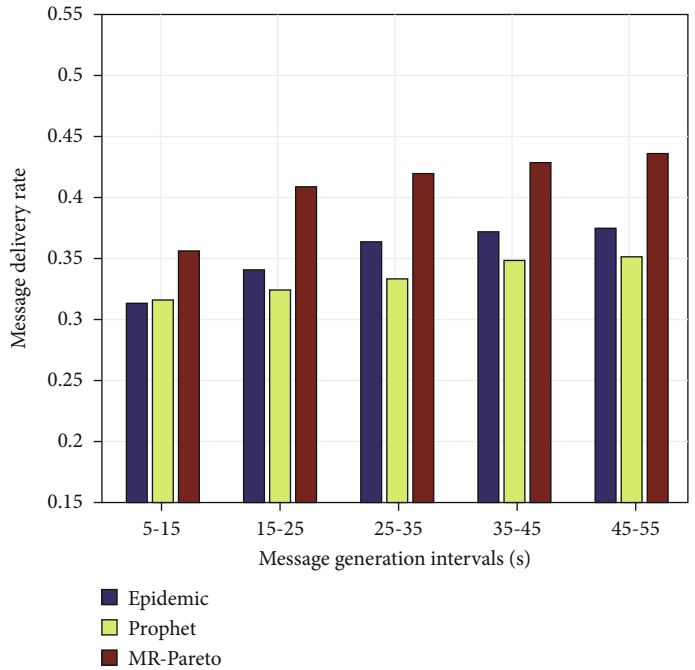



Figure 11: Message delivery rates changes with the message generation intervals.

makes the message delivery rate higher than the prophet routing strategy, but also results in partial meaningless forwarding, which produce a smaller number of successful message delivery to destination nodes than the MR-Pareto routing strategy.

Figure 7–11 presents the residual energy, network overhead, packet loss number, and message delivery rate of the proposed MR-Pareto routing algorithm under various message generation intervals. In order to ensure the efficiency of the simulation, referencing to the simulation parameters set in the table, the MR-Pareto routing algorithm with the other two traditional routing policy performance evaluation, in this paper, consider the MCS intensity produced by the network news; time interval will be generated from 5 to 15 s which gradually increased to 45 to 55 s, to prove that in different time intervals, MR-Pareto routing algorithm compared with the epidemic and prophet traditional routing strategy performance is better.

According to Figure 7, the average residual energy of the three routing strategies increases progressively under the conditions of different message generation intervals, due to the fact that the sensing nodes minimize the generation of new messages and decrease the energy consumption of the nodes during message forwarding and sharing. The epidemic routing strategy is slightly higher than the prophet routing strategy when the message generation density is high but lower than the prophet when the message generation rate is slow. The average MR-Pareto routing strategy is 6.5% and 6.1% higher than epidemic and prophet, respectively.

Figure 9 shows the variation of network overhead with message generation interval between the MR-Pareto, epidemic, and prophet routing strategies. With the decrease of message generation density, the network overhead of the three comparison routing strategies increases gradually. When the time interval of message generation is set at 15-25 s, the gap between epidemic routing strategy and prophet routing strategy increases gradually and remains at a high level. Compared with the other two routing strategies, the network overhead of the MR-Pareto routing strategy is 46.2% lower than that of the traditional epidemic and prophet routing strategies.

It can be seen from Figure 10 that under the circumstance of different message generation intervals, the number of lost packets of the MR-Pareto routing strategy tends to be stable, indicating that it has good stability. The number of lost packets of the three routing strategies decreases gradually because the number of messages generated by the network decreases. The comparison of MR-Pareto routing strategies with respect to epidemic and prophet routing strategies decreased by 34.7% and 32.6%, respectively.

Figure 11 depicts the message delivery rates for different message generation intervals. When the message generation time is 5-15 s, the delivery rate of the MR-Pareto routing strategy is significantly higher than that of the other two routing strategies. It can be concluded that under the condition of relatively large message generation density, the delivery rate of the message is higher. Compared to epidemic and prophet, the MR-Pareto routing strategy has an average improvement of 16.7% and 25.4%, respectively.

## 6. Conclusion

Given the problem that existing routing strategies fail to combine the spatiotemporal characteristics of nodes, single consideration factors, and subjective intention to balance all attributes in the resource-constrained environment of mobile group-intelligence sensing network, this paper proposed a multiattribute routing method based on Pareto optimal solution. By predicting the location of nodes in communication time, this method evaluates the possibility of nodes becoming node diversity and takes the relationship between nodes as a measuring factor. Pareto optimal method is adopted to calculate the measurement value of nodes becoming candidate nodes. Finally, the message forwarding strategy is determined by combining node energy and cache to realize the optimization of routing method. The simulation results show that, compared to the epidemic and prophet routing strategies, the residual network energy of this routing method increases by 8.3% on average. Moreover, the network performance, such as network overhead, packet loss quantity, and message delivery rate, has also improved. In the case of different message generation intervals, in addition to the overall residual energy of the network increased by 6.3% on average, the network overhead, packet loss quantity, and message delivery rate were significantly improved comparing to the two traditional routing strategies. The MR-Pareto routing method has obvious effect on prolonging network life and improving data forwarding performance.

## Data Availability

There is no data set used in this article; the nodes used in the experiment were randomly generated in the simulation environment. In this paper, the specific parameter setting method is given.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] X. Zhang, Z. Yang, W. Sun et al., "incentives for mobile crowd sensing: a survey," *Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 54–67, 2015.

[2] B. Guo, Z. Wang, Z. Yu, and X. Zhou, "Mobile crowd sensing and computing," *ACM Computing Surveys (CSUR)*, vol. 48, no. 1, pp. 1–31, 2015.

[3] D. Zhao, H. Ma, L. Liu, and X. Y. Li, "Opportunistic coverage for urban vehicular sensing," *Computer Communications*, vol. 60, pp. 71–85, 2015.

[4] H. L. Krauss, C. W. Bostian, and F. H. Raab, "Opportunistic networking: data forwarding in disconnected mobile ad hoc networks," *IEEE Communications Magazine*, vol. 44, no. 11, pp. 134–141, 2006.

[5] H. Ma, D. Zhao, and P. Yuan, "Opportunities in mobile crowd sensing," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 29–35, 2014.

[6] S. Chen, Z. Chen, J. Wu, and K. Liu, "An adaptive delay-tolerant routing algorithm for data transmission in opportunistic social networks," *Electronics*, vol. 9, no. 11, p. 1915, 2020.

[7] K. Liu, Z. Chen, J. Wu, and L. Wang, "FCNS: a fuzzy routing-forwarding algorithm exploiting comprehensive node similarity in opportunistic social networks," *Symmetry*, vol. 10, no. 8, p. 338, 2018.

[8] C. Liu, D. Fang, X. Liu et al., "Low-cost and robust geographic opportunistic routing in a strip topology wireless network," *ACM Transactions on Sensor Networks*, vol. 15, no. 2, pp. 1–27, 2019.

[9] T. Fu, Z. Wen, J. Liu, Z. Zheng, and W. Li, "An Adaptive Energy Saving Scheme of DTN based on Geographic Grid for Human Living Areas," in *2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, Chongqing, China, 2018.

[10] A. Ali, A. B. Abdel, and H. Hong, "FQ-AGO: fuzzy logic Q-learning based asymmetric link aware and geographic opportunistic routing scheme for MANETs," *Electronics*, vol. 9, no. 4, p. 576, 2020.

[11] S. Hang, G. Bai, and Z. Tang, "QMOR: QoS-aware multi-sink opportunistic routing for wireless multimedia sensor networks," *Wireless Personal Communications*, vol. 75, no. 2, pp. 1307–1330, 2014.

[12] N. Li, M. O. Jose-Fernan, and D. V. Hernandez, "Probability prediction-based reliable and efficient opportunistic routing algorithm for VANETs," *IEEE/ACM Transactions on Networking*, vol. PP, pp. 1933–1947, 2018.

[13] X. Zhang, X. Cao, Y. Long, and K. S. Dan, "A street-centric opportunistic routing protocol based on link correlation for urban VANETs," *IEEE Transactions on Mobile Computing*, vol. 15, no. 7, pp. 1586–1599, 2016.

[14] H. Wu and H. Ma, "Opportunistic routing for live video streaming in vehicular ad hoc networks," in *Proceeding of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014*, Sydney, NSW, Australia, 2014.

[15] X. Zhang, L. Tao, F. Yan, and D. K. Sung, "Shortest-latency opportunistic routing in asynchronous wireless sensor networks with independent duty-cycling," *IEEE Transactions on Mobile Computing*, vol. 19, no. 3, pp. 711–723, 2020.

[16] E. Montijano, G. Oliva, A. Gasparri, F. Yan, and D. K. Sung, "Distributed estimation and control of node centrality in undirected asymmetric networks," *IEEE Transaction on Automatic Control*, vol. 66, no. 5, pp. 2304–2311, 2021.

[17] S. Kim, J. Lee, and S. Yang, "A social overlay-based forwarding scheme for mobile social networks," *Wireless Networks*, vol. 22, no. 7, pp. 2439–2451, 2016.

[18] M. E. NewmanS, J. Kim, and S. Y. Lee, "Clustering and preferential attachment in growing networks," *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 64, no. 2, article 025102, 2001.

[19] C. Chou and M. Chen, "Learning multiple factors-aware diffusion models in social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1268–1281, 2018.

[20] A. Jaszkiewicz and T. Lust, "ND-tree-based update: a fast algorithm for the dynamic nondominance problem," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 5, pp. 778–791, 2018.

[21] M. Sumit, V. Prakash, and M. Buzdalov, "Labeling-oriented non-dominated sorting is $\Theta(MN^3)$," in *GECCO '21: Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2021.

[22] H. Jin and J. Zhao, "Content-centric heterogeneous fog networks relying on energy efficiency optimization," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13579–13592, 2020.

[23] H. Jin and J. Zhao, "Real-time energy consumption detection simulation of network node in internet of things based on artificial intelligence," *Sustainable Energy Technologies and Assessment*, vol. 44, no. 3, article 101004, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 8911651, 21 pages
https://doi.org/10.1155/2022/8911651

*Research Article*

# Metaheuristic Load-Balancing-Based Clustering Technique in Wireless Sensor Networks

**Sandip K. Chaurasiya,[1] Arindam Biswas,[2] Prasit Kumar Bandyopadhyay,[3] Amit Banerjee,[4] and Rajib Banerjee[5]**

[1]*University of Petroleum & Energy Studies (UPES), Department of Cybernetics, School of Computer Science, Energy Acres Building Bidholi, Dehard un-248007, Uttarakhand, India*
[2]*School of Mines, Kazi Nazrul University, Asansol, West Bengal, India*
[3]*Department of Electronics & Communication Engineering, School of Engineering, Sister Nivedita University, DG 1/2, New Town, Action Area 1, Kolkata, West Bengal, India*
[4]*Physics Department, Bidhan Chandra College, Asansol, 713303 West Bengal, India*
[5]*Department of Electronics & Communication Engineering, Dr. B. C. Roy Engineering College, Durgapur, West Bengal, India*

Correspondence should be addressed to Arindam Biswas; mailarindambiswas@yahoo.co.in
and Amit Banerjee; amitbanerjee.nus@gmail.com

Received 11 August 2021; Accepted 6 December 2021; Published 21 January 2022

Academic Editor: Rajesh Kaluri



The resource-constrained nature of wireless sensor networks engenders the development of energy-efficient network operations. To mitigate the prime concern of developing an energy-efficient network, clustering of the nodes has emerged as a very effective tool. If executed intelligently, clustering can not only help in obtaining even load distribution among the network nodes but also help in having the enhanced network lifetime and scalability. In this work, a Metaheuristic Load-Balancing-Based Clustering Technique (MLBCT) in wireless sensor networks has been proposed which formulates the energy-balanced clusters based on the differential evolution technique to improve the network lifetime. To ensure the formation of balanced clusters, several metrics like nodes' proximity, nodes' distribution, and energy distribution across the sensing field have been considered. Moreover, to facilitate the even load distribution among the cluster members, a randomized rotation of cluster head is implemented. The supremacy of the proposed scheme is confirmed through an extensive set of simulations against the state-of-art schemes. Simulation results reflect an average gain of 51.85% in network lifetime under the variable network configurations in an ideal environment. Moreover, a thorough statistical analysis is performed to prove the efficacy of the proposed fitness function by obtaining confidence intervals under two different network scenarios with variable node counts.

## 1. Introduction

A wireless sensor network (WSN) comprises a large number of tiny devices capable of sensing the surrounding, processing the collected data as per the application, and communicating the processed field information to the centralized base station (BS) [1]. However, the sensor nodes deployed (either randomly or deterministically) in the sensing field suffer from several constraints. They are limited in processing abilities, storage abilities, power, and other allied restrictions [2]. Among all these restrictions, limited power is the most severe one as the node drained of all the energy and frequent recharging and replacement cannot be facilitated, especially in remote applications of WSN like habitat monitoring, environmental monitoring, industrial monitoring, and military surveillance systems [3, 4].

Typically, transmission and route allocation consume most of the nodes' energy and are very much responsible for the power drainage of the sensor nodes. Thus, to solve this issue, energy-efficient network layer operations have been targeted by researchers for many years. Routing is the main functionality of the network layer, and hence,

designing an energy-efficient routing protocol is consistently captivating the attention of the community. To the aforementioned, clustering has evolved as a very significant tool that not only eases the task of routing and distributes the load evenly within a cluster but also, through the use of data aggregation, results in substantial saving of nodes' energy to be consumed in other significant network operations.

Clustering has been defined as the grouping of nodes based on some common attributes. In a clustering-based architecture, the network nodes are partitioned into some groups termed clusters. Within the cluster, a node is designated as cluster head (CH) which carries out more energy heavy tasks such as data aggregation and long-distance communication to the sink on behalf of the entire cluster. The rest of the nodes, called cluster members, perform the basic task of sensing and short-distance communication to the CH [5]. To effectively improve the WSN performance, balancing the clusters is a prerequisite. Thus, the formation of clusters in the WSN can be seen as an optimization problem involving multiple variables to be brought into consideration like nodes' proximity, nodes' residual energy, and size of the tentative clusters. The optimization problems can be classified into two major categories—heuristic and metaheuristic.

The primary motivation behind this work is to pursue the problem of clustering through metaheuristic algorithms. As mentioned above, since the formation of balanced clusters leading to the energy-efficient network operation requires the adequate consideration of various parameters such as nodes' proximity and cluster size, optimization techniques can help a lot in having a suitable solution. With the obtainment of balanced clusters and rotation of cluster head's role among the nodes over the network rounds, the foremost goal of network lifetime improvement can be achieved effectively. In this paper, a novel energy-efficient clustering protocol, Metaheuristic Load-Balancing-Based Clustering Technique (MLBCT), is proposed for the wireless sensor networks based on the idea of differential evolution, a metaheuristic technique. The proposed scheme defines a suitable fitness function to formulate the balanced network partitioning. Once the clusters are finalized, the scheme freezes those and enables the CH-role rotation among the cluster members. To prove the scheme's efficacy, an extensive set of simulations demonstrate the showcasing of the improved network lifetime and network energy consumption.

*1.1. Major Contributions and Organization of the Paper.* The major contributions of the proposed MLBCT are as follows:

(i) Design of an appropriate fitness function leading to
  (a) balanced cluster formation
  (b) reduced intracluster communication cost
(ii) Development of a differential evolution-based energy-efficient clustering scheme on the basis of the devised fitness function
(iii) Performance analysis of the proposed scheme, MLBCT
  (a) under varying network configurations to showcase its adaptability and scalability
  (b) with comparison to the state-of-art schemes in terms of network performance
  (c) with statistically justified results

The rest of the paper is organized into five descriptive sections. Section 2 outlines the literature review of the existing works in the same context to identify the technical gaps. Section 3 presents the adopted network model, an introductory discussion on differential evolution, and the terminology to be used throughout the work. Section 4 describes the proposed scheme detailing each of its constituent phases. Section 5 discusses the performance in detail to confirm the supremacy and efficacy of the MLBCT, and finally, Section 6 concludes the work by mentioning the future scope for the same.

## 2. Literature Review

As mentioned in Section 1, the optimization techniques can be majorly categorized as heuristic and metaheuristic schemes. Heuristic techniques utilize the complete set of particulars of a given problem and, being greedy in nature, generate solutions that might get trapped into local maxima/minima instead of producing the global maxima/minima.

On the other hand, metaheuristic techniques, also termed guided random search algorithms, are problem-independent, providing the optimal solution without getting stuck into the local maxima/minima. Metaheuristic algorithms compute the optimal solution by thoroughly exploring and exploiting the available search space in multiple iterations. The general working of the metaheuristic techniques is summarized in Figure 1.

The metaheuristic scheme starts working with a randomly selected set of solution vectors that improve over the iteration. Once the application-specific parameters such as scaling factor and crossover rate are defined, the fitness of the current solution set is evaluated through a carefully designed fitness function. Then, the counter which keeps track of the iterations is initialized. Afterward, a selection from the population chosen is made, and the selected vectors undergo a variation phase (mutation/crossover). Thus, updated vectors are again evaluated for their current fitness, and through a survivor function, a greedy selection strategy, the population for the next generation is finalized. The process of updating the set of solutions is repeated for a predefined number of iteration, and at last, the most recent population is selected as the final solution. An intelligently and carefully designed fitness function plays the most significant role in obtaining further improved offspring in metaheuristic techniques.

Here, we present a brief review of such schemes based on the approaches known as heuristic and metaheuristic.

*2.1. Heuristic Schemes.* In one work [6], the authors proposed the most popular clustering-based routing protocol, Low-Energy Adaptive Clustering Hierarchy (LEACH), for the wireless sensor networks, which features a probabilistic

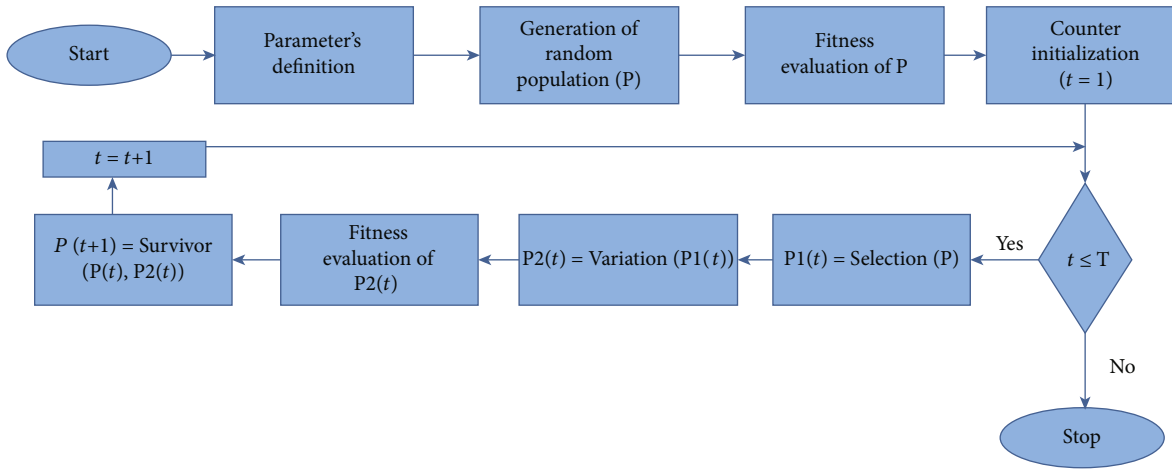



Figure 1: General scheme of metaheuristic techniques.

selection of cluster heads. It implements the localized coordination for various network operations and randomized rotation of the role of the cluster heads for load balancing among the nodes. However, since the selection of cluster heads does not count the residual energy of the nodes, nodes with low residual energy might suffer from early death if frequently selected as cluster heads.

In another work [7], the authors of the LEACH proposed an extension of the [6] requiring the nodes to send their location and energy status to the base station for the selection of cluster heads in a centralized manner and the formation of appropriate clusters via the application of simulated annealing algorithm.

The authors proposed a chain-based scheme in which, instead of forming multiple clusters [8], the nodes were provisioned to develop chains in a way that each sensor could exchange data with the neighbor nodes. At last, the chain leader concludes the entire data flow and forwards it to the base station. However, the scheme proved to be more energy-efficient than LEACH, but the significant delay in the delivery and dynamic topological adjustments appeared as the major issues of the scheme.

In [9], the authors proposed a static clustering scheme that eradicated the energy costing of the dynamic cluster formation in every round of the network operation as in LEACH, etc. In this scheme, distance-based clustering is executed via the base station. Once the clusters are decided, two important parameters—residual energy of the nodes and the nodes' spatial distribution—are considered to select cluster heads. However, the scheme only targeted energy consumption minimization.

In one scheme [10], the authors proposed a centralized scheme that treated coverage in the sensing field as equally important as the energy efficiency. The scheme starts with the distance-based clustering as in the [9]. It selects the cluster heads based on the weighted mean of the contribution factor of the nodes, where the contribution factor is defined as the ratio of the node's residual energy to that of the native grid in the sensing field. The main objective of the scheme is to assure network-wide coverage for the maximum network operation time.

In [11], the authors proposed a LEACH-based clustering protocol that mainly targets the energy efficiency and the fault tolerance in the network. To improve the network lifetime, the network nodes are provisioned to send their data to their respective cluster heads only when the current data is distinct from the previous data. At the end of every network round, noncluster head nodes forward their current energy status to the respective cluster heads to get classified as faulty (nodes with lower residual energy level) and live nodes (nodes with sufficient residual energy). The identification of faulty nodes facilitates the fault tolerance in the network.

In [12], the authors proposed a Fault-Tolerant Clustering-based Multipath algorithm (FTCM) to address the problems of energy efficiency and fault tolerance in the wireless sensor networks. The scheme calls the hybrid energy-efficient distributed clustering (HEED) [13] scheme to partition the network into an appropriate number of clusters. It also appoints a backup CH (BCH) for a cluster head to improve the fault tolerance. The BCH consistently monitors the performance of CH and keeps a copy of CH's data until delivered to the base station. In case of any mishap at the CH end, the BCH can instantly transmit data to the base station without asking the member nodes to send their data again. In addition to the regular responsibilities of CH, the CH is also responsible for the removal of the majority of faulty nodes via hypothesis testing and majority voting. The proposed scheme enables three paths to transfer data from the source node to the base station based on the parameters—residual energy of the nodes, number of hops, propagation speed, and path reliability.

In [14], the authors proposed a clustering-based Hierarchical Fault-Management Framework (HFMF) to address energy management and fault management jointly. For the minimization of energy consumption, the sleep/active method is used. For the management of faults, that is, faults' detection and recovery, backup CH (BCH) is appointed along with every CH to take care of acting CH in the event of its malfunctioning or failure. Later by measuring the data correlation among the cluster members, nodes are grouped virtually to further achieve the energy and fault

management. The authors have successfully demonstrated that the proposed scheme not only manages the transient faults, intermittent faults, and permanent hardware faults but also the link faults are detected.

*2.2. Metaheuristic Schemes.* A wide variety of metaheuristic techniques such as genetic algorithm (GA), genetic programming (GP), evolutionary programming (EP), evolution strategies (ES), differential evolution (DE), particle swarm optimization (PSO), ant colony optimization (ACO), and teaching-learning-based optimization (TLBO) exist in the literature. Such metaheuristic techniques with the virtue of being problem-independent have already imparted a lot in almost every field of engineering like [15]. In the context of wireless sensor networks, some contributions are noticed especially for the selection of cluster heads and the effective formulation of the clusters like in [16–25].

Due to its simplicity, robustness, and fast convergence, differential evolution has proved its worth over the algorithms like GA and PSO [26]. Several contributions have already been proposed based on this outstanding differential evolution technique in search of suitable clusters of the nodes in WSN. This subsection discusses some of the prime contributions in this regard as follows:

In one work [27], a differential evolution-based routing scheme, DE-LEACH, is proposed for environmental monitoring wireless sensor networks. DE-LEACH applies the fast and straightforward converging search technique of differential evolution to produce the clusters by considering the nodes' residual energy status and spatial distribution. The scheme consists of four phases: partitioning initial clusters, collecting status information of the nodes within the clusters through the auxiliary cluster heads, determining optimized cluster heads with differential evolution, and forming optimized clusters. The phases are to be executed in every round of the network operation. The scheme outperforms the traditional LEACH, and LEACH-C [7]. However, the nodes are burdened with heavy computational responsibilities.

In another work [28], a differential evolution-based clustering algorithm (DECA) is proposed, which provisions specialized nodes enriched with the additional amount of initial energy to act as cluster heads. These specialized nodes are called relay nodes or gateways. In DECA, besides providing a suitable fitness function (to measure the health of the tentative clusters), a new local improvement phase has also been proposed that carefully prevents early death of the gateways. DECA utilizes the DE/best/1/bin scheme for the differential evolution. In addition to a novel scheme for the vector representation, a fitness function is designed by considering the standard deviation of the lifetime of gateways and average cluster distance. The scheme outperforms the [29–31] traditional differential evolution and genetic algorithm-based scheme in terms of network lifetime; however, the scheme gives only a little attention to the cluster balancing via its local improvement phase.

A hybrid differential evolution and simulated annealing (DESA) scheme for the improvement of network lifetime in wireless sensor networks is proposed in [32]. The scheme utilizes a hybrid of differential evolution and simulated annealing for local and global optimal solutions, respectively. There are four phases in the scheme—population vector initialization, mutation, crossover, and selection as in the traditional differential evolution. However, instead of using a random selection of population vectors, a more effective, "opposite point method" [33] technique is used for the initialization of population vectors. The mutation scheme is decided randomly at run time based on a chosen threshold value (here, it is 0.5) in such a way that a random number belonging to (0, 1) is observed, and if it is below the threshold, the mutation scheme is DE/rand/1; otherwise, it is DE/target – to – best/1. The fitness function is designed by considering the ratio of nodes' energy to that of the respective clusters. And for crossover, a blending rate based on Gaussian distribution is used. The scheme outperforms the traditional differential evolution scheme in terms of network lifetime, energy consumption, throughput, etc.; however, it converges slowly.

In [34], the authors proposed Multiobjective Load-Balancing Clustering (MLBC) which is a multiobjective optimization technique that addresses two significant problems in WSN—energy efficiency and reliability. It utilizes the Multiobjective Particle Swarm Optimization (MOPSO). MLBC targets energy efficiency by appropriately considering the average residual energy of the cluster heads and reliability by reducing the intercluster communication cost among the nodes in a cluster. It also provisions the load balancing via shuffling the roles of the next-hop node and CH in every iteration. However, it considers only the average residual energy of cluster heads in formulating the objective function for energy efficiency.

In a scheme [35], efficient energy consumption in wireless sensor networks using an improved differential evolution algorithm is highlighted. The scheme is an improvement of [28], in which the mutation strategy has been updated to accommodate the target vector along with the prior best and two random population vectors. Also, the fitness function has been upgraded to accommodate the total energy of the gateways and nodes in addition to the existing network lifetime standard deviation component. However, nothing has been mentioned concerning the load balancing among the clusters.

In one work [36], the authors proposed a hybrid metaheuristic clustering algorithm that exploits the best of Artificial Bee Colony and differential evolution optimization techniques. In their proposed Artificial Bee Colony (ABC) with differential evolution (DE) scheme, known as ABC-DE-based clustering scheme, the objective function is designed by taking into account the three network parameters—average intracluster distance, average energy of cluster heads, and data transmission delay to ensure the load-balanced cluster heads. In addition to this, an ABC-based metaheuristic algorithm has also been proposed to facilitate the dynamic repositioning of the mobile sink within the cluster-based network to achieve further energy efficiency.

In [37], the authors have addressed the problem of energy optimization in an Internet-of-Thing-based WSN (IoT-based WSN). In pursuance of the problem, as mentioned earlier, a hybrid of the Whale Optimization

Algorithm (WOA) and simulated annealing (SA) metaheuristic algorithms have been employed to select the most suitable cluster heads in their respective clusters. For choosing the most appropriate cluster heads, the fitness function of the proposed scheme considers a set of five node-specific parameters: residual energy, load, delay, distance, and temperature. The fitness function ensures that the node with the highest residual energy but the least load, delay, distance, and temperature is selected as the cluster heads in every network round.

In one work [38], the authors proposed an Artificial Intelligence- (AI-) based quorum system to address the issue of energy conservation in the wireless sensor networks. The primary motivation behind the proposed AI-based was to fasten the neighbor discovery process in order to minimize the network latency. Moreover, the scheme facilitates a quorum-based grid system that allows a substantial increase in the number of nodes in the quorum without mandating the increase in the number of quorums to reduce the effective network delay. In addition to the aforesaid, the feature of weighted load balancing reduces the network energy consumption to improve the network lifetime. Through the various experimentation, the authors have established the outperformance of their proposed scheme over the state-of-the-art quorum algorithms in terms of latency, improved coverage, energy efficiency, and network lifetime.

In [39], the authors proposed a genetic algorithm- (GA-) inspired clustering-based approach to address the problem of node's localization in wireless sensor networks. To find the accurate position of unknown nodes with respect to the anchors or known nodes, the authors used the Euclidean distance objective function in their proposed scheme. Through various simulation results, the supremacy of the GA-based localization scheme with an extended clustering approach has been established over the state-of-the-art schemes like Centroid and Distance Vector-Hop (DV-Hop) in terms of improved location accuracy.

In a scheme [40], the author proposed a genetic algorithm-based energy-efficient clustering scheme which addressed the localization problems in wireless sensor networks. The authors utilized parameters like node's residual energy, distance estimation, and coverage connection in the formulation of fitness function for their proposed scheme, Energy-Efficient Clustering in Genetic Algorithm Localization (EECGL). Through various experimentation, the authors have shown that EECGL approximates the unknown node's location with the least localization error and extends the effective network lifetime by minimizing the overall network energy consumption.

In a work [41], the authors proposed a metaheuristic energy-efficient clustering technique which is inspired by the Brain Storm Optimization (BSO). The BSO is a swarm-based metaheuristic technique exploiting the human brainstorming process in search of the best possible solutions. In their proposed scheme, Energy-Efficient Clustering-Brain Storm Optimization (EEC-BSO), the authors have focused on deciding energy-efficient clusters in a way that nonparticipating nodes in the information transmission process are sent to sleep mode minimizing the overall network consumption. In the formulation of such clusters, the fitness function is designed by considering the parameters like node's residual energy, coverage, and packet data rate. Moreover, the outperformance of EEC-BSO has been established over the state-of-the-art schemes such as LEACH, LEACH-Centralized, Energy-Efficient Clustering Scheme (EECS), and LEACH-BSO in terms of reduced energy consumption, improved coverage, and data packet rate.

In a proposed scheme [42], a differential evolution-based clustering routing protocol (DEBCRP) for wireless sensor networks. DEBCRP is a base station-dependent scheme that applies DE/best/1/bin scheme for the network partitioning into some clusters. The fitness function devised by the authors considers the nodes' residual energies with respect to the probable cluster heads and the distance between the nodes and the cluster heads for the formulation of clusters. At last, to communicate the data from the sensing field to the base station, a PEGASIS [8] like a chain of the cluster heads is formed. The scheme DEBCRP is reported to outperform the S-DE [43] in terms of network lifetime. However, no adequate consideration is given for the formulation of load-balanced clusters, which is the most prime key to network lifetime improvement. Also, PEGASIS like chain of the cluster heads suffers from similar problems as in [8], for example, delayed communication, and since data from one CH is to be aggregated with that of the others in the direction to the sink, there might be introduced some inaccuracy in the information being sent to the base station.

From the aforementioned analysis, it can be easily concluded that despite being the most important factor for the formulation of clusters in the network, cluster balancing has been addressed the least. Thus, the work being presented here serves the following objectives:

(i) Balanced cluster formulation to contribute effectively towards the enhancement of network lifetime

(ii) Adaptable clustering solution to perform consistently well in any network configuration

## 3. Preliminaries

This section describes the network model for the scheme. In addition to this, it also discusses the basics of the differential evolution metaheuristic technique and the entire set of notations used throughout the work.

*3.1. Network Model.* MLBCT assumes the wireless sensor network with the following characteristics:

(1) All the sensor nodes are deployed randomly across the sensing field and are static. More illustratively, nodes once deployed cannot change their location

(2) The sensor nodes are homogeneous and equipped with a definite amount of initial energy

(3) The sensor nodes are facilitated with the power control features to introduce variations in the transmission power as and when needed

(4) The base station is also static and can be placed at any point in the network accordingly

(5) The continuous data flow model is used here to define the working mode of the sensor nodes

*3.2. Differential Evolution: An Overview.* The differential evolution has evolved as a prevalent stochastic metaheuristic multimodal optimization technique over the continuous search space. Similar to the general scheme of metaheuristic techniques as discussed in Section 1, it starts with the definition of the initial parameters where the values of scaling factor and crossover rate are defined along with the randomized set of initial solutions (initial population) and the number of iterations. Here, each solution vector (equivalently known as chromosome or genome) termed as a target vector undergoes the mutation phase followed by the recombination. This mutation followed by the recombination is nothing but the variation phase of Figure 1. As depicted in Figure 2, the target vector, once it passes through the mutation phase, becomes the donor/mutant vector. After the recombination or crossover phase, the donor vector is known as the trial vector.

In the differential evolution scheme, obtainment of the next-generation solutions is performed only after the generation of all trail vectors when compared to particle swarm optimization, and teaching-learning-based optimization [44, 45]. In other words, the greedy selection towards the next-generation solution is performed between the pair of target and trial vectors once all the target vectors have been converted into trial vectors. A variety of mutation strategies exist, such as random, best, and target-to-best, along with the two types of crossover techniques—binomial and exponential crossovers. The binomial and exponential crossover can be defined as follows:

*3.2.1. Binomial Crossover.*

$$u_j = \begin{cases} v_j & \text{if } r \le C_{\mathrm{p}} \text{ OR } j = \delta, \\ x_j & \text{if } r > C_{\mathrm{p}} \text{ AND } j \ne \delta, \end{cases} \tag{1}$$

where $C_{\mathrm{p}}$ is the crossover probability, $\delta$ is the randomly selected variable location from the set $\{1, 2, 3, \cdots, |\text{decision variable}|\}$, $r$ is the random number between 0 and 1, $u_j$ refers to the $j^{\text{th}}$ variable of the trial vector, $v_j$ refers to the $j^{\text{th}}$ variable of donor/mutant vector, and $x_j$ refers to the $j^{\text{th}}$ variable of the target vector.

*3.2.2. Exponential Crossover.* In the exponential crossover, at very first, the $n^{\text{th}}$ variable from the donor vector is copied into the trial vector. Afterward, every subsequent variable from the donor vector is copied into the trial vector as long as the $r \le C_{\mathrm{p}}$. Once $r > C_{\mathrm{p}}$, variables from the target vector are copied into the trial vector.

Based on the adapted mutation strategy and crossover type, various schemes have been proposed for differential evolution, and to discriminate among them, a standard notation, **DE/x/y/z**, is used. Here, **DE** refers to the differential evolution,

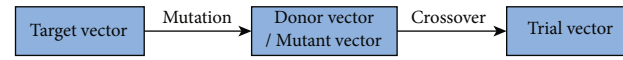



Figure 2: Vector transformation in differential evolution.

**x** denotes the mutation strategy, **y** denotes the number of difference vectors to be used in the mutation operation, and **z** refers to the crossover scheme selected. Some of the variants of the DE schemes are listed here in Table 1.

Here, in Table 1, $V$ is the donor vector, $F$ is the scaling factor such that $F \in (0, 2)$, $X_{\text{best}}$ is the target vector with best fitness value, $X_i$ is the $i^{\text{th}}$ target vector, and $X_{r_j}$ is the $j^{\text{th}}$ target vector chosen randomly where $j \in [1, N]$, $N$ being the number of target vectors in the population. Once the trail vectors are generated for all the target vectors of current generation, say $G$, offsprings are chosen based on the fitness value of the corresponding pairs of target and trial vectors, i.e., $<X_{i,G}, U_{i,G}>$ for $i \in [1, N]$ as follows:

$$X_{i,G+1} = \begin{cases} U_i & \text{if fitness}(U_{i,G}) \ge \text{fitness}(X_{i,G}), \\ X_i & \text{otherwise.} \end{cases} \tag{2}$$

*3.3. Terminology.* The notations used throughout the work have been listed as follows:

(i) $S$ denotes the set of sensor nodes such that $S = \{s_1, s_2, s_3, \cdots, s_N\}$ where $N$ is the number of nodes deployed in the sensing field

(ii) $\Theta$ denotes the set of cluster heads such that $\Theta = \{\mathrm{CH}_1, \mathrm{CH}_2, \mathrm{CH}_3, \cdots, \mathrm{CH}_k\}$ where $k$ is the number of cluster heads

(iii) $\mathrm{RE}_i$ denotes the residual energy of the $i^{\text{th}}$ node in the network

(iv) $\mathrm{Cluster}^i_{\mathrm{RE}}$ denotes the residual energy of the $i^{\text{th}}$ cluster such that $\mathrm{Cluster}^i_{\mathrm{RE}} = \sum_{j=1}^{m} \mathrm{RE}_j$ where $m$ refers to the cluster size

(v) $\mathrm{CS}_i$ denotes the cluster size of the $i^{\text{th}}$ cluster

(vi) AvgCS refers to the average cluster size, i.e., average number of nodes in a cluster

(vii) ACE refers to the average cluster energy such that $\mathrm{ACE} = \sum_{i=1}^{N} \mathrm{RE}_i / k$

(viii) $d(i, j)$ denotes the Euclidean distance between the $i^{\text{th}}$ and $j^{\text{th}}$ nodes in the network

(ix) $\mathrm{dist}_m(i, j)$ denotes the Euclidean distance between the $i^{\text{th}}$ and $j^{\text{th}}$ members of the $m^{\text{th}}$ cluster. This parameter is basically used to measure the nodes' proximity

(x) $R_{\mathrm{C}}$ denotes the communication range of the nodes

(xi) $\mathrm{ComCH}(s_i)$ refers to the set of cluster heads within the communication range of the node $s_i$, i.e., $\mathrm{ComCH}(s_i) = \{\mathrm{CH}_j \mid d(s_i, \mathrm{CH}_j) \le R_{\mathrm{C}}\}$

Table 1: Differential evolution schemes.

| DE scheme | Mutation strategy | Mutation expression | Crossover type |
|---|---|---|---|
| DE/rand/1/bin | Random | $V = X_{r_1} + F(X_{r_2} - X_{r_3})$ | Binomial |
| DE/rand/2/exp | Random | $V = X_{r_1} + F(X_{r_2} - X_{r_3}) + F(X_{r_4} - X_{r_5})$ | Exponential |
| DE/best/1/bin | Best | $V = X_{\text{best}} + F(X_{r_1} - X_{r_2})$ | Binomial |
| DE/best/2/bin | Best | $V = X_{\text{best}} + F(X_{r_1} - X_{r_2}) + F(X_{r_3} - X_{r_4})$ | Binomial |
| DE/target-to-best/1/exp | Target-to-best | $V = X_i + F(X_{\text{best}} - X_i) + F(X_{r_1} - X_{r_2})$ | Exponential |
| DE/target-to-best/2/exp | Target-to-best | $V = X_i + F(X_{\text{best}} - X_i) + F(X_{r_1} - X_{r_2}) + F(X_{r_3} - X_{r_4})$ | Exponential |

The main objective of the present work is to formulate the balanced clusters within the network for the even distribution of load among the nodes. To ensure this, it is attempted that the clusters are equipped with an almost similar count of member nodes situated close to one another. Also, the clusters are left with an approximately equal amount of residual energy at the end of every network round.

## 4. Proposed Scheme: Metaheuristic Load-Balancing-Based Clustering Technique (MLBCT)

This section describes the proposed scheme, Metaheuristic Load-Balancing-Based Clustering Technique (MLBCT) in wireless sensor network. The MLBCT is a base station-(BS-) assisted scheme which calls the BS for the differential evolution-based cluster formation. Once the optimized and balanced clusters come into existence, it hands over the responsibility of further network operations to the network nodes.

The scheme starts with a bootstrapping phase in which all the nodes are assigned unique IDs, which in turn communicate their IDs and location information to the BS. The BS then applies the differential evolution with a well-established fitness function (detailed below) and formulates the balanced clusters. The selected cluster heads are then informed of their specific roles and their members' information by the base station. Thus, selected cluster heads then provide their IDs to the respective members along with the TDMA schedules. Afterward, the overall network operation is divided into rounds where each round consists of the steady-state phase and the responsible node selection phase. In the steady-state phase, cluster members send their data to their respective cluster heads, which aggregate the received data and forward it to the base station. In the responsible node selection phase, the current cluster head in a cluster, select a node randomly to act as head for the next round and broadcast into the concerned cluster. The entire workflow is summarized in Figure 3 and has been detailed into the subsequent subsections and algorithm as follows:

*4.1. Bootstrapping.* In bootstrapping, differential evolution is applied by the base station to divide the entire network into $k$ number of balanced clusters where $k$ is a user-defined parameter. It starts with the sharing of node-specific information such as identity, residual energy, and location information to the base station by the nodes deployed. Based on the information received, BS performs the following to determine the required partitioning.

*4.1.1. Generation of the Random Population.* The population vectors are generated as per the [28]. Each population vector is chosen in such a way that it indicates the assignment of every network node to one of the cluster heads. The notation adopted to represent the $i^{\text{th}}$ population vector of the $G^{\text{th}}$ generation is as follows:

$$\vec{X}_{i,G} = [x_{1,i,G}, x_{2,i,G}, x_{3,i,G}, \cdots, x_{N,i,G}], \tag{3}$$

where $x_{1,i,G}, x_{2,i,G}, x_{3,i,G}, \cdots, x_{N,i,G}$ are the random numbers between 0 and 1. $x_{j,i,G}$ denotes the assignment of the node $s_j$ to one of cluster heads, say $k$, as follows:

$$l = \text{ceiling}\left(x_{j,i,G} * \left|\text{ComCH}\left(s_j\right)\right|\right), \tag{4}$$

$$CH_k = \text{index}\left(\text{ComCH}\left(s_j\right), l\right). \tag{5}$$

Here, the length of the population vectors is definite and determined by the number of nodes deployed in the field.

Thus, corresponding to every population vector, say $\vec{X}_{i,G} = [x_{1,i,G}, x_{2,i,G}, x_{3,i,G}, \cdots, x_{N,i,G}]$, we have another vector, say $\vec{Y}_{i,G} = [y_{1,i,G}, y_{2,i,G}, y_{3,i,G}, \cdots, y_{N,i,G}]$ such that

$$f\left(x_{j,i,G}\right) = y_{k,i,G}, \tag{6}$$

where $y_k \in \Theta$ is assigned to the node $x_j$ in the $i^{\text{th}}$ vector of $G^{\text{th}}$ generation as per equations (4) and (5).

*4.1.2. Fitness Function.* It can be easily intuited that if the clusters are balanced in the clustered network architecture, they might have an almost similar level of residual energy and a similar count of member nodes. With this conception, to meet our primary objective of network partitioning into some balanced clusters, nodes' residual energy and cluster size have been taken as the decision parameters. In addition to this, nodes' proximity has also been taken into account, ensuring the reduced energy consumption in intracluster communication.

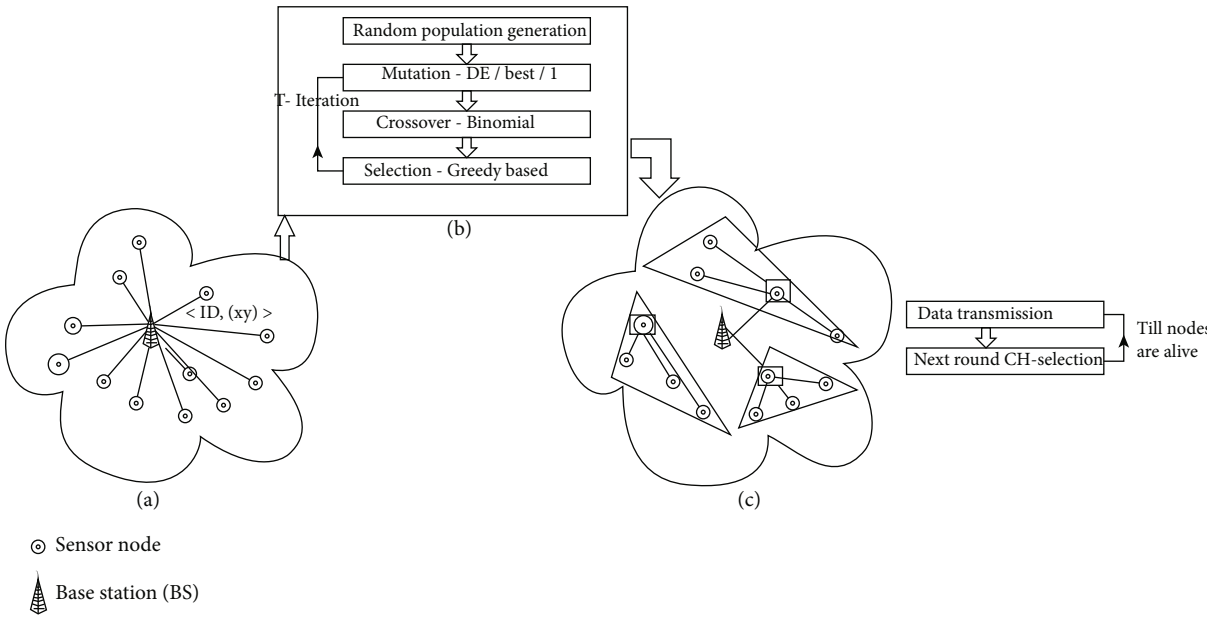



Figure 3: MLBCT: operation.

A suitable fitness function always contributes the most to the differential evolution to converge. Thus, the fitness function has been derived in such a way that it characterizes all the aforementioned requirements as follows:

(i) Standard deviation of average cluster energy

If the clusters have been formed in an optimized way, ensuring the entire network energy is distributed evenly across the clusters formed in the network, each cluster is supposed to have an almost similar level of residual energy. In other words, it can be said that in terms of average cluster energy (ACE), each cluster should have the approximately same amount of energy, and hence, the standard deviation accords to the following:

$$\sigma_{\mathrm{CE}} = \sqrt{\frac{1}{k} * \sum_{i=1}^{k} \left(\mathrm{ACE} - \mathrm{Cluster}_{\mathrm{RE}}^{i}\right)^2}, \tag{7}$$

where $k$ is the number of clusters. It is quite obvious that the lower the value of $\sigma_{\mathrm{CE}}$, the higher the value of fitness, i.e.,

$$\text{Fitness Value} \propto \frac{1}{\sigma_{\mathrm{CE}}}. \tag{8}$$

(ii) Standard deviation of average cluster size

The balanced clusters must have an approximately equal number of members. In other words, it can be said that the average cluster size (AvgCS) of each cluster should have the almost same count of cluster members.

With this, the standard deviation and the fitness value accord to equations (9) and (10), respectively.

$$\sigma_{\mathrm{CS}} = \sqrt{\frac{1}{k} * \sum_{i=1}^{k} (\mathrm{AvgCS} - \mathrm{CS}_i)^2}, \tag{9}$$

where $k$ is the number of clusters. It can be intuited again that the lower the value of $\sigma_{\mathrm{CS}}$, the higher the value of fitness, i.e.,

$$\text{Fitness Value} \propto \frac{1}{\sigma_{\mathrm{CS}}}. \tag{10}$$

(iii) Nodes' proximity within the cluster

This is the metric that ensures that when there comes to decide on the nodes to be a part of a cluster, the one who is located at a shorter distance from the other members gets priority. The central idea behind having this metric is to reduce the cost of communication within the cluster. The lower the value of this metric, the higher the value of fitness. More illustratively,

$$\text{Fitness Value} \propto \frac{1}{\sum_{m=1}^{k} \mathrm{dist}_m(i,j)}. \tag{11}$$

From equations (8), (10), and (11), we can have the following:

$$\text{Fitness Value} \propto \frac{1}{\sigma_{\mathrm{CE}}} * \frac{1}{\sigma_{\mathrm{CS}}} * \frac{1}{\sum_{m=1}^{k} \mathrm{dist}_m(i,j)}, \tag{12}$$

i.e.,

$$\text{Fitness Value} = \frac{K}{\sigma_{\text{CE}} * \sigma_{\text{CS}} * \sum_{m=1}^{k} \text{dist}_m(i,j)}, \tag{13}$$

where "$K$" is proportionality constant which can be set as $K = 1$ without loss of generality.

And, hence,

$$\text{Fitness Value} = \frac{1}{\sqrt{1/k * \sum_{i=1}^{k} \left(\text{ACE} - \text{Cluster}_{\text{RE}}^{i}\right)^2} * \sqrt{1/k * \sum_{i=1}^{k} (\text{AvgCS} - \text{CS}_i)^2} * \sum_{m=1}^{k} \text{dist}_m(i,j)} \tag{14}$$

or

$$\text{Fitness Value} = \frac{k}{\sqrt{\sum_{i=1}^{k} \left(\text{ACE} - \text{Cluster}_{\text{RE}}^{i}\right)^2} * \sqrt{\sum_{i=1}^{k} (\text{AvgCS} - \text{CS}_i)^2} * \sum_{m=1}^{k} \text{dist}_m(i,j)}. \tag{15}$$

*4.1.3. Mutation Strategy.* Like in [28, 42], **DE/best/1/bin** scheme is adapted here in this work which refers to the application of the **DE/best/1** mutation strategy. As depicted in Figure 2, each target vector of the population (say, of the size $P$) will go through this scheme to get transformed into a donor vector. From Table 1, the mutation expression for the selected strategy is

$$\vec{V}_{i,G} = \vec{X}_{\text{best},G} + F\left(\vec{X}_{r_1,G} - \vec{X}_{r_2,G}\right), \tag{16}$$

where $\vec{X}_{\text{best},G}$ and $\vec{X}_{r_1,G}, \vec{X}_{r_2,G}$ refer to the best vector, and any two randomly selected vectors from the $G^{\text{th}}$ generation of the population such that $i$, $r_1$, and $r_2$ are the three random integers $\in [1, P]$ and $i \neq r_1 \neq r_2$, respectively. $F$ is the scaling factor that may have any value between $(0, 2)$.

From equation (3), it is quite obvious that the components of the vectors in equation (16)—$\vec{X}_{\text{best},G}$, $\vec{X}_{r_1,G}$, and $\vec{X}_{r_2,G}$—are the random values $\in (0, 1)$. In order to ensure that the components of the vector $\vec{V}_{i,G}$ are also the values $\in (0, 1)$, a few amendments are being introduced as in [28].

Let

$$\vec{D}_{i,G} = \vec{X}_{r_1,G} - \vec{X}_{r_2,G}, \tag{17}$$

then,

$$d_{j,i,G} = \begin{cases} 1 + \left(x_{j,r_1,G} - x_{j,r_2,G}\right), & \text{if} \quad \left(x_{j,r_1,G} - x_{j,r_2,G}\right) \leq 0, \\ \left(x_{j,r_1,G} - x_{j,r_2,G}\right), & \text{otherwise.} \end{cases} \tag{18}$$

Also, for the computation of $v_{j,i,G}$ contributing to $\vec{V}_{i,G}$, the following can be referred to

$$v_{j,i,G} = \begin{cases} \left(x_{j,\text{best},G} + F * d_{j,i,G}\right) - 1, & \text{if} \quad \left(x_{j,\text{best},G} + F * d_{j,i,G}\right) > 1, \\ x_{j,\text{best},G} + F * d_{j,i,G}, & \text{otherwise.} \end{cases} \tag{19}$$

*4.1.4. Crossover Scheme.* The crossover schemes in terms of the binomial and exponential crossover are already described in Section 3. A binomial crossover scheme is used in this work to convert the donor vector into the trial vector.

*4.1.5. Selection or Offspring Generation.* Once all the trial vectors are generated following the above-mentioned steps, the next generation can be obtained on basis of the comparison of fitness values of the corresponding pair of target and trial vectors as given in

$$\vec{X}_{i,G+1} = \begin{cases} \vec{U}_{i,G} & \text{if} \quad \text{fitness}\left(\vec{U}_{i,G}\right) \geq \text{fitness}\left(\vec{X}_{i,G}\right), \\ \vec{X}_{i,G}, & \text{otherwise.} \end{cases} \tag{20}$$

*4.1.6. Complexity Analysis.* Throughout the proposed scheme, fitness function would be evaluated for $N_P + N_P * T$ times where $N_P$ refers to the size of population and $T$ refers to the number of iterations known a priori.

Moreover, exploiting solution space in search of the most optimal solution is a continuous process in the metaheuristic scheme. For this reason, even in the best case, the complexity of the fitness function will be $O(n^2)$ as each newly generated solution has to be compared with its predecessor in terms of its fitness value. Similarly, complexity of the fitness function in the worst case will be $O(n^2)$ due to successive fitness value computation and comparison. Thus, the average-case complexity for the fitness function can be concluded as $O(n^2)$.

As explained at the beginning of this section, once the clusters are formed, and members are notified of their respective initial heads, further network operations can be

divided into two rounds—the steady-state phase and the responsible node selection phase.

*4.2. Steady-State Phase.* This phase refers to the data transmission in which cluster members send their data to their respective cluster heads in the designated time slots. After receiving the data from its members, cluster heads aggregate the collected data and forward it to the base station on behalf of their entire cluster.

*4.3. Responsible Node Selection Phase.* After executing the steady-state phase, a cluster head in its respective cluster selects a node randomly as the head for the next round and communicates the same to its members. The members note the same and communicate their data to that newly selected cluster head in the upcoming round accordingly. The process is carried out in each of the clusters in the network.

## 5. Performance Analysis

This section deals with the various experimental processes conducted throughout the work and analyses the obtained results thoroughly.

*5.1. Experimental Environment.* In conducting the experiments, different network configurations with varying node densities have been examined. More illustratively, experiments have been performed with the different number of nodes, say 50, 100, 150, and 200 in an area of $100 \times 100\,m^2$ with two different sink placements—one at the center of the sensing field (50 m, 50 m) and another beyond the network precisely at (50 m, 150 m). An instance of clustering with 50 nodes and 5 and 10 cluster heads, respectively, is demonstrated in Figure 4. The base station is situated at (50 m, 150 m) in this exemplary instance.

An extensive set of experiments have been performed for the proposed scheme using MATLAB.

Mainly, the experiments have been performed to

(1) Prove the efficacy of the proposed fitness function

In this set of experiments, the proposed fitness function as in equation (15) has been tested for the quality of clusters being produced. It has been verified that the proposed fitness function yields balanced clusters in terms of cluster size. The clusters generated as per equation (15) have been compared with the clusters produced by the fitness function given in [42] under two different clustering scenarios. The network is divided into 5 clusters and 10 clusters, respectively.

(2) Prove the supremacy of the proposed scheme, MLBCT in terms of network lifetime and network stability

In the second set of experiments, the performance of MLBCT is compared to that of DEBCRP [42] and improved differential evolution-LEACH (ImDE-LEACH) [46], majorly in terms of network lifetime and network stability with respect to the number of alive nodes in the network, network energy consumption, average residual energy per network nodes over the network rounds, and data packets delivered to the base station under the variable network configurations. Moreover, for the sake of experimentation, the performance of the LEACH [6] has also been recorded into the same context as that of MLBCT, DEBCRP, and ImDE-LEACH.

*5.2. Simulation Parameters.* To compare the performance of the proposed scheme, MLBCT, with that of DEBCRP and ImDE-LEACH, simulation parameters have been adopted here as listed in Table 2. However, to prove the scalability and adaptability of the proposed scheme, the performance has also been tested under variable network configurations.

In addition to the parameters listed in Table 2, the following performance criteria have been used for the evaluation of schemes:

(i) Network lifetime: the network lifetime is generally measured as the time when the first node dies, or when the last node dies in the network [28–31, 42]. In this work, both definitions have been considered to demonstrate the supremacy of the MLBCT over DEBCRP, and ImDE-LEACH

(ii) Network stability: network stability refers to how smoothly the network operations are going on. It can be measured in terms of the rate of the network energy consumption and the average residual energy per network node. The lower the rate of energy consumption, the more stable the network is, resulting in improved network lifetime. Similarly, the higher the value of average residual energy per network node, the more stable and durable the network is

To further compare the performance of the schemes—MLBCT, DEBCRP, and ImDE-LEACH, packet delivery at the base station can also be considered as a criterion.

The success in this regard can be judged by the higher number of successfully delivered packets to the base station.

To find the energy consumption by the nodes in the network operation, the widely adopted first-order radio model [13, 28, 42, 46–52] has been used here in this work.

*5.3. Results and Discussion.* As stated in point 1 of Section 5.1, the suitability of the proposed fitness function equation (15) is manifested in the first set of experiments. Since the scheme is a metaheuristic one, a suitable fitness function might contribute a lot to decide the best possible clusters. The main objective of this work is to formulate the clusters which are balanced in the sense that the clusters are having an almost similar count of member nodes and the member nodes are located close to one another to have minimized intracluster communication.

In this experimentation, variable node counts as in Table 2 have been considered for two instances of clustering such as 5 clusters and 10 clusters as shown in Figure 5.

The success of the fitness proposal mentioned above is evident in Figure 5. When implemented in the scheme DEBCRP, the proposed fitness function has been found more effective in having more balanced clusters. In other

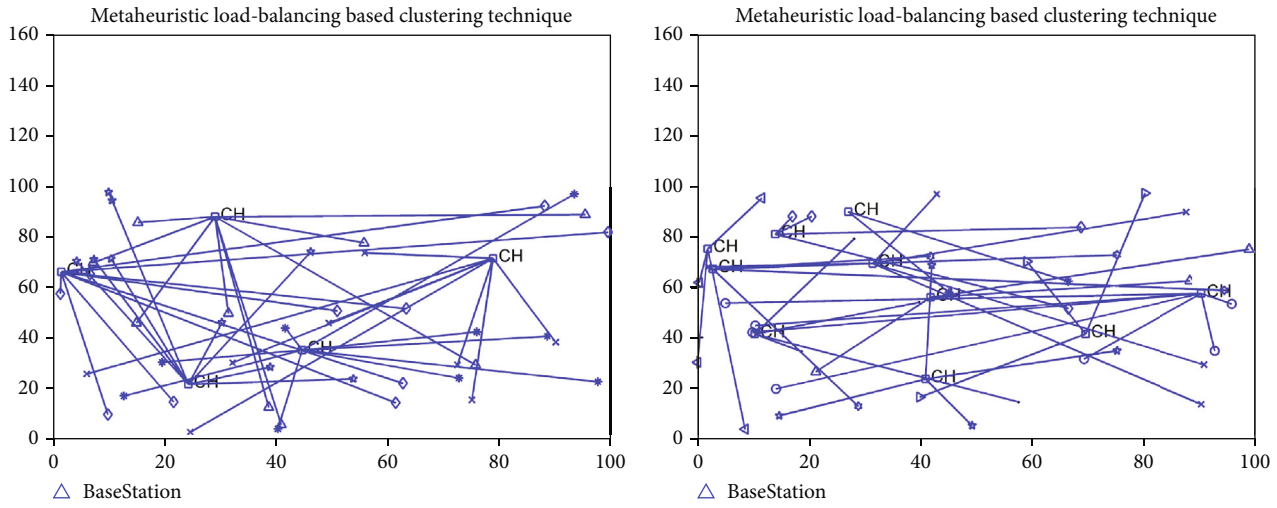
**Input:**
∗ N: No. of randomly deployed sensor nodes.
$*ff^n()$: Fitness function.
∗ F: Mutation/Scaling factor.
∗ T: No. of iteration
∗ k: No. of user-specified clusters
$*C_r$: Crossover rate
◊ BEGIN
%% BOOTSTRAPPING PHASE %%.
◊ for i ⟵1: N
◊ Status Transmission($Node_i \longrightarrow$ BS)
◊ end for
◊ Random population generation (P) where each vector ($X_i$) refers to the complete assignment of all the nodes to the $k$ cluster heads
◊ for i ⟵1: size(P)
◊ $ff^n(X_i)$
◊ end for
◊ for i ⟵1: T
◊ for j ⟵1: size(P)
◊ $V_j = X_{best} + F(X_{r_1} - X_{r_2})$
%% $V_j$ is the $j^{th}$ donor vector
◊ $U_j = [u_j^l]$ where $u_j^l$ is the $l^{th}$ component of $U_j$ defined as follows:

$$u_j^l = \begin{cases} v_j^l & \text{if } r \le C_r \text{ OR } l = \delta \\ x_j^l & \text{if } r > C_r \text{ AND } l \ne \delta \end{cases}$$

%% $U_j$ is the $j^{th}$ trial vector
◊ end for (inner loop)
◊ for j ⟵1: size(P)
◊ $ff^n(U_j)$
%% i.e. fitness function evaluation of the $j^{th}$ trial vector
◊ if ($ff^n(U_j) > ff^n(X_j)$)
◊ Update P
%% Greedy approach for the update of population vector
◊ end if
◊ end for (inner loop)
◊ end for (outer loop)
%% STEADY STATE PHASE %%
◊ while(nodes are alive)
◊ for i ⟵1:
%% i.e. for every cluster
◊ for j ⟵1:
%% m ⟶ no. of members in the $i^{th}$ cluster
◊ DataTransmission($Node_j^i \longrightarrow CH^i$)
◊ end for (inner loop)
◊ DataTransmission($CH^i \longrightarrow BS$)
%% here, aggregated data by the $CH^i$ to the base station
◊ end for (outer loop)
%% RESPONSIBLE NODE SELECTION PHASE %%
◊ for i ⟵1: k
◊ Random selection of $CH_{next}^i$ from within the $i^{th}$ cluster by $CH^i$.
◊ New CH's Information dissemination by the $CH^i$
◊ end for
◊ end while
◊ END

Algorithm 1: MLBCT.

words, clusters are obtained with an approximately similar count of member nodes, leading to the even distribution of load throughout the network nodes. In Figure 5(a), the efficacy of the proposed scheme is demonstrated with five clusters being formed in the network, whereas Figure 5(b) presents the same while partitioning the network into 10 clusters. It can be easily observed from the figure that the members recorded in the clusters do not vary to the extent




(a) Clustering instance with 5 clusters

(b) Clustering instance with 10 clusters

Figure 4: Simulation interface for network operation.

Table 2: Parameters used in the simulation.

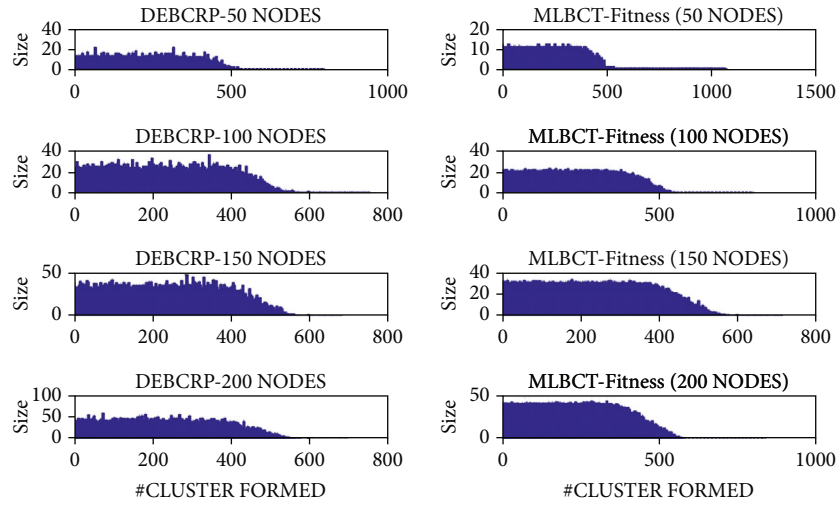
| Parameter | Parameter's value |
|---|---|
| Network area | $100 \times 100\ m^2$ |
| Base station's position | $\{(50\ m, 50\ m), (50\ m, 150\ m)\}$ |
| Node deployment strategy | Random deployment |
| Number of nodes deployed in the network | {50,100,150,200} |
| Initial energy of the normal nodes | 0.1 J |
| Size of data packet | 4000 bits |
| Size of data packet header | 200 bits |
| Energy consumed in data aggregation ($\varepsilon_{da}$) | 5 nJ/bits/signal |
| Energy consumed in the transceivers' circuitry ($E_{elec}$) | 50 nJ/bit |
| Amplification factor in free space model ($\varepsilon_{fs}$) | $10\ pJ/bit/m^2$ |
| Amplification factor in multipath fading model ($\varepsilon_{mp}$) | $0.0013\ pJ/bit/m^4$ |
| Population size | 10 |
| Mutation factor | 0.5 |
| Crossover rate | 0.7 |

as it is there in DEBCRP over the network rounds. Also, it has been verified that the scheme for the fitness evaluation of the clusters works invariably well irrespective of node density present in the network.

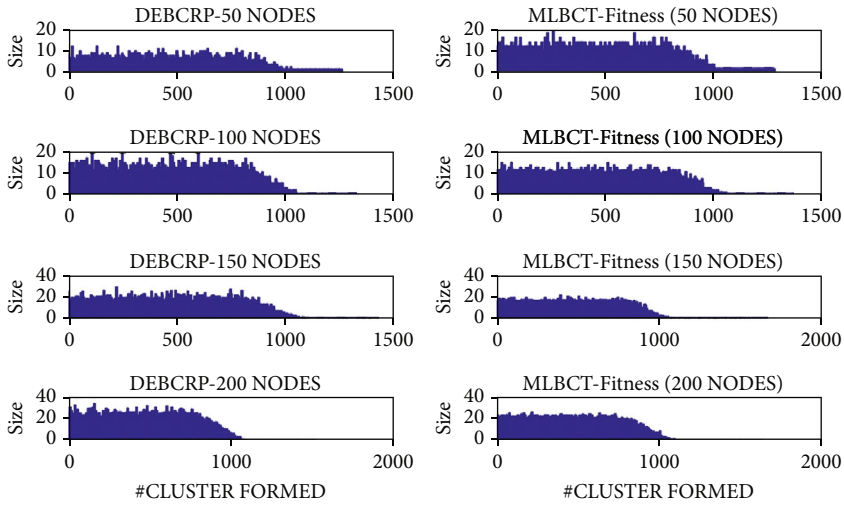
*5.3.1. Statistical Analysis.* Statistical analysis is performed to further explain the efficacy of the proposed fitness function (MLBCT-fitness) as in equation (15) in producing the balanced clusters. This is done by finding out the standard deviation of average cluster size, $\sigma_{CS}$ following equation (9) along with the confidence interval. Standard deviation is defined as the measurement of how the clusters being produced deviate from the ideal distribution of the nodes among the specified number of clusters. The ideal distribution refers to the clusters with $(N/k)$ nodes if $N$ nodes are to be distributed among $k$ clusters.

For this very purpose, as explained above, the proposed fitness function is fitted into the scheme of DEBCRP, and the performance of such a modified scheme is compared with that of DEBCRP with respect to the formation of clusters. This is achieved by recording the clusters' length in both cases until the first node dies. Afterward, standard deviations of the average cluster size are measured in both of the cases—with its own fitness function ($\sigma_{D-\text{Fitness}}$) and MLBCT-fitness function ($\sigma_{M-\text{Fitness}}$).

Figures 6(a) and 6(b) demonstrate the standard deviations of the average cluster size for the different network deployments with 50, 100, 150, and 200 nodes with the requirements mentioned above of having 5 clusters and 10 clusters, respectively. It can be explicitly observed that the standard deviations and the MLBCT-fitness function are quite low compared to the standard deviations obtained via




(a) Cluster formation with the 5 clusters



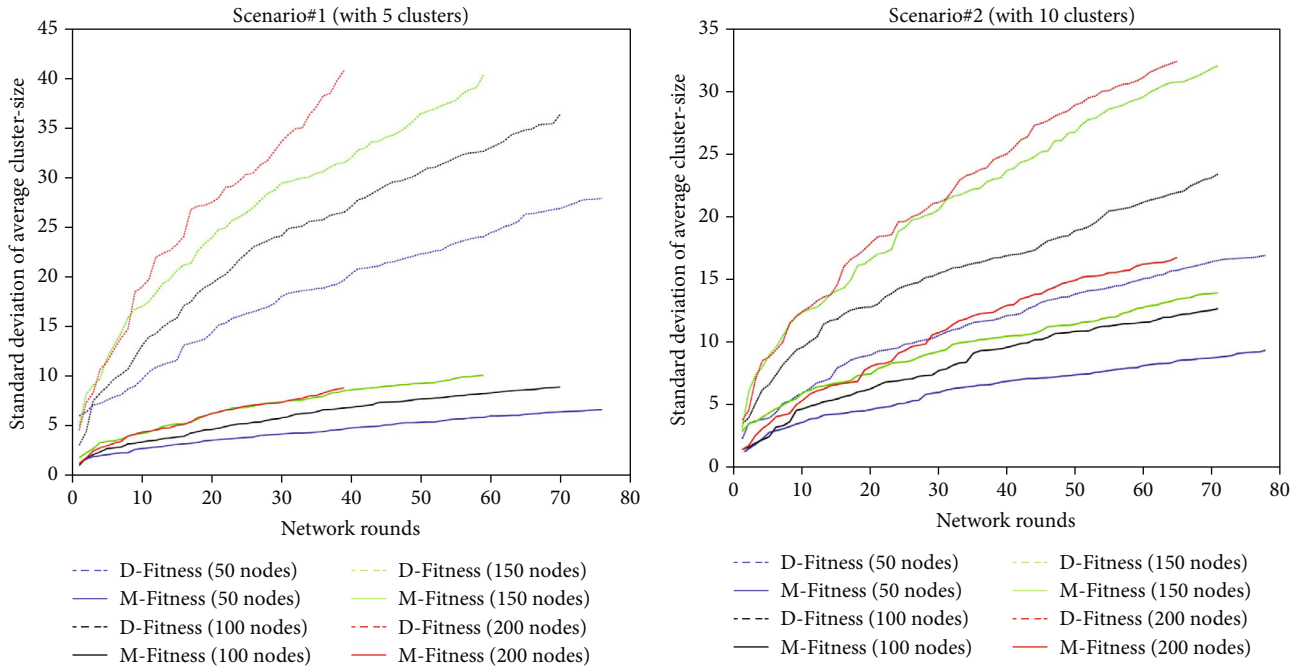

(b) Cluster formation with the 10 clusters

Figure 5: Efficacy of the proposed fitness function.

the application of the DEBCRP-fitness function for all the node deployments under both the specified requirements of 5 clusters and 10 clusters. This also justifies the efficacy of the scheme.

Another statistical analysis known as confidence interval justifies the probability of the deployment of the nodes within a range of the values of the cluster. In this case, the confidence intervals with the confidence levels 95% and 99%, respectively, are measured for both cases of the clustering scenarios with variable node counts. Table 3 clearly explains the efficacy of the MLBCT-fitness function over the fitness function used in DEBCRP in every possible network configuration. For example, when 100 nodes are deployed to be distributed among 5 clusters, ideally, each cluster should have 20 nodes. Here, the proposed fitness function ensures that each cluster has a node count in the range [18.8245, 21.1755] with 95% confidence and in the range [18.4526, 21.5474] with 99% confidence, whereas the fitness function of DEBCRP finds the same as in the ranges [15.2210, 24.7790] and [13.7093, 26.2907] with 95% and 99% confidences, respectively. It can be easily intuited that the node count in each cluster is much closer to the ideal node count (20 here) with the MLBCT-fitness function when compared to that with the DEBCRP-fitness function. The consistency of the MLBCT-fitness function in terms of balanced clusters' formation can be seen in Table 3.

*5.3.2. Experimental Analysis.* In this second set of experiments, as stated in point 2 of Section 5.1, MLBCT is compared to DEBCRP, ImDE-LEACH, and LEACH concerning the metrics—network lifetime, network energy consumption rate, and average residual energy per network node under two different network configurations, say WSN#1 and WSN#2. In WSN#1, the sink has been placed at the center of the $100 \times 100\,m^2$ sensing field, precisely at (50 m, 50 m) whereas, in WSN#2, the sink is located outside the sensing field at (50 m, 150 m). Moreover, to validate the adaptability of the scheme, simulations have been conducted with variable node deployments, say with 50 nodes, 100 nodes, 150 nodes, and 200 nodes.




(a) Scenario#1-5 clusters

(b) Scenario#2-10 clusters

Figure 6: Standard deviation of average cluster size for the clusters formed over the network rounds.

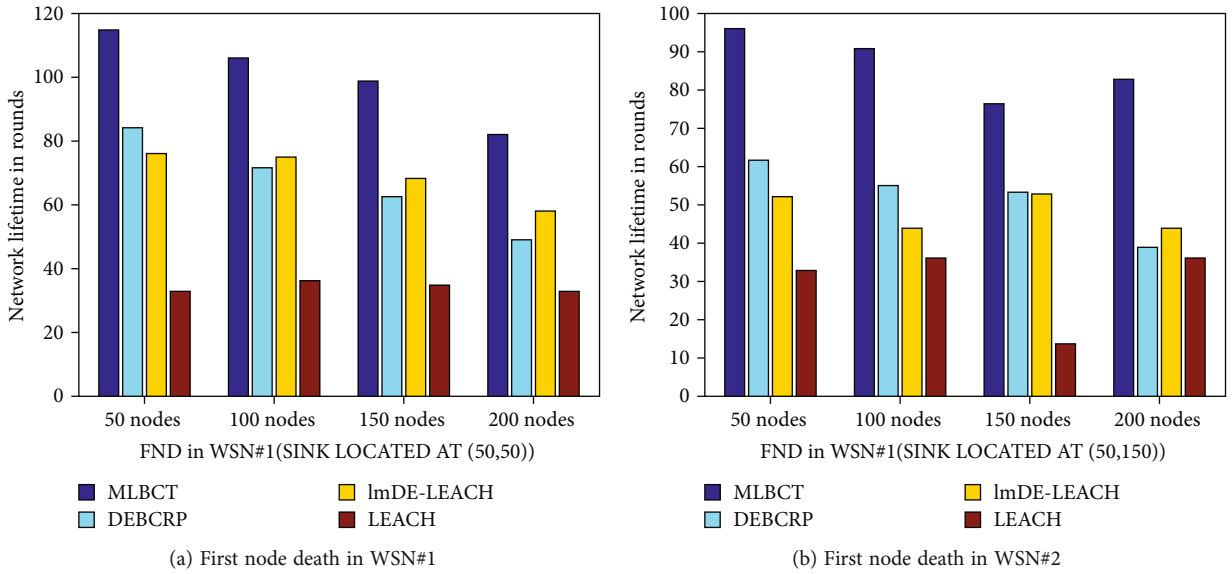
Table 3: Mean of standard deviations and confidence intervals for the clusters generated.

| Clustering scenario | #nodes | $\text{Mean}_{\sigma D-\text{Fitness}}$ | $\text{Mean}_{\sigma M-\text{Fitness}}$ | Interval estimate with 95% confidence | | Interval estimate with 99% confidence | |
|---|---|---|---|---|---|---|---|
| | | | | With *D*-fitness | With *M*-fitness | With *D*-fitness | With *M*-fitness |
| Scenario#1 (with 5 clusters) | 50 | 18.9254 | 4.4628 | [4.7541, 15.2459] | [8.7630, 11.2370] | [3.0947, 16.9053] | [8.3717, 11.6283] |
| | 100 | 24.3826 | 5.9975 | [15.2210, 24.7790] | [18.8245, 21.1755] | [13.7093, 26.2907] | [18.4526, 21.5474] |
| | 150 | 27.1672 | 6.9578 | [25.6523,34.3477] | [28.8865, 31.1135] | [24.2771, 35.7229] | [28.5343, 31.4657] |
| | 200 | 25.9534 | 5.7280 | [36.4030, 43.5970] | [39.2061, 40.79939] | [35.2652, 44.7348] | [38.9550, 41.0450] |
| Scenario#2 (with 10 clusters) | 50 | 11.6163 | 6.3429 | [1.7801, 8.2199] | [3.2419, 6.7581] | [0.7616, 9.2384] | [2.6857, 7.3143] |
| | 100 | 15.8541 | 8.4076 | [6.8926, 13.1074] | [8.3521, 11.6479] | [5.9096, 14.0904] | [7.8308, 12.1692] |
| | 150 | 21.6239 | 9.5944 | [11.5395, 18.4605] | [13.4646, 16.5354] | [10.4448, 19.5552] | [12.9789, 17.0211] |
| | 200 | 21.7987 | 10.7536 | [16.9789, 23.0211] | [18.5096, 21.4904] | [16.0232, 23.9768] | [18.0382, 21.9618] |

*(1) Network Lifetime.* As mentioned earlier in this section that the network lifetime can be defined as the time when the first node dies in the network or the time when the last node dies in the network. In Figures 7 and 8, both strategies have been followed separately.

Figures 7(a) and 7(b) describe the death of the first node that is FND (first node death) in the schemes MLBCT, DEBCRP, ImDE-LEACH, and LEACH under the network scenarios WSN#1 and WSN#2.

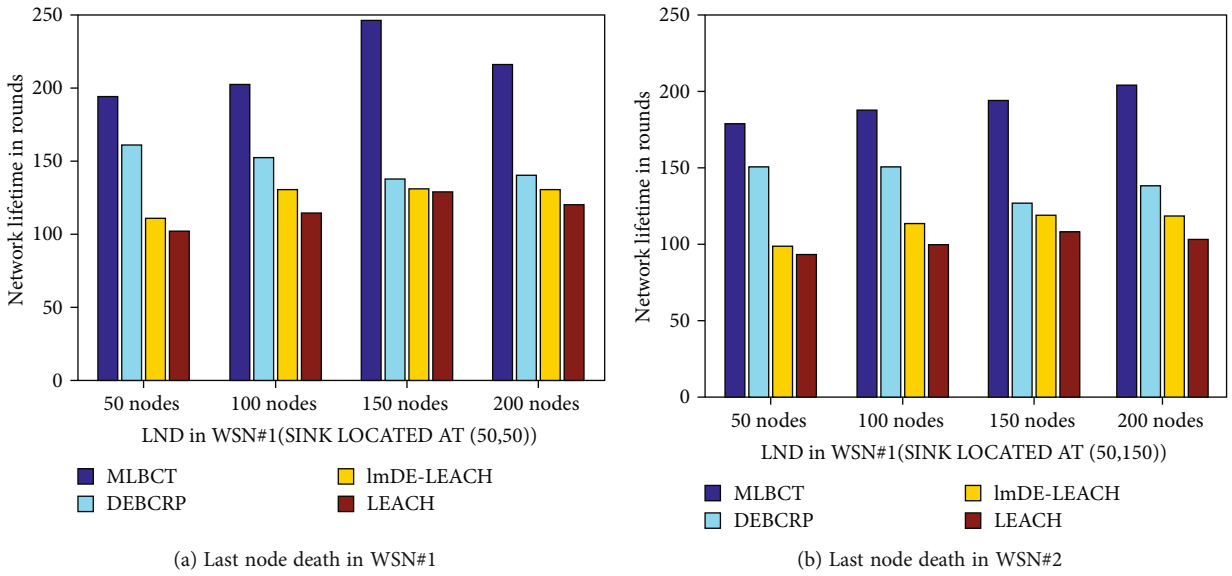
In WSN#1 (Figure 7(a)), when the number of nodes deployed are 50, 100, 150, and 200, the events of the first node's death (FND) occur at the round no. 115, 106, 99, and 82 in the proposed scheme; at 84, 72, 63, and 49 in DEBCRP; at 76, 75, 68, and 58 in ImDE-LEACH; and 33, 36, 35, and 33 in LEACH, respectively. Similarly, in




(a) First node death in WSN#1

(b) First node death in WSN#2

Figure 7: Network lifetime comparison in terms of FND.




(a) Last node death in WSN#1

(b) Last node death in WSN#2

Figure 8: Network lifetime comparison in terms of LND.

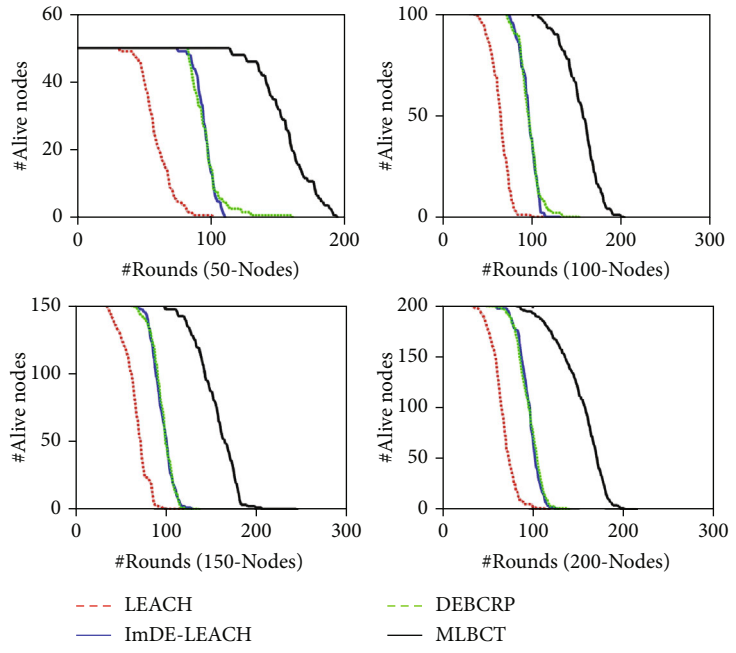
WSN#2 (Figure 7(b)), FNDs occur at round no. 96, 91, 77, and 83 in MLBCT; at 62, 55, 53, and 59 in DEBCRP; at 52, 44, 53, and 44 in ImDE-LEACH; and at 33, 36, 14, and 36 in LEACH, respectively, for the aforementioned nodes' count.

On the other hand, if the network lifetime is taken as the time when the last node dies that is LND (last node death) in the network, Figures 7(a) and 7(b) describe the outcomes of experiments conducted in this regard with the variable number of nodes as above, say 50, 100, 150, and 200, respectively.

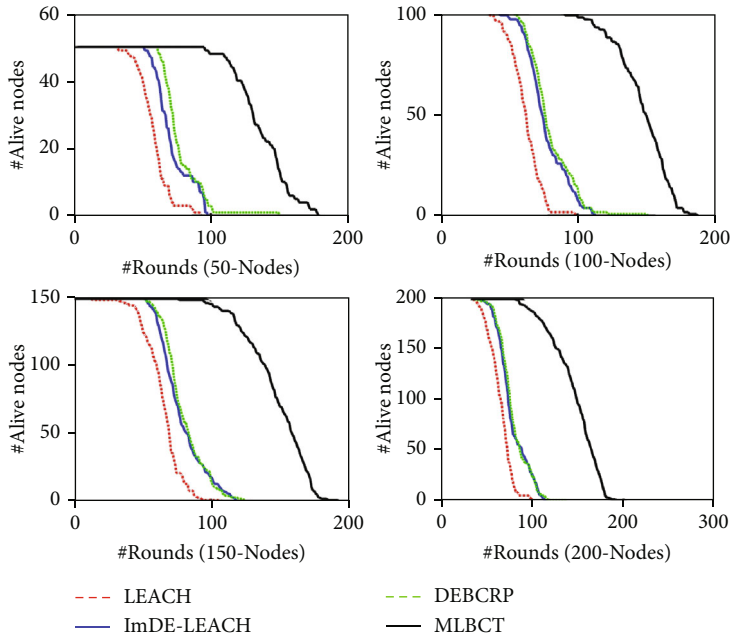
In WSN#1 (Figure 8(a)), the last node's death events occur at round no. 194, 202, 246, and 216 in the MLBCT; at 161, 152, 138, and 141 in DEBCRP; at 111, 131, 131, and 130 in ImDE-LEACH; and at 102, 114, 129, and 119 in LEACH, respectively. Likewise, in WSN#2 (Figure 8(b)), LNDs occur at round no. 178, 187, 193, and 203 in MLBCT; at 150, 151, 126, and 138 in DEBCRP; at 98, 113, 119, and 118; and at 93, 99, 108, and 103 in LEACH, respectively, for the aforementioned nodes' count. The appreciable results due to FND and LND calculation state the supremacy of using the proposed MLBCT over other schemes.

Moreover, the comparative performance of the schemes MLBCT, DEBCRP, ImDE-LEACH, and LEACH with respect to the nodes' death rate can also be observed from Figure 9.

Figure 9(a) describes the performance of the MLBCT against that of DEBCRP, ImDE-LEACH, and LEACH in




(a) Alive nodes over the network rounds in WSN#1

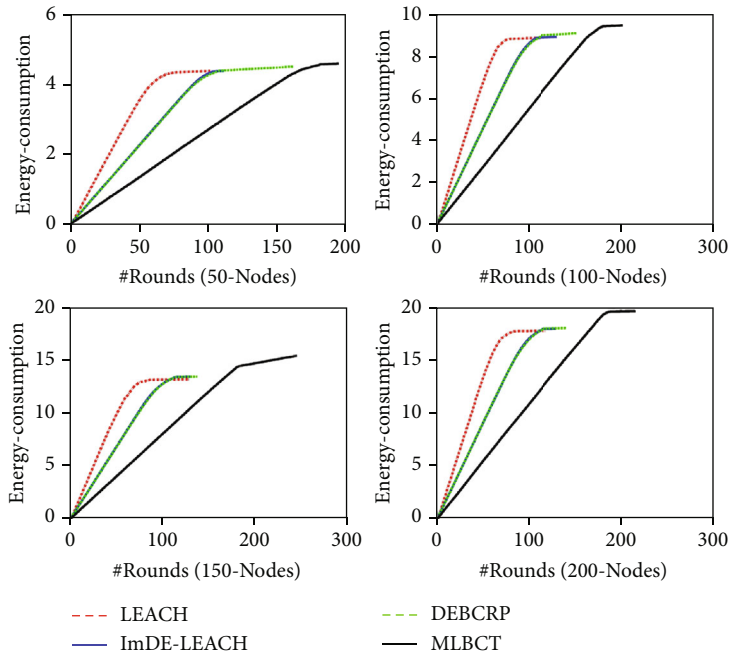



(b) Alive nodes over the network rounds in WSN#2

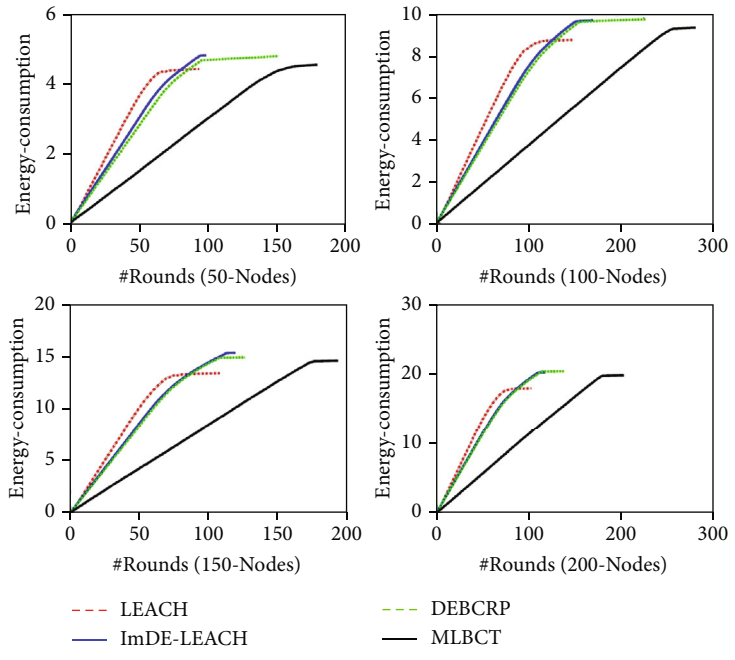
Figure 9: Network lifetime comparison in terms of alive nodes/round.

variable node population under the first network scenario WSN#1. Similarly, Figure 9(b) describes the same but for WSN#2. It is evident from Figure 9 that irrespective of the network configuration and nodes' population in the sensing field, MLBCT performs consistently well as the nodes' death rate is low in MLBCT, and hence, the number of alive nodes is high at any point of network operation in MLBCT when compared to DEBCRP, ImDE-LEACH, and LEACH. Thus, it can be concluded here that the MLBCT outperforms DEBCRP, ImDE-LEACH, and LEACH in terms of the first performance criterion—network lifetime.

*(2) Network Energy Consumption*. From Figure 10, it can be concluded that at any point of the network operation, the energy consumption in MLBCT is less than that in DEBCRP, ImDE-LEACH, and LEACH in both of the scenarios implemented that is in WSN#1 (Figure 10(a)) and




(a) Network energy consumption over the network rounds in WSN#1

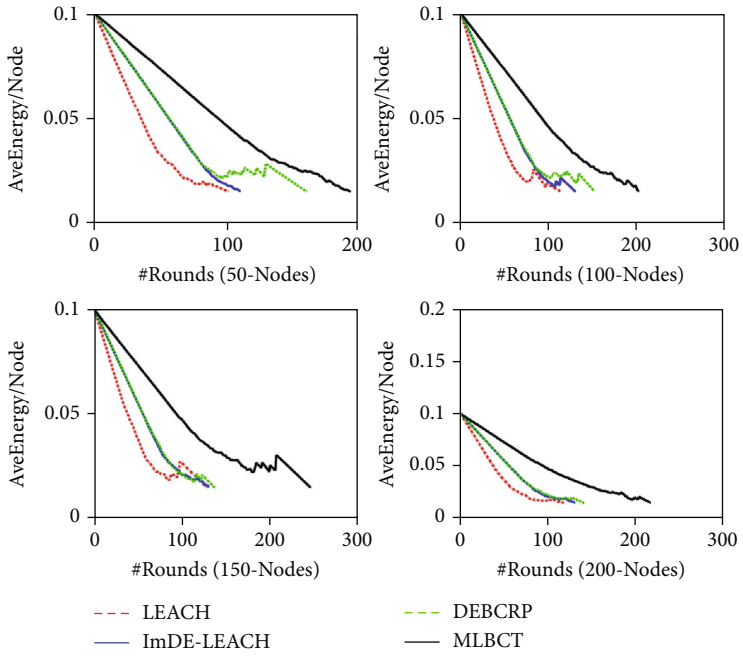



(b) Network energy consumption over the network rounds in WSN#2

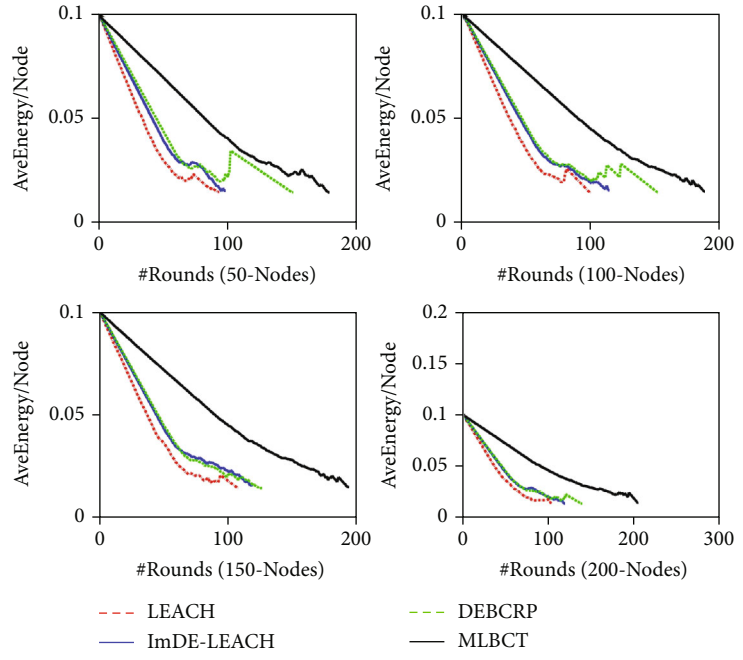
Figure 10: Comparison of network energy consumption over the network rounds.

WSN#2 (Figure 10(b)). Moreover, to demonstrate the consistency in the performance, variable counts of sensor nodes have been deployed here too.

*(3) Average Residual Energy/Node*. In this next set of experiments, the performance of MLBCT is measured in terms of the average residual energy that a network node has at any point in the network operation for the schemes DEBCRP, ImDE-LEACH, and LEACH. It can be explicitly observed that the nodes are always equipped with a larger amount of residual energy if being operated with MLBCT in comparison to DEBCRP, ImDE-LEACH, and LEACH (Figure 11). It is noticed not only in WSN#1 (Figure 11(a)) but also in WSN#2 (Figure 11(b)); average residual energy for a network node is higher at any point in network operation if implemented with MLBCT.




(a) Average residual energy/node in WSN#1

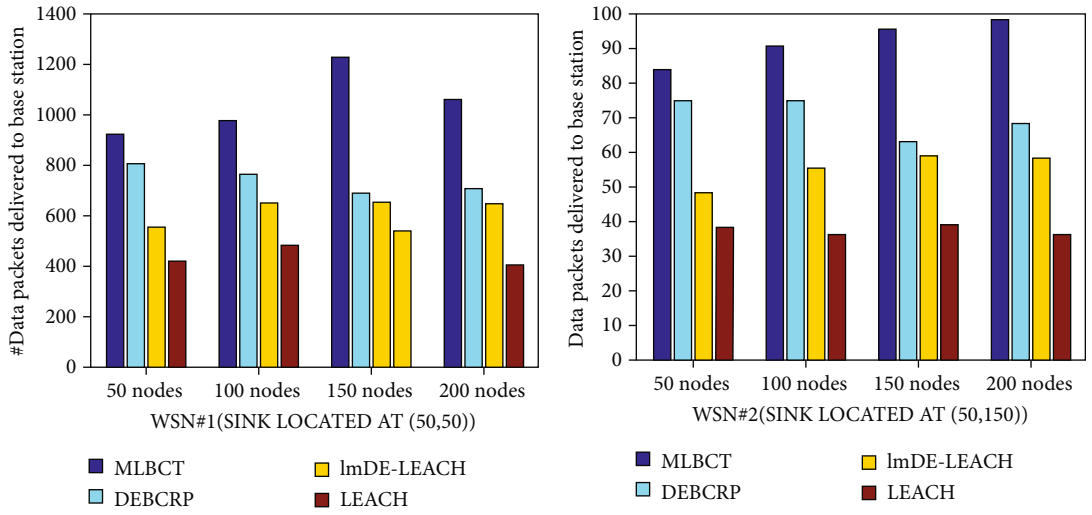



(b) Average residual energy/node in WSN#2

Figure 11: Comparison of average energy/node over the network rounds.

This depicts that a network utilizing MLBCT saves energy and keeps its resource intact for future usage, which is the desired criteria for sensor networks.

*(4) Data Packet Delivery at Base Station*. In the final set of experiments, the performance of MLBCT against the DEBCRP, ImDE-LEACH, and LEACH with respect to the number of data packets delivered to the base station is compared. The predominance of the proposed scheme, MLBCT, can be read for both the network scenarios WSN#1 and WSN#2 in Figures 12(a) and 12(b), respectively. For the 50, 100, 150, and 200 nodes, MLBCT enriches the base station with 915, 969, 1221, and 1054 data packets, respectively. However, DEBCRP results into 800, 755, 685, and 700 data packets, ImDE-LEACH results into 550, 650, 650, and 645 data packets, and LEACH results into 416, 477, 533, and




(a) Data packet delivery at base station in WSN#1

(b) First node death in WSN#2

Figure 12: Data packet delivery at base station.

401 data packets, respectively, for the aforesaid network nodes into WSN#1. Similarly, for the WSN#2, in comparison to 745, 750, 625, and 685 data packets due to DEBCRP, 98, 113, 119, and 118 data packets through ImDE-LEACH, and 382, 360, 388, and 372 data packets via LEACH, MLBCT results into 838, 903, 950, and 983 data packets at the base station, respectively, for the node deployment mentioned above. This suggests that MLBCT successfully transmits more packets depicting its dominance in terms of successful transmission.

Based on the outcomes of the various simulations conducted so far, it can be concluded that the MLBCT outperforms the DEBCRP, ImDE-LEACH, and LEACH in terms of the chosen criteria of network lifetime, network stability, average residual energy, and data packet delivery.

## 6. Conclusion and Future Works

In this work, a Metaheuristic Load-Balancing-Based Clustering Technique has been proposed for wireless sensor networks. To achieve the prime objective of load-balanced clusters, a fitness function has been proposed that offers balanced clusters in terms of their size and energy and ensures the members to be in close proximity to one another reducing the cost of intracluster communication. Through an extensive set of simulations and experimentation, the supremacy of the proposed scheme MLBCT has been proved over the existing ones DEBCRP, and ImDE-LEACH in terms of improved network lifetime and network stability, average residual energy, and data packet delivery.

Statistical analysis also justifies and supports the feasibility of the scheme. Moreover, the scheme's adaptability and scalability have also been established by varying the network configuration with the different number of nodes and different placement of the base station.

As a future extension of this work, a heterogeneous wireless sensor network (HWSN) would be investigated to device a clustering-based scheme induced by metaheuristic techniques to consistently contribute to the network operations without being affected by the heterogeneity present in the network.

## Data Availability

Extensive analysis, method, and result data has been fully provided.

## Conflicts of Interest

The authors declare that they have no competing interests.

## Acknowledgments

This work is partially supported by DST/TDT/DDP-38/2021, Device Development Programme (DDP), by the Department of Science & Technology (DST), Ministry of Science and Technology, Government of India.

## References

[1] I. F. Akyildiz, Weilian Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communication Magazine*, vol. 40, no. 8, pp. 102–114, 2002.

[2] C. O. Iwendi and A. R. Allen, "Enhanced security technique for wireless sensor network nodes," in *IET Conference on Wireless Sensor Systems (WSS 2012)*, pp. 1–5, London, 2012.

[3] A. Dumka, S. K. Chaurasiya, A. Biswas, and H. L. Mandoria, *A Complete Guide to Wireless Sensor Networks: From Inception to Current Trends*, CRC Press, Boca Raton, Florida, USA, 1st edition, 2019.

[4] S. J. Ramson and D. J. Moni, "Applications of wireless sensor networks –survey," in *2017 International Conference on Innovations in Electrical, Electronics, Instrumentation and Media Technology (ICEEIMT)*, pp. 325–329, Coimbatore, India, 2017.

[5] S. Ghiasi, A. Srivastava, X. Yang, and M. Sarrafzadeh, "Optimal energy aware clustering in sensor networks," *Sensors Journal*, vol. 2, no. 7, pp. 258–269, 2002.

[6] W. R. Heinzelman, A. Chandrakasan, and H. Balkrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of 33rd Hawaii international conference on system science*, pp. 1–10, Maui, HI, USA, 2000.

[7] W. R. Heinzelman, A. Chandrakasan, and H. Balkrishnan, "An application -specific protocol architecture for wireless microsensor networks," *IEEE Transactions on wireless communications*, vol. 1, no. 4, pp. 660–670, 2002.

[8] S. Lindsey and C. S. Raghavendra, "PEGASIS: power-efficient gathering in sensor information systems," in *Proceedings, IEEE Aerospace Conference*, pp. 1125–1130, Big Sky, MT, USA, 2002.

[9] S. K. Chaurasiya, T. Pal, and S. D. Bit, "An enhanced energy-efficient protocol with static clustering for WSN," in *The International Conference on Information Networking 2011 (ICOIN2011)*, pp. 58–63, Kuala Lumpur, Malaysia, 2011.

[10] S. K. Chaurasiya, J. Sen, S. Chatterjee, and S. D. Bit, "EBLEC: an energy-balanced lifetime enhancing clustering for WSN," in *2012 14th International Conference on Advanced Communication Technology (ICACT)*, pp. 189–194, PyeongChang, Korea (South), 2012.

[11] M. N. Cheraghlou and M. Haghparast, "A novel fault-tolerant leach clustering protocol for wireless sensor networks," *Journal of Circuits, Systems and Computers*, vol. 23, no. 3, article 1450041, 2014.

[12] E. Moridi, M. Haghparast, M. Hosseinzadeh, and S. Jafarali Jassbi, "Novel fault-tolerant clustering-based multipath algorithm (FTCM) for wireless sensor networks," *Telecommunication Systems*, vol. 74, no. 4, pp. 411–424, 2020.

[13] O. Younis and S. Fahmy, "HEED: a hybrid energy-efficient distributed clustering approach for ad hoc sensor networks," *IEEE Transaction on Mobile Computing*, vol. 3, no. 4, pp. 366–379, 2004.

[14] E. Moridi, M. Haghparast, M. Hosseinzadeh, and S. Jafarali Jassbi, "A novel hierarchical fault management framework for wireless sensor networks: HFMF," *Peer-to-Peer Networking and Applications*, vol. 85, 2021.

[15] T. R. Gadekallu, M. Alazab, R. Kaluri et al., "Hand gesture classification using a novel CNN-crow search algorithm," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1855–1868, 2021.

[16] T. Shankar, S. Shanmugavel, A. Karthikeyan, A. M. Gupte, and S. Sarkar, "Load balancing and optimization of network lifetime by use of double cluster head clustering algorithm and its comparison with various extended LEACH versions," *International Review on Computers and Software*, vol. 8, no. 3, pp. 795–803, 2013.

[17] T. Shankar and S. Shanmugavel, "Energy optimization in cluster based wireless sensor networks," *Journal of Engineering Science and Technology*, vol. 9, no. 2, pp. 246–260, 2014.

[18] A. M. Zungeru, L. M. Ang, and K. P. Seng, "Classical and swarm intelligence based routing protocols for wireless sensor networks: a survey and comparison," *Journal of Network and Computer Applications*, vol. 35, no. 5, pp. 1508–1536, 2012.

[19] M. Saleem, A. Gianni, and M. F. Di Caro, "Swarm intelligence based routing protocol for wireless sensor networks: survey and future directions," *Information Sciences*, vol. 181, no. 20, pp. 4597–4624, 2011.

[20] R. V. Kulkarni and G. K. Venayagamoorthy, "Particle swarm optimization in wireless-sensor networks: a brief survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 2, pp. 262–267, 2011.

[21] R. Mukherjee, S. Debchoudhury, R. Kundu, S. Das, and P. Suganthan, "Adaptive differential evolution with locality based crossover for dynamic optimization," in *2013 IEEE Congress on Evolutionary Computation*, pp. 63–70, Cancun, Mexico, 2013.

[22] S. Das, A. Konar, and U. K. Chakraborty, "Annealed differential evolution," in *2007 IEEE Congress on Evolutionary Computation*, pp. 1926–1933, Singapore, 2007.

[23] P. Kuila, S. K. Gupta, and P. K. Jana, "A novel evolutionary approach for load balanced clustering problem for wireless sensor networks," *Evolutionary Computation*, vol. 12, pp. 48–56, 2013.

[24] P. Kuila and P. K. Jana, "Energy efficient clustering and routing algorithms for wireless sensor networks: particle swarm optimization approach," *Engineering Applications of Artificial Intelligence*, vol. 33, pp. 127–140, 2014.

[25] D. Kim, S. Song, and B. Y. Choi, "Energy-efficient adaptive geosource multicast routing for wireless sensor networks," *Journal of Sensors*, vol. 2013, 14 pages, 2013.

[26] J. Vesterstrom and R. Thomsen, "A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems," in *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No.04TH8753)*, pp. 1980–1987, Portland, OR, USA, 2004.

[27] X. Li, L. Xu, H. Wang, J. Song, and S. X. Yang, "A differential evolution-based routing algorithm for environmental monitoring wireless sensor networks," *Sensors*, vol. 10, no. 6, pp. 5425–5442, 2010.

[28] P. Kuila and P. K. Jana, "A novel differential evolution based clustering algorithm for wireless sensor networks," *Applied Soft Computing*, vol. 25, pp. 414–425, 2014.

[29] C. P. Low, C. Fang, J. M. Ng, and Y. H. Ang, "Efficient load-balanced clustering algorithms for wireless sensor networks," *Computer Communications*, vol. 31, no. 4, pp. 750–759, 2008.

[30] G. Gupta and M. Younis, "Load-balanced clustering of wireless sensor networks," in *IEEE International Conference on Communications, 2003. ICC '03*, vol. 3, pp. 1848–1852, Anchorage, AK, USA, 2003.

[31] K. Pratyay and P. K. Jana, "Energy efficient load-balanced clustering algorithm for wireless sensor networks," *Procedia Technology*, vol. 6, pp. 771–777, 2012.

[32] T. Sweta Potthuri and A. R. Shankar, "Lifetime improvement in wireless sensor networks using hybrid differential evolution and simulated annealing (DESA)," *Ain Shams Engineering Journal*, vol. 9, no. 4, pp. 655–663, 2018.

[33] J. Brest, S. Greiner, B. Boskovic, M. Mernik, and V. Zumer, "Self-adapting control parameters in differential evolution: a comparative study on numerical benchmark problems," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 6, pp. 646–657, 2006.

[34] S. Randhawa and S. Jain, "MLBC: multi-objective load balancing clustering technique in wireless sensor networks," *Applied Soft Computing*, vol. 74, pp. 66–89, 2019.

[35] M. Ghahramani and A. Laakdashti, "Efficient energy consumption in wireless sensor networks using an improved differential evolution algorithm," in *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 18–23, Mashhad, Iran, 2020.

[36] G. P. Gupta and B. Saha, "Load balanced clustering scheme using hybrid metaheuristic technique for mobile sink based wireless sensor networks," *Journal of Ambient Intelligence and Humanized Computing*, vol. 2020, 2020.

[37] C. Iwendi, P. K. R. Maddikunta, T. R. Gadekallu, K. Lakshmanna, A. K. Bashir, and M. J. Piran, "A metaheuristic optimization approach for energy efficiency in the IoT networks," *Software: Practice and Experience*, vol. 51, pp. 1–14, 2021.

[38] S. Ponnan, A. K. Saravanan, C. Iwendi, E. Ibeke, and G. Srivastava, "An artificial intelligence-based quorum system for the improvement of the lifespan of sensor networks," *IEEE Sensors Journal*, vol. 21, no. 15, pp. 17373–17385, 2021.

[39] S. H. Sackey, J. Chen, A. J. Henry, and X. Zhang, "A clustering approach based on genetic algorithm for wireless sensor network localization," in *2019 15th International Conference on Computational Intelligence and Security (CIS)*, pp. 54–58, Macao, China, 2019.

[40] J. Chen, S. H. Sackey, J. H. Anajemba, X. Zhang, and Y. He, "Energy-efficient clustering and localization technique using genetic algorithm in wireless sensor networks," *Complexity*, vol. 2021, 12 pages, 2021.

[41] S. H. Sackey, J. A. Ansere, J. H. Anajemba, M. Kamal, and C. Iwendi, "Energy efficient clustering based routing technique in WSN using brain storm optimization," in *2019 15th International Conference on Emerging Technologies (ICET)*, Peshawar, Pakistan, 2019.

[42] M. Kaddi, Z. Khalili, and M. Bouchra, "A differential evolution based clustering and routing protocol for WSN," in *2020 International conference on mathematics and information technology*, pp. 190–195, Adrar, Ageria, 2020.

[43] A. Gaur and T. Kumar, "Switching-differential evolution (S-DE) for cluster head election in wireless sensor network," *IJARIIE*, vol. 2, no. 5, pp. 2395–4396, 2016.

[44] R. V. Rao, V. J. Savsani, and D. P. Vakharia, "Teaching-learning-based optimization: a novel method for constrained mechanical design optimization problems," *Computer-Aided Design*, vol. 43, no. 3, pp. 303–315, 2011.

[45] R. V. Rao, V. J. Savsani, and D. P. Vakharia, "Teaching-learning-based optimization: an optimization method for continuous non-linear large scale problems," *Information Sciences*, vol. 183, no. 1, pp. 1–15, 2012.

[46] M. Ramadas and A. Abraham, "Clustering wireless sensor networks using ImDE algorithm with LEACH protocol," in *2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)*, pp. 1–4, Manama, Bahrain, 2019.

[47] S. K. Chaurasiya, J. Mondal, and S. Datta, "Field-of-view based hierarchical clustering to prolong network lifetime of WMSN with obstacles," in *2014 International Conference on Electronics, Communication and Computational Engineering (ICECCE)*, pp. 72–77, Hosur, India, 2014.

[48] S. K. Singh, P. Kumar, and J. P. Singh, "A survey on successors of LEACH protocol," *IEEE Access*, vol. 5, pp. 4298–4328, 2017.

[49] A. Mehmood, S. Khan, B. Shams, and J. Lloret, "Energy-efficient multi-level and distance-aware clustering mechanism for WSNs," *International Journal of Communication Systems*, vol. 28, no. 5, pp. 972–989, 2015.

[50] R. Banerjee, S. Chatterjee, and S. Das Bit, "An energy saving audio compression scheme for wireless multimedia sensor networks using spatio-temporal partial discrete wavelet transform," *Computers and Electrical Engineering*, vol. 48, pp. 389–404, 2015.

[51] R. Banerjee and S. Das Bit, "Low-overhead video compression combining partial discrete cosine transform and compressed sensing in WMSNs," *Wireless Networks*, vol. 25, no. 8, pp. 5113–5135, 2019.

[52] R. Banerjee and S. Das Bit, "An energy efficient image compression scheme for wireless multimedia sensor network using curve fitting technique," *Wireless Networks*, vol. 25, no. 1, pp. 167–183, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9065768, 26 pages
https://doi.org/10.1155/2022/9065768

*Review Article*

# Blockchain Technology on Smart Grid, Energy Trading, and Big Data: Security Issues, Challenges, and Recommendations

**Mohammad Kamrul Hasan**,[1] **Ali Alkhalifah**,[2] **Shayla Islam**,[3] **Nissrein B. M. Babiker**,[4] **A. K. M. Ahasan Habib**,[1] **Azana Hafizah Mohd Aman**,[1] **and Md. Arif Hossain**[1]

[1]*Center for Cyber Security, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Malaysia*
[2]*Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia*
[3]*Institute of Computer Science and Digital Innovations, UCSI University, 56000 Kuala Lumpur, Malaysia*
[4]*Information System Department, College of Science and Arts-Tathleeth, University of Bisha, P.O. Box 551, Bisha 61922, Saudi Arabia*

Correspondence should be addressed to Mohammad Kamrul Hasan; hasankamrul@ieee.org and Shayla Islam; shayla@ucsiuniversity.edu.my

Received 12 August 2021; Accepted 12 November 2021; Published 18 January 2022

Academic Editor: Rajesh Kaluri



The smart grid idea was implemented as a modern interpretation of the traditional power grid to find out the most efficient way to combine renewable energy and storage technologies. Throughout this way, big data and the Internet always provide a revolutionary solution for ensuring that electrical energy linked intelligent grid, also known as the energy Internet. The blockchain has some significant features, making it an applicable technology for smart grid standards to solve the security issues and trust challenges. This study will present a rigorous review of blockchain implementations with the cyber security perception and energy data protections in smart grids. As a result, we describe the major security issues of smart grid scenarios that big data and blockchain can solve. Then, we identify a variety of recent blockchain-based research works published in various literature and discuss security concerns on smart grid systems. We also discuss numerous similar practical designs, experiments, and items that have recently been developed. Finally, we go through some of the most important research problems and possible directions for using blockchain to address smart grid security concerns.

## 1. Introduction

Internet of Things (IoT) is considered the most uncontrollable innovation in today's world; this improves our ordinary life by reworking the bodily items that surround us into an ecosystem of facts. IoT and big data have numerous applications in day-to-day life, i.e., security, transportation, industrial, retail, healthcare, home automation, military, agriculture, surveillance, and good infrastructure. Indeed, IoT and big data have heavily driven nowadays smart grid developments, and smart meters are progressing by featuring more vital sensing abilities and higher connectivity [1, 2]. The smart electricity generation, transmission, and distribution system and smart buildings/homes are all controlled and explicitly maintained by ICT devices like WAMS, IEDS, and RTUs for service systems, as well as AMIs for smart building/home management in the smart grid (SG) [3]. The IoT-enabled field measurement data can be safely and automatically collected by including the blockchain control and field measurement with smart communication to these ICT devices in HAN/SN, NAN, and WAN [4]. Furthermore, blockchain-enabled AMIs can use DAPPS services to conduct decentralized system capacity, local power management, and trading in a cyber-secured environment [5].

The days are passing. Civilization is also progressing rapidly. Science is becoming more powerful as time passes. Simultaneously with the problems that the modern world presents, scientists and engineers face difficulties in satisfying market demand at various levels for various reasons. The total electricity generation, transmission, and

distribution system are becoming a loss project because of a lack of raw electricity generation raw material supply, corruption on both the transmitting and receiving ends, transmission line and distribution system losses, and other factors. As a result, the SG technology was developed to meet consumer demand, improve the electricity generation and distribution system efficiency, ensure customer protection, and monitor and regulate the entire system through communication (generating and receiving end). As a result, the critical focus of the paper is to include an overview of blockchain (BC) in smart grid and energy trading presented in Figure 1.

Beyond the area of computer vision, this article will contribute as a supplement to an adversarial attacks' summary and protections for SG IoT and big data linked devices and networks. The contribution of this study is discussed below.

(1) *General Working Flow*. We review an overall working flow to describe the Internet-connected devices, protocols, and network infrastructure and its adversarial attacks in SG big data/IoT networks. The integration of the potential BC technology in SG IoT networks is presented. Based on this, a robust classification is provided to organize and structure existing attacks intricately and effectively where the defenses can be possibly accomplished in SG IoT-connected devices and networks.

(2) *Systematic and Comparable Studies*. We classify current attacks based on the above taxonomy into three standard sensor data types: textual, audio, and surveillance sensor data. Here, we also did a quantitative comparison between them based on six technical factors. In addition, we define as well as outline three possible defense strategies for aggressive attacks in CPSs.

(3) *Open Issues and Opportunities*. We highlight several existing research prospects which should be pursued in the future in order to inspire and enhance future follow-up on this research topic.

The overview of this paper's structure is presented in Figure 2. Section 2 discusses the study methodology and the relevant research on IoT security specifications. Section 3 examines the findings, highlighting critical characteristics for understanding IoT and general security criteria related to the entire lot system. Section 4 presents the overview of blockchain technology and its application. Finally, in Section 5, we present the findings of this study, and the conclusion is in Section 6.

## 2. Research Methodology

*2.1. Research Questions. Research questions*. The following research questions are to be analyzed and accomplished throughout the paper:

(1) What are the recent features technologies in SG?

(2) What are the security vulnerabilities, threats, and their counter measurement in SG?

(3) What are the blockchain technology and the security mechanism that attracted the researchers to security solutions for SG?

(4) What are the critical success factors of the blockchain that can ensure the security of SG systems (smart metering, energy trading, SG communication systems, etc.)?

(5) What are the issues and challenges of the blockchain- (BC-) based security solutions, and what are the possible enhancements of the blockchain framework that strengthen the security in SG?

*2.2. Review Protocol.* The specific review protocol of the procedures can be followed during the studies. It is necessary to make assessments almost the review issue, data extractions, data synthesis, inclusion criteria, quality assessment, search strategy, research collection, and dissemination plans. Only full published conferences and journals in the English from 2010 and 2021 were considered. Data sources, data extraction, research collection, and selection strategy process are the main components of the review protocol.

*2.3. Data Sources.* To assist in answering the research questions, research papers related to blockchain, IoT, and big data were chosen. Related research articles which are not addressed or even endorse with the research questions were rejected. Our primary resources for looking the published research publications are in the following libraries:

(i) Science Direct

(ii) IEEE Xplore Digital Library

(iii) MDPI

(iv) Taylor and Francis

(v) Springer Link

(vi) ACM Digital Library

(vii) Google Scholar

*2.4. Search Process.* Based on the research methodology, we focused on IoT and blockchain-based keyword patterns to find any research queries. We apply Boolean operators and symbols like "AND," "OR" to find out the following keywords: (block chain OR (block chain technology) OR (block chain security AND block chain issues) OR (IoT security)) OR (big data in SG) AND (study OR Adoption) AND ((requirements AND solution) OR (benchmark AND regulation)) AND ((block chain application AND fields). Figure 3 presents this process.

*2.5. Data Selection.* The data collection is the deciding process of the appropriate data source and type and perfect implements to collect the data. Data selection precedes the actual repetition of data collection. Data selection criteria were as follows:

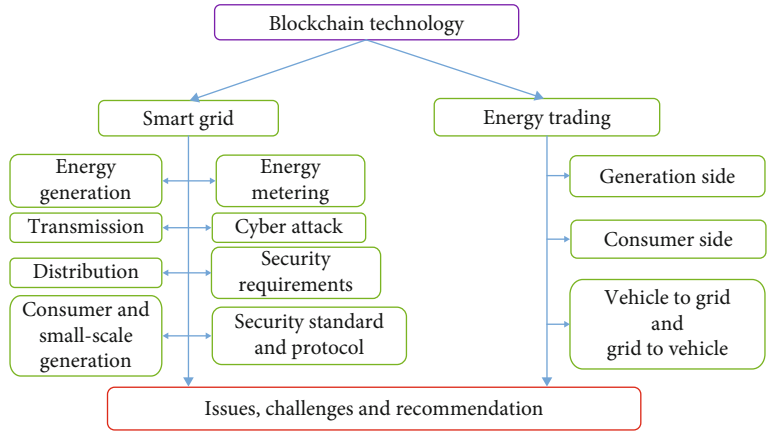



Figure 1: Overview of study.

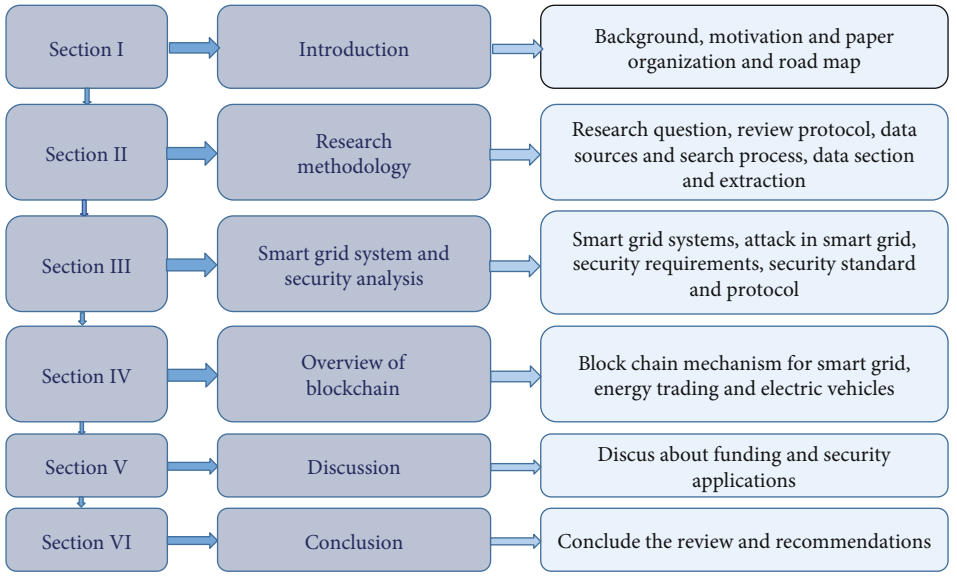



Figure 2: Paper organization

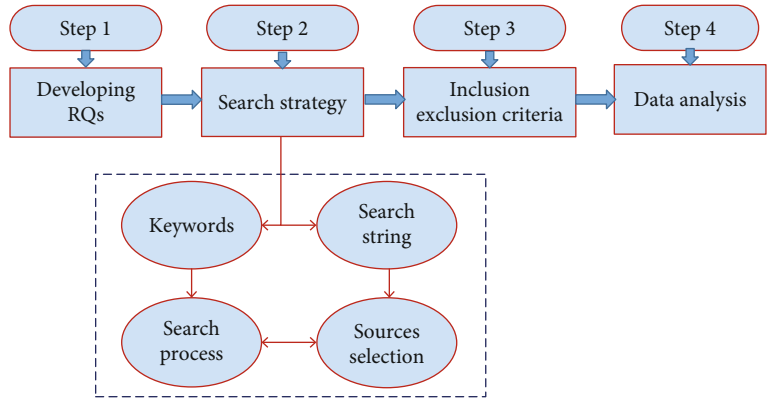



Figure 3: Proposed systematic search process.

(i) Has been the study paper written between 2010 and July 2021?

(ii) Is the research paper published in the well-known/referred data sources?

(iii) Does the research paper reference or address BC/IoT/big data?

(iv) Is there any discussion of security, requirements, or practice in the research paper?

*2.6. Data Extraction.* In July 2021, we completed the search process and discovered 269 publications and websites. Related research papers were carefully extracted by following the collection and rejection criteria as part of the search process. Finally, preliminary results were found from 142 abstract studies and 57 full-length reviews and research work for studies. The data synthesis and extraction of the selected review papers to find the research question answers and classify the studies shown in Tables 1 and 2 present the IoT and blockchain application field, respectively.

Table 1: Selective articles extraction from primary study sources.

| Sources | Found | Candidate | Selected |
|---|---|---|---|
| IEEE Xplore | 187 | 108 | 64 |
| Elsevier | 132 | 83 | 43 |
| ACM Digital Library | 25 | 8 | 3 |
| Hindawi | 21 | 14 | 9 |
| Google Scholar | 200 | 56 | 37 |
| Science Direct | 87 | 39 | 26 |
| Springer Link | 42 | 27 | 13 |
| MDPI | 51 | 22 | 14 |
| Website | 4 | 4 | 4 |
| Total | 772 | 321 | 213 |

## 3. Smart Grid System and Security Analysis

The term "SG" states to a concept that encapsulates the entire electricity generation, transmission, and distribution system in a single edging. In other words, an SG makes smarter the entire system more competent or safer. Clean energy is now in high demand all over the world. As a result, clean energy is also called smart energy. The word "smart grid" was first used in the year 2003 [6]. That was the first time Michael T. Burr used the word in a document. He clarified how the power grid's flaws could be detected and fixed to improve the power flow mechanism from generation to delivery across the whole transmission lines. This SG idea is now a reality, and the SG design objectives are presented in Figure 4. It was becoming a fact through the excellence of executing some one-of-a-kind function that makes things simpler. The SG is prepared smart by exhausting the national grid's security mechanism and central control via the supervisory control and data acquisition (SCADA), transmission equipment monitoring and diagnostic, grid computing, handling the whole power system as a hybrid adaptive power system, and using distributed computer agents to make the self-healing power system network [7].

*3.1. Smart Grid Systems.* The development of a highly secure, dependable, and eco-friendly national power grid system, termed the SG, is being driven by rising concerns about greenhouse gas emissions like carbon dioxide ($CO_2$) and the demand for additional efficient and dependable power transmission and distribution [8]. An SG uses two-way digital technology to transmit power between providers and consumers. It monitors and regulates smart appliances in users' homes or buildings to conserve energy, save costs, and improve dependability, efficiency, and transparency (Figure 5) [9]. The legacy power network is intended to be modernized by a smart grid. It automatically monitors, protects, and optimizes the function of the associated pieces. Several of the SG technologies are already in use in different industrial regions, like manufacturing process of wireless and sensor networks in telecommunications, and are starting to be modified for application in the different intelligent fields and linked scenarios such as energy distributions, communication systems, energy metering, and energy trading. The conventional power delivery system focuses on designing technology that improves the power supply's integrity, availability, and secrecy. Until recently, modern communication technology and equipment were thought to be boosting the dependability of the power industry. Nonetheless, the growing connection is becoming more critical for the power system's cyber security. In particular, securing the electrical grid system protects, arranges for, recovers, responds, and mitigates from unexpected cyber system incidents or natural catastrophes [10].

The integration of security system/protocol/algorithm with smart grid (SG) technology is becoming so sophisticated key solutions for facilitating comprehensive security functionality SG technology. The core related interfaces, components, and applications of SG that are critically security dependent are discussed in analyzing the key RQs. The feature of SG is presented in Figure 6.

*3.1.1. Smart Meters for Energy Trading.* Smart meters (SMs) are distinctive characteristics of SG technology that become a most reliable device for data measurement in electricity generation, transmission, and consumption. The SMs combine use with digital meters and communication systems to allow real-time monitoring of the consumers' energy [2]. In simple terms, a SM is a meter that calculates the amount of electricity used by customers. It usually records the reading at several times during the day. A typical SM assists the customers to understand electricity consumption and billing procedures; therefore, they can easily manage their usage electricity inside their desired budget/billing limit.

On the other hand, the SM measurement aids the suppliers and consumers in calculating accurate bills for customers. SM acts as a contact point between households and the Distribution System Operator (DSO), part of energy transactions. It is crucial to have a secure connection between SM and utility servers because it can affect transactions and billing information. It is essential to maintain track of the transactions in terms of planning the operation and compute invoices. When several parties engage in any kind of trading, trust is a big challenge. Initially, the record of transactions is maintained by a responsible third party. For SM-DSO transactions, blockchain technology can be used to maintain a distributed ledger. As a result, they were implementing blockchain technology to trade energy required to be trusted on their third party [11–15].

*3.1.2. Distributed Generations.* Smart grid technology relies heavily on distributed generation (DG). The term

Table 2: The use of IoT blockchain technology application field.

| Sector | The use or application field |
|---|---|
| Smart city | Smart transaction and data maintenance, facilitating digital data and data transactions, pollution control data, water management data, and energy management data are all examples of smart service offerings. |
| Healthcare | Health costs, hospital information systems, medication records, digitizing old medical records, digital case notes, electronic medical records, genome data, and vital signs are all examples of genomic data. |
| Agriculture | Agriculture big data used for seed processing, agro-seed marketing, sales data, soil data, yields, agro-product shipping, and analytics |
| Energy | Data on energy generation, demand, resource availability and raw material information, utility condition monitoring, resource tracking, and tariff data maintenance |
| Manufacturing | Manufacturing packaging data, product output data and management, actuators/sensors, automation, raw material, supplier data monitoring, and transaction data monitoring for product distribution |
| Transport and logistics | Transport records, vehicle tracking, toll data management, logistic service identifiers, shipment data, and container tracking and accurate distribution are all examples of transportation records. |
| Others | Virtual nations, voting and government, space development, precious and jewels metals, ownership, economic sharing, and digital content |
| Distribution | Mining chips, digital currencies, marketplace, sales records, storage records, transport records, used sales, and goods |
| Business | Import and export data, tech industry digital documents, and transaction processing data have been used for financial analytics. |
| Finance | Money trading, money deposits, money transfer, crowd funding, smart securities, smart contracts, social banking, digital transaction assets, and crypto currency are all examples of digital transaction assets. |

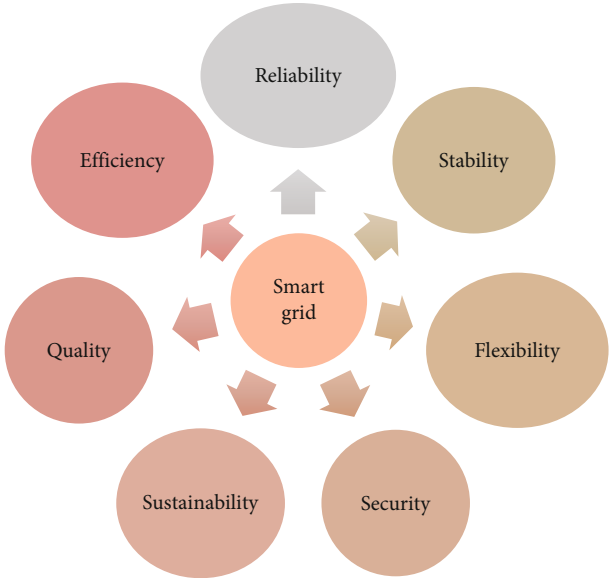



Figure 4: Smart grid design objectives.

"distributed generation" refers to the production of electricity from various small energy sources. Massive power plant generation has inevitable consequences, such as environmental effects on transmission and distribution and a very stable electricity supply via the grid [6]. The present electricity networks are becoming more overburdened as demand rises regularly. As a result, traditional strategies contribute to the complexity of existing networks. To meet customer expectations on the distribution side, such as lower power bills, increased comfort, reliability, and data security, a comprehensive analysis of SG components such as distributed generation is necessary [9, 16]. Integrated minor nonconventional power resources can be utilized to produce electricity at the load end in distributed generation. This technology improves power quality, efficiency, reliability, and security while lowering operational costs and environmental impact [10, 17].

*3.1.3. Integration of the Renewable Energy.* The interconnection of renewable energy is another critical function of the SG system. Improving the grid's IRE (Integration Renewable Energy) capability allows the national power grid to address customers' increased demand while maintaining future security. Like the DG (distributed generation), IRE will face some difficulties as it integrates into the smart grid.

*3.1.4. Two-Way Communication System.* The SG system is more straightforward for both suppliers and consumers when the bidirectional communication system is activated. The SG communicates in two-way communication with consumer's alert of the price and energy consumption as well as electricity generation, and suppliers are aware of the simple billing system of usage electricity. Cyberphysical security employs communication interfaces such as Universal Asynchronous Receiver-Transmitter (UART), Ethernet, and WLAN for the complex Internet Connected SACADA and PMU device WAMS in SG networks. The IEEE C37.118, Gateway Exchange Protocol (GEP), SIEGate has been designed and presented to secure cyberphysical communication interface, gateway, and control systems. Furthermore, only this communication device allows for central control of the entire grid. However, one thing to keep in mind is that privacy must be protected when interacting in the SG system, whether multidirectional or bidirectional.

*3.1.5. Automatic Healing Capability.* Since SG system is a cognitive approach for electricity generating and distributing

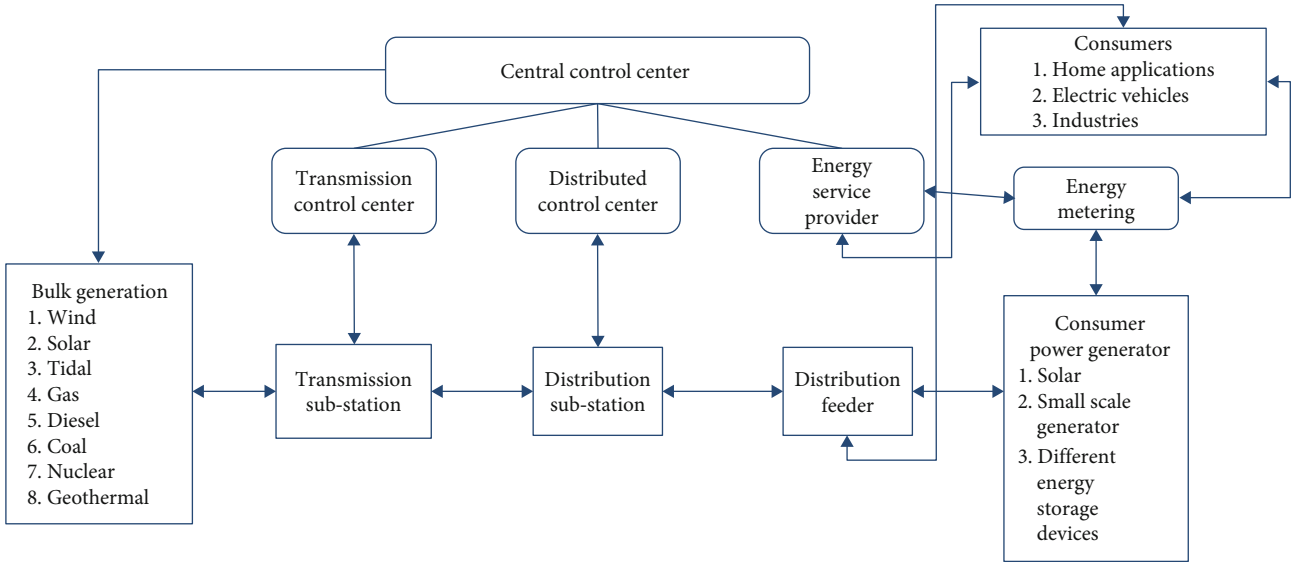



Figure 5: Smart grid communication infrastructures [9].

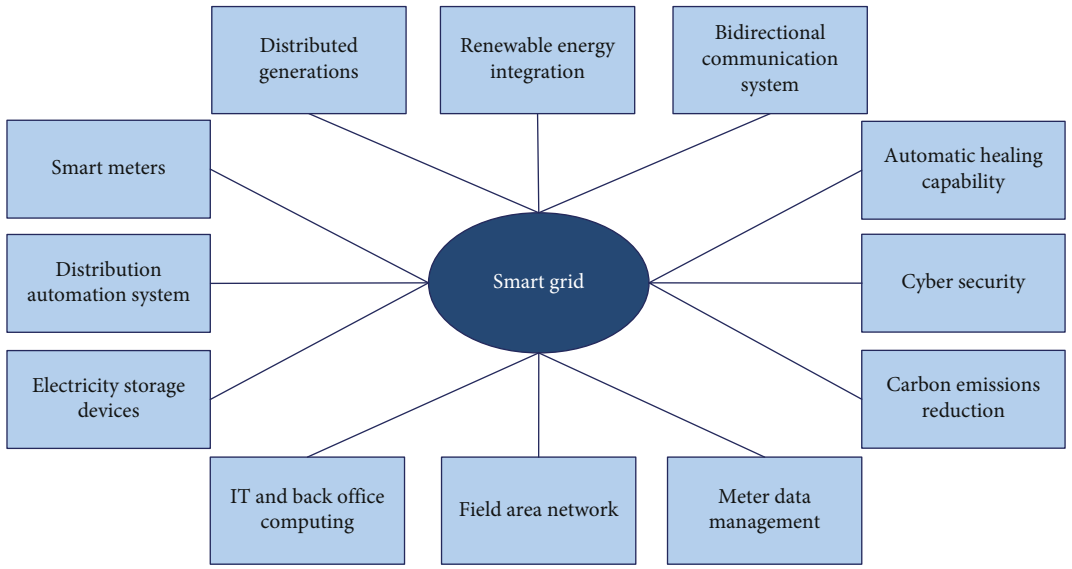



Figure 6: Smart grid features.

with a high level of data protection, convenience, and robustness, the SG must provide one feature: Automatic Healing Capability (AHC). This function comprises automatic identification of unstable system conditions, such as overcurrent, fault current and surge voltage, and information transmission from the central control room and fault or disruption healing/recovery capability.

*3.1.6. Carbon Emission Reduction.* The SG is called Green Grid. Since it can integrate renewable energy sources into the grid and efficient energy production and distribution, SG technology can help reduce carbon emissions by a significant amount.

*3.1.7. Meter Data Management.* The key component of Advanced Metering Infrastructure (AMI) is data management systems of the meter [18]. Meter data management (MDM) is a software that stores and manages enormous amounts of generated data by SM systems over time.

*3.1.8. Field Area Networks.* In power delivery, field networks help build impregnable connectivity between various field equipment, such as transformers, distributors, and smart electronic devices. Near field instruments, several electrical sensors are mounted [19].

*3.1.9. Electricity Storage Devices.* Energy storage systems in many mobile devices have found excellent applications. Therefore, the environmentally safe products replace the standard battery-acid metal storage equipment, requiring more charging time and less acid use. Based on the SG feature and application, Table 3 presents the contribution of some published work.

Table 3: Comparison survey paper related to smart grid.

| Ref | Country | Year | Publication | Sources | Contribution |
|---|---|---|---|---|---|
| [6] | USA | 2012 | J | Science Direct | This article presents an overview of the various network technologies and their contexts. In addition, networking methods, quality of service (QoS), and various optimization problems were briefly addressed. |
| [8] | Mexico | 2012 | J | IEEE | The authors discussed the study on the intelligent delivery system and intelligent metering system in this article. In addition, the authors presented feature analysis of distributed asset optimization, alignment, and connectivity sensors of large-scale deployment of AMIs. |
| [9] | Russia | 2013 | J | IEEE | The authors of this survey looked at the connectivity required for SG and renewable energy sources. They believe that by using smart grid technology, consumers can reduce the peak to average ratio. |
| [20] | USA | 2011 | J | IEEE | The author discus the uses of SMs, benefits, difficulties, and drawbacks according to the energy sector. |
| [21] | Netherlands | 2015 | J | Science Direct | The survey proposes the role of SG power electronics component role, control techniques, renewable energy resource integration, intelligent communication, and metering technologies. Additionally, the authors developed the idea and application of electricity inversion systems with the intelligent grid. |
| [22] | USA | 2012 | J | Science Direct | A survey of smart grid connectivity architecture routing protocols was addressed. Authors have also outlined QoS, security problems, benefits, and drawbacks of current communication protocols. |
| [23] | USA | 2010 | J | Science Direct | This study contains a comprehensive review of integration and the problems associated with hybrid electric vehicles. Different energy control schemes are proposed to mitigate problems relating to hybrid plug-in vehicles (PHEVs). |
| [24] | USA | 2013 | J | Science Direct | A discussion of cyber security issues and risks in SG infrastructure is discussed. Additionally, they conducted on contact, security, and protocol specifications. |
| [25] | Russia | 2010 | J | IEEE | The paper's various approaches for improving the electric grid are proposed. Different methods are used for a stable power grid system, real-time information transfer, reporting, SM using and automation, and improved transmission control system. |
| [26] | China | 2015 | J | Google Scholar | The authors looked at key problems such as communication systems and cyber security issues and possible solutions. They also contribute to potential smart grid research directions. |
| [27] | USA | 2016 | J | Science Direct | The authors examined privacy approaches and issues to achieve the most safety of SG system data transfer. They have discussed several issues on vehicle-to-grid (V2G) and their solutions. |
| [28] | Pakistan | 2016 | J | Science Direct | A thorough examination of demand side management (DSM) and load forecasting are discussed in this paper. To minimize the peak to average ratio, models and forms of dynamic pricing models and load forecasting are briefly discussed. |
| [29] | USA | 2015 | J | energies | This study is focused on V2G network authentication protocols, home and wide area network authentication protocols, and different access control patterns. Also, they discussed the problems of reliable authentication in the smart grid. |
| [30] | Italy | 2014 | J | Science Direct | This survey article examines demand response (DR) and SG technologies in depth. The DR would aid in the reduction of capital and operating costs for smart grid technologies. |
| [31] | Austria | 2017 | C | IEEE | This study focused on start-up approaches in different technical characteristics and blockchain technology-based standard revelation on microgrid and peer-to-peer trading. |
| [32] | Pakistan | 2017 | J | Elsevier | This survey presents the smart grid communication network by a multilayer approach like as Home Area Network (HAN), Neighborhood Area Network (NAN), and Wide Area Network (WAN). |
| [33] | Pakistan | 2018 | J | Elsevier | The goal of this review is to reveal and investigate the current technologies of the smart grid.<br>This study highlights the Architectural Model focusing DR Program (DRP), DSM, and consumer empowerment (CE). Additionally, the study presents a detail discussion on the communication technologies for the power systems like virtual, intergrid, picogrid, nanogrid, and microgrid system. |
| [34] | China and Singapore | 2018 | J | MDPI | Blockchain energy Internet and their challenges |

Table 3: Continued.

| Ref | Country | Year | Publication | Sources | Contribution |
|---|---|---|---|---|---|
| [35] | China and USA | 2019 | J | MDPI | Energy trading in blockchain |
| [36] | Australia and China | 2019 | J | IEEE | Theoretical framework and testbed study of blockchain in intelligent grid |
| [37] | Singapore | 2020 | J | IEEE | Identify the significant challenges and issues on smart grid addressing blockchain technology |
| [38] | India | 2020 | J | Elsevier | Smart grid applicable various technologies, communication system, future, and opportunity of smart grid |
| [39] | UK | 2021 | J | IET | Multidimensional blockchain technology in smart-grid |
| [40] | China | 2021 | J | Springer | Hybrid blockchain technology (public and private) using 5G network in smart grid |

*3.2. Attack in Smart Grid.* Scanning, surveillance, maintaining, and manipulation are the major four access and measures to use by hackers to target the devices and gain access and control [12]. The attacker collected and gathered information to their target through the first phase, reconnaissance. In the second stage, they take attempts to locate the system's vulnerabilities. These movements are designed to learn and identify the service methods on the open port operating system individually and their flaws. They make an attempt to gain and concession the complete control system during the goal exploitation period. When the target administration access is gained, then the final move must be complete and continuously can access. This is consummate by installing an undetectable and stealthy program, consenting them to simply back to the target system. In SG, security criteria are a concession with attackers [1] following the same steps. They apply various methods to compromise a specific system in the SG at each level. As a result, these steps can be used to classify attacks. The types of attacks that occur during each stage are presented in Figure 7. It depicts the variety of attacks that could occur during the exploitation process. The attacks and the malicious activities have occurred during every step.

*3.2.1. Reconnaissance.* Attacks such as traffic analysis and social engineering are part of the reconnaissance process. Instead of technological skills, emphasize human interaction and social skills in social engineering (SE). An attacker applies persuasion and communication gain to legitimate the user's confidence to obtain private and credential information, i.e., PIN or passwords to log in to the server. For instance, password and phishing attacks have become well-known methods used in SE [41]. The traffic analysis listens to the attack and analyzes network traffic to decide which computers and hosts connect to the network and their IP addresses. The security of information is primarily jeopardized by social engineering and traffic analysis.

*3.2.2. Scanning.* The scanning attack is the next move in discovering all of the computers and hosts on the network that are still alive. Scans can be divided into four categories: IP addresses, ports, utilities, and security flaws, which are all things that need to be considered [42]. An intruder usually begins to identify the network with an IP scan in the hosts connected with their won IP addresses. Then, they explore a little deeper by port, checking to see which ones are available. This scan process is run on any host network that has been discovered. After that, the attacker performs a service scan to determine which service or device is running behind each opened port [41]. The final stage is vulnerability scanning to find the flaws, aims, and vulnerabilities associated with every service system on the target devices to be exploited later. Industrial protocols Modbus and DNP3 are also susceptible to scan attacks. The TCP/Modbus was developed to protect the communication system rather than hack by using the scanning Modbus network technique. Attacker entails sending a harmless message to all network-connected computers to collect their information. Mods scan is a well-known scanner on the SCADA Modbus network that can detect and open TCP/Modbus connections and identify system IP addresses and slave IDs.

*3.2.3. Exploitation.* The SG system components are exploited by malicious activities and attempt to gain control and vulnerabilities over it are included in the third phase, exploitation [41, 43]. Viruses, worms, and Trojan horses popping the human-machine interface (HMI), privacy violations, channel jamming, and integrity breaches, as well as different attacks like denial of service (DOS), man-in-the-middle (MITM), and replay attacks, are all examples of these activities [24]. In the SG, viruses are a program that infects a particular computer or machine. A worm is a program that replicates itself. It spreads across the network, copies itself, and infects the system and other devices. A Trojan horse is a computer program that pretends to do something useful on the target machine. In the context, however, it executes malicious code. An attacker uses this form of malware to infect a computer with a virus or worm.

*3.2.4. Maintaining Access.* In the final stage, the attacker applies a specific attack form for retaining access, such as backdoors, viruses, and Trojan horses, to obtain permanent access to the target. An undetectable program like a backdoor is mounted for the target invisibly to be quickly and easily accessed. Suppose the attacker is successful in surrounding a backdoor into the SCADA server control. In that case, they will be able to initiate a series of attacks in contradiction of the system, which will significantly affect the

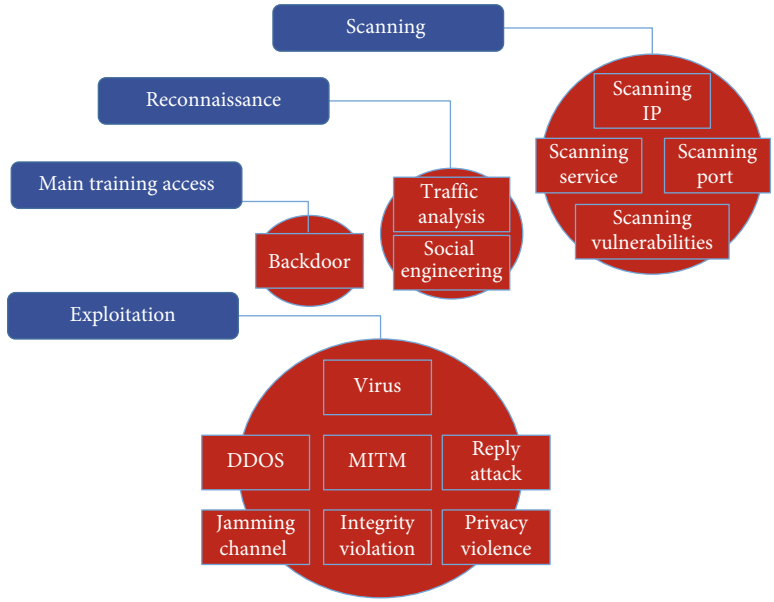



Figure 7: Smart grid-based attacking cycle.

power system [41]. The security criteria on the IT network is defined in the following order based on their importance: confidentiality, honesty, transparency, and availability. They are known as availability, honesty, transparency, and confidentiality in the SG. As a result, attacks that compromise the availability of smart grid networks are considered to be of high severity.

In contrast, attacks that target confidentiality are considered to be of low severity. Every attack has a degree of probability of being carried out in addition to its intensity. Attacks like Stuxnet and Duqu, for example, have a high intensity because they can vandalize industrial control systems and circumvent all security barriers; however, they are complex and complicated [44]. As a result, these viruses are hazardous, but their chances of being spread are poor. The popping HMI attack is another example. It is a highly severe attack requiring specialized networking expertise or extensive experience with industrial and security control systems. While the vulnerability documentation for the devices is publicly accessible, a hacker may quickly launch an attack using open source tools, including Meterpreter and Metasploit. As a result, this attack has a high probability of being carried out.

*3.2.5. Impact of the Cyber Attack.* Significant impacts can cause cyber attack (CA) on economic and physical/technical impact in SG. Though recent research has concentrated on cyber technical/physical attacks on SG, it is also essential to focus more on CA economic risk. The SG has faced a significant economic problem for CA [45, 46], specifically renewable energy resources with high penetration grid-connection mode. The electricity market is a combination of real-time and day-ahead markets [47, 48]. Mainly, the day-ahead market focused on solving the optimization and load forecasting problem at a minimum cost. The optimization problem explains the location marginal price (LMP) of electricity in different locations at each bus (economic dispatch) since load forecasting is affected by false data injection (FDI) CAs in the day-ahead market.

In contrast, the real-time market estimates the generated power and load power for each bus/line. Each line power is required to calculate to achieve the congestion pattern (when estimated line power exceeds the maximum power limit congested), and real-time LMP can estimate this. Thus, FDI state estimation on CA significantly impacts the real-time market that is briefly discussed in [46–51].

The FDI attacks have significant technical/physical impacts on SG. Typically, SG faces steady-state stability and transient impact for FDI attacks. The FDI attacks on steady-state stability significantly impact SG voltage control (AC/DC voltage control in AC-DC SG), demand current/voltage/power management, and energy management [52–56]. Additionally, the CA has an adverse effect on SG steady-state operation; the FDI attacks have impacted SG dynamic and transient stability. Currently, the SG frequency control system can be affected by FDI, but rotor angle stability will be the target [52, 57–61]. Moreover, all of the attacks were occurred in SG protection system.

*3.2.6. Cyber Security.* SG infrastructure must be protected against a variety of threats and attacks. Hackers, attackers, organized crimes and cyber terrorists, certain criminal elements, poorly or careless workers, and industrial rivals may all attack the SG. To abuse the vulnerability system, individual criminals, a group of hackers, attackers, organized criminals, and cyber terrorists may target SG systems and networks. Poorly qualified workers running the system carelessly will create the entire system susceptible to physical/cyber security attacks. Since infrastructure is interrelated across the system, if one part of the SG cyber security (CS) network is targeted, the whole system is at risk, resulting in a complete blackout or system failure. As a result, CS must

be robust sufficient to ensure the system's smooth and effective operation. Data privacy, secrecy, and verification are essential for the infrastructure's security and performance of SG applications. Disregarded cyber security strategies must be implemented to protect data security and supervise the infrastructure to prevent unwanted alterations across the infrastructure [41, 62]. There are several security flaws in SG applications, and each has its unique features. SG applications are vulnerable to diversity of cyber threats that might harm the moderate to more comprehensive level [24, 63]. A jamming attack can only be carried out by accessing the data transmission channel. Stuxnet and other zero-day attacks pose the risk of undiscovered data breaches within control systems. These data breaches may only be identified after the attack is executed [42, 64].

The attacker interrogates the communications between the nodes on the data transmission in an eavesdropping attack [20, 65]. Privacy can be compromised by password theft, traffic analysis on MITM, spoofing attacks, and overhearing. Reliability might be affected by data injection, wormhole data injection, task scheduling, spoofing attacks, and data manipulation. DoS, puppet, buffer overrun, wormhole, jammer, and flooding attacks cause security breaches [20, 65]. Services, applications, end nodes, and networks are the four levels of IoT-based information security solutions for smart infrastructures. Cyber attack (CA) countermeasures include intrusion detection systems (IDSs), sensor verification, compact cryptography, causal inference, and antijamming at the application level. Authorization, anti-DoS, pattern detection, intrusion prevention, cryptography, load balancing [47], ant jamming, and packet filtering are all elements of CA remedies at the network layer. Access control, encryption, pattern detection, authentication, information manipulation, controlled disclosures, and session identifiers are all components of cyber attack solutions at the service layer. CA solutions comprise verification, encrypting, and analysis of the anomaly behavior of software and systems at the end-node layer [43, 66]. Figure 8 shows the security solutions for IoT-based information security applications.

*3.3. Security Requirements in Smart Grid.* CS is a crucial concern due to the risk of CAs and accidents beside this critical industry as it associated with interconnected, according to the EPRI report. Not just malicious threats by malicious workers, corporate espionage, and hackers, but even accidental breaches to the communication system due to software errors, computer faults, and natural disasters must be addressed. Vulnerabilities may enable the attacker to break into a network system, manipulate load conditions and control the gain access in the software to disrupt the power grid in unexpected ways. In the SG system, there are two kinds of data that are shared. Specifically, data and functional data [67]. The logging system, energy trending, power billing, marking, geographical areas' historical reporting, customers' records, and emails are all examples of information. Real-time voltage and current values, capacitor banks, load current, transformer feeder, transformer tap changers, relay position, circuit breakers, and fault positions status are examples of operational information. To secure smart grid networks against any weakness or attack resulting in a power outage, operational data demands a high degree of protection. The smart grid's security criteria and goals are as follows:

*3.3.1. Availability.* The term "availability" discusses the right to use the information and obtain appropriately and accurately. If the SG's contact information is dislocated, that leads to a loss of availability, so the maximum security criteria are necessary [68, 69]. For example, a lack of availability will disrupt the control system's functioning by preventing network information, and operator systems prevent the network's availability. Availability attacks potentially distort, restrict, or hinder data transmission [11, 70]. Additionally, availability attacks in the smart grid prohibit and may disrupt authorized access. It was challenging to target asset availability in the large-scale conventional power grid. ICT is embedded into the power grid's information assets in the smart grid, allowing them to be attacked and completely inaccessible [12, 71].

The DoS attacks are called availability attacks [13, 72]. DoS attacks attempt to interrupt data transmission by obstructing, corrupting, or stalling it. This makes network sources inaccessible. Availability attacks are designed using several methods to overburden networks to ensure that the system does not operate correctly [14, 73]. Attackers transmit significant volumes of traffic to overwhelm the network's transmission connections. For this, the valid data package's presence is lost and not processed in network traffic. IEC 61850 and IP/TCP are IP-based protocol system, which are subject to availability attacks [15, 74]. The most important security prerequisite in SG technology, robust, and comprehensive remedies against availability attacks must be executed. Some successful methods include traffic filtering, big pipes, air-gapped networks, and anomaly detection methods [16, 75]. In SG system, attacks by DoS pose the biggest threat to big data; integrating software solutions in different network layers may prevent DoS attacks significantly.

*3.3.2. Integrity.* In the SG, integrity states securing data against unauthorized modification or degradation. The absence of integrity ensues when data is destroyed, modified, or altered deprived of existence identified [43]. For example, power injection is a destructive attack by an opponent who intelligently modifies calculations and state estimator from the power flow and injection meters. To protect the dignity, material authenticity or nonrepudiation is necessary. Integrity threats are not limited to unauthorized data alteration or injection. Integrity attacks include device impersonation, sparse, and replay attacks. Data integrity threats are prevented via cryptography techniques and approaches [17, 76]. SQL injection and MITM attacks use gaps in the SG to alter, takeover, or corrupt authorized operations.

In SG application system, the data concentrators are linked to SM HAN's. On the other hand, an attacker can use unauthorized data alteration or MITM to impair data transmission among the SM and the data concentrator unit. One of the subdivisions of integrity attacks is load-drop

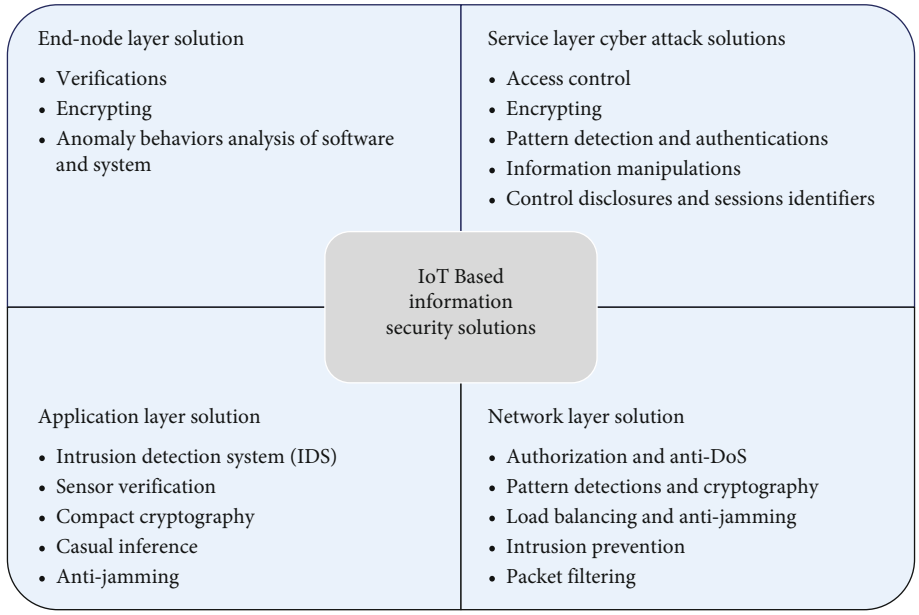



Figure 8: Security solutions for IoT-based applications.

attacks [18, 77]. MITM attacks threaten the systems, and CIA's tried accountability. Security gateways enable the authentication of both target nodes and sources and the confidentiality of data transfer [62, 78]. TLS protocols also include inbuilt asymmetric cryptographic features that can uncover and resolve vulnerabilities efficiently, preventing MITM attacks [63, 75]. By inserting script commands into databases, SQL injection attacks attempt to manipulate databases. SQL injection attacks insert fraudulent demands into the database system to maintain the control system, erase or alter current information, and insert falsified data. SQL injection attacks in the SG network system can be mitigated by using techniques like input type checking, matching the positive pattern, verified static code, database access prevention for remote users, dynamic SQL prevention, and conducting vulnerability scanning. Attackers may use characters such as semicolons; therefore, these characters should be monitored and excluded during type verification [63, 75]. Other kinds of integrity attacks include tampering SCADA systems [64, 79], replay attacks [65, 80], and time synchronization attacks (TSA) [66, 81]. To prevent the mentioned integrity attacks on the SG networks, authentication methods and end-to-end encryption recommendations were used. To launch a confidentiality or integrity attack, attackers can be verified the communication network access and confidential data [19, 24]. As a result, authentication and access control are key to reduce integrity attacks on the SG system.

*3.3.3. Confidentiality.* In particular, confidentiality protects permitted limits on access to and dissemination of records. In other words, the confidentiality criterion includes preventing unauthorized persons, organizations, or systems from disclosing or accessing proprietary or sensitive details [82]. Confidentiality is compromised if materials are released deprived of permission. For instance, information transmitted among the customer and multiple agencies, i.e., metering use, meter management, and billing information, can be private and protective; else, the customer's information can be exploited and changed, or other uses nefarious purposes [24]. Confidentiality attacks have a negative impact on the communication network's functionality. Confidentiality attacks seek toward obtaining the data that should be kept or disclosed confidential between trusted parties. Accessing device memory unlawfully, replay attacks, spoofing payload, and altering the software control of SG are some instances of confidentiality attacks. Password attacks commonly include the social manipulation, dictionary attacks, password sniffing, and password guessing. Social manipulation is a technique of breaking into a scheme utilizing social skills relatively technical skills [15, 74].

Eavesdropping is a kind of passive attack that also compromises data confidentiality [20, 65]. Eavesdropping attacks on local area networks (LANs) in SG networking systems sniff IP packets or intercept wireless transmissions, causing harm to the system's accountability and transparency. Encryption protects sensitive information from eavesdropping attacks [83]. Traffic analysis attacks are passive confidentiality attacks. Interpreting and sniffing the messages permit the attackers to get crucial data around the communication pattern among the networks. Masquerading attacks, also known as impersonating or identity spoofing, are other confidentiality attacks [84]. Other confidentiality attacks include unauthorized access, MITM, and data injection attacks [10, 63, 78, 84, 85]. To prevent confidentiality attacks, smart grid equipment must include authentication, data encryption, and awareness of privacy protocols.

*3.3.4. Authentication.* Machine and human authentication is of high importance; besides this, it is also a weakness because it can lead and cause the attacker to gain access to personal and confidential information or illegitimate devices creating

Table 4: General application and assessment details of SG standards and protocols.

| No | Standard | Scope | Type | Range | Applicability | CT | Pby | Ref |
|---|---|---|---|---|---|---|---|---|
| 1 | ISO/IEC 27001 & 27002 | IS management | General and technical | Worldwide | All components | Yes | 2000 | [101, 102] |
| 2 | The State Grid Corporation of China (SGCC) Framework | Management in electric sector | General and technical | China | All components | Yes | 2002 | [111] |
| 3 | IEEE 1686 | Cyber security | Technical | Worldwide | Substations | Yes | 2007 | [105] |
| 4 | IEC 62351 | Security of communication protocols | Technical | Worldwide | All components | Yes | 2007 | [96, 97] |
| 5 | AMI-SER | CS requirements for procurement | Technical | US | AMI | Yes | 2008 | [94] |
| 6 | GB/T 22239 | IS management | Technical | China | All components | Yes | 2008 | [103] |
| 7 | NIST SP 800-64 | CS | Technical | US | Systems in development | yes | 2008 | [109] |
| 8 | NIST SP 800-115 | CS testing and assessment | Technical | US | All components | Yes | 2008 | [108] |
| 9 | ISO/IEC 15408 and 18045 | Security evaluation criteria | Technical | Worldwide | IT products (hardware and software) | No | 2008 (2012) | [104] |
| 10 | DHS catalog | IACS security | Technical | US | IACS (SCADA) | Yes | 2009 | [109] |
| 11 | IEC 62443 (ISA99) | Security of IACS | Technical | Worldwide | All components | Yes | 2009 | [106] |
| 12 | IEC Strategic Group 3 SG | Security of communication protocols and IACS | Technical | Worldwide | All components | Yes | 2009 | [105] |
| 13 | SG Interoperability Panel | Communication protocols | Technical | US | All components | Yes | 2009 | [107, 109] |
| 14 | NIST | Cyber and information security, risk management | General and technical | US* | Enterprise and systems in development | Yes | 2010 | [99, 100] |
| 15 | NRC RG 5.71 | CS of nuclear infrastructure | General | US | All components | Yes | 2010 | [107] |
| 16 | German Standardization Roadmap E-Energy/SG | Energy storage systems' interoperability | Technical | German | Storage | No | 2010 | [107, 110] |
| 17 | ITU-T Smart Grid Focus Group | Security of communication protocols | Technical | Worldwide | All components | Yes | 2010 | [107] |
| 18 | ISO/IEC 27005 | Risk management | General | Worldwide | Enterprise | Yes | 2011 | [101, 102] |
| 19 | European Commission SG Mandate Standardization M/490 | Management in electric sector | General and technical | Europe | All components | Yes | 2011 | [112] |
| 20 | Japanese Industrial Standards Committee Roadmap to International Standardization for SG | Management in electric sector | General and technical | Japan | All components | Yes | 2012 | [107] |
| 21 | CEN-CENELEC-ETSI SG Coordination Group | Management in electric sector | Technical | Worldwide | All components | Yes | 2012 | [95] |
| 22 | NIST SP 800-53 | Information security management | General | US* | Enterprise | Yes | 2013 | [109] |
| 23 | NIST SP 800-82 | IACS security | Technical | US* | IACS (SCADA) | Yes | 2013 | [99] |
| 24 | NERC-CIP | Bulk power system cyber security | General | US | All components | Yes | 2013 | [98] |
| 25 | IEEE Std 2030-2011 | | Technical | Worldwide | Storage | No | 2015 | [108] |

Table 4: Continued.

| No | Standard | Scope | Type | Range | Applicability | CT | Pby | Ref |
|---|---|---|---|---|---|---|---|---|
| | | Energy storage systems' interoperability | | | | | | |
| 26 | Open SG Security Working Group | Security and communication | General and technical | - | All components | Yes | - | [7] |

∗Also it is used world-wide.

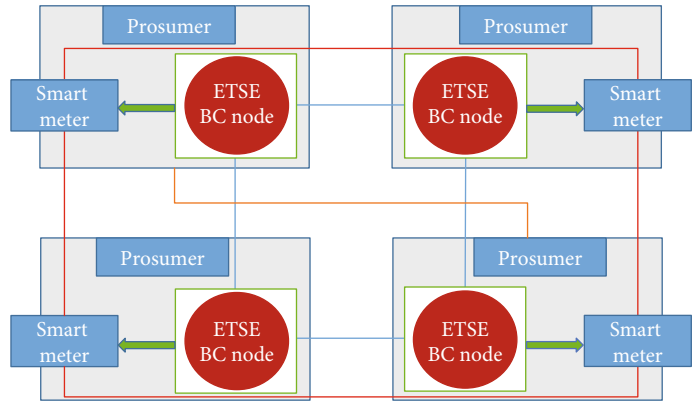



Figure 9: A smart contract enabled permissioned BC in SG [119].

procedure of the SG resources [29]. One of the most challenging aspects of SG communication is authentication. The SCADA systems with communication standards and protocol systems are used in modern SG applications. However, these networks' protocols are often sensitive to MITM attacks, impersonation attacks, and replay attacks. Also, cryptographic keys are applied in the system's different devices that can be exploited. Integrating the SCADA system into Internet communication infrastructure raises security and privacy risks considerably [86]. Mutual authentication between smart devices can be achieved using identity-based authentication and critical public infrastructure (PKI) methods [87]. To avoid authentication attacks, the late-launch dynamic root of trust for measurement (DRTM) technology can be applied to secure the cryptographic key of a specific device [88]. Moreover, to prevent authentication attacks on a mobile RFID-based SG network, an authentication technique can be developed; however, it adds cost and memory [89].

*3.3.5. Authorization.* They are granting access and permission to the computer (also known as access control). Because of the large number of devices besides these people involved in a SG network system, an authorization system is needed to ensure that information and resources are adequately controlled [29]. Unauthorized individuals or systems are prohibited from access to the system without authorization [79]. For this CS, required authorization refers to a decision differentiating between authorized and illegitimate parties based on authentication. If authorization is breached, it may result in security risks. Access control ensures that resources in the smart grid are only accessible by appropriate personnel and entities who have been properly identified [80]. Strict authentication measures should be established to prevent unwanted access to sensitive data and vital assets [24]. Flexible access control, compulsory access control, and role-based access control are examples of authentication techniques that can improve system performance and minimize security risks. As a result, access controls are required to limit the device's network and user's access.

*3.3.6. Nonrepudiation.* Attempt to convince that a device or user's operation cannot be reversed later. For example, an IoT system cannot deny sending a message it has already received. When sensitive resources and knowledge are involved, nonrepudiation becomes a major problem [82, 90]. Data integrity relies heavily on nonrepudiation and legitimacy. Accountability attacks are aimed at changing client information such as account information, payment information, or network operation data such as device status and voltage measurements. Such attacks try to interfere with the source information in the communication network process to interrupt vital communication process in the smart grid [91].

*3.4. Security Standard and Protocol.* There are many security algorithms, standards, and protocols presented to provide the security in the SG system. In an overall standard, generation companies and customers/consumers are linked in distributed ledger and peer-to-peer communication with a trusted third party (TTP) [92]. In SG application, the widely applicable protocol is open smart grid protocol (OSGP) via encryption techniques. However, the rigorous study shows that OSGP encryption mechanism has some weaknesses.

Table 5: Summary of blockchain-based solutions in smart grid.

| Ref | Year | Publication | Country | Sources | Methods/Fields | Finding |
|---|---|---|---|---|---|---|
| [15] | 2017 | J | USA | IEEE | Transactive energy applications | Using a blockchain-based AMI to allow stable and quick energy exchange between DERs |
| [120] | 2018 | J | Romania | MDPI | Decentralized demand response program management | A distributed ledger with blockchain technology for storing SM data (also known as energy transactions) and balancing energy demand and supply |
| [121] | 2019 | J | Norway | IEEE | Traceable and transparent energy usage | In the SG, a permissioned blockchain ensures anonymity and energy protection (traceable and open energy usage). |
| [122] | 2019 | J | China | MDPI | Energy demand and supply information | For energy service providers, a blockchain-based privacy-preserving energy scheduling model |
| [123] | 2017 | J | China | IEEE | Peer-to-peer energy trading | A consortium-based energy blockchain |
| [124] | 2020 | J | Australia | IEEE | Peer-to-peer energy trading | Next-generation energy management technique with reducing peak demand |
| [125] | 2018 | J | China | Springer | Smart grid power trading | A consortium that uses blockchain technology to make energy trading more effective, scalable, and secure |
| [126] | 2019 | C | Canada | IEEE | Energy trading in V2G setup | A hierarchical authentication scheme based on the blockchain for privacy-preserving and transferring the energy from V2G and awarding EVs |
| [127] | 2019 | J | USA | IEEE | Crowdsourced energy system, P2P energy trading, and energy market | A crowdsourced energy infrastructure and energy sharing model aided by blockchain technology |
| [128] | 2019 | J | Singapore | Science Direct | Interconnected cyberphysical systems (CPSs) | ICS-BlockOpS is a blockchain-based industrial control system architecture that ensures organizational data immutability, consistency, and redundancy. |
| [129] | 2018 | J | China | IEEE | Smart grid monitoring | Blockchain is being used to map smart grids between power utilities and consumers for data transparency. |
| [130] | 2019 | J | | IEEE | Industrial CPS | A completely decentralized, blockchain-oriented architecture for a more stable and efficient industrial CPS infrastructure while still addressing the existing shortcomings of cloud-based systems |
| [131] | 2018 | J | China | IEEE | Electric vehicle (EV) charging services in smart community | A stable electric vehicle charging system using a blockchain-based approach combined with contract theory, including an efficient scheduling algorithm and innovative energy allocation in the Internet of Things |
| [132] | 2019 | C | USA | IEEE | ESU charging coordination in smart grid | A decentralized, open, and privacy-preserving synchronize charging platform for ESUs like EVs and batteries built on blockchain technology |
| [133] | 2018 | J | China | IEEE | IoE for EVs and their charging pile management | LNSC, a decentralized security model based on blockchain, has been developed to improve the protection of transactions between EV and charging stations. |
| [134] | 2018 | J | Austria | Springer | EVs charging management | To find the best charging station, energy costing and distance to the EVs were used in a blockchain-assisted automated and privacy-preserving protocol. |
| [135] | 2017 | C | USA | IEEE | Microgrid optimization and control | To solve market price coercion and privacy leakage issues by microgrid aggregators or operators, a decentralized microgrid operating architecture is built on blockchain and the alternating path system of multipliers (ADMM). |
| [136] | 2018 | C | Denmark | IEEE | Voltage control in microgrid | A blockchain-based commensurate management system to reward DERs for their contributions to microgrid voltage regulation |
| [137] | 2019 | J | | IEEE | Grid operation services for TES | A distributed voltage control algorithm for transactive energy systems (TESs) based on blockchains |
| [138] | 2017 | J | China | MDPI | Electricity transactions in microgrid | A decentralized microgrid energy transaction mode based on blockchain and continuous double auction (CDA) to provide autonomous transactions between distributed generations (DGs) and consumers |

Table 5: Continued.

| Ref | Year | Publication | Country | Sources | Methods/Fields | Finding |
|---|---|---|---|---|---|---|
| [139] | 2017 | C | USA | IEEE | Resilient networked microgrids | A decentralized transactive microgrid model |
| [140] | 2020 | C | USA | IEEE | Utilize cognitive methods and tiered blockchain architecture | Scalable blockchain distributed adaptive protection and network architecture with data exchange security |
| [141] | 2020 | C | India | IEEE | Challenges of blockchain technology for rural electrification | Make revenue as per budget using blockchain peer-to-peer trading; smart metering continues accuracy and transparency of energy transactions. |
| [142] | 2020 | J | Algeria | Elsevier | Fog-based SCADA systems for cyber security | The security vulnerability and requirement solution system are classified, i.e., authentication solution, intrusion detection and management systems, and privacy-preserving solutions. |
| [143] | 2021 | J | China | Hindawi | Authentication and authorization protocol | Blockchain-based power system protocol integrates resource authorization and immutable ledger characteristics and identifies decentralized authentication in smart grid systems. |
| [144] | 2021 | J | China | IEEE | Peer-to-peer energy trading | Five-layer design of local energy market based on blockchain |

The weakness is generating and transmitting messages by stream cypher encryption with a new key to another customer where only difference is the first 8 bytes of the key. Another issue is using this generate encryption key to use in authentications. After that, when the BC was introduced and used in the SG application, the security issues are solved. Using the existing security protocol in BC architecture became more secure [93]. The existing SG standards and protocols are as follows: AMI-SER [94], CEN-CENELEC-ETSI SG Coordination Group [95], IEC 62351 [96, 97], NERC-CIP [98], IST [99, 100], ISO/IEC 27001 and 27002 [101, 102], GB/T 22239 [103], ISO/IEC 15408 and 18045 [104], IEC Strategic Group 3 SG [105], IEC 62443 (ISA99) [106], IEC 62443 (ISA99) [107], IEEE Std 2030-2011 [108], IEEE 1686 [105], DHS catalog [109], German Standardization Roadmap E-Energy/SG [108, 110], NIST SP 800-82 [99], NRC RG 5.71 [107], NIST SP 800-53 & 800-64 [110], NIST SP 800-115 [108], Open SG Security Working Group [108], ITU-T SG Focus Group [107], SG Interoperability Panel [108, 109], The State Grid Corporation of China (SGCC) Framework [111], European Commission SG Mandate Standardization M/490 [112], and Japanese Industrial Standards Committee (JISC) Roadmap to International Standardization for SG [107].

The SG standard and protocol summary is presented in Table 4, which presents the applicable scope, types, ranges, applicability, communication technologies (CT), and publication year (Pby).

## 4. Overview of Blockchain (BC)

The BC is a computer network-based archives (big data system), where hackers can access any place worldwide. This is a fully transparent system, where if provisioned for public BC, all service providers and consumers can see the change made and transactions [113]. For this, BC focused on responsiveness in many industries. This is significantly applied in the energy industry, communication, data exchanges, e-treading, and authorization and authentication tamper-proof mechanism. In the point of energy trading, BC technology adopts the grid energy [114].

The block transection of BC is achieved by secure and integrated consensus algorithms [115]. In 2008, the first cryptocurrency, Bitcoin, was introduced in the market and this is the peer-to-peer electronic currency transfer process. In this transection process, without authorization from one party to another party, currency was securely done online transaction by a trusted third party and was first applied in BC technology. This BC technology is significantly and successfully applied in the financial industry, SG, electric vehicle (EV) system, healthcare, IoT, supply chain, etc. [116].

*4.1. Blockchain Mechanism for Smart Grid.* The integration of BC with SG technology is becoming so sophisticated key solutions for facilitating comprehensive security functionality SG technology. The core related interfaces, components, and applications of SG that are critically security dependent are discussed in analyzing the key RQs. The existing centralized ledger system may be transferred by BC technology into a distributed ledger because of the public key algorithm. It also has end-to-end encryption technology and, due to the distribution processing structure, guarantees low costs. The idea of blockchains is generating a lot of research and functional attention right now. A BC is a cryptographic collection of node blocks, where the headers, corresponding transaction data, and auxiliary protection metadata are secured for each block. Intrinsically, the BC supports free connectivity, incorruptibility, openness and secure storage, and transfer of data [117, 118]. In recent years, several BC implementations have arisen beyond initial cryptocurrency applications, like Bitcoins.

Bitcoin's BC system is a public data database that saves the history of Bitcoin value transfers updated regularly. To avoid forgery, this ledger is created using cryptographic technology.

Table 6: Key findings of the IoT-based paper and their primary studies.

| Ref | Publication type | Year | Finding | Types of security applications |
|---|---|---|---|---|
| [150] | J | 2016 | Proof of the pseudonymous concept protocol for secure communications among the IoT devices using Bitcoin in blockchain technology | IoT |
| [151] | J | 2016 | A broad review of the advantages of blockchain-based IoT devices. For example, instead of distributing firmware patches from the middle, IoT devices from one vendor are connected to the same blockchain firmware and spread peer to peer. It is acknowledged that a token is needed. Alternative solutions are presented. | IoT |
| [152] | J | 2018 | Deprived of relying on an essential service like Notary, a blockchain-based system for ensuring the authenticity of Docker images has been developed (offers to defend against denial of service). The importance of a robust blockchain has been recognized. | Internet of Things & Docker |
| [153] | C | 2017 | Blockchain is used to build a multilevel network of IoT computers. Rather than entirely autonomous nodes and miners, the blockchain's security is managed by coordination between layers. | IoT |
| [154] | C | 2017 | A concept for low-power IoT devices can connect with a proper gateway to allow Ethereum blockchain node communication. | IoT |
| [155] | C | 2018 | Introducing "ControlChain," a blockchain-based access power system for IoT devices. Using the same concepts as the Bitcoin blockchain, multiples are proposed. In blockchains, IoT control could be used to handle different aspects. | IoT |
| [156] | C | 2018 | IoT data privacy, access, and trading are the main topics of discussion, suggesting a blockchain solution for each to provide anonymity. The Ethereum platform is being used. | IoT |
| [157] | J | 2017 | Discussion on blockchain strengths and security, mainly with IoT. Highlighting IoT supply chain from manufacturer to end-user security benefits | IoT |
| [158] | J | 2021 | Authentication methods for fog cloud IoT architecture | IoT |
| [159] | J | 2021 | Provides a data mining technique based on Fischer linear discrimination and quadratic discrimination analysis | Big data and IoT |
| [160] | J | 2018 | A thorough examination of IoT protection. What role could blockchain play in addressing the challenges of reducing current security threats to such devices? Mentioning Ethereum as a possible medium for developing smart contracts in an alternative manner | IoT |
| [161] | C | 2018 | Suggestion to create "IoT Chain," a blockchain-based system that allows the authentication to IoT devices and secure access. The Ethereum platform was used to assess the viability of their plan. Authentication servers, key servers, and clients are the three full nodes used by the researchers. The latter serves as the transaction miner, storing data on the blockchain through proof-of-stake and proof-of-work consent mechanisms. The researchers create their Proof-of-Possession system for IoT Chain. | IoT |
| [162] | C | 2019 | Blockchain technology based on various technology in smart grid is discussed in this paper, i.e., cost reduction, communication between provider and consumer, machine-to-machine interaction, and security. | IoT |
| [163] | J | 2019 | User-friendliness and energy optimization in terms of electronics devices controlling and monitoring. This study focused on the interdisciplinary domain that will be helpful for new researchers. | IoT |
| [164] | J | 2020 | Build a green IoT ecosystem based on blockchain technology and discussed the crucial factors and future research direction for a sustainable green IoT ecosystem. | IoT |
| [165] | J | 2020 | Blockchain-based IoT architecture for HANs and NANs in SG system | IoT |
| [166] | J | 2021 | Blockchain-based access control protocol in IoT-enabled SG system | IoT |
| [167] | J | 2021 | IoT-based energy conversion process and inquire on future energy demand on SG system | IoT |

The BC technology could help solve a numerous complex matters relating to the transparency and trustworthiness of fast, distributed, and complex data exchanges and energy transactions. Smart contracts built on the BC often exclude the need to negotiate with third parties, constructing it easier toward monetizing distributed and implementing energy transfers and connections, containing both energy flows and financial transactions (Figure 9). Table 5 presents some BC-based SG applicable methods and findings.

*4.2. Blockchain Mechanism for Energy Trading.* In BC technology, energy trading is necessary for academic research and industrial application with emergency SG electricity generation and distribution. The BC technology is used to reduce the fraudulent act. A certificate is issued for achieving the generators/consumers' trust/guarantee in this energy trading. Implementing BC technology makes the energy trading system easy and helps to reduce the marketing effort and minimize the time. Conventional fossil fuels are diminishing

Table 7: The field of application of the paper.

| Title | Year | Topic | Publication type | Ref |
|---|---|---|---|---|
| Block chain-based… | 2016 | Smart city | J | [168] |
| Building a Robust… | 2016 | Smart energy | J | [169] |
| Security and privacy in decentralized…. | 2016 | Smart property/smart city/smart energy | J | [170] |
| PB-PKI: A Privacy-aware…….. | 2017 | Generic application/smart home | J | [171] |
| Block chain platform for…… | 2016 | Smart manufacturing/smart city/generic application | J | [172] |
| Securing smart cities…….. | 2016 | Generic application/smart home/smart city | C | [173] |
| A block chain connected…… | 2018 | Smart manufacturing/generic application | J | [174] |
| Peer-to-Peer Approaches…… | 2017 | Smart city/smart home | C | [175] |
| Block chain in Internet of Things…… | 2016 | Generic application/smart home | J | [176] |
| Block chain for IoT……. | 2017 | Generic application/smart home | C | [177] |
| CertCoin:A Name Coin …….. | 2014 | Generic application | J | [178] |
| A review on block chain…… | 2017 | Smart property | J | [179] |
| Cloud-based commissioning…….. | 2016 | Smart city/smart manufacturing | C | [180] |
| Ethernam blockchain technology | 2021 | Industry 5.0 | J | [181] |
| A novel method for…….. | 2015 | Smart property | C | [182] |
| Managing IoT devices……. | 2017 | Smart home | C | [183] |
| Authcoin: Validation and Authentication… | 2016 | Generic application | J | [184] |
| Integration of the…… | 2017 | Smart city/smart home/smart energy | C | [185] |
| Towards a novel…… | 2018 | Generic application | J | [186] |
| Converging block chain…….. | 2019 | Generic application | j | [187] |
| Towards block chain…….. | 2017 | Generic application | J | [188] |
| Block chain technology……. | 2017 | Smart manufacturing | J | [189] |
| A Peer-to-Peer…….. | 2014 | Smart home | J | [190] |
| When your sensor……. | 2017 | Generic Application | C | [191] |
| A block chain-based…….. | 2018 | Generic application | J | [192] |
| An IoT electric……. | 2015 | Smart property | C | [193] |
| Decentralized Computation……. | 2015 | Generic application | J | [194] |
| Decentralized Access……. | 2019 | Others | J | [195] |
| Hybrid-IoT: Hybrid…….. | 2018 | Generic application | J | [196] |
| Managing computation…… | 2018 | Generic application | C | [197] |
| An out-of-band…….. | 2018 | Smart home | C | [198] |
| OSCAR: Object security……. | 2015 | Generic application | J | [199] |
| Privacy-preserving and…….. | 2018 | Generic application/smart energy | J | [200] |
| Block chain technologies……. | 2016 | Generic application | C | [201] |
| Digital supply chain…… | 2017 | Generic application/ smart manufacturing | C | [202] |
| Block chain technology…… | 2016 | Generic application | C | [203] |
| Semantic block chain…….. | 2017 | Generic application | J | [204] |
| Blockchain in the construction……... | 2019 | Constriction sector | J | [205] |
| DeliveryCoin: An IDS and blockchain……. | 2019 | Automotive industry | J | [206] |
| Demonstrating blockchain…… | 2019 | Energy trading | J | [207] |
| Phase offset analysis of asymmetric……. | 2019 | Smart grid | J | [208] |
| Dynamic pricing in industrial…….. | 2020 | Smart city | J | [209] |
| Blockchain outlook for…… | 2020 | Smart home | J | [210] |
| HSIC bottleneck based distributed……. | 2020 | Smart grid | J | [211] |
| A blockchain-enabled secure…….. | 2020 | Energy trading | J | [212] |
| An approach for applying blockchain…….. | 2021 | Energy trading | J | [213] |
| Emergence of blockchain-technology……. | 2021 | Energy trading | J | [214] |
| Lightweight Cryptographic Algorithms…... | 2021 | Cyber security | J | [215] |

Table 7: Continued.

| Title | Year | Topic | Publication type | Ref |
|---|---|---|---|---|
| Machine Learning Technologies…… | 2021 | Automotive industry | J | [216] |
| ElStream: An Ensemble Learning…… | 2021 | Machine learning techniques | J | [217] |
| Blockchain and ANFIS empowered…..….. | 2021 | Privacy tracing | J | [218] |
| An improved dynamic thermal……. | 2021 | Big data | J | [219] |
| Toward Blockchain for Edge-of-Things….. | 2021 | Smart grids/big data | J | [220] |
| A peer-to-peer blockchain…….. | 2021 | Smart grid | J | [221] |
| Sustainable Security for…… | 2021 | Cyber security/big data | J | [222] |

rapidly, and researchers and governments worldwide are looking for suitable alternative energy sources like renewable energy. For this, many smaller generated companies produce energy for smaller grid scale and need to connect in the national grid so that consumers can buy [145, 146].

Additionally, the consumer also produces the energy and sells it on the market. The BC system gives an efficient peer-to-peer trading process for local consumers, which generates a small amount of energy. The peer-to-peer topology automatically handles this data and stores it on the public ledger, where all copies are reflected over the network. The BC technology transmits the data and communicates with SG network in a block node. All nodes are connected where every device shares the address and information with previous devices [115].

*4.3. Blockchain Mechanism for Electric Vehicles.* Last few years, EV connection with SGs has been an important and hot topic. Primarily, EV charging systems make more concern to connect with the SG. The power grid system can face severe stress for EV irrelevant charging. Thus, BC technology adopted this problem with several approaches. The BC technology in EV charging integration was discussed in [147–149]. Researchers recommended integrating the EV charging system with the BC technology to be able to find out a near charging station so that EVs can charge. Using this BC technology, EV was used easily to find out the low cost and best location for EV charging station to ensure the privacy and security system.

## 5. Discussion

Since smart grid technology is the most incredible tool for dealing with the complexities of rising energy demand in the future, we should be more mindful of how to use it specifically and wisely. Both underdeveloped and emerging countries, like developed countries, should begin developing policies to make their grid systems smarter and cleaner. There is an adage that says, "cleaner electricity is smarter electricity." And, in this age of environmental degradation, we need a reasonable amount of renewable energy. Smart grid infrastructure assists in the interconnection of national networks. Smart grid systems can transmit energy through a smart web infrastructure, with far-flung transmission and delivery guaranteeing the system's perfection. Under the English Channel, an IF 2000 Under Sea connection creates 2000MW HVDC submarine interconnection that ties up the national grids of France and the United Kingdom. Via a bidirectional transmission and delivery network, this interconnection network assists all countries in meeting their increased energy demand as peak demand rises. Tables 6 and 7 present some funding and application of BC-related published work.

## 6. Conclusion

A SG infrastructure attack does not only affect consumers; it also affects energy providers' profitability. There are several risks to the SG networks that could turn into attacks depending on the adversary's benefit. To make identifying and analyzing such attacks easier, we have divided them into five categories. The paper also looks at and reports on countermeasures for all types of assaults. Extensive research is also required to ensure that IoT and big data on the SG system can protect against adversarial threats without compromising customer trust in the utility provider or dramatically inconvenience. Based on the survey, we still found some research gaps; those are required more concern and improvement for a sustainable BC-based SG and energy trading system. To address issues and challenges, the further improvements and recommendations are as follows:

(i) The BC in different SG systems needs efficient cryptographic schemes

(ii) BC network required penalty and incentive mechanisms

(iii) Advance privacy, security, and data communication exchanges

(iv) The BC-based SG system is required to keep penalty/reward policies

(v) Interoperability limitation among the SG process

(vi) Game theory, cognitive modeling, and deep learning need to add a standard processing technique for benchmark and validation

(vii) SG energy sources required optimal allocation

(viii) Renewable/storage energy system required communication and advance metering for integration with SG, control, and monitoring

(ix) Energy management systems are required considering the burden and computational complexity to design and implement

(x) SG required more focus to handle uncertainty: source intermittency, weather condition, electric vehicle/plug-in-electric vehicle driving pattern, impulsive human behavior during the load connection, and disconnection

## Data Availability

All related data is available in the manuscript.

## Conflicts of Interest

The authors declare no conflict of interest.

## Acknowledgments

This work was supported by the Universiti Kebangsaan Malaysia (UKM) under the FRGS/1/2020/ICT03/UKM/02/6 and GP-2021-K023208.

## References

[1] S. Paul, M. S. Rabbani, R. K. Kundu, and S. M. R. Zaman, "A review of smart technology (smart grid) and its features," in *2014 1st International Conference on Non Conventional Energy (ICONCE 2014)*, pp. 200–203, Kalyani, India, 2014.

[2] I. Mistry, S. Tanwar, S. Tyagi, and N. Kumar, "Blockchain for 5G-enabled IoT for industrial automation: a systematic review, solutions, and challenges," *Mechanical Systems and Signal Processing*, vol. 135, article 106382, 2020.

[3] N. Z. Aitzhan and D. Svetinovic, "Security and privacy in decentralized energy trading through multi-signatures, blockchain and anonymous messaging streams," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 5, pp. 840–852, 2018.

[4] D. Zheng, K. Deng, Y. Zhang, J. Zhao, X. Zheng, and X. Ma, *Smart Grid Power Trading Based on Consortium Blockchain in Internet of Things*, vol. 11336 LNCS, Springer International Publishing, 2018.

[5] V. Hassija, V. Chamola, S. Garg, D. N. G. Krishna, G. Kaddoum, and D. N. K. Jayakody, "A blockchain-based framework for lightweight data sharing and energy trading in V2G network," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 5799–5812, 2020.

[6] J. Gao, Y. Xiao, J. Liu, W. Liang, and C. L. P. Chen, "A survey of communication/networking in smart grids," *Future Generation Computer Systems*, vol. 28, no. 2, pp. 391–404, 2012.

[7] B. Zheng, W. Wei, Y. Chen, Q. Wu, and S. Mei, "A peer-to-peer energy trading market embedded with residential shared energy storage units," *Applied Energy*, vol. 308 Article I.D. 118400, 2022.

[8] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid - the new and improved power grid: a survey," *IEEE Communication Surveys and Tutorials*, vol. 14, no. 4, pp. 944–980, 2012.

[9] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on smart grid communication infrastructures: motivations, requirements and challenges," *IEEE Communication Surveys and Tutorials*, vol. 15, no. 1, pp. 5–20, 2013.

[10] J. Liu, Y. Xiao, S. Li, W. Liang, C. L. P. Chen, and C. L. P. Chen, "Cyber security and privacy issues in smart grids," *IEEE Communication Surveys and Tutorials*, vol. 14, no. 4, pp. 981–997, 2012.

[11] S. M. S. Hussain, S. M. Farooq, and T. S. Ustun, "Implementation of blockchain technology for energy trading with smart meters," *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2019, pp. 1–5, Vellore, India, 2019.

[12] J. A. Abdella and K. Shuaib, "An architecture for blockchain based peer to peer energy trading," in *2019 Sixth International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, pp. 412–419, Granada, Spain, 2019.

[13] F. S. Ali, M. Aloqaily, O. Alfandi, and Ö. Özkasap, "Cyber-physical blockchain-enabled peer-to-peer energy trading," *Computer*, vol. 53, no. 9, pp. 56–65, 2020.

[14] R. K. Kodali, S. Yerroju, and B. Y. K. Yogi, "Blockchain based energy trading," in *TENCON 2018-2018 IEEE Region 10 Conference*, vol. 1, pp. 1778–1783, Jeju, Korea (South), October 2018.

[15] A. A. Habib, M. K. Hasan, M. Mahmud, S. M. A. Motakabber, M. I. Ibrahimya, and S. Islam, "A review: energy storage system and balancing circuits for electric vehicle application," *IET Power Electronics*, vol. 14, no. 1, pp. 1–13, 2021.

[16] S. Kakran and S. Chanana, "Smart operations of smart grids integrated with distributed generation: a review," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 524–535, 2018.

[17] T. Q. D. Khoa, P. T. T. Binh, and H. B. Tran, "Optimizing location and sizing of distributed generation in distribution systems," in *2006 IEEE PES Power Systems Conference and Exposition*, pp. 725–732, Atlanta, GA, USA, 2006.

[18] M. Z. Gunduz and R. Das, "Cyber-security on smart grid: threats and potential solutions," *Computer Networks*, vol. 169, p. 107094, 2020.

[19] S. Ahmed, T. M. Gondal, M. Adil, S. A. Malik, and R. Qureshi, "A survey on communication technologies in smart grid," in *2019 IEEE PES GTD Grand International Conference and Exposition Asia (GTD Asia)*, pp. 7–12, Bangkok, Thailand, 2019.

[20] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni, and N. Gudi, "Smart meters for power grid - challenges, issues, advantages and status," in *2011 IEEE/PES Power Systems Conference and Exposition*, pp. 1–7, Phoenix, AZ, USA, 2011.

[21] I. Colak, E. Kabalci, G. Fulli, and S. Lazarou, "A survey on the contributions of power electronics to smart grid systems," *Renewable and Sustainable Energy Reviews*, vol. 47, no. 1, pp. 562–579, 2015.

[22] N. Saputro, K. Akkaya, and S. Uludag, "A survey of routing protocols for smart grid communications," *Computer Networks*, vol. 56, no. 11, pp. 2742–2771, 2012.

[23] H. E. Brown, S. Suryanarayanan, and G. T. Heydt, "Some characteristics of emerging distribution systems considering the smart grid initiative," *The Electricity Journal*, vol. 23, no. 5, pp. 64–75, 2010.

[24] W. Wang and Z. Lu, "Cyber security in the smart grid: survey and challenges," *Computer Networks*, vol. 57, no. 5, pp. 1344–1371, 2013.

[25] G. F. Reed, P. A. Philip, A. Barchowsky, C. J. Lippert, and A. R. Sparacino, "Sample survey of smart grid approaches and technology gap analysis," in *2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe)*, pp. 1–10, Gothenburg, Sweden, 2010.

[26] M. H. F. Wen, K. C. Leung, V. O. K. Li, X. He, and C. C. J. Kuo, "A survey on smart grid communication system," *APSIPA Transactions on Signal and Information Processing*, vol. 4, p. 2015, 2015.

[27] W. Han and Y. Xiao, "Privacy preservation for V2G networks in smart grid: a survey," *Computer Communications*, vol. 91-92, pp. 17–28, 2016.

[28] A. R. Khan, A. Mahmood, A. Safdar, Z. A. Khan, and N. A. Khan, "Load forecasting, dynamic pricing and DSM in smart grid: a review," *Renewable and Sustainable Energy Reviews*, vol. 54, pp. 1311–1322, 2016.

[29] N. Saxena and B. J. Choi, "State of the art authentication, access control, and secure integration in smart grid," *Energies*, vol. 8, no. 10, pp. 11883–11915, 2015.

[30] P. Siano, "Demand response and smart grids–a survey," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461–478, 2014.

[31] A. Goranović, M. Meisel, L. Fotiadis, S. Wilker, A. Treytl, and T. Sauter, "Blockchain applications in microgrids an overview of current projects and concepts," in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, pp. 6153–6158, Beijing, China, October 2017.

[32] S. Alam, M. F. Sohail, S. A. Ghauri, I. M. Qureshi, and N. Aqdas, "Cognitive radio based smart grid communication network," *Renewable and Sustainable Energy Reviews*, vol. 72, pp. 535–548, 2017.

[33] N. Shaukat, S. M. Ali, C. A. Mehmood et al., "A survey on consumers empowerment, communication technologies, and renewable generation penetration within smart grid," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1453–1475, 2018.

[34] J. Wu and N. K. Tran, "Application of blockchain technology in sustainable energy systems: an overview," *Sustainability*, vol. 10, no. 9, p. 3067, 2018.

[35] N. Wang, X. Zhou, X. Lu et al., "When energy trading meets blockchain in electrical power system: the state of the art," *Applied Sciences*, vol. 9, no. 8, p. 1561, 2019.

[36] A. S. Musleh, G. Yao, and S. M. Muyeen, "Blockchain applications in smart grid–review and frameworks," *Ieee Access*, vol. 7, pp. 86746–86757, 2019.

[37] M. B. Mollah, J. Zhao, D. Niyato et al., "Blockchain for future smart grid: a comprehensive survey," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 18–43, 2021.

[38] G. Dileep, "A survey on smart grid technologies and applications," *Renewable Energy*, vol. 146, pp. 2589–2625, 2020.

[39] C. Liu, X. Zhang, K. K. Chai, J. Loo, and Y. Chen, "A survey on blockchain-enabled smart grids: advances, applications and challenges," *IET Smart Cities*, vol. 3, no. 2, pp. 56–78, 2021.

[40] D. Wang, H. Wang, and Y. Fu, "Blockchain-based IoT device identification and management in 5G smart grid," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, 19 pages, 2021.

[41] Z. El Mrabet, H. El Ghazi, N. Kaabouch, and H. El Ghazi, "Cyber-security in smart grid: Survey and challenges," *Computers & Electrical Engineering*, vol. 1, no. 67, pp. 469–482, 2018.

[42] Y. Deng and S. Shukla, "Vulnerabilities and countermeasures – a survey on the cyber security issues in the transmission subsystem of a smart grid," vol. 1, pp. 251–276, 2012.

[43] S. Wang, C. Zhang, and Z. Su, "Detecting nondeterministic payment bugs in Ethereum smart contracts," *Proceedings of the ACM on Programming Languages*, vol. 3, no. OOPSLA, pp. 1–29, 2019.

[44] N. Komninos, E. Philippou, A. Pitsillides, and S. Member, "Survey in smart grid and smart home security : issues , challenges and countermeasures," *IEEE Communication Surveys and Tutorials*, vol. 16, no. 4, pp. 1933–1954, 2014.

[45] C. Zhao, J. He, P. Cheng, and J. Chen, "Analysis of consensus-based distributed economic dispatch under stealthy attacks," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 6, pp. 5107–5117, 2017.

[46] P. Li, Y. Liu, H. Xin, and X. Jiang, "A robust distributed economic dispatch strategy of virtual power plant under cyber-attacks," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4343–4352, 2018.

[47] R. M. S. Priya, S. Bhattacharya, P. K. R. Maddikunta et al., "Load balancing of energy cloud using wind driven and firefly algorithms in internet of everything," *Journal of Parallel and Distributed Computing*, vol. 142, pp. 16–26, 2020.

[48] L. Xie, Y. Mo, and B. Sinopoli, "Integrity data attacks in power market operations," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 659–666, 2011.

[49] M. K. Hasan, M. Mahmud, A. K. M. Ahasan Habib, S. M. A. Motakabber, and S. Islam, "Review of electric vehicle energy storage and management system: standards, issues, and challenges," *Journal of Energy Storage*, vol. 41, p. 102940, 2021.

[50] L. Xie, Y. Mo, and B. Sinopoli, "False data injection attacks in electricity markets," in *2010 First IEEE International Conference on Smart Grid Communications*, pp. 226–231, Gaithersburg, MD, USA, October 2010.

[51] L. Jia, R. J. Thomas, and L. Tong, "Impacts of malicious data on real-time price of electricity market operations," in *2012 45th Hawaii International Conference on System Sciences*, pp. 1907–1914, Maui, HI, USA, January 2012.

[52] S. Sahoo, S. Mishra, J. C. Peng, and T. Dragicevic, "A stealth cyber-attack detection strategy for DC microgrids," *IEEE Transactions on Power Electronics*, vol. 34, no. 8, pp. 8162–8174, 2019.

[53] X. Liu, M. Shahidehpour, Y. Cao, L. Wu, W. Wei, and X. Liu, "Microgrid risk analysis considering the impact of cyber attacks on solar PV and ESS control systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 3, pp. 1330–1339, 2017.

[54] S. Gholami, S. Saha, and M. Aldeen, "A cyber attack resilient control for distributed energy resources," in *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, pp. 1–6, Torino, Italy, September 2017.

[55] O. A. Beg, T. T. Johnson, and A. Davoudi, "Detection of false-data injection attacks in cyber-physical DC microgrids," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2693–2703, 2017.

[56] J. Hao, E. Kang, J. Sun et al., "An adaptive Markov strategy for defending smart grid false data injection from malicious attackers," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2398–2408, 2018.

[57] A. Farraj, E. Hammad, and D. Kundur, "On the impact of cyber attacks on data integrity in storage-based transient stability control," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 3322–3333, 2017.

[58] A. Farraj, E. Hammad, and D. Kundur, "A systematic approach to delay adaptive control design for smart grids," in *Proceedings of the IEEE International Conference on Smart Grid Communications*, pp. 768–773, Miami, FL, USA, November 2015.

[59] A. Farraj, E. Hammad, and D. Kundur, "Enhancing the performance of controlled distributed energy resources in noisy communication environments," in *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering*, pp. 1–4, Vancouver, BC, Canada, May 2016.

[60] A. Farraj, E. Hammad, and D. Kundur, "A cyber-physical control framework for transient stability in smart grids," *IEEE Transactions on Smart Grid*, vol. 9, pp. 1205–1215, 2018.

[61] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye, "Detecting false data injection attacks on DC state estimation," in *Proceedings of the Preprints 1st Workshop Secure Control Systems (CPSWEEK)*, pp. 1–9, Stockholm, Sweden, April 2010.

[62] A. O. Otuoze, M. W. Mustafa, and R. M. Larik, "Smart grids security challenges: classification by sources of threats," *Journal of Electrical Systems and Information Technology*, vol. 5, no. 3, pp. 468–483, 2018.

[63] A. Procopiou and N. Komninos, "Current and future threats framework in smart grid domain," in *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 1852–1857, Shenyang, China, 2015.

[64] P. Eder-Neuhauser, T. Zseby, J. Fabini, and G. Vormayr, "Cyber attack models for smart grid environments," *Sustainable Energy, Grids and Networks*, vol. 12, pp. 10–29, 2017.

[65] M. Z. Gunduz and R. Das, "Analysis of cyber-attacks on smart grid applications," in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, pp. 1–5, Malatya, Turkey, 2018.

[66] J. Pacheco and S. Hariri, "IoT security framework for smart cyber infrastructures," in *2016 IEEE 1st International Workshops on Foundations and Applications of Self-Systems (FAS-W)*, pp. 242–247, Augsburg, Germany, 2016.

[67] T. Bhatt, C. Kotwal, and N. Chaubey, "Survey on smart grid : threats , vulnerabilities and security protocol," *International Journal of Electrical, Electronics and Computer Systems(I-JEECS)*, vol. 6, p. 340, 2017.

[68] N. Nurelmadina, M. K. Hasan, I. Memon et al., "A systematic review on cognitive radio in low power wide area network for industrial IoT applications," *Sustainability*, vol. 13, no. 1, p. 338, 2021.

[69] M. F. Ali, N. A. Abu, and N. Harum, "A novel session payment system via Internet of Things (IOT)," *International Journal of Applied Engineering Research*, vol. 12, no. 23, pp. 13444–13450, 2017.

[70] R. K. Pandey and M. Misra, "Cyber security threats — smart grid infrastructure," in *2016 National Power Systems Conference (NPSC)*, pp. 1–6, Bhubaneswar, India, 2016.

[71] C. Bekara, "Security issues and challenges for the IoT-based smart grid," *Procedia Computer Science*, vol. 34, pp. 532–537, 2014.

[72] K. I. Sgouras, A. D. Birda, and D. P. Labridis, "Cyber attack impact on critical smart grid infrastructures," in *ISGT 2014*, pp. 1–5, Washington, DC, USA, 2014.

[73] R. Kaur, A. L. Sangal, and K. Kumar, "Modeling and simulation of DDoS attack using omnet++," in *2014 International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 220–225, Noida, India, 2014.

[74] Y. Yang, T. Littler, S. Sezer, K. McLaughlin, and H. F. Wang, "Impact of cyber-security issues on smart grid," in *2011 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies*, pp. 1–7, Manchester, UK, 2011.

[75] G. Bedi, G. K. Venayagamoorthy, R. Singh, R. R. Brooks, and K. C. Wang, "Review of Internet of Things (IoT) in electric power and energy systems," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 847–870, 2018.

[76] S. Shapsough, F. Qatan, R. Aburukba, F. Aloul, and A. R. Al Ali, "Smart grid cyber security: challenges and solutions," in *2015 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE)*, pp. 170–175, Offenburg, Germany, 2015.

[77] E. B. Rice and A. AlMajali, "Mitigating the risk of cyber attack on smart grid systems," *Procedia Computer Science*, vol. 28, pp. 575–582, 2014, 28.

[78] D. Acarali, K. R. Rao, M. Rajarajan, D. Chema, and M. Ginzburg, "Modelling smart grid IT-OT dependencies for DDoS impact propagation," *Computers & Security*, vol. 12, p. 102528, 2022.

[79] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on cyber security for smart grid communications," *IEEE Communication Surveys and Tutorials*, vol. 14, no. 4, pp. 998–1010, 2012.

[80] D. B. Rawat and C. Bajracharya, "Cyber security for smart grid systems: status, challenges and perspectives," in *SouthEastCon 2015*, pp. 1–6, Fort Lauderdale, FL, USA, 2015.

[81] Z. A. Baig and A. R. Amoudi, "An analysis of smart grid attacks and countermeasures," *The Journal of Communication*, vol. 8, no. 8, pp. 473–479, 2013.

[82] D. Minoli and B. Occhiogrosso, "Blockchain mechanisms for IoT security," *Internet of Things*, vol. 1-2, pp. 1–13, 2018.

[83] V. Delgado-gomes, J. F. Martins, C. Lima, and P. N. Borza, "Smart grid security issues," in *2015 9th International Conference on Compatibility and Power Electronics (CPE)*, pp. 534–538, Costa da Caparica, Portugal, 2015.

[84] Carlos Lopez, Arman Sargolzaei, Hugo Santana, and Carlos Huerta, "Smart grid cyber security: an overview of threats and countermeasures," *Journal of Energy and Power Engineering*, vol. 9, no. 7, pp. 632–647, 2015.

[85] N. Komninos, E. Philippou, and A. Pitsillides, "Survey in smart grid and smart home security: issues, challenges and countermeasures," *IEEE Communication Surveys and Tutorials*, vol. 16, no. 4, pp. 1933–1954, 2014.

[86] G. N. Ericsson, "Cyber security and power system communication—essential parts of a smart grid infrastructure," *IEEE Transactions on Power Delivery*, vol. 25, no. 3, pp. 1501–1507, 2010.

[87] S. Lee, J. Bong, S. Shin, and Y. Shin, "A security mechanism of smart grid AMI network through smart device mutual authentication," in *The International Conference on Information Networking 2014 (ICOIN2014)*, pp. 592–595, Phuket, Thailand, 2014.

[88] A. J. Paverd and A. P. Martin, "Hardware Security for Device Authentication in the Smart Grid," in *Smart Grid Security. SmartGridSec 2012*, vol. 7823Springer, Berlin, Heidelberg.

[89] B. Vaidya, D. Makrakis, and H. T. Mouftah, "Authentication mechanism for mobile RFID based smart grid network," in *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 1–6, Toronto, ON, Canada, 2014.

[90] M. Hossain, R. Hasan, and A. Skjellum, "Securing the Internet of Things: a meta-study of challenges, approaches, and open problems," in *2017 IEEE 37th International Conference*

*on Distributed Computing Systems Workshops (ICDCSW)*, pp. 220–225, Atlanta, GA, USA, 2017.

[91] B. Khelifa, "Security concerns in smart grids: threats, vulnerabilities and countermeasures," in *2015 3rd International Renewable and Sustainable Energy Conference (IRSEC)*, pp. 1–6, Marrakech, Morocco, 2015.

[92] R. Das and M. Z. Gündüz, "Analysis of cyber-attacks in IoT-based critical infrastructures," *International Journal of Information Security Science*, vol. 8, no. 4, pp. 122–133, 2020.

[93] M. Shrestha, C. Johansen, J. Noll, and D. Roverso, "A methodology for security classification applied to smart grid infrastructures," *International Journal of Critical Infrastructure Protection*, vol. 28, p. 100342, 2020.

[94] F. Nejabatkhah, Y. W. Li, H. Liang, and R. Reza Ahrabi, "Cyber-security of smart microgrids: a survey," *Energies*, vol. 14, no. 1, p. 27, 2021.

[95] CEN-CENELEC-ETSI Smart Grid Coordination Group, "Smart Grid Reference Architecture," pp. 1–107, 2012.

[96] R. Schlegel, S. Obermeier, and J. Schneider, "Assessing the security of IEC 62351," in *3rd International Symposium for ICS & SCADA Cyber Security Research 2015 (ICS-CSR 2015)*, pp. 11–19, Germany, 2015.

[97] S. S. Hussain, T. S. Ustun, and A. Kalam, "A review of IEC 62351 security mechanisms for IEC 61850 message exchanges," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 9, pp. 5643–5654, 2020.

[98] North American Electric Reliability Corporation, "Critical Infrastructure Protection," April 2021https://www.nerc.com/pa/stand/Pages/ReliabilityStandardsUnitedStates.aspx?jurisdiction=United%20States.

[99] Archived NIST Technical Series Publication, "Smart Grid Cyber Security," 2021, https://nvlpubs.nist.gov/nistpubs/ir/2010/NIST.IR.7628.pdf.

[100] A. Gopstein, A. R. Goldstein, D. Anand, and P. A. Boynton, "Summary report on NIST smart grid testbeds and collaborations workshops," 2021.

[101] E. Kurniawan and I. Riadi, "Security level analysis of academic information systems based on standard ISO 27002: 2003 using SSE-CMM," 2018, https://arxiv.org/abs/1802.03613.

[102] V. Diamantopoulou, A. Tsohou, and M. Karyda, "From ISO/IEC 27002: 2013 information security controls to personal data protection controls: guidelines for GDPR compliance," in *Computer Security*, pp. 238–257, Springer, Cham, 2019.

[103] M. A. Li, Z. H. U. Guobang, and L. U. Lei, "Baseline for classified protection of cybersecurity (GB/T 22239-2019) standard interpretation," *Netinfo Security*, vol. 19, no. 2, p. 77, 2019.

[104] S. Dotsenko, O. Illiashenko, S. Kamenskyi, and V. Kharchenko, "Integrated model of knowledge management for security of information technologies: standards ISO/IEC 15408 and ISO/IEC 18045," *Information & Security*, vol. 43, no. 3, pp. 305–317, 2019.

[105] R. Leszczyna, "A review of standards with cybersecurity requirements for smart grid," *Computers & Security*, vol. 77, pp. 262–276, 2018.

[106] D. Dolezilek, D. Gammel, and W. Fernandes, "Cybersecurity based on IEC 62351 and IEC 62443 for IEC 61850 systems," in *15th International Conference on Developments in Power System Protection (DPSP 2020)*, pp. 1–16, Liverpool, UK, 2020.

[107] R. Leszczyna, "Standards on cyber security assessment of smart grid," *International Journal of Critical Infrastructure Protection*, vol. 22, pp. 70–89, 2018.

[108] IEEE Guide for Smart Grid, "IEEE Guide for Smart Grid Interoperability of Energy Technology and Information Technology Operation with the Electric Power System (EPS), End-Use Applications, and Loads," 2021, https://www.techstreet.com/standards/ieee/2030_2011?product_id=1781311#full.

[109] R. Leszczyna, "Standards with cybersecurity controls for smart grid—a systematic analysis," *International Journal of Communication Systems*, vol. 32, no. 6, article e3910, 2019.

[110] R. Leszczyna, "Cybersecurity and privacy in standards for smart grids - a comprehensive survey," *Computer Standards & Interfaces*, vol. 56, pp. 62–73, 2018.

[111] X. Yi-chong, "China's giant state-owned enterprises as policy advocates: the case of the state grid corporation of China," *The China Journal*, vol. 79, no. 1, pp. 21–39, 2018.

[112] M. Sanduleac, "Unbundled Smart meters in the new smart grid era: assessment on compatibility with European standardisation efforts and with IoT features," in *2018 19th IEEE Mediterranean Electrotechnical Conference (MELECON)*, pp. 35–41, Marrakech, Morocco, May 2018.

[113] S. Höhne and V. Tiberius, "Powered by blockchain: forecasting blockchain use in the electricity market," *International Journal of Energy Sector Management*, vol. 14, no. 6, pp. 1221–1238, 2020.

[114] M. Mylrea and S. N. G. Gourisetti, "Blockchain for smart grid resilience: exchanging distributed energy at speed, scale and security," in *2017 Resilience Week (RWS)*, pp. 18–23, Wilmington, DE, USA, September 2017.

[115] O. Samuel, N. Javaid, T. A. Alghamdi, and N. Kumar, "Towards sustainable smart cities: A secure and scalable trading system for residential homes using blockchain and artificial intelligence," *Sustainable Cities and Society*, vol. 76, p. 103371, 2022.

[116] A. Hajizadeh and S. M. Hakimi, "Blockchain in Decentralized Demand-Side Control of Microgrids," in *Blockchain-Based Smart Grids*, pp. 145–167, Academic Press, 2020.

[117] N. Kshetri, "Blockchain's roles in strengthening cybersecurity and protecting privacy," *Telecommunications Policy*, vol. 41, no. 10, pp. 1027–1038, 2017.

[118] A. Panarello, N. Tapas, G. Merlino, F. Longo, and A. Puliafito, "Blockchain and iot integration: a systematic survey," *Sensors*, vol. 18, no. 8, p. 2575, 2018.

[119] F. Lombardi, L. Aniello, S. De Angelis, A. Margheri, and V. Sassone, "A blockchain-based infrastructure for reliable and cost-effective IoT-aided smart grids," in *Living in the Internet of Things: Cybersecurity of the IoT - 2018*, p. 6, London, UK, 2018.

[120] C. Pop, T. Cioara, M. Antal, I. Anghel, I. Salomie, and M. Bertoncini, "Blockchain based decentralized management of demand response programs in smart energy grids," *Sensors*, vol. 18, no. 2, p. 162, 2018.

[121] K. Gai, Y. Wu, L. Zhu, L. Xu, and Y. Zhang, "Permissioned blockchain and edge computing empowered privacy-preserving smart grid networks," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7992–8004, 2019.

[122] S. Tan, X. Wang, and C. Jiang, "Privacy-preserving energy scheduling for ESCOs based on energy blockchain network," *Energies*, vol. 12, no. 8, p. 1530, 2019.

[123] Z. Li, J. Kang, R. Yu, D. Ye, Q. Deng, and Y. Zhang, "Consortium blockchain for secure energy trading in industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3690–3700, 2017.

[124] W. Tushar, T. K. Saha, C. Yuen, D. Smith, and H. V. Poor, "Peer-to-peer trading in electricity networks: An overview," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3185–3200, 2020.

[125] D. Zheng, K. Deng, Y. Zhang, J. Zhao, X. Zheng, and X. Ma, "Smart grid power trading based on consortium blockchain in Internet of Things," in *Algorithms and Architectures for Parallel Processing. ICA3PP 2018*, vol. 11336 of *Lecture Notes in Computer Science*, Cham, 2018.

[126] S. Garg, K. Kaur, G. Kaddoum, F. Gagnon, and J. J. P. C. Rodrigues, "An efficient blockchain-based hierarchical authentication mechanism for energy trading in V2G environment," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, Shanghai, China, 2019.

[127] S. Wang, A. F. Taha, J. Wang, K. Kvaternik, and A. Hahn, "Energy crowdsourcing and peer-to-peer energy trading in blockchain-enabled smart grids," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 8, pp. 1612–1623, 2019.

[128] A. Maw, S. Adepu, and A. Mathur, "ICS-BlockOpS: blockchain for operational data security in industrial control system," *Pervasive and Mobile Computing*, vol. 59, p. 101048, 2019.

[129] J. Gao, K. O. Asamoah, E. B. Sifah et al., "GridMonitoring: secured sovereign blockchain based monitoring on smart grid," *IEEE Access*, vol. 6, pp. 9917–9925, 2018.

[130] J. Wan, J. Li, M. Imran, D. Li, and Fazal-e-Amin, "A blockchain-based solution for enhancing security and privacy in smart factory," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3652–3660, 2019.

[131] Z. Su, Y. Wang, Q. Xu, M. Fei, Y. C. Tian, and N. Zhang, "A secure charging scheme for electric vehicles with smart communities in energy blockchain," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4601–4613, 2019.

[132] M. Baza, M. Nabil, M. Ismail, M. Mahmoud, E. Serpedin, and M. A. Rahman, "Blockchain-based charging coordination mechanism for smart grid energy storage units," in *2019 IEEE International Conference on Blockchain (Blockchain)*, pp. 504–509, Atlanta, GA, USA, 2019.

[133] X. Huang, C. Xu, P. Wang, and H. Liu, "LNSC: a security model for electric vehicle and charging pile management based on blockchain ecosystem," *IEEE Access*, vol. 6, pp. 13565–13574, 2018.

[134] F. Knirsch, A. Unterweger, and D. Engel, "Privacy-preserving blockchain-based electric vehicle charging with dynamic tariff decisions," *Computer Science - Research and Development*, vol. 33, no. 1–2, pp. 71–79, 2018.

[135] E. Munsing, J. Mather, and S. Moura, "Blockchains for decentralized optimization of energy resources in microgrid networks," in *2017 IEEE Conference on Control Technology and Applications (CCTA)*, pp. 2164–2171, Maui, HI, USA, 2017.

[136] P. Danzi, M. Angjelichinoski, C. Stefanovic, and P. Popovski, "Distributed proportional-fairness control in microgrids via blockchain smart contracts," in *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pp. 45–51, Dresden, Germany, 2017.

[137] S. Saxena, H. Farag, H. K. Turesson, and H. Kim, "Blockchain based grid operation services for transactive energy systems," 2019, https://arxiv.org/abs/1907.08725.

[138] J. Wang, Q. Wang, N. Zhou, and Y. Chi, "A novel electricity transaction mode of microgrids based on blockchain and continuous double auction," *Energies*, vol. 10, no. 12, p. 1971, 2017.

[139] M. Sabounchi and J. Wei, "Towards resilient networked microgrids: blockchain-enabled peer-to-peer electricity trading mechanism," in *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*, pp. 1–5, Beijing, China, 2017.

[140] D. Sikeridis, A. Bidram, M. Devetsikiotis, and M. J. Reno, "A blockchain-based mechanism for secure data exchange in smart grid protection systems," in *2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC)*, pp. 1–6, Las Vegas, NV, USA, January 2020.

[141] V. Kulkarni and K. Kulkarni, "A Blockchain-based smart grid model for rural electrification in India," in *2020 8th International Conference on Smart Grid (icSmartGrid)*, pp. 133–139, Paris, France, June 2020.

[142] M. A. Ferrag, M. Babaghayou, and M. A. Yazici, "Cyber security for fog-based smart grid SCADA systems: solutions and challenges," *Journal of Information Security and Applications*, vol. 52, p. 102500, 2020.

[143] Y. Zhong, M. Zhou, J. Li et al., "Distributed blockchain-based authentication and authorization protocol for smart grid," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 5560621, 15 pages, 2021.

[144] Z. Zeng, M. Dong, W. Miao, M. Zhang, and H. Tang, "A data-driven approach for blockchain-based smart grid system," *IEEE Access*, vol. 9, pp. 70061–70070, 2021.

[145] I. Kouveliotis-Lysikatos, I. Kokos, I. Lamprinos, and N. Hatziargyriou, "Blockchain-powered applications for smart transactive grids," in *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, pp. 1–5, Bucharest, Romania, September 2019.

[146] I. Petri, M. Barati, Y. Rezgui, and O. F. Rana, "Blockchain for energy sharing and trading in distributed prosumer communities," *Computers in Industry*, vol. 123, p. 103282, 2020.

[147] C. Liu, K. K. Chai, X. Zhang, E. T. Lau, and Y. Chen, "Adaptive blockchain-based electric vehicle participation scheme in smart grid platform," *IEEE Access*, vol. 6, pp. 25657–25665, 2018.

[148] S. Chen, J. Ping, Z. Yan, and W. Wei, "Blockchain for decentralized optimization of energy sources: EV charging coordination via blockchain-based charging power quota trading," in *Blockchain-Based Smart Grids*, pp. 169–179, Academic Press, 2020.

[149] P. W. Khan and Y. C. Byun, "Blockchain-based peer-to-peer energy trading and charging payment system for electric vehicles," *Sustainability*, vol. 13, no. 14, p. 7962, 2021.

[150] A. Ouaddah, A. Abou Elkalam, and A. Ait Ouahman, "FairAccess: a new blockchain-based access control framework for the Internet of Things," *Security and Communication Networks*, vol. 9, 5964 pages, 2016.

[151] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the Internet of Things," *IEEE Access*, vol. 4, pp. 2292–2303, 2016.

[152] Q. Xu, C. Jin, M. F. B. M. Rasid, B. Veeravalli, and K. M. M. Aung, "Blockchain-based decentralized content trust for

docker images," *Multimedia Tools and Applications*, vol. 77, no. 14, pp. 18223–18248, 2018.

[153] C. Li and L. J. Zhang, "A blockchain based new secure multi-layer network model for Internet of Things," in *2017 IEEE International Congress on Internet of Things (ICIOT)*, pp. 33–41, Honolulu, HI, USA, 2017.

[154] K. R. Oezyilmaz and A. Yurdakul, *Integrating low-power IoT devices to a blockchain-based infrastructure: work-in-progress*, EMSOFT Companion, 2017.

[155] O. J. A. Pinno, A. R. A. Gregio, and L. C. E. De Bona, "Controlchain: blockchain as a central enabler for access control authorizations in the IoT," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1–6, Singapore, 2017.

[156] Z. Huang, X. Su, Y. Zhang, C. Shi, H. Zhang, and L. Xie, "A decentralized solution for IoT data trusted exchange based-on blockchain," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, pp. 1180–1184, Chengdu, China, 2018.

[157] M. Banerjee, J. Lee, and K. K. R. Choo, "A blockchain future for Internet of Things security: a position paper," *Digital Communications and Networks*, vol. 4, no. 3, pp. 149–160, 2018.

[158] S. Amanlou, M. K. Hasan, and K. A. Bakar, "Lightweight and secure authentication scheme for IoT network based on publish- subscribe fog computing model," *Computer Networks*, vol. 199, p. 108465, 2021.

[159] M. K. Hasan, T. M. Ghazal, A. Alkhalifah et al., "Fischer linear discrimination and quadratic discrimination analysis–based data mining technique for Internet of Things framework for healthcare," *Frontiers in Public Health*, vol. 9, 2021.

[160] O. Alphand, M. Amoretti, T. Claeys et al., "IoTChain: a blockchain security architecture for the Internet of Things," *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, Barcelona, Spain, 2018.

[161] E. F. Jesus, V. R. L. Chicarino, C. V. N. De Albuquerque, and A. A. D. A. Rocha, "A survey of how to use blockchain to secure Internet of Things and the stalker attack," *Security and Communication Networks*, vol. 2018, 27 pages, 2018.

[162] D. Orazgaliyev, Y. Lukpanov, I. A. Ukaegbu, and H. S. K. Nunna, "Towards the application of blockchain technology for smart grids in Kazakhstan," in *2019 21st International Conference on Advanced Communication Technology (ICACT)*, pp. 273–278, PyeongChang, Korea, February 2019.

[163] S. Mugunthan and T. Vijayakumar, "Review on IoT based smart grid architecture implementations," *Journal of Electrical Engineering and Automation*, vol. 10, no. 1, pp. 12–20, 2019.

[164] P. K. Sharma, N. Kumar, and J. H. Park, "Blockchain technology toward green IoT: opportunities and challenges," *IEEE Network*, vol. 34, no. 4, pp. 263–269, 2020.

[165] S. Garlapati, "Blockchain for IOT-based NANs and HANs in smart grid," 2020, https://arxiv.org/abs/2001.00230.

[166] B. Bera, S. Saha, A. K. Das, and A. V. Vasilakos, "Designing blockchain-based access control protocol in iot-enabled smart-grid system," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5744–5761, 2021.

[167] N. Renugadevi, S. Saravanan, and C. M. Naga Sudha, "IoT based smart energy grid for sustainable cites," *Materials Today: Proceedings.*, 2021.

[168] J. Sun, J. Yan, and K. Z. K. Zhang, "Blockchain-based sharing services : what blockchain technology can contribute to smart cities," *Financial Innovation*, vol. 2, no. 1, p. 26, 2016.

[169] Y. Symey, S. Sankaranarayanan, and S. S. N. Binti, *Building a Robust Value Mechanism to Facilitate TransActive Energy*, Energy, 2016.

[170] N. Z. Aitzhan and D. Svetinovic, "Security and privacy in decentralized energy trading through multi-signatures , blockchain and anonymous messaging streams," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 5, pp. 840–852, 2016.

[171] L. Axon and M. Goldsmith, "PB-PKI : a privacy-aware blockchain-based PKI PB-PKI : a privacy-aware blockchain-based PKI," in *Proceedings of the 14th International Joint Conference on e-Business and Telecommunications (ICETE 2017)*, vol. 4, pp. 311–318, Madrid, Spain, 2017.

[172] A. Bahga and V. K. Madisetti, "Blockchain Platform for Industrial Internet of Things," *Journal of Software Engineering and Applications*, pp. 533–546, 2016.

[173] K. Biswas and V. Muthukkumarasamy, "Securing smart cities using blockchain technology," in *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pp. 5–7, Sydney, NSW, Australia, 2016.

[174] S. C. Cha, J. F. Chen, C. Su, and K. H. Yeh, "A blockchain connected gateway for BLE-based devices in the Internet of Things," *IEEE Access*, vol. 6, pp. 24639–24649, 2018.

[175] M. Conoscenti and J. C. De Martin, "Peer-to-Peer Approaches for a Decentralized Private-By-Design Internet of Things," in *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*, Buenos Aires, Argentina, 2017.

[176] A. Dorri, S. S. Kanhere, and R. Jurdak, "Blockchain in Internet of Things : challenges and solutions," 2016, https://arxiv.org/abs/1608.05187.

[177] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Blockchain for IoT security and privacy : the case study of a smart home," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 618–623, Kona, HI, USA, 2017.

[178] B. Qin, J. Huang, Q. Wang, X. Luo, B. Liang, and W. Shi, "Cecoin: A decentralized PKI mitigating MitM attacks," *Future Generation Computer Systems*, vol. 107, pp. 805–815, 2020.

[179] P. Ghuli, U. P. Kumar, and R. Shettar, "A review on blockchain application for decentralized decision of ownership of IoT devices," *Advanced Computer Science & Technology*, vol. 10, no. 8, pp. 2449–2456, 2017.

[180] T. Hardjono and N. Smith, "Cloud-based commissioning of constrained devices using permissioned blockchains," in *IoTPTS '16: Proceedings of the 2nd ACM International Workshop on IoT Privacy, Trust, and Security*, pp. 29–36, 2016.

[181] C. Rupa, D. Midhunchakkaravarthy, M. Kamrul Hasan, H. Alhumyani, and R. A. Saeed, "Industry 5.0: Ethereum blockchain technology based DApp smart contract," *Mathematical Biosciences and Engineering*, vol. 18, no. 5, pp. 7010–7027, 2021.

[182] J. Herbert and A. Litchfield, "A novel method for decentralised peer-to-peer software license validation using

cryptocurrency blockchain technology," in *Proc. 38th Australasian Computer Science Conference (ACSC 2015)*, vol. 159no. January, pp. 27–35, Sydney, Australia, 2015.

[183] S. Huh, S. Cho, and S. Kim, "Managing IoT devices using blockchain platform2017 19th International Conference on Advanced Communication Technology (ICACT), pp. 464–467, PyeongChang, Korea (South), 2017.

[184] B. Leiding, C. H. Cap, T. Mundt, and S. Rashidibajgan, "Authcoin: validation and authentication in decentralized networks," 2016, https://arxiv.org/abs/1609.04955.

[185] A E Commission, "Integration of the blockchain in a smart grid model," in *The 14th International Conference of Young Scientists on Energy Issues (CYSENI) 2017*, pp. 127–134, Kaunas, Lithuania, 2017.

[186] A. Ouaddah, "Towards a Novel Privacy-Preserving Access Control Model Based on Blockchain Technology in IoT," in *Europe and MENA Cooperation Advances in Information and Communication Technologies*, vol. 520 of Advances in Intelligent Systems and Computing,, Springer, Cham.

[187] D. Nettikadan, R. T. Raphael, and B. D. Paul, "Converging blockchain and Internet of Things," *The International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 7, pp. 662–667, 2019.

[188] H. Shafagh, L. Burkhalter, A. Hithnawi, and S. Duquennoy, "Towards blockchain-based auditable storage and sharing of IoT data," in *CCSW '17: Proceedings of the 2017 on Cloud Computing Security Workshop*, pp. 45–50, New York, United States, November 2017.

[189] J. J. Sikorski, J. Haughton, M. Kraft, P. Street, and P. F. Drive, *Blockchain technology in the chemical industry : machine-to-machine electricity market*, University of Cambridge, 2016.

[190] S. Wilkinson, *A Peer-to-Peer Cloud Storage Network*, Finance Magnates, 2014.

[191] S. E. E. Profile, "When Your Sensor Earns Money: Exchanging Data for Cash with Bitcoin," in *2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle Washington, USA, 2017.

[192] Q. Xu, K. M. M. Aung, Y. Zhu, K. L. Yong, and K. L. Yong, "A Blockchain-Based Storage System for Data Analytics in the Internet of Things," *Studies in Computational Intelligence*, vol. 715, pp. 119–138, 2018.

[193] Y. Zhang and J. Wen, "An IoT electric business model based on the protocol of bitcoin," in *2015 18th International Conference on Intelligence in Next Generation Networks*, pp. 184–191, Paris, France, 2015.

[194] G. Zyskind, O. Nathan, and A. Pentland, "Enigma: decentralized computation platform with guaranteed privacy," 2015, https://arxiv.org/abs/1506.03471.

[195] R. Deters, "Decentralized access control with distributed ledgers using blockchain to manage IoT access," in *IEEE International Conference on Industrial Internet (IEEE ICII)*, pp. 248–257, Orlando, FL, USA, November 2019.

[196] G. Sagirlar, B. Carminati, E. Ferrari, J. D. Sheehan, and E. Ragnoli, "Hybrid-IoT: hybrid blockchain architecture for Internet of Things - PoW sub-blockchains," in *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 1007–1016, Halifax, NS, Canada, 2018.

[197] R. B. Chakraborty, M. Pandey, and S. S. Rautaray, "Managing computation load on a blockchain - based multi - layered Internet - of - Things network," *Procedia Computer Science*, vol. 132, pp. 469–476, 2018.

[198] L. Wu, X. Du, W. Wang, and B. Lin, "An out-of-band authentication scheme for Internet of Things using blockchain technology," in *2018 International Conference on Computing, Networking and Communications (ICNC)*, pp. 769–773, Maui, HI, USA, 2018.

[199] M. Vučinić, B. Tourancheau, F. Rousseau, A. Duda, L. Damon, and R. Guizzetti, "OSCAR: object security architecture for the Internet of Things," *Ad Hoc Networks*, vol. 32, pp. 3–16, 2015.

[200] Z. Guan, G. Si, X. Zhang et al., "Privacy-preserving and efficient aggregation based on blockchain for power grid communications in smart communities," *IEEE Communications Magazine*, vol. 56, no. 7, pp. 82–88, 2018.

[201] M. English, S. Auer, and J. Domingue, "Block chain technologies & the semantic web: a framework for symbiotic development," in *Bonn-Aachen International Center for Information Technology Dahlmannstrasse 2, 53113 Bonn*, pp. 47–61, North Rhine-Westphalia, Germany, May 2016.

[202] K. Korpela, J. Hallikas, and T. Dahlberg, "Digital supply chain transformation toward blockchain integration," in *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*, vol. 50, Big Island, Hawaii, USA, January 2017.

[203] M. Mettler and M. A. Hsg, "Blockchain technology in healthcare the revolution starts here," in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pp. 1–3, Munich, Germany, 2016.

[204] M. Ruta, F. Scioscia, S. Ieva, G. Capurso, and E. Di Sciascio, "Semantic blockchain to improve scalability in the Internet of Things," *Open Journal of Internet Of Things (OJIOT)*, vol. 3, no. 1, pp. 46–61, 2017.

[205] J. Li, D. Greenwood, and M. Kassem, "Blockchain in the construction sector: a socio-technical systems framework for the construction industry," in *Advances in informatics and computing in civil and construction engineering*, pp. 51–57, Springer, Cham, 2019.

[206] M. A. Ferrag and L. Maglaras, "DeliveryCoin: An IDS and blockchain-based delivery framework for drone-delivered services," *Computer*, vol. 8, no. 3, p. 58, 2019.

[207] O. Jogunola, M. Hammoudeh, B. Adebisi, and K. Anoh, "Demonstrating blockchain-enabled peer-to-peer energy trading and sharing," in *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pp. 1–4, Edmonton, AB, Canada, May 2019.

[208] M. K. Hasan, S. H. Yousoff, M. M. Ahmed, A. H. A. Hashim, A. F. Ismail, and S. Islam, "Phase offset analysis of asymmetric communications infrastructure in smart grid," *Elektronika ir Elektrotechnika*, vol. 25, no. 2, pp. 67–71, 2019.

[209] H. A. Khattak, K. Tehreem, A. Almogren, Z. Ameer, I. U. Din, and M. Adnan, "Dynamic pricing in industrial Internet of Things: blockchain application for energy management in smart cities," *Journal of Information Security and Applications*, vol. 55, p. 102615, 2020.

[210] H. Hosseinian, H. Shahinzadeh, B. G. Gharehpetian, Z. Azani, and M. Shaneh, "Blockchain outlook for deployment of IoT in distribution networks and smart homes," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 3, pp. 2787–2796, 2020.

[211] M. Akhtaruzzaman, M. K. Hasan, S. R. Kabir, S. N. H. S. Abdullah, M. J. Sadeq, and E. Hossain, "HSIC bottleneck based distributed deep learning model for load forecasting in smart grid with a comprehensive survey," *IEEE Access*, vol. 8, pp. 222977–223008, 2020.

[212] Z. Liu, D. Wang, J. Wang, X. Wang, and H. Li, "A blockchain-enabled secure power trading mechanism for smart grid employing wireless networks," *IEEE Access*, vol. 8, pp. 177745–177756, 2020.

[213] M. Aybar-Mejía, D. Rosario-Weeks, D. Mariano-Hernández, and M. Domínguez-Garabitos, "An approach for applying blockchain technology in centralized electricity markets," *The Electricity Journal*, vol. 34, no. 3, p. 106918, 2021.

[214] M. K. Thukral, "Emergence of blockchain-technology application in peer-to-peer electrical-energy trading: a review," *Clean Energy*, vol. 5, no. 1, pp. 104–123, 2021.

[215] M. K. Hasan, M. Shafiq, S. Islam et al., "Lightweight cryptographic algorithms for guessing attack protection in complex Internet of Things applications," *Complexity*, vol. 2021, Article ID 5540296, 13 pages, 2021.

[216] E. S. Ali, M. K. Hasan, R. Hassan et al., "Machine learning technologies for secure vehicular communication in internet of vehicles: recent advances and applications," *Networks*, vol. 2021, article 8868655, 23 pages, 2021.

[217] A. Abbasi, A. R. Javed, C. Chakraborty, J. Nebhen, W. Zehra, and Z. Jalil, "ElStream: an ensemble learning approach for concept drift detection in dynamic social big data stream learning," *IEEE Access*, vol. 9, pp. 66408–66419, 2021.

[218] B. Aslam, A. R. Javed, C. Chakraborty, J. Nebhen, S. Raqib, and M. Rizwan, "Blockchain and ANFIS empowered IoMT application for privacy preserved contact tracing in COVID-19 pandemic," *Personal and Ubiquitous Computing*, vol. 1-17, 2021.

[219] M. K. Hasan, M. M. Ahmed, S. S. Musa et al., "An improved dynamic thermal current rating model for PMU-based wide area measurement framework for reliability analysis utilizing sensor cloud system," *IEEE Access*, vol. 9, pp. 14446–14458, 2021.

[220] S. N. Ghorpade, M. Zennaro, B. S. Chaudhari, R. A. Saeed, H. Alhumyani, and S. Abdel-Khalek, "A novel enhanced quantum PSO for optimal network configuration in heterogeneous industrial IoT," *IEEE Access.*, vol. 9, pp. 134022–134036, 2021.

[221] M. M. Ahmed, M. K. Hasan, M. Shafiq et al., "A peer-to-peer blockchain based interconnected power system," *Energy Reports*, vol. 7, pp. 7890–7905, 2021.

[222] C. Iwendi, S. U. Rehman, A. R. Javed, S. Khan, and G. Srivastava, "Sustainable security for the Internet of Things using artificial intelligence architectures," *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 3, pp. 1–22, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 2930366, 10 pages
https://doi.org/10.1155/2021/2930366

*Research Article*

# Optimization of VRR for Cold Chain with Minimum Loss Based on Actual Traffic Conditions

**Lishuan Hu**,[1] **Caihong Xiang**,[2] **and Chengming Qi**[1]

[1]*College of Urban Rail Transit and Logistics, Beijing Union University, Beijing, China*
[2]*TAIJI Computer Corporation Limited, Beijing, China*

Correspondence should be addressed to Lishuan Hu; hulishuan@buu.edu.cn

Received 13 August 2021; Accepted 21 October 2021; Published 10 December 2021

Academic Editor: Rajesh Kaluri



Recently, fresh agricultural cold-chain logistics have been greatly developed with the increasing needs of people's life. Reducing costs of cold-chain distribution has become the main object of loss control in logistics enterprises. The objective of this research is to find a set of optimal routes that minimize the total loss, including fuel cost, refrigeration cost, soft time window penalty cost, and cargo damage cost over transit time. In this paper, the definition and model construction of vehicle routing problem (VRP) with multiobjective minimum lost are introduced first. Then, an ant colony optimization (ACO) algorithm combined with Pareto local search (PLS) is put forward to solve the minimum loss model. In order to avoid the influence of complex road conditions during distribution, the distance matrix and the transit time matrix are both derived from the recommended navigation road based on E-map API. At last, a compare experiment between the traditional method and our proposed method is performed. The results indicate that our method has strong applicability and potential advantages in cold-chain logistic and has important practical significance and application value.

## 1. Introduction

According to relevant media reports, rotting fruit, vegetables, and other foods transported by truck alone are worth about 70 billion yuan each year, causing a huge economic waste. Agricultural products especially for fruit, vegetables, fish, meat, etc. require strictly limited temperature, humidity, and time in the process of transportation and storage. As a branch of the logistics industry, cold-chain logistics provides a guarantee for the safe transportation of fresh agricultural products. In the process of transportation from cold storage to customers, a complete cold-chain logistics realizes the temperature control of the whole process of refrigerated and frozen food, as well as the closed environment, storage, and transportation during loading and unloading of goods. By selecting the optimal path in the process of cold-chain logistics transmission, the circulating rate of fruits and vegetables, meat, and aquatic products is reduced; the waste of resources and the cost of logistics costs are reduced too. The vehicle routing problem (VRP) introduced by Dantzig and Ramser [1] plays a central role in the optimization of distribution networks.

Recently, the loss problem of cold-chain distribution has attracted the attention of many scholars and experts. For distribution problem of fresh agricultural products, fuel cost, refrigeration cost, cargo damage cost, and soft time window penalty cost have become the important components of loss control in distribution companies [2]. Li et al. [3] proposed a low-carbon model for fresh food and a genetic simulated annealing algorithm to solve the model in the cold-chain distribution. Yao et al. [4] proposed a minimizing fuel consumption solution to time-dependent VRP with time window. Kim et al. [5] proposed a Markov decision process method to solve a dynamic VRP (DVRP) model with nonstationary transit times under actual traffic congestion. Chen et al. [6] developed a hybrid heuristic algorithm including harmony search and neighborhood descent to solve DVRP with

time window. Abidi et al. [7] proposed a GA with a simple heuristic to solve a variant of RVPR (Rich VPR) with time widows and dynamically changing orders. Dongdong and Yinzhen [8] proposed a green multitype VRP with time windows to reduce the wastes of fuel consumption and carbon emission and use an improved tabu search algorithm to solve the G-MVRPTW. Fan et al. [9] pointed out that in fresh agricultural product cold-chain logistics, the total costs are composed of five kinds: fixed, transportation, damage, penalty, and energy consumption. Fang and Ai [10] proposed a mathematical model with soft time window penalty cost, refrigeration cost, cargo damage cost, and a hybrid ant colony algorithm to minimize the total costs. The DVRP mainly deal with the time-variation information of customer demands and road conditions. To share the traffic information and classify traffic conditions, Big Data and classification techniques are used in logistics and transportation. If the drivers receive the information of the traffic congestion or poor weather, they can change their way to reduce the costs and save time. Dimensionality reduction must be performed first in Big Data transmitted and stored process. Thippa et al. [11] proposed a machine learning (ML) algorithm with PCA to reduce the dimension of Big Data when the data sets are high. Gadekallu et al. [12] have studied the hybrid PCA-whale optimization algorithm to extract features and used a deep neural network to classify the diseases of tomato. Guha et al. [13] proposed an ANN-based content classification in combination with $n$ -grame TF-IDF feature descriptor to classify the documents with accurate, sensitive information. The ant colony optimization (ACO) proposed by Colorni et al. [14] has been widely used in solving NP-hard vehicle routing problems. Many scholars used ACO algorithms to solve multiobjective combinatorial optimization problems [15–19]. These papers considered the costs of fuel, refrigeration, cargo damage, and delay simultaneously. However, their studies are not comprehensive enough in factors affecting the loss, such as the cost of energy consumption and rotting consumption when the door of compartment is opening.

We construct a total cost model, including soft time window penalty cost, fuel cost, refrigeration cost in transit and during unloading, and cargo damage cost in transit and during unloading. In our model, to take into account the actual traffic condition, the distance matrix and the transit time matrix are both archived from the navigation functions based on E-Map API. The traditional heuristic algorithm ACO has some limitations, such as easy stagnation in the initial stage and slow search speed. We perform an ACO algorithm with Pareto local search (PLS) on ants to obtain the uniform Pareto-optimal frontier and keep the diversity of Pareto solution set.

The paper is organized as follows. Section 1 introduces the definition and structure of VRP for cold-chain logistics. Section 2 constructs a minimum cost model based on actual traffic conditions. Then, an improved ACO algorithm with PLS is presented in Section 3. In Section 4, a compare experiment is performed to indicate our proposed model and the improved method. At last, we make a conclusion of our contributions in Section 5.

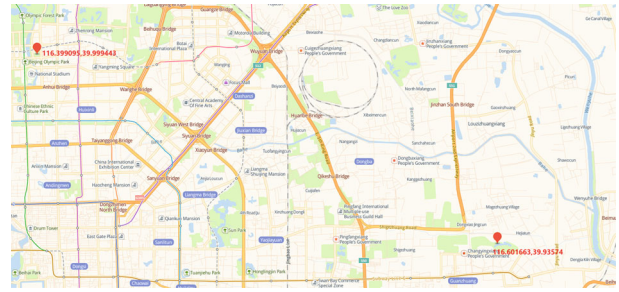


Figure 1: Two points on the E-map.

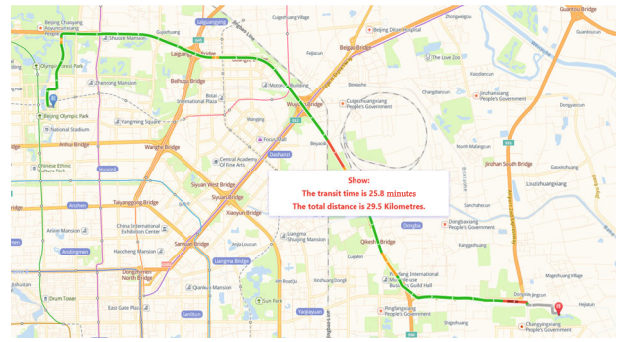



Figure 2: The output of navigation result.

## 2. Problem Description

A cold-chain supplier has a warehouse with a certain number of transportation vehicles that delivers a variety of fresh products to a certain number of customers. The capacity of the transportation vehicles is limited, and the vehicles are all the same type with a total capacity at the starting point. The object of vehicle routing problem is to distribute goods to each customer with correct goods, limited time, minimum cost, and so on. VRP with time window means that a vehicle has to visit a customer within a certain time window. So, the service time, the transit time, and the total time must be calculated when dealing with the VRP. The mathematical model of VRPTW in cold-chain distribution is defined as follows [20].

There are $K$ transportation trucks in this distribution center. The maximum load and the maximum transit time of per vehicle are $Q$ and T. The $k$th distribution vehicle is responsible for path $k$. $N$ is the total customers. The needs and the service time of customer point $i$ are $q_i$ and $u_i$; $t_{ij}$ means the delivery time from customer points $i$ to $j$. $x_i^k$ is a 0-1 variable, and $x_i^k = 1$ means the point $i$ in the path $k$. $x_{ij}^k$ is a 0-1 variable, and $x_{ij}^k = 1$ means that the $k$th distribution vehicle travels from $i$ to $j$.

## 3. Model

Minimizing the loss cost is the target of this paper. Therefore, the following paper introduces the structure of the loss cost of cold chain.

Table 1: Parameter setting.

| Parameter of the symbol | Meaning | Value | Parameter of the symbol | Meaning | Value |
|---|---|---|---|---|---|
| $K$ | The total vehicles | 4 | $p$ | Unit price of goods | 5 |
| $N$ | The total customer points | 20 | $m$ | Number of ants | 50 |
| $Q$ | Full load of one vehicle | 2000 | $n$ | Number of iterations | 60 |
| $C_{11}$ | Penalty coefficients earlier than TW | 2.0 | Alpha | The pheromone important factor | 1 |
| $C_{12}$ | Penalty coefficients later than TW | 3.0 | Beta | Heuristic function important factor | 3 |
| $C_{21}$ | Unit fuel cost of no load | 2.5 | Rho | The pheromone factor | 0.85 |
| $C_{22}$ | Unit fuel cost of full load | 3.0 | $k_1$ | Positive constant | 0.75 |
| $C_{31}$ | Unit refrigeration cost in transit | 2.0 | | | |
| $C_{32}$ | Unit refrigeration cost in unloading | 2.5 | | | |
| $C_{41}$ | Coefficient of damage in transit | 0.01 | | | |
| $C_{42}$ | Coefficient of damage in unloading | 0.015 | | | |

Table 2: The data of the distribution center and customers.

| Number of customers | Latitude and longitude on E-map | | The demand of customer (kg) | Starting time (min) | Ending time (min) | Server time (min) |
|---|---|---|---|---|---|---|
| | Latitude (easting) | Longitude (northing) | | | | |
| 0 | 116.4843 | 39.8768 | 0 | 0 | 0 | 0 |
| 1 | 116.486081 | 39.801535 | 300 | 912 | 967 | 15 |
| 2 | 116.43287 | 39.851657 | 1100 | 825 | 870 | 30 |
| 3 | 116.571474 | 39.857011 | 125 | 65 | 146 | 10 |
| 4 | 116.51774 | 39.86957 | 100 | 727 | 782 | 10 |
| 5 | 116.258165 | 39.896287 | 200 | 15 | 67 | 10 |
| 6 | 116.598001 | 39.920533 | 150 | 621 | 702 | 10 |
| 7 | 116.475892 | 39.92811 | 150 | 170 | 225 | 10 |
| 8 | 116.659393 | 39.9282 | 450 | 255 | 324 | 20 |
| 9 | 116.110548 | 39.943414 | 300 | 534 | 605 | 20 |
| 10 | 116.50121 | 39.967727 | 100 | 357 | 410 | 10 |
| 11 | 116.450433 | 39.971126 | 950 | 448 | 505 | 30 |
| 12 | 116.408206 | 39.973734 | 125 | 652 | 721 | 10 |
| 13 | 116.339988 | 39.978354 | 150 | 30 | 90 | 10 |
| 14 | 116.468832 | 40.006183 | 150 | 567 | 620 | 10 |
| 15 | 116.584485 | 40.007985 | 550 | 384 | 429 | 20 |
| 16 | 116.601421 | 40.054617 | 150 | 475 | 528 | 10 |
| 17 | 116.439297 | 40.057803 | 100 | 99 | 148 | 10 |
| 18 | 116.670363 | 40.140651 | 150 | 179 | 254 | 10 |
| 19 | 116.226626 | 40.228101 | 400 | 278 | 345 | 20 |
| 20 | 116.653284 | 40.332406 | 300 | 10 | 73 | 20 |

*3.1. Distance and Transit Time Analysis.* In traditional methods, the distance between two points is achieved by Euclidean distance formula and the transit time is achieved by the distance divided by assumed average speed. But, the actual transit time is influenced greatly by road congestion, traffic accidents, traffic control, raining, snowing, etc. Collecting real-time traffic data will result in large amounts of data, and Big Data analysis technology should be considered.

In our literature, the navigation function of E-map is used to avoid the mentioned shortcomings. The distance matrix and the transit time matrix under actual traffic conditions among customer points (including the distribution center) are derived from the E-map public platform [21].

As Figure 1 shows, there are two points with their latitudes and longitudes on the E-map.

```
Initialize the coefficient variables and pheromone;
for iteration 1,…,M{
Initialize a taboo table and a start position for each ant.
for ant, 1,…,N{
Select the next node according to rules
The selected node is stored in the taboo table
If all nodes are stored in taboo tables, the iteration is completed, break;
The total cost is calculated
The local pheromone is updated
}
The global pheromone is updated
All ants' total costs are compared
The current optimal solution is stored
}
The process is stopped and the current path with the shortest cost is output.
```

Algorithm 1

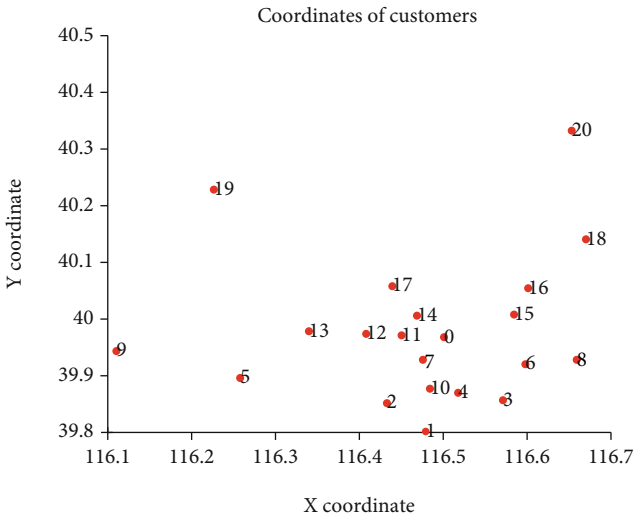



Figure 3: Distribution of warehouse and customers in the $XY$ system.

As Figure 2 shows, the distance and the transit time between the two points are shown by the navigation function of the E-map platform.

In our study, $t_{ij}$ represents transit time between customer $i$ and customer $j$, and $d_{ij}$ is the distance between customer $i$ and customer $j$. As abovementioned, $t_{ij}$ and $d_{ij}$ are both achieved from the navigation result based on E-map API.

### 3.2. Loss Cost Target Construction

*3.2.1. Soft Time Window Penalty Cost.* Each customer has different numbers of good needs and a soft time window. Penalty cost will be incurred if the vehicle arrives beyond the time window boundary. The service time of each customer point is $u_i(i = 1, 2, \cdots, N)$, $[Et_i, Lt_i]$ is the range of the time window of customer point $i$, $C_{11}$ is the penalty coefficient when vehicle $k_i$ arrives at customer $i$ earlier than $Et_i$, and $C_{12}$ is the penalty coefficient when vehicle $k_i$ leaves from customer $i$ later than $Lt_i$. The arrival time of vehicle $k$ at customer $i$ is $T_{ki}$. According to the analysis above, the penalty cost of cold chain can be defined as the following:

$$\begin{aligned} C_1 = {} & C_{11} \sum_{k=1}^{K} \left[ \sum_{i=1}^{N} x_i^k \max\left(Et_i - T_{ki}, 0\right) \right] \\ & + C_{12} \sum_{k=1}^{K} \left[ \sum_{i=1}^{N} x_i^k \max\left(T_{ki} + u_i - Lt_i, 0\right) \right], \\ & \text{s.t.} \quad T_{ki} \geq T_{k0} + t_{0i}, \\ & T_{k(i+1)} \geq T_{ki} + t_{i(i+1)} + u_i. \end{aligned} \tag{1}$$

*3.2.2. Fuel Cost.* The fuel consumption of the distribution vehicle is inevitable for completing the distribution task. We assume that the fuel cost is proportional to the load of the vehicle pert unit distance. Let $Q_k$ be the load of the $k$ vehicle and $P_k$ be the load rate of the $k$th vehicle.

$$\begin{aligned} Q_k &= \sum_{i=1}^{n} x_i^k q_i, \\ P_k &= \frac{Q_k}{Q}. \end{aligned} \tag{2}$$

Sets $C_{21}$ and $C_{22}$ represent no load and full load fuel consumption costs per unit distance of the distribution vehicle. $C_2^k$ means the fuel consumption cost per unit distance of the $k$th vehicle.

$$C_2^k = C_{21} + p_k(C_{22} - C_{21}). \tag{3}$$

The total cost of fuel consumption $C_2$ can be expressed as

$$C_2 = \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j=1}^{N} x_{ij}^k d_{ij}(C_{21} + P_k(C_{22} - C_{21})). \tag{4}$$

```
.....
var output_t = "The transit time:";
var output_d = "The traveling distance:";
var distance_time_search = function (customerPoints){.
if (vehicle_transit.getStatus() != BMAP_STATUS_SUCCESS){.
return ;
}
var t_d_result = customerPoints.getPlan(0);
travel_time += t_d_result.getDuration(true) + "\n"; //get transit time of two points.
trave_distance+= t_d_result.getDistance(true)+"\n"; //get travel distance of two points.
}
var vehicle_transit = new BMapGL.DrivingRoute(map, {renderOptions: {map: map},
onSearchComplete: distance_time_search;
}});
for(var i=0;i<=NumberOfCustomers;i++)
for( var j=i+1;j<= NumberOfCustomers;j++)
{
var start=new BMapGL.Point(vertex[i][0], vertex[i][1]);
var end=new BMapGL.Point(vertex[j][0], vertex[j][1]);
vehicle_transit.search(start, end);
}
.....
```

Algorithm 2

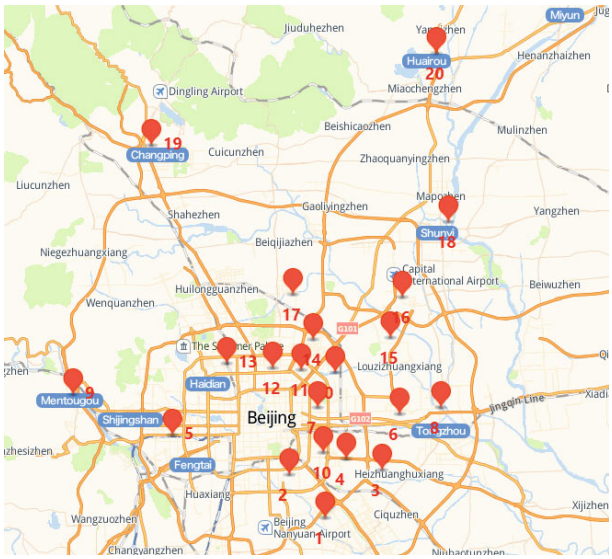



Figure 4: Geographic locations of warehouse and customers in E-map.

*3.2.3. Refrigeration Cost.* The temperature of the cold-chain logistic must be maintained at a certain low level to keep the freshness of goods. The refrigeration function of vehicles must be operated to achieve the required temperature, which causes the refrigeration cost immediately.

(1) Refrigeration cost in transit

The refrigeration cost generated by maintaining the required low temperature per unit time in transit is $C_{31}$; the total cooling cost $C_3^1$ in transit can be expressed as

$$C_3^1 = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}^k t_{ij} C_{31}. \tag{5}$$

(2) Refrigeration cost during unloading

Opening the door of the compartment will cause the cool air inside to flow out and the hot air from outside to flow in. To maintain the temperature inside, more energy will be consumed. $C_{32}$ is set to be the cooling cost per unit time during opening the door, and the total refrigeration cost of unloading $C_3^2$ is

$$C_3^2 = \sum_{k=1}^{K}\sum_{i=1}^{N} x_i^k u_i C_{32}. \tag{6}$$

$C_3$ represents the total cooling cost, which consists of the total cooling cost in transit, and the total cooling cost of unloading can be expressed as

$$C_3 = \sum_{k=1}^{K}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} x_{ij}^k t_{ij} C_{31} + x_i^k u_i C_{32}\right). \tag{7}$$

*3.2.4. Cargo Damage Cost.* The most cargo of the cold chain is fresh goods; with the increase of transit time even in low temperature, the growth of microorganisms will happen. When opening the door of compartment during unloading process, the temperature inside will be unstable which results in the damage of fresh goods more significantly. As mentioned above, $p$ is the unit price of goods, and $Q_{ki}$ is the load of the $k$th vehicle when it leaves from point $i$.

Table 3: Transit time matrix under actual traffic conditions (minutes).

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 22 | 25 | 25 | 26 | 35 | 13 | 16 | 24 | 43 | 25 | 13 | 18 | 23 | 25 | 21 | 24 | 15 | 43 | 46 | 40 |
| 1 | 22 | 0 | 17 | 19 | 28 | 27 | 21 | 26 | 38 | 17 | 28 | 34 | 51 | 34 | 37 | 32 | 43 | 40 | 59 | 58 | 55 |
| 2 | 25 | 17 | 0 | 25 | 20 | 26 | 32 | 30 | 18 | 43 | 33 | 28 | 35 | 34 | 38 | 37 | 45 | 63 | 42 | 61 | 61 |
| 3 | 25 | 19 | 25 | 0 | 16 | 24 | 24 | 14 | 26 | 36 | 34 | 28 | 43 | 57 | 34 | 27 | 51 | 55 | 53 | 41 | 32 |
| 4 | 26 | 28 | 20 | 16 | 0 | 9 | 23 | 24 | 21 | 22 | 28 | 38 | 52 | 35 | 29 | 39 | 36 | 29 | 57 | 51 | 53 |
| 5 | 35 | 27 | 26 | 24 | 9 | 0 | 26 | 28 | 35 | 27 | 50 | 42 | 33 | 51 | 36 | 50 | 65 | 49 | 54 | 49 | 71 |
| 6 | 13 | 21 | 32 | 24 | 23 | 26 | 0 | 13 | 27 | 20 | 26 | 25 | 33 | 31 | 38 | 41 | 19 | 59 | 53 | 35 | 50 |
| 7 | 16 | 26 | 30 | 14 | 24 | 28 | 13 | 0 | 9 | 23 | 17 | 14 | 24 | 25 | 20 | 27 | 28 | 48 | 47 | 42 | 49 |
| 8 | 24 | 38 | 18 | 26 | 21 | 35 | 27 | 9 | 0 | 24 | 27 | 32 | 42 | 35 | 24 | 63 | 32 | 37 | 41 | 58 | 57 |
| 9 | 43 | 17 | 43 | 36 | 22 | 27 | 20 | 23 | 24 | 0 | 39 | 40 | 46 | 46 | 51 | 60 | 63 | 42 | 48 | 69 | 78 |
| 10 | 25 | 28 | 33 | 34 | 28 | 50 | 26 | 17 | 27 | 39 | 0 | 18 | 29 | 23 | 25 | 34 | 26 | 33 | 51 | 48 | 49 |
| 11 | 13 | 34 | 28 | 28 | 38 | 42 | 25 | 14 | 32 | 40 | 18 | 0 | 8 | 16 | 10 | 23 | 40 | 20 | 26 | 34 | 38 |
| 12 | 18 | 51 | 35 | 43 | 52 | 33 | 33 | 24 | 42 | 46 | 29 | 8 | 0 | 16 | 13 | 22 | 34 | 26 | 43 | 24 | 40 |
| 13 | 23 | 34 | 34 | 57 | 35 | 51 | 31 | 25 | 35 | 46 | 23 | 16 | 16 | 0 | 22 | 34 | 36 | 36 | 32 | 52 | 50 |
| 14 | 25 | 37 | 38 | 34 | 29 | 36 | 38 | 20 | 24 | 51 | 25 | 10 | 13 | 22 | 0 | 18 | 24 | 35 | 39 | 26 | 37 |
| 15 | 21 | 32 | 37 | 27 | 39 | 50 | 41 | 27 | 63 | 60 | 34 | 23 | 22 | 34 | 18 | 0 | 19 | 32 | 49 | 36 | 49 |
| 16 | 24 | 43 | 45 | 51 | 36 | 65 | 19 | 28 | 32 | 63 | 26 | 40 | 34 | 36 | 24 | 19 | 0 | 25 | 31 | 45 | 46 |
| 17 | 15 | 40 | 63 | 55 | 29 | 49 | 59 | 48 | 37 | 42 | 33 | 20 | 26 | 36 | 35 | 32 | 25 | 0 | 43 | 38 | 41 |
| 18 | 43 | 59 | 42 | 53 | 57 | 54 | 53 | 47 | 41 | 48 | 51 | 26 | 43 | 32 | 39 | 49 | 31 | 43 | 0 | 33 | 45 |
| 19 | 46 | 58 | 61 | 41 | 51 | 49 | 35 | 42 | 58 | 69 | 48 | 34 | 24 | 52 | 26 | 36 | 45 | 38 | 33 | 0 | 42 |
| 20 | 40 | 55 | 61 | 32 | 53 | 71 | 50 | 49 | 57 | 78 | 49 | 38 | 40 | 50 | 37 | 49 | 46 | 41 | 45 | 42 | 0 |

Table 4: The distance matrix derived from the E-map navigation function (KM).

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 22 | 13.6 | 20.7 | 17.6 | 6.5 | 19.2 | 5.6 | 11.3 | 21.2 | 31.6 | 16.8 | 11.4 | 7.2 | 17.4 | 41.6 | 22 | 14.2 | 53.4 | 41.6 | 35.1 |
| 1 | 22 | 0 | 9.1 | 12.7 | 11.3 | 21 | 19.5 | 26.6 | 13.1 | 29.5 | 45.4 | 23.5 | 27.8 | 32.7 | 37.9 | 27.3 | 39.5 | 35.4 | 62.2 | 86.4 | 64.7 |
| 2 | 13.6 | 9.1 | 0 | 12.3 | 19.1 | 11.7 | 22.4 | 22.9 | 6.5 | 29.3 | 16.2 | 40 | 36.3 | 27 | 27 | 15.4 | 37.2 | 55.6 | 67.3 | 34 | 69.1 |
| 3 | 20.7 | 12.7 | 12.3 | 0 | 9.8 | 8.8 | 15.2 | 18.2 | 10.3 | 24.6 | 39.6 | 27.4 | 21.5 | 24.6 | 29.8 | 31.5 | 55.6 | 34.1 | 52.5 | 74.2 | 59.1 |
| 4 | 17.6 | 11.3 | 19.1 | 9.8 | 0 | 3.8 | 16.9 | 10.5 | 17.9 | 19 | 28.9 | 18.7 | 33.6 | 26.1 | 49.6 | 21.9 | 25.7 | 28.3 | 53.1 | 79.8 | 55.3 |
| 5 | 6.5 | 21 | 11.7 | 8.8 | 3.8 | 0 | 22.6 | 24.1 | 26.8 | 32.2 | 17.8 | 37.5 | 33.9 | 19 | 29.8 | 45 | 57.8 | 33 | 74.8 | 44.6 | 60.5 |
| 6 | 19.2 | 19.5 | 22.4 | 15.2 | 16.9 | 22.6 | 0 | 6 | 17.1 | 11.5 | 21.9 | 32.3 | 24.4 | 20.6 | 12.3 | 17.6 | 64.4 | 55.6 | 30.4 | 31.3 | 64.1 |
| 7 | 5.6 | 26.6 | 22.9 | 18.2 | 10.5 | 24.1 | 6 | 0 | 18.3 | 8.2 | 11.9 | 16.9 | 6 | 13.5 | 21 | 21.4 | 57.2 | 40.9 | 18.5 | 44 | 38 |
| 8 | 11.3 | 13.1 | 6.5 | 10.3 | 17.9 | 26.8 | 17.1 | 18.3 | 0 | 21.4 | 26.2 | 28.7 | 14.9 | 36.6 | 26.5 | 18.7 | 25.4 | 59.8 | 34.4 | 52.4 | 65.2 |
| 9 | 21.2 | 29.5 | 29.3 | 24.6 | 19 | 32.2 | 11.5 | 8.2 | 21.4 | 0 | 29.4 | 32.6 | 55.4 | 42.1 | 36.9 | 40.2 | 85.5 | 44.6 | 41.9 | 86.9 | 75 |
| 10 | 31.6 | 45.4 | 16.2 | 39.6 | 28.9 | 17.8 | 21.9 | 11.9 | 26.2 | 29.4 | 0 | 17.7 | 12.5 | 27.8 | 26.9 | 17 | 24.6 | 81.2 | 27.2 | 52 | 59.4 |
| 11 | 16.8 | 23.5 | 40 | 27.4 | 18.7 | 37.5 | 32.3 | 16.9 | 28.7 | 32.6 | 17.7 | 0 | 5.8 | 6.1 | 11.6 | 24.2 | 13.2 | 26.7 | 36.8 | 51.1 | 38 |
| 12 | 11.4 | 27.8 | 36.3 | 21.5 | 33.6 | 33.9 | 24.4 | 6 | 14.9 | 55.4 | 12.5 | 5.8 | 0 | 11.7 | 22.1 | 9.3 | 7.7 | 26.6 | 38.4 | 35.7 | 52.7 |
| 13 | 7.2 | 32.7 | 27 | 24.6 | 26.1 | 19 | 20.6 | 13.5 | 36.6 | 42.1 | 27.8 | 6.1 | 11.7 | 0 | 14 | 28.4 | 37.1 | 59 | 44.7 | 21 | 32.9 |
| 14 | 17.4 | 37.9 | 27 | 29.8 | 49.6 | 29.8 | 12.3 | 21 | 26.5 | 36.9 | 26.9 | 11.6 | 22.1 | 14 | 0 | 8.7 | 21.2 | 16.6 | 37.6 | 48 | 33.7 |
| 15 | 41.6 | 27.3 | 15.4 | 31.5 | 21.9 | 45 | 17.6 | 21.4 | 18.7 | 40.2 | 17 | 24.2 | 9.3 | 28.4 | 8.7 | 0 | 8.5 | 16.4 | 23.3 | 57.3 | 50.5 |
| 16 | 22 | 39.5 | 37.2 | 55.6 | 25.7 | 57.8 | 64.4 | 57.2 | 25.4 | 85.5 | 24.6 | 13.2 | 7.7 | 37.1 | 21.2 | 8.5 | 0 | 18.8 | 51.9 | 17.6 | 50.3 |
| 17 | 14.2 | 35.4 | 55.6 | 34.1 | 28.3 | 33 | 55.6 | 40.9 | 59.8 | 44.6 | 81.2 | 26.7 | 26.6 | 59 | 16.6 | 16.4 | 18.8 | 0 | 36 | 44.1 | 29.7 |
| 18 | 53.4 | 62.2 | 67.3 | 52.5 | 53.1 | 74.8 | 30.4 | 18.5 | 34.4 | 41.9 | 27.2 | 36.8 | 38.4 | 44.7 | 37.6 | 23.3 | 51.9 | 36 | 0 | 25 | 47.5 |
| 19 | 41.6 | 86.4 | 34 | 74.2 | 79.8 | 44.6 | 31.3 | 44 | 52.4 | 86.9 | 52 | 51.1 | 35.7 | 21 | 48 | 57.3 | 17.6 | 44.1 | 25 | 0 | 56.3 |
| 20 | 35.1 | 64.7 | 69.1 | 59.1 | 55.3 | 60.5 | 64.1 | 38 | 65.2 | 75 | 59.4 | 38 | 52.7 | 32.9 | 33.7 | 50.5 | 50.3 | 29.7 | 47.5 | 56.3 | 0 |

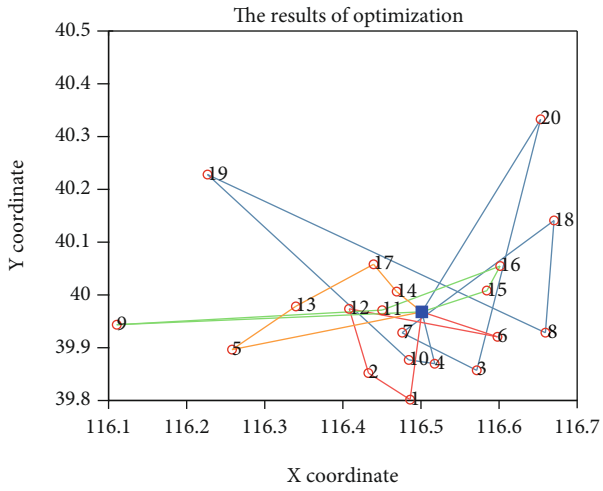



Figure 5: The optimal path plan based on the traditional method.

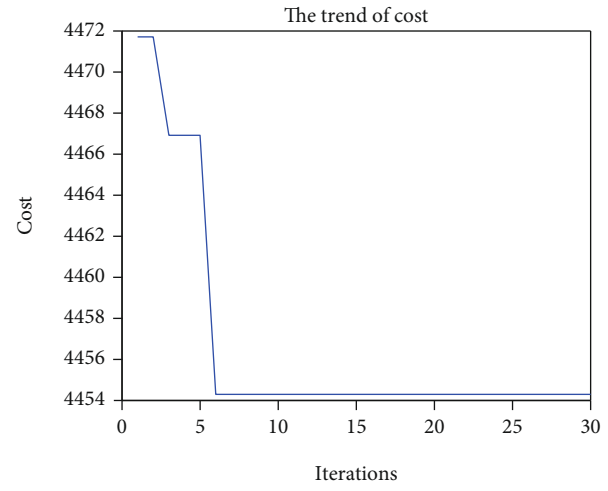



Figure 6: The trend of cost during the iteration.

(1) Cost of cargo damage in transit

$C_{41}$ is given as the loss coefficient of goods per unit weight per unit time in transit; the total cost of cargo damage in transit is as follows:

$$C_4^1 = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}^k t_{ij} p Q_{ki} C_{41}. \tag{8}$$

(2) Cost of cargo damage during unloading

During unloading at a customer point, the door of the compartment is open and the cost of cargo damage will increase obviously. Let the cargo damage coefficient of goods per unit weight per unit time when unloading be $C_{42}$; the cost of cargo damage during unloading can be expressed as

$$C_4^2 = \sum_{k=1}^{K}\sum_{i=1}^{N} x_i^k u_i p Q_{ki} C_{42}. \tag{9}$$

The total cost of damage is

$$C_4 = pQ_{ki}\sum_{k=1}^{K}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} x_{ij}^k t_{ij} C_{41} + x_i^k u_i C_{42}\right). \tag{10}$$

In conclusion, the mathematical model of minimum loss target can be expressed as

$$\min Z = C_1 + C_2 + C_3 + C_4. \tag{11}$$

*3.3. Mathematical Model of Minimum Loss Analysis.* According to the above analysis, the minimum loss model of the cold-chain distribution problem is as follows:

$$\begin{aligned}\min\ Z = {} & C_{11}\sum_{k=1}^{K}\left[\sum_{i=1}^{N} x_i^k \max\left(Et_i - T_{ki}, 0\right)\right] \\ & + C_{12}\sum_{k=1}^{K}\left[\sum_{i=1}^{N} x_i^k \max\left(T_{ki} + u_i - Lt_i, 0\right)\right] \\ & + \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}^k\left(C_{21} + P_k\left(C_{22} - C_{21}\right)\right) \\ & + \sum_{k=1}^{K}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} x_{ij}^k t_{ij} C_{31} + x_i^k u_i C_{32}\right) \\ & \cdot pQ_{ki}\sum_{k=1}^{K}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} x_{ij}^k t_{ij} C_{41} + x_i^k u_i C_{42}\right)\end{aligned} \tag{12}$$

$$\text{s.t.}\quad x_i^k = \begin{cases} 1, & \text{point } i \text{ is serviced by car } K, \\ 0, & \text{other}, \end{cases} \tag{13}$$

$$x_{ij}^k = \begin{cases} 1, & \text{vehicle } K \text{ travels from point } i \text{ to point } j, \\ 0, & \text{other}, \end{cases} \tag{14}$$

$$\sum_{k=1}^{K} x_i^k = 1, \quad \forall i \in N, \tag{15}$$

$$\sum_{i=1}^{N} x_i^k q_i \le Q, \quad \forall k \in K, \tag{16}$$

$$\sum_{j=1}^{N} x_{0j}^k = \sum_{i=1}^{N} x_{i0}^k \le 1, \quad \forall k \in K. \tag{17}$$

Equation (12) is the objective optimization function which is aimed at finding the minimum total cost during the whole distribution process. Equation (13)–Equation (17) are the constraint conditions, where Equation (13) is a 0-1 variable; the value 1 means the given customer is

Table 5: Distribution paths of 4 vehicles based on the traditional method.

| Vehicles | The route of each vehicle | The realistic delivery on E-map (minutes) |
|---|---|---|
| Fist vehicle | 0 – >20 – >3–>7–>18–>8–>19–>10–>4–>0 | 334 |
| Second vehicle | 0 – >6 – >12–>2–>1–>0 | 120 |
| Third vehicle | 0 – >15 – >16–>11–>9–>0 | 163 |
| Fourth vehicle | 0 – >5 – >13–>17–>14–>0 | 182 |
| The total time | | 799 |
| The total cost | | 4504.5 |

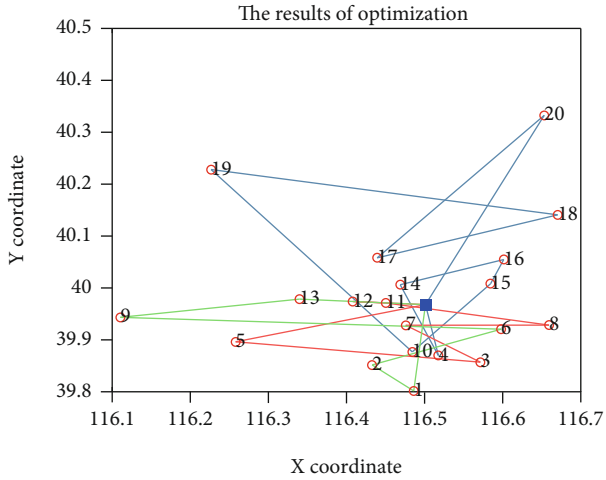



Figure 7: The optimal path plan based on the proposed method.

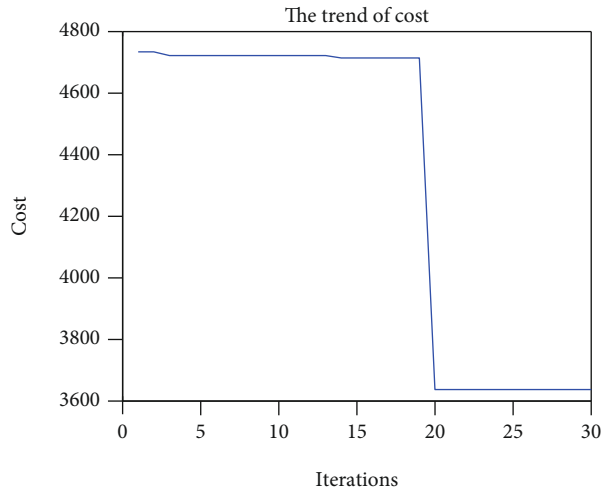



Figure 8: The trend of cost during the iteration.

serviced by the given vehicle. Equation (14) is a 0-1 variable too; the value 1 means the given route is travelled by the given vehicle. Equation (15) means each customer point is serviced by one and only one vehicle for distribution. Equation (16) means that the total demand served by the vehicle cannot exceed its maximum load. Equation (17) means that the vehicle can only start or end once.

## 4. Ant Colony Optimization Algorithm

*4.1. Ant Colony Optimization Algorithm.* The ant colony optimization (ACO) algorithm with parallelism, positive feedback, and strong robustness is used to solve the NP-hard and highly constrained problem. In the 1990s, Italian scholars Dorigo and Maniezzo found that ants will release a certain amount of substance called pheromone in the path when they are searching for food. When they need to choose a path, they prefer to choose the path with high concentration of pheromone. At last, the route with the highest concentration will be selected as an optimist route between their home and the food source.

The ACO algorithm is defined as follows:

$$p_{ij}^{k}\begin{cases}\dfrac{\tau_{ij}^{\alpha}(t)\cdot\eta_{ij}^{\beta}(t)}{\sum_{s\in \text{allowed}_k}\tau_{is}^{\alpha}(t)\cdot\eta_{is}^{\beta}(t)}, & \text{if } j\in \text{allowed}_k,\\ 0, & \text{otherwise,}\end{cases}$$
$$\tau_{ij}(t+n)=\rho\cdot\tau_{ij}(t)+\Delta\tau_{ij},$$
$$\Delta\tau_{ij}=\sum_{k=1}^{m}\Delta\tau_{ij}, \tag{18}$$

where $p_{ij}^{k}$ is probability of ant $k$ transferring from point $i$ to point $j$ at time $t$; $\eta_{ij}(t)$ is the heuristic function, and $\beta$ is the heuristic function important factor; $\tau_{ij}(t)$ means the pheromone concentration on the path at time $t$, and $\alpha$ is the pheromone important factor. $allowed_k$ represents points that ant $k$ is allowed to select in the next step; $\tau_{ij}(t+n)$ is the pheromone update over time; $\rho$ is the pheromone factor; and $\Delta\tau_{ij}^{k}$ represents the pheromone enhancer.

The rough process of the algorithm is as follows:

*4.2. Combining ACO Algorithm with PLS.* We applied Pareto local search (PLS) in our paper. After finding a path, PLS is used to further optimize the path. PLS proposed by Paquete et al. is a heuristic algorithm for tackling NP-hard multiobjective combinatorial optimization problems in the Pareto sense [22]. Pareto dominance defines a partial order on the set of feasible solutions. The goal when tackling the multiobjective problems in the Pareto sense is to find the set of Pareto-optimal solutions. The weak component-wise ordering is used as a mutually nondominated criterion in PLS. Let a solution $s\in A$ and its neighborhood $s'\in N(s)$. If each $s'$ in $N(s)$ is

Table 6: Distribution paths of 3 vehicles based on the proposed method.

| Vehicles | The route of each vehicle | The realistic delivery on E-map (minutes) |
|---|---|---|
| Fist vehicle | 0 – >20 – >17–>18–>19–>10–>15–>16–>14–>4–>0 | 331 |
| Second vehicle | 0 – >5 – >3–>7–>8–>11–>12–>0 | 113 |
| Third vehicle | 0 – >13 – >9–>6–>2–>1–>0 | 160 |
| The total time | | 593 |
| The total cost | | 3673 |

not dominated by any solution in $A$, $s$ is marked visited and is added into $A$. When $A$ contains only solutions that have been visited, a Pareto local optimum is achieved [23].

An improved ACO with PLS is proposed in our research to deal with the mode. In our approach, ants are used to optimize solutions and generate new Pareto solutions, and the new pheromone updating strategy is used to control the path selection. The procedure is described as follows:

*Step 1.* Compute $p_{ij}^k$ of $n$ customers and update the trail level $\tau_{ij}$_new for each $k$.

*Step 2.* Run PLS, and update the current solution according to calculating results.

*Step 3.* Compare (Cost, CostBest), and record the costBest and BestSolution.

*Step 4.* For each move $(i, j)$ in BestSolution to update the trail level $\tau_{ij}$_new.

*Step 5.* Repeat step 1 until the maximum computation time reaches.

The proportion of the pheromone update model of the ACO algorithm with PLS is as follows:

$$\omega = 1 - k_1 \frac{S_{\max} - S(t)}{S_{\max}}, \tag{19}$$

$$S(t) = -K \sum_{j \in D} p_{ij}(t) \log p_{ij}(t), \tag{20}$$

where $S(t)$ means information entropy value and $p_{ij}(t)$ is transition probability. Equation (19) means that the proportion of the pheromone update will decreases with the increase of the pheromone value. $k_1$ is the positive constant, and its value range is [0,1]. The value 1 means the proportion of the pheromone update is high influenced by the information entropy value, and 0.75 is usually used as the value of $k_1$. This definition combined a sequence of arithmetic with information entropy to adjust the pheromone adaptively.

## 5. Experimental Results and Discussion

*5.1. Data Set Description.* In our experiments, the distribution situation of JingKeLong in Beijing City with one distribution center and 20 customer points is selected to verify our proposed mathematical model and algorithm. The distribution center has 4 vehicles of the same type with an average speed of 60 km/h, a maximum driving time of 1200 minutes, and a maximum load of 2000 kg. Table 1 shows the settings of parameters in our given minimum loss model and ACO+PLS algorithm.

The data of the distribution center and customers are shown in Table 2. The distribution center is defined as 0, and customer points are defined as 1 to 20. Figure 3 shows the latitude and longitude of each point in the $XY$ system, and Figure 4 shows the actual locations of them in E-map.

Our proposed method needs an actual transit time matrix and distance matrix both derived from E-map API. We supposed that the coordinates of points are stored into vertex [$N$] [2]; the core codes with E-map API to achieve the transit time matrix and distance matrix are described as follows.

Table 3 shows the transit time matrix among customer points and distribution center obtained by E-map API.

Table 4 shows the distance matrix among customer points and distribution center obtained by E-map API.

*5.2. Experimental Result and Discussion.* In the traditional method, the distance matrix is achieved by the Euclidean distance formula. The Euclidean distance formula of two points on earth is defined as the following:

$$D_{AB} = R * \arccos\left(\cos\left(\text{lon}A\right) * \cos\left(\text{lon}B\right) * \cos\left(\text{lat}A - \text{lat}B\right) + \sin\left(\text{lon}A\right) * \sin\left(\text{lon}B\right)\right) * \frac{\pi}{180}, \tag{21}$$

where $A$ and $B$ are two points on the surface of earth and $R$ is the radius of the earth. Lat$A$ and lon$A$ represent the longitude and latitude of point $A$.

First, the traditional method ACO with the Euclidean distance matrix is performed to solve the VRP with minimum cost. Figures 5 and 6 are the optimal distribution routes and the trend of cost during iteration processed. The optimal distribution routing scheme, the total transit time, and the total cost based on the traditional method are shown in Table 5.

Second, our proposed method ACO+PLS with the actual transit time matrix and travel distance matrix is performed to solve the VRP with minimum cost. Figures 7 and 8 are the calculated distribution routes on map and the trend of cost during iteration processed. The optimal distribution routing scheme, the total transit time, and the total cost based on the proposed method are shown in Table 6.

From the above compare experimental results, we can see that ACO+PLS with actual traffic conditions can get smaller total cost and save more transit time. Our proposed method can be used to get the optimal solution effectively for VRP of cold-chain logistics.

## 6. Conclusion

This paper studies the model of VPR with minimum cost of cold-chain distribution. Minimizing the total cost of distribution can maximize the economic benefit of distribution enterprises. To consider the dynamic change of transit speed under actual traffic conditions, the transit time matrix and distance matrix are both derived from the navigation function based on E-map API. We proposed a heuristic approach ACO+PLS to solve the minimum loss model of cold-chain logistics. The experiments show that our proposed method has strong applicability and potential advantages in cold-chain distribution. In future studies, a combination of multiple spatial information technologies such as geographic information technology and remote sensing technology can be realized to make the problem more practical.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no competing interest.

## Acknowledgments

This work was supported by Scientific Research Project of Beijing Municipal Education Commission (KM201911417006).

## References

[1] G. B. Dantzig and J. H. Ramser, "The truck dispatching problem," *Management Science*, vol. 6, no. 1, pp. 80–91, 1959.

[2] Z. Rao, "Common distribution path of cold chain logistics of fresh agricultural products," *Agronomia*, vol. 36, no. 5, 2019.

[3] L. Li, Y. Yang, and G. Qin, "Optimization of integrated inventory routing problem for cold chain logistics considering carbon footprint and carbon regulations," *Sustainability*, vol. 11, no. 17, p. 4628, 2019.

[4] E. Yao, Z. Lang, Y. Yang, and Y. Zhang, "Vehicle routing problem solution considering minimising fuel consumption," *IET Intelligent Transport Systems*, vol. 9, no. 5, pp. 523–529, 2015.

[5] G. Kim, Y. S. Ong, T. Cheong, and P. S. Tan, "Solving the dynamic vehicle routing problem under traffic congestion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 8, pp. 2367–2380, 2016.

[6] S. Chen, R. Chen, and J. Gao, "A modified harmony search algorithm for solving the dynamic vehicle routing problem with time windows," *Scientific Programming*, vol. 2017, Article ID 1021432, 13 pages, 2017.

[7] H. Abidi, K. Hassine, and F. Mguis, "Genetic algorithm for solving a dynamic vehicle routing problem with time windows," in *2018 International Conference on High Performance Computing & Simulation (HPCS)*, pp. 782–788, Orleans, France, 2018.

[8] H. E. Dongdong and L. I. Yinzhen, "Optimization model of green multi-type vehicles routing problem," *Journal of Computer Applications*, vol. 38, no. 12, pp. 3618–3624, 2018.

[9] S. Q. Fan, D. Lou, and Y. Sun, "Research on vehicle distribution path optimization of fresh agricultural products cold-chain logistics," *Storage Process*, vol. 17, no. 6, pp. 106–111, 2017.

[10] W. T. Fang and S. Z. Ai, "Research on cold chain logistics distribution path optimization based on hybrid ant colony algorithm," *Chinese Journal of Management Science*, vol. 27, no. 11, pp. 108–115, 2020.

[11] R. G. Thippa, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 99, 2020.

[12] T. R. Gadekallu, D. S. Rajput, M. P. K. Reddy et al., "A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU," *Journal of Real-Time Image Processing*, vol. 18, no. 4, pp. 1383–1396, 2021.

[13] A. Guha, D. Samanta, A. Banerjee, and D. Agarwal, "A deep learning model for information loss prevention from multi-page digital documents," *IEEE Access*, vol. 9, pp. 80451–80465, 2021.

[14] A. Colorni, M. Dorigo, and V. Mariiezzo, "Distributed optimization by ant colonies," in *Proceedings of the first European conference on artificial life*, pp. 134–142, Cambridge, MA, 1991.

[15] B. Chandra Mohan and R. Baskaran, "A survey: ant colony optimization based recent research and implementation on several engineering domain," *Expert Systems with Applications*, vol. 39, no. 4, pp. 4618–4627, 2012.

[16] D. Angus and C. Woodward, "Multiple objective ant colony optimisation," *Swarm Intelligence*, vol. 3, no. 1, pp. 69–85, 2009.

[17] M. López-Ibáñez and T. Stützle, *The impact of design choices of multiobjective antcolony optimization algorithms on performance: an experimental study on the biobjective TSP*, ACM, 2010.

[18] D. M. Chitty, "Applying ACO to large scale TSP instances," 2017, https://arxiv.org/abs/1709.03187.

[19] J. Li, P. Fu, X. Li, J. Zhang, and D. Zhu, "Study on vehicle routing problem and tabu search algorithm under low-carbon environment," *Chinese Journal of Management Science*, vol. 23, pp. 98–106, 2015.

[20] N. Labadie, C. Prins, and C. Prodhon, *Metaheuristics for Vehicle Routing Problems*, John Wiley & Sons, 2016.

[21] http://lbsyun.baidu.com/.

[22] L. L. Paquete, *Pareto Local Optimum Sets in the Biobjective Traveling Salesman Problem: An Experimental Study*, Springer, Berlin Heidelberg, 2004.

[23] A. Jaszkiewicz and T. Lust, "ND-tree: a fast online algorithm for updating a Pareto archive and its application in many-objective Pareto local search," 2016, https://arxiv.org/abs/1603.04798.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 4700204, 10 pages
https://doi.org/10.1155/2021/4700204

*Research Article*

# Research on Subway Pedestrian Detection Algorithm Based on Big Data Cleaning Technology

**Zhuoyang Lyu**

*College of Science, Purdue University, USA*

Correspondence should be addressed to Zhuoyang Lyu; lyurobin914@gmail.com

Received 6 August 2021; Revised 13 October 2021; Accepted 20 October 2021; Published 7 December 2021

Academic Editor: Rajesh Kaluri



The pedestrian detection model has a high requirement on the quality of the dataset. Concerning this problem, this paper uses data cleaning technology to improve the quality of the dataset, so as to improve the performance of the pedestrian detection model. The dataset used in this paper is obtained from subway stations in Beijing and Nanjing. The data images' quality is subject to motion blur, uneven illumination, and other noisy factors. Therefore, data cleaning is very important for this paper. The data cleaning process in this paper is divided into two parts: detection and correction. First, the whole dataset goes through blur detection, and the severely blurred images are filtered as the difficult samples. Then, the image is sent to DeblurGAN for deblur processing. 2D gamma function adaptive illumination correction algorithm is used to correct the subway pedestrian image. Then, the processed data is sent to the pedestrian detection model. Under different data cleaning datasets, through the analysis of the detection results, it is proved that the data cleaning process significantly improves the detection model's performance.

## 1. Introduction

Researches regarding data cleaning are first appeared in the United States, on the correction of social security number errors. The early research on data cleaning mainly focused on information data. The main research contents are as follows: (1) detection and elimination of abnormal data; (2) detection and elimination of approximate duplicate data; (3) data integration; (4) domain specific data cleaning.

Big data is the symbolic representation of this information-driven world. It has four characteristics: volume, variety, value, and velocity. It is gradually independent of software products and even dominated the development of some software products, such as Hadoop, Oracle, Hive, and Spark. Today, people can obtain massive amount of data from a variety of ways. After obtaining data, we often need to process them differently according to our specific purpose and extract valuable information from them. In order to get valuable information to meet people's needs, the data obtained should be reliable and accurate in reflecting the actual situation. However, the first-hand data we are able to collect is often dirty. Dirty data refers to inconsistent, inaccurate data resulting from human errors. Dirty data itself has the characteristics of inconsistency and inaccuracy, which directly affect its explicit and implicit value, that is, directly affect its quality [1].

The steps of data cleaning can be divided into the following steps:

(1) Demand analysis. The purpose of this stage is to clarify the format of effective data by analyzing the application field and application environment of the data and then the goal of data cleaning [2]

(2) Preprocessing. Through data analysis technology, we identify the quality problems existing in the dataset and summarize information regarding data's quality

(3) Determination of cleaning rules. This part analyzes the root causes of noise to define data cleaning rules. Different datasets have different characteristics, so the rules need to be selected to suit specific dataset [3]

(4) Cleaning and correction. This part involves cleaning the data according to the defined cleaning rules, using related technologies to correct the dirty data, and meeting the requirements of demand analysis. There are two general divisions among common data cleaning methods: repeated data detection and outlier data detection [4]. Repeated data detection includes field-based detection algorithm Levenshtein distance algorithm [5] and cosine similarity function algorithm. Levenshtein distance algorithm is easy to implement. Cosine similarity algorithm is more used to detect text similarity. The smaller the value of similarity measure obtained by this algorithm, the more similar the individuals are. Record-based detection algorithms include N-grams algorithm, clustering algorithm, SNM algorithm, and MPN algorithm [6]. N-grams algorithm generates a hash table and then judge the similarity between records according to the hash table; clustering algorithm classifies similar data into one class through calculation. The implementation of SNM algorithm is relatively easy, but it depends on keywords to a large extent and has strong dependency. The advantage of MPN algorithm is that it can collect the repeated data more comprehensively, but it is more cumbersome to use. Outlier detection is used to detect objects that are significantly different from other data points—outliers. Outlier detection algorithms mainly include the algorithm based on aggregate model, the algorithm based on proximity, the algorithm based on density, and the algorithm based on clustering [7]. The detection steps based on the statistical model algorithm are as follows: first, the data model is established, then the detection algorithm conducts analysis according to the model to identify outliers. Proximity-based algorithms define the proximity between objects. The core of density-based algorithm is to detect the local density of an object. When its local density is lower than that of most objects in the neighborhood, it is judged as an outlier. Cluster-based algorithms are used to find groups of objects that are locally strongly related, while outliers are objects that are not strongly related to other objects. After the test is completed, correct the wrong data according to the test results to achieve the purpose of cleaning

(5) Verification. Finally, the corresponding calibration operation is used to verify whether the cleaned data meets the requirements. If it does not meet the task requirements, the cleaning rules needs to be modified, the data cleaning process should be repeated, and the results can be verified and evaluated again. R-CNN [8] (region-based convolutional neural networks) algorithm, which was proposed in 2013, is a region-based CNN, which can be applied to the industrial field. Later, the region-based CNN has been further optimized, resulting in many better performance region-based convolutional neural networks, for example, the current mainstream detector: faster R-CNN [9]. The detector based on deep learning learns the features of the target autonomously through the backbone network in the training process while the traditional algorithm needs manually set features. The method based on deep learning is more robust and is easier to generalize

With a large number of scholars dedicated to this field, the algorithm is being improved continuously at present. On the contrary, the performance of the model is limited at the data level. Right now, the data quality of pedestrian datasets KITTI [10], Caltech [11], and CityPersons [12] published is relatively general, which means it is usually affected by uneven illumination and motion blur, the two prominent problems.

This paper designs the following steps through the data cleaning of the collected mass subway pedestrian pictures.

(1) Demand analysis. In view of these two prominent problems, this paper collects, cleans, and makes a dataset of subway pedestrians from real life scenes. Aiming at the image quality requirements of subway pedestrian detection task, we produce a high-quality subway pedestrian dataset

(2) Preprocessing. In the preprocessing step, the variance of the image is calculated according to the Laplace operator, the degree of blur of the image is identified, and the distribution of fuzzy image and clear image in the collected image is statistically analyzed

(3) Set cleaning rules. In this paper, a threshold is set according to the preprocessing results. If the variance of the image is less than the threshold, it will be regarded as a fuzzy image and its data will be cleaned

(4) Cleaning and calibration. For blurred images, this paper uses a DeBlurgan network for deblurring; for images with uneven illumination distribution, the illumination intensity is adjusted adaptively by using two-dimensional gamma function

(5) Check. In this paper, the dataset obtained by using different data cleaning rules will be sent into the classical YOLOV3 network to test the performance of the model and analyze the effectiveness of the data cleaning method used in this paper in the pedestrian detection task

The structure of this paper is as follows: the second section is the introduction and quality analysis of the dataset. The third section is the method of data cleaning and verification we used. The fourth section is the experimental design and results, and the fifth section is the conclusion of this paper.

## 2. Dataset

*2.1. Subway Pedestrian Dataset.* Due to the relatively dense number of passengers in the subway station and the height

and angle of the monitoring camera, when the crowd is dense, pedestrian's trunk is easy to block each other, and the head-shoulder positions are generally relatively complete. Therefore, the detection model based on the head-shoulder positions is established for the detection of pedestrians. Subway pedestrian dataset was collected by monitoring video of Beijing subway station. First, the video data was read in frame by frame, and the generated pictures were stored locally in JPG format. With reference to the format of VOC2007 dataset, a total of 17774 original pictures of multiple scenes were processed and annotated.

Our dataset is a pedestrian dataset obtained from subway station, which contains a large number of occlusion scenes. It can effectively evaluate the robustness of the detector to occlusion problems. It contains a total of 9,000 images in training set. These pictures are all from some subway stations in Beijing and Nanjing. The average number of pedestrians per picture is 13.36, more than Caltech and CityPersons. As shown in Table 1, our dataset is more challenging than the Caltech and CityPersons benchmark datasets.

Subway pedestrian dataset was made, and labelme software was used to mark the pedestrian head-shoulder positions with a rectangular frame. The marking box should contain as much pedestrian head and shoulder positions as possible while containing as little background information as possible. The obtained subway pedestrian dataset contains the passenger flow situation at different times and places in the subway station. When annotated, XML files are generated in the same folder; as shown in Figure 1, the labelme software and the information are contained in the XML file.

*2.2. Quality Analysis of Dataset.* The collected subway pedestrian images are inevitably affected by various factors in the process of acquisition, storage, and transmission. And then produce different types of distortion and different degrees of distortion, in which blur distortion is the most common. Blur distortion leads to the degradation of image quality, which affects the accuracy of pedestrian detection. So, we use the method of blur detection to analyze the image quality of the dataset.

This paper uses Laplacian function of OpenCV to detect image blur. Because Laplacian operator is used to measure the second derivative of the image, it can emphasize the region with fast changing density in the image, that is, the boundary region. In the general picture, the boundary is clear, so the variance will be larger; however, there is little boundary information in the blurred image, so the variance will be small. Firstly, a channel of the image is selected and convoluted with $3 * 3$ convolution kernel to calculate the variance of the output. The formula of convolution is shown as

$$\int_{+\infty}^{-\infty} f(t)y(x-t)dt. \tag{1}$$

$f(t)$ represents the convolution kernel, and $y(x-t)$ represents a channel of the image.

If the variance is less than the set threshold, it is regarded as a blurred image. We set the threshold value to 40. If the function return value is less than 40, it is considered as a blurred image. If the function value exceeds 40, it is a clear image. At the same time, we divide the blurred image into different grades. The range of 0-10 is regarded as level 1 blur; 10-20 is two-level blur; 20-30 is grade 3, and 30-40 is grade 4.

We analyze the image quality of all the datasets. The images with variance greater than 40 are regarded as clear images, and the images with variance less than 40 are processed for subsequent deblurring.

As shown in Figure 2, there are four levels of blur distribution in the subway pedestrian dataset. It can be seen from the observation that a large number of images are gathered in the first and second levels of blur, so it is necessary to deblur the dataset. More than 60% of the images in subway pedestrian dataset are blurred. A large number of blur samples will affect the training effect of the network, so it is necessary to deblur the images in the dataset.

## 3. Data Cleaning Algorithm

*3.1. DeblurGAN.* Motion blur image is generated by relative movement of equipment and target during image acquisition process. In subway station, the monitoring camera is generally on the high place with a certain inclination angle. When making datasets by intercepting monitoring video frames, the blurred picture will appear when passengers move quickly, which not only affects the image quality, but also makes it difficult to detect pedestrians.

DeblurGAN, an end-to-end learning method for generating network and content loss based on conditional antagonism, removes blurring of images due to pedestrian movements. When the image to be detected is blurred, first, blurring the image can improve the accuracy of pedestrian detection.

The blurring of the image can be seen as the convolution of the original image and the convolution kernel plus additive noise. It can be expressed as

$$I_B = k * I_S + N. \tag{2}$$

$I_B$ and $I_S$ represent blurred image and clear image, respectively; $k$ is unknown blur kernel; $N$ is additive noise. Most algorithms rely on the classic Lucy Richardson algorithm [13] and Wiener or Tikhonov filter to perform deconvolution operation and obtain the estimation [14], which is to restore the blurred image with known blur kernel. But usually, the blur function is unknown, so it is uncertain to find the blur function for each pixel. DeblurGAN processes the blurred image $I_B$ as input without information about the blur kernel to get a clear image $I_S$. In DeblurGAN training phase, CNN is trained as generator network $G_\theta$ and discriminator network $D_\theta$ by constructing Generative Adversarial Network, with pairs of blur image and clear image as input, and finally, clear image is reconstructed through the means of adversarial. After the training, the whole process of deblurring is completed by the trained

Table 1: Our dataset compare with public benchmark datasets.

| | Caltech | KITTI | CityPersons | COCOPersons | Our datasets |
|---|---|---|---|---|---|
| Images | 42,782 | 3,172 | 2,975 | 64,115 | 9,000 |
| Persons | 13,674 | 2,322 | 19,238 | 257,252 | 120,325 |
| Person/image | 0.32 | 0.63 | 6.47 | 4.01 | 13.36 |

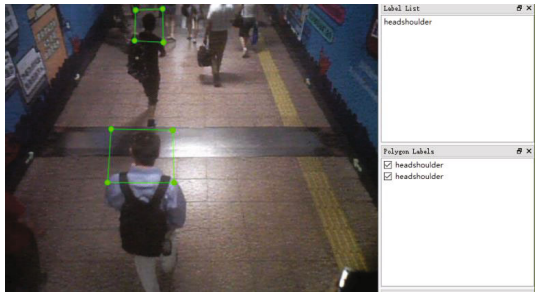



Figure 1: labelme software.

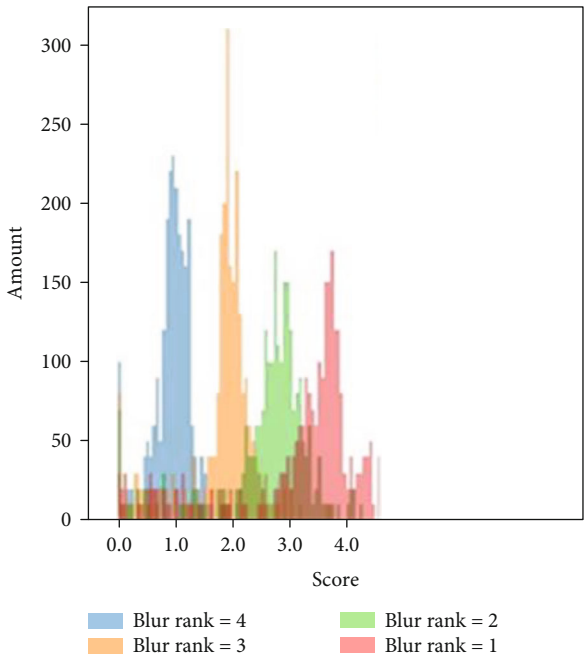



Figure 2: Image blur level distribution in dataset.

network $G_\theta$. In this case, only the input blurred image $I_B$ is restored to get a clear image $I_S$, so as to achieve the result of motion blur removal.

DeblurGAN's loss function consists of two parts: antiloss $L_{\text{GAN}}$ and content loss $L_x$. It can be expressed as

$$L = L_{\text{GAN}} + \alpha L_x. \tag{3}$$

$\alpha$ is 100 in this deblurring experiment, which means the weight of content loss. WGAN-GP is used as the antiloss function, which is robust to the training of the generated network. It can be expressed as

$$L_{\text{GAN}} = \sum_{n=1}^{N} -D_\theta(G_\theta(I_B)). \tag{4}$$

Content loss uses the perceptual loss function, which is a simple L2 loss. The difference of each layer's feature map between the generated image CNN and the target image CNN is calculated, and the final cumulative error is the perceptual loss. The calculation in Equation (5) shows that $\phi_{i,j}$ is the feature map obtained by the $i$ activated convolution layer before the $j$ largest pooling layer of vgg19 network trained on ImageNet dataset; $W_{i,j}H_{i,j}$ represents the dimension of the feature map.

$$L_x = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W} \sum_{y=1}^{H} \left(\phi_{i,j}(I_S)_{x,y} - \phi_{i,j}(G_\theta(I_B))_{x,y}\right)^2. \tag{5}$$

*3.2. 2D Gamma Function.* In subway stations, pedestrians block each other, and different areas have different light irradiation intensity, which often leads to uneven illumination around pedestrians. This is mainly reflected in the insufficient illumination in some areas of the image, and the excessive illumination in some areas of the image. Some image details cannot be extracted in the test, which seriously affects the pedestrian detection results. Therefore, it is necessary to correct the uneven illumination of subway pedestrian image to eliminate the influence caused by uneven illumination as far as possible.

Generally speaking, the digital image can be regarded as a 2D function $f(x, y)$, which is obtained by multiplying the incident light component $i(x, y)$ and the object surface reflection component $r(x, y)$,

$$f(x, y) = i(x, y) * r(x, y). \tag{6}$$

The spatial relationship is shown in Figure 3. For images with uneven illumination, it is the uneven distribution of incident illumination component that causes the image brightness value to be too large in areas with strong illumination, while the image brightness value in areas with weak illumination is too small. It is very important to extract the incident light component from the illumination correction of the image with uneven illumination. The illumination component is extracted by multiscale Gaussian function and the Gaussian function formula as shown

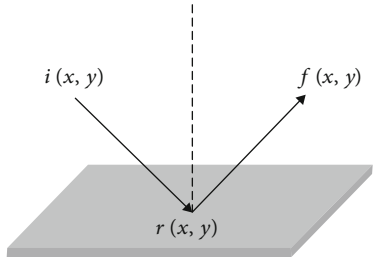



Figure 3: Spatial relation of object illumination.

in Equation (7).

$$G(x,y) = \lambda \exp\left(-\frac{x^2+y^2}{c^2}\right), \tag{7}$$

where $c$ is the scale factor and $\lambda$ is the normalization constant, and the Gaussian function is required to meet the normalization condition $\iint G(x,y)dxdy = 1$. The Gaussian function is convolved with the input image $F(x,y)$ to obtain the estimated value of the illumination component $I(x,y) = F(x,y)G(x,y)$. The multiscale Gaussian function method is adopted to extract the illumination component by using the Gaussian function of different scales, and then, the illumination component is weighted. Finally, the estimated value of the illumination component is obtained. The formula is shown in Equation (8), which $\omega_i$ represents the weight coefficient of the illumination component corresponding to the Gaussian function of the $i$ scale.

$$I(x,y) = \sum_{i=1}^{N} \omega_i [F(x,y)G_i(x,y)]. \tag{8}$$

After the illumination component is extracted, an adaptive brightness correction method based on the 2D gamma function is constructed. According to the distribution characteristics of the illumination component, the parameters of the 2D gamma function are adjusted adaptively, and the image with uneven illumination is corrected, so as to reduce the brightness value of the area with too strong illumination and increase the brightness value of the area with too low illumination, so as to achieve the effect of processing the image with uneven illumination. This allows the model to learn more details about the dark parts of the image. For the input image $F(x,y)$, assuming that the extracted illumination component is $I(x,y)$, the improved 2D gamma function expression is shown in Equation (9), which $O(x,y)$ represents the brightness value of the corrected image, $\gamma$ represents the index value of brightness enhancement, and $m$ represents the mean brightness value of the illumination component.

$$O(x,y) = 255\left(\frac{F(x,y)}{255}\right)^{\gamma}, \gamma = \left(\frac{1}{2}\right)^{(m-I(x,y))/m}. \tag{9}$$

*3.3. Pedestrian Detection Based on YOLOV3.* Object detection algorithms based on deep learning mainly include two types, one based on anchor frame and divided into two stages and one stage. Two-stage detection methods, such as RCNN series et al. [15, 16], first generate a group of candidate bounding boxes that may contain targets by using the region proposal module and then classify and regression these borders by using deep convolutional neural network [17, 18]. One-stage detection methods, such as YOLO series [19, 20] and SSD [21], unify all modules of target detection into a single convolutional network, enabling it to simultaneously predict the probability of multiple bounding boxes and categories. The other is anchor-free detection method, such as CornerNet [22] and ExtremeNet [23]. As a one-stage object detection method, YOLOV3 can locate the object in the input image and predict its category at the same time, thus transforming the object detection problem into a regression problem. The overall detection process of its network is shown in Figure 4.

We use the PyTorch framework, and the resolution of the input image is 416 ∗ 416. After passing through multiple convolution layers, the data of three scales will be output. If we use the COCO dataset, there are 80 categories, namely, (N,255,13,13), (N,255,26,26), and (N,255,52,52). Since there is only one type of target to be detected in subway pedestrian detection process (marked with head-shoulder), the number of output categories of YOLOV3 network is 1 by modifying the length of network prediction tensor is 18, and the three scales are (N,18,13,13), (N,18,26,26), and (N,18,52,52), respectively. Each figure is divided into 3 priori box positions on the grid of 13, 13, 26, 26, 52, and 52.

## 4. Experimental Results and Analysis

*4.1. DeBlurGAN Removes Blur.* The entire structure of the DeBlurGAN training network for motion blur removal is shown in Figure 5, where the generator network takes the blur image as input and produces the reconstructed image. During training, the discriminant network takes the reconstructed image and the original clear image as input and estimates the distance between them. The generator network structure, shown in Figure 6, consists of two step convolution blocks with one half of the stride size, nine residual blocks (ResBlocks), and two transpose convolution blocks. Each ResBlock consists of a convolution layer, an instance normalization layer, and a ReLU activation layer. Add a missing regularization with a probability of half after the first convolution layer in each ResBlock. In addition, there is a global skip connection called ResOut. The DeBlurGAN discriminator network architecture still uses Patch-GaN from Pix2Pix. In this paper, through the use of GoPro dataset (part), a total of 1146 pairs of 720 ∗ 720 blur-clear image pairs were taken from different scenes, 200 iteration training was carried out in the TensorFlow framework of Linux system, and the training result model was saved every 20 times by modifying the network settings. For the blur image of subway pedestrians, there is no image processing and no corresponding clear image, so the supervised method cannot

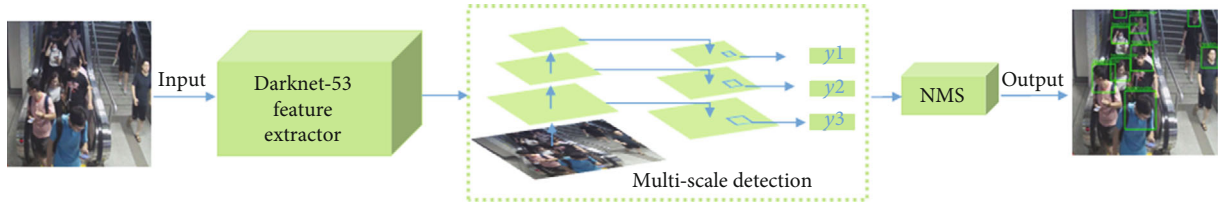



Figure 4: Detection process of YOLOv3.

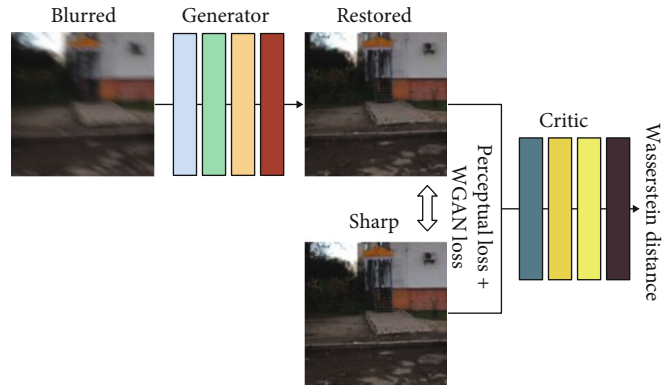



Figure 5: The architecture of DeblurGAN training network.

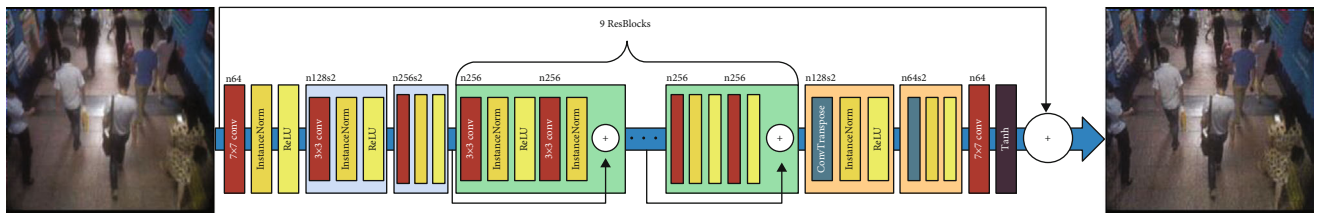
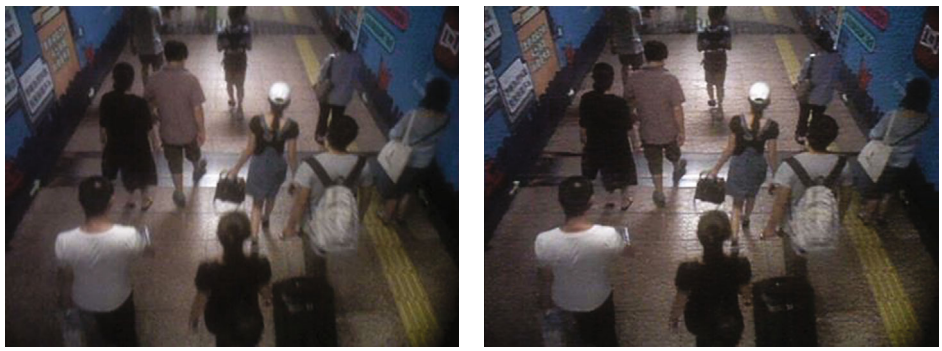



Figure 6: The generator architecture of DeblurGAN.

be used to conduct deblurring training on this dataset. However, the dataset is derived from the actual subway scene that needs to be deblurred and has practical application significance, so it can be used as a test dataset. By calling the training model of subway pedestrian to deal with the blur dataset, to the whole process of blur network of training alone, because the original network output picture image resolution is relatively low, the changes to the network are not reducing image characteristics to the original size to save to the pedestrian subway after blur images.

DeblurGAN was used to process the subway pedestrian dataset to obtain the deblurted image (as shown in Figure 7). It can be seen that compared with the original Figure 7(a), the deblurted image in Figure 7(b) is clearer, the detailed texture in the image is more prominent, and the pedestrian contour on the left of the image is more obvious. It is convenient to detect the head-shoulder of pedestrians in the image. In Figure 8, the model obtained from image training before and after deblurring is detected through the pedestrian detection network. As shown in Figure 8(a), the two pedestrians at the bottom of the image are not detected. As shown in Figure 8(b), the texture details are more visible in the deblurred image, so they are successfully detected.

*4.2. Adaptive Luminance Correction Algorithm for Two-Dimensional Gamma Function.* Using multiscale Gaussian function to extract the subway dataset nonuniform illumination image of light weight, structure based on 2D adaptive brightness adjustment function of the Gamma function, and using the distribution characteristics of light weight adaptively adjust the 2D gamma function parameter and adaptive correction in nonuniform illumination image processing. On the premise of effectively retaining the effective information of the original image, the purpose of correcting the image with uneven illumination can not only effectively improve the visual effect of the pedestrian detection image but also find more details of the dark place in the image. The RGB color space of the input pedestrian detection image is transferred to the HSV space, and the V (brightness) component of the HSV space is operated without affecting the color information of the image. The multiscale Gaussian filter of Retinex is used to obtain the incident light component, and then, the 2D gamma function is used. The image



(a) Before deblurring (b) After deblurring

Figure 7: Comparison between the images before and after deblurring.

(a) Detect before deblurring (b) Detect after deblurring

Figure 8: Comparison between the detection images before and after deblurring.

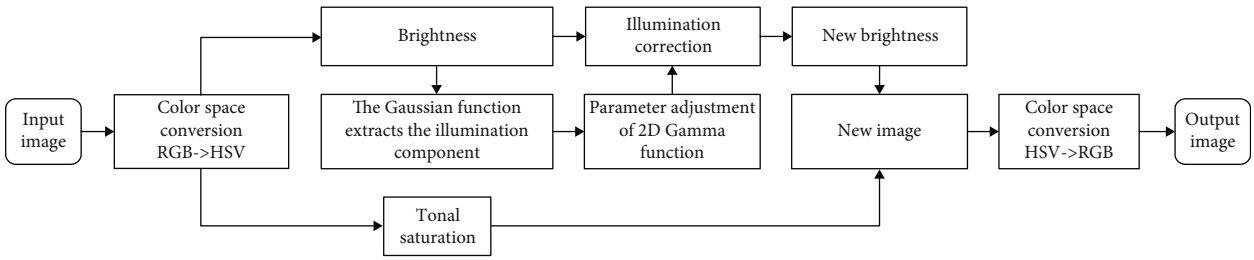



Figure 9: Flow chart of nonuniform illumination correction algorithm.

brightness is corrected by changing the brightness, and then, the image is synthesized with T(tonal) and S(saturation) components, and then, the image is returned to the RGB color space to output the corrected image of uniform illumination. In this paper, the illumination correction program was written by MATLAB under Windows to deal with the image dataset of subway pedestrians with uneven illumination in batches. In order not to affect the subsequent entry into the object detection network, the illumination correction pictures were saved in full size. The algorithm flow chart is shown in Figure 9.

The illumination component is extracted from Figure 10(a) of subway pedestrians with uneven illumination to obtain the Figure 10(b) of the corresponding light component. As shown in Figure 10(a), the brightness of the middle part of the original image is larger due to the illumination of subway lights, while the brightness is darker if there is no direct illumination around. The middle part of the Figure 10(b) after the illumination component is also larger. Figure 10(c) of illumination correction processing was obtained by self-adaptive correction processing. Compared with the original image, the brightness of the middle part decreased, and the brightness of the four corners increased significantly.

After testing the model obtained from the training of the image dataset before and after the illumination correction treatment, the comparison of the detection images before and after the illumination correction treatment is shown in Figure 11. Figures 11(a) and 11(c)are the preillumination models to detect the images before illumination processing,

(a) Before illumination correction

(b) Illumination component

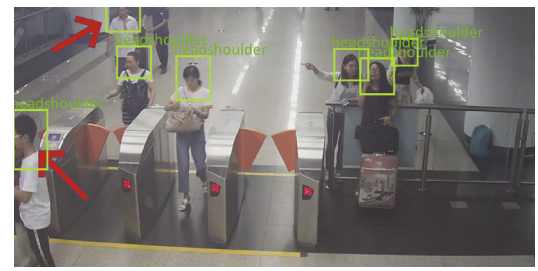
(c) After illumination correction

Figure 10: Comparison between the images before and after illumination correction.


(a) Detect before illumination correction


(b) Detect before illumination correction


(c) Detect after illumination correction




(d) Detect after illumination correction

Figure 11: Comparison between the detection images before and after illumination correction.

Table 2: The default anchor box size before and after $K$-means clustering.

| Feature maps of different sizes | $13 \times 13$ | $26 \times 26$ | $52 \times 52$ |
|---|---|---|---|
| Anchor before clustering | (116, 90) | (30, 61) | (10, 13) |
| | (156, 198) | (62, 45) | (16, 30) |
| | (373, 326) | (59, 119) | (33, 23) |
| Anchor after clustering | (135, 120) | (99, 92) | (48, 48) |
| | (159, 166) | (105, 113) | (63, 71) |
| | (202, 230) | (119, 146) | (80, 83) |

Table 3: Datasets and models corresponding to different image processing.

| Pedestrian image processing | Dataset | Model |
|---|---|---|
| The original image | Dataset I | Model I |
| Deblurring | Dataset II | Model II |
| Illumination correction | Dataset III | Model III |
| Deblurring+ illumination correction | Dataset IV | Model IV |
| Illumination correction + deblurring | Dataset V | Model V |

Table 4: Comparison of test results of corresponding models in each dataset.

| Model | Number of test-set images | Number of pedestrian instances | Recall | Precision | mAP |
|---|---|---|---|---|---|
| Model I | 3555 | 30717 | 0.83 | 0.68 | 0.78 |
| Model II | 3555 | 30717 | 0.85 | 0.69 | 0.79 |
| Model III | 3555 | 30717 | 0.85 | 0.83 | 0.83 |
| Model IV | 3555 | 30717 | 0.87 | 0.81 | 0.84 |
| Model V | 3555 | 30717 | 0.88 | 0.78 | 0.85 |

and it is found that there are false detection and redundant detection frames, etc. After illumination correction, the brightness of pedestrians in the dark environment around the picture will increase. Figures 11(b) and 11(d) accurately detect pedestrians without false detection and redundant detection frames.

*4.3. Pedestrian Detection Based on YOLOv3.* In this paper, Yolov3 network is used to train and detect the subway pedestrian dataset. Three anchor frames are set at each scale. Before training, $K$-means clustering is performed on the label frame of the subway pedestrian dataset in this paper to calculate the initial value of the anchor frame in the training set, making the size of the anchor frame more consistent with the size of the pedestrian head and shoulder. The size of the default Anchor box before and after clustering is shown in Table 2.

The original dataset of subway pedestrians is named dataset I.

Dataset II was obtained from dataset I after DeblurGAN deblurring.

Dataset III was obtained from dataset I after illumination correction.

Dataset IV was obtained from dataset I after DeblurGAN deblurring and illumination correction.

Dataset V was obtained from dataset I after illumination correction and DeblurGAN deblur processing.

As shown in Table 3, the deep convolutional neural network YOLOV3 was used for multiple rounds of training under the framework of PyTorch. The detection models obtained from the corresponding training of five datasets were named as model I, model II, model III, model IV, and model V, respectively.

The same YOLOV3 detection network under the PyTorch framework was used to test the models I, II, III, IV, and V obtained by training, respectively. The model file sizes of the five models were almost the same. When the speed was tested, 10 of the targets in the video had speeds of around 17-19 fps below them, and 10 of the targets had speeds of around 13-16 fps above them. The number of test pictures is 3555, including 30,717 subway pedestrian head-shoulder targets. The model detection results are shown in Table 4. It can be seen that image DeblurGAN deblurring, uneven illumination adaptive correction, first DeblurGAN deblurring followed by uneven illumination adaptive correction, first DeblurGAN deblurring followed by uneven illumination adaptive correction, and then DeblurGAN deblurring will all improve the mean detection accuracy (mAP) of the model. Among them, model IV and model V are higher than model II and model III in mAP, indicating that

the combined operation effect of two treatment methods of DeblurGAN deblurring and illumination uneven adaptive correction is better than that of one treatment without any sequence.

## 5. Conclusion

In this paper, it is considered that the metro pedestrian dataset with large data volume and low data quality is the main reason for the poor performance of pedestrian detection model. Therefore, the data cleaning technology is introduced into the subway pedestrian detection system. We first use Laplace operator to carry out blur detection on subway pedestrian images and divide the images in the dataset into clear pictures and blur pictures. We also used the DeblurGAN network to deblur the blurred image and further used the 2D gamma function to equalize the light in the image. Through the use of different combination of data cleaning methods and the verification of YOLOV3 algorithm, the rationality of our hypothesis is verified, and the performance of pedestrian detection algorithm is significantly improved by data cleaning.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no competing interest.

## References

[1] F. Wenfei, "Extending dependencies with conditions for data cleaning," in *8th IEEE International Conference on Computer and Information Technology*, pp. 185–190, Sydney, NSW, Australia, 2008.

[2] D. Aebi and L. Perrochon, *Towards Improving Data Quality*, In CiSMOD, 1993.

[3] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[4] G. Salon and M. J. Mcgill, *Introduction to Modern Information Retrieval*, McGraw-Hill Book Co, New York, 1983.

[5] Y. Li and L. Bo, "A normalized Levenshtein distance metric," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[6] H. Galhardas and D. Florescu, "An extensible framework for data cleaning," in *Proceedings of the 16 th IEEE International Conf erence on Data Engineering*, pp. 312–312, San Diego, California, 2000.

[7] N. Deepa, Q. V. Pham, D. C. Nguyen et al., "A survey on blockchain for big data: approaches, opportunities, and future directions," 2020, https://arxiv.org/abs/2009.00858.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference oncomputer vision and pattern recognition*, pp. 580–587, Columbus, Ohio, the United States, 2014.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," 2015, https://arxiv.org/abs/1506.01497.

[10] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision Meets Robotics: The KITTI Dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[11] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[12] S. Zhang, R. Benenson, and B. Schiele, "City persons: a diverse dataset for pedestrian detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, the United States, 2017.

[13] M. K. Singh, U. S. Tiwary, and Y. H. Kim, "An adaptively accelerated Lucy-Richardson method for image deblurring," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, 10 pages, 2007.

[14] E. Nursultanov, M. Ruzhansky, and S. Tikhonov, "Nikolskii inequality and Besov, Triebel-Lizorkin, Wiener and Beurling spaces on compact homogeneous manifolds," 2014, https://arxiv.org/abs/1403.3430.

[15] R. Girshick, "Fast r-cnn," in *In proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Piscataway, NJ, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[17] A. R. Javed, M. Usman, S. U. Rehman, M. U. Khan, and M. S. Haghighi, "Anomaly detection in automated vehicles using multistage attention-based convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[18] A. Rehman, S. U. Rehman, M. Khan, M. Alazab, and T. Reddy, "CANintelliIDS: detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-based GRU," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 1456–1466, 2021.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, Las Vegas, Nevada, the United States, 2016.

[20] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[21] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, Cham, 2016.

[22] H. Law and J. Deng, "Cornernet: detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 734–750, Munich, Germany, 2018.

[23] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 850–859, California, the United States, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6567346, 11 pages
https://doi.org/10.1155/2021/6567346

*Research Article*

# Effective Passive Multitarget Localization Using Maximum Likelihood

**Yasir Munir**,[1] **Muhammad Umar Aftab**,[2] **Danish Shehzad**,[2] **Ali M. Aseere**,[3] **and Habib Shah**[3]

[1]*Department of Electrical Engineering, Government College University Faisalabad, 38000, Pakistan*
[2]*Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad, Chiniot-Faisalabad Campus, Chiniot 35400, Pakistan*
[3]*Department of Computer Science, College of Computer Science, King Khalid University, Abha 62529, Saudi Arabia*

Correspondence should be addressed to Muhammad Umar Aftab; ms.umaraftab@yahoo.com

Received 4 August 2021; Revised 19 September 2021; Accepted 7 October 2021; Published 30 November 2021

Academic Editor: Rajesh Kaluri



Localization of multiple targets is a challenging task due to immense complexity regarding data fusion received at the sensors. In this context, we propose an algorithm to solve the problem for an unknown number of emitters without prior knowledge to address the data fusion problem. The proposed technique combines the time difference of arrival (TDOA) and frequency difference of arrival (FDOA) measurement data fusion which further uses the maximum likelihood of the measurements received at each sensor of the surveillance region. The measurement grids of the sensors are used to perform data association. The simulation results show that the proposed algorithm outperforms the multipass grid search and further effectively eliminated the ghost targets created due to the fusion of measurements received at each sensor. Moreover, the proposed algorithm reduces the computational complexity compared to other existing algorithms as it does not use repeated steps for convergence or any biological evolutions. Furthermore, the experimental testing of the proposed technique was executed successfully for tracking multiple targets in different scenarios passively.

## 1. Introduction

In the modern era of wireless technologies, the localization of the target sensors such as aircraft, ships, or unmanned vehicles is challenging. The challenge is magnified further when positions and velocities of sensor targets cannot be estimated precisely which results in the inaccuracy of sensor's locations and massive consequences in the practical environment [1]. Most of the previous research studies utilized localization methods that mainly depend on the accuracy and robust estimation, in which the time difference of arrival (TDOA) has significant importance. This TDOA is based on receiving the same signal on different sensors or receivers varying in time. Further, it is used to calculate the difference in the signal's arrival with respect to the reference sensor. Emitter devices have the benefit of providing the frequency difference of arrival (FDOA) which is the result of the relative motion of the source and the target, which improves the accuracy with the estimation of the velocity of the target [2]. Passive localization of a target or multiple targets has one advantage over active localization, that passive localization is stealth in nature, which localizes the target without letting the target know about the existence of the sensors.

Consider passive localization using TDOA/FDOA measurement to initially estimate a single target and develop an algorithm to estimate multitargets. Compare the performances using different metrics, including geometric dilution of precision (GDOP) [3], measurement covariance, sensor's self-navigation error, data-link transmission delay, and sensor geometry to help in better understanding of the research topic. Also, there is a need for a well-developed, efficient, and effective algorithm to localize the multitargets which should

be able to deal with the complexity of the measurement fusion and localize multitargets and further track the targets without any prior information. So, consider the increasing interest to have such an algorithm that can localize the increasing number of emitter devices.

A previous work using passive localization extensively focuses on a single target, whereas very few studies in the literature are found related to multitarget localization. Localization of multiple targets is very complicated and difficult to estimate their positions than a single target as sensors receive multiple signals, and it becomes a fusion of signals, therefore becoming complex to associate the exact signal to the target. It becomes worst in the presence of high noise. Also, the data reduction can cause the loss of possible target detection in a specific area due to noise [4].

The localization problem is challenging. However, this problem can be solved using different approaches—the Taylor series [5] with a good initial guess. Chan and Ho [6] and Ho and Xu [7] proposed a two-step weighted least square (WLS) [8] and total least square and the semidefinite programming (SPD) [9] method based on ML, whereas localization of the target can also be achieved using a reconfigurable intelligent surface (RIS) supported with millimeter-wave multi-input multi-output radar system [10] and based on link analysis in passive UHF RFID to identify real targets and eliminate false targets [11].

Sensor geometry along with the number of sensors is essential in localization. Strong sensor geometry results in a small GDOP [3] value and will cause low position uncertainty, where the number of sensors is the means of measurements. As the number of sensors increased, the estimate's accuracy and efficiency increase, but the problem in localization is to decrease the number of sensors without decreasing the accuracy, or say reasonable accuracy should be maintained [12]. In the future, the proposed algorithm could be extended in vehicle identification using the RF signals for traffic surveillance [13, 14].

The proposed algorithm is able to localize the unknown number of targets having the complexity of the multiple sensors with multiple grids; the algorithm is not complex as it is not using repeated steps for convergence, not using any biological evolutions used in the existing algorithms; in a single scan, it computed likelihood for all sensors using only one parameter grid and passes the results for multitarget estimation. It also eliminates the localization of the ghost target which becomes an issue when the measurement data fusion is received; here, we call them possible candidates when considering the combinations of the measurements from different sensors.

This paper is organized to easily understand the research purpose from the background and primary literature to the research topic. Including the introduction, it is divided into five sections. The second section describes the related work to the proposed algorithm, while the third section explains the proposed algorithm step by step using mathematical equations and block diagrams. The fourth section shows the results of the algorithm, and the last section concludes this research paper.

## 2. Applications Using Maximum Likelihood for Multitargets

Few applications make use of maximum likelihood for tracking multiple targets. But there is not much in the literature. Following are the three optimization algorithms:

(1) Multipass grid (MPG)

(2) Genetic algorithm (GA)

(3) Directed subspace (DSS)

All the above algorithms use a set of certain threshold measurements over several frames (data window length) from a detection processor. The time to calculate a track estimate is primarily a function of the maximization routine used on the LLR (Log-Likelihood Ratio). This, in short, depends upon the number of data frames involved in the estimate, the detections in each data frame, and the number of LLR calculations required by the maximization algorithm [15].

The parameter space is the measurements comprised of bearing, range, range rate, and amplitude values. Dissimilarities in the dimensionality of the observation space, window length, detector $P_{fa}$, and target SNR (signal to noise ratio) are considered. Let us have a brief overview of each one by one.

*2.1. Multipass Grid (MPG).* In this method, $K$ steps are involved. A set of values $c_k$ are monotonically decreasing, here $k = 1, 2, \cdots, K$, and are established with $c_k = 1$. A grid search is implemented over the parameter space using the artificially improved measurement noise standard deviations. The standard deviation of every measurement component is amplified by multiplying it by a parameter defined $c_1$. From the grid search, the best value resulting is forwarded to a local optimization routine, for example, a Newton-Raphson or Davidon-Fletcher-Powell.

During each successive step, the smaller and new values of measurement noise standard deviation are used. The local optimization routine is started in steps from the parameter value it had converged in the previous step [16]. Repeat the process until $k = K$ at which the measurement noise standard deviations are restored to their actual value. Once it has converged by the local optimization routine at the final step, the track estimate is obtained [17].

Using a multipass grid search has an advantage as it requires less computation than a comprehensive search. However, for the more composite difficulty, the multipass grid requires a large number of evaluations and additional calculations to achieve the results.

*2.2. Genetic Algorithm (GA).* Over a discrete parameter space, a stochastic search is performed using a set of rules based on biological evolutionary development. Theory suggests that when using the GA, one is essentially searching more of the parameter space than that is reflected in the number of LLR evaluations corresponding to a purely

random search computing up to the cube of the number of LLR evaluations used in the genetic search.

Genetic algorithm is effective and capable of searching the global maximum which has been shown for a varied class of objective functions [18] and in many cases has performed the best of other optimization methods [19].

In practice, the employment of GA is not able to find the global maximum with probability 1 of the random neutral function. Results in [20] show the asymptotic convergence beyond the critical population size. However, it is challenging to limit the essential population size in a particular problem. Table 1 lists the steps which are performed to generate one generation using a genetic algorithm [18].

*2.3. Directed Subspace (DSS).* This algorithm is motivated by a methodology of using the data information of the measurements to guide the search. Directed subspace utilizes the measurement information itself to select and search regions of parameter space that might contain the LLR global maximum while avoiding those parameter space regions that cannot have this maximum.

In many tracking applications, the space of measurements is defined as the subspace of the parameter space that can be of any dimensional measurement space. For example, range, range rate, and bearing become a 3-dimensional measurement space. Range and bearing can map to the Cartesian positions and range rate is a counterpart to radial velocity. Table 2 lists the steps in the direct subspace algorithm search [19].

Once the LLR values are computed over the grid using the measurements, a local optimization algorithm [21] is used for the final converged parameter.

*2.3.1. Window-Based ML-PDA Algorithm.* The window-based ML-PDA algorithm is developed in a way to use in real-time applications; to compute the track estimate, a subset of the $N_w$ most recent data frames is used [22]. When a new frame is received with data, the ML-PDA algorithm is recurred, as it adds the new data frame; the oldest data frame is removed from the data set, which creates a sliding window for localization and tracking update.

The existing work requires a large number of evaluations and computations to achieve results where the employment of GA is not able to find the global maximum with probability 1 of random neutral function. It is challenging for existing work to limit the essential population size in a particular problem, whereas the proposed algorithm requires no convergence and less computation. Some of the productive work regarding the multitarget localization and object recognition algorithms are experimented using different distributed algorithms, and some of them are introduced for tracking of vehicles along with other items of interest [23–26].

Table 1: Steps of genetic algorithm.

| Step | Action |
|---|---|
| 1 | Calculate the fitness function for every population fellow |
| 2 | Selection of the reproduction population |
| 3 | Selection of the reproduction population mates |
| 4 | Child population production (cloning or crossover) |
| 5 | Apply alterations to the child population |
| 6 | Trial experiment for convergence |

Table 2: Steps of the direct subspace algorithm.

| Step | Action |
|---|---|
| 1 | Setting grid density for the free parameter(s) |
| 2 | Mapping one measurement to parameter space |
| 3 | Using the measurements, calculate LLR over the grid of free parameter(s) |
| 4 | Repeat steps 2 and 3 for all measurements in the data set |
| 5 | Forward the finest result in the local optimization routine |

## 3. Proposed Multitarget Localization and Tracking

In conventional passive localization and tracking, localization of multiunknown targets is a hard problem. Unlike in single-target localization, sensors receive multiple measurements that depend on the number of targets in the specific region, which causes a problem for the receiver to associate measurements to the specific targets.

Figure 1 shows the arrangement of three sensors along with three targets in a surveillance region. Localization of multiple targets is complex due to the data association ambiguity; i.e., there is no information of which TDOA/FDOA measurement is associated with the specific target [27]. The more complex problem occurs when the measurements from the target are not received by the sensor, and there might be few measurements that are considered false. Such problems exist in a real scenario [17]. Receiving multiple measurements also creates ghost targets which become more challenging for the ghost targets in the surveillance region to be eliminated [28]. To solve this issue, an algorithm is proposed to address the data association ambiguity by using the maximum likelihood of the measurements, which is further processed to localize the multiple targets using the least number of sensors. Maximum likelihood estimation (MLE) is extensively used for a single target for minimizing the estimation problem [29]. Multiple grids are created considering the complexity of the multiple sensors, whereas the algorithm is not complex as it is not using repeated steps for convergence or any biological evolutions of a genetic algorithm [19].

Consider the simplicity of this algorithm that is solving the fusion of measurements. Moreover, it can accurately localize the targets using a computed likelihood for all sensors. It has only one parameter grid and passes the results for multitarget estimation. Our proposed algorithm confirms the unknown number of targets that can be localized correctly when the number of false alarms is low and the high probability of detection.

*3.1. Main Steps Summary.* The proposed algorithm is divided into different steps to understand the functionality

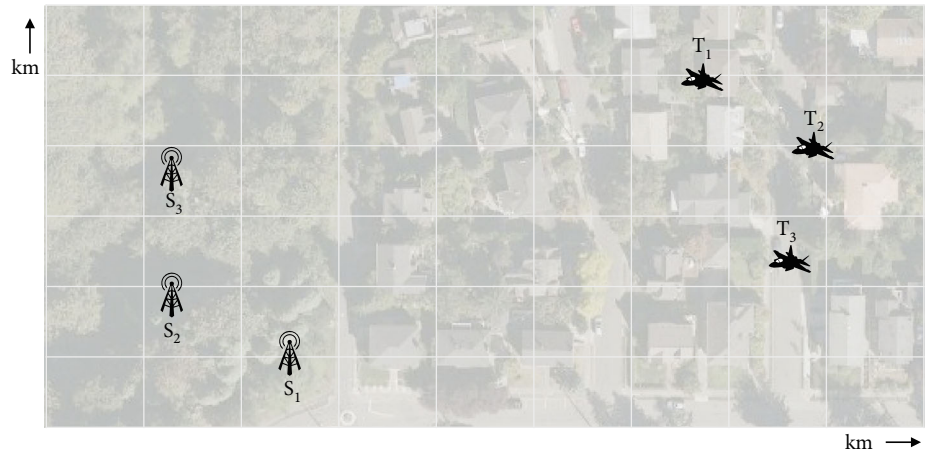



Figure 1: Multitarget scenario with three sensors and targets in a surveillance region.

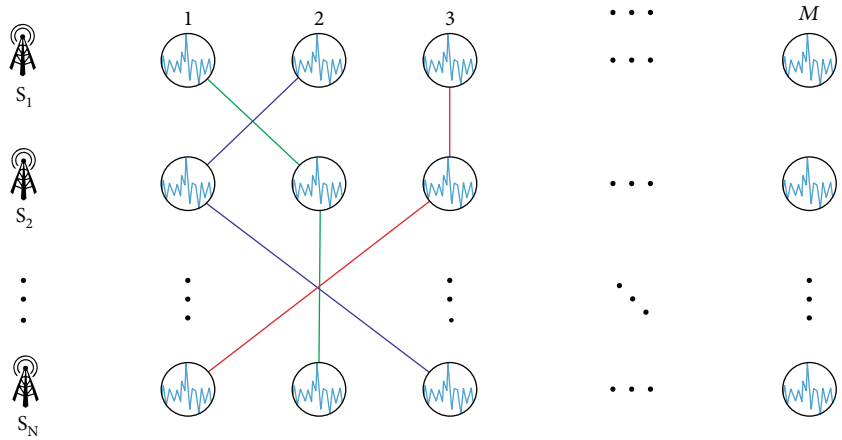



Figure 2: Graphical representation of $M$ measurements received at $N$ sensors and possible combinations.

of this work in a significant way. Consider the $N$ number of sensors $S_N$ positioned $(x_N, y_N)$ in the 2-dimensional surveillance region that received $M$ number of measurements from an unknown number of targets $U_M$ positioned $(x_M, y_M)$ in 2 dimensions.

*3.1.1. First Step.* This step involves the receiving of measurements and processing for the next step. In addition, each sensor receives $M$ measurements at a time, which is equal to $N \times M$ measurements, having $M^N$ possible combinations. Each possible combination is responsible for generating a candidate. Measurements can be written in vector form for each sensor.

$$S_N = \left[m_1^N, m_2^N, m_3^N, \cdots, m_M^N\right]. \tag{1}$$

Here, $m_M^N = \tilde{m}_M^N + n_M$, where $\tilde{m}_M^N = d_M - d_N$.

According to equation (1), $\tilde{m}_M^N$ is the $M^{\text{th}}$ range difference measurement received by sensor $N$. $d_N$ is the range of the target at sensor $N$. Assume that the received measurements contain the measurement noise which is considered to be independent zero-mean Gaussian random noise $n_M \sim N(0, \sigma^2)$. To deal with the measurement association to the targets, it is essential to create the combination of all the measurements received on all the sensors and obtain the candidates from each combination as visualized in Figure 2, which can be written as

$$C_k = \left[m_M^1, m_M^3, \cdots, m_M^N\right]_{1\times N}. \tag{2}$$

$C_k$ is the set of measurements that are associated with the $k^{\text{th}}$ candidate.

Considering all the combinations, the possible candidate can be given by

$$\text{Possible candidates} = k = M^N. \tag{3}$$

*3.1.2. Second Step.* Sensors are placed in a surveillance region; a grid of measurements is created for each sensor before scanning so that it can monitor the region for targets. The grid depends on the defined surveillance region. To reduce the computation, a 350 km × 350 km surveillance region is considered. This means each grid for a specific sensor is 350 km × 350 km. In an example considering 3 sensors,

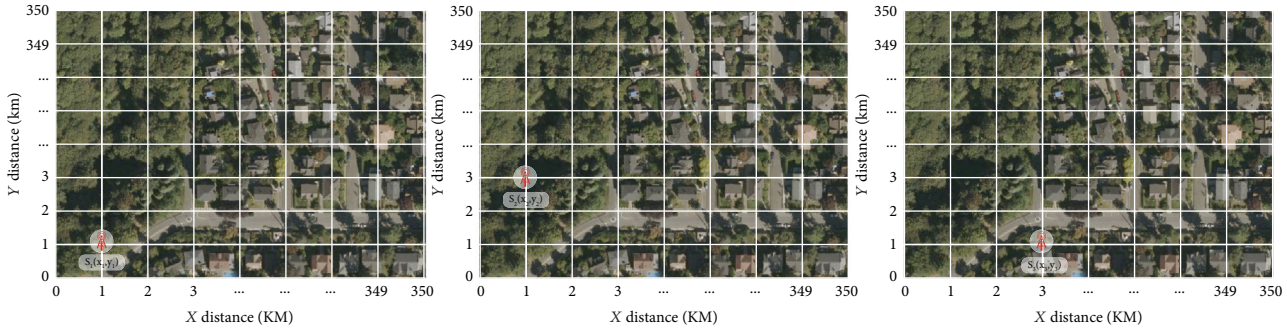



Figure 3: Graphical representation of measurement grids for three sensors.

each is having a grid of measurements between different points that are shown in Figure 3.

Each grid of the sensor can be written as

$$\mathrm{SG}_i = \begin{bmatrix} m^i_{0,0} & m^i_{1,0} & \cdots & m^i_{350,0} \\ m^i_{0,1} & m^i_{1,1} & \cdots & m^i_{350,1} \\ \vdots & \vdots & \ddots & \vdots \\ m^i_{0,350} & m^i_{1,350} & \cdots & m^i_{350,350} \end{bmatrix}, \quad (4)$$

where $\mathrm{SG}_i$ is a grid of measurements for $i^{\text{th}}$ sensor of the surveillance region and $m^i_{x,y}$ is the measurement at $i^{\text{th}}$ sensor received from the $(x, y)$ location in the region.

*3.1.3. Third Step.* It is considered that each sensor receives only one measurement maximum produced by each target at a particular time instant under consideration. For every measurement received at the sensor, the likelihood is measured with respect to the specific sensor's grid created in step 2.

The individual likelihood at sensor $i$ as a result of associating the $M^{\text{th}}$ measurement with location $X$. The likelihood function is given by

$$L_i(X_k) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}\left[C_k - \mathrm{SG}_i\right]'\left[C_k - \mathrm{SG}_i\right]\right\}, \quad (5)$$

where $L_i(X_k)$ is the likelihood of the $k$th measurement set at the $i^{\text{th}}$ sensor in (5). If the likelihood of the received measurement is more than the threshold, this likelihood result is shortlisted along with all the associated measurements of different sensors for each grid. To associate and localize the presence of the target, it is assumed that the maximum likelihood for all sensors is greater than the probability $\xi$, that is,

$$\mathrm{ML}(S_1)\&\mathrm{ML}(S_2)\&\cdots\&\mathrm{ML}(S_N) > \xi. \quad (6)$$

Here, the value $\xi$ is critical in precising the candidates to be shortlisted in the next steps; its value results in a narrow and wide likelihood curve. $\mathrm{ML}(S_N)$ is the maximum likelihood of $N^{\text{th}}$ sensor. Shortlisted measurements are forwarded to the next step.

Table 3: Proposed algorithm procedure.

| Step | Action |
|---|---|
| 1 | Obtain the possible candidates using the received measurements |
| 2 | Generate the grids for each sensor of the surveillance region |
| 3 | Shortlist the maximum-likelihood of the measurements |
| 4 | Using WLS to estimate the positions of the candidate and shortlisted measurements |
| 5 | Find the nearest measurement to candidates for potential targets |
| 6 | Remove duplicates, and average all estimates with the same tags |
| 7 | Repeat steps 3 to 6 for tracking for static sensors, otherwise repeat steps 2 to 6 |

*3.1.4. Fourth Step.* The shortlisted likelihood measurements are used to estimate the positions $X_{\text{est}}$ for each measurement using a weighted least square (WLS). Meanwhile, for each candidate measurement $C_k$, using the WLS candidate position $X_{\text{c}}$ is estimated, where $S_N$ is the positions of the sensors in 2-D $(x_N, y_N)$; $U_M$ is the position of the targets in 2-D $(x_M, y_M)$; $\tilde{m}^N_M$ is the $M^{\text{th}}$ measurement received by sensor $N$ without noise; $n_M$ is the zero-mean Gaussian random noise of $M^{\text{th}}$ measurement $n_M \sim N(0, \sigma^2)$; $C_k$ is the set of measurements belonging to the $k$th candidate; $k = M^N$, where $N$ is the number of sensors and $M$ is the number of targets; $SG_i$ is the grid of measurements of surveillance area at $i^{\text{th}}$ sensor; $m^i_{x,y} = \|u_{x,y} - S_i\|$; $m^i_{x,y}$ is the measurement value in between the $i^{\text{th}}$ sensor and $(x, y)$ location of the monitoring area; $u_{x,y}$ is the instantaneous value for the grid at $(x, y)$ in 2-D; $L_i(X_k)$ is the likelihood of the $k^{\text{th}}$ measurement at $i^{\text{th}}$ sensor; $\mathrm{ML}(S_i)$ is maximum likelihood of the $i^{\text{th}}$ sensor; $X_{\text{C}}$ is the position in $(x, y)$ of candidate measurements $C_k$ using WLS; and $X_{\text{est}}$ is the position in $(x, y)$ of shortlisted measurements by ML using WLS.

*3.1.5. Fifth Step.* The generated position of candidates $X_{\text{C}}$ and the shortlisted likelihood measurements $X_{\text{est}}$ in step 4 are compared in a way that, for each candidate, the root mean square (RMS) value is calculated to each of the shortlisted likelihood measurements and search for the minimum

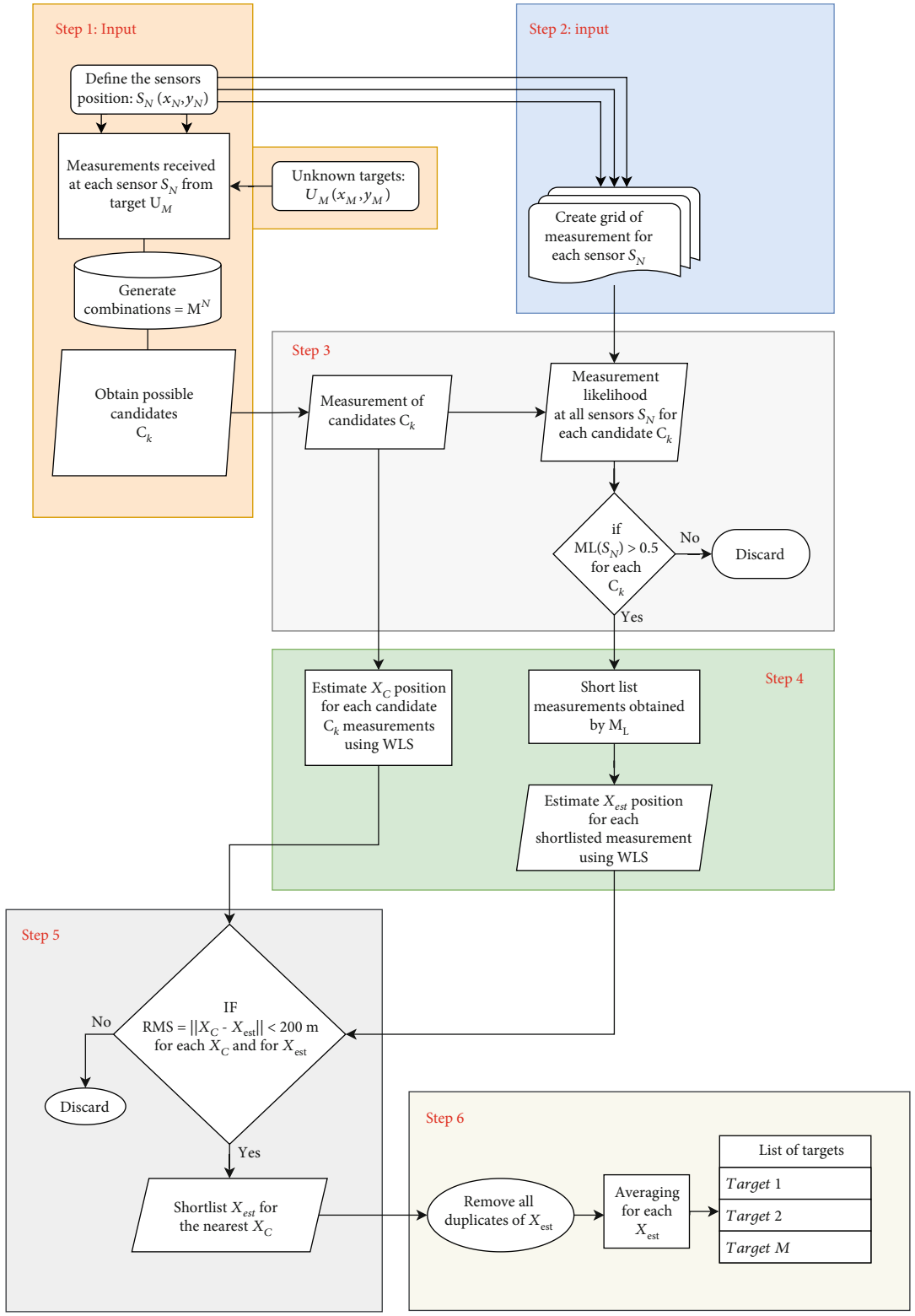



Figure 4: Flow diagram of proposed algorithm for multitarget localization and tracking.

value of RMS, if the minimum value is found as per the threshold $\beta$ set for accuracy.

$$\text{RMS} = \|X_C - X_{est}\| < \beta. \quad (7)$$

In equation (7), if for specific $X_C$ the above condition satisfies, then it is shortlisted and $X_{est}$ is associated with the $X_C$ candidate by assigning a tag number, which helps to count the number and increment the index number of $X_C$ and forwarded to step 6.

*3.1.6. Sixth Step.* Shortlisted results can have duplicate measurements that produce the same result. These duplicate measurements are removed, which improves the computation and takes an average of all the shortlisted results associated with $X_C$, and the process is repeated according to the number of targets localized.

$$T_j = \frac{1}{L}\sum_{i=1}^{L} X^i_{\text{est}_j}. \tag{8}$$

$T_j$ is the $j^{\text{th}}$ target localized, $L$ is the count of shortlisted associated measurements, and $X^i_{\text{est}_j}$ is the shortlisted measurements associated with a $j^{\text{th}}$ target.

*3.2. Algorithm Procedure.* The algorithm steps are explained in the last section, and its procedure is explained in Table 3. The proposed algorithm reduces the computation to localize and track the multiple targets and also reduces the complexity as compared to the existing work whereas the result produced using the proposed algorithm is effective and efficient as shown in the simulation section.

## 4. Simulations

The flowchart of the proposed localization and tracking scenario with details is described in Figure 4. To better understand the algorithm, simulation is performed in MATLAB to explain the methodology and compare it with existing methods. For simulation, three sensors and six targets were considered in 2-dimensional space. We consider two targets and repeat this step for three to six targets. The surveillance region is defined to be 350 km by 350 km considering far-field targets. Tables 4 and 5 show the positions and velocities of the sensors and targets, respectively, where the sensors are considered to be static, and both targets are moving with constant speed.

The three sensors and two targets can generate a maximum of eight candidates that can be the potential targets. Using equation (3), it is explained in Figure 5 with the green points:

$$k = M^N = 2^3 = 8, \tag{9}$$

as $k$ is the number of possible candidates. So, at this point, eight measurements are received by all the sensors and grids of the surveillance region are created and the further likelihood of the measurement is calculated.

The computation increases as the value increases which depends on the number of targets and sensors. Figure 6 shows the curves for all the individual likelihood without any condition of threshold comparison. In Figure 6, the blue box represents the reference sensor. The past algorithm used the grid with different dimensions of the parameter, whereas our algorithm can localize using only a single dimension of parameter measurement. A multidimensional parameter grid can produce more accurate results as it has more information to process, next. Figure 7 shows the result of the localized target areas in the surveillance region. The localized areas can have an additional area other than the actual target due to the measurement noise in the received measurements.

Table 4: Multitarget algorithm—sensor positions and velocity.

| Sensors | Position | | Velocity | |
|---|---|---|---|---|
| | $X$ (km) | $Y$ (km) | $\dot{X}$ (m/s) | $\dot{Y}$ (m/s) |
| Sensor 1 | 75 | 200 | 0 | 0 |
| Sensor 2 | 50 | 25 | 0 | 0 |
| Sensor 3 | 112 | 100 | 0 | 0 |

Table 5: Multitarget algorithm—target positions and velocity.

| Target | Position | | Velocity | |
|---|---|---|---|---|
| | $X$ (km) | $Y$ (km) | $\dot{X}$ (m/s) | $\dot{Y}$ (m/s) |
| Target 1 | 300 | 190 | -20 | -15 |
| Target 2 | 200 | 300 | -20 | -15 |
| Target 3 | 250 | 150 | -20 | -15 |
| Target 4 | 300 | 300 | -20 | -15 |
| Target 5 | 300 | 50 | -20 | -15 |
| Target 6 | 250 | 250 | -20 | -15 |

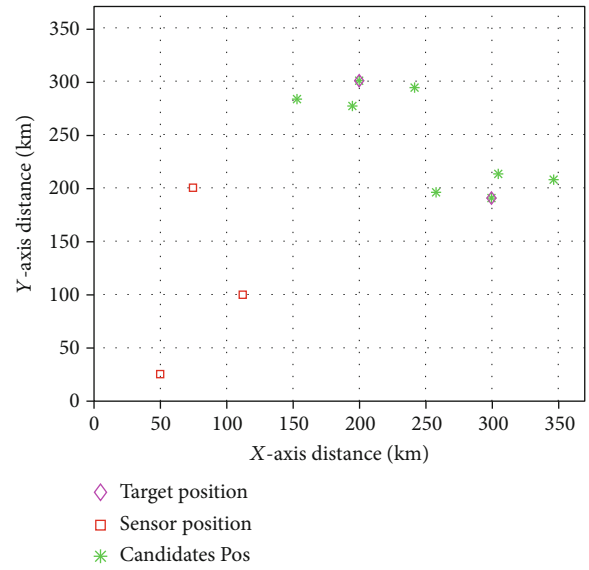



Figure 5: Candidate positions for two targets.

Figure 8 shows the results of the maximum likelihood of the measurements in red, these are the shortlisted measurements, and the rest of the measurements which do not satisfy the condition are discarded. Furthermore, the accuracy of the localization for the unknown targets depends on the thresholds, which are adjusted by performing multiple runs and compare the results to get more accuracy. $\beta$ is set to 200 meters, and $\xi$ is equal to 0.5. Both can be varied depending on the results and monitoring environment.

Figure 9 shows the localization of two targets, and Figure 10 is the zoomed plots, in which the algorithm is

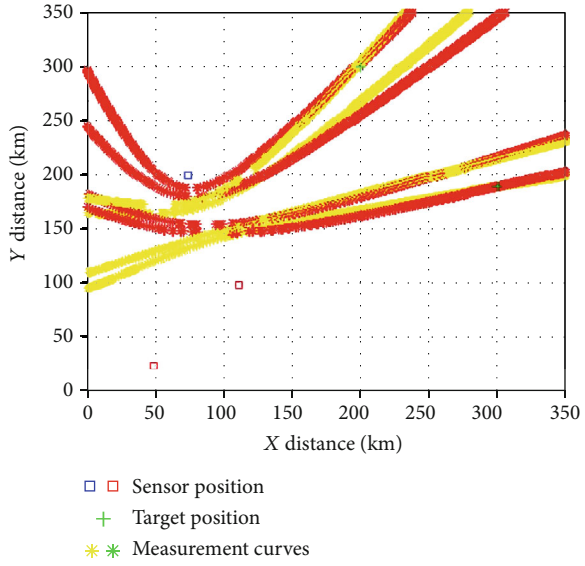



Figure 6: Likelihood curves of received measurements.

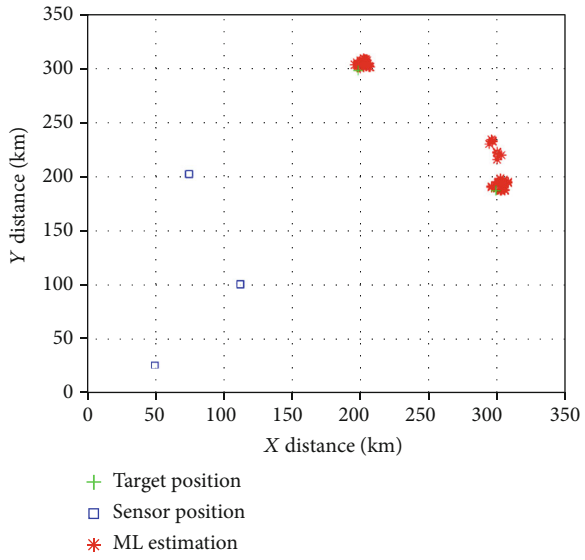



Figure 7: Shortlisted likelihood measurements (red).

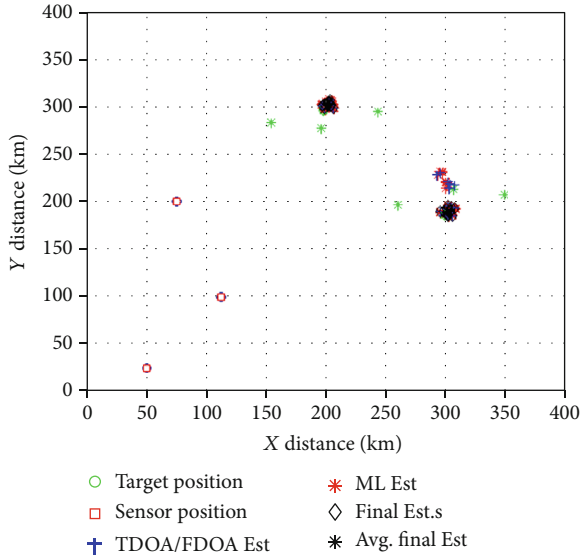



Figure 8: Localization of two targets using TDOA/FDOA.

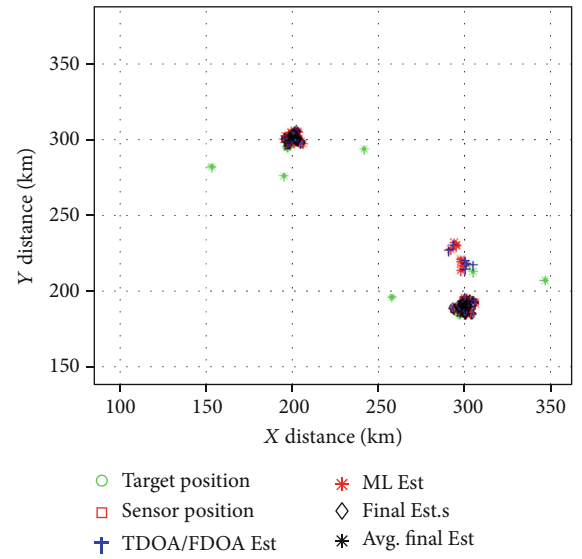



Figure 9: Two localized targets (dense black) with candidates (green).

not only able to localize the targets but also can rectify the ghost created in the maximum likelihood process.

If the sensors are not moving, the measurement grid will not be regenerated. On the other hand, if sensors are in motion, the sensor's measurement grid has to regenerate continuously to estimate the high position accuracy of the targets. The increase in the received measurements indicates an increased number of targets, which causes an increase in the possible candidates also. Figure 11 shows the localization of three targets with a different number of possible candidates in each plot, and Figure 12 shows the localization of five targets. Once the locations are localized, the proposed algorithm can be used in a simulation to track the targets which are successfully tested using two targets.

If the sensors are not moving, the measurement grid will not be regenerated. On the other hand, if sensors are in motion, the sensor's measurement grid has to regenerate continuously to estimate the high position accuracy of the targets. The increase in the received measurements indicates an increased number of targets, which causes an increase in the possible candidates also. Figure 11 shows the localization of three targets with a different number of possible candidates in each plot, and Figure 12 shows the localization of five targets. Once the locations are localized, the proposed algorithm can be used in a simulation to track the targets which are successfully tested using two targets.

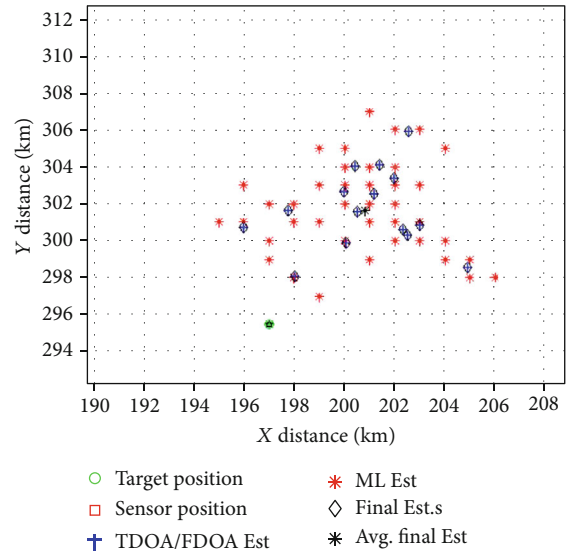



Figure 10: Zoomed at one target estimates (black asterisk).

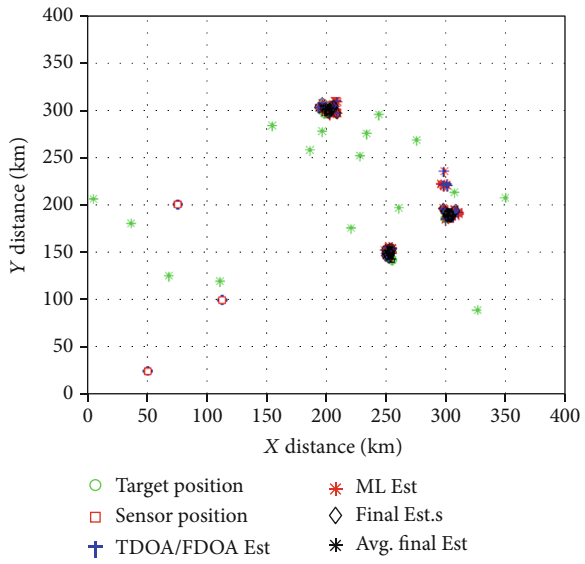



Figure 11: Localization using the algorithm for three targets.

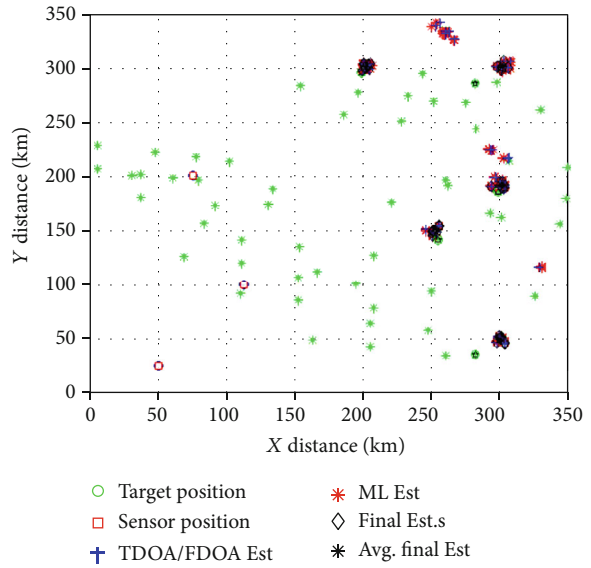



Figure 12: Localization using the algorithm for five targets.

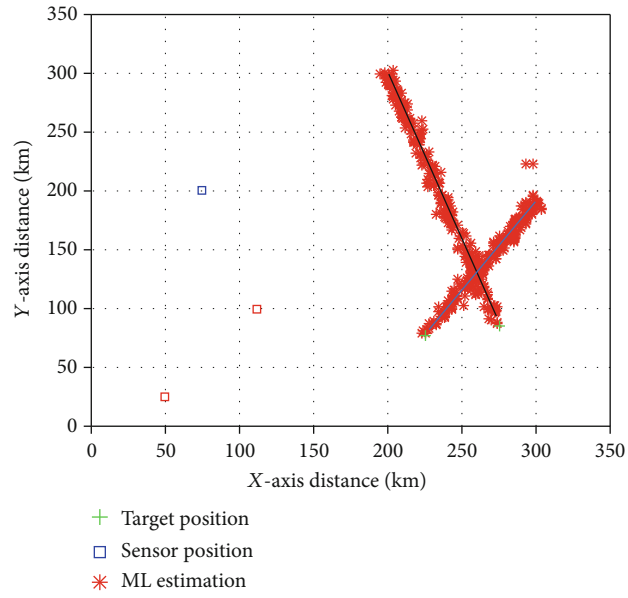



Figure 13: Tracking of two moving targets intersecting at a point.

Once the algorithm runs and localizes the targets, the procedure is repeated for newly received measurements on different sensors and localize the targets again and keep track of old and new shortlisted measurements which provides the complete track of targets. Figures 13 and 14 show the results in which their updated positions are tracked using maximum likelihood, where Figure 13 shows the scenario of intersecting two targets and Figure 14 shows the simple scenario of two targets moving in the straight path.

The proposed algorithm does not have any window size, as it gives a different approach to solve the multitarget localization problem. To compare the proposed algorithm's performance to other algorithms, i.e., multipass grid search and directed subspace search, we consider the results from [26] and calculate the mean runtime of the simulation of the single estimate for window size $N_w$ from 5 to 10. The results show in Figure 15 [26] that the proposed algorithm (red curve) performed better than multipass grid search overall but for directed subspace search, the proposed algorithm results better after $N_w = 8$. Directed subspace is ineffective due to a search in a specific area. Further, in Table 6, the simulation shows that the runtime of the proposed algorithm increased very slightly and gradually.

In a single scan, this algorithm as compared to other existing work can localize and track the unknown number

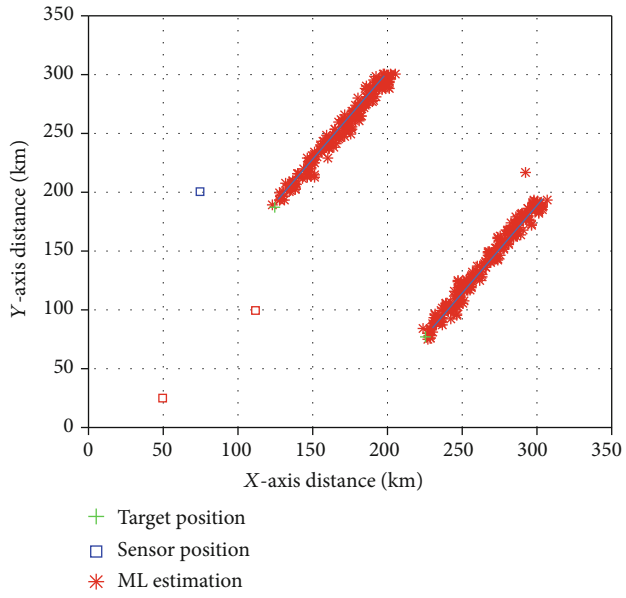



Figure 14: Tracking of two moving targets in parallel.

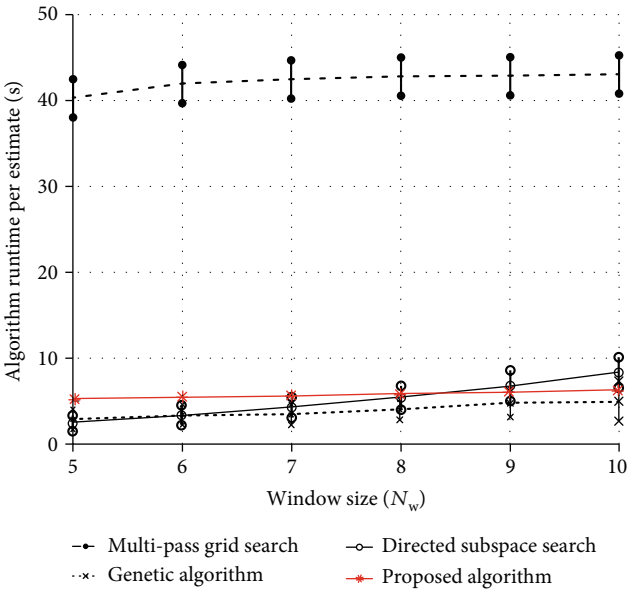



Figure 15: Single estimate mean runtime with one standard deviation in 2-D measurement space.

Table 6: Runtime for different window sizes of the proposed algorithm.

| Window size $N_w$ | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|
| Algorithm runtime (s) | 5.23 | 5.47 | 5.6 | 5.81 | 6.06 | 6.26 |

of targets without repeated steps for convergence or biological evolution. Hence, the proposed algorithm is effective and simple to implement as compared to the existing work which reduces the computation.

## 5. Concluding Remarks

This research proposed an algorithm based on TDOA/FDOA and optimization of the shortlisted measurement. This proposed technique provides localization and tracking of multitargets using the maximum likelihood of the measurement and further addresses the data association problem of the received measurements and targets.

This algorithm can localize the unknown number of targets having the complexity of multiple sensors with multiple grids. The proposed algorithm reduces complexity as it does not use repeated steps for convergence and any biological evolutions. A single scan computes the likelihood for all sensors using only one parameter grid and passes the results for multitarget estimation.

It also reduces the localization of the ghost target by eliminating the data fusion issue. Therefore, this method will be the best-suited candidate when considering the combinations of the measurements from different sensors. In the future, an improvement can be made in the algorithm by considering multiparameter approach and including the maximum likelihood of those measurements to increase the position accuracy. Also, this algorithm can work in meters for considerably distant fields. Other optimization models can be combined to authenticate the algorithm's efficiency and effectiveness.

## Data Availability

All the data used to support the findings of this study are available in this paper.

## Conflicts of Interest

The authors declare that there is no conflict of interest.

## Acknowledgments

The authors would like to thank King Khalid University of Saudi Arabia for supporting this research under the grant number R.G.P. 2/177/42.

## References

[1] K. W. Cheung, H.-C. So, W.-K. Ma, and Y.-T. Chan, "Least squares algorithms for time-of-arrival-based mobile location," *IEEE Transactions on Signal Processing*, vol. 52, no. 4, pp. 1121–1128, 2004.

[2] C. Berry, D. J. Bucci, and S. W. Schmidt, "Passive multi-target tracking using the adaptive birth intensity PHD filter," in *2018 21st International Conference on Information Fusion (FUSION)*, pp. 353–360, Cambridge, UK, 2018.

[3] R. B. Langley, "Dilution of precision," *GPS World*, vol. 10, no. 5, pp. 52–59, 1999.

[4] M. Ahmad and I. U. Haq, "Linear unmixing and target detection of hyperspectral imagery using OSP," *proc. of IPCSIT*, vol. 10, pp. 179–183, 2011.

[5] W. H. Foy, "Position-location solutions by Taylor-series estimation," *IEEE Transactions Aerospace and Electronic Systems*, vol. 12, no. 2, pp. 187–194, 1976.

[6] Y.-T. Chan and K. C. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Transactions on Signal Processing*, vol. 42, no. 8, pp. 1905–1915, 1994.

[7] K. C. Ho and W. Xu, "An accurate algebraic solution for moving source location using TDOA and FDOA measurements," *IEEE Transactions on Signal Processing*, vol. 52, no. 9, pp. 2453–2463, 2004.

[8] H.-C. Shin, "Weighted least squares estimation with sampling weights," *JSM*, pp. 1523–1530, 2013.

[9] K. Yang, L. Jiang, and Z.-Q. Luo, "Efficient semidefinite relaxation for robust geolocation of unknown emitter by a satellite cluster using TDOA and FDOA measurements," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2584–2587, Prague, Czech Republic, 2011.

[10] E. Čišija, A. M. Ahmed, A. Sezgin, and H. Wymeersch, "Ris-aided mmWave MIMO radar system for adaptive multi-target localization," in *2021 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 196–200, Rio de Janeiro, Brazil, 2021.

[11] Y. Ma, Y. Zhang, B. Wang, and W. Ning, "SCLA-RTI: a novel device-free multi-target localization method based on link analysis in passive UHF RFID environment," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3879–3887, 2020.

[12] M. U. Aftab, Y. Munir, A. Oluwasanmi et al., "A hybrid access control model with dynamic COI for secure localization of satellite and IoT-based vehicles," *IEEE Access*, vol. 8, pp. 24196–24208, 2020.

[13] J. Zakria, J. Cai, M. Deng, S. Khokhar, and M. U. Aftab, "Vehicle classification based on deep convolutional neural networks model for traffic surveillance systems," in *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 224–227, Chengdu, China, 2019.

[14] Zakria, J. Cai, J. Deng, M. Aftab, M. Khokhar, and R. Kumar, "Efficient and deep vehicle re-identification using multi-level feature extraction," *Applied Sciences*, vol. 9, no. 7, p. 1291, 2019.

[15] Y. Lee, T. S. Wada, and B.-H. Juang, "Multiple acoustic source localization based on multiple hypotheses testing using particle approach," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2722–2725, Dallas, TX, USA, 2010.

[16] X. Dang, H. Zhu, and Q. Cheng, "Multiple Sound Source Localization Based on a Multi-Dimensional Assignment Model," in *2018 21st International Conference on Information Fusion (FUSION)*, pp. 1732–1737, Cambridge, UK, July 2018.

[17] T. Kirubarajan and Y. Bar-Shalom, "Low observable target motion analysis using amplitude information," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 4, pp. 1367–1384, 1996.

[18] D. E. Goldberg, *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley, Boston, MA, 1989.

[19] W. R. Blanding, P. K. Willett, Y. Bar-Shalom, and R. Lynch, "Directed subspace search ML-PDA with application to active sonar tracking," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, no. 1, pp. 201–216, 2008.

[20] R. Cerf, "Asymptotic convergence of genetic algorithms," *Advances in Applied Probability*, vol. 30, no. 2, pp. 521–550, 1998.

[21] W. H. Press, B. P. Flannery, and S. A. Teukolsky, *T VW: Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, New York, 2002.

[22] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Systems Magazine*, vol. 29, no. 6, pp. 82–100, 2009.

[23] J. Chen and P. Dames, "Collision-free distributed multi-target tracking using teams of mobile robots with localization uncertainty," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6968–6974, Las Vegas, NV, USA, 2020.

[24] A. Oluwasanmi, M. U. Aftab, E. Alabdulkreem, B. Kumeda, E. Y. Baagyere, and Z. Qin, "CaptionNet: automatic end-to-end Siamese difference captioning model with attention," *IEEE Access*, vol. 7, pp. 106773–106783, 2019.

[25] A. Oluwasanmi, E. Frimpong, M. U. Aftab, E. Y. Baagyere, Z. Qin, and K. Ullah, "Fully convolutional CaptionNet: Siamese difference captioning attention model," *IEEE Access*, vol. 7, pp. 175929–175939, 2019.

[26] H. Qin, W. Chen, W. Chen, N. Li, M. Zeng, and Y. Peng, "A collision-aware mobile tag reading algorithm for RFID-based vehicle localization," *Computer Networks*, vol. 199, article 108422, 2021.

[27] F. Meyer, A. Tesei, and M. Z. Win, "Localization of multiple sources using time-difference of arrival measurements," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3151–3155, New Orleans, LA, USA, 2017.

[28] S. C. K. Herath, P. N. Pathirana, and N. L. de Boer, "Localization with ghost elimination of emitters via time-of-arrival measurements," in *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 2231–2235, Guangzhou, China, 2012.

[29] Y. M. Chen, C.-L. Tsai, and R.-W. Fang, "TDOA/FDOA mobile target localization and tracking with adaptive extended Kalman filter," in *2017 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*, pp. 202–206, Prague, Czech Republic, 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5291528, 10 pages
https://doi.org/10.1155/2021/5291528

*Research Article*

# An Optimized Machine Learning and Big Data Approach to Crime Detection

**Ashokkumar Palanivinayagam**,[1] **Siva Shankar Gopal**,[1] **Sweta Bhattacharya**,[2] **Noble Anumbe**,[3] **Ebuka Ibeke**,[4] **and Cresantus Biamba**[5]

[1]*Sri Ramachandra Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Tamil Nadu, India*
[2]*School of Information Technology and Engineering, VIT, Tamil Nadu, India*
[3]*Department of Mechanical Engineering, University of South Carolina, Columbia, SC, USA*
[4]*School of Creative & Cultural Business, Robert Gordon University, Aberdeen, UK*
[5]*Faculty of Education and Business Studies, University of Gavle, Sweden*

Correspondence should be addressed to Cresantus Biamba; cresantus.biamba@hig.se

Received 23 June 2021; Accepted 10 October 2021; Published 13 November 2021

Academic Editor: Vishal Sharma



Crime detection is one of the most important research applications in machine learning. Identifying and reducing crime rates is crucial to developing a healthy society. Big Data techniques are applied to collect and analyse data: determine the required features and prime attributes that cause the emergence of crime hotspots. The traditional crime detection and machine learning-based algorithms lack the ability to generate key prime attributes from the crime dataset, hence most often fail to predict crime patterns successfully. This paper is aimed at extracting the prime attributes such as time zones, crime probability, and crime hotspots and performing vulnerability analysis to increase the accuracy of the subject machine learning algorithm. We implemented our proposed methodology using two standard datasets. Results show that the proposed feature generation method increased the performance of machine learning models. The highest accuracy of 97.5% was obtained when the proposed methodology was applied to the Naïve Bayes algorithm while analysing the San Francisco dataset.

## 1. Introduction

In the last few decades, there has been an exceptional growth in urban population which has led to the demand for a secured, hospitable, and sustainable society. With the ever-expanding growth of city, engulfing suburbs and rural spaces, the management of urbanization remains a major challenge for administrative authorities. Cities are getting overpopulated, compelling governments to undertake smart city initiatives that would help achieve better management of infrastructure and overcome the major challenges of security, sustainability, and development. Although smart city initiatives have gained immense momentum with promises to enhance quality of life, it does have its own challenging aspects as well. One of the major challenges in smart city life is public safety. Various studies have been conducted to help understand crime patterns and its relationship to the social economic development of particular regions, the human characteristics, their level of education, and family bonding [1].

Crime investigating organizations have identified various types of crimes. The four main categories include killing, molestation, looting, and intensive attacks. Killing or murder refers to the willful assassination of a person by another. Molestation means the sexual abuse of a woman, man, or child against their wish. This crime is as heinous as rape, having significant consequences. Looting refers to the act of stealing goods from a human domain, using excessive physical force or violence. Finally, intensive attacks refer to illegal confrontation by one person against another to achieve something or to

simply harm the individual [2]. Crime detection is a necessity in urban life, and machine learning is a popular crime detection and prevention technique. Several organizations across the globe have been experimenting with these techniques.

It has been observed that crimes are often predictable, and it just requires the processing of high volumes of data that would reveal interesting patterns suitable for law enforcement. In many of the instances, crimes conducted often remain unreported due to external pressures from all verticals of the society. Intelligent systems can promptly detect crimes and help eradicate such manipulative activities by bypassing individuals and automatically informing relevant authorities. As an example, the research by Borges et al. [1] discussed the case study of San-Francisco, USA, and Natal in Brazil where criminal activities were prevalent. The various attributes of urbanization in these two cities were analysed, and then, machine learning models were implemented to detect criminal activity hotspots. As per [2], they created a regression model to predict crime rates in various Indian states. Supervised and unsupervised learning techniques were also deployed to achieve enhanced accuracy in crime prediction. In [3], fuzzy $C$-means algorithm was used for the clustering of crime data for various cognizable crimes, namely, kidnapping, murder, theft, robbery, and crimes against women. Similarly, $K$ nearest neighbour methods have been deployed for the observation of crime rates which have helped to understand crime types and time/place of occurrence.

Considering the various studies conducted, it is observed that most of the existing works emphasize the use of crime history and population density for the crime prediction. The present work presents four attribute generation methods for the detection of crimes. The dataset holds various crime locations in an area where $K$-means clustering is applied, yielding crime hotspots. Then, a crime ratio matrix is constructed leading to the prediction of crime probability when subjected to a machine learning model. As part of the proposed methodology, crime monitoring is performed with the help of the following methods:

(i) Crime transition probability computes the connection of one crime to another

(ii) Vulnerability of an area indicates how safe an area is

Many existing works use artificial intelligence and machine learning to extract crime patterns and to detect and prevent crime incidents. Most of the existing works have few limitations which include incompetence of finding links between different crime incidents and vulnerability analysis. In this paper, we propose four unique stages of crime detection which uses the combination of locations, vulnerability, correlation, and temporal patterns.

The unique contributions of the proposed method are highlighted below:

(i) Ability to analyse the relationships between time zones, namely, morning, evening, and night for each type of crime

(ii) Prediction of crime probability for the following day considering the present-day crime history

(iii) Generation of crime hotspots in the form of geolocations indicating occurrence of a greater number of crimes

(iv) Performing vulnerability analysis to identify locations more prone to criminal activities in the future

This paper contains five sections: Section 2 discusses previous studies. Section 3 describes the four-feature generation process used in the proposed work. The results of our work are discussed in Section 4. Finally, Section 5 contains the conclusion and future work.

## 2. Related Works

Various studies have been conducted that are relevant to crime detection, analysis of the various factors that contribute significantly towards crime occurrence and its impact on the socio-economic status of various regions. Machine learning approaches have been a predominant and popular area of research interest in the crime detection domain. This section summarizes some of the interesting studies conducted in crime detection, analysis, and prediction. The overview will help highlight research gaps or limitations in this field.

A research proposed by [4] implemented deep learning approaches on CCTV camera images to detect crimes, eliminating traditional (manual) monitoring systems that rely on human supervision. In the traditional system, the CCTV cameras are installed at various positions in the public and private surroundings which capture videos and images with the prime objective of monitoring and preventing incidences of crime. However, the detection of crimes does not happen automatically as it requires human supervision and constant monitoring of CCTV screens. This physical monitoring system is often prone to errors due to the chance of missing important incidents since effectively monitoring multiple screens at the same time is often difficult. To overcome the challenge, [4] developed a pretrained deep learning model VGGNet19 that detected criminal events in real time and generated an alert for the human supervisor to ensure immediate action is taken. The results were evaluated against GoogleNet and InceptionV3, with the VGGNet19 model yielding higher training accuracy. However, the model detects criminal intentions but does not provide any insight on crime hotspots nor does it highlight the probabilities of crime occurrence.

One more research [5] proposes a visual surveillance system that would detect hostile intent and behaviour inside the elevators. The surveillance camera that captures images of the small, confined elevator space based on the illumination of the opening and closing of the elevator doors was used for this study. The implementation involved a three-layered approach for the detection of violent events. The low-level feature-segmented foreground blobs from the background and their motion velocities were captured using an optical flow method. In the second or midlevel feature, the velocity and directions were computed to analyse motions of the

images captured. Sequences of image frames having more than one person in the elevator were analysed, and whenever an average velocity magnitude exceeded a threshold value, a violent event occurrence was assumed to have been detected.

The methodology proposed by [6] is aimed at predicting crime without human intervention using computer vision and machine learning approaches. The paper implements rectified linear unit (ReLU) and convolutional neural networks (CNN) for the detection of weapons such as knives or guns from a particular image. This helped to validate the occurrence of a crime and identify the location of occurrence as well. The accuracy of the results seemed quite promising, which achieved almost 92% accuracy for a testing dataset.

Another interesting research [7] discusses the excessive surge in document forging incidents using powerful photo editing software used as a tool for creating fake documents. Such fake documents are scanned and forgotten in minutes with the help of automated editing tools used exclusively for the said purpose. The study involves the use of a GUI which is designed to detect if an image is manipulated or not. The GUI helps to load and preprocess the image, enhancing its global contrast. The image is then partitioned into three segments using the $K$-means clustering approach. The segment containing most of the information is further analysed, extracting its features. These are compared with the scanned images in the database to identify the occurrence of tampering. Support vector machine (SVM) and ANN were implemented, but SVM yielded better accuracy and thus was considered the most suitable.

A ML model [8] proposed a fraud detection system using a hybrid machine learning approach emphasizing on electronic transactions. It has been observed that most economic frauds involve business transactions relating to credit cards. The paper uses feature engineering approach on the dataset and then SVM and random forest implementation as a hybrid technique to detect fraudulent transactions.

One more research work by [9] developed a machine learning-based approach for the detection of spam images. In the present day and age, email is one of the vital modes of communication almost among all stakeholders in the society. Email not only acts as digital letters but also enable the attachment of documents, pictures, videos, and music to be sent to recipients. There are certain miscreants who send unsolicited emails to users to weaken the internet traffic. The spammers also sent such emails to users attracting them to buy products which are prohibited. The study involved using chi-square test for feature engineering and sequential minimal optimization (SMO) algorithm. Post feature selection method, multilayer perceptron (MLP) algorithm is used for the detection of spams. Both SMO and MLP yielded an $F$-score of 98.5% and 98.4%, respectively.

The work done by [10] developed a machine learning model that would help to predict potential crimes in a geographic location, analysing the existing crime and repeating incident occurrence datasets. The paper used the Chicago Police Department CLEAR dataset and selected 9 features from the dataset for further analysis. Finally, Naïve Bayes- and decision tree-based approaches were used to predict potential crimes. This was intended to help create contingency plans and keep the society safe, promoting hospitable and secured living. The results highlighted the superiority of the decision tree-based approach considering 7, 8, and 9 features for the matrices: correctly classified instances (CCI), accuracy (AC), ROC, precision, and recall, respectively.

The study in [11] focused on comparing two images by identifying the query image from the source image, which would help in the recognition of a particular person or object in the image. The frames that matched were generated as an output after implementation of the scale invariant feature transformation (SIFT) method. SIFT was used to extract features that were invariant to image scaling, rotation, presence of noise, or all changes in the image lighting. Once the feature points in an image were identified, they were compared with the feature points in the frame implementing homographic estimation. The Euclidian distance formula was used for the comparison.

The work by [12] targeted the occurrences of road transport crimes and identified methods to reduce them. Road transport is often used by criminals for escaping after conducting heinous crimes. Moreover, a lot of crimes remain unregistered and unresolved due to lack of evidence on the roads. To eliminate such occurrences, a machine learning algorithm was deployed in the study using text and facial recognition techniques. The system extracts characters from the vehicle number plates using a text recognition mechanism. On the other hand, the facial recognition algorithm helps in the identification of the face of the suspects. The extracted feature is mapped to the relevant features of the images saved in the database, and in case of mismatch, an alert is generated. In the same way, the facial images are compared with criminal face images available in the database, and in case of anomaly, an alert is generated. KNN and SVM in association with face detection classifier were used to achieve the proposed objective [13].

In [14], news is analysed using machine learning algorithms and provides a report on the classified crime news. The traditional system involves reading the complete news and manually analysing the same which is prone to errors. Moreover, the approach is quite time consuming. To overcome this challenge, a machine learning-based classification approach is implemented involving the use of three classifiers. The result segregates crime-related data and noncrime-related data. The website or newspaper contents are fed into the system, a crawling program is implemented written in Python, and the data is finally stored in a temporary database. The result generated display crime and noncrime data presented in a tabular format to the user. Table 1 shows the summary of related works performed.

Another research work [15] concentrates on crime hotspot detection. They have used data from 2 million crime data between 2006 and 2018 to train GAN model. Their research work proposes a new city plan based on the crime distribution. The simulated new city plan seems to have much lower crime rate than the original city.

The crime data is imbalanced most of the time. [16] uses data argumentation and loss function to develop samples

Table 1: Summary of related works.

| References | Dataset | Methods used | Evaluation metrics | Limitations |
|---|---|---|---|---|
| Navalgund and K. (2018) [4] | YouTube and Google | VGGNet -19 | Accuracy, recall, F1-score and support | Detection of crime hotspots and probability of occurrences not included. |
| Younghyun Lee et al. (2011) [5] | Real-time elevator data collected using surveillance camera of 320 ∗ 240 pixels | Violent frame detector, motion vector extraction, and foreground segmentation | Detection rate, no. of people in the elevator, false-positive rates (FPR) | Includes only detection but not prediction or probabilities of occurrence results |
| Nakib et al. (2018) [6] | Real-time data | Softmax regression model, CNN | Accuracy | The size of the dataset was relatively small. The model was not evaluated against the other classical models. |
| Ranjan et al. (2018) [7] | Image collected from various internet sources and then morphed to test the methods | SVM and ANN | Accuracy, sensitivity and specificity | Comparison of the results with other traditional approaches were not included. Availability of larger dataset also is a challenge |
| Vynokurova et al. (2020) [8] | Real-time dataset | SVM and random forest-based hybrid approach | Accuracy | Comparison of the results with other traditional approaches were not included. Availability of larger dataset also is a challenge |

and improve the minority class. They have used neural network to enhance the crime detection problem.

## 3. Preparing the Model

In this section, we present the working of the proposed model and the four attribute generation methods such as fraction of day, crime growth factor, distance from crime hotspot, and vulnerability analysis. The overall flow of the proposed method is shown in Figure 1.

*3.1. Fraction of the Day.* Crimes are more likely to occur at certain times of the day, for example, more crimes occur between 6 p.m. and 12 a.m. (next day) than between 6 a.m. and 12 p.m. Hence, to increase the prediction success rate, it will be better to consider a fraction of the day instead of the day as a whole [17, 18].

Consider 100 crimes that happened on day $X$. Since most of the crimes are more likely to occur at night, in the proposed model, we consider the impact of different fractions of the day instead of the whole day. In this case, we divide a single day into four fractions such as

(i) Fraction 1: between 00:00 AM and 06:00 AM

(ii) Fraction 2: between 06:01 AM and 12:00 PM

(iii) Fraction 3: between 12:01 PM and 06:00 PM

(iv) Fraction 4: between 06:01 PM and 11:59 PM

For each crime, the number of crime events is noted and stored in crime counter (CC) as per Figure 2.

$C_1, C_2, , \cdots C_n$ are different crimes and $F_1$, $F_2$, $F_3$, and $F_4$ are the four fractions, respectively. $N_{ci,Fj}$ represents the number of crimes $i$ that occurred at fraction $j$. The time fractions can be made dynamic; however, dividing a day into four fractions makes the segregation of crimes simpler and more meaningful.

*3.2. Crime Growth Vector.* The most important aspect of crime forecasting system is detecting the probability of crime each day [19, 20]. The probability of crime $i$ can be found by calculating the percentage of the number of crime $i$ events in the total number of all crimes. The crime vector CV stores the probability of all crimes. Equation (1) shows the structure of the CV. Each value in the vector is calculated by Equation (2)

$$\text{Crime Vector}(\text{CV}) = (P_{c1}, P_{c2}, \cdots, P_c), \tag{1}$$

$$P_{C_i} = \frac{\text{Number of crimes } i}{\text{Total number of crimes}}. \tag{2}$$

Transition probability matrix (TPM) is one of the methods which can help to forecast the probabilities of future days. TPM needs a vector (to denote the initial probability) and a matrix (to represent the Markov chain). In this context, we use the crime vector as the initial probability matrix. The crime growth factor can be used as Markov chains. A crime growth factor between two crimes A and B is how much likely a crime B is to happen on day $d + 1$ when crime A has happened on day $d$. Equations (3) and (4) can be used to calculate the likelihood of two crimes happening on day $d$ and day $d - 1$. The values are normalized so that

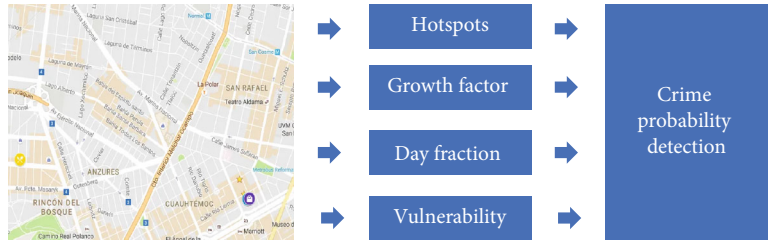



Figure 1: The working of the proposed model.

$$\mathrm{CC} = \begin{bmatrix} N_{C_1F_1} & N_{C_1F_2} & \cdots & N_{C_1F_m} \\ N_{C_2F_1} & N_{C_2F_2} & \cdots & N_{C_2F_m} \\ \cdots & \cdots & \ddots & \cdots \\ N_{C_nF_1} & N_{C_nF_2} & \cdots & N_{C_nF_m} \end{bmatrix}$$

Figure 2: The crime counter.

the factors are turned into probability values. The crime growth factor is calculated for each day fraction separately and finally merged into a single matrix as per Equation (5).

$$GF_{i,j}^{d} = \frac{g_{ij}^{d}}{\sum_{i=1}^{n} g_{ik}^{d}}, \tag{3}$$

$$g_{i,j}^{d} = \frac{\text{Number of crime } j \text{ on day } d}{\text{Number of crimes } i \text{ on day } d * 1}, \tag{4}$$

$$\text{Final Value}\left(FV_{i,j}\right) = \sqrt[n]{\prod_{k=1}^{n} g_{i,j}^{k}}, \tag{5}$$

Next Day Crime Probability Vector

$$=: \left[P_{C_1}, P_{C_2}, \cdots, P_{C_n}\right] \begin{bmatrix} FV_{1,1} & FV_{1,2} & \cdots & FV_{1,n} \\ FV_{2,1} & FV_{2,2} & \cdots & FV_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ FV_{n,1} & FV_{n,2} & \cdots & FV_{n,n} \end{bmatrix}. \tag{6}$$

Using this TPM, the next day probability can be easily calculated by multiplying the CV and the final value matrix. The calculation is mentioned in Equation (6).

*3.3. Determining Hotspots.* Hotspot identification is an important factor to consider for crime detection. A hotspot represents highly frequent crime locations; hence, accurate prediction of the crime hotspots increases the accuracy of the crime detection process. Hotspot represents a spatial relationship between the occurrences of crime.

The calculation of hotspots is as follows: first, the coordinates of all crime reporting are grouped based on the type of crime. For example, the coordinates of "VEHICLE–STOLEN" are grouped into a separate list; second, the $X$ and $Y$ locations are clustered using $K$-means clustering. Finally, the distance from the nearest cluster is found. The working of hotspot identification is presented in Algorithm 1. The algorithm converges when there are no more additional changes in the clusters. Figure 3 illustrates the working of hotpot identification.

*3.4. Vulnerability Analysis.* In this subsection, we present the vulnerability analysis, which can detect the possible areas where there are more chances for a crime to occur. Suppose we consider an area $X$, a crime $Y$ has happened, that means the area is open to attacks or there are fewer or insufficient security measures. Hence, the area surrounding $X$ is more likely to become vulnerable to $Y$. We use kNN to analyse the vulnerability. Let us say, there is a vehicle theft at a place $X$, that means $X$ has less security for monitoring the crime $Y$ ; hence, the same area or the surrounding areas are too likely to become a vulnerable point. Link-based algorithms such as [21] will be helpful in creating a graph; the latter kNN algorithm can easily predict the crime spots. Figure 4 shows a visualization of crime in San Francisco; the visualization shows which areas are vulnerable and lack security monitoring.

We have considered 5 as the value for $k$ and the kNN used in this model produces 86.61% accuracy.

The proposed crime detection model works as follows: Firstly, a day is fragmented into four sections because it enhances the identification of temporal patterns of crimes. Few crimes such as robbery and chain snatching mostly occur at night, whereas other crimes such as hit and run and kidnapping occur during the day. Segregation of the day into various time quantum can help the prediction process. Secondly, the relationship between various crimes is established, i.e., how different crimes are linked to each other. The proposed method uses the crime correlation and growth rate to increase the prediction of the crime events. Thirdly, the hotspots of crime are identified. A hotspot represents a small geographical location where many crime incidents have been reported. Finally, vulnerability identification allows the proposed method to recommend an area where crime events are likely to occur in the future. By using both temporal and spatial inputs, the proposed model develops an increased ability to correctly predict crime events.

## 4. Results and Discussion

The proposed algorithm is to predict the probability of a given crime for a given area. We performed a comparison of our results with other machine learning algorithms such

**procedure** HOTSPOT Generation
  **K**←The number of hotspot
  Output ←**S**{}, the crime location in each hotspot
begin:
  Initialize the midpoints $m^{(1)}$={**R**andom **K** points}
  **for***i*=1 to *k***do**
    Add respective midpoint to the hotspot. $C_i^{(1)} = m_j^{(1)}$
  iter=1
  **while** True **do**
    **for** each point P**do**
      min=0
      Cluster=None
      **for***i* = 1 to *k***do**
        dist=$\|p - m_i^{iter}\|^2$
        **if** min<dist **then**
          min=dist
          cluster=*i*
      $S_{\text{cluster}} = \mathbf{S}_{cluster}^{iter} \cup \{\boldsymbol{P}\}$
      iter=iter+1
      **for***i* = 1 to *k***do**
        $m_i^{iter} = 1/|S_i^{iter-1}| \sum x \in S_i^{iter=1x}$
    break when past 2 **S** values are same
  return **S**

Algorithm 1: Identification of hotspots.

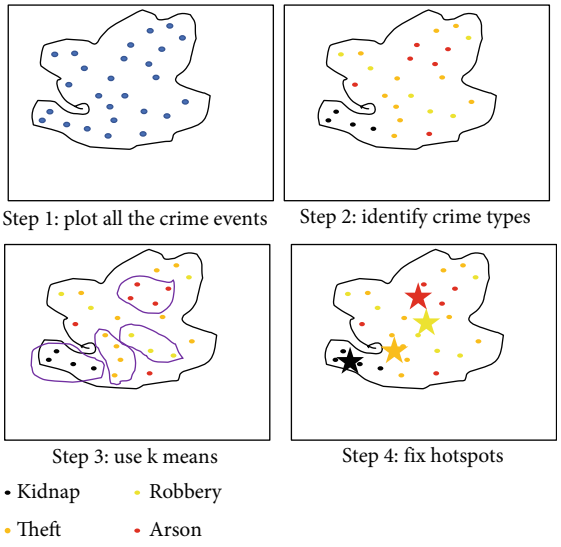



Figure 3: Illustration of hotspot identification.

as Naive Bayes, kNN, random forest, and support vector machines.

*4.1. Dataset Description.* We have used four attributes present in the Los Angeles dataset and six attributes in San Francisco dataset. The attributes used for the dataset are shown in Table 2. These attributes are used to train the existing machine learning algorithms.

In addition to the attributes present in the dataset, we have added four new attributes as discussed in Section 3 and fed into the proposed method.

*4.2. Evaluation Metrics.* Our evaluation metrics include accuracy, precision, and recall. The outputs of all classifiers are binary; hence, we can define the terms true positive (TP), true negative (TN), false positive (FP), and false negative (FN) as follows.

(i) TP: when a crime event is predicted as a crime event

(ii) TN: when a noncrime event is predicted as a noncrime event

(iii) FP: when a noncrime event is predicted as a crime event

(iv) FN: when a crime event is predicted as a noncrime event

We used three parameters (i.e., accuracy, precision, and recall) to test and evaluate the performance of the proposed model using existing machine learning algorithms.

*Accuracy:* accuracy is defined as the quality of correctness, and it is calculated by using the formula given by the equation

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{7}$$

*Precision:* precision explains how many positives out of the total positives predicted are. Precision is calculated based on

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{8}$$


Figure 4: Crime incidents and the vulnerable areas (data taken from San-Francisco crime dataset: Mohan [22, 23]).

Table 2: Dataset description.

| Dataset | Attributes used | Attributes generated |
|---|---|---|
| Crime in Los Angeles | (1) Crime code<br>(2) Date occurred<br>(3) Time occurred<br>(4) Location | (1) Predicted probability<br>(2) $X$<br>(3) $Y$<br>(4) Day of week<br>(5) Fraction of day |
| San Francisco Crime Dataset | (1) Category<br>(2) Day of week<br>(3) Date<br>(4) Time<br>(5) $X$<br>(6) $Y$ | (1) Predicted probability<br>(2) Fraction of day |

Table 3: Performance evaluation: San Francisco Dataset.

| Classifier | 15 days average | | | 7 days average | | | 2 days average | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| NB | 95 | 94 | 95.91 | 92.5 | 93.68 | 90.8 | 90.5 | 89.98 | 90.81 |
| NB (with proposed features) | 97.5 | 97.97 | 97 | 94.5 | 91.91 | 96.8 | 92.5 | 91.91 | 92.85 |
| RF | 93.5 | 94.05 | 93.13 | 89.5 | 87.87 | 90.6 | 90 | 87.87 | 91.57 |
| RF (with proposed features) | 97 | 96.03 | 97.97 | 93 | 88.11 | 97.8 | 91.5 | 88.11 | 94.68 |
| kNN | 95 | 93 | 96.87 | 94 | 92.85 | 94.79 | 90 | 89.21 | 91 |
| kNN (with proposed features) | 95.5 | 95.95 | 95 | 94 | 92.92 | 94.84 | 91.5 | 92.92 | 90.19 |
| SVM | 96.5 | 97 | 96.03 | 91.5 | 91.57 | 90.62 | 90 | 91.57 | 87.87 |
| SVM (with proposed features) | 97 | 98 | 96.07 | 92.5 | 97.8 | 87.25 | 91.5 | 95.69 | 87.25 |

*Recall:* the recall is a measure to calculate how many actual positives from all predicted positives found by the classifier. The recall is calculated by

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{9}$$

*4.3. Day History Analysis.* We considered three different values for the construction of Final Value $\text{FV}_{ij}$, i.e., 15, 7, and 2. The performance evaluations for both datasets are shown in Tables 3 and 4, respectively. We found that classifier performed better given more historical data. We also found that Naive Bayes resulted in the best performance when the number of days was 15 (i.e., 15-day average) compared to other classifiers.

Crime predictions based on patterns were performed by the classifiers. Additionally, we input four new attributes as mentioned in Section 3 into the classifiers. This allowed

Table 4: Performance evaluation: Los Angeles Dataset.

| Classifier | 15 days average | | | 7 days average | | | 2 days average | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| NB | 91 | 93.75 | 88.24 | 90 | 91.67 | 88 | 89 | 89.58 | 87.76 |
| NB (with proposed features) | 93 | 92.93 | 92.93 | 92.5 | 91.92 | 92.86 | 90.5 | 88.35 | 92.86 |
| RF | 90.5 | 88.24 | 92.78 | 89.5 | 88.24 | 90.91 | 88 | 87.88 | 87.88 |
| RF (with proposed features) | 94.5 | 92.08 | 96.88 | 93.5 | 90.29 | 96.88 | 92.5 | 90.1 | 94.79 |
| kNN | 88 | 83.33 | 93.75 | 88.5 | 85.05 | 92.86 | 85.5 | 81.98 | 91 |
| kNN (with proposed features) | 92 | 92.86 | 91 | 91.5 | 92.78 | 90 | 90.5 | 92.78 | 88.24 |
| SVM | 87 | 85.71 | 87.5 | 88.5 | 88.42 | 87.5 | 89.5 | 90.53 | 87.76 |
| SVM (with proposed features) | 92.5 | 97.8 | 87.25 | 91 | 97.8 | 84.76 | 89.5 | 95.7 | 83.96 |

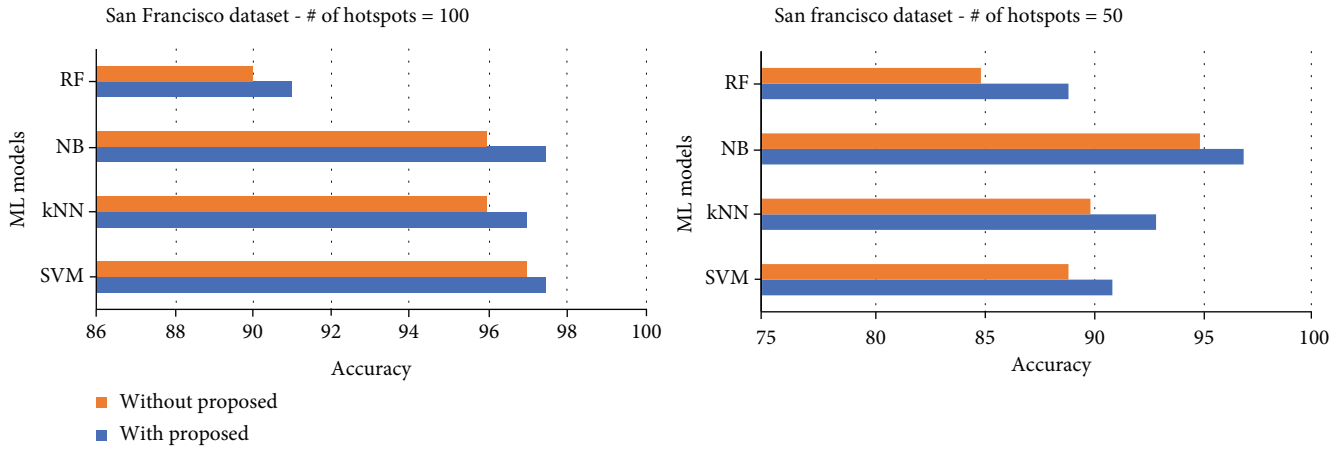



Figure 5: Accuracy of San Francisco Dataset, when # of hotspots is 100 and 50.

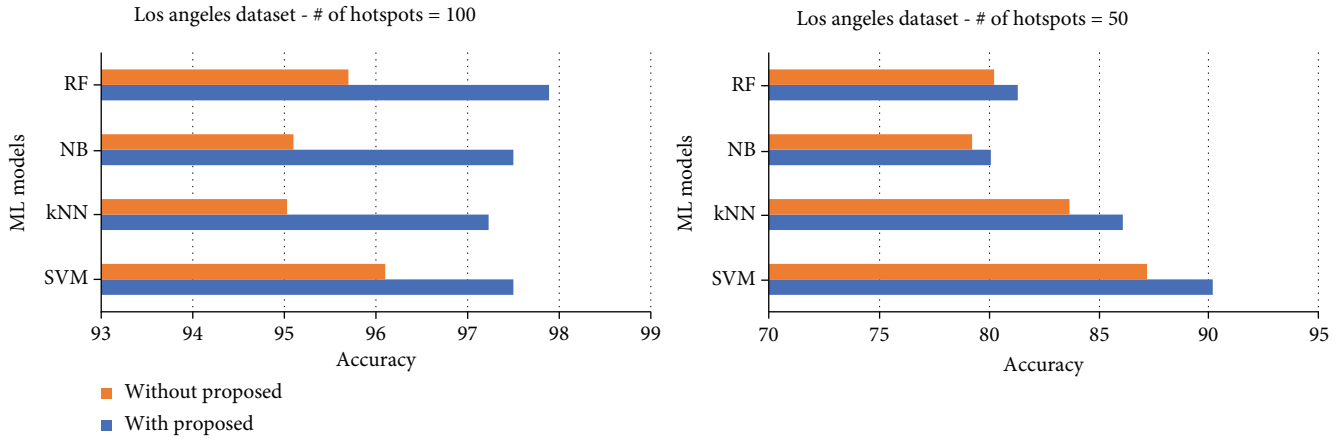



Figure 6: Accuracy of Los Angeles Dataset, when # of hotspots is 100 and 50.

the consideration of crime probabilities, hotspots, and vulnerability analysis. This information helped the classifier to analyse the time series and predict the crime rate better.

*4.4. Analysis of Hotspots.* Similar crimes are likely to happen frequently at the same place, which includes highly dense areas or low secured places and so on. This information can be captured using a hotspot cluster. Thus, the distance from a cluster is an important factor to consider for crime prediction. If the distance is very low, then it is more likely for a crime to happen.

A hotspot represents a spatial relationship with high frequent crimes [24]. The accurate prediction of crime hotspots helps the police department to take timely action to avoid crime at specific locations. Determining the number of clusters is an important criterion [25, 26]. We have assumed a

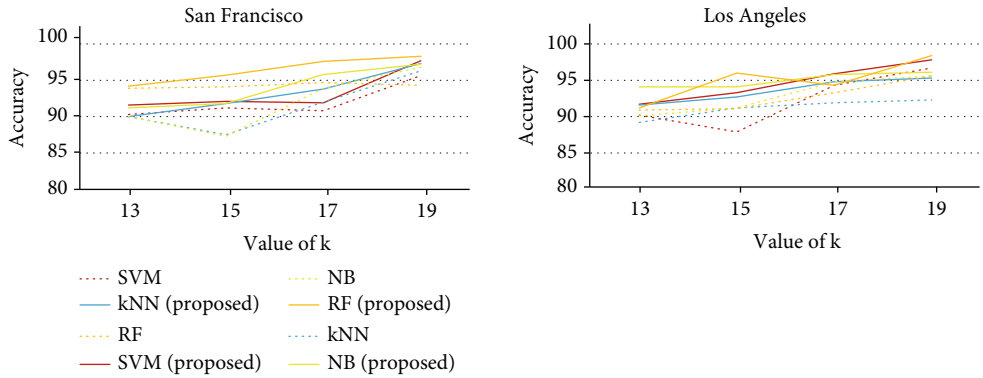



Figure 7: Accuracy measure when $n$ (neighbour) value is 13, 15, 17, and 19.

cluster head per 1 $KM^2$ or 2 $KM^2$ and fixed the number of cluster heads as 50 and 100. Figures 5 and 6 show the accuracy comparison of the classifiers with and without the proposed method for the databases San Francisco and Los Angeles, respectively.

*4.5. Analysis of Vulnerability.* A place is vulnerable for crime when any neighbouring area witness a crime event [27, 28]. We tested the performance of our proposed method using the number of neighbours 13, 15, 17, and 19 [19]. The graph in Figure 7 shows the accuracy of the different classifiers when the value of $k$ changes.

## 5. Conclusion

Despite many preventive measures, crime rates increase day by day in several regions. This paper concentrates on feature generation methods such as time zone classification, crime probability calculation, analysis of crime hotspots, and vulnerability analysis. The recommended features are fed into four machine learning models which comprises random forest, $K$ nearest neighbour, support vector machines, and Naïve Bayes. The results show that Naïve Bayes produced successful results in predicting the crime incidents.

## Symbols

$K$: How many unique crime events
$S$: crime locations
$m$: crime hotspot location
$C$: clusters
$P$: temporary points.

## Data Availability

Data are available in San Francisco open data https://github.com/ashok0501/ResearchPaperCodes.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] J. Borges, D. Ziehr, M. Beigl et al., "Feature engineering for crime hotspot detection," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp. 1–8, San Francisco, CA, USA, 2017.

[2] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma, and N. Yadav, "Crime pattern detection, analysis & prediction," in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 1, pp. 225–230, Coimbatore, India, 2017.

[3] B. Sivanagaleela and S. Rajesh, "Crime analysis and prediction using fuzzy c-means algorithm," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 595–599, Tirunelveli, India, 2019.

[4] U. V. Navalgund and K. Priyadharshini, "Crime intention detection system using deep learning," in *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, pp. 1–6, Kottayam, India, 2018.

[5] Y. Lee, T. Song, H. Kim, D. K. Hant, and H. Ko, "Hostile intent and behaviour detection in elevators," in *4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*, pp. 1–6, London, 2011.

[6] M. Nakib, R. T. Khan, M. S. Hasan, and J. Uddin, "Crime scene prediction by detecting threatening objects using convolutional neural network," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pp. 1–4, Rajshahi, Bangladesh, 2018.

[7] S. Ranjan, P. Garhwal, A. Bhan, M. Arora, and A. Mehra, "Framework for image forgery detection and classification using machine learning," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1872–1877, Tirunelveli, India, 2018.

[8] O. Vynokurova, D. Peleshko, O. Bondarenko, V. Ilyasov, V. Serzhantov, and M. Peleshko, "Hybrid machine learning system for solving fraud detection tasks," in *2020 IEEE Third International Conference on Data Stream Mining Processing (DSMP)*, pp. 1–5, Lviv, Ukraine, 2020.

[9] E. E. Eryilmaz, D. O. Ahin, and E. Kl, "Machine learning based spam e-mail detection system for Turkish," in *2020 5th International Conference on Computer Science and Engineering (UBMK)*, pp. 7–12, Diyarbakir, Turkey, 2020.

[10] B. S. Aldossari, F. M. Alqahtani, N. S. Alshahrani et al., "A comparative study of decision tree and naive bayes machine learning model for crime category prediction in Chicago," in *Proceedings of 2020 the 6th International Conference on Computing and Data Engineering, ser. ICCDE 2020*, p. 3438, New York, NY, USA, 2020.

[11] A. Chowdhary and B. Rudra, "Video surveillance for the crime detection using features," in *Advances in Intelligent Systems and Computing Advanced Machine Learning Technologies and Applications*, pp. 61–71, Cairo, Egypt, 2020.

[12] R. Jain, A. Nayyar, and S. Bachhety, "Factex: a practical approach to crime detection," in *Data Management, Analytics and Innovation Advances in Intelligent Systems and Computing*, p. 503516, Pune, India, 2019.

[13] S. Afzal, M. Asim, A. R. Javed, M. O. Beg, and T. Baker, "URLdeepDetect: a deep learning approach for detecting malicious URLs using semantic vector models," *Journal of Network and Systems Management*, vol. 29, no. 3, 2021.

[14] P. Ashokkumar, N. Arunkumar, and S. Don, "Intelligent optimal route recommendation among heterogeneous objects with keywords," *Computers & Electrical Engineering*, vol. 68, pp. 526–535, 2018.

[15] J. He and H. Zheng, "Prediction of crime rate in urban neighborhoods based on machine learning," *Engineering Applications of Artificial Intelligence*, vol. 106, p. 104460, 2021.

[16] S. A. Chun, V. A. Paturu, S. Yuan, R. Pathak, V. Atluri, and N. R. Adam, "Crime prediction model using deep neural networks," in *Proceedings of the 20th Annual International Conference on Digital Government Research*, pp. 512–514, Dubai United Arab Emirates, 2019.

[17] A. Palanivinayagam and S. Nagarajan, "An optimized iterative clustering framework for recognizing speech," *International Journal of Speech Technology*, vol. 23, no. 4, pp. 767–777, 2020.

[18] A. R. Javed and Z. Jalil, "Byte-level object identification for forensic investigation of digital images," in *2020 International Conference on Cyber Warfare and Security (ICCWS)*, Islamabad, Pakistan, 2020.

[19] N. Deepa, Q. Pham, D. C. Nguyen et al., "A survey on blockchain for big data: approaches, opportunities, and future directions," 2020, http://arxiv.org/abs/2009.00858.

[20] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[21] P. Ashokkumar and S. Don, "Link-based clustering algorithm for clustering web documents," *Journal of Testing and Evaluation*, vol. 47, no. 6, p. 20180497, 2019.

[22] N. Mohan, "Crime analysis with San Francisco open data interactive data visualization with Python and Plotly," June 2019. https://medium.com/@navaneeth.mohan94/crimeanalysis-with-san-francisco-open-data-interactive-data-visualizationwith-python-and-plotly-7b7db7e65d72.

[23] ResearchPaperCodes, "Python code," March 2021, https://github.com/ashok0501/ResearchPaperCodes.

[24] P. Ashok Kumar, G. Shiva Shankar, P. K. R. Maddikunta, T. R. Gadekallu, A. Al-Ahmari, and M. H. Abidi, "Location based business recommendation using spatial demand," *Sustainability*, vol. 12, no. 10, p. 4124, 2020.

[25] A. R. Javed, M. O. Beg, M. Asim, T. Baker, and A. H. Al-Bayatti, "AlphaLogger: detecting motion-based side-channel attack using smartphone keystrokes," *Journal of Ambient Intelligence and Humanized Computing*, vol. 1, 2020.

[26] M. Mittal, C. Iwendi, S. Khan, and A. Rehman Javed, "Analysis of security and energy efficiency for shortest route discovery in low-energy adaptive clustering hierarchy protocol using Levenberg-Marquardt neural network and gated recurrent unit for intrusion detection system," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 6, 2021.

[27] C. Iwendi, Z. Jalil, A. R. Javed et al., "KeySplitWatermark: zero watermarking algorithm for software protection against cyber-attacks," *IEEE Access*, vol. 8, pp. 72650–72660, 2020.

[28] A. Palanivinayagam and D. Sasikumar, "Drug recommendation with minimal side effects based on direct and temporal symptoms," *Neural Computing and Applications*, vol. 32, no. 15, pp. 10971–10978, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8752022, 9 pages
https://doi.org/10.1155/2021/8752022

*Research Article*

# K-Nearest Robust Active Learning on Big Data and Application in Epitope Prediction

**Tianchi Lu**

*School of Mathematics and Statistics, Lanzhou University, Lanzhou, China*

Correspondence should be addressed to Tianchi Lu; lutch17@lzu.edu.cn

Received 11 August 2021; Revised 10 September 2021; Accepted 17 September 2021; Published 11 November 2021

Academic Editor: Rajesh Kaluri



B-cells that induce antigen-specific immune responses in vivo produce large numbers of antigen-specific antibodies by recognizing subregions (epitopes) of antigenic proteins, in which they can inhibit the function of antigen protein. Epitope region prediction facilitates the design and development of vaccines that induce the production of antigen-specific antibodies. There are many diseases which are difficult to treat without vaccines. And the COVID-19 has destroyed many people's lives. Therefore, making vaccines to COVID-19 is very important. Making vaccines needs a large number of experiments to get labeled targets. However, obtaining tremendous labeled data from experiments is a challenge for humans. Big data analysis has proposed some solutions to deal with this challenge. Big data technology has developed very fast and has been applied in many areas. In the bioinformatics area, big data analysis solves a large number of problems, particularly in the area of active learning. Active learning is a method of building more predictive models with less labeled data. Active learning establishes models with less data by asking the oracle (human) for the most valuable samples to train models. Hence, active learning's application in making vaccines is meaningful that the scientists do not need to do tremendous experiments. This paper proposed a more robust active learning method based on uncertainty sampling and K-nearest density and applies it to the vaccine manufacture. This paper evaluates the new algorithm with accuracy and robustness. In order to evaluate the robustness of active learners, a new robustness index is designed in this paper. And this paper compares the new algorithm with a pool-based active learning algorithm, density-weighted active learning algorithm, and traditional machine learning algorithm. Finally, the new algorithm is applied to epitope prediction of B-cell data, which is significant to making vaccines.

## 1. Introduction

Big data analysis is a thriving field. The branch of big data analysis, artificial intelligence, has greatly promoted the team's understanding of life science in the field of bioinformatics [1, 2]. We can use machine learning to predict major disease problems for the benefit of human beings, such as vaccine manufacturing.

Now, people are fully aware of the importance of health. At the same time, with the development of the Internet and big data, many mobile applications with collaborative systems have been developed to detect people's health [3–5]. As we can see from previous work, many collaborative systems have begun to work with machine learning methods [3–5]. These systems can use machine learning to detect body states. At the same time, COVID-19 disease has destroyed many people's lives, so it is necessary to use such a system to detect whether people have COVID-19. And making vaccines to COVID-19 is also an emergency. Therefore, we choose the B-cell [6, 7] data which is so relative to the immune system. Antibodies inhibit the function of antigen proteins by identifying antigen epitope that can be seen as "vaccines," because B-cells are immune cells that can recognize antigens when producing antibodies. Therefore, predicting epitopes using B-cell data [6, 7] is important for the preparation of experimental vaccines. Since it is not difficult to get the specimen of B-cell, using the epitope prediction to the health collaborative systems is a good way to assess whether people are suffering from COVID-19. Several physical and computational methods [8–12] have been proposed to predict epitopes. In the physical methods, the features used are limited to those associated with the target

amino acid sequence, so the representations of these models are inadequate [6]. And using the physical methods to predict epitope requires tremendous experiments which need labor, establishment, and money. Although the computational methods have achieved better, it still requires tremendous samples to train. Therefore, these methods are expensive.

There are several ways to cope with the big data problem when reducing the burden of data and experiments. Dimensionality reduction [1] is one of the most important methods to reduce the complexity of models and select the most important variables. However, dimensionality reduction still requires tremendous samples. Active learning [13] is a solution to this problem. Active learning which aims to reduce the number of samples required by asking the oracle is a subfield of machine learning with the same name [14] in educational literature. And the area of active learning is booming: many active learning methods [15, 16] have been proposed. These algorithms are based on uncertainty sampling. In addition to uncertainty sampling, many sampling processes for active learning are proposed. Based on label changes [17], committee queries [18], representative changes [19], and density-based sampling [20, 21] are some of the processes. More importantly, the active learning method has been successfully applied to speech recognition [22], information extraction [23], and bioinformatics [24–26].

We believe that using active learning is of great significance to predict epitope, and this paper mainly concentrates on the uncertainty sampling [27, 28] and density sampling method [20, 21]. The uncertainty sampling method usually selects the outliers as the most uncertain and informative samples to ask the oracle. Outliers are not so valuable and may result in less robust classifiers when new samples are added to the training data. To solve this problem, density-weighted sampling has been proposed. Density-weighted sampling [20, 21] is a good way to solve the outlier problem. But the density-weighted method does not provide the same information as the uncertainty sampling method. Recently, some methods have developed new loss functions by integrating uncertainty sampling and K-nearest density weighting to improve the performance of active learning [29–31]. However, these methods may still cause loss of information, just like the density-weighted method. And calculating the density of samples in the pool samples is difficult since the computing complexity is great when there are many samples [32]. In order to use the most valuable data and make more robust query strategies without high complexity, this paper establishes a new algorithm. Specifically, the work uses uncertainty sampling to find the most informative points firstly, then uses K-nearest density in the uncertainty data with $L_1$ norm (Manhattan distance) to eliminate outliers to improve pool-based active learners' robustness. This paper calls the new algorithm K-nearest robust active learning (KRAL). Compared to the density-weighted method like SUD [31], the KRAL is with less complexity. This is because not many most uncertainty samplers are generated in each step; calculating density in this dataset does not result in computing complexity being too high. At the same time, using the K-nearest density method, we eliminate the outliers, which guarantees the maximization of information utilization and does not increase computing complexity excessively.

Our proposal is to make a new algorithm which predicts the epitope with less labeled data and higher accuracy when compared to the existed pool-based active learning and density-weighted active learning algorithms in epitope prediction problem. Hence, this paper uses B-cell data with epitope to do the experiments. The data comes from the immune epitope database (IEDB) which is a public database of immune epitope [7]. By experimenting and comparing the KRAL with pool-based active learning and density-weighted method on B-cell data, we finally get a more accurate and robust model with less complexity. Therefore, the results of this study may be helpful in the production of the COVID-19 vaccine.

## 2. Data and Methodology

*2.1. Data and Task Description.* The world is suffering from a pandemic in which COVID-19 has destroyed a large number of people's lives. Substances that mimic the structure and function of epitopes can be thought of as "vaccines" of organisms designed to induce specific antibodies in vivo. Therefore, the B-cell data [6] is selected for this study. B-cells are immune cells that recognize antigens when producing antibodies. Antibodies can inhibit the function of antigen proteins by binding to antigen epitope regions. Hence, it is very helpful to find a good prediction model of epitope for this problem. There are some physical methods to predict the epitope. For instance, the three-dimensional structural analysis of antibody-antigen complexes by X-ray [9] or nuclear magnetic resonance (NMR) spectroscopy [10] is considered to identify the epitope.

But these methods are quite expensive and require a lot of time and labor to predict epitope. Recently, various big data analysis methods were proposed based on machine learning [11, 12]. Under this circumstance, the performance of epitope prediction has improved by machine learning methods. But we still need a lot of data for training, which is still expensive. Hence, the task is still challenging for humans. Next, we describe this task in detail.

The data and variables description:

Independent variables:

(i) start_position: start position of peptide

(ii) end_position: end position of peptide

(iii) chou_fasman: peptide feature, $\beta$ turn

(iv) emini: peptide feature, relative surface accessibility

(v) kolaskar_tongaonkar: peptide feature, antigenicity

(vi) parker: peptide feature

(vii) isoelectric_point: protein feature

(viii) aromaticity: protein feature

(ix) hydrophobicity: protein feature

(x) stability: protein feature

Dependent variable:

(i) Antibody valence (target value)

The task is a binary classification problem with 10 independent variables, and the target was antibody valence, where 0 stands for negative and 1 stands for positive. There are 14387 samples in the data. The structure of the dataset is shown in Figure 1 and Table 1. Figure 1 and Table 1 illustrate that about 3/4 samples are negative and 1/4 are positive. From the skewness and kurtosis from Table 1, we can see that some of the independent variables do not follow the normal distribution, and some are sparsely distributed. Particularly, the end_position, start_position, and emini are not obeying the normal distribution. And some others, like the hydrophobicity, are sparsely distributed. Therefore, the dataset may have some outliers that may affect the performance of active learning algorithms. Therefore, traditional machine learning and active learning methods may not work well.

*2.2. Methodology Description.* In this paper, we propose a new big data analysis method to predict epitope, and B-cell data were used to establish the model. The detailed steps of this work are shown in Figure 2. More specifically, this paper uses KRAL to predict targets and incorporates the new algorithm with traditional pool-based active (PBL) learners, density-weighted active learning method (SUD), and basic algorithms (random forest [33] and SVM [34]) with random selection (RS) in both accuracy and robustness. In order to evaluate the active learners' robustness, this paper designs a new index called sequential robust index (SRI).

## 3. Active Learning Process

This paper is interested in big data and pool-based active learning based on uncertainty sampling [15]. That is, active learners have the least confidence in the samples with the greatest uncertainty, while pool-based active learners have two-stage samples. There are a small number of labeled samples and a large number of unlabeled samples. Pool-based active learners require oracle to provide the most uncertainty samples and add them to the labeled samples for the next training. The full algorithm is illustrated as follows [16]. The algorithm results are shown in Algorithm 1.

Pool-based active learning:

## 4. Uncertainty Measures

There is too much useless information when dealing with big data. Therefore, choosing the sample with the most useful information is important. In the uncertainty sampling scheme, the unlabeled sample with the largest uncertainty is considered the one with the largest amount of information. Therefore, it is significant to find a good evaluation method of measurement sample uncertainty.

The well-known entropy [27] has been widely used in previous studies in evaluating uncertainty [35, 36].

$$\mathbf{H}(\mathbf{x}) = -\sum_{\mathbf{y}\in\mathbf{Y}} \mathbf{P}(\mathbf{y} \mid \mathbf{x})\mathbf{logP}(\mathbf{y} \mid \mathbf{x}), \tag{1}$$

where $P(y \mid x)$ is the a posteriori probability, target (label) $\mathbf{y} \in \mathbf{Y} = \{y_1, y_2, \cdots y_k\}$. $\mathbf{H}(\mathbf{x})$ is the uncertainty measurement function based on the entropy estimation of the classifier's posterior distribution.

Entropy is a baseline method for measuring uncertainty, which involves a large amount of information. Therefore, this paper uses the entropy as the uncertainty sampling measurement.

## 5. K-Nearest Robust Active Learning

Although entropy contains the most useful information, it has some drawbacks. Entropy is to find the nearest point to the classification boundary, that is, outliers are usually used as the sample with the least confidence. Outliers may contain too much noise, resulting in poor robustness of the model. Therefore, dealing with outliers is a feasible way to improve the model's performance.

*5.1. K-Nearest Neighbor Classification.* K-Nearest Neighbor Classification (KNN) [32] learners are the basic way to deal with classification problems. KNN is characterized by estimating sample density using a distance function. Therefore, using K-nearest density can help us find outliers and pass them out of the training data. However, the KNN model is a method with great complexity. Therefore, how to use the K-nearest neighbor algorithm in active learning is a challenge.

*5.2. Distance Function.* There are many distance functions. Hence, selecting a fit distance function is fundamental to estimate the K-nearest density of samples. Considering the effect of outliers and the complexity of big data, this paper uses the Manhattan distance to calculate the K-nearest density.

Manhattan distance:

$$d_{12} = \sum_{k=1}^{n} |X_{1k} - X_{2k}|. \tag{2}$$

The Manhattan distance method is not affected so much by outliers. It evaluates whether two points are close or not. And the Manhattan distance is also not so complex. Therefore, calculating the Manhattan distance will not add too much complexity in big data. Hence, using Manhattan distance to calculate the distance between vectors is a good choice.

*5.3. K-Nearest Robust Active Learning.* Using the K-nearest classification to calculate the density of samples can help us to find the outliers. The density function $\mathbf{Den}(\mathbf{X_i})$ is

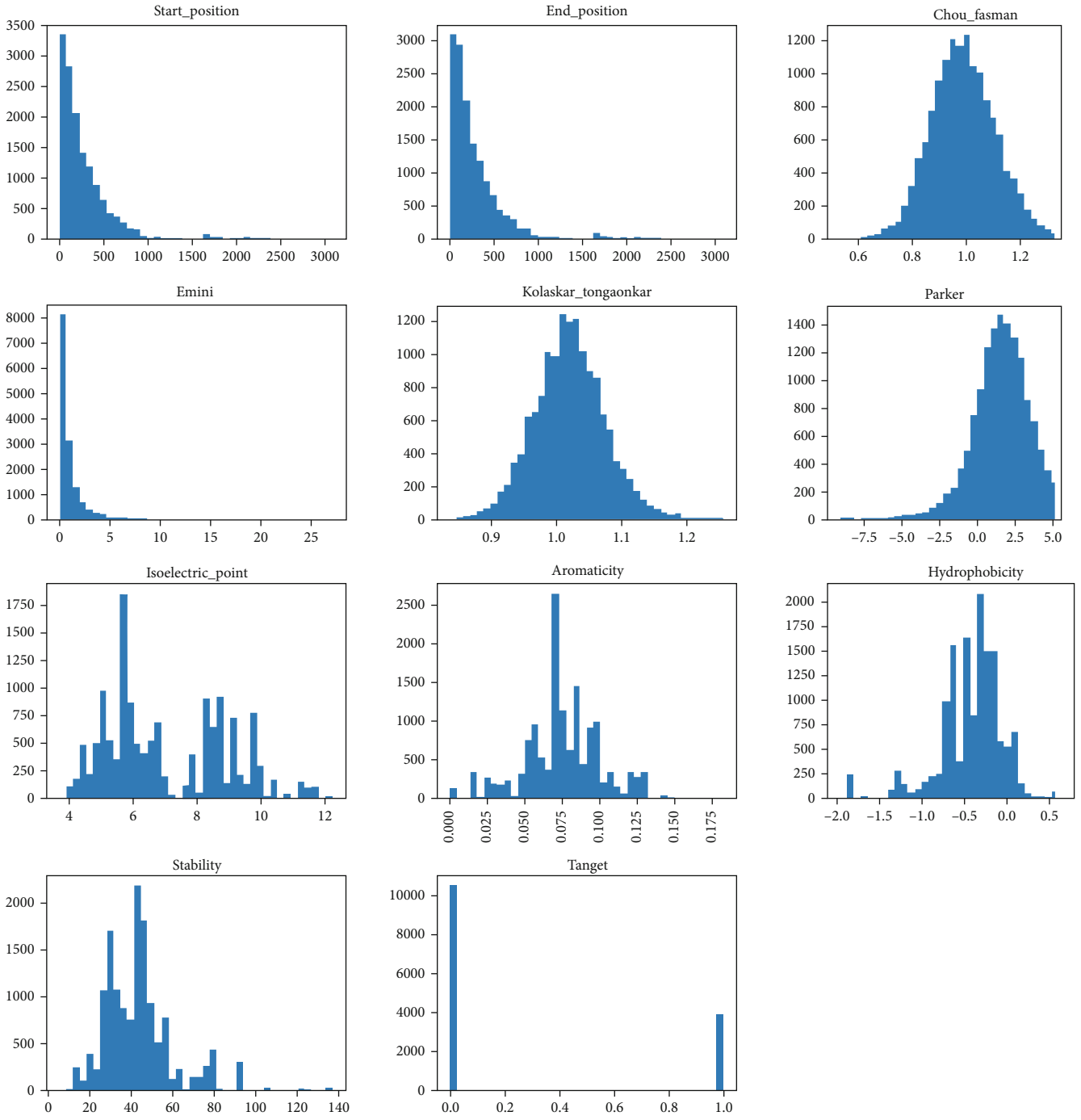



Figure 1: The data describing.

defined as

$$\mathbf{Den}(\mathbf{X_i}) = \frac{\mathbf{1}}{\sum_{\mathbf{J=1}}^{\mathbf{m}} \mathbf{d}(\mathbf{X_i}, \mathbf{X_j})}, \tag{3}$$

where $\mathbf{m}$ is the number of uncertainty samples. $\{\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_m}\}$ are the most $\mathbf{m}$th uncertainty samples. In the KRAL, the $\mathbf{K}$ is defined as the $\mathbf{m}$, which means that we calculate the distance among the most uncertainty samples at every step. From the form of density function, the sample with the smallest density function is the point which needs to be threaded out. Compared to a naive opinion which is to apply the K-nearest procedure like SUD for all unlabeled data, the new density functions do not add much complexity to the algorithm. This is because calculating the K-nearest density on a big data will increase the complexity greatly. And it is difficult to decide the $\mathbf{K}$ value because the whole unlabeled data is concluded too much samples. Hence, we cannot simply use $\mathbf{m}$ (number of samples) as the $\mathbf{K}$.

Table 1: An error and statistical analysis to the data.

| | Descriptive statistics | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *N* | Mean | Std. deviation | Variance | Skewness | | Kurtosis | |
| | Statistic | Statistic | Statistic | Statistic | Statistic | Std. error | Statistic | Std. error |
| Start_position | 14387 | 297.68 | 353.741 | 125133.014 | 3.009 | .020 | 11.607 | .041 |
| End_position | 14387 | 308.09 | 353.733 | 125127.245 | 3.005 | .020 | 11.574 | .041 |
| Chou_fasman | 14387 | .994705915000000 | .124772254000000 | .016 | .248 | .020 | .398 | .041 |
| Emini | 14387 | 1.059787725000000 | 1.621931429000000 | 2.631 | 5.051 | .020 | 40.411 | .041 |
| Kolaskar_tongaonkar | 14387 | 1.021188364000000 | .053804291800000 | .003 | .186 | .020 | .380 | .041 |
| Parker | 14387 | 1.767136582000000 | 1.968984865000000 | 3.877 | -.362 | .020 | 1.266 | .041 |
| Isoelectric_point | 14387 | 7.067471661000001 | 1.888708170000000 | 3.567 | .439 | .020 | -.915 | .041 |
| Aromaticity | 14387 | .075726787200000 | .025767473200000 | .001 | -.131 | .020 | .570 | .041 |
| Hydrophobicity | 14387 | -.406096679000000 | .394618135000000 | .156 | -.706 | .020 | 3.058 | .041 |
| Stability | 14387 | 43.703902170000000 | 16.682362480000002 | 278.301 | 1.366 | .020 | 3.248 | .041 |
| Valid *N* (listwise) | 14387 | | | | | | | |

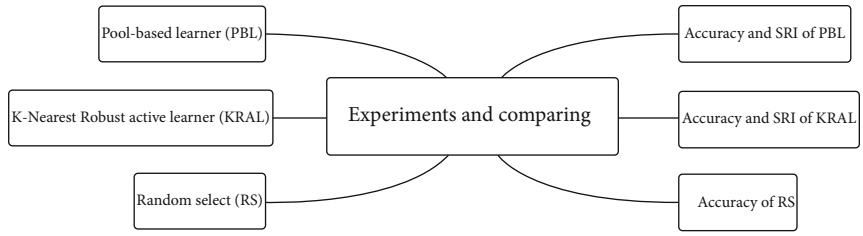



Figure 2: The step of this paper's work.

**Require:** A set of labeled samples L, a set of unlabeled samples U
  **while** Termination condition not satisfied do
    Train a classifier $\varphi c(\cdot|L)$ based on labeled samples;
    **for** i =1 : |U| **do**
      Calculate the uncertainty(Entropy) of the sample, $Un(x_i^u)$;
      Select the top nth uncertainty sample $y_i$ as a new set N to query;
      L=L∪N;
      U=U-N;
    **end for**
  **end while**

Algorithm 1: Pool-based active learning process

Therefore, how to choose the **K** is a challenge. And the most uncertainty samples are more like to be outliers. Hence, calculating the density among the most uncertainty samples is a good idea. Hence, this paper can exclude outliers from the least confidence level sample with K-nearest density.

Then, we detail the new algorithm steps. The new algorithm (KRAL) uses entropy sampling in the first step to select the most indeterminate sample, which is the same as pool-based active learning. Next, KRAL uses Manhattan distance to calculate the K-nearest density in the most uncertain samples at the next step. Then, the KRAL selects the sample with the lowest density for threading, because the big data is so complex and large that many methods cannot be used to cope with big data. From the form of KRAL, we can see that KRAL both consider the computing complexity and accuracy. Hence, the KRAL can be applied in big data analysis. And the new full algorithm of KRAL is illustrated as follows. The algorithm results are shown in Algorithm 2.

K-nearest robust active learning:

## 6. Experiments and Numeric Research

In this section, experimental and numerical studies are performed using B-cell data. More specifically, this paper compares the accuracy and robustness of KRAL and pool-based active learning.

*6.1. Sequential Robust Index (SRI).* There is no method that has been proposed to evaluate the robustness of active learning when adding new samples to training data. In order to

**Require:** A set of labeled samples L, a set of unlabeled samples U
  **while** Termination condition not satisfied do
    Train a classifier $\varphi c(\cdot|\mathbf{L})$ based on labeled samples;
  **for** i =1 : |U| **do**
    Calculate the uncertainty(Entropy) of the sample, $\mathbf{Un}(\mathbf{x_i^u})$;
    Select the top nth uncertainty sample $\mathbf{y_i}$ as a new set N ;
    Calculate the density function $\mathbf{Den}(\mathbf{y_i})$ in set N;
    p= **argmin** $\mathbf{Den}(\mathbf{y_i})$;
    N=N-p;
    use N's samples to query the oracle;
    L=L∪N;
    U=U-N;
  **end for**
  **end while**

Algorithm 2: K-nearest robust active learning process

evaluate the robustness of the algorithms, a new robustness index (SRI) for the sequence of robustness evaluation indexes is presented. The sequential robust index is defined as

$$\sum \boldsymbol{I}(\boldsymbol{a}_i - \boldsymbol{a}_{i-1} < \mathbf{0}) \sum |\boldsymbol{a}_i - \boldsymbol{a}_{i-1}| \boldsymbol{I}(\boldsymbol{a}_i - \boldsymbol{a}_{i-1} < \mathbf{0}), \quad (4)$$

where the $\mathbf{a_i}$ is the accuracy of one-step test data and $\mathbf{I}(\mathbf{x})$ is the indication function. When $x < 0$, $\mathbf{I}(\mathbf{x}) = 1$; otherwise, $\mathbf{I}(\mathbf{x}) = 0$.

We expect that when new samples are added to the training set, the prediction accuracy of the test set will increase. Through this way, we can reduce the computational complexity when facing big data. However, sometimes, adding new samples into the training set in active learning process will result in a lower prediction accuracy. The SRI measures the number of times predictive accuracy decreases and the total amount of decline when new samples are added to the training set. Because the fewer times the accuracy is reduced and the fewer the accuracy is reduced when adding samples into training data, the more valuable the data is added to the training set each time. Hence, it can be seen from the form of SRI that the smaller the index, the better the model. We can see that if a good query strategy is stable, the new data it queries will make the proactive learning prediction accuracy increasing. However, if a query strategy is unstable, the queried data may reduce the prediction accuracy, so SRI can measure the stability of a query strategy greatly at some step, that is, SRI evaluates the robustness of the query strategy. Therefore, the SRI can estimate the robustness of active learning during the query process. And the computation complexity of SRI is not high. So SRI can very deal with big data.

*6.2. Experimental Settings.* We use random forest (RF) and support vector machine (SVM) as base learners. And the query strategy is based on maximum entropy. Cross-validation is a good way to examine the performance of models in big data analysis. Therefore, in order to ensure the rationality of the experiment, we randomly select samples as labeled data by cross-validation and repeat the experiment 100 times and use the mean value to record in results. To be more specific, we randomly divide the data into 50 parts using 50-fold cross-validation and randomly select one of them as training set and the rest as pools for active learning queries. The data is a public dataset (IEDB) [7], which we will use for epitope prediction. And we use the pool samples as test set at every query step.

We compare KRAL with pool-based active learning and SUD. And our evaluation metrics are the test set accuracy, SRI, and the running time. Among them, test set accuracy is used to directly measure the effectiveness of several methods, SRI is used to evaluate the query robustness of several methods, that is, to evaluate the stability of the query, and running time is to evaluate the computational complexity of the model. We mainly compare the effectiveness and computational complexity of every algorithm.

In every query step, we let the most uncertain dataset includes 40 samples. Under this circumstance, we continue our experiments. And my computer setting is GPU: RTX 3060 and CPU: 16G, I7, 11th generation.

The IDE is Spyder.

## 7. Result and Analysis

This paper records the accuracy and SRI when the number of samples increases. The results are recorded in Figures 3–6. Figures 3 and 5 record the accuracy of each learner, and Figures 4 and 6 record the SRI of each active learner. From Figures 3–6, we can see the results of each model: random selection sampling is the weakest in both random forests and support vector machine models. As the number of samples increases, the sensitivity of random selection sampling decreases. In the SVM model, adding new samples to the training data does not significantly improve the accuracy. Both pool-based active learning, KRAL and SUD methods, improve the performance of basic learners. Figures 3–6 show that when new samples are added, the active learning's accuracy is higher than the basic learners. Therefore, using the active learning method can reduce the complexity when coping with big data. And when the basic learner is random forest, the performance of KRAL is 12.1% better than that of the basic learner. Therefore, the effectiveness of KRAL was

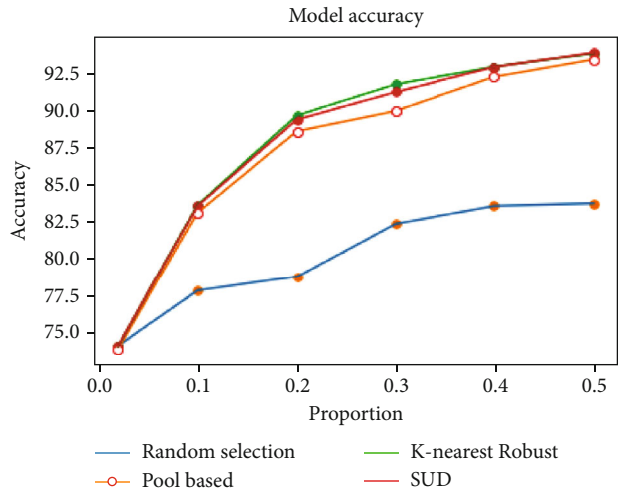



Figure 3: The accuracy in the random forest model.

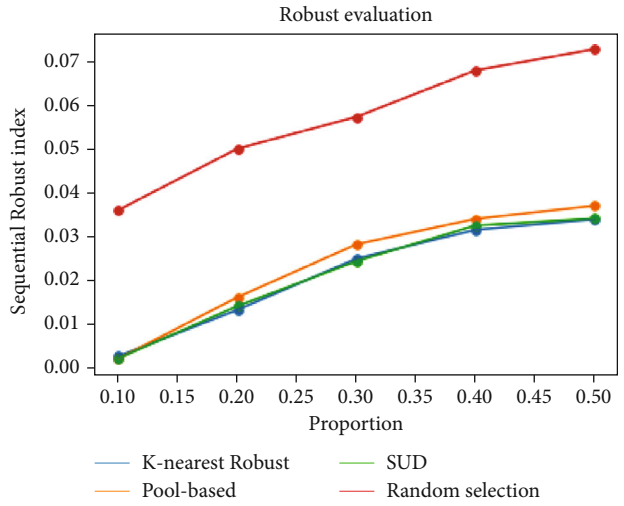



Figure 4: The robustness in the random forest model.

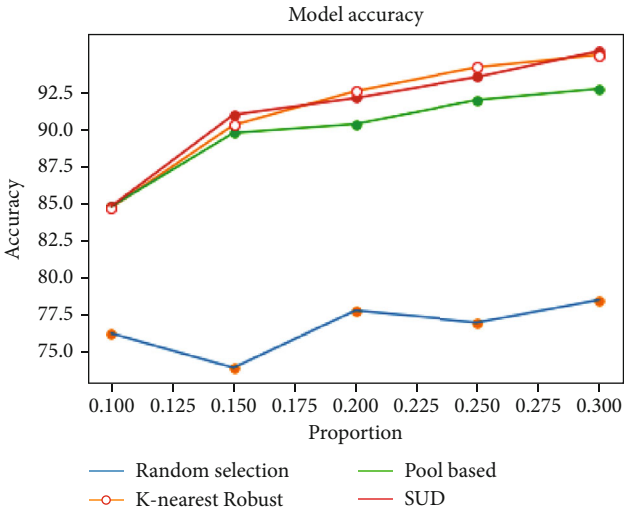



Figure 5: The accuracy in the SVM model.

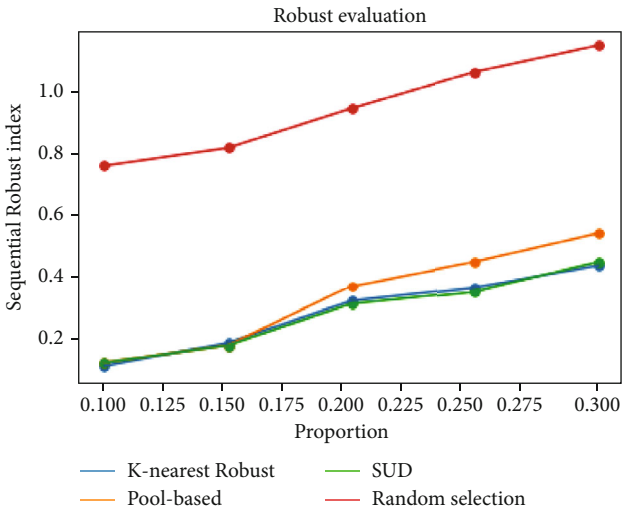



Figure 6: The robustness in the SVM model.

examined. So active learning methods can be used to reduce the computational complexity in big data analysis and improve the accuracy.

Meanwhile, the SRI of KRAL and SUD is smaller than pool-based active learning and random selection strategy, indicating that KRAL and SUD are more robust than pool-based active learning. And the SRI among the active learners is much smaller than random selection. More specifically, the KRAL algorithm achieves at least 6% and 15% higher in SRI evaluation and the SUD algorithm achieves at least 5.8% and 15.1% higher in SRI evaluation than pool-based active learning algorithm in random forest and SVM. And KRAL and SUD algorithms are more accurate than pool-based active learning algorithm in SVM and RF models. When the sample scale is 5/10, the learning accuracy of KRAL is 0.5% higher than that of pool-based active learners in both two basic learners (RF and SVM). Therefore, SUD and KRAL can use less data to establish a better model than pool-based active learners.

And we can see that the prediction accuracy of RF and SVM is different. This is because RF is an ensemble learning method, and its base learner is a decision tree. A decision tree is not a linear regression or classification method, it can be applied to different types of datasets. At the same time, the use of ensemble learning and certain randomness make RF have stronger generalization ability. In this experiment, we use linear SVM, which form is simple and cannot deal with the complex data structure. And SVM does not use ensemble methods. Therefore, the effect of SVM is weaker than that of RF in this experiment.

However, if we only use the accuracy and SRI to evaluate SUD and KRAL, we cannot tell the difference between the two algorithms. However, as we mentioned, the SUD uses the whole unlabeled data to calculate its K-nearest density. But using whole unlabeled data to calculating density will cause the increase of computational complexity. In order to evaluate the complexity, we use the time consuming of every algorithm. Figure 7 shows that the KRAL's complexity is strongly lower than SUD. SUD is the time consumed

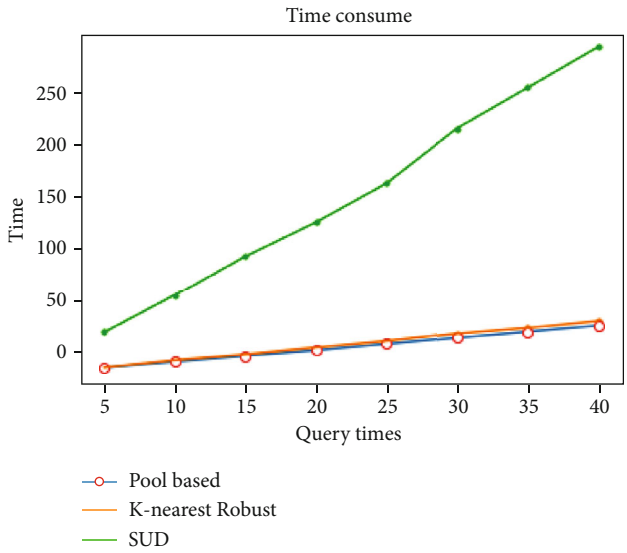



Figure 7: Time consumption in random forest.

when compared to the KRAL and the basic pool-based active learners, which means KRAL is more fit to deal with the big data problem than the SUD.

It can be seen from the experiment that the performance of the KRAL algorithm is better than pool-based active learning in accuracy and robustness and SUD in computational complexity. Therefore, a more robust and accurate method with less computational complexity is obtained, especially when KRAL is applied to outlier-sensitive models like SVM. This is because KRAL can select information data instances with fewer outliers. And the RF is not so sensitive to the outliers, which may reduce the effect of KRAL and SUD. Therefore, considering accuracy, robustness, and computation complexity in big data analysis, this paper uses the KRAL algorithm with random forest to predict B-cell data epitopes. The accuracy of the model obtained in this paper is 93.8%, and only 4/10 of the samples are used.

## 8. Conclusion and Discussion

The contribution of this paper to big data analysis is to propose a new more robust active learning method with higher accuracy and a new active learning robustness evaluation metric SRI. The new algorithm can also reduce the complexity of density-weighted pool-based active learners like SUD when facing the big data. And the effectiveness and robustness of KRAL, SUD, and pool-based active learning are evaluated experimentally by the SRI. Through the experiments, a more robust and accurate algorithm with less complexity is obtained. Apart from the computational complexity, KRAL has some advantages in big data area when compared with the SUD algorithm. More specifically, KRAL eliminates outliers by estimating sample density for better performance. However, SUD only uses a new loss function to change the structure of the model. Hence, KRAL has greater potential. This is because scholars can change the proportion of deleted samples before adding them to the training data. Specifically, using a dynamic greedy algorithm with a reasonable loss function to improve KRAL's performance is a prospective direction. Therefore, when the basic learner is not sensitive to outliers, the algorithm can achieve better results. However, SUD cannot use this method to improve performance. But the KRAL also has some disadvantages: KRAL still uses the uncertainty query strategies for searching the most valuable samples. However, this may be not fit in many areas such as the natural language processing (NLP). Therefore, changing the query strategies to fit these areas is a good direction. In the future, the author will look for a good loss function to improve the performance of KRAL and look for some new query strategies for active learning and make more contributions in big data and artificial intelligence area. To be more specific, the author will devote himself into the NLP area and find more suitable query strategies to let the active learning method more effective in such as Neural Machine Translation (NMT) problem. And the author will conduct some research in bioinformatics to find some cure to kinds of diseases.

## Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Conflicts of Interest

The author declares that there are no conflicts of interest concerning the publication of this paper.

## Acknowledgments

The author would like to thank Jiatong Shi, who gives some specific and significant suggestions to this paper.

## References

[1] D. Guru and S. Perumal, "Approaches towards blockchain innovation: a survey and future directions," *Electronics*, vol. 10, no. 10, pp. 1219–1219, 2021.

[2] N. Deepa, Q. V. Pham, and D. C. Nguyen, "A survey on blockchain for big data: approaches, opportunities, and future directions," 2020.

[3] A. R. Javed, M. U. Sarwar, M. O. Beg, M. Asim, T. Baker, and H. Tawfik, "A collaborative healthcare framework for shared healthcare plan with ambient intelligence," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–21, 2020.

[4] M. U. Sarwar and A. R. Javed, "Collaborative health care plan through crowdsource data using ambient application," in *2019 22nd International Multitopic Conference (INMIC)*, pp. 1–6, Islamabad, Pakistan, 2019.

[5] A. Arj, B. Lgf, and B. Aaf, "Automated cognitive health assessment in smart homes using machine learning," *Sustainable Cities and Society*, vol. 65, 2021.

[6] T. Noumi, S. Inoue, H. Fujita et al., "Epitope prediction of antigen protein using attention-based LSTM network," *Journal of Information Processing*, vol. 29, pp. 321–327, 2021.

[7] R. Vita, J. A. Overton, J. A. Greenbaum et al., "The immune epitope database (IEDB) 3.0," *Nucleic Acids Research*, vol. 43, no. D1, pp. D405–D412, 2015.

[8] H. M. Regenmortel, "The concept and operational definition of protein epitopes," *Philosophical Transactions of the Royal Society of London*, vol. 323, no. 1217, pp. 451–466, 1989.

[9] J. Rux, "Type-specific epitope locations revealed by X-ray crystallographic study of adenovirus type 5 hexon," *Molecular Therapy the Journal of the American Society of Gene Therapy*, vol. 1, no. 1, pp. 18–30, 2000.

[10] M. Mayer and B. Meyer, "Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor," *journal of the american chemical society*, vol. 123, no. 25, pp. 6108–6117, 2001.

[11] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili, "BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes," *Nucleic Acids Research*, vol. 45, no. W1, pp. W24–W29, 2017.

[12] H. Singh, H. R. Ansari, and G. P. S. Raghava, "Improved method for linear B-cell epitope prediction using antigen's primary sequence," *PLoS One*, vol. 8, no. 5, 2013.

[13] D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, no. 4, pp. 319–342, 1988.

[14] C. C. Bonwell, *Active Learning: Creating Excitement in the Classroom. ERIC Digest*, ERIC Clearinghouse on Higher Education, Washington, DC, 1991.

[15] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," Heidelberg: Springer Verlag, Berlin, 1994.

[16] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[17] P. Juszczak and R. Duin, "Selective sampling based on the variation in label assignments," in *ICPR 2004. Proceedings of the 17th International Conference on. 2004*, 2004.

[18] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," Association for Computing Machinery, New York, NY, USA, 1992.

[19] F. Sebastiani, "Representative sampling for text classification using support vector machines," *Lecture Notes in Computer Science Advances in Information Retrieval*, vol. 2633, pp. 393–407, 2003.

[20] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 2001.

[21] S. C. H. Hoi, R. Jin, and J. Zhu, "Batch mode active learning and its application to medical image classification," Association for Computing Machinery, New York, NY, USA, 2006.

[22] X. Zhu, *Semi-Supervised Learning with Graphs*, PhD thesis, Carnegie Mellon University, 2005.

[23] B. Settles, M. Craven, and L. Friedland, "Active learning with real annotation costs," *Proceedings of the NIPS Workshop on Cost-Sensitive Learning.*, vol. 1, 2008.

[24] A. W. Naik, J. D. Kangas, D. P. Sullivan, and R. F. Murphy, "Active machine learning driven experimentation to determine compound effects on protein patterns," *eLife*, vol. 5, 2016.

[25] T. Nakano, S. Takeda, and J. Brown, "Active learning effectively identifies a minimal set of maximally informative and asymptotically performant cytotoxic structure–activity patterns in nci-60 cell lines," *RSC Medicinal Chemistry*, vol. 11, no. 9, pp. 1075–1087, 2020.

[26] M. Hafner, M. Niepel, K. Subramanian, and P. K. Sorger, "Designing drugresponse experiments and quantifying their results," *Current Protocols in Chemical Biology*, vol. 9, no. 2, pp. 96–116, 2017.

[27] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[28] C. Campbell, N. Cristianini, and A. J. Smola, "Query learning with large margin classifiers," Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 2000.

[29] Y. J. Gu, D. Zydek, and Z. Jin, "Active learning based on random forest and its application to terrain classification," in *Progress in Systems Engineering*, Advances in Intelligent Systems and Computing, volume 366, Springer, Cham, 2015.

[30] J. Zhu, H. Wang, B. K. Tsou, and M. Ma, "Active learning with sampling by uncertainty and density for data annotations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1323–1331, 2010.

[31] J. Zhu, H. Wang, and T. Yao, "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification," Coling 2008 Organizing Committee, Manchester, UK, 2008.

[32] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1953.

[33] Breiman, "Random forests," *MACH LEARN*, vol. 45, no. 1, pp. 5–32, 2001.

[34] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[35] M. Tang, X. Luo, and S. Roukos, "Active learning for statistical natural language parsing," *USA: Association for Computational Linguistics.*, pp. 120–127, 2002.

[36] J. Zhu and E. H. Hovy, "Active learning for word sense disambiguation with methods for addressing the class imbalance problem," in *Conference on Emnlp-conll*, DBLP, 2007.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 7439744, 15 pages
https://doi.org/10.1155/2021/7439744

*Research Article*

# Industrial Efficiency Algorithm Based on Spatio-Temporal-Data-Driven

**Hongqu Lv and Wensi Cheng**

*Shandong University of Political Science and Law, Jinan, China*

Correspondence should be addressed to Wensi Cheng; chengwensi777@163.com

Received 9 August 2021; Revised 3 September 2021; Accepted 6 September 2021; Published 8 November 2021

Academic Editor: Rajesh Kaluri



Stochastic frontier model is an important and effective method to calculate industry efficiency. However, when dealing with temporal and spatial data from the industry, it is difficult to accurately calculate the industrial production efficiency due to the influence of spatial correlation and time lag effect. If the traditional spatial statistical method is used, the setting method of spatial weight matrix is often questioned. To solve this series of problems, one possible idea is to design a spatial data mining process based on stochastic frontier analysis. Firstly, the stochastic frontier model should be improved to analyze spatio-temporal data. In order to accurately measure the technical efficiency in the case of dual correlation between time and space, a more effective spatio-temporal stochastic frontier model method is proposed. Meanwhile, based on the idea of generalized moment estimation, an estimation method of spatiotemporal stochastic frontier model is designed, and the consistency of estimators is proved. In order to ensure that the most appropriate spatial weight matrix can be selected in the process of model construction, the $K$-fold crossvalidation method is adopted to evaluate the prediction effect under the data-driven idea. This set of spatio-temporal data mining methods will be used to measure the technical efficiency of high-tech industries in various provinces of China.

## 1. Introduction

Stochastic frontier analysis (SFA) is an important method to measure technical efficiency and calculate total factor productivity. The whole process is divided into two steps: the first step is the model estimation process, which can be regarded as a supervised learning process; the second step is to use the estimated model to calculate the technical efficiency, which can be regarded as an unsupervised learning process.

From the perspective of machine learning, supervised processes have three main objectives: (a) feature selection and reduction of the dimension of feature variables; (b) selecting the optimal one from multiple classifiers or prediction models; (c) model evaluation, which estimates the prediction error of the selected classifier or prediction model on the new data.

The paper found that the traditional stochastic frontier analysis method has the following defects: (a) it is not suitable for the special structure of spatial data or spatio-temporal data; (b) the modeling process lacks variety. The traditional analysis process is knowledge-driven and completely relies on a single theoretical model for estimation and testing. The above two characteristics lead to the large deviation of the traditional stochastic frontier model when analyzing the spatio-temporal data, and it is impossible to make an accurate measure of the production efficiency with spatial relationship, either. To solve the two problems above, this study considers two improvements to the industrial efficiency calculation process based on temporal and spatial data: (1) improve the existing stochastic frontier model and make it suitable for spatial data or spatio-temporal data; (2) turn the modeling process into a spatial data mining process. In view of the unique structure of spatio-temporal data, a more suitable crossvalidation method is proposed for the selection of prediction model.

Stochastic frontier analysis (SFA) was successively proposed by Aigner et al. (1977) [1], Meeusen and Broeck (1977) [2], and Battese and Corra (1977) [3]. Over the past

40 years, its theoretical system and methods have been continuously expanded and innovated; it is widely used to measure the operating efficiency of different industries.

The development of spatial statistics provides a theoretical basis for studying spatial interactions in stochastic frontier models. Druska and Horrace (2004) [4] first applied the method of spatial econometrics to the analytical framework of stochastic frontier model and started the research of spatial stochastic frontier model. Affuso (2010) [5] established the spatial stochastic frontier model and gave the maximum likelihood estimation in the empirical study. Tonini and Pede (2011) [6] applied maximum entropy method to parameter estimation of spatial stochastic frontier model. Vidolia et al. (2016) [7], Tsionas and Michaelides (2016) [8], Carvalho (2018) [9], and Adetutu et al. (2015) [10] consider SF models with local spatial dependence. Jin and Lee (2020) [11] proved the asymptotic properties of a maximum likelihood estimator of a spatial autoregressive stochastic frontier model. Kutlu et al. (2020) [12] proposed a spatial autoregressive stochastic frontier model, which allows for the endogeneity in both the frontier and environmental variables, and discussed a single-stage control function approach to estimate the parameters.

Because spatial stochastic frontier analysis methods can fully consider the impact of spatial correlation, they can obtain more accurate results in efficiency analysis of data with spatial spillover effect and thus have been more widely used in recent years. Bergantino et al. (2020) [13] analyses the potential impact of airport competition on technical efficiency by applying the spatial stochastic frontier. Graaff (2020) [14] used spatial stochastic frontier model to estimate spatially correlated technical efficiencies within a European regional production function context. At present, some literatures have studied panel spatial stochastic frontier model, for example, Druska and Horrace (2004) [4], Tonini and Pede (2011) [6], and Lin Jia-Xian (2014) [15]. These literatures all focus on the static panel space stochastic frontier model, and the model utilizes two-dimensional information from panel data; formally, the spatial lag term of the explained variable and the spatial lag term of the error are used to capture the spatial correlation of the production unit. The time lag term is not included in the model, which means that the model still cannot fit well when there is significant inertia in the research problem. In input-output analysis, current behavior is largely dependent on past behavior, for example, the adjustment of capital stock is often influenced by previous capital. Therefore, a dynamic stochastic frontier model should be established, and the model should describe the double lag effect of space and time, so as to reflect the influence relationship between economic variables more objectively. The spatial weight matrix in spatial statistics is often considered to be "subjective." Moreover, due to the various setting methods of spatial weight matrix, the selection of different spatial weight matrix may lead to the difference of model estimation results. In addition, the selection of spatial weight matrix has not formed a unified principle. Based on the above three points, the spatial weight matrix is often questioned. But in the era of "big data," such skepticism may end [16].

This paper proposes the spatiotemporal stochastic frontier model; considering that the model may be endogenous in time and space dimensions, a generalized method of moments (GMM) estimation process is designed to estimate the model. When Druska and Horrace (2004) [4] studied the static panel space stochastic frontier model, a generalized moment estimation process was proposed by referring to Kelejian and Prucha (1999) [17] for spatial error correlation. In this paper, Druska and Horrace (2004) [4] is used to deal with model's error space autocorrelation, which is different from that of the stochastic frontier model. According to the method of Kapoor et al. (2007) [18], the compound error term was processed, and the moment condition was constructed to estimate the distribution parameters of the error term. In this paper, Jacobs et al. (2009) [19] was used as a reference to construct the moment condition, and Anselin (1988) [20] was used as a reference for the selection of tool variables to obtain the generalized moment estimator. Furthermore, the consistency of the obtained structural parameter estimators is proved by using the extreme value consistency theorem and the law of uniform large numbers (ULLN). To solve the problem of selecting spatial weight matrix, we can consider a crossvalidation method suitable for spatio-temporal data. Fortunately, a series of methods such as dimensionality reduction, feature selection, and model generalization has been provided by machine learning methods. The earliest crossvalidation method was called hold-out, which relied on only one partition of the data, and there was no crossover process, so it was also called the verification method [21]. Noting that the hold-out method relies on a partition of data and is easily affected by contingency factors, Geisser (2010) [22] proposed a crossvalidation method that includes the average of multiple hold-out estimates, realizing the transition from verification estimation to crossvalidation estimation. In order to reduce the combination number of data partition in crossvalidation, Shao (1993) [23] proposed the leave-$P$-out crossvalidation (LPOCV) in which the number of test samples in each data partition was the same. Especially in the special case when $P = 1$, the method is evolved to leave-one-out crossvalidation (LOOCV). LOOCV is the simplest and most widely used crossvalidation in traditional analysis. Compared with the LPOCV considering all data partitioning, Geisser (2010) also proposed a crossvalidation based on only partial data partitioning, which is called RLT method. $K$-folded crossvalidation is proposed as an alternative to LOOCV which has a large computational overhead and relies on a basic partition of data divided into $K$-fold, each of which has a data capacity of $N/K$. In the case of limited samples, $k$-fold crossvalidation is the simplest and most widely used method of generalization error estimation. From the various crossvalidation methods that have appeared in the past, each method fully considers the randomness of the validation set to ensure the generalization ability of the test model. However, for the special panel data such as spatio-temporal data, there is usually an internal connection between spatial individuals, and the overall data also tends to have time trend. This problem is not taken into account by the previous crossvalidation methods, which may break the inherent regularity of spatio-

temporal data. Based on the above considerations, this paper designs a kind of crossvalidation scheme suitable for spatiotemporal data. It is used to select stochastic frontier models, especially models with different weight matrices.

Finally, the technology efficiency of China's high-tech industry is analyzed by establishing a spatiotemporal stochastic frontier model.

## 2. Methodology

Previous studies on panel spatial stochastic frontier models mainly involved static panel spatial stochastic frontier models. Spatial lag effect is considered in the process of model building, but the influence of time lag effect is not included. If the time lag term and time-spatial lag term are added into the model, this kind of model can be called spatiotemporal stochastic frontier model. Obviously, the time-space double lag effect will produce stronger endogeneity, and new estimation methods should be considered to solve it.

### 2.1. Model Specification and Assumption

*2.1.1. Model Specification.* The general form of the spatiotemporal stochastic frontier model can be stated in matrix form as

$$\begin{aligned} Y_t &= \lambda_1 W Y_t + \lambda_2 W Y_{t-1} + \gamma Y_{t-1} + \mathrm{X}_t B + E_t, \\ E_t &= \rho M E_t + \varepsilon_t, \\ \varepsilon_t &= \nu_t - u, \end{aligned} \tag{1}$$

where $Y_t$, $E_t$, $\varepsilon_t$, and $\nu_t$ are $N$-dimensional vectors, whose components at time $t = 1, \cdots, T$ are given by $Y_t = [y_{1t}, \cdots, y_{Nt}]'$, $E_t = [E_{1t}, \cdots, E_{Nt}]'$, $\varepsilon_t = [\varepsilon_{1t}, \cdots, \varepsilon_{Nt}]'$, and $\nu_t = [\nu_{1t}, \cdots, \nu_{Nt}]'$. The vector $Y_t$ consists of the outputs of the $N$ production units, $E_t$ and $\varepsilon_t$ are the composite error vectors corresponding to $Y_t$, $\nu_t$ is the heterogeneous error vector, and $u = [u_1, \cdots, u_N]'$ is the vector of time-invariant inefficiency terms. This kind of setting is appropriate when the time span is not large. As $u$ is time invariant, it can be regarded as the individual effect, and thus, this paper primarily considers $u$ as a fixed effect. $X_t$ is an $N \times K$-dimensional matrix consisting of the $K$ exogenous input variables of the $N$ production units at time $t$. $W$ and $M$ are $N \times N$ spatial weight matrices which are usually assumed to be different. If $W = M$, $\lambda_1$ and $\rho$ cannot be distinguished by means of the maximum likelihood method although they can be effectively distinguished by the GMM method [19]. The variables are stacked according to the section and time series in the following matrix form:

$$\begin{aligned} Y &= \lambda_1 (I_T \otimes W) Y + \lambda_2 (I_T \otimes W) Y_{-1} + \gamma Y_{-1} + XB + E, \\ E &= \rho (I_T \otimes M) E + \varepsilon, \\ \varepsilon &= \nu - (e_T \otimes I_N) u, \end{aligned} \tag{2}$$

where $\otimes$ represents the Kronecker product of matrices, $I_T$ and $I_N$ are, respectively, the identity matrices of orders $T$ and $N$, and $e_T$ is a $T$-dimensional column vector with all the entries equal to 1. The parameter vector of the model to be estimated is $(\lambda_1, \lambda_2, \gamma, B, \rho, \sigma_\nu^2, \sigma_u^2)$, and its dimension is $(K + 6)$, where $B$ is the parameters corresponding to the $K$-dimension exogenous explanatory variables $X_t$. $B$ and $\lambda_1$, $\lambda_2$, $\gamma$ together constitute the structural parameters, and $\rho$, $\sigma_\nu^2$, $\sigma_u^2$ are the error term parameters.

*2.1.2. Model Assumption.* The assumptions of the spatiotemporal stochastic frontier model are the following:

*Assumption 1.* The distribution of the error vector $\nu$ is given by $\nu \sim N(0, \sigma_\nu^2 I_{NT})$.

*Assumption 2.* The inefficiency term $u$ is time invariant, with distribution $u \sim N^+(0, \sigma_u^2 I_{NT})$.

*Assumption 3.* $0 < \sigma_\nu^2 < b_\nu < \infty$, $0 < \sigma_u^2 < b_u < \infty$, and $\nu$ and $u$ have finite fourth moment.

*Assumption 4.* $\nu$ and $u$ are uncorrelated with $X$.

*Assumption 5.* The spatial weight matrices $W$ and $M$ satisfy $w_{ii} = 0$ and $m_{ii} = 0$; $i = 1, 2, \cdots, N$. For arbitrary $|\lambda_1| < 1$, $|\lambda_2| < 1$, and $|\rho| < 1$, the matrices $1 - \lambda_1 W$, $1 - \lambda_2 W$, and $1 - \rho M$ are all nonsingular matrices. For each of the matrices $W$, $M$, $1 - \lambda_1 W$, $1 - \lambda_2 W$, $1 - \rho M$, $(1 - \lambda_1 W)^{-1}$, $(1 - \lambda_2 W)^{-1}$, and $(1 - \rho M)^{-1}$, the row sums and column sums are all absolutely uniformly bounded.

Assumption 1 is a classic assumption of the spatial error autocorrelation model. By Assumption 2, the same individual inefficiency term remains constant at different times. When using GMM to estimate the structural parameters of the model, the distribution of error terms can be ignored; nevertheless, in order to improve the efficiency of computation, the half normal distribution for the inefficiency term is usually assumed. Assumption 3 ensures the boundedness of the variance of the error term in this model, which is an important condition for the consistency of the estimator. Assumption 4 is a classical assumption commonly used in traditional regression analysis methods, and the moment condition is set according to this assumption in the generalized moment estimation of this model. Assumption 5 is set according to the space weight matrix of this model and the properties of space station autoregressive coefficient and space-time autoregressive coefficient, which also ensures the consistency of parameter estimators.

### 2.2. Parameter Estimation

*2.2.1. Estimation strategy.* There is an endogeneity problem in the model. For the spatial lag term $\lambda_1 W Y_t$ in the model, there is

$$\begin{aligned} COV[(WY_t), E_t] &= COV\{[WA^{-1}(\lambda_2 WY_{t-1} \\ &\quad + \gamma Y_{t-1} + X_t B + E_t)], E_t\} \\ &= \lambda_2 COV\{[WA^{-1}(WY_{t-1})], E_t\} + \gamma COV \\ &\quad \cdot \{[WA^{-1}Y_{t-1}], E_t\} + COV\{[WA^{-1}]E_t, E_t\}, \end{aligned} \tag{3}$$

where $A = I - \lambda_1 W$, and $COV\{[WA^{-1}(WY_{t-1})], E_t\} = 0$, $COV\{[WA^{-1}Y_{t-1}], E_t\} = 0$, and $COV\{WA^{-1}[X_t B], E_t\} = 0$ can be obtained from the assumptions of the regression model, and $COV\{[WA^{-1}]E_t, E_t\}$ is quadratic. In spatial econometrics, $W$ is usually not a zero matrix, and so, $WA^{-1}$ is not a zero matrix. While taking into account the expected value of the compound error term cannot be 0, it can be considered that the quadratic form $COV\{[WA^{-1}]E_t, E_t\}$ is almost impossible to be equal to 0 (see Appendix A for proof). Therefore, in the dynamic panel spatial stochastic frontier models, there is an endogeneity problem which will lead to the inconsistency of traditional estimators. So, we considered GMM as a good way to solve the endogeneity problem.

The parameter vector to be estimated in the model is $(\lambda_1, \lambda_2, \gamma, B, \rho, \sigma_v^2, \sigma_u^2)'$, where $\lambda_1$, $\lambda_2$, $\gamma$, and $B$ are the structural parameters of the model, and $\rho$, $\sigma_v^2$, and $\sigma_u^2$ are the error distribution parameter of the model. The estimation of the model is completed in three steps:

*Step 1.* Using the GMM to estimate the structure parameter $(\lambda_1, \lambda_2, \gamma, B)$ in the model.

*Step 2.* Making a moment estimation of the parameter $(\rho, \sigma_v^2, \sigma_u^2)$ that is included in the error term.

*Step 3.* Using the estimator obtained in Step 2 to modify the result of Step 1.

*2.2.2. Estimation of Structural Parameter* $(\lambda_1, \lambda_2, \gamma, B)$

*(1) Difference Model and Level Model.* Anderson and Hsiao (1981) [24] proposed to use $y_{i,t-2}$ as the instrumental variable of $\Delta y_{i,t-1}$, and then, 2SLS estimation is carried out. This estimator is called "Anderson-Hsiao estimator." According to the same logic, lag variables of higher order are also valid IV. Arellano and Bond (1991) [25] used all possible lag variables as IV (the number of IVs is more than the number of endogenous variables) to conduct GMM estimation. This GMM estimator is called Arellano-Bond estimator or difference GMM. The disadvantage of difference GMM is that the variable which does not change with time is eliminated, and its coefficient cannot be estimated. If the series $\{y_{i,t}\}$ has a strong persistence, that is, the first-order autoregressive coefficient is close to 1, then the correlation may be very weak and lead to the problem of weak instrumental variables. In order to solve the above two problems, Arellano and Bver (1995) [26] returned to the level equation and used $\{\Delta y_{i,t-1}, \Delta y_{i,t-2}, \cdots\}$ as IV to estimate the GMM of the level equation, which was called "level GMM." Blundell and Bond (1998) [27] combined difference GMM with level GMM and estimated the difference equation and level equation as one equation system for GMM, which was called "system GMM." The advantage of system GMM is that it can improve the efficiency of estimation (small sample properties are better), and it can estimate the variable that does not change with time (the system GMM contains the level equation). In order to solve the endogenous problem of dynamic panel data model, Arellano and Bond (1991) [25], Arellano and Bover (1995) [26], and Blundell and Bond (1998) [27], respectively, considered from the perspective of difference model and level model, and different instrumental variables were selected.

The corresponding difference model and level model of Equation (1) are simplified as

$$\Delta Y_t = \Delta Z_t \theta + \Delta E_t, \tag{4}$$

$$Y_t = Z_t \theta + E_t. \tag{5}$$

(4) and (5) can also be collectively called spatial system model, where Equation (4) is the difference model, and Equation (5) is the level model, $Z_t = [WY_t, WY_{t-1}, Y_{t-1}, X_t]'$ is the vector composed of all explanatory variables, and $\theta = [\lambda_1, \lambda_2, \gamma, B]'$ is the vector composed of structural parameters. The expansion of Equation (5) is Equation (1); the expansion of Equation (4) can be expressed as follows:

$$\begin{aligned} \Delta Y_t &= \lambda_1 W \Delta Y_t + \lambda_2 W \Delta Y_{t-1} + \gamma \Delta Y_{t-1} + \Delta X_t B + \Delta E_t, \\ E_t &= \rho M \Delta E_t + \Delta \varepsilon_t, \\ \Delta \varepsilon_t &= \Delta \nu_t. \end{aligned} \tag{6}$$

*(2) Moment Condition and Instrumental Variable.* Since $\Delta X_t$ is a strictly exogenous variable, it is not related to the compound error term $\Delta E_t$, nor is it related to $E_t$. The moment conditions for identifying $B$ in the difference model and the level model are as follows:

$$\begin{aligned} E\left(\Delta X_t{}' \Delta E_t\right) &= 0, \quad t = 3, \cdots, T, \\ E\left(\Delta X_t{}' E_t\right) &= 0, \quad t = 3, \cdots, T. \end{aligned} \tag{7}$$

The moment condition structure for identifying $\lambda_2$ and $\gamma$ in the two models is as follows: since the spatial lag term and time lag term of the dependent variable $\Delta Y_t$ are both endogenous variables, therefore, it is necessary to find a set of instrumental variables that is related to time lag and space lag and exogenous explanatory variables, but not related to the difference error term $\Delta E_t (t = 3, \cdots, T)$. Arellano and Bond (1991) [25] uses all possible level lag variables $(y_{t-2}, \cdots, y_1)$ of $Y_t$ as instrumental variables for the time-lag first-order difference term $(\Delta Y_{t-1})$ of the dependent variable. These instrumental variables are related to $(\Delta Y_{t-1})$, but not to $\Delta E_t$. The moment conditions corresponding to the difference model and the level model are as follows:

$$E\left(Y_{t-s}{}'\Delta E_t\right)=0,\quad t=3,\cdots,T;s=2,\cdots,t-1, \tag{8}$$

$$E\left(\Delta Y_{t-s}{}'E_t\right)=0,\quad t=3,\cdots,T;s=2,\cdots,t-1. \tag{9}$$

The moment conditions for identifying $\lambda_1$ in the two models are as follows.

Construct a spatial lag item $WY_t$ as follows; Jacobs et al. (2009) [19] provided a method of finding instrumental variables, that is time lag terms of spatial lag dependent variables, who also proved that the moment condition obtained by this method was as valid as Equation (8). So, corresponding to the difference model and the level model, the following moment conditions can be listed:

$$\begin{aligned}&E\left(\left\{W^lY_{t-s}\right\}'\Delta E_t\right)=0,\quad t=3,\cdots,T;s=2,\cdots,t-1;l=1,\cdots,L,\\&E\left(\left\{W^l\Delta Y_{t-s}\right\}'E_t\right)=0,\quad t=3,\cdots,T;s=1,\cdots,t-2;l=1,\cdots L,\end{aligned} \tag{10}$$

where $l$ is the exponential of matrix $W$ and the integer $L$ is the maximum order of spatial lag that can be used as the instrumental variable.

In addition, based on the method provided by Kelejian and Robinson (1993) [28], formula (1) shows that $WY_t$ depends on $WX_t$, so the instrumental variable $W\Delta X_t$ can be selected by the first-order difference method for $W\Delta Y_t$. Since $\Delta X_t$ is a strictly exogenous variable, it is not related to the compound error term $\Delta E_t$, so corresponding to the difference model and the level model, the instrumental variables satisfy the following moment conditions:

$$\begin{aligned}E\left(\left\{W^l\Delta X_t\right\}'\Delta E_t\right)=0,\quad t=3,\cdots,T,\\E\left(\left\{W^l\Delta X_t\right\}'E_t\right)=0,\quad t=3,\cdots,T.\end{aligned} \tag{11}$$

*(3) GMM Estimation*. When we estimate the parameters of the spatio-temporal stochastic frontier model, we use the system GMM method similar to the general dynamic panel model to construct the spatial system GMM estimation. Unlike the system GMM, the IVs of the spatial system GMM are composed of time lag variable and spatial lag variable.

For each period of $t$, the moment condition of $J \geq K+2$ can be given. The moment conditions corresponding to the difference model and the level model can be abbreviated as

$$\begin{aligned}E\left(H_{N,ABt}{}'\Delta E_t\right)=0,\\E\left(H_{N,Lt}{}'E_t\right)=0.\end{aligned} \tag{12}$$

The matrices $H_{N,ABt}$ and $H_{N,Lt}$ are expressed as follows

$$\begin{aligned}Y_{t-s}&=\begin{bmatrix}y_1&0&0&0&0&0&\cdots&\cdots&0&0&\cdots&0\\0&y_2&y_1&0&0&0&\cdots&\cdots&0&0&\cdots&0\\0&0&0&y_3&y_2&y_1&\cdots&\cdots&0&0&\cdots&0\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\cdots&\cdots&0&0&\cdots&0\\0&0&0&0&0&0&\cdots&\cdots&y_{t-2}&y_{t-3}&\cdots&y_1\end{bmatrix},X_t=\begin{bmatrix}x_3\\x_4\\\vdots\\x_t\end{bmatrix},\\\Delta Y_{t-s}&=\begin{bmatrix}\Delta y_1&0&0&0&0&0&\cdots&\cdots&0&0&\cdots&0\\0&\Delta y_2&\Delta y_1&0&0&0&\cdots&\cdots&0&0&\cdots&0\\0&0&0&\Delta y_3&\Delta y_2&\Delta y_1&\cdots&\cdots&0&0&\cdots&0\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\cdots&\cdots&0&0&\cdots&0\\0&0&0&0&0&0&\cdots&\cdots&\Delta y_{t-2}&\Delta y_{t-3}&\cdots&\Delta y_1\end{bmatrix},\Delta X_t=\begin{bmatrix}\Delta x_3\\\Delta x_4\\\vdots\\\Delta x_t\end{bmatrix}.\end{aligned} \tag{13}$$

Then, $H_{N,ABt}=[Y_{t-s},WY_{t-s},W^2Y_{t-s},\cdots,W^LY_{t-s},W\Delta X_t,\Delta X_t]$ and

$$H_{N,Lt}=\left[\Delta Y_{t-s},W\Delta Y_{t-s},W^2\Delta Y_{t-s},\cdots,W^L\Delta Y_{t-s},WX_t,X_t\right](t=3,\cdots,T). \tag{14}$$

That is, $H_{N,ABt}$ and $H_{N,Lt}$ are matrices with instrumental variables as column vectors, and the subscript $N$ means that the matrix depends on the unit number of individuals. Let $H_{N,AB}$ and $H_{N,L}$ be block diagonal matrices composed of block $H_{N,ABt}$ and $H_{N,Lt}$, respectively. In order to define the GMM (Spatial Blundell Bond, SBB) estimator of the spatial dynamic panel stochastic frontier model, the difference variables and level variables are combined to define the matrix as follows:

$$Y*_N=\left[\Delta Y'_N,Y'_N\right],Z*_N=\left[\Delta Z'_N,Z'_N\right],E*_N=\left[\Delta E'_N,E'_N\right]. \tag{15}$$

The instrument variable matrix is

$$H_{N,BB} = \text{diag}\left\{H_{N,AB}, H_{N,L}\right\} = \begin{bmatrix} H_{N,AB} & 0 \\ 0 & H_{N,L} \end{bmatrix}, \tag{16}$$

where $H_{N,AB}$ is the instrumental matrix of spatial difference GMM estimation, and $H_{N,L}$ is the instrumental matrix of spatial level GMM estimation. The weight matrix is

$$G_{N,BB} = \text{diag}\left\{G_{N,AB}, I_{T-2} \otimes I_N\right\} = \begin{bmatrix} G_{N,AB} & 0 \\ 0 & I_{T-2} \otimes I_N \end{bmatrix}. \tag{17}$$

This diagonal of the matrix is composed of the weight matrix defined in the process of spatial difference GMM estimation and an identity matrix, where $G_{N,AB} = I_N \otimes G$ is $N(T-2) \times N(T-2)$ weight matrix which elements are

$$G_{ij} = \begin{cases} 2, & i = j, \\ -1, & i = j+1, \\ -1, & j = i+1, \\ 0, & others. \end{cases} \tag{18}$$

This weight matrix is proposed by Arellano and Bond (1991) [25] which is further define the weight matrix:

$$A_{N,BB} = \left[\frac{H'_{N,BB} G_{N,BB} H_{N,BB}}{N}\right]^{-1}. \tag{19}$$

Through the above process, combining the spatial difference equation with the spatial level equation, we get the spatial system GMM estimation process. Get the objective function of generalized moment estimation for spatial system as follows:

$$\begin{aligned} &\frac{1}{N}\left(H'_{N,BB}\Delta E*_N\right)' A_{N,BB}\left(H'_{N,BB}\Delta E*_N\right) \\ &= \frac{1}{N}\left[H'_{N,BB}(\Delta Y*_N - \Delta Z*_N\theta)_N\right]' A_{N,BB} \\ &\cdot \left[H'_{N,BB}(\Delta Y*_N - \Delta Z*_N\theta)_N\right]. \end{aligned} \tag{20}$$

The one-stage SBB estimator of $\theta$ can be obtained by minimizing Equation (20):

$$\widehat{\theta}_{\text{SBB}} = \left(Z*'_N H'_{N,BB} A*_N H'_{N,BB} Z*_N\right) Z*'_N H'_{N,BB} A*_N H'_{N,BB} Y*_N. \tag{21}$$

Equation (21) can also be called the spatial system GMM estimator.

*(4) Improvement of Instrumental Variable Matrix.* The instrumental variable matrix constructed in accordance with the above method has a high dimension and grows exponentially as the values of $T$ and $L$ increase. In order to reduce the dimension of the instrumental variable matrix and avoid overfitting the instrumental variable, we can simplify it by using the "condensing instrumental variable matrix" proposed by Beck and Levine (2004) [29].

We still set $s = 2$ and $L = 1$ in the GMM instrumental variable matrix of the space system, and the corresponding condensed instrumental variable matrix is

$$H'_{N,BB} = \begin{bmatrix} H^1{}_{AB} & \cdots & H^N{}_{AB} \\ H^1{}_L & \cdots & H^N{}_L \end{bmatrix}, \tag{22}$$

where $H^i{}_L$ $(i = 1, \cdots, N)$ is the instrumental variable quantum matrix of the level model corresponding to the $i$ individual.

$$H^i{}_L = \begin{bmatrix} \Delta y_1 & 0 & 0 & \cdots & W\Delta y_1 & 0 & 0 & \cdots & W\Delta X_3 & \Delta X_3 \\ \Delta y_2 & \Delta y_1 & 0 & \cdots & W\Delta y_2 & W\Delta y_1 & 0 & \cdots & W\Delta X_4 & \Delta X_4 \\ \Delta y_3 & \Delta y_2 & \Delta y_1 & \cdots & W\Delta y_3 & W\Delta y_2 & W\Delta y_1 & \cdots & W\Delta X_5 & \Delta X_5 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \end{bmatrix}. \tag{23}$$

### 2.3. Estimation of Error Term Parameters

*2.3.1. Estimation of $\rho, \sigma_v^2$.* After the estimator of the structural parameter $\theta_N = (\lambda_1, \lambda_2, \gamma, B)'$ in model (1) is obtained in the first stage, the model residual $\widehat{E}_t = Y_t - Z_t\widehat{\theta}_N$ can be further obtained, in which the $Z_t = (WY_t, WY_{t-1}, Y_{t-1}, X_t)$ is the vector set composed of all explanatory variables in model (1). Consistent GMM estimation can be obtained by using residual $\widehat{E}_t$ and modifying the moment condition proposed by Kapoor et al. (2007) [18]. The specific process is as follows.

According to the assumptions of the model, the individual effect of the model is the inefficiency term. According to the covariance structure of the compound error term, it can be known that

$$Cov\left(\varepsilon\varepsilon'\right) = \sigma_u^2(I_N \otimes J_T) + \sigma_v^2 I_{NT}. \tag{24}$$

Introduce transformation matrix:

$$Q_0 = \left(I_T - \frac{J_T}{T}\right) \otimes I_N, Q_1 = \left(\frac{J_T}{T}\right) \otimes I_N, \tag{25}$$

where $I_T$ and $I_N$ are the identity matrix of order $T$ and order $N$, $J_T = e_T e_T{}'$ is the matrix of order $T \times T$, and the elements of that are all 1. Properties of transformation matrix $Q_0$: (i) $tr(Q_0) = N(T-1)$; (ii) $Q_0(e_T \otimes I_N) = 0$; (iii) $Q_0(I_T \otimes R_N) = (I_T \otimes R_N)Q_0$ (where $R_N$ is any $N \times N$ matrix). From the properties (ii), we can further deduce the special properties (iv) $Q_0\varepsilon = Q_0\nu$ of matrix $Q_0$ in this paper, where $\varepsilon$ and $\nu$ are the corresponding error terms in model (1).

Let

$$\bar{E} = (I_T \otimes M)E; \bar{\bar{E}} = (I_T \otimes M)\bar{E}; \bar{\varepsilon} = (I_T \otimes M)\varepsilon. \tag{26}$$

Then,

$$\varepsilon = E - \rho\bar{E}, \bar{\varepsilon} = \bar{E} - \rho\bar{\bar{E}}. \tag{27}$$

Based on the above transformation and referring to the first three of the six moment conditions given by Kapoor et al. (2007) [18] and related properties, the following three moment conditions are given in this paper:

$$\begin{aligned} E\left(\varepsilon' Q_0 \varepsilon\right) &= \sigma_v^2 N(T-1), E\left(\bar{\varepsilon}' Q_0 \bar{\varepsilon}\right) \\ &= \sigma_v^2 N(T-1) tr\left(M'M\right), E\left(\bar{\varepsilon}' Q_0 \varepsilon\right) = 0. \end{aligned} \tag{28}$$

To further integrate the above moment conditions, we can get that

$$E\begin{pmatrix} \frac{1}{N(T-1)}\varepsilon' Q_0 \varepsilon \\ \frac{1}{N(T-1)}\bar{\varepsilon}' Q_0 \bar{\varepsilon} \\ \frac{1}{N(T-1)}\bar{\varepsilon}' Q_0 \varepsilon \end{pmatrix} = \begin{pmatrix} \sigma_v^2 \\ \frac{1}{N}\sigma_v^2 tr\left(M'M\right) \\ 0 \end{pmatrix}. \tag{29}$$

Substitute Equations (26) and (27) into Equation (29) to obtain that

$$\begin{aligned} &\frac{1}{N}E\begin{pmatrix} \frac{2}{(T-1)}\bar{E}' Q_0 E & -\frac{1}{(T-1)}\bar{E}' Q_0 \bar{E} & 1 \\ \frac{2}{(T-1)}\bar{E}' Q_0 \bar{\bar{E}} & -\frac{1}{(T-1)}\bar{\bar{E}}' Q_0 \bar{\bar{E}} & tr\left(M'M\right) \\ \frac{1}{(T-1)}\left(\bar{E}' Q_0 \bar{\bar{E}} + \bar{E}' Q_0 \bar{E}\right) & -\frac{1}{(T-1)}\bar{E}' Q_0 \bar{\bar{E}} & 0 \end{pmatrix}\begin{pmatrix} \rho \\ \rho^2 \\ \sigma_v^2 \end{pmatrix} \\ &= \frac{1}{N}\begin{pmatrix} \frac{1}{(T-1)}E' Q_0 E \\ \frac{1}{(T-1)}\bar{E}' Q_0 \bar{E} \\ \frac{1}{(T-1)}E' Q_0 \bar{E} \end{pmatrix}. \end{aligned} \tag{30}$$

The residual $\widehat{E}_t = Y_t - Z_t\widehat{\theta}_N$ estimated in the first stage is substituted into $\bar{E}$ and $\bar{\bar{E}}$ in Equation (30) to obtain the sample moment equation. In the sample moment equation, the estimated value $\widehat{\rho}, \widehat{\sigma}_v^2$ of $\rho$ and $\sigma_v^2$ can be solved by the following objective function:

$$\begin{aligned} \left(\widehat{\rho}, \widehat{\sigma}_v^2\right) = \arg\min &\left\{\left[G_N\left(\rho, \rho^2, \sigma_v^2\right) - g_N\right]'\left[G_N\left(\rho, \rho^2, \sigma_v^2\right) - g_N\right]\right\}, \\ &\rho \in [-a, a], \sigma_v^2 \in [0, b_v], \end{aligned} \tag{31}$$

where

$$\begin{aligned} G_N &= \frac{1}{N}\begin{pmatrix} \frac{2}{(T-1)}\bar{E}\wedge' Q_0 \widehat{E} & -\frac{1}{(T-1)}\bar{E}\wedge' Q_0 \bar{\bar{E}} & 1 \\ \frac{2}{(T-1)}\bar{E}\wedge' Q_0 \bar{\bar{E}} & -\frac{1}{(T-1)}\overline{\bar{E}\wedge}' Q_0 \bar{\bar{E}} & tr\left(M'M\right) \\ \frac{1}{(T-1)}\left(\bar{E}\wedge' Q_0 \bar{\bar{E}} + \bar{E}\wedge' Q_0 \bar{\bar{E}}\right) & -\frac{1}{(T-1)}\bar{E}\wedge' Q_0 \bar{\bar{E}} & 0 \end{pmatrix}, g_N \\ &= \frac{1}{N}\begin{pmatrix} \frac{1}{(T-1)}E\wedge' Q_0 \widehat{E} \\ \frac{1}{(T-1)}\bar{E}\wedge' Q_0 \bar{\bar{E}} \\ \frac{1}{(T-1)}E\wedge' Q_0 \bar{\bar{E}} \end{pmatrix}. \end{aligned} \tag{32}$$

*2.3.2. Estimation of $\sigma_u^2$.* The fourth moment condition given by Kapoor et al. (2007) [18] is

$$E\left(\varepsilon' Q_1 \varepsilon\right) = NT\sigma_u^2 + N\sigma_v^2 = N\sigma_1^2, \tag{33}$$

where $\sigma_1^2 = T\sigma_u^2 + \sigma_v^2$. However, considering the characteristics of the stochastic frontier model, the compound error term $\varepsilon$ obeys the asymmetric distribution of the expected nonzero, so the moment condition (33) cannot be directly applied, and the following formula can be proved:

$$COV\left(\varepsilon' Q_1 \varepsilon\right) = NT\sigma_u^2 + N\sigma_v^2 = N\sigma_1^2. \tag{34}$$

The estimator of parameter $\sigma_1^2$ can be obtained by combining Equation (34) with Equation (31):

$$\begin{aligned}\widehat{\sigma}_1^2 &= \frac{1}{N} COV\left[\left(E\wedge - \rho\wedge\bar{E}\wedge\right)' Q_1\left(\widehat{E} - \widehat{\rho}\bar{\widehat{E}}\right)\right] \\ &= \frac{1}{(T-1)} COV\left(E\wedge' Q_1 \widehat{E}\right) - \frac{2}{(T-1)} COV\left(\bar{E}\wedge' Q_1 \widehat{E}\right) \\ &\quad \cdot \widehat{\rho} - \frac{1}{(T-1)} COV\left(\bar{E}\wedge' Q_1 \bar{\widehat{E}}\right)\rho\wedge^2.\end{aligned} \tag{35}$$

Substitute $\widehat{\sigma}_v^2$ and $\widehat{\sigma}_1^2$ into (33) to obtain the moment estimator $\widehat{\sigma}_u^2$ of $\sigma_u^2$.

*2.4. Spatial Correction of Estimators.* Although it can be proved that the estimator (21) is a consistent estimator, it can also be proved that the consistency of the GMM estimator can be guaranteed even if the model has spatial error autocorrelation. However, the estimator (21) cannot solve the spatial dependence of the error term, and the variance of the estimator is relatively large. After obtaining the consistent estimator of $\rho$ by Equation (31), the consistent estimator can be obtained by a correcting transformation. According to the spatial correction method given by Jacobs et al. (2009) [19], the estimator obtained in the first step was corrected.

The estimator $\widehat{\rho}$ obtained from Equation (31) is used to construct matrix $I - \widehat{\rho}M$, and left the difference GMM and the explained variables and the instrumental variables matrix of the system GMM estimation, if

$$\begin{aligned}&\tilde{Y} = (I - \widehat{\rho}M)Y, \Delta\tilde{Y}_N = (I - \widehat{\rho}M)\Delta Y_N, \tilde{Y}_{-1} \\ &= (I - \widehat{\rho}M)Y_{-1}, \tilde{W}Y = (I - \widehat{\rho}M)WY, \tilde{X} = (I - \widehat{\rho}M)X.\end{aligned} \tag{36}$$

The corresponding explanatory variable set $Z_t = [WY_t, WY_{t-1}, Y_{t-1}, X_t]$ is corrected as

$$\tilde{Z} = (I - \widehat{\rho}M)Z*_N, \Delta\tilde{Z} = (I - \widehat{\rho}M)\Delta Z_N. \tag{37}$$

The instrumental variable matrix and weight matrix corresponding to the GMM estimation of the spatial system are corrected as follows:

$$\begin{aligned}\tilde{H}_{\mathrm{SBB}} &= (I - \widehat{\rho}M)H_{N,BB}, \\ \tilde{A}_{\mathrm{SBB}} &= \left(\tilde{H}'_{\mathrm{SBB}}\tilde{H}_{\mathrm{SBB}}\right)^{-1}.\end{aligned} \tag{38}$$

Then, the corrected system GMM estimator is

$$\widetilde{\theta}_{\mathrm{SBB}} = \left(\tilde{Z}'_{\mathrm{SBB}}\tilde{H}_{\mathrm{SBB}}\tilde{A}_{\mathrm{SBB}}\tilde{H}'_{\mathrm{SBB}}\tilde{Z}_{\mathrm{SBB}}\right)\tilde{Z}'_{\mathrm{SBB}}\tilde{H}_{\mathrm{SBB}}\tilde{A}_{\mathrm{SBB}}\tilde{H}'_{\mathrm{SBB}}\tilde{Y}_{\mathrm{SBB}}. \tag{39}$$

## 3. Results and Discussion

### 3.1. Properties of the Estimators

*3.1.1. Properties of the Estimator $\widetilde{\theta}_{SBB}$.* According to the Extremum Consistency Theorem [30] (see Appendix B), the estimators $\widehat{\theta}_{\mathrm{SBB}}$ and $\widetilde{\theta}_{\mathrm{SBB}}$, obtained by Equations (21) and (27), are consistent.

*Proof.* see Appendix C. □

*3.1.2. Properties of the Estimators $\widehat{\rho}$, $\widehat{\sigma}_v^2$, $\widehat{\sigma}_u^2$, and $\widetilde{\theta}_{SBB}$.* It can be proved that the estimators $\widehat{\rho}$, $\widehat{\sigma}_v^2$, $\widehat{\sigma}_u^2$, and $\widetilde{\theta}_{\mathrm{SBB}}$ are consistent. The proof of the consistency of $\widehat{\rho}$, $\widehat{\sigma}_v^2$, and $\widehat{\sigma}_u^2$ is similar to that in Kapoor et al. (2007) [18], and it is omitted here. The consistency of $\widetilde{\theta}_{\mathrm{SBB}}$ can be derived from the consistency of $\widehat{\rho}$, $\widehat{\sigma}_v^2$, and $\widehat{\theta}_{\mathrm{SBB}}$. Similar to the consistency of GLS estimates, the correcting transformation does not affect the consistency of the estimator $\widetilde{\theta}_{\mathrm{SBB}}$.

*3.2. Crossvalidation Scheme and Selection of Spatial Weight Matrix.* In order to avoid affecting the accuracy of model estimation due to the choice of spatial weight matrix, the optimal spatial weight matrix was selected by crossvalidation. This is a widely used model selection and generalization method in machine learning. However, since the data used is panel data and the model used is spatio-temporal model, the structural features of spatio-temporal data may be destroyed if the training set and validation set are generated by hold-out method and LOOCV or $K$-folded crossvalidation. Therefore, this paper considers a stratified crossvalidation approach. For the spatio-temporal data, if $N$ and $T$ are assumed to be the number of spatial individuals and the number of periods contained in the observed samples, respectively, and the rest of the conventions on independent variables and dependent variables are the same as Equation (1), stratified crossvalidation can include the following three forms.

*3.2.1. Leave-One-Out Crossvalidation for the Time Dimension (TLOOCV).* Select the date $t$ as the validation set and the rest $T-1$ of the date as the training set. Let $C_1, C_2, \cdots, C_T$ denote, respectively, the index values of the observations contained in period $t$ ($t = 1, 2, \cdots, T$), and $N_1, N_2, \cdots, N_T$ the number of observations contained in period $t$ ($t = 1, 2, \cdots, T$). Let $n_1, n_2, \cdots, n_T$ denote the number of the observations in part $t$. Do the above for each period $t = 1, 2, \cdots, T$ in turn and calculate

$$CV_T = \frac{1}{T}\sum_{t=1}^{T} \mathrm{MSE}_t, \tag{40}$$

where $\mathrm{MSE}_t = \sum_{i=1}^{N_t}(y_{it} - y\wedge_{it})^2/N_t$, and $\widehat{y}_{it}$ is the fitting value of the $i$th observed value in period $Ty_{it}$.

This crossvalidation method is suitable when the number of periods $T$ is not too large.

*3.2.2. K-Fold Pooled Crossvalidation for the Spatial Dimension (SK-Fold PCV).* When the total number of

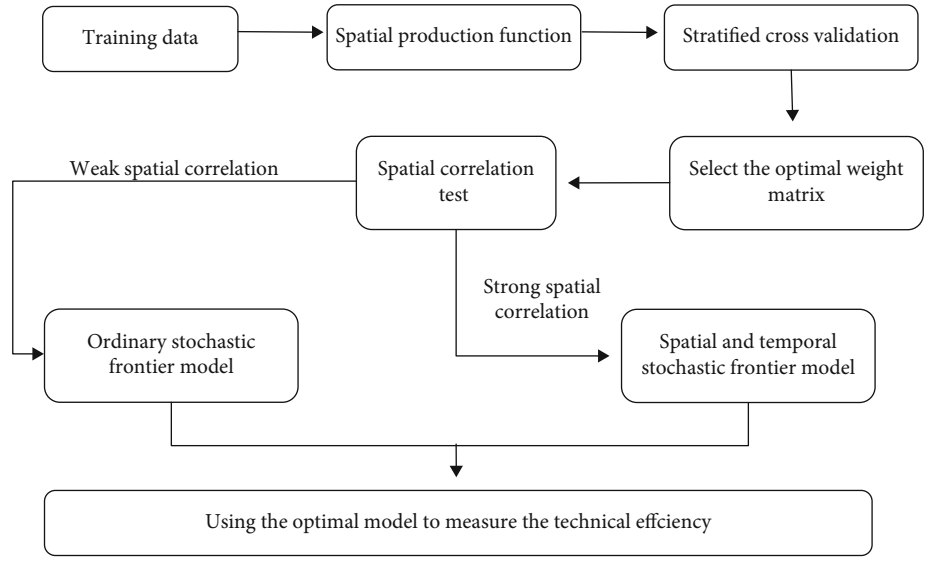



Figure 1: Spatio-temporal industrial efficiency measurement process.

periods $T$ is large and the number of spatial individuals $N$ is also large, this method is suitable for use. All observed values in each period $t$ ($t = 1, 2, \cdots, T$) were randomly divided into $K$ groups of equal size (i.e., the subsample size of each group was $N/K$), and a group was randomly selected from each period $t$ ($t = 1, 2, \cdots, T$) to obtain $NT/K$ observed values combined as the validation set and $N(K-1)T/K$ remaining observed values in each period combined as the training set. Do the above for each period sequence and calculate

$$CV_K = \frac{1}{K}\sum_{k=1}^{K} \mathrm{MSE}_k, \tag{41}$$

where $\mathrm{MSE}_k = \sum_{i=1,t=1}^{N_t,T} K(y_{it} - y\wedge_{it})^2/TN_t$, and $\widehat{y}_{it}$ is the fitting value of the $i$th observed value in period $Ty_{it}$.

*3.2.3. Leave-One-Out Crossvalidation for the Spatial Dimension (SLOOCV).* When the crossvalidation method presented in Section 3.2.2 and the condition $K = N$ attached, it can be called $K$-fold pooled crossvalidation for the spatial dimension (SN-fold PCV) or stratified leave-one crossvalidation (SLOOCV). When the total number of periods $T$ is large and the number of spatial individuals $N$ is small, this method is suitable for use.

*3.2.4. Determination of Weight Matrix and Industrial Efficiency Measure.* To discuss industrial efficiency from the perspective of spatial statistics or spatial data mining, a good spatial weight matrix should be determined first. In this paper, the spatial lag production function is chosen as the basic model, and the spatial weight matrix involved in the construction of the model can take various alternative forms. In a data-driven way, the training samples were imported into the model for parameter estimation, and then, the most appropriate spatial weight matrix was determined by stratified crossvalidation. To determine whether the spatial model is selected for analysis, the spatial correlation test is further carried out. If there is a strong spatial correlation, a spatial stochastic frontier model or a spatio-temporal stochastic frontier model will be established; if the spatial correlation is weak, an ordinary panel stochastic frontier model will be selected. After the estimation is completed, the best performing model is used to measure the technical efficiency. The flow chart of the entire analysis process is shown in Figure 1.

## 4. The Efficiency of the High and New Technology Industries in China

China has developed high and new technology industries for many years in order to transform the economic growth mode, cultivating knowledge and technology intensive new companies with great growth potential and low resources consumption that provide a sustainable development. As the technology of such industries disseminates, some issues emerge, such as spatial technology spillover, continuity of technological upgrade, and delay from research and development to market acceptance. In this paper, we analyze the efficiency of this kind of industries by the above analysis process.

*4.1. Introduction to the Model and Data.* In the framework of spatio-temporal model analysis, the spatial lag production function model based on the Cobb-Douglas production function is chosen as the basic model for weight selection. The matrix form of the model is as follows:

$$\ln Y = \rho W_N \ln Y + [\ln k, \ln l]B + u, \tag{42}$$

where $\ln Y$, $\ln k$, and $\ln l$ are, respectively, the logarithm vectors of the main business income, the assets investment, and the mean number of employees of the high and new technology industries in every province in China; $W_N$ is the spatial

Table 1: Various weight matrices to be selected.

| Expression | Matrix name | Meaning |
|---|---|---|
| $W_1$ | Position adjacency weight matrix | The weight matrix is constructed by rook, bishop, and queen position adjacence, and queen adjacence matrix is selected in this paper. Matrix element $w_{ij} = \begin{cases} 0 & \text{Region } i \text{ and } j \text{ are not adjacent} \\ 1 & \text{Region } i \text{ and } j \text{ are adjacent} \end{cases}$ |
| $W_2$ | Geographical distance weight matrix | The weight matrix is constructed by geographical distance between regions, and this paper constructs the weight matrix by reciprocal distance between the centers of provincial capitals in China. Matrix element $w_{ij} = 1/d_{ij}$, where $d_{ij}$ is the geographical distance between regions $i$ and $j$ |
| $W_3$ | Economic distance weight matrix | The weight matrix is constructed by the difference of economic level among different regions, that is, the smaller the economic gap, the stronger the spatial correlation. In this paper, the GDP of each province in China is used as the proxy variable of economic development level to construct the matrix Matrix element $w_{ij} = \begin{cases} 1/\lvert\overline{\text{GDP}}_i - \overline{\text{GDP}}_j\rvert & i \neq j \\ 0 & i = j \end{cases}$ ($\overline{\text{GDP}}_i$ represents the annual average GDP of region $i$) |
| $W_4$ | Adjacency and distance combination weight matrix | Matrix element $w_{ij} = \begin{cases} 0 & \text{Region } i \text{ and } j \text{ are not adjacent} \\ 1/d_{ij} & \text{Region } i \text{ and } j \text{ are adjacent} \end{cases}$ |
| $W_5$ | Adjacency and economic combination weight matrix | Matrix element $w_{ij} = \begin{cases} 0 & \text{Region } i \text{ and } j \text{ are not adjacent} \\ 1/\lvert\overline{\text{GDP}}_i - \overline{\text{GDP}}_j\rvert & \text{Region } i \text{ and } j \text{ are adjacent} \end{cases}$ |
| $W_6$ | Weight matrix of distance and economy combination | Matrix element $w_{ij} = (1/d_{ij}) \cdot (\overline{\text{GDP}}_i/\overline{\text{GDP}}_j)$ (where $d_{ij}$ is the geographical distance between region $i$ and $j$; $\overline{\text{GDP}}_i$ represents the annual average GDP of region $i$) |

weight matrix of the 31 provinces in China. Among many spatial weight matrices, we choose the three most commonly used spatial weight matrices in economic problems and their combination forms. Various weight matrices and their interpretations are shown in Table 1.

Before the fitting, each weight matrix was row-standardized, and the optimal weight matrix determined by crossvalidation was implemented to establish the stochastic frontier model. Starting from the ordinary panel stochastic frontier model and considering the spatial correlations, we construct the static panel spatial stochastic frontier model and the spatiotemporal stochastic frontier model. In order to determine if the model variables present spatial correlation, we let them go through a spatial correlation test, and to determine if there should be a time lag term in the model, we test the significance. Comparing the results of the three models and selecting the one that provides the best fit, we estimate the technology efficiency of the high and new technology industries in every province.

The matrix forms of the three models are as follows:

Ordinary panel stochastic frontier model:

$$\ln Y = [\ln k, \ln l]B + v - u. \tag{43}$$

Static panel space stochastic frontier model:

$$\begin{aligned} \ln Y &= \lambda(I_T \otimes W_N) \ln Y + [\ln k, \ln l]B + v - u, \\ v &= \rho(I_T \otimes M_N)v + \xi. \end{aligned} \tag{44}$$

Spatiotemporal stochastic frontier model:

$$\begin{aligned} \ln Y &= \lambda(I_T \otimes W) \ln Y + \pi(I_T \otimes W) \ln Y_{-1} + [\ln k, \ln l]B + E, \\ v &= \rho(I_T \otimes M_N)v + \xi, \end{aligned} \tag{45}$$

where $v$ is a general vector of stochastic error, $u$ is the inefficiency term, $I_T$ is the identity matrix of order $T$, $\lambda$ and $\rho$ are the spatial correlation coefficients of the corresponding equations, $\pi$ is the spatiotemporal time lag coefficient of the spatiotemporal stochastic frontier model, and $B$ is the regression coefficient vector.

The development plan of high and new technology industry in China started in 1988, but due to the relatively slow progress in the beginning, the scale development of this industry did not start until the beginning of the twenty-first century. For this reason, we have chosen as research sample the panel data of the high and new technology industries of the 31 provinces in China from 2001 to 2018. The data of capital and labor input factors have been taken from "China high-tech industry yearbook." Descriptive statistics are shown in Table 2.

Figure 2 is the histogram drawn by taking the intragroup mean of data in each region according to year.

Figure 2 shows the difference in investment and average development level of high-tech industries in different provinces of China from 2001 to 2018. As can be seen from the figure, Guangdong, Jiangsu, and Shandong provinces have

Table 2: Descriptive statistics of output value and factor input index of high-tech industry.

| Variable | Average | Standard deviation | Median | Maximum |
|---|---|---|---|---|
| $Y$ | 30.283 | 53.712 | 13.615 | 237.170 |
| $k$ | 31.682 | 54.757 | 15.884 | 241.980 |
| $l$ | 39.477 | 75.530 | 18.817 | 372.000 |

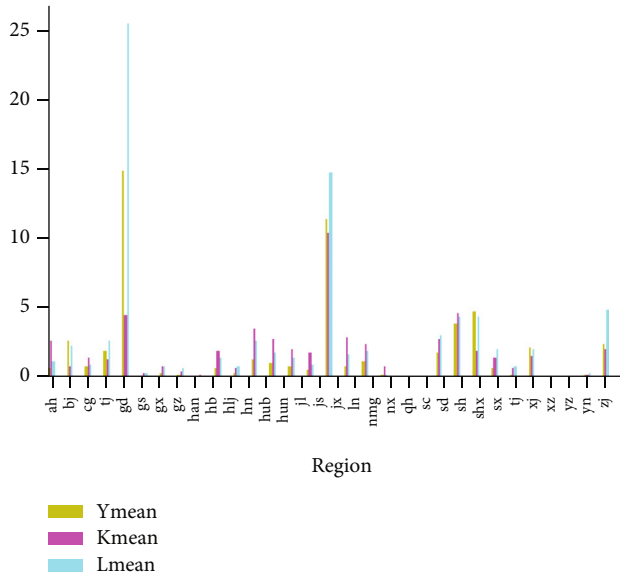



Figure 2: Histogram of the mean value of output value and factor input of high-tech industries in each province.

the highest input and output levels of high-tech industries, and the differences among these three provinces are also very large. In the past 18 years, the average output value of the high-tech industry of Guangdong, which ranks first, reached 23.7 billion yuan, while that of Shandong, which ranks third, reached only 6.997 billion yuan, less than one third of that of Guangdong. In terms of the development and distribution of high-tech industries nationwide, the gap between provinces is even more obvious. The average output value of the high-tech industries in Tibet, which ranks the last, is only 0.085 billion yuan, less than 1/1000 of that of Guangdong.

*4.2. Crossvalidation Results and the Selection of Weight Matrix.* According to the characteristics of the data obtained, the time limit contained in the data is 18 years, which is relatively short and smaller than the number of regions. Therefore, the leave-one-out crossvalidation for the time dimension (TLOOCV) method was chosen. Each weight matrix in Table 1 was introduced into model (31), the training set data were imported into the model one by one for fitting, and then, Equation (40) was calculated to obtain the CV statistics corresponding to each weight matrix. The calculation results are shown in Table 3.

By comparing the calculation results of CV statistics of validation set, it can be found to be the optimal spatial weight matrix required by this paper.

*4.3. Empirical Results.* To ensure that spatial econometrics is applicable to the problem we are studying, we need to test the spatial correlation of the variables we are interested in. The most popular method to measure spatial autocorrelation is Moran's index $I$ (Moran's $I$):

$I = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})/S^2 \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, where $S^2$ is the sample variance, $w_{ij}$ is the $(i, j)$ element of the spatial weight matrix (used to measure the distance between region $i$ and region $j$), and $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ is the sum of all spatial weights.

The value of Moran's $I$ is generally between -1 and 1, and its greater than 0 indicates positive autocorrelation. That is, the high value is adjacent to the high value and the low value is adjacent to the low value. Less than 0 means negative autocorrelation. That is, a high value is adjacent to a low value. If the Moran's $I$ is close to 0, then the spatial distribution is random, and there is no spatial autocorrelation.

To test the existence of spatial correlations in the variables of the high and new technology industry, we calculate the global Moran's $I$ indices of the production values of the industry from 2011 to 2018 (Table 4).

From the results of Moran's $I$ index calculation, we found that the $P$ value of the index is smaller than 0.01 for every year, demonstrating that the index is significant below 1% for every year, and the average Moran's $I$ index is also significant for every year. The Moran's $I$ index reached the minimum value 0.286 in 2011 and the maximum value 0.340 in 2013. We observe that the production value of the high and new technology industries of every province shows significant spatial correlation for every year and conclude that the production values of these industries of the provinces in China have apparent spatial aggregating effects. Furthermore, the testing of spatial correlation and the location quotient calculation both demonstrated that the high and new technology industries in different regions of China have apparent spatial correlation. We therefore choose the panel space stochastic frontier model for the analysis.

Table 5 presents the estimation results of the static, spatiotemporal stochastic frontier model, where the static model was further analyzed by considering fixed and stochastic effects.

The spatial autoregressive coefficients $\lambda$ and $\rho$ of the three models can all pass the significance test. From this and the spatial correlation test, one can conclude that the spatial stochastic frontier model is more reasonable. The estimated spatial autoregressive coefficients of the three models are all positive, which implied that the spatial effects have a positive impact on the development of the high and new technology industries. The negative value of the Hausman statistics of the static panel space model implies that the random effect model should be chosen. The random effect $\sigma_u^2$ value of the static panel space stochastic frontier model is far greater than the $\sigma_v^2$ value, and the value of $\gamma$ is 0.719, implying that there apparently exists technical inefficiency. Comparing the spatiotemporal stochastic frontier model with the estimation of the static panel spatial random effect, the spatiotemporal lag coefficient of the former can

Table 3: CV statistics corresponding to various weight matrices.

| Weight matrices | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_5$ | $W_6$ |
|---|---|---|---|---|---|---|
| CV statistics | 526.932 | 403.9296 | 333.7308 | 157.5527 | 320.589 | 104.8576 |

Table 4: Moran's $I$ indices among the production values of the high and new technology industry.

| | Index | $P$ value | Index | $P$ value | Index | $P$ value | Index | $P$ value |
|---|---|---|---|---|---|---|---|---|
| Year | 2011 | | 2012 | | 2013 | | 2014 | |
| Moran $I$ | 0.286 | $P \le 0.01$ | 0.321 | $P \le 0.001$ | 0.340 | $P \le 0.001$ | 0.302 | $P \le 0.01$ |
| Year | 2015 | | 2016 | | 2017 | | 2018 | |
| Moran $I$ | 0.312 | $P \le 0.01$ | 0.298 | $P \le 0.01$ | 0.305 | $P \le 0.01$ | 0.317 | $P \le 0.01$ |

Table 5: Estimation results by the static panel space and spatiotemporal stochastic frontier model.

| Dependent variable | Ln $y$ | | | | | |
|---|---|---|---|---|---|---|
| Model | Static panel spatial stochastic frontier | | | | Spatiotemporal stochastic frontier | |
| Variable | Fixed effects | | Random effects | | | |
| | Coefficients | $P$ value | Coefficients | $P$ value | Coefficients | $P$ value |
| ln $k$ | 0.397*** | 0.001 | 0.401** | 0.001 | 0.405** | 0.008 |
| ln $l$ | 0.686*** | 0.008 | 0.713** | 0.017 | 0.718*** | 0.005 |
| $\lambda$ | 0.435** | 0.012 | 0.576*** | 0.002 | 0.369*** | 0.005 |
| $\pi$ | — | — | — | — | 0.205** | 0.011 |
| $\rho$ | 0.551*** | 0.007 | 0.698** | 0.013 | 0.585** | 0.017 |
| $\sigma_u^2$ | 2.65$E$-05 | | 0.337 | | 0.656 | |
| $\sigma_v^2$ | 0.387 | | 0.132 | | 0.180 | |
| $\gamma$ | 6.850$E$-04 | | 0.719 | | 0.786 | |
| *Hausman*statistics | -39282.77 | | | | $D$ statistics | 10.668** |

In the table, ** indicates those that can pass the test with 5% significance level.

Table 6: Average technological efficiency of high-tech industries in 31 provinces.

| Province | TE | Province | TE | Province | TE | Province | TE |
|---|---|---|---|---|---|---|---|
| Anhui | 0.817 | Guizhou | 0.818 | Hunan | 0.812 | Ningxia | 0.786 |
| Beijing | 0.985 | Hainan | 0.876 | Jilin | 0.835 | Qinghai | 0.733 |
| Chongqing | 0.912 | Hebei | 0.786 | Jiangsu | 0.951 | Sichuan | 0.851 |
| Fujian | 0.936 | Heilongjiang | 0.835 | Jiangxi | 0.792 | Shandong | 0.876 |
| Gansu | 0.932 | Henan | 0.752 | Liaoning | 0.931 | Shanghai | 0.966 |
| Guangdong | 0.755 | Hubei | 0.851 | Neimenggu | 0.872 | Shǎnxi | 0.795 |
| Guangxi | 0.785 | Shānxi | 0.787 | Tianjin | 0.975 | Xinjiang | 0.651 |
| Tibet | 0.668 | Yunnan | 0.815 | Zhejiang | 0.952 | | |

Data source: calculated based on pattern II stochastic effects model of a function.

pass the 10% significance test, obtaining that the spatiotemporal lag term in the model has a significant function. The distance statistic of the spatiotemporal stochastic frontier model pass the 5% significance distance test, proving that the spatiotemporal stochastic frontier model is globally significant. Moreover, the estimation of the $\gamma$ value of the spatiotemporal stochastic frontier model is higher than that of the static panel spatial stochastic frontier model, demonstrating that the inefficiency term of the spatiotemporal stochastic frontier model has a more significant function. All the impact factor variables of the technical inefficiency terms of the analysis of the two models can at least pass the 5% significance test, and the signs of the regression coefficient of the two models are consistent, the numerical values are relatively close. Taken together, the above results all demonstrate that the analysis of the spatiotemporal stochastic frontier model is more reasonable and the development of the high and new technology industry has positive correlations in space and time. Due to space limitation, this paper does not report the annual technical efficiency of the high-

tech industries of each province. The spatial-temporal stochastic frontier model is used to calculate the average technical efficiency of the high-tech industries of each province from 2001 to 2018 as follows.

It can be seen from the calculation results in Table 6 that the average technical efficiency value of the high-tech industries in all provinces in China is less than 1, which indicates that the actual output of the high-tech industries in all provinces has not reached the most effective output level, and there is technological inefficiency in production. The five-year national average technical efficiency level was 0.837, and there are obvious regional differences in the technical efficiency values presented in Table 4. Nine provinces (Beijing, Chongqing, Fujian, Gansu, Jiangsu, Liaoning, Shanghai, Tianjin, and Zhejiang) achieved an average technical efficiency of more than 0.9, seven provinces are located in the eastern region, one in the central region, and one in the western region. There are 11 provinces with average technical efficiency below 0.8, namely, Guangdong, Guangxi, Guizhou, Hebei, Henan, Jiangxi, Ningxia, Qinghai, Shaanxi, Xinjiang, and Tibet. Only one of the provinces is in the east, six in the central region, and four in the west.

## 5. Conclusion

In this paper, taking into account that the variables to be explained might be affected by the time lag term and the space-time interaction, we develop a dynamic model within the framework of the panel spatial stochastic frontier model. Due to the apparent endogeneity of the model, we use the systematic GMM method to estimate the parameters, choose suitable tool variables according to the model assumptions and variable characteristics, and construct the suitable spatiotemporal stochastic frontier model. We use the extreme value consistence theorem and the uniform law of large numbers (ULLN) to prove the consistency of the structural parameter estimators and of the estimators of the error term distribution parameters. Aiming at the selection of spatial weight matrix of spatio-temporal model, a stratified crossvalidation method is designed to select the most appropriate spatial weight matrix in a data-driven way according to the characteristics of spatio-temporal data. Although the spatial weight matrix selected by supervised learning may not be suitable for analyzing all problems, this data-driven model selection method is undoubtedly valuable and efficient.

From the analysis of the stochastic frontier model of high-tech industries in China and the measurement of their technical efficiency, we can draw the following conclusions.

There is a spatial positive correlation in the development of high-tech industries between different regions of China. The positive correlation between the output values of these industries in different regions has been obtained by calculating the Global Moran's $I$ index in each year. The estimation of the spatial panel stochastic frontier model also indicates that the spatial autoregressive coefficient is positive, proving the existence of such a positive correlation which has a positive impact on the development of high-tech industries. There is also a spatial agglomeration effect and a spatial and temporal lag effect in these industries, illustrating that both static spill over and dynamic continuity occur in the development of the high-tech industries in China. The technical efficiency of high-tech industries is relatively low. The strategic emerging industries started earlier in eastern region, but developed more slowly than in the central and western regions.

The Chinese economy is at a critical stage of replacing old drivers of growth with new ones and transforming and upgrading industries. The new round of technological and industrial revolution 5.0 has given rise to new technologies, new industries, new forms of business, and new models. In this study, the data mining algorithm based on stochastic frontier is used to calculate industrial efficiency, which is not only suitable for high-tech industry but also helpful to further enrich the research on the efficiency of new industry and new mode and has certain practical significance to promote the steady development of the new round of scientific and technological revolution of industry 5.0.

## Appendix

## A. Proof

$E_t = [E_{1t}, \cdots, E_{Nt}]'$ is an $n$-dimensional nonzero composite random error term vector, and $E_t$ is set as interindividual nonautocorrelation according to classical econometric assumptions for simplicity.

$$\begin{aligned} COV\left\{\left[WA^{-1}\right]E_t, E_t\right\} &= \sum_i \sum_j \delta_{ij} \operatorname{cov}\left(E_{it}, E_{jt}\right) \\ &= \sum_i \delta_{ii} \operatorname{var}\left(E_{it}\right), \end{aligned} \tag{A.1}$$

where $\delta_{ij}$ is the element of the matrix $WA^{-1}$. For any $E_{it}$, $\operatorname{var}(E_{it}) > 0$, then the sufficient and necessary condition for the above formula to be 0 is $WA^{-1} = 0$, which is obviously inconsistent with the assumption of spatial weight matrix in this paper. Therefore, it is proved that

$$COV\left\{\left[WA^{-1}\right]E_t, E_t\right\} \neq 0. \tag{A.2}$$

## B. Extremum Consistency Theorem

If (1) (identification) $Q(\theta)$ is uniquely maximized at $\theta_0$, (4) (compactness) the parameter space $\Theta$ is compact, (8) (continuity) $Q(\theta)$ is continuous, and (9) (uniform convergence in probability) $\sup_{\theta \in \Theta} |\widehat{Q}(\theta) - Q(\theta)| \longrightarrow^p 0$, then $\widehat{\theta} = \arg\max_{\theta \in \Theta} \widehat{Q}(\theta) \longrightarrow^p \theta_0$.

## C. Proof

From (25), we let $\widehat{H}(\theta) = (1/N)[H'_{N,BB}(\Delta Y *_N - \Delta Z *_N \theta)]$ and $H(\theta) = E[H'_{BB}(\Delta Y * - \Delta Z * \theta)]$, which are continuous functions of the corresponding parameters constructed by sample and global moment conditions, respectively. $\widehat{A} = A_{N,BB}$ and $A = A_{BB} = E[H'_{BB} G_{BB} H_{BB}]^{-1}$ are the weight matrix sample constructed and global constructed, respectively,

where $\theta = (\lambda_1, \lambda_2, \gamma, B)'$ is the structural parameter vector of the model. The above settings satisfy the following conditions:

(i) When $\theta = \theta_0$, $H(\theta) = E[H'_{BB}(\Delta Y * - \Delta Z * \theta)] = 0$, and $\widehat{A}$ and $A$ are both positive definite matrices

(ii) $\boldsymbol{\Theta}$, the parameter space of $\theta$, is compact, and $\theta_0 \in \boldsymbol{\Theta}$

(iii) For an arbitrary point $\theta \in \boldsymbol{\Theta}$, $H(\theta)$ is a continuous function

(iv) $E \sup_{\theta \in \boldsymbol{\Theta}} \|H(\theta)\| < \infty$. Let

$$\begin{aligned} \widehat{Q}(\theta) &= -\widehat{H}(\theta)' \widehat{A} \widehat{H}(\theta), \\ Q(\theta) &= -H(\theta)' A H(\theta) \end{aligned} \tag{C.1}$$

We first prove uniqueness of the maximum value of $Q(\theta)$.

As $H(\theta_0) = 0$, one has $Q(\theta_0) = 0$, which is the maximum value of $Q(\theta)$ given that $A$ is a positive definite matrix. From the uniqueness of the true parameter value, for $\theta \neq \theta_0$, it is satisfied that

$$H(\theta) \neq 0, Q(\theta) = -H(\theta)' A H(\theta) < 0, \tag{C.2}$$

and therefore, $\theta = \theta_0$ is the unique maximum of $Q(\theta)$.

According to the uniform law of large numbers (ULLN), when conditions (ii), (iii), and (iv) hold, $E[H(\theta)]$ and $Q(\theta)$ are continuous functions, and $\sup_{\theta \in \boldsymbol{\Theta}} \|\widehat{H}(\theta) - H(\theta)\| \longrightarrow^p 0$. Noticing the structure of the weight matrices $\widehat{A}$ and $A$, and the fact that the instrumental variable matrices $H'_{N,BB}$ and $H'_{BB}$ can be regarded as $\widehat{H}(\theta)$ and $H(\theta)$ under the condition $\Delta Y * - \Delta Z * \theta \equiv 1$, one has $\widehat{A} \longrightarrow^p A$.

$$\begin{aligned} \left|\widehat{Q}(\theta) - Q(\theta)\right| &= \left|\widehat{H}(\theta)' \widehat{A} \widehat{H}(\theta) - H(\theta)' A H(\theta)\right| \\ &= \left|\widehat{H}(\theta)' \left(\widehat{A} - A + A\right) \widehat{H}(\theta) - H(\theta)' A H(\theta)\right| \\ &= \left|\widehat{H}(\theta)' \left(\widehat{A} - A\right) \widehat{H}(\theta) + \widehat{H}(\theta)' A \widehat{H}(\theta) \right. \\ &\quad \left. - H(\theta)' A H(\theta)\right| = \left|\widehat{H}(\theta)' \left(\widehat{A} - A\right) \widehat{H}(\theta) \right. \\ &\quad + [H\wedge(\theta) - H(\theta) + H(\theta)]' A \left[\widehat{H}(\theta) \right. \\ &\quad \left. - H(\theta) + H(\theta)\right] - H(\theta)' A H(\theta)\Big| \\ &= \left|\widehat{H}(\theta)' \left(\widehat{A} - A\right) \widehat{H}(\theta) + [H\wedge(\theta) \right. \\ &\quad - H(\theta)]' A \left[\widehat{H}(\theta) - H(\theta)\right] + 2[H\wedge(\theta) \\ &\quad - H(\theta)]' A H(\theta) + H(\theta)' A H(\theta) \\ &\quad \left. - H(\theta)' A H(\theta)\right| = \left|\widehat{H}(\theta)' \left(\widehat{A} - A\right) \widehat{H}(\theta) \right. \\ &\quad + [H\wedge(\theta) - H(\theta)]' A \left[\widehat{H}(\theta) - H(\theta)\right] \\ &\quad \left. + 2[H\wedge(\theta) - H(\theta)]' A H(\theta)\right|, \end{aligned} \tag{C.3}$$

where in the last line, we have applied the triangle inequality.

$$\begin{aligned} &\leq \left|\widehat{H}(\theta)' \left(\widehat{A} - A\right) \widehat{H}(\theta)\right| + \left|[H\wedge(\theta) - H(\theta)]' A \left[\widehat{H}(\theta) \right.\right. \\ &\quad \left.\left. - H(\theta)\right]\right| + 2\left|[H\wedge(\theta) - H(\theta)]' A H(\theta)\right| \end{aligned} \tag{C.4}$$

Taking supreme on both sides of the above inequality, we obtain

$$\begin{aligned} \sup_{\theta \in \boldsymbol{\Theta}} \left|\widehat{Q}(\theta) - Q(\theta)\right| &\leq \sup_{\theta \in \boldsymbol{\Theta}} \left|\widehat{H}(\theta)' \left(\widehat{A} - A\right) \widehat{H}(\theta)\right| \\ &\quad + \sup_{\theta \in \boldsymbol{\Theta}} \left|[H\wedge(\theta) - H(\theta)]' A \left[\widehat{H}(\theta) - H(\theta)\right]\right| \\ &\quad + 2 \sup_{\theta \in \boldsymbol{\Theta}} \left|[H\wedge(\theta) - H(\theta)]' A H(\theta)\right| \\ &\leq \sup_{\theta \in \boldsymbol{\Theta}} \left\|H\wedge(\theta)'\right\|^2 \left\|\widehat{A} - A\right\| \\ &\quad + \sup_{\theta \in \boldsymbol{\Theta}} \|H\wedge(\theta) - H(\theta)\|^2 \|A\| \\ &\quad + \sup_{\theta \in \boldsymbol{\Theta}} \left\|\widehat{H}(\theta) - H(\theta)\right\| \\ &\quad + 2 \sup_{\theta \in \boldsymbol{\Theta}} \left\|\widehat{H}(\theta) - H(\theta)\right\| \cdot \|A\| \\ &\quad \cdot \sup_{\theta \in \boldsymbol{\Theta}} \|H(\theta)\| \xrightarrow{p} 0 \end{aligned} \tag{C.5}$$

According to the Extremum Consistency Theorem, we have $\widehat{\theta}_{\mathrm{SBB}} \longrightarrow^p \theta$.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the Shandong Provincial Social Science Planning Key Project "New Infrastructure" to Promote the Shandong Provincial Economic High-Quality Development Path Research (21BJJJ07), Special Research Project of Shandong Social Science Fund on the Conversion of Old and New Driving Forces "Study on the Space-time Evaluation Mechanism of Production Efficiency of New Industry and New Format in Shandong" (19CDNJ37), Social Science Research Project of Shandong University of Political Science and Law "Study on The Conversion of Old And New Driving Forces and Efficiency Index of Cultural Industry" (2019Q14B), Shandong Province Humanities Social Science Finance Application Key Project "Research on Rural Supply Chain Finance Model and Credit Risk Early Warning in Shandong Province under the Environment of Internet +" (2020-JRZZ-11), and National Natural Science Foundation of China-Shandong Joint Fund (No. U1806203). This material is also based upon work supported by Program for Big

Data and Artificial Intelligence Legal Research Collaborative Innovation Center in Shandong University of Political Science and Law.

## References


[1] D. Aigner, K. C. A. Lovell, and P. Schmidt, "Formulation and estimation of stochastic frontier production function models," *Journal of Econometrics*, vol. 6, no. 1, pp. 21–37, 1977.

[2] W. Meeusen and J. van den Broeck, "Technical efficiency and dimension of the firm: some results on the use of frontier production functions," *Empirical Economics*, vol. 2, no. 2, pp. 109–122, 1977.

[3] G. E. Battese and G. S. Corra, "Estimation of a production frontier model: with application to the pastoral zone of Eastern Australia," *Australian Journal of Agricultural and Resource Economics*, vol. 21, no. 3, pp. 169–179, 1977.

[4] V. Druska and W. C. Horrace, "Generalized moments estimation for spatial panel data: Indonesian rice farming," *American Journal of Agricultural Economics*, vol. 86, no. 1, pp. 185–198, 2004.

[5] E. Affuso, "Spatial Autoregressive Stochastic Frontier Analysis: An Application to an Impact Evaluation Study," 2010, SSRN 1740382.

[6] A. Tonini and V. Pede, "A generalized maximum entropy stochastic frontier measuring productivity accounting for spatial dependency," *Entropy*, vol. 13, no. 11, pp. 1916–1927, 2011.

[7] F. Vidoli, C. Cardillo, E. Fusco, and J. Canello, "Spatial nonstationarity in the stochastic frontier model: an application to the Italian wine industry," *Regional Science and Urban Economics*, vol. 61, pp. 153–164, 2016.

[8] E. G. Tsionas and P. G. Michaelides, "A spatial stochastic frontier model with spillovers: evidence for Italian regions," *Scottish Journal of Political Economy*, vol. 63, no. 3, pp. 243–257, 2016.

[9] A. Carvalho, "Efficiency spillovers in Bayesian stochastic frontier models: application to electricity distribution in New Zealand," *Spatial Economic Analysis*, vol. 13, no. 2, pp. 171–190, 2018.

[10] M. Adetutu, A. J. Glass, K. Kenjegalieva, and R. C. Sickles, "The effects of efficiency and TFP growth on pollution in europe: a multistage spatial analysis," *Journal of Productivity Analysis*, vol. 43, no. 3, pp. 307–326, 2015.

[11] F. Jin and L.-f. Lee, "Asymptotic properties of a spatial autoregressive stochastic frontier model," *Journal of Spatial Econometrics*, vol. 1, no. 1, pp. 1–40, 2020.

[12] L. Kutlu, K. C. Tran, and M. G. Tsionas, "A spatial stochastic frontier model with endogenous frontier and environmental variables," *European Journal of Operational Research*, vol. 286, no. 1, pp. 389–399, 2020.

[13] A. S. Bergantino, M. Intini, and N. Volta, "Competition among airports at worldwide level: a spatial analysis," *Transportation Research Procedia*, vol. 45, no. 1, pp. 621–626, 2020.

[14] T. D. Graaff, "On the estimation of spatial stochastic frontier models: an alternative skew-normal approach," *The Annals of Regional Science*, vol. 64, no. 2, pp. 267–285, 2020.

[15] L. Jia-Xian, L. Zhi-He, and L. Guang-Ping, "Spatial Panel Stochastic Frontier Model and Technical Efficiency Estimation," *Journal of Business Economics*, vol. 223, no. 5, pp. 71–78.

[16] N. Deepa, Q. V. Pham, D. C. Nguyen et al., "A survey on blockchain for big data: approaches, opportunities, and future directions," 2020, https://arxiv.org/abs/2009.00858.

[17] H. H. Kelejian and I. R. Prucha, "A generalized moments estimator for the autoregressive parameter in a spatial model," *International Economic Review*, vol. 40, no. 2, pp. 509–533, 1999.

[18] M. Kapoor, H. H. Kelejian, and I. R. Prucha, "Panel data models with spatially correlated error components," *Journal of Econometrics*, vol. 140, no. 1, pp. 97–130, 2007.

[19] J. P. Jacobs, J. E. Ligthart, and H. Vrijburg, "Dynamic panel data models featuring endogenous interaction and spatially correlated errors," *SSRN Electronic Journal*, vol. 92, pp. 1–37, 2009.

[20] L. Anselin and D. A. Griffith, "Do spatial effecfs really matter in regression analysis?," *Papers in Regional Science*, vol. 65, no. 1, pp. 11–34, 1988.

[21] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[22] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[23] S. Jun, "Linear model selection by cross-validation," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 486–494, 1993.

[24] T. W. Anderson and C. Hsiao, "Estimation of Dynamic Models With Error Components," *Journal of American Statistical Association*, vol. 76, no. 375, pp. 598–606, 1981.

[25] M. Arellano and S. Bond, "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *The Review of Economic studies*, vol. 58, no. 2, pp. 277–297, 1991.

[26] M. Arellano and O. Bover, "Another look at the instrumental variable estimation of error-components models," *Journal of Econometrics*, vol. 68, no. 1, pp. 29–51, 1995.

[27] R. Blundell and S. Bond, "Initial conditions and moment restrictions in dynamic panel data models," *Journal of Econometrics*, vol. 87, no. 1, pp. 115–143, 1998.

[28] H. H. Kelejian and D. P. Robinson, "A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model," *Papers in Regional Science*, vol. 72, no. 3, pp. 297–312, 1993.

[29] T. Beck and R. Levine, "Stock markets, banks, and growth: Panel evidence," *Journal of Banking & Finance*, vol. 28, no. 3, pp. 423–442, 2004.

[30] W. Newey and D. McFadden, "Chapter 36 Large sample estimation and hypothesis testing," in *Handbook of Econometrics, vol. 4*, R. Engle and D. McFadden, Eds., Elsevier B.V, New York: North-Holland, 1994.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 2456629, 11 pages
https://doi.org/10.1155/2021/2456629

*Research Article*

# Scientific Impact of Sports on Human Health and Physique Based on Optimization Big Data Ant Colony Algorithm

**Lin Wu**

*School of Physical Education, Xi'an Peihua University, Xi'an, Shaanxi 710100, China*

Correspondence should be addressed to Lin Wu; wulin@peihua.edu.cn

Received 13 August 2021; Revised 29 September 2021; Accepted 1 October 2021; Published 2 November 2021

Academic Editor: Rajesh Kaluri



With the continuous improvement of living standards, people began to pay more and more attention to sports, and the impact of sports on human health and physique has been paid more and more attention. This study mainly analyzes the scientific impact of sports on human health and physique under the background of big data. Firstly, the big data analytic hierarchy process is used to construct the comprehensive evaluation structure system of sports on human health and physique. Then, an improved big data adaptive ant colony classification rule algorithm is proposed. Finally, the performance evaluation and physical impact analysis of the improved big data algorithm are carried out. The results show that compared with other algorithms, ACA ∗ (ant colony algorithm) based on big data has more obvious advantages in stability, optimization ability, running time, and convergence speed and is more suitable for practical application. In general, the improvement of the physical fitness level of the association members in 2019 mainly depends on the results of the improvement of the physical fitness level. In the future, we need to strengthen physical exercise, change living habits and traffic habits, and other methods to optimize the overall physical fitness.

## 1. Introduction

Sports play an important role in promoting human health. This study attempts to use big data to optimize big data ant colony algorithm to analyze the scientific impact of sports on human health and physique. Big data ant colony algorithm has been widely used in various fields because of its strong positive feedback mechanism, self-organization, and distributed computing characteristics [1, 2]. de Santis' team proposed a metaheuristic routing algorithm (FW ACO) based on ant colony optimization (ACO) metaheuristic algorithm and Floyd-Warshall (FW) algorithm. The results show that FW ACO algorithm can provide better results than heuristic algorithm and metaheuristic algorithm and has high computational efficiency, which is suitable for determining the path of selector in real time [3]. Babanezhad and other scholars use the big data ant colony intelligence method to learn CFD data of different parts of BCR. The results show that the algorithm does not need to solve the Navier-Stokes equation and complex solution process but can be used for prediction process [4]. Sutar and other scholars proposed a dynamic energy-saving virtual machine migration method, which applies big data ant colony optimization algorithm to analyze the load on the physical machine and start the sleep mode of an idle physical machine to reduce power consumption, so as to achieve the purpose of energy saving [5]. Kusumahardhini's team proposed to use the $K$-means method and cross ant colony optimization (ACO) to solve the medium-term strategic planning problem of travel agencies, compared and analyzed the results of the $K$-means method and cross ant colony optimization, and analyzed the impact of selecting cities as parking lots on the total travel distance [6].

Asghari and Navimipour proposed the inverse ant colony optimization (IACO) algorithm to improve the load balance between nodes. The results show that the IACO algorithm is superior to the ACO algorithm in load balancing, waiting time, and resource utilization [7]. Ji's team combined the heuristic operator of big data ant colony optimization (ACO) with decomposition-based multiobjective evolutionary algorithm (MOEA/D) to propose a

multiobjective community detection algorithm mocd ACO. In the experiment, the performance of mocd ACO is evaluated by using comprehensive network data set and real network data set. The results show that compared with the five most advanced methods, mocd ACO is effective in standardizing mutual information and modularization [8]. Sinwar et al. proposed the application of swarm intelligence (SI) technology based on big data ant colony optimization (ACO) and particle swarm optimization (PSO). Simulation results show that the performance of the ACOP protocol is better than other protocols [9]. Ning and other scholars designed a new pheromone smoothing mechanism to improve the global search ability. When the search process of big data ant colony algorithm approaches a fixed stagnation state, the pheromone matrix is reinitialized. The results show that the improved big data ant colony algorithm is superior to the traditional big data ant colony algorithm in solution diversity and convergence speed [10]. Moeini and Afshar combined ACoA with NLP to optimize the design of sewage pipe network. The results show that acoa-nlp2 is an effective method to solve the problem of optimal design of sewage pipe network [11]. Yuan and other scholars use big data ant colony algorithm to determine the boundary information of the foreground target and fuse different pheromone images at the superpixel level to generate three accurate bitmaps. Experimental results show that the generated three high-quality images effectively improve the performance of the algorithm and achieve accurate segmentation $\alpha$ mask estimation [12]. Wang's team proposed a big data ant colony algorithm based on multiobjective optimization and carried out simulation experiments using cloudsim cloud simulation platform. The results show that compared with other commonly used algorithms, this algorithm has reached a better level in SLA violation rate, power consumption, and resource loss of cloud platform [13]. Yu and other scholars studied the logistics terminal distribution mode and path optimization, combined with the application of big data ant colony algorithm in traveling salesman problem, and analyzed the basic principle and implementation process of big data ant colony algorithm. The results show that the application of big data ant colony algorithm in logistics terminal distribution path optimization is of great significance [14]. Zhao and other scholars proposed a multiobjective optimization model based on cost, carbon emission, and customer satisfaction and designed an improved big data ant colony algorithm (acomo) with multiobjective heuristic function. The results show that acomo algorithm can effectively solve the vehicle routing problem under the multiobjective optimization model, is better than the traditional big data ant colony algorithm, and obtains more Pareto optimal solutions [15]. According to the equipment parameters of the national hot strip mill, the national D team adopted the load distribution optimization method and the max-min big data ant colony algorithm to optimize the load distribution of the rolling mill. After using this method, the compression ratio of upstream and downstream stands gradually decreased, and the change of proportional crown width was less than that of the traditional energy method. The actual fluctuation amplitude is reduced by nearly 50%, which is conducive to shape control, and the total dissipated power is lower than that of the energy consumption method [16].

Based on the above research results, this study will focus on the optimization process based on big data ant colony algorithm, combined with big data analytic hierarchy process to analyze the scientific impact of sports on human health and physique.

This study mainly analyzes the scientific impact of sports on human health and physique under the background of big data. This paper is divided into three parts. The first part constructs a comprehensive evaluation structure system of sports on human health and physique by using big data analytic hierarchy process. In the second part, an improved big data adaptive ant colony classification rule algorithm is proposed. Finally, the performance evaluation and physical impact analysis of the improved big data algorithm are carried out.

## 2. Construction of the Influence Model of Sports on Human Health and Physique Based on Big Data Ant Colony Algorithm

*2.1. Comprehensive Evaluation Structure System of Sports on Human Health and Physique.* Under the guidance of the national fitness program, people pay more and more attention to the impact of sports on human health, among which experts and scholars who have been engaged in sports work for a long time have a say. As people participate in sports, the level of physical fitness often has a certain dynamic, so we should consider the actual data and the opinions of experts and scholars. Taking the physique monitoring results of a sports association in Chengdu from 2015 to 2019 as the data sample, this study analyzes the comprehensive physique level of young people aged 20 to 39 and uses the big data analytic hierarchy process (AHP) to construct the evaluation standard of human health and physique after sports, as shown in Figure 1.

Big data analytic hierarchy process (AHP) is a multiobjective decision-making method, which decomposes the problem into different elements according to the nature of the problem and the solving goal and constructs a multilevel structure model by combining the membership relationship between the elements. In the hierarchical structure model of this study, there are three primary indexes in the middle layer: body shape, body function, and physical fitness. The lowest secondary indexes include BMI index, chest circumference index, waist circumference index, vital capacity, quiet pulse, systolic blood pressure, diastolic blood pressure, push up/sit up, grip strength/back strength, vertical jump, and choice reaction time.

After building the hierarchical structure model, it is necessary to allocate the weight and check the consistency of the indicators in the model. The big data analytic hierarchy process (AHP) is to decompose the problem into multiple levels and take the highest level as the decision-making objective of the problem. According to the relative scale, it calculates the weight of the next level relative to the high level, so as to make the best decision. After building the model judgment

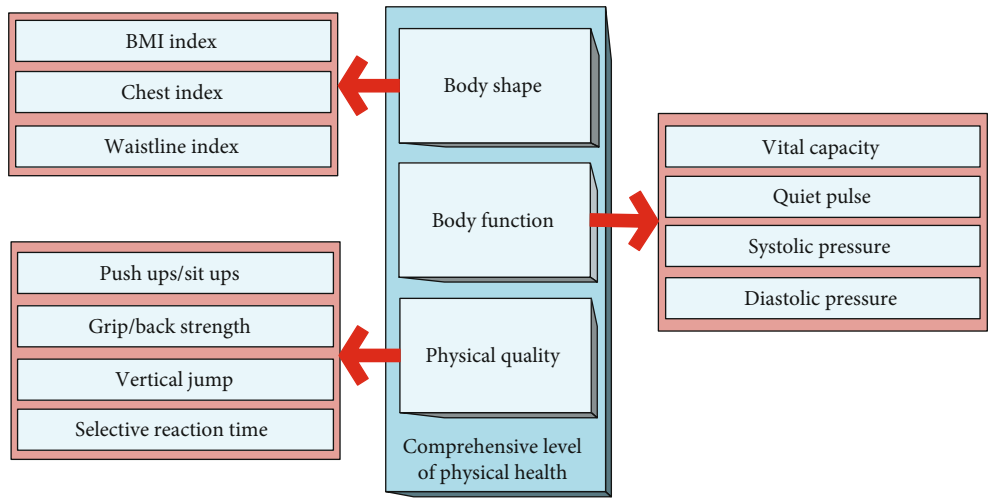



Figure 1: Comprehensive evaluation structure of sports on human health and physique.

matrix, we need to scale the importance of the indicators according to the relative scale, which is to reduce the impact of the differences in the nature of the indicators as far as possible, so as to improve the accuracy of the judgment matrix. By calculating the maximum eigenvalue of the model judgment matrix, we can get the eigenvector of the indicators at the same level. By normalizing the eigenvector, we can get the relative weight of the indicators at the same level. For the judgment matrix $A$, its maximum eigenvalue can be obtained by the following formula, where $n$ represents the number of indicators in the judgment matrix of this level, and the elements in $W$ represent the weight of indicators after normalization.

$$\lambda_{\max} = \sum_{i=1}^{n} \frac{(AW)_i}{W_i} \cdot \frac{1}{n}. \tag{1}$$

Consistency test is to test the deviation of judgment matrix. The calculation formula of consistency index (CI) is shown in formula (2). When CI = 0, the judgment matrix has complete consistency; when CI is close to 0, the matrix has relatively satisfactory consistency; when the value of CI is larger, the consistency of the judgment matrix is worse.

$$\mathrm{CI} = \frac{\lambda_{max} - n}{n - 1}. \tag{2}$$

Although the consistency index can reflect the inconsistency of the judgment matrix, it can not accurately describe the inconsistency degree of the judgment matrix. Therefore, the consistency ratio of the judgment matrix can be calculated by referring to equation (3), and the judgment matrix can be adjusted in different degrees according to the inconsistency of the judgment matrix.

$$\mathrm{CR} = \frac{\mathrm{CI}}{\mathrm{RI}}. \tag{3}$$

Consistency ratio defines the rationality of nonconsistency of judgment matrix. According to the calculation result of formula (3), when $\mathrm{CR} \le 0.1$, the judgment matrix is considered to be consistent. Since the normalized result of the maximum eigenvalue of the judgment matrix represents the index weight vector of the model, it can be seen that the weight distribution of the corresponding model index of the consistent judgment matrix is reasonable. Therefore, the weight of each indicator relative to the superior indicator is shown in Table 1.

*2.2. Adaptive Ant Colony Classification Rule Mining Algorithm Model.* This research combines big data ant colony algorithm and classification rule algorithm to get a simplified graph of adaptive ant colony classification rule mining algorithm to find the best path, as shown in Figure 2. Among them, Figure 2(a) shows that the ant colony finds the best path at the starting point and the target two points; Figure 2(b) shows that the number of ants on both sides of the obstacle is almost the same in a short time after placing the obstacle between the starting point and the target; Figure 2(c) shows that with the increase of the time of placing the obstacle, the ants will pass through the path with high pheromone concentration, and the number and number of ants are almost the same. The concentration of pheromone is proportional to the concentration of pheromone; Figure 2(d) shows that most ant colonies can find food in the fastest way after a period of time.

Classification rule algorithm is a data classification technology based on rule classifier. Among them, ant miner algorithm is a traditional classification rule algorithm, which uses information entropy theory to establish heuristic function. The algorithm has high classification accuracy and simple rules [17]. The disadvantage of ant miner algorithm is that the calculation of entropy is complex, the running speed is slow, and it is easy to local stagnation. Based on the original ant miner algorithm, the heuristic factor and pheromone updating method are optimized, and the adaptive mechanism is added to construct the optimization algorithm. For the decision classification problem in the covering algorithm, set $k$ data in the data set, and the attribute variables

Table 1: Comprehensive level of physical health.

| Target layer | Criterion layer | Weight | Index layer | Weight |
|---|---|---|---|---|
| Comprehensive level of physical health | Body shape | 0.24 | BMI index | 0.69 |
| | | | Chest index | 0.20 |
| | | | Waistline index | 0.11 |
| | Body function | 0.28 | Vital capacity | 0.50 |
| | | | Quiet pulse | 0.20 |
| | | | Systolic pressure | 0.16 |
| | | | Diastolic pressure | 0.14 |
| | Physical quality | 0.48 | Physical quality | 0.30 |
| | | | Grip/back strength | 0.30 |
| | | | Vertical jump | 0.12 |
| | | | Selective reaction time | 0.28 |

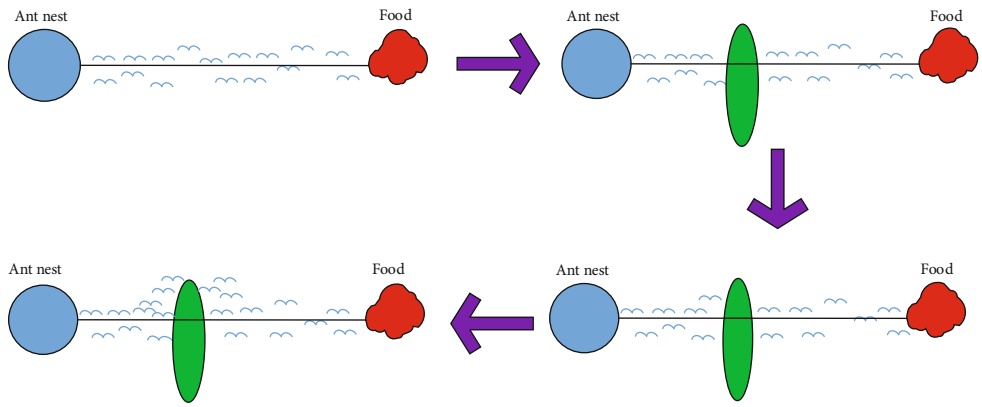



Figure 2: Schematic diagram of big data ant colony algorithm seeking optimal method.

in the data set are $A_i$, $i = 1, 2, \cdots a$, $V_{ij}$, and $b_i$, which, respectively, represent the $j$th attribute value and the number of $A_i$ attributes. Attribute $\text{term}_{ij}$ means that $V_{ij}$ and $A_i$ are the same. In the covering algorithm, the best rule is selected. In the process of rule construction, the $\text{term}_{ij}$ selected each time is applied to the current rule. The probability calculation formula of selecting $\text{term}_{ij}$ attribute is

$$P_{ij}(t) = \frac{\tau_{ij}(t)^{\alpha} \cdot \eta_{ij}(t)^{\beta}}{\sum_{i}^{a}\sum_{j}^{b_i}\tau_{ij}(t)^{\alpha} \cdot \eta_{ij}(t)^{\beta}}. \tag{4}$$

In equation (4), $\eta_{ij}(t)$ and $\tau_{ij}(t)$ represent pheromone concentration and heuristic factor of $t$ path $ij$ at time, respectively, $\alpha$ and $\beta$ represent the relative importance of pheromone concentration and heuristic factors, respectively, and $\alpha$ and $\beta$ are greater than 0. The higher the $\alpha$ value is, the higher the pheromone concentration is. The higher the $\alpha$ value is, the higher the pheromone concentration is. The larger the $\beta$ value is, the greater the influence of the current attribute of the covering sample on the ant's selection path [18]. In order to accelerate the convergence speed, it is necessary to initialize the information concentration, and the pheromones of all paths are equal at the initial time. After the new classification rules are constructed, the data information attributes need to be updated iteratively, and the method is shown in

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \tau_{ij}(t) \cdot Q. \tag{5}$$

In equation (5), $Q$ is the validity of the rule, $\rho$ is the information exertion factor, and the value is in the interval [0,1]. The relationship between pheromone and rule validity is positive effect. The attribute update method without rules is shown in

$$\tau_{ij}(t+1) = \frac{\tau_{ij}(t)}{\sum_{i}^{a}\sum_{j}^{b_i}\tau_{ij}(t)}. \tag{6}$$

Due to the complexity of the traditional information entropy updating method for attribute items, the heuristic factor based on density is used for calculation, as shown in

$$\eta_{ij}(t) = \frac{|T_{ij}(t)|}{|T(t)|}. \tag{7}$$

In equation (7), $|T_{ij}(t)|$ and $|T(t)|$ represent the number and total number of attribute item data contained in the data set, respectively. Then, the pruning strategy is used to avoid overfitting the rules, and the rules to judge the effectiveness before and after pruning are shown in

$$Q = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \cdot \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{8}$$

In equation (8), TP refers to the number of data samples that meet the requirements before and after the rule change, FP refers to the number of samples that meet the requirements before the rule change but do not meet the requirements after the rule change, and FN and TN are just opposite to TP and FP. When $Q$ increases after pruning, pruning operation can be carried out, and the rules after pruning are effective rules. To clear the data set contained in the effective rule in the data set, the rule can be added to the rule set only if the number of training set samples of the rule is greater than or equal to the specified number of samples [19]. In view of the existing problems of big data ant colony algorithm, deterministic selection and random selection are applied to optimize, and information volatility is dynamically adjusted. Under the condition of large number of iterations, the gap between the highest and lowest pheromone concentration paths is narrowed, and the probability of random selection is increased. Some scholars pointed out that only setting a small $\rho$ value at the initial stage of path search and increasing the $\rho$ value at the later stage can effectively prevent local stagnation [20]. Therefore, the adaptive adjustment mechanism used in this study is shown in

$$\rho(t) = \frac{3}{2}\int_0^t f(\tau)d\tau. \tag{9}$$

In equation (9), $f(\tau)$ is the normal distribution function with $\mu = 0$, the maximum value of $\rho$ is 0.75, and the standard deviation is 10. Compared with the standard deviation value of 4, the $\rho$ value decreases more smoothly with time, which makes the convergence speed faster and helps to find the best path. The probability expression of path selection is shown in

$$\mathrm{term}_{ij} = \begin{cases} \mathrm{argmax}\left\{\tau_{ij}(t)^{\alpha} \cdot \eta_{ij}(t)^{\beta}\right\}, & \text{if } r \le P_0, \\ \text{select } P_{ij} \text{ according to probability } \mathrm{term}_{ij}, & \text{else}, \end{cases} \tag{10}$$

where $P_0 \in (0, 1)$ and $r$ are random numbers uniformly distributed in (0,1) interval. The improved algorithm is shown in Figure 3, which is divided into three processes: initialization, iterative process, and comparison rule effectiveness. $k$ and $m$ are the number of ants and the total number of ants.

*2.3. Construction of Health Constitution Influence Model Based on Improved Adaptive Ant Colony Classification Rule Algorithm.* In the process of adaptive classification, the difference of the initial change source will lead to the coupling phenomenon of the algorithm, which increases the difficulty of the optimal classification path [21]. Therefore, an improved adaptive big data ant colony algorithm (ACA*) is proposed in this study. By introducing a guide factor in the probability transfer strategy, the predictability of the big data ant colony algorithm to the target node is improved, the blind selection of ants in the state transition is avoided, and the speed of ant search for feasible solution is accelerated [22]. Let the length from the current node $i$ to the target node $E$ be $iE$, and the calculation of the guiding factor is shown in

$$\lambda_{iE} = \frac{m-k}{m}\frac{N_{\max} - N}{N_{\max}}\frac{1}{d_{iE}}. \tag{11}$$

In equation (11), $m$ is the total number of ants, $k$ is the current number of ants, $N$ is the current number of iterations, and $N_{\max}$ is the maximum number of iterations. The transition probability of introducing the guidance factor from node $i$ to $j$ is shown in

$$p_{ij}^{k}(t) = \begin{cases} \dfrac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}\lambda_{jE}^{\gamma}(t)}{\sum_{s\in \mathrm{allowed}_k}\tau_{\mathrm{is}}^{\alpha}(t)\eta_{\mathrm{is}}^{\beta}\lambda_{jE}^{\gamma}(t)}, & j \in \mathrm{allowed}_k, \\ 0, & \text{otherwise}. \end{cases} \tag{12}$$

In equation (12), $\lambda_{jE}^{\gamma}(t)$ is the guiding function of the current node, $\lambda_{iE}^{\gamma}(t)$ is the guiding function of selecting the next node, $\eta_{ij}$ is the heuristic factor, that is, the expected degree of ants from node $i$ to node $j$, and $\eta_{ij} = 1/I_{ij}$. The improved big data ant colony algorithm is used to solve the path optimization problem. In the established change propagation analysis network, the objective function and constraints are defined, as shown in

$$\begin{cases} \min\left(\sum_{1}^{k} I_{ij}\right), & k = 1, 2, 3, \cdots, \\ \text{s.t.} \sum_{i=1}^{p} \rho_i \ge \Delta\rho_u, & p = 1, 2, 3, \cdots, \\ \text{s.t.} \sum_{j=1}^{q} \rho_j \ge \Delta\rho_v, & q = 1, 2, 3, \cdots, \\ \text{s.t.} \sum_{l=1}^{r} \rho_l \ge \Delta\rho_w, & r = 1, 2, 3, \cdots, \\ \cdots. \end{cases} \tag{13}$$

In equation (13), $\Delta\rho_u$, $\Delta\rho_v$, and $\Delta\rho_w$ denote the change impact of the initial change node (ICI), $\rho_i$, $\rho_j$, and $\rho_l$ denote the change over absorption capacity of the node they represent, $p$, $q$, and $r$ denote the propagation steps of each initial change node, and $k$ represents the sum of the propagation steps of all change nodes. The optimal adaptive optimization algorithm framework based on the improved big data ant colony algorithm is shown in Figure 4 [11].

As can be seen from Figure 4, the first step of the improved adaptive big data ant colony algorithm path is to establish a network model of influencing factors, map different influencing factors to the network as different nodes, and map the relationship between influencing factors to the edges in the network model. The second step

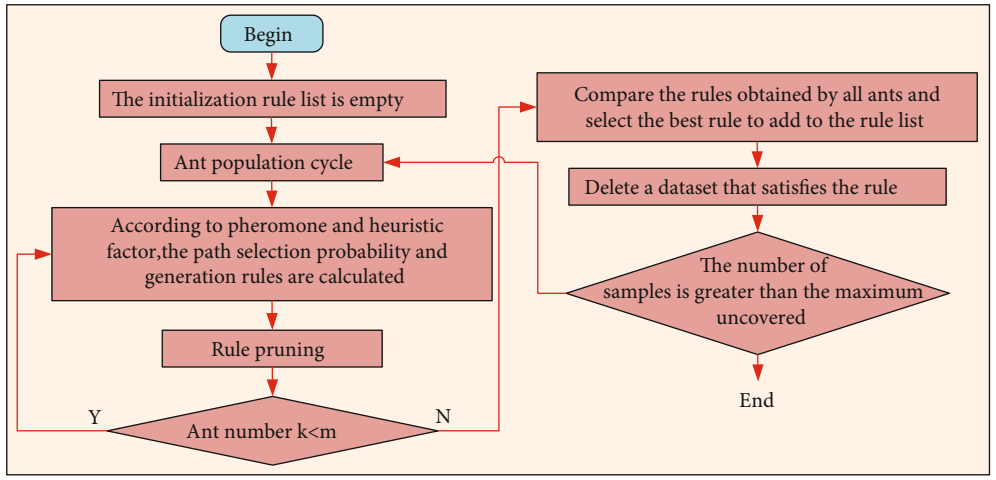



Figure 3: Decision-making model of adaptive ant colony classification rules.

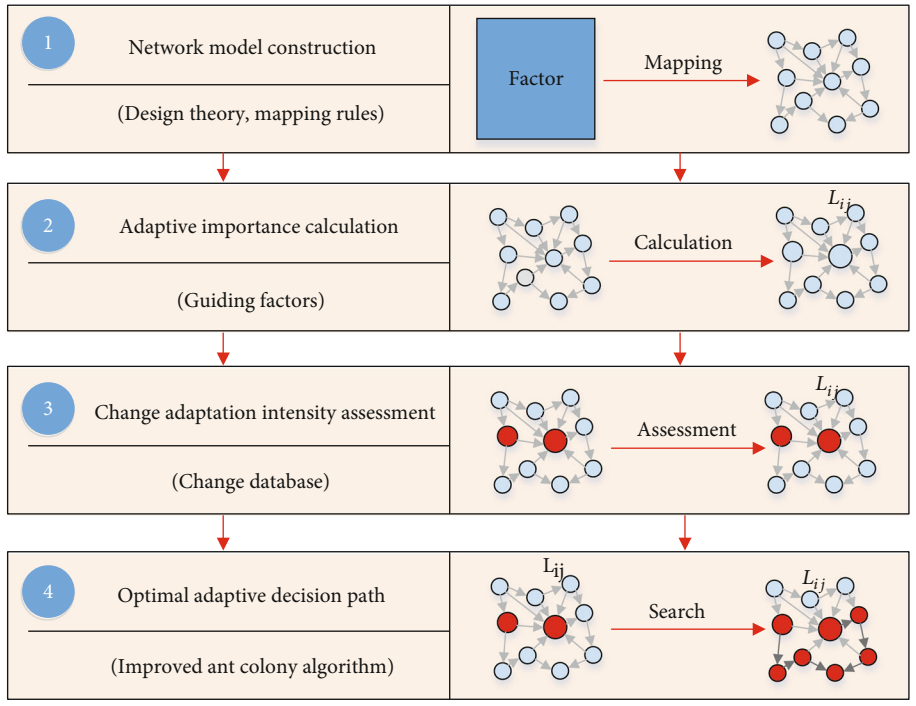
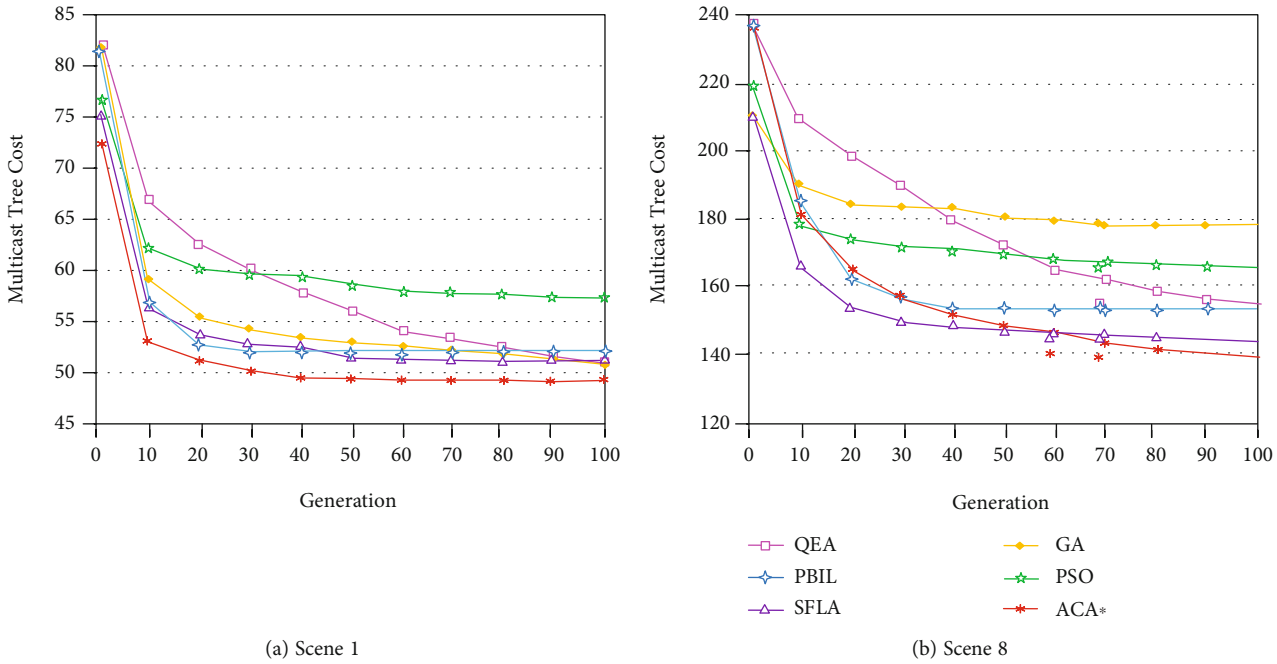



Figure 4: Framework of optimal adaptive optimization algorithm.

is to evaluate the connection importance of each edge by using the guiding factor. The third step is to obtain the data in the historical change database for analysis and calculate the change propagation probability of each side, then calculate the design change index according to the data in the historical change database, finally calculate the node connection importance, and calculate and evaluate the change propagation intensity of each side. The fourth step is to set change nodes of multiple sources. Firstly, the ability of each node to absorb changes is evaluated. Secondly, the ICI of the initial change node is set. Finally, the improved adaptive big data ant colony algorithm is used to solve the path, and the optimal propagation path is obtained.

## 3. Performance Evaluation and Physique Impact Analysis of Improved Adaptive Big Data Ant Colony Algorithm

*3.1. ACA* Performance Evaluation and Analysis.* This study uses the route convergence method to evaluate the performance of ACA*. This method refers to the process of route reestablishment, sending, learning, and stability when the topology of the network changes and notifies all related routes of the network of the change. In the experiment, the convergence of ACA* is compared with other five algorithms in eight scenarios. The statistical comparison between scenario 1 and scenario 8 is shown in Figure 5.




(a) Scene 1 (b) Scene 8

Figure 5: Convergence statistics of different algorithms in different scenarios.

It can be seen from Figure 5 that each algorithm tends to converge after iteration to 100 generations. The convergence process of the QEA algorithm is relatively slow, and there is a convergence sign only at 100 generations, while PBIL and SFLA algorithms tend to converge at 40 generations. ACA* is still searching for optimization when most other algorithms begin to converge. Therefore, ACA* can converge to a higher quality solution than other algorithms, and with the increase of algebra, ACA* can find a better solution than other algorithms. This is mainly due to the classification rule strategy used by ACA*, which makes it have stronger path propagation ability.

Figure 6 shows the average running time of several algorithms in the path optimization, and the results retain two decimal places. The running time of GA algorithm is the shortest, while that of ACA* is about twice that of GA algorithm. However, since the time-consuming data of the two systems are of the same order of magnitude, the time difference in actual operation is at most 2 seconds. If the performance of the computer used in the detection is higher, the time-consuming gap between them will be smaller. Combined with the data in Figure 7, the convergence speed of ACA* is significantly faster than GA algorithm, and the quality of the solution is higher. Therefore, ACA* outperforms the other five algorithms.

Figure 7 shows the standard deviation data of ACA* and other five algorithms after 20 times of optimization, with two decimal places reserved. As can be seen from Figure 7, the standard deviation of PSO algorithm is relatively large, while the standard deviation of ACA* in each scenario is lower than other algorithms, and the data difference is large. It can be proved that ACA* has a higher degree of stability.

Figure 8 shows the box chart statistics of ACA* and other five algorithms in each scenario. It can be seen from

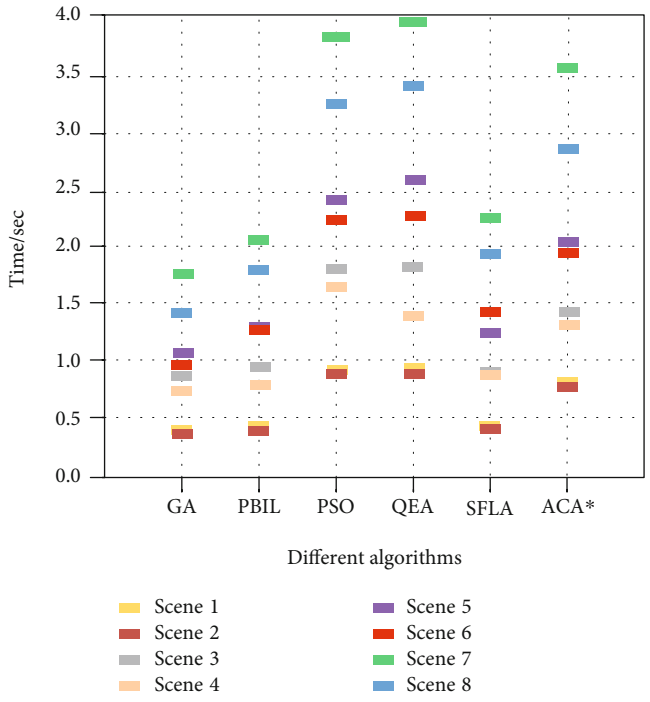



Figure 6: Statistical curve of average running time of six algorithms.

Figure 8 that the "data box block" and nonabnormal data area of ACA* are the narrowest in most scenarios, which proves that the result data of ACA* is relatively concentrated. ACA* has a lot of abnormal data in individual scenarios, mainly because most of the data is too concentrated; in reality, the gap between abnormal data and nonabnormal data of

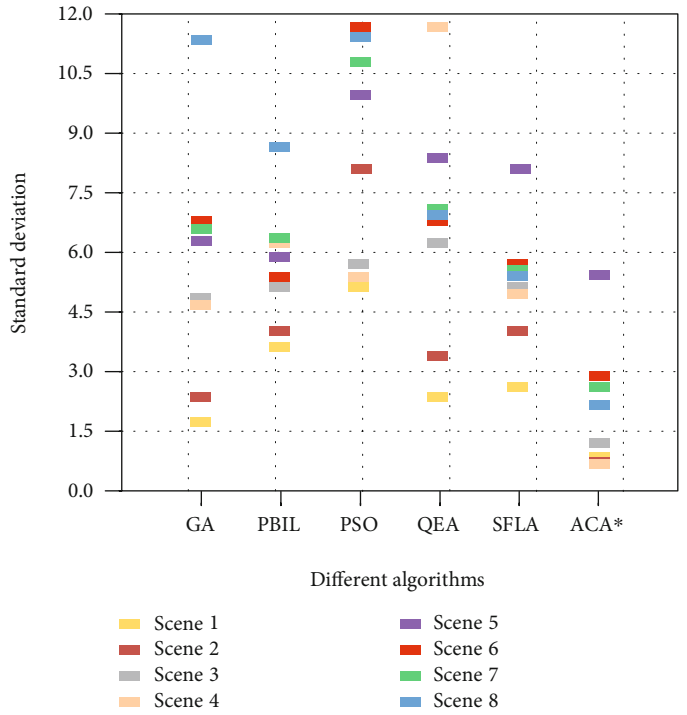



Figure 7: Six algorithms run in each scenario, and the standard deviation of the output multicast tree overhead value is 20.

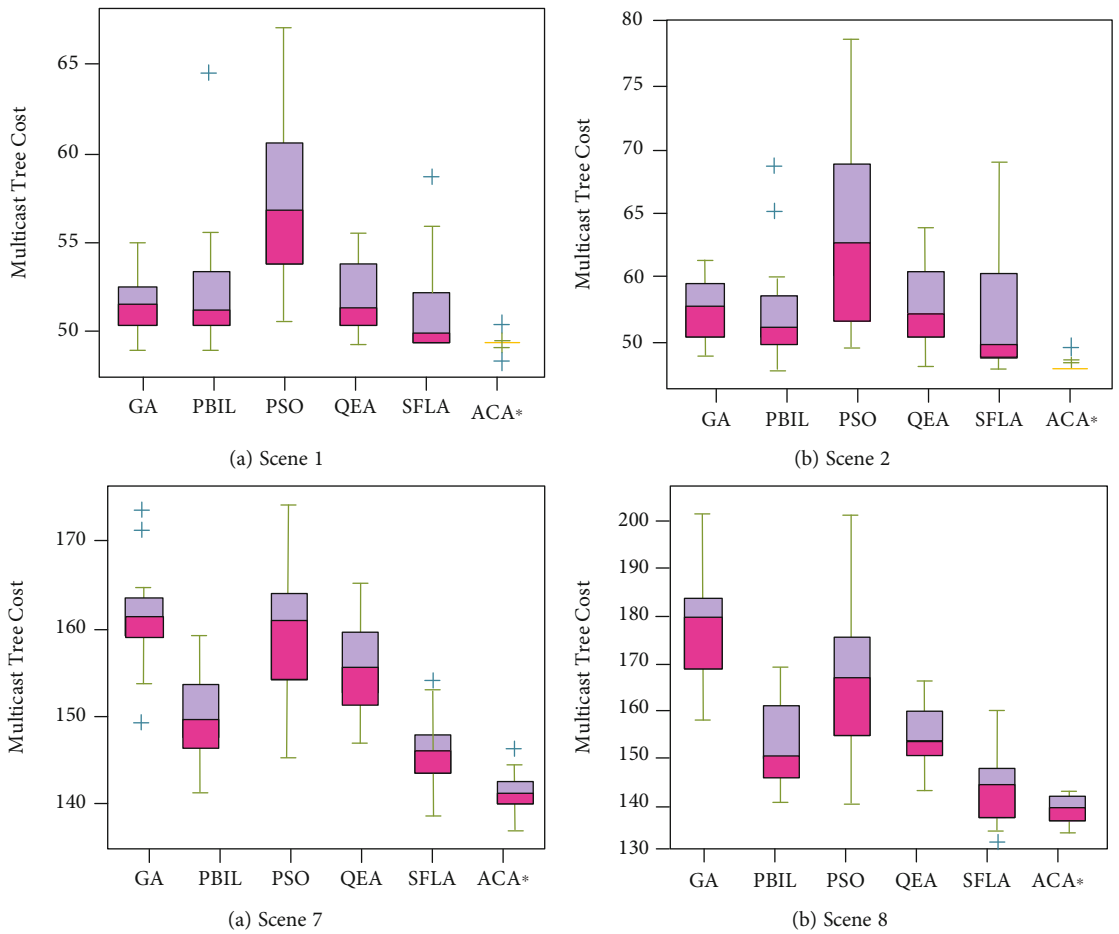



Figure 8: Box-plot statistics of six algorithms in different scenes.

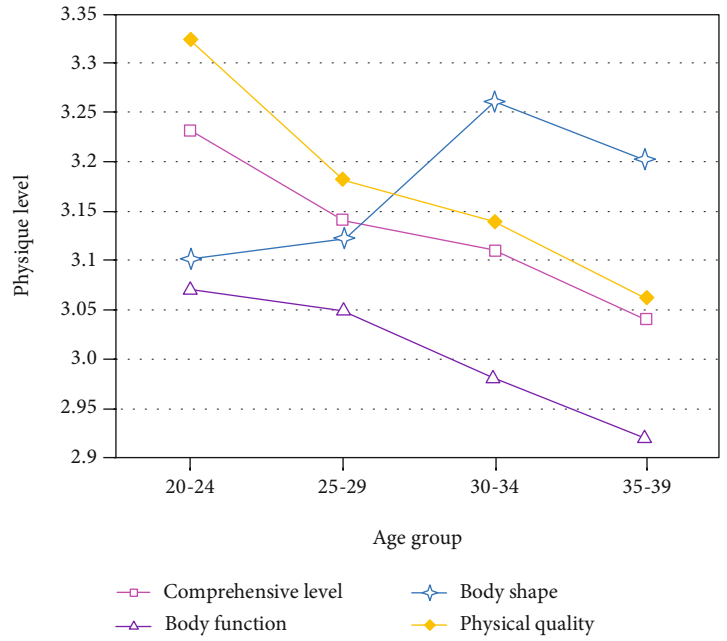



Figure 9: Average level of physical fitness of young people of different ages.

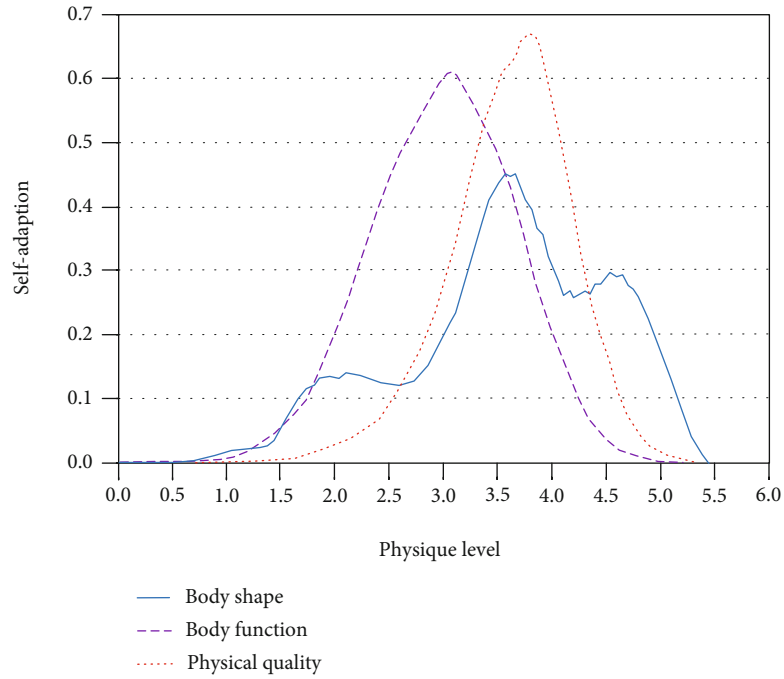



Figure 10: Comprehensive evaluation of physique level of body shape, body function, and body quality.

ACA* is very small. Combined with the data information of standard deviation in Figure 7, it can be concluded that ACA* is more stable than other algorithms.

*3.2. Scientific Evaluation of Human Health Constitution.* Firstly, the comprehensive scores of body shape, body function, and body quality of the members of a sports association in Chengdu in 2019 are obtained by big data analytic hierarchy process. From the distribution of physical fitness indexes, the skewness of all comprehensive indexes is slightly less than 0, reflecting that less than half of the members' physical fitness level after exercise is higher than the average level; the body shape kurtosis coefficient is also less than 0, and the distribution is steep, reflecting the difference before and after exercise; and the body quality coefficient kurtosis coefficient is not less than 0, and the distribution is flat, reflecting the difference before and after exercise. After that, there are significant differences in the quality level

among individuals, so that there are obvious differences in the comprehensive physique level. Among the members aged 20-39, the physical fitness level of each age group is shown in Figure 9.

With the growth of age, the physical fitness level of young people also continues to decline, mainly due to the greater influence of the law of physiological function development, but also closely related to the corresponding life and work. In terms of body shape, the average level of 20-29 years old youth is low, while the average level of 30-34 years old youth is high; the physical quality and skill level also decrease significantly with the increase of age.

The ACA* proposed in this study is applied to evaluate the comprehensive level of physical fitness of members of a sports association in Chengdu in 2019. The prediction of the physical fitness level of the members in three aspects of body shape, function, and quality in 2019 is shown in Figure 10. On the whole, the distribution of physical fitness reflects a stable trend, while the physical function and shape are on the right and steep, reflecting that the improvement of the physical fitness level of the members of the association in 2019 mainly depends on the result of the improvement of physical fitness level. In the future, we need to optimize the overall physical fitness by strengthening sports, changing living habits, and transportation modes.

## 4. Conclusion

This study mainly uses the improved big data adaptive ant colony classification rule algorithm to analyze the impact of sports on human health and physique. The results show that under the background of big data, compared with other algorithms, ACA* can converge to higher quality solutions, and with the increase of algebra, ACA* can find better solutions than other algorithms, which is mainly due to the classification rule strategy adopted by ACA*, which makes it have stronger path propagation ability; the convergence speed of ACA* is significantly faster than that of the other five algorithms, and the solution quality is high. The standard deviation of PSO algorithm is relatively large, while the standard deviation of ACA* in each scene is lower than that of other algorithms, and the data difference is large. It can be proved that ACA* has high stability; compared with other algorithms, ACA* has more obvious advantages in stability, optimization ability, running time, and convergence speed and is more suitable for practical application; in general, the improvement of the physical fitness level of the association members in 2019 mainly depends on the results of the improvement of the physical fitness level. In the future, it is necessary to strengthen the physical fitness, optimize the overall physical fitness, carry out sports, and change the living habits and transportation modes. Due to the limitation of time and ability, this study only selects eight scenes as the background to compare the six algorithms. However, this paper does not reflect the scientific impact of the algorithm on human health and physique. In the future, we need to test the performance of ACA* in more scenarios.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The author declares that no competing interests.

## References

[1] E. D. Koch, H. Tost, U. Braun et al., "Relationships between incidental physical activity, exercise, and sports with subsequent mood in adolescents," *Scandinavian Journal of Medicine and Science in Sports*, vol. 30, no. 11, pp. 2234–2250, 2020.

[2] A. A. Musawi, W. Al-Ani, and M. Al-Aghbari, "Impact of using m-health app on improving undergraduate students' sports and health habits and their attitudes toward its use," *E-Health Telecommunication Systems and Networks*, vol. 8, no. 1, pp. 1–9, 2019.

[3] R. de Santis, R. Montanari, G. Vignali, and E. Bottani, "An adapted ant colony optimization algorithm for the minimization of the travel distance of pickers in manual warehouses," *European Journal of Operational Research*, vol. 267, no. 1, pp. 120–137, 2018.

[4] M. Babanezhad, I. Behroyan, A. T. Nakhjiri, A. Marjani, A. Heydarinasab, and S. Shirazian, "Liquid temperature prediction in bubbly flow using ant colony optimization algorithm in the fuzzy inference system as a trainer," *Scientific Reports*, vol. 10, no. 1, p. 21884, 2020.

[5] S. G. Sutar, P. J. Mali, and A. Y. More, "Resource utilization enhancemnet through live virtual machine migration in cloud using ant colony optimization algorithm," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 79–85, 2020.

[6] N. Kusumahardhini, G. F. Hertono, and B. D. Handari, "Implementation of K-means and crossover ant colony optimization algorithm on multiple traveling salesman problem," *Journal of Physics: Conference Series*, vol. 1442, no. 1, article 012035, 2020(5pp).

[7] S. Asghari and N. J. Navimipour, "Resource discovery in the peer to peer networks using an inverted ant colony optimization algorithm," *Peer-to-Peer Networking and Applications*, vol. 12, no. 1, pp. 129–142, 2019.

[8] P. Ji, S. Zhang, and Z. P. Zhou, "A decomposition-based ant colony optimization algorithm for the multi-objective community detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 1, pp. 173–188, 2020.

[9] D. Sinwar, N. Sharma, S. K. Maakar, and S. Kumar, "Analysis and comparison of ant colony optimization algorithm with DSDV, AODV, and AOMDV based on shortest path in MANET," *Journal of Information and Optimization Sciences*, vol. 41, no. 2, pp. 621–632, 2020.

[10] J. Ning, Q. Zhang, C. Zhang, and B. Zhang, "A best-path-updating information-guided ant colony optimization algorithm," *Information Sciences*, vol. 433-434, pp. 142–162, 2018.

[11] R. Moeini and M. H. Afshar, "Hybridizing ant colony optimization algorithm with nonlinear programming method for effective optimal design of sewer networks," *Water environment research: a research publication of the Water Environment Federation*, vol. 91, no. 4, pp. 300–321, 2019.

[12] G. Yuan, J. Li, and Z. Hua, "Image matting trimap optimization by big data ant colony algorithm," *Multimedia Tools and Applications*, vol. 23, pp. 1–27, 2020.

[13] Z. Wang, J. du, Q. Miao, and R. Cheng, "Research of ant colony algorithm based on multi-objective optimization in cloud platform virtual machine initialization," *Journal of Physics: Conference Series*, vol. 1213, article 032005, 2019.

[14] M. Yu, G. Yue, Z. Lu, and X. Pang, "Logistics terminal distribution mode and path optimization based on ant colony algorithm," *Wireless Personal Communications*, vol. 102, no. 4, pp. 2969–2985, 2018.

[15] B. Zhao, H. Gui, H. Li, and J. Xue, "Cold chain logistics path optimization via improved multi-objective ant colony algorithm," *IEEE Access*, vol. 8, pp. 142977–142995, 2020.

[16] D. Jing-Guo, M. Geng-Sheng, and P. Wen, "Load distribution optimization based on max-min ant colony algorithm in hot strip rolling process," *Metallurgist*, vol. 62, no. 7-8, pp. 837–846, 2018.

[17] J. Kuidong, M. Zhanli, C. Haonan, and O. Eryao, "Optimization of evacuation route selection for personnel in smoke environment based on big data ant colony algorithm," *Safety production science and technology in China*, vol. 14, no. 11, pp. 133–137, 2018.

[18] W. Wang and Z. Cai, "Optimization of linear consecutive-k-out-of-n systems with Birnbaum importance based ant colony optimization algorithm," vol. 25, no. 2, pp. 253–260, 2020.

[19] J. Li, Y. Xia, B. Li, and Z. Zeng, "A pseudo-dynamic search ant colony optimization algorithm with improved negative feedback mechanism," *Cognitive Systems Research*, vol. 62, pp. 1–9, 2020.

[20] H. Zhao, C. Zhang, and B. Zhang, "A decomposition-based many-objective ant colony optimization algorithm with adaptive reference points," *Information Sciences*, vol. 540, pp. 435–448, 2020.

[21] T. R. Gadekallu, D. S. Rajput, M. Reddy et al., "A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU," *Journal of Real-Time Image Processing*, vol. 18, no. 4, pp. 1383–1396, 2021.

[22] G. V. R. Bright, K. Raimond, and J. Lovesum, "A novel approach for automatic remodularization of software systems using extended ant colony optimization algorithm," *Information and Software Technology*, vol. 114, pp. 107–120, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9601078, 12 pages
https://doi.org/10.1155/2021/9601078

*Research Article*

# Energy-Efficient Enhancement for the Prediction-Based Scheduling Algorithm for the Improvement of Network Lifetime in WSNs

**Md. Khaja Mohiddin**,[1] **Rashi Kohli**,[2] **V. B. S. Srilatha Indira Dutt**,[3] **Priyanka Dixit**,[4] **and Gregus Michal**[5]

[1]*Department of ET&T, Bhilai Institute of Technology, Raipur, Chhattisgarh, India*
[2]*Senior Member (IEEE), New York, USA*
[3]*Department of ECE, GITAM (Deemed to Be University), Visakhapatnam, India*
[4]*Department of CSE, UIT, RGPV Bhopal, Madhya Pradesh, India*
[5]*Department of Info. Systems, Faculty of Management, Comenius University, Bratislava, Slovakia*

Correspondence should be addressed to Md. Khaja Mohiddin; khwaja7388@gmail.com, Rashi Kohli; rashikohli@ieee.org, and Gregus Michal; michal.gregusml@fm.uniba.sk

Received 6 July 2021; Accepted 15 September 2021; Published 31 October 2021

Academic Editor: Rajesh Kaluri



In wireless sensor networks, due to the restricted battery capabilities of sensor nodes, the energy issue plays a critical role in network efficiency and lifespan. In our work, an upgraded long short-term memory is executed by the base station to frequently predict the forecast positions of the node with the help of load-adaptive beaconing scheduling algorithm. In recent years, new technologies for wireless charging have offered a feasible technique in overcoming the WSN energy dilemma. Researchers are deploying rechargeable wireless sensor networks that introduce high-capacity smartphone chargers for sensor nodes for charging. Nearly all R-WSN research has focused on charging static nodes with relativistic routes or mobile nodes. In this work, it is analysed how to charge nondeterministic mobility nodes in this work. In this scenario, a new mechanism is recommended, called predicting-based scheduling algorithm, to implement charging activities. In the suggested technique, it directs them to pursue the mobile charger and recharge the sensor, which is unique for the present work. The mobile charger will then choose a suitable node, utilizing a scheduling algorithm, as the charging object. A tracking algorithm based on the Kalman filter is preferred during energy transfer to determine the distance needed for charging between the destination node & mobile charger. Here, the collecting & processing of data are performed through the big data collection in WSNs. The R-WSN charging operations of nondeterministic mobility nodes will be accomplished using the proposed charging strategy.

## 1. Introduction

WSN is a well-organized environment consisting of a large number of microactive nodes spread dynamically across the monitoring area through wireless broadcasting. With its significant importance in the analysis of armed forces, heath support, inventory control, atmospheric tracking, horticulture & effective promotional fields, WSNs have been found out to be the most significant computer research & communication technology. Sensor nodes rely on the supply of battery power, their broadcast functionality, and very limited energy storage capacity, so the need to synchronize the energy usage of the network along with the improvement of network lifetime came into being to use the vitality of the nodes more precisely.

The sensor framework comprises of multiple nodes, only a small number of which have various functions like detecting, transmission, and networking, which are spread around

the nodes within or very close to the sensor. This is where each of the above-listed nodes gathers the data and routes it back to either a sink or base station (BS). At present, WSN is the most widely used current networking solutions to provide sensed retrieved information to the base station with insufficient power capacity. Care needs to be taken in order to check all potential climate conditions (including the abovementioned ones), as well as temperature, moisture, lightning scenario, pressure, soil composition, vehicular development, noise densities, target field imaging in military areas, emergency management, fire alarm sensors, criminal, and surveillance hunting are all important in WSN environments.

The WSN consists of one or perhaps many BS's & several sensor nodes that are positioned in a wide area for cooperative monitoring of a physical environment. The WSN has different application scenarios such as ecological surveying, smart city, and wildlife surveillance because of its characteristics of low cost & strong self-organization. The capacity of node batteries also limits the operating duration of nodes, affecting WSN lifespan and efficiency, which has become a critical issue limiting WSN's growth [1–4].

For data collecting, there is an increasing trend toward large-scale sensor networks. Despite the fact that they are a new generation of sensor networks, their application is constrained by a number of factors, including adaptability to conventional network scaling methods. To allow a network to effectively play its role, a number of significant difficulties must be overcome. It might cite the appropriate placement of sinks, as well as the reduction of sensor energy consumption and lifetime, as examples of these issues. Large-scale WSNs (LS-WSNs) may give valuable solutions for big data collecting in a big data setting, with a massive volume of high-speed, regularly changing, and changeable data [5].

WSN has done a lot of research and analysis on how to overcome the energy dilemma in recent decades. Most of the previous study can be classified into three techniques: energy saving systems, energy harvesting techniques, and approaches to wireless charging. Only energy efficiency can be enhanced by energy saving systems, without making up for node energy consumption. Therefore, the energy saving systems in WSN can only mitigate the energy crisis in the sensor environment by considering various approaches and algorithms to minimize the utilization of the energy consumption [6–8].

The performance of energy converters varies substantially according to environmental conditions. Wireless power transmission has been shown in investigations over the last two decades, allowing sensor nodes to be recharged wirelessly. Magnetic resonance coupling has gotten a lot of attention from researchers because of its high transmission power and lengthy transmission distance. Wireless power transfer is accomplished by developing a sophisticated network architecture in which sensor nodes are wirelessly charged utilizing chargers with large batteries, such as (R-WSN) [9].

The majority of R-WSN investigators are interested in scheduling algorithms based on a mobile charger (MC), which configure the energizing sequence of stable sensors.

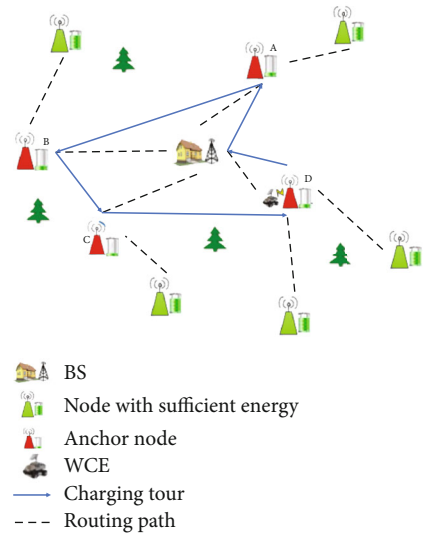



Figure 1: Network model.

Some latest researches have looked into the R-WSN using mobile sensor nodes. A tree-based method for charging mission-critical robots that decreases the MC's path length without causing robot energy depletion is suggested [10]. The complexity of the location of the static loading pile has been studied so that mobile sensing nodes can come closer when energy is exhausted [11]. Sensor node projections are stochastic, so that node activities can be monitored, according to these investigations. Wildlife tracking is a well-known scenario, in which sensor nodes are connected to each tracked species [5]. As a consequence, nodal mobility is uncontrollable & nonstochastic, even though the network's model of mobility is the goal.

In this scenario, an appropriate strategy is to identify a few static spots in the domain region that are known as hubs that are routinely contacted by sensors and then position the mobile charger to stay and load sensors at these hubs using an RL-based process. Although hotspots are never changed once they have been selected, this technique is unable to adapt to changes in nodal movement patterns. Furthermore, in this method, the BS directs the MC to approach the hubs, implying that the MC must stay for a minimum duration in each hub, even though all of the sensors that may connect directly have been charged [12].

Here, it investigates, how nonrandom mobility nodes can be energized. It also assists the MC in seeking the sensor immediately in any charging activity, rather than depending on sensors at fixed positions as in [12], so that it may respond to potential changes in nodal trajectories. It must first tackle the three concerns in order to have chargeable functionality in this case listed as follows:

*1.1. Node Finding Issue (NFI).* The MC must first locate nodes before charging them. The MC does not know the present position of the individual sensor due to its nondeterministic mobility of sensors. Furthermore, due to R-WSN's constrained resources, it is unable to use the framework for cellular networks throughout cloud environments in our

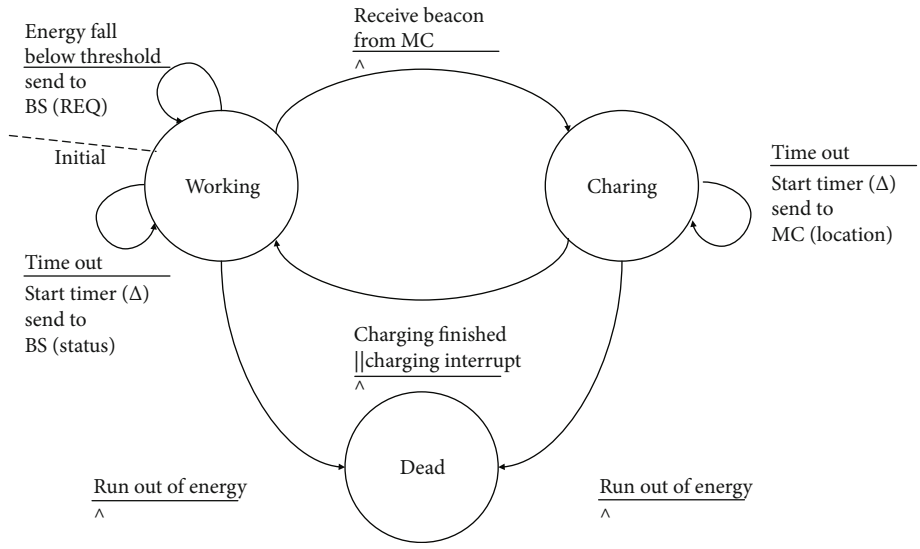



Figure 2: FSM for sensor nodes.

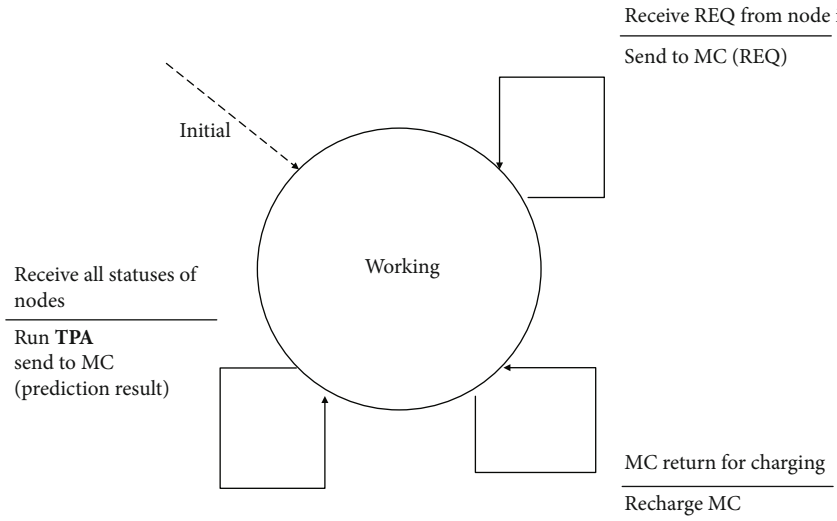



Figure 3: FSM for base station.

study. If data packets are interacting and the MC interchanges to the indicated positions, sensor nodes can depart the sites, causing efficiency to deteriorate in a short span of time. The lack of knowledge on the positions of nodes is the most significant problem in our case. It needs to figure out how to find out where nodes will be in the future so that the MC can reach & charge them [13].

*1.2. Node Election Issue (NEI).* Following the acquisition of sensor node positions, the MC must choose a suitable source to charge within the constraints of battery available power. The NEI problem is similar to the static WRSN scheduling problem. In this situation, however, the MC only charges each node at a row for NEI, but in static R-WSN, the BS directs the MC to connect a cluster of nodes at once. Once the present charge target is fulfilled, the MC initiates an unique charge task.

*1.3. Node Protection Issue (NPI).* It is well known that two basic criteria for wireless power transmitting are that the MC and target node must be relatively close together and the transmission must be able to continue for a period of time. In order to fulfil these two conditions, the MC must accompany the target node while it's being energetically stimulated. In spite of the nondeterministic mobility of the nodes and position, the positions of the nodes can be changed; however, it's impossible to know from where the nodes are. To find a moving node, the MC must track it.

With nondeterministic versatility, there exists a multitude of problems. In the method proposed by the NFI, it is suggested that a predictive method is used to forecast potential node positions. However, in the method prescribed by NEI, it evaluates charges depending on predicted results. During the energy transmission, the MC must also follow the destination node. For this reason, charging nondeterministic nodes must be done on an assumption that they will switch unpredictably.

A wireless sensor network (WSN) is an ad hoc network made up of a series of sensor nodes that are arbitrarily fixed

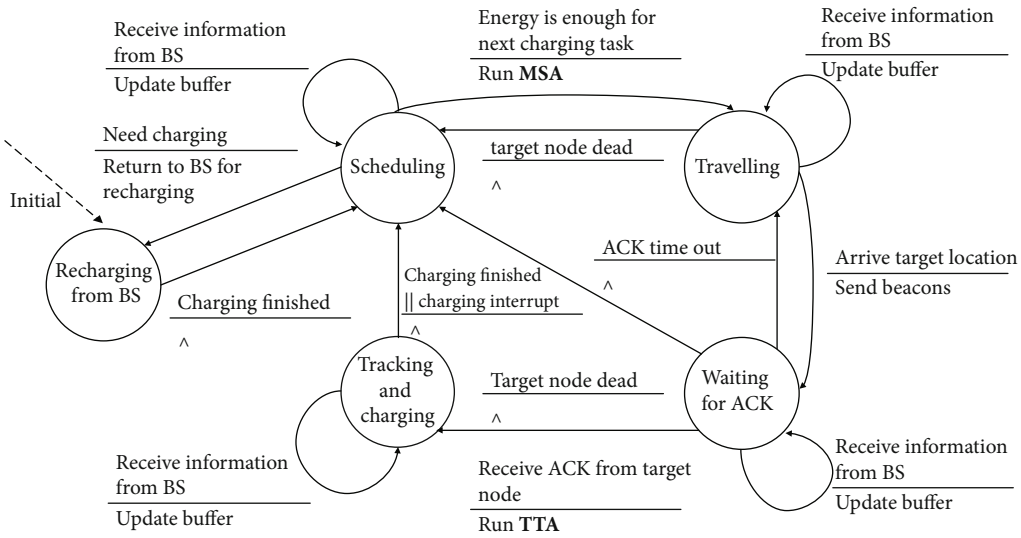



Figure 4: FSM for mobile charger.

**1st Algorithm**: -Choosing the most similar segment
i $T_Y \leftarrow T_X$ & $\Phi_t \leftarrow 0$
ii Calculates the tiniest region $RGN_X$ which includes all the positions in $T_X$
iii for every node $n_Y \in N - \{n_X\}$ do
iv for ($t_0 = 0$; $t_0 < t - J + 1$, $t_0 = t_0 + \Delta t$) do
v if $l_Y(t_0)$ in $RGN_X$ then
vi Calculates the tiniest region which includes all the positions in $T_Y^{t_0}$ starting with $l_Y(t_0)$ of length J
vii if $RGN_X = RGN_Y^{t_0}$ then
viii Estimate the Spearman's Coefficient $\Phi_{spearman}$ of $T_X$ & $T_Y^{t_0}$
ix if $\Phi_{spearman} > \Phi_t$ then
x $T_Y \leftarrow T_Y^{t_0}$ and $\Phi_t \leftarrow \Phi_{spearman}$
xi end if
xii end if
xiii end if
xiv end for
xv end for
xvi Result: $T_Y$ and $\Phi_t$

Algorithm 1: Trajectory prediction algorithm.

or scattered in a specific geographical area and communicate over a wireless link to gather, evaluate, and transfer data in their area to a special node called a sink. The region refers to the geographical area in which the sensor nodes function. Sensor nodes in a WSN can self-organize to gather data about the environment in which they are installed. Based on the nature of the installed application, the collected data might be transmitted on a regular basis or on an as-needed basis. The sink is a node with two or more network interfaces that connect the WSN to the end-network, user's (for example, a local area network or the Internet). The user can use the sink to request access to other nodes in the network, such as specifying the type of data to be collected. The fundamental architecture of a WSN is shown in Figure 1 for example purposes [5].

Here, it suggests a charging scheme called predicting-based scheduling algorithm (PSA) based on the above study, which includes three algorithms to solve the current research challenges. This is the first time that a chase technique has been deployed to energize nodes with nonrandom mobility, to our knowledge. To be more precise, the following are the major factors that may affect our work:

(1) The issues of energizing nodes with nonrandom mobility have been resolved and are now coordinated

(2) It first offers a new LSTM methodology for evaluating the precise positions of sensor nodes, which uses the prior trajectory to forecast the forthcoming positions of each sensor node

(3) In order to maximize the network's efficiency, it offers a node identification method for selecting the optimal node as the destination node based on the forecasting findings and the energy level of each node

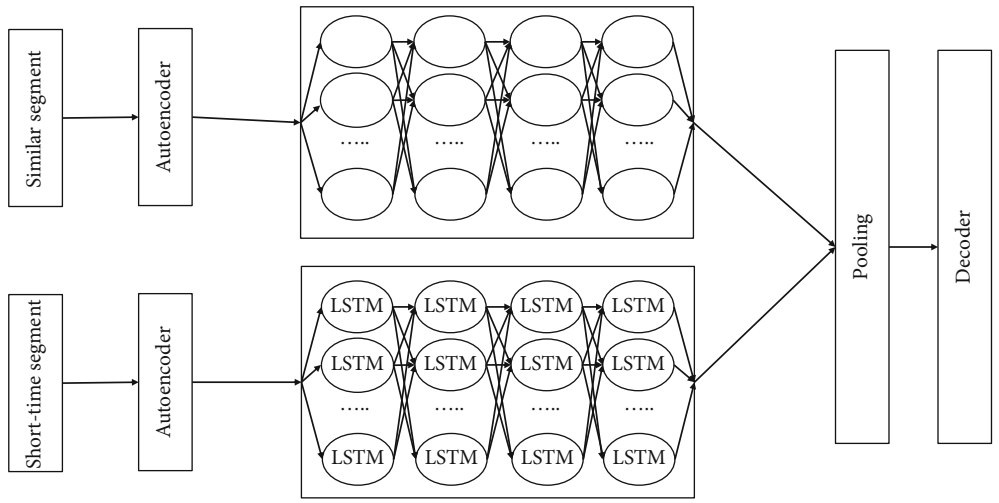



Figure 5: Model of the hybrid LSTM scheduling algorithm.

**2nd Algorithm**:-Choosing the destination sensor for charging
i **while** the Mobile Charger barrier for RREQ is not void **do**
ii Search the sensor $S_P$'s RREQ in the mobile charger barrier with minimum energy status & remove $S_P$'s RREQ, $S_J \leftarrow S_P$
iii **if** $S_J$ satisfies **then**
iv Return $S_J$
v **end if**
vi **end while**
vii The mobile charger beams
viii if send $S_J$ reacts to the mobile charger & recent charged sensors aren't on the list **then**
ix $S_J \leftarrow S_Y$
x **else:**
xi Search the sensor node $S_X$ with high energy necessity $E_{gain}(x)$in its accessibility
xii $S_J \leftarrow S_X$
xiii **end if**
xiv **if** $S_J$ satisfies **then**
$S_J \leftarrow$ Base Station
xv Output: $S_J$

Algorithm 2: Mobile charger-based scheduling algorithm.

(4) Ultimately, in order to fulfil the demands of wireless charging, it implements a Kalman filter-based tracking technique that directs the MC to monitor mobile target nodes during energy conversion

## 2. Related Work

Current charging schemes can be divided into two groups based on sensor nodes' motion states in R-WSN: static sensor nodes and mobile sensor nodes.

During the last ten years, investigators have suggested a number of charging techniques for R-WSN stable sensor nodes to determine the best node sequence for recharging MCs. A periodic system was proposed, for example, in circumstances where sensor nodes are distributed evenly or nonevenly [14]. The MC determines the minimum round route that connects all of the sensor nodes, then regulates the routes & energizes the sensor node with the most energy. The charging problem was developed as a vehicle routing problem in [15], with the minimum Hamiltonian cycle as the solution, taking into account the placement and condition of sensor nodes. Another key study proposes a methodology of distributing many MCs to sustain performance levels of R-WSN of life-critical sensor nodes [16]. It is also proposed a novel idea called "shuttling," as well as an optimal charging mechanism that used charger collaboration to reduce the number of chargers and increase the sensing range [17].

In following work on the optimization of load features and reduction of the transmission range of MC [18], a mechanism was presented to load several sensor nodes sequentially entirely inside the same transmission range. This strategy enhances the charging effectiveness and device effectiveness of WRSNs compared to standard techniques. Research in this paper compares the chargeability of unique-frequency loaders with diverse-frequency loaders when determining the position of wireless chargers that simultaneously load a distinct WSN

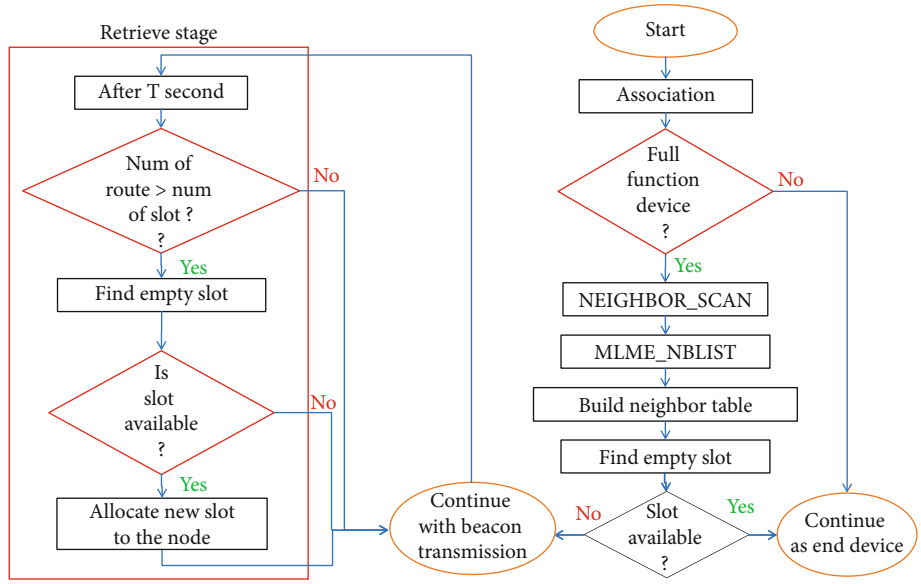



Figure 6: LAB scheduling algorithm.

[19]. This paper provides a smooth logic-based approach that is suggested for on-demand charging in a dense R-WSN [20]. They have used a foggy logical method to develop an ideal on-demand charging package for a dense R-WSN, taking account of the available energy, range to the MC, and key node density. According to the previous study [20, 22], many MCs were established for on-demand WRSNs by load-balanced network dividing and dynamic recharge limitations were determined for sensor nodes.

In this, most previous research concluded that sensor node trajectories are random. For illustration, a way for regulating the operating power of the stable charging stack for charging sensor nodes is identified [21, 22]. The researchers proposed a way of assigning the static recharge batteries to mobile nodes with essential power, while flexible nodes take over the recharge node duties effectively [11]. Research in this paper submits a tree-based operating robotic charge plan that places the MC on a path of depleted sensor nodes with zero reduction of robot energy to keep R-WSN operating properly [10. 24].

This paper suggests a primary study to have an overview of the usage of an energy-restricted MC to charge nonrandom mobile sensors [12]. The objective of this survey was to identify certain fixed spots in which sensors are often located and then to deploy the MC to stay and load at these spots. Although, in order to find these positions, this method necessitates a significant amount of historical trajectory data and is unable to adjust to changes in the nodal mobility model.

## 3. Methodologies

So far, in order to energize nodes with a nondeterministic behaviour, three concerns have been explored and formulated. It proposes our accounting strategy in this chapter, which will be utilized to solve three problems at a time. Three algorithms are included in the proposed charge technique for addressing these problems: the trajectory prediction algorithm, the MC-based scheduling algorithm & target tracking algorithm (TTA). Each time, the BS performs TPA to estimate possible sensor placements and provides the forecasting results to the MC in the required manner. The BS information is buffered in the MC. Using the protected data to estimate the next loading destination or direct to the BS, MC conducts MSA during the departure or the

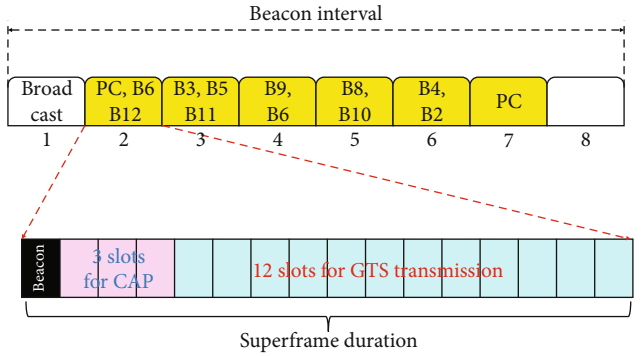



Figure 7: Superframe specification of LABS.

Table 1: Simulation parameters.

| Simulation parameters | |
|---|---|
| Parameter | Value |
| $L_{grid}$ | 30 m |
| $N$ | 20 |
| $E_{sensor}$ | 8000 J |
| $E_{mc}$ | 300 kJ |
| $\Delta t$ | 25 s |
| $\Phi$ | 0.3 |
| $R_{com}$ | 250 m |
| $\Delta$ | 3 s |
| $\lambda$ | 0.6 |

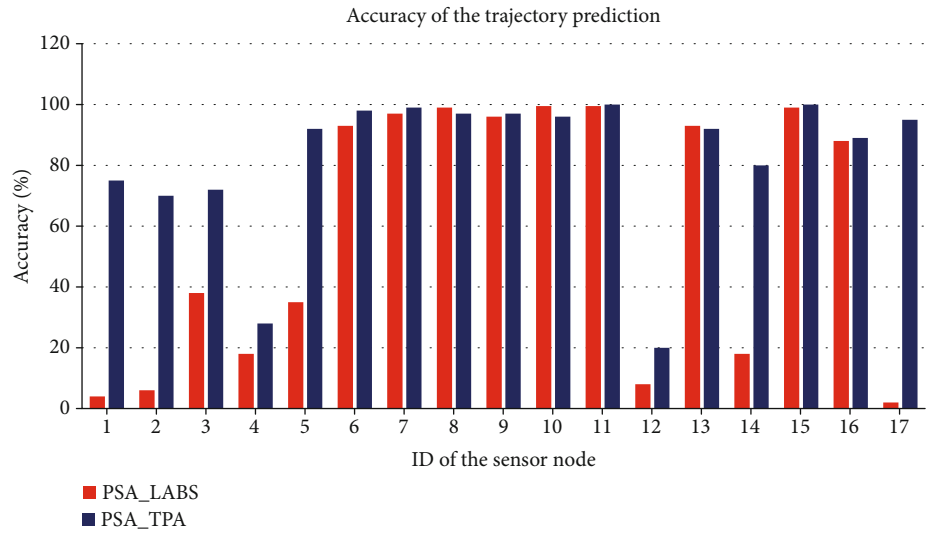



Figure 8: Accuracy of the LABS & TPA.

completion of a charging task. The MC advances to the forecasting position of the destination node once the destination node is located. When the MC reaches the target node, the TTA is used for tracking it during energy transfer.

Figures 2–4 display the FSM concepts for the BS, MC, and sensor nodes, respectively, to better explain how the developed system operates. The FSM in Figure 2 depicts sensor node operations, whereas the FSMs in Figures 3 and 4 depict BS and MC operations, individually. The axes in the FSM specifications represent the conversions of the systems from one phase to the next. The event that activates the switch is indicated above the straight axis that denotes the transition, while the actions conducted during the event are indicated under the straight axis.

*3.1. Trajectory Prediction Algorithm (TPA).* To deal with NDP, it is needed to develop a node trajectory time sequence forecasting approach. There are several time series prediction experiments available today, such as exponential smoothing (ES), moving average (MA) & autoregressive integrated moving average model (ARIMA). Unfortunately, due to the difficulty of node mobility, these methods might not work well in our case. As a consequence, it turns to neural network techniques and the algorithm is shown in Algorithm 1.

*3.2. Hybrid LSTM Predicting Algorithm.* The LSTM technique is used in TPA because of its integral gain in analyzing sequence data and predicting [24–26]. Calculating an integrated mobility prototype for the entire sensor nodes would be exceedingly time consuming because individual nodes have distinct movement patterns. It is currently attempting to determine the mobility prototype for each sensor node individually. Because all sensor nodes are part of the same sensor network, they must have some spatial and temporal consistency. The forecasting model's convergence will be accelerated if this type of regularity is applied.

The neural framework primarily consists of an LSTM block that extracts mobility patterns from short sections and a highly interconnected block that differs substantially regarding future moves from the furthermost identical segment [27–30]. The LSTM block is made up of an autoencoder layer & 4 LSTM layers, each with 32 LSTM cells, while the completely integrated block is made up of an autoencoder layer & four highly integrated layers, each with 32 neurons as shown in Figure 5. The performance from the LSTM and completely connected blocks is mixed in the correct ratio using a pooling layer. A decoder layer predicts the performance of the pooling layer.

*3.3. Mobile Charger-Based Scheduling Algorithm (MSA).* As the NEI is a top-down multiobjective recursive challenge with the first goal of reducing the percentage of dead nodes, MSA seeks to decrease the count of dead nodes initially and then the charging transit time. As a result, the node election method looks like this:

First, MSA examines the RREQ messages it has issued, since $REQ_X$ indicates that $S_X$ is experiencing an energy alert. In the buffered $REQ_X$, the MC selects the $S_X$ sensor with the least power [31]. This will then be monitored by the mobile charger. The $S_X$ sensor is loaded by the MC when the limit is reached. Otherwise, in buffered messages, the MC selects the RREQ node with minimum energy as shown in Algorithm 2.

When no appropriate node is chosen based on RREQ information, MSA attempts to choose the destination node for the lowest requested journey distance. MC casts beacons to check for neighbouring sensor nodes and selects the basic station with minimum remaining energy between those who react on MC. A list is being used to record the sensors powered by the MC currently, in order to prevent charging the sensors around frequently. As a result, the messages from these awaiting sensor nodes will be ignored by the MC.

*3.4. Target Tracking Algorithm (TTA).* The MC will interact with the destination node for current movement information until it is associated with the target node [32]. The MC must guide the target node throughout the entire energy

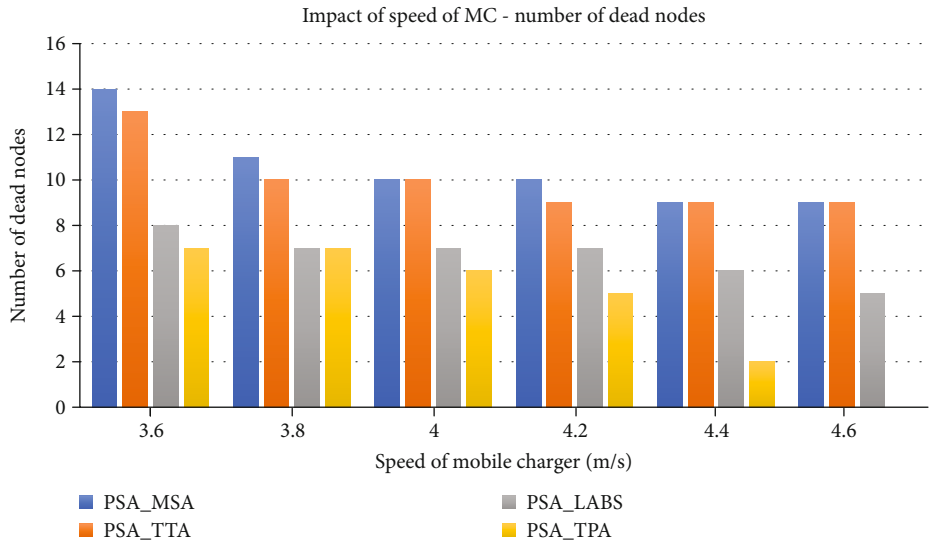



Figure 9: Impact of speed of MC—no. of dead nodes.

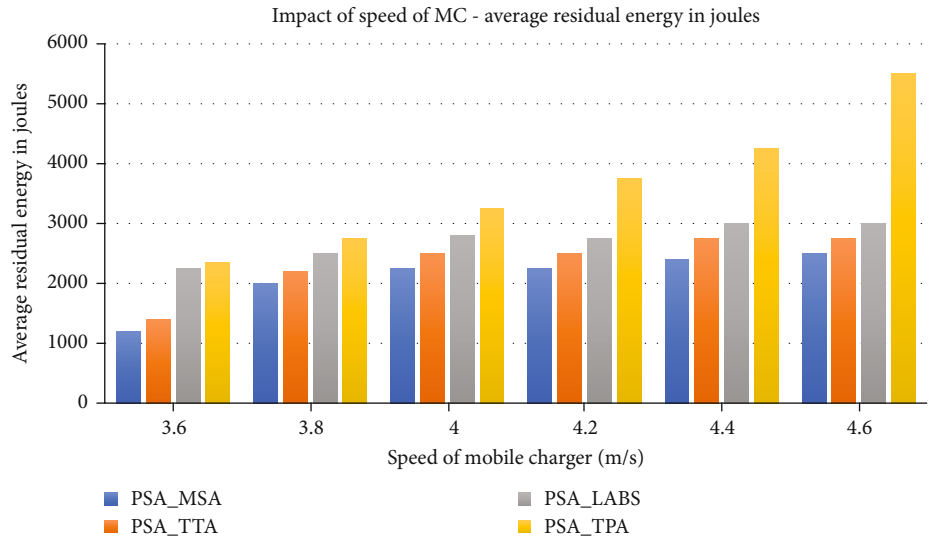



Figure 10: Impact of speed of MC—average residual energy.

transfer process in order to complete the mission. This goal is recommended to be accomplished using the target tracking algorithm.

Since the prediction phase "$\Delta$" is tiny, it typically considers the target node's mobility to be a linear trace. The Gaussian noise can be approximated when it comes to GPS position error. To predict the position of the target node, the Kalman filter is used, which has excellent short-time prediction efficiency.

*3.5. LAB Scheduling Algorithm (LABS).* This algorithm is important for mesh topology and preventing beacon collisions. There are two sections of this algorithm as shown in Figure 6. Initially, the nodes will be connected with one another, after which the neighbouring nodes will be examined to determine the energy information required for proper scheduling. The retrieval stage follows, in which the time slot is mapped to a variety of routes between the source and sink nodes as shown in Figure 7.

## 4. Results and Discussion

Table 1 refers to the simulation scenario preferred for the result parameter calculation. Figure 8 represents the accuracy comparison between LABS and TPA, where the prediction-based scheduling algorithm corresponding to the LABS performs better. Figures 9 and 10 represent the impact of the speed of the mobile charger relevant to the number of dead nodes and average residual energy where MSA, TTA, LABS & TPA are compared. Figures 11 and 12 represent the impact of battery capacity relevant to the number of dead nodes as well as average residual energy in which prediction-based scheduling algorithm based on TPA provides good results. Finally, Figures 13 and 14 represent the

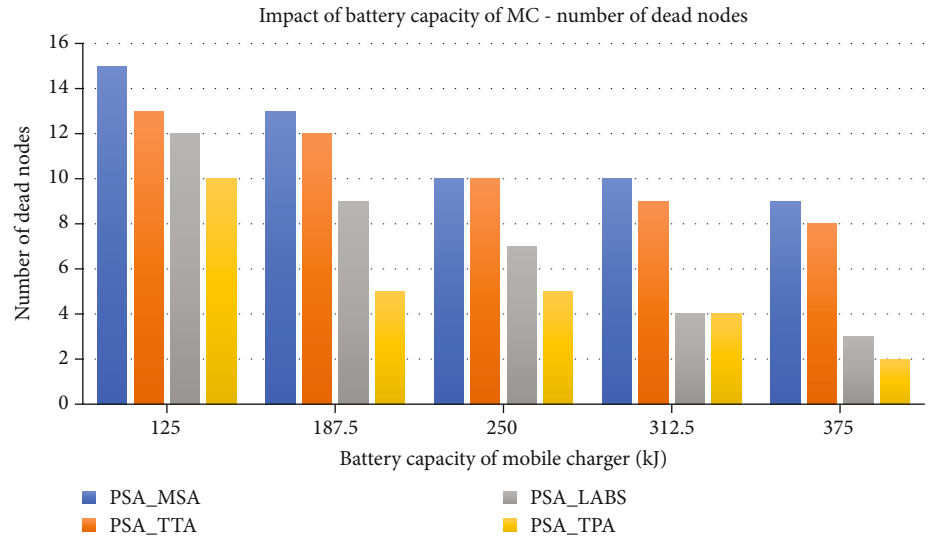



Figure 11: Impact of battery capacity of MC—no. of dead nodes.

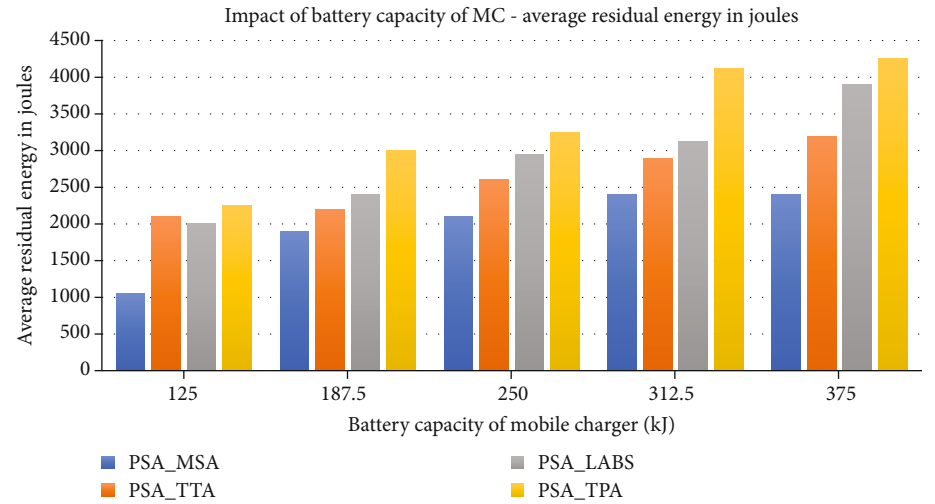



Figure 12: Impact of battery capacity of MC—average residual energy.

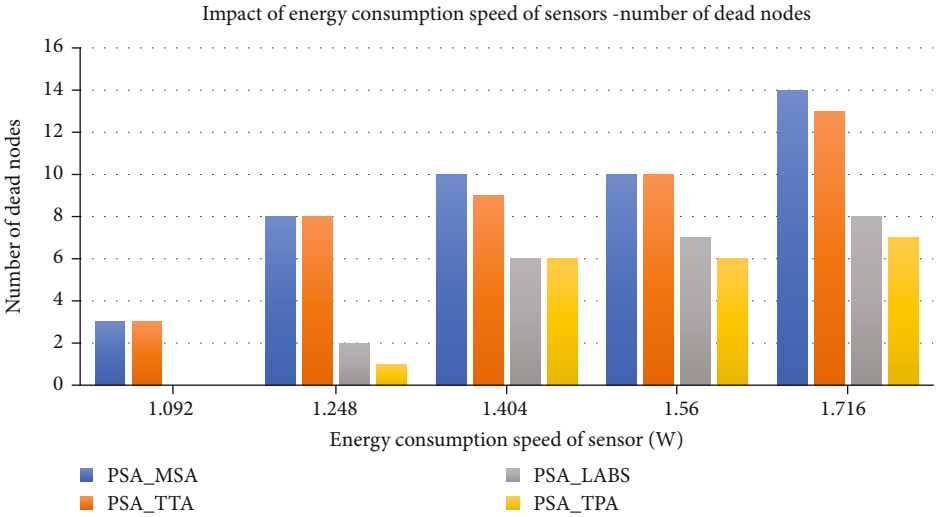



Figure 13: Impact of energy consumption speed of sensors—no. of dead nodes.

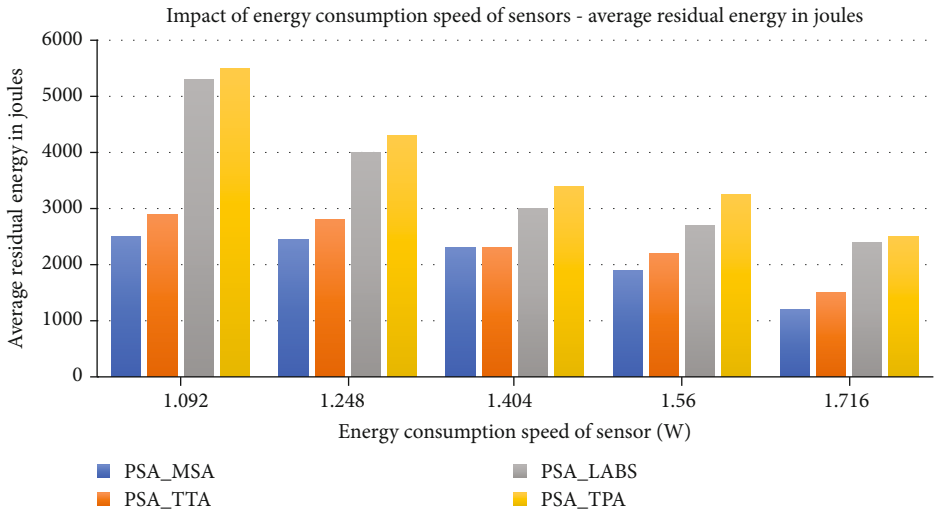



Figure 14: Impact of energy consumption speed of sensors—average residual energy.

impact of energy consumption of the speed of sensors based on the number of dead nodes and average residual energy which are compared and result in good performance of TPA.. It has been observed that TPA performs good when compared to the other scheduling techniques.

## 5. Conclusions

In this work, how the mobile charger of restricted energy charger can charge the nodes with undetermined mobility has been observed. Three major issues have also been discussed for supplying the charging service like node finding issue, node election issue, and node protection issue. A prediction-based scheduling algorithm has been suggested which consists of trajectory prediction algorithm, mobile charger-based scheduling algorithm, load-adaptive beaconing scheduling, and target tracking algorithm to address the abovementioned issues. Finally, the performance of the prediction-based scheduling algorithm has been evaluated with respect to the existing through simulations. The proposed scheme of the prediction-based scheduling algorithm in interfacing with the trajectory prediction algorithm outperforms good and monitors the rechargeable-wireless sensor network system in an efficient state. However, the trajectory prediction algorithm technique depends on a neural system approach and it takes a very long time to develop the mobility patterns in individual sensor nodes. Consequently, a network with large numbers of nodes in an environment cannot achieve the proposed technique at all. The trajectory prediction algorithm can therefore be preferred as a future work, together with several mobile chargers by processing the data collection through the big data in WSNs. Also, the proposed technique can be applied in the application health care monitoring in patients, earth/environmental sensing to detect natural disasters, industrial monitoring, threat detection for detecting ground-based nuclear devices, area monitoring to detect enemy intrusion, data transfer in power systems, etc.

## Acronyms

WSN: Wireless sensor networks
R-WSN: Rechargeable wireless sensor networks
PSA: Prediction-based scheduling algorithm
LSTM: Long short-term memory
LABS: Load-adaptive beaconing scheduling
BS: Base station
LS-WSN: Large-scale wireless sensor networks
MC: Mobile charger
NFI: Node finding issue
NEI: Node election issue
NPI: Node protection issue
TTA: Target tracking algorithm
MSA: Mobile charger-based scheduling algorithm
TPA: Trajectory prediction algorithm
ES: Exponential smoothing
MA: Moving average
ARIMA: Autoregressive integrated moving average
RREQ: Route request
$L_{grid}$: The total distance of the grid
$N$: The number of sensor nodes
$E_{sensor}$: The battery capacity of the sensors
$E_{mc}$: The battery capacity of the MC
$R_{com}$: The communication range between the MC and the nodes
$_{\Delta t}$: The time interval between the node's request to the BS
$_{\Phi}$: The node's energy level
$_{\Delta}$: The time interval between the node's request to the MC
$\lambda$: Wavelength.

## Data Availability

All data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References


[1] Y. Li, L. Zhong, and F. Lin, "Prediction-scheduling-tracking: charging nodes with non-deterministic mobility," *IEEE Access*, vol. 9, pp. 2213–2228, 2020.

[2] H. Yetgin, K. T. K. Cheung, M. el-Hajjar, and L. Hanzo, "Network lifetime maximization of wireless sensor networks," *IEEE Access*, vol. 3, pp. 2191–2226, 2015.

[3] W. Xu, W. Liang, X. Lin, and G. Mao, "Efficient scheduling of multiple mobile chargers for wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 7670–7683, 2016.

[4] M. Khaja Mohiddin and V. B. S. Srilatha Indira Dutt, "An optimum energy consumption hybrid algorithm for XLN strategic design in WSNs," *International Journal of Computer Networks and Communications*, vol. 11, no. 4, pp. 61–80, 2019.

[5] A. C. Djedouboum, A. Abba Ari, A. M. Gueroui, A. Mohamadou, and Z. Aliouat, "Big data collection in large-scale wireless sensor networks," *Sensors*, vol. 18, no. 12, p. 4474, 2018.

[6] P. Dixit, R. Kohli, and A. Acevedo-Duque, "Comparing and analyzing applications of intelligent techniques in cyberattack detection," *Security and Communication Networks*, vol. 2021, Article ID 5561816, 23 pages, 2021.

[7] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[8] A. Naeem, A. R. Javed, M. Rizwan, S. Abbas, J. C.-W. Lin, and T. R. Gadekallu, "DARE-SEP: a hybrid approach of distance aware residual energy-efficient SEP for WSN," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 2, pp. 611–621, 2021.

[9] A. Poniszewska-Maranda, R. Matusiak, N. Kryvinska, and A.-U.-H. Yasar, "A real-time service system in the cloud," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 3, pp. 961–977, 2020.

[10] A. Lavric, V. Popa, and S. Sfichi, "Street lighting control system based on large-scale WSN: a step towards a smart city," in *2014 International Conference and Exposition on Electrical and Power Engineering (EPE)*, pp. 673–676, Iasi, Romania, 2014.

[11] M. Alazab, K. Lakshmanna, G. Thippa Reddy, Q.-V. Pham, and P. K. R. Maddikunta, "Multi-objective cluster head selection using fitness averaged rider optimization algorithm for IoT networks in smart cities," *Sustainable Energy Technologies and Assessments*, vol. 43, article 100973, 2021.

[12] A. Naureen, N. Zhang, S. Furber, and Q. Shi, "A GPS-less localization and mobility modelling (LMM) system for wildlife tracking," *IEEE Access*, vol. 8, pp. 102709–102732, 2020.

[13] A. Verma, Y. Kumar, and R. Kohli, "Study of AI techniques in quality educations: challenges and recent progress," *SN Computer Science*, vol. 2, no. 4, pp. 1–7, 2021.

[14] M. K. Mohiddin and V. B. S. S. I. Dutt, "Mobility error prediction based LAB scheduling algorithm for optimizing system throughput in wireless sensor networks," *International Journal on Emerging Technologies*, vol. 11, no. 2, pp. 1087–1092, 2020.

[15] M. T. R. Khan, S. H. Ahmed, and D. Kim, "AUV-aided energy-efficient clustering in the Internet of underwater things," *IEEE Transactions on Green Communication Networking*, vol. 3, no. 4, pp. 1132–1141, 2019.

[16] A. H. Dehwah, S. Elmetennani, and C. Claudel, "UD-WCMA: an energy estimation and forecast scheme for solar powered wireless sensor networks," *Journal of Network and Computer Applications*, vol. 90, pp. 17–25, 2017.

[17] Z. Fan, Z. Jie, and Q. Yujie, "A survey on wireless power transfer-based charging scheduling schemes in wireless rechargeable sensor networks," in *2018 IEEE 4th International Conference on Control Science and Systems Engineering (ICCSSE)*, pp. 194–198, Wuhan, China, 2018.

[18] C. Wang, J. Li, Y. Yang, and F. Ye, "Combining solar energy harvesting with wireless charging for hybrid wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 3, pp. 560–576, 2018.

[19] A. Poniszewska-Maranda, D. Kaczmarek, N. Kryvinska, and F. Xhafa, "Studying usability of AI in the IoT systems/paradigm through embedding NN techniques into mobile smart service system," *Computing*, vol. 101, no. 11, pp. 1661–1685, 2019.

[20] R. Kohli and S. Gupta, "A nascent approach for noise reduction via EMD thresholding," in *Ambient Communications and Computer Systems*, vol. 904 of Advances in Intelligent Systems and Computing, , pp. 55–65, Springer, 2019.

[21] A. Kurs, A. Karalis, R. Moffatt, J. D. Joannopoulos, P. Fisher, and M. Soljacic, "Wireless power transfer via strongly coupled magnetic resonances," *Science*, vol. 317, no. 5834, pp. 83–86, 2007.

[22] L. He, P. Cheng, Y. Gu, J. Pan, T. Zhu, and C. Liu, "Mobile-to-mobile energy replenishment in mission-critical robotic sensor networks," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, pp. 1195–1203, Toronto, ON, Canada, 2014.

[23] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan, "The pothole patrol: using a mobile sensor network for road surface monitoring," in *Proceeding of the 6th international conference on Mobile systems, applications, and services - MobiSys '08*, pp. 29–39, Breckenridge, USA, 2008.

[24] C.-F. Cheng and C.-C. Wang, "The energy replenishment problem in mobile WRSNs," in *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pp. 143-144, Chengdu, China, 2018.

[25] T. Liu, B. Wu, W. Xu, X. Cao, J. Peng, and H. Wu, "Learning an effective charging scheme for mobile devices," in *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 202–211, New Orleans, LA, USA, 2020.

[26] V. Balasubramanian, F. Zaman, M. Aloqaily, I. Al Ridhawi, Y. Jararweh, and H. B. Salameh, "A mobility management architecture for seamless delivery of 5G-IoT services," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pp. 1–7, Shanghai, China, 2019.

[27] W. Yao, M. Li, and M. Y. Wu, "Inductive charging with multiple charger nodes in wireless sensor networks," in *Advanced Web and Network Technologies, and Applications. APWeb 2006,*, pp. 262–270, Springer, 2006.

[28] L. Xie, Y. Shi, Y. T. Hou, W. Lou, and H. D. Sherali, "On traveling path and related problems for a mobile station in a rechargeable sensor network," in *Proceedings of the fourteenth ACM international symposium on Mobile ad hoc networking and computing - MobiHoc '13*, pp. 109–118, Bangalore, India, 2013.

[29] W. Liang, W. Xu, X. Ren, X. Jia, and X. Lin, "Maintaining large-scale rechargeable sensor networks perpetually via multiple mobile charging vehicles," *ACM Transactions on Sensor Networks*, vol. 12, no. 2, pp. 1–26, 2016.

[30] T. Liu, B. Wu, H. Wu, and J. Peng, "Low-cost collaborative mobile charging for large-scale wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 8, pp. 2213–2227, 2017.

[31] Y. Ma, W. Liang, and W. Xu, "Charging utility maximization in wireless rechargeable sensor networks by charging multiple sensors simultaneously," *IEEE/ACM Transactions on Networking*, vol. 26, no. 4, pp. 1591–1604, 2018.

[32] P. Guo, X. Liu, M. Chen, and K. Zhang, "Should interference be avoided? Charging WSNs with efficient placement of wireless chargers," *IEEE Access*, vol. 6, pp. 54876–54883, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9820259, 1 page
https://doi.org/10.1155/2023/9820259

*Retraction*

# Retracted: The Importance of Traditional Sports into College Physical Education Based on Big Data Dynamic Programming Algorithm

**Wireless Communications and Mobile Computing**

Received 3 October 2023; Accepted 3 October 2023; Published 4 October 2023



This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

In addition, our investigation has also shown that one or more of the following human-subject reporting requirements has not been met in this article: ethical approval by an Institutional Review Board (IRB) committee or equivalent, patient/participant consent to participate, and/or agreement to publish patient/participant details (where relevant).

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

## References

[1] Z. Zheng, "The Importance of Traditional Sports into College Physical Education Based on Big Data Dynamic Programming Algorithm," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 2996940, 13 pages, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 2996940, 13 pages
https://doi.org/10.1155/2021/2996940





*Research Article*

# The Importance of Traditional Sports into College Physical Education Based on Big Data Dynamic Programming Algorithm

**Zhibin Zheng**

*Jilin University, Changchun, Jilin 130012, China*

Correspondence should be addressed to Zhibin Zheng; zhengzb@jlu.edu.cn

Received 13 August 2021; Accepted 20 September 2021; Published 29 October 2021

Academic Editor: Thippa Reddy G



Physical education teaching is conducive to the cultivation of students' lifelong sports consciousness, which can improve students' health and enhance their physique. In order to explore the importance of traditional sports based on big data dynamic programming algorithm into college physical education, the video action recognition and segmentation technology based on big data dynamic programming algorithm is designed. The complex actions in traditional sports teaching video are divided into a series of atomic actions with single semantics. The human action results are modeled according to the relationship between complex actions and atomic actions, and the actions are completed, and the changes of students' sports level were compared under different teaching modes. Compared with the no segment method, the average accuracy of the experimental design method increased by 2.80% and 3.50%, respectively, and the action recognition rate increased by 11.50%, 8.40%, 13.60%, 13.50%, and 13.60%, respectively. Before and after the experiment, there was a significant difference in the performance of the experimental group ($P = 0.021 < 0.05$). The results show that the traditional sports teaching mode based on video action recognition technology of big data dynamic programming algorithm can effectively improve the teaching quality of sports teaching. This research has a certain reference value to promote the current physical education teaching reform policy.

## 1. Introduction

Physical education can effectively enhance students' physique, cultivate their sports skills, and improve their health. Physical education is the main way for students to acquire sports technology and skills [1]. Professional sports video learning is an essential part of college physical education; students watch professional technical video repeatedly, analyze the athletes' actions in the video, and then establish the correct action representation, so as to lay a solid foundation for the cultivation of follow-up sports skills [2]. Aiming at the long PE teaching video, the big data dynamic programming algorithm is used to segment the video optimally and recognize the specific action of a single video segment [3]. For complex action video, a discriminant model with hidden variables is established to detect complex action and atomic action (single semantic meaning action decomposed from complex action), and the mapping matrix is used to reflect the many to one relationship between video segment and atomic action, so as to realize the goal of accurately identifying complex action in video segment and reduce the difficulty of students' video learning [4].

For long and complex videos, Liu et al. put forward a time boundary regression method based on time series segmentation, which uses clustering algorithm to deal with the boundary regions of high-probability behaviors in time domain and combines the maximum inhibition method to formulate the segmentation scheme. Each behavior is described by the characteristics of three subsegments (proposal segment, start sub segment, and end subsegment), and the recognition rate of behavior actions reaches 30.1% [5]. Moving object recognition in video survey is a potential leap forward development opportunity among different PC vision system applications. Thangaraj and Monikavasagom have designed a robust video object detection and tracking technology, which is composed of detection stage, tracking stage, and evaluation stage. In the evaluation stage, video segment feature extraction and classification are realized, and texture-based features are obtained from the processed frames [6]. Semantic segmentation is a research hotspot in

the field of computer vision. Lyu et al. plan to capture the urban scene from the angle of inclined UAV and propose a new high-resolution UAV semantic segmentation data set UAVid data set, which is composed of 30 video sequences. At the same time, the feature extraction of corresponding images is realized by multiscale extended network [7]. Image segmentation is the main part of target recognition in image analysis. Its purpose is to identify the notification regions of images belonging to different targets. Combined with the region merging algorithm, Lian et al. proposed a noise robust edge detection technology based on anisotropic Gaussian kernel and obtained high-quality edge detection results. The experimental results show that it has good noise robustness and positioning accuracy [8]. Huang et al. expressed the object as shape and appearance and used it as the constraint of segmentation, so as to ensure that the object segmentation mask is consistent with the object area and the knowledge in the image [9].

Online segmentation and skeleton-based gesture recognition are very difficult, especially for incomplete gestures, whose early recognition easily falls into local optimum. Chen et al. use a temporal hierarchical dictionary to guide the decoding process of hidden Markov model and propose a measure called "relative entropy mapping," which guides HMM decoding according to time context [10]. Gao et al. proposed a fiber recognition framework based on image segmentation, deep convolution neural network, and vision to segment overlapping and adhesive translucent fibers. The results show that the accuracy of multi fiber recognition strategy can reach 99.5% [11]. Saifuddin Saif et al. use convolutional neural network combined with compression function to extract spatial and temporal information features in image segmentation, which significantly improves the performance of human behavior recognition [12]. Reddy et al. studied two important dimensionality reduction techniques, linear discriminant analysis (LDA) and principal component analysis (PCA), on four popular machine learning (ML) algorithms [13]. Patil and Sunitha believe that in the dynamic video survey system, the location and detection of moving items in the video scene are very important, which are affected by obstacles, shadows, and noise. The basic development direction of the multicamera video survey system is the horizontal coordination and rerecognition of moving object detection and tracking on multiple cameras [14]. Ma and Song proposed a moving object detection method in H.264/AVC compressed domain for video surveillance applications, which uses H.264 to compress the information in the bit stream, while reducing the computational complexity and memory requirements, and completes the detection and segmentation of moving objects through motion vectors and quantization parameters [15].

It can be seen from the above research results that there are a lot of researches on dynamic video segmentation, human behavior recognition in video, video monitoring, and other different directions in the current society, but the research on big data dynamic programming algorithm for segmentation, recognition, and annotation of sports teaching video or professional athletes' skill display video is limited. At the same time, most scholars ignore the role of video action recognition technology based on big data dynamic programming algorithm in traditional sports teaching to a certain extent and lack of related research on the connection between video action recognition and college physical education. Therefore, the importance of video action recognition technology based on big data dynamic programming algorithm in traditional sports teaching in colleges and universities will be discussed. This paper is performed to discuss and analyze.

This paper is divided into four parts. The first part expounds the importance of traditional sports based on big data dynamic programming algorithm in college sports teaching, and the second part expounds the sports segmentation and recognition in sports video. The third part expounds the analysis content of the application effect of sports teaching video. The last part summarizes that the traditional physical education teaching mode of video action recognition technology based on big data dynamic programming algorithm can effectively improve the teaching quality of physical education. This study has a certain reference value for promoting the current physical education reform policy.

## 2. Motion Segmentation and Recognition in Sports Video

*2.1. Action Segmentation and Recognition of Long Video.* The complex actions in traditional physical education teaching video are divided into a series of atomic actions with single semantics. In this section, a structured discriminant model with hidden variables is designed, by which the continuous video stream is divided into a series of video segments with only a single action, and the action type of each video segment is marked.

The different states in the implicit state sequence in Figure 1 can reflect the potential semantic concepts in the corresponding unit video segment. In the model, an action is represented by the interaction among the learning video segment features, the contained latent semantic concepts, and action categories, and the corresponding temporal context between learning video segments is mined from the latent semantic level [16]. For the test video containing multiple actions, the big data dynamic programming algorithm is used to find the optimal video segment mode, and the video segment action category is judged [17].

The long video $X$ is divided into a series of unit video segments, and the spatiotemporal feature $x_t$ of each unit video segment is extracted. At this time, the long video $X = \{x_1, x_2, \cdots, x_T\}$, the number of unit video segments $T$, and the video length are proportional [18]. If $S = \{s_1, s_2, \cdots, s_M\}$ is a long video, $X$ is divided into $M$ video segments, and $Y = \{y_1, y_2, \cdots, y_M\}$ belongs to the action tag of video segment, and $s_M = T$ exists, then the $m$th video segment is $X_M = X_{(s_{m-1}, s_m)} = \{x_{s_{m-1}+1}, \cdots, x_{s_m}\}$, and its corresponding action tag is $y_m$. The state variable $H = \{h_1, h_2, \cdots, h_T\}$ is introduced into the model, and the semantic meaning of unit video segment $x_t$ is represented by $h_t$ [19]. In this paper, we expect to learn a discriminant function $f_w(X)$, which can

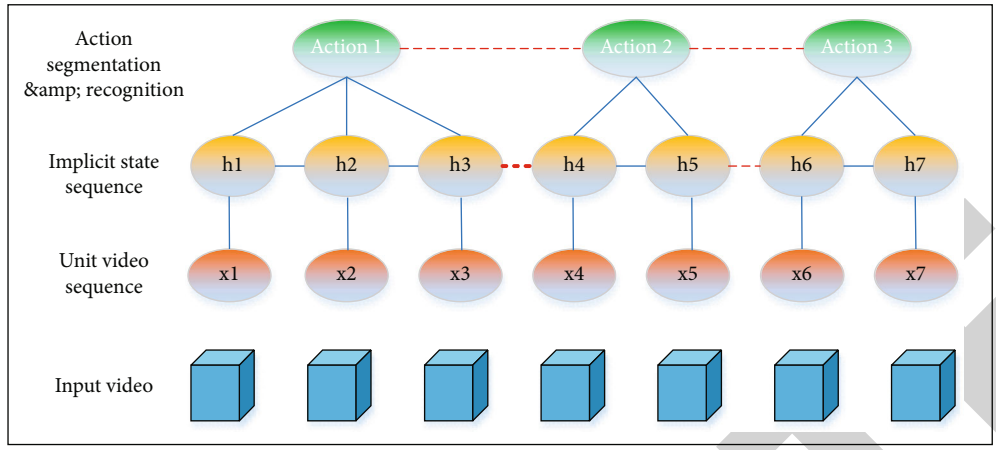



FIGURE 1: Schematic diagram of action segmentation and recognition method.

segment the video $X = \{x_1, x_2, \cdots, x_T\}$ in $S = \{s_1, s_2, \cdots, s_M\}$ mode and label the action tag $Y = \{y_1, y_2, \cdots, y_M\}$:

$$f_w(X) = \max_{Y,S,H} F(X, Y, S, H) = \max_{Y,S,H} w^T \cdot \Phi(X, Y, S, H). \quad (1)$$

In equation (1), $w$ is the model parameter, and the feature vector of the interaction among $X$, CC, DD, and EE is described by FF. GG is defined as the sum of several potential energies.

In equation (1),$w$is the model parameter; video$X$, video segment division$S$, video segment corresponding label$Y$, implied variable$H$, and the eigenvector of the interaction among the four variables are described by$\Phi(X, Y, S, H)$. $w^T \cdot \Phi(X, Y, S, H)$ is defined as the sum of several potential energies.

$$\begin{aligned} w^T \cdot \Phi(X, Y, S, H) = & \sum_{m=1:M} \alpha^T \cdot \phi_1(X_m, y_m) \\ & + \sum_{m=1:M} \beta^T \cdot \phi_2(X_m, H_m) \\ & + \sum_{m=1:M} \gamma^T \cdot \phi_3(y_m, H_m) \\ & + \sum_{m=1:M} \eta^T \cdot \phi_4(H_{m-1}, H_m) \\ & + \sum_{m=1:M} \delta^T \cdot \phi_5(y_{m-1}, y_m). \end{aligned} \quad (2)$$

In equation (2), $w = \{\alpha; \beta; \gamma; \eta; \delta\}$ is the model parameter. The potential energy functions shown in equations (3), (4), and (5) are only related to the video segment containing a single action.

$$\alpha^T \cdot \phi_1(X_m, y_m) = \sum_{y^{'}} \alpha_{y^{'}}^T \cdot X_m \cdot I\left(y_m = y^{'}\right). \quad (3)$$

Equation (3) evaluates the matching degree between the global feature $X_m$ of the whole video segment and the action category template of the video segment, where $\alpha_y$ is the classification template of action $y$ and $I(y_m = y^{'})$ is the indicator function; when $y_m = y^{'}$, then $I(y_m = y^{'}) = 1$, otherwise $I(y_m = y^{'}) = 0$.

$$\beta^T \cdot \phi_2(X_m, H_m) = \sum_{t=s_{m-1}+1:s_m} \sum_{h^{'}} \beta_{h^{'}}^T \cdot x_t \cdot I\left(h_t = h^{'}\right), \quad (4)$$

$$\begin{aligned} \gamma^T \cdot \phi_3(y_m, H_m) = & \sum_{y^{'}} \sum_{t=s_{m-1}+1:s_m} \sum_{h^{'},h^{''}} \gamma_{y^{'}h^{'}h^{''}}^T \cdot I\left(y_m = y^{'}\right) \\ & \cdot I\left(h_t = h^{'}\right) \cdot I\left(h_{t+1} = h^{''}\right). \end{aligned} \quad (5)$$

Formula (4) reflects the constraint relationship between local features and latent semantic state of video segment. Formula (5) is the modeling of cooccurrence relationship between the overall action category and the corresponding implied state of a video segment, reflecting the semantic constraints between two unit video segments in the sequential adjacent state of a certain action [20].

$$\eta^T \cdot \phi_4(H_{m-1}, H_m) = \sum_{h^{'}} \sum_{h^{'}} \eta_{h^{'}h^{''}}^T \cdot I\left(h_{s_{m-1}} = h^{'}\right) \cdot I\left(h_{s_{m-1}+1} = h^{''}\right), \quad (6)$$

$$\delta^T \cdot \phi_5(y_{m-1}, y_m) = \sum_{y^{'}} \sum_{y^{''}} \delta_{y^{'}y^{''}}^T \cdot I\left(y_{m-1} = y^{'}\right) \cdot I\left(y_m = y^{''}\right). \quad (7)$$

Formula (6) is the potential energy function based on the relationship between adjacent video segments $X_m$ and $X_{m+1}$ in the semantic concept level, and formula (7) is the potential energy function based on the relationship between adjacent video segments $X_m$ and $X_{m+1}$ in the action category level. Before segmenting the whole long video $X$, first, understand the best segmentation method of the first $u$ unit video segments $\{x_1, x_2, \cdots, x_u\}$ of $X$ [21]. If the action label of the $U$ th unit video segment $\{x_1, x_2, \cdots, x_u\}$ is $y$ and the corresponding hidden state is $h$, the best segmentation method can be

described by the function $g(X, h, y, u)$, and the maximum potential energy function value can be taken at this time [22].

$$\begin{aligned} g(X, h, y, u) &= \max_{h_{u\_}, y\_, u\_, H_{u\_+1:u-1}} \{g(X, h_{u\_}, y\_, u\_) \\ &\quad + \theta(X_{u\_+1:u}, H_{u\_+1:u}, y, y\_, h_{u\_})\} \\ &= \max_{h_{u\_}, y\_, u\_} \{g(X, h_{u\_}, y\_, u\_) \\ &\quad + \max_{H_{u\_+1:u-1}} \theta(X_{u\_+1:u}, H_{u\_+1:u}, y, y\_, h_{u\_})\}, \end{aligned} \tag{8}$$

$$\begin{aligned} &\theta(X_{u\_+1:u}, H_{u\_+1:u}, y, y\_, h_{u\_}) \\ &= \alpha^T \cdot \phi_1(X_{u\_+1:u}, y) + \beta^T \cdot \phi_2(X_{u\_+1:u}, H_{u\_+1:u}) \\ &\quad + \gamma^T \cdot \phi_3(y, H_{u\_+1:u}) + \eta^T \cdot \phi_4(h_{u\_}, h_{u\_+1}) \\ &\quad + \delta^T \cdot \phi_5(y\_, y). \end{aligned} \tag{9}$$

Equation (8) gives the incremental form of the function $g(X, h, y, u)$ and defines the relationship between it and the function value $g(X, h_{u\_}, y\_, u\_)$ of the previous video segment. Suppose that the last unit video in the previous video segment is $x_{u\_}$, which meets the condition $l_{\min} \le u - u\_ \le l_{\max}$. $h_{u\_}$ is used to reflect the hidden state of $x_{u\_}$, and $y\_$ is used to represent the action tag of $x_{u\_}$, where $l_{\min}$ and $l_{\max}$ are the minimum length and the maximum length of an action in turn [23]. The model is trained by a set of videos $\{(X^i, S^i, Y^i)\}_{i=1:N}$ marked with action segments and their categories, including long video $X^i$, real video segmentation $S^i$, and action label $Y^i$ of each video segment. The training framework based on maximum interval is used to learn the model parameter $w$:

$$\begin{aligned} &\min_{w} \quad 0.5\|w\|^2 + C \cdot \sum_i \xi_i \\ &\text{s.t.} \quad \max_H w^T \cdot \Phi\left(X^i, S^i, Y^i, H\right) - \max_H w^T \cdot \Phi\left(X^i, Y, S, H\right) \\ &\quad \ge \Delta\left[\left(Y^i, S^i\right), (Y, S)\right] - \xi_i, \forall i, Y, S. \end{aligned} \tag{10}$$

In equation (10), the standard maximum interval constraint condition is used to constrain, and the model parameter $W$ should reasonably divide the video and accurately label the action category of each video segment. $\xi_i$ is the penalty factor, which acts on action segmentation and action recognition, and $C$ is the penalty factor coefficient.

$$\Delta\left[\left(Y^i, S^i\right), (Y, S)\right] = \frac{1}{T} \cdot \sum_{t=1:T} I\left(\chi_t \ne \chi_t^i\right). \tag{11}$$

Equation (11) is the definition of loss function, where $\chi_t^i$ is the action tag of unit video segment $x_t$, determined by $(Y^i, S^i)$, and $\chi_t$ is the action tag of unit video segment $x_t$, determined by $(Y, S)$. In this case, the definition of function $g(X, h, y, u)$ is shown in the following equation:

$$\begin{aligned} g(X, h, y, u) &= \max_{\substack{h_{u\_}, y\_, u\_, \\ H_{u\_+1:u-1}}} \left\{ g(X, h_{u\_}, y\_, u\_) + \frac{1}{T} \sum_{t=u\_+1:u} \right. \\ &\quad \left. \cdot I\left(y \ne \chi_t^i\right) + \theta(X_{u\_+1:u}, H_{u\_+1:u}, y, y\_, h_{u\_}) \right\} \\ &= \max_{h_{u\_}, y\_, u\_} \left\{ g(X, h_{u\_}, y\_, u\_) + + \frac{1}{T} \sum_{t=u\_+1:u} I\left(y \ne \chi_t^i\right) \right. \\ &\quad \left. + \max_{H_{u\_+1:u-1}} \theta(X_{u\_+1:u}, H_{u\_+1:u}, y, y\_, h_{u\_}) \right\}. \end{aligned} \tag{12}$$

*2.2. Complex Action Analysis Based on Semantic Decomposition.* Sports teaching video contains a series of complex movements, such as "three-step layup" and "triple jump." Complex actions consist of a series of simple actions with a single semantic, which are called "atomic actions" in this section.

Figure 2 shows the expression of an action video through complex actions, atomic actions contained in complex actions, and video segments of atomic actions. A discriminant model with hidden variables is used to show the relationship among high-level action categories, middle-level atomic actions, and low-level video segments. While detecting atomic actions in video, the temporal structure of atomic actions is discussed [24]. Then, the mapping matrix is introduced to associate the video segment with the atomic action, and the many to one correspondence between them is established; that is, multiple video segments show the same atomic action.

Let the training sample set be $\{(x^n, y^n, h^n) | n = 1, 2, \cdots N\}$, $x^n$ and $y^n$ to represent the $n$th video and the complex action category that the video belongs to. The atomic action of the video is marked as $h^n = [h_1^n, h_2^n, \cdots, h_v^n]$. When the $i$th atomic action appears in the video, there is $h_i^n = 1$ and vice versa $h_i^n = 0$. The mapping matrix between video segment and atomic action is used as the hidden variable $\{g^n | n = 1, 2, \cdots, N\}$ in the model [25].

$$f_w(X) = \max_{y,h,g} F(x, y, h, g) = \max_{y,h,g} w^T \cdot \Phi(x, y, h, g). \tag{13}$$

Equation (13) is the prediction function of expected learning, and the eigenvectors of the relationships among video $x$, complex action $y$, atomic action $h$, and mapping matrix $g$ are described by $\Phi(x, y, h, g)$. Firstly, the video segment atomic action mapping matrix is established, the atomic action $h$ of video $x$ is labeled as $h$, and the video $x$ is divided into equal size $r$ subvideo segment $x = [x_1, x_2, \cdots, x_R]$, and the mapping matrix $g = \{g_{ij}, i = 1, 2, \cdots, N; j = 1, 2, \cdots, V\}$ is introduced [26]. If the $i$th video segment is labeled as the $j$th atomic action, there is $g_{ij} = 1$, otherwise $g_{ij} = 0$. Then, the model is built according to the relationship among video segment, atomic action, and complex action category, as shown in the following equation:

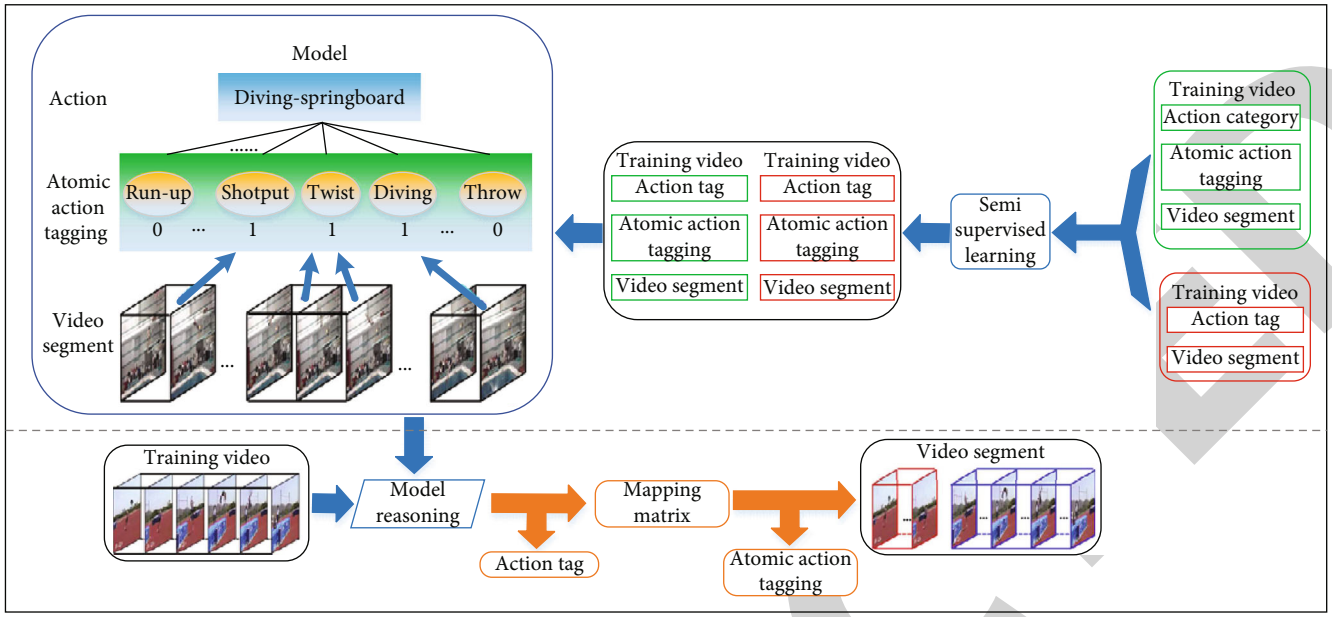



Figure 2: Complex action analysis method based on semantic decomposition.

$$w^T \cdot \Phi(x, y, h, g) = \alpha^T \phi(x, g) + \beta^T \psi(x, y) + \gamma^T \varphi(h, y). \quad (14)$$

Equation (14) is the potential energy function $w^T \cdot \Phi(x, y, h, g)$ and the model parameter $w = (\alpha; \beta; \gamma)$.

$$\alpha^T \phi(x, g) = \sum_{i=1}^{R} \sum_{j=1}^{V} \alpha_j^T \cdot x_i \cdot g_{ij}. \quad (15)$$

Equation (15) is a video segment atomic action interaction model, which reflects the matching degree between video segment and atomic action. The template of the $j$th atomic action is $\alpha_j$, the feature of the $i$th video segment is $x_i$, and there is a constraint $\max_i g_{ij} = h_j$.

$$\beta^T \psi(x, y) = \beta_y^T \cdot x. \quad (16)$$

Equation (16) shows the matching degree between the video $x$ and the action template and uses the standard linear model to predict the possibility that the video $x$ belongs to the action $y$.

$$\gamma^T \varphi(h, y) = \sum_{j=1}^{V} \left[ \gamma_{j,1}^{y} \cdot h_j + \gamma_{j,0}^{y} \cdot (1 - h_j) \right]. \quad (17)$$

Formula (17) shows the semantic relationship between atomic actions and complex actions. Different complex actions have different decomposition modes, and different videos of the same complex action have different decomposition modes. The training set $\{(x^n, y^n, h^n) | n = 1, 2, \cdots N\}$ and learning parameter $W$ are given to train the model, and the SVM framework with hidden variables is used to learn the model parameter $W$:

$$\min_{w, \xi^n} \quad 0.5\|w\|^2 + C \sum_{n=1}^{N} \xi^n$$

$$\text{s.t.} \quad \max_g w^T \cdot \Phi(x^n, y^n, h^n, g) - \max_g w^T \cdot \Phi(x^n, y, h, g) \geq \Delta[(y^n, h^n), (y, h)] - \xi^n, \forall n, \forall y, \forall h. \quad (18)$$

In equation (18), $\Delta[(y^n, h^n), (y, h)]$ predicts the loss function of tag $(y, h)$ with video $x^n$, and the definition of the loss function is $\Delta[(y^n, h^n), (y, h)] = \ell(y^n, y) + \sum_{j=1}^{V} \ell(h_j^n, h_j)$, with

$$\ell(a, b) = \begin{cases} 1, & \text{if } a \neq b, \\ 0, & \text{if } a = b. \end{cases} \quad (19)$$

In the test stage, the model is used to predict the complex action category $y*$ of video $X$ and the atomic action annotation $h*$, with $(y*, h*) = \arg \max_{y,h} [\max_g w^T \Phi(x, y, h, g)]$ [27]. The big data dynamic programming algorithm is used to complete the segmentation and recognition of each action in sports video. Students can improve their mastery of sports professional skills and enrich the content of college sports teaching by marking the unit video segment.

As shown in Figure 3, in the teacher's position, teachers can play sports-related action videos during the teaching process and explain the key points and difficulties of corresponding professional sports actions to students according to the video segment labels, so as to clarify the demonstration points of professional sports. From the perspective of students' position, students follow the teacher's explanation, master the way and focus of watching the video, mark the main points of action, and then communicate with each other and practice in groups [28].

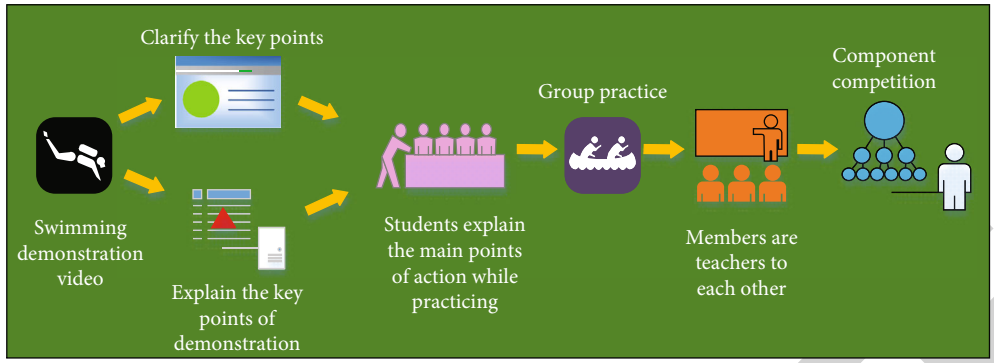



Figure 3: High-efficiency physical education teaching with sports video.

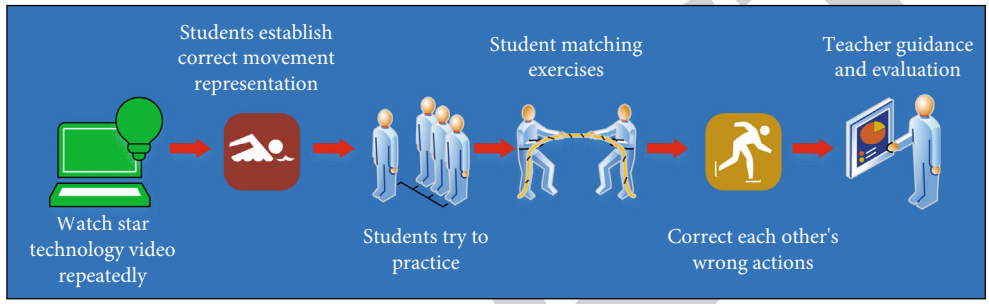



Figure 4: The cultivation of students' ability to correct mistakes.

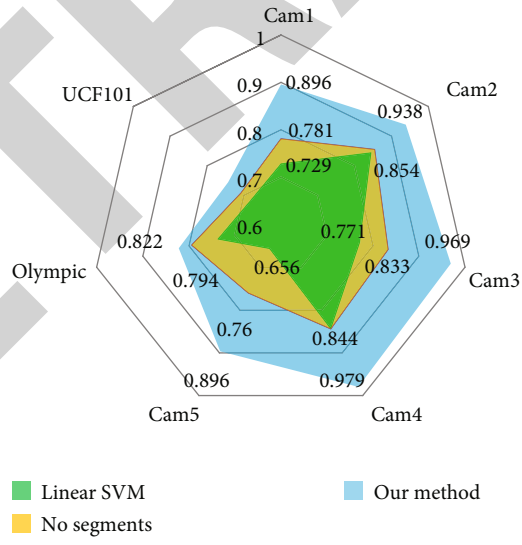
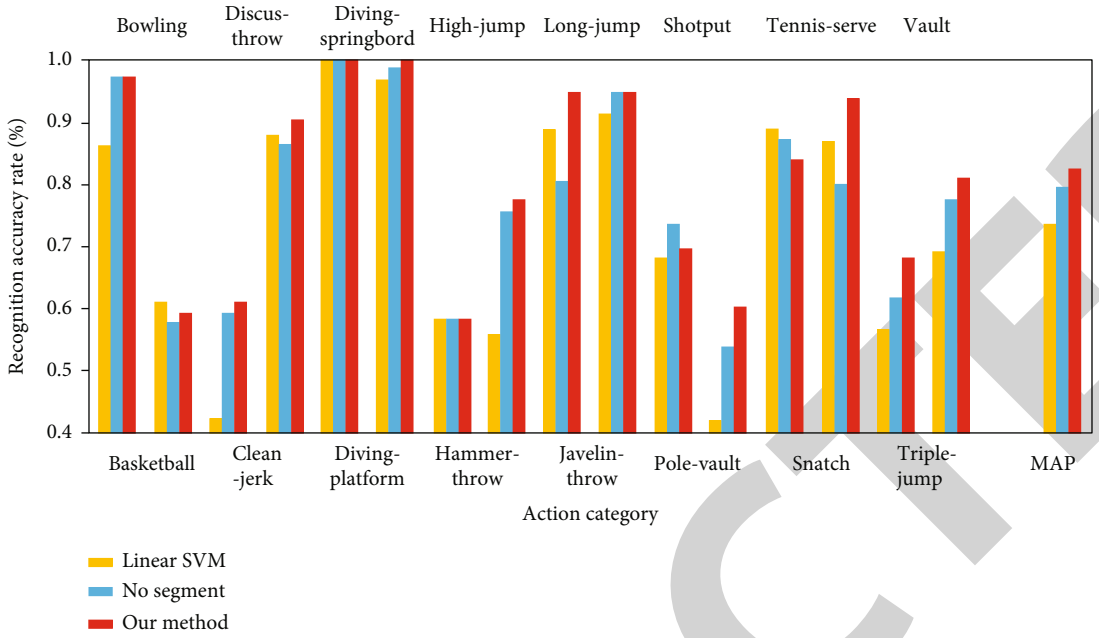



Figure 5: Comparison of action recognition results on three databases.

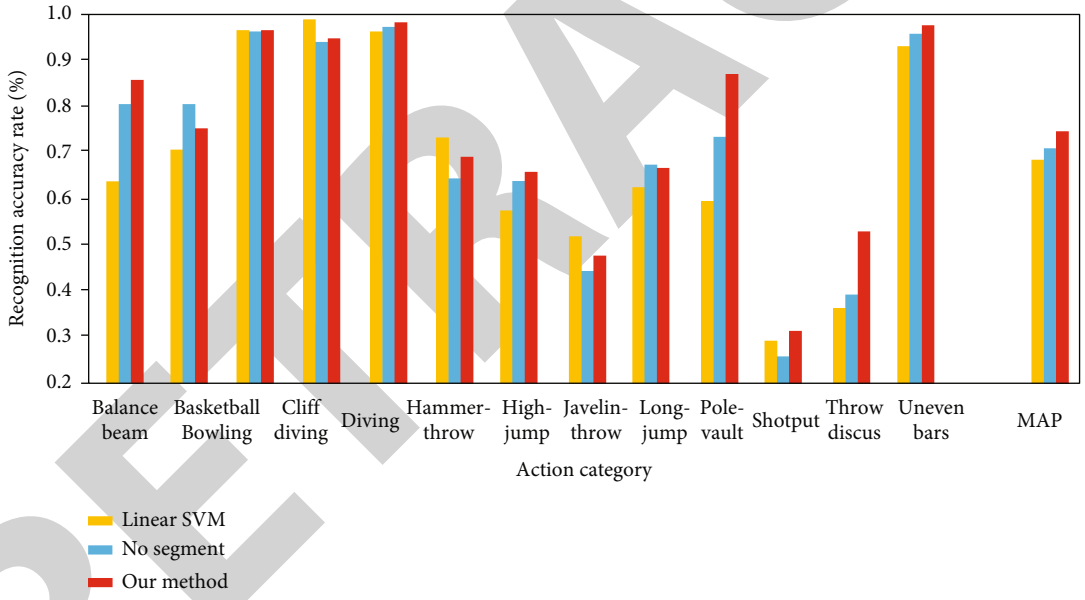
As shown in Figure 4, students can watch and learn professional sports videos repeatedly and initially establish correct action representation. The annotation of single action in video segments, combined with mutual communication and action practice between students, can gradually correct the action errors in the learning process. At the same time, according to the content of the teaching video, teachers pay for the wrong actions of students [29].

## 3. Analysis on the Application Effect of Sports Teaching Video

*3.1. Sports Video Action Recognition Effect.* Two reference algorithms are set up. The first one uses linear SVM classifier to classify video features, which is called linear SVM algorithm for short. The second is the simplification of the algorithm described in Section 2, which constructs the model of




(a) Effect comparison of several methods for action recognition on the Olympic dataset

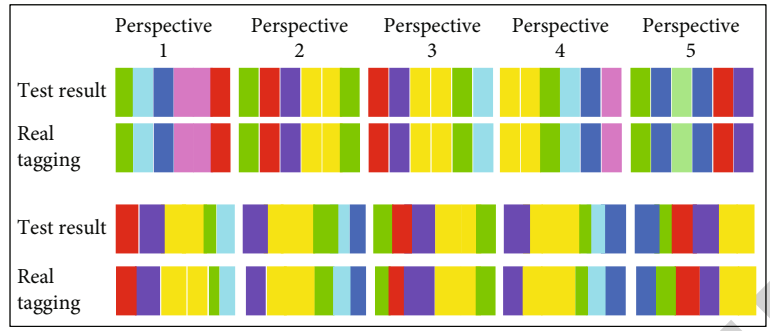



(b) Effect comparison of several methods for action recognition on the UCF101 dataset

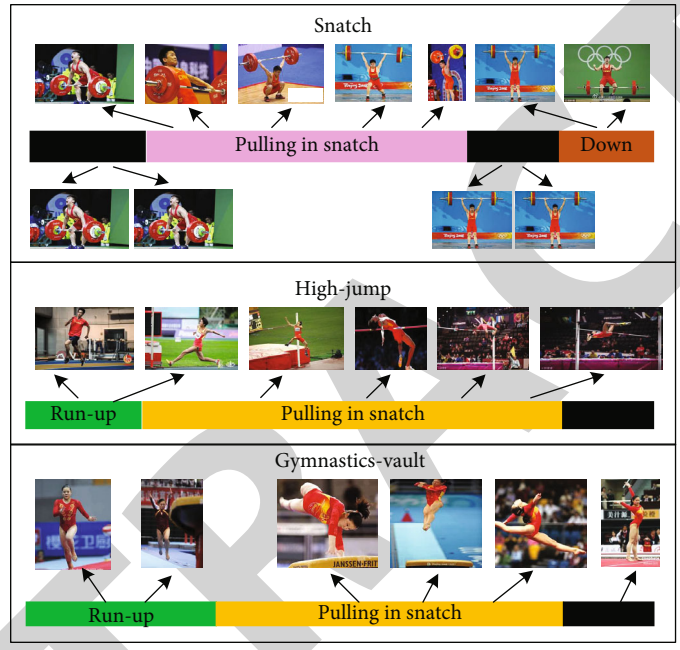
Figure 6: Comparison of recognition effects of different types of actions on two data sets.

the relationship between complex actions and atomic actions through the structured SVM framework. This method does not consider the video segmentation features and ignores the relationship between video segments and atomic actions, which is called no segments for short. The synthetic data set, Olympic data set, and ucf101 data set were selected to compare the action recognition rate (synthetic data set) and mean average precision (map) (Olympic data set and ucf101 data set) of the two reference algorithms and the design algorithm.

Figure 5 shows that the overall action recognition rate of no segment method is higher than that of linear SVM method on the composite data set, and the action recognition rates of different subdata sets are increased by 5.20% (CAM1), 1.00% (CAM2), 6.20% (CAM3), 0.00% (CAM4), and 10.40% (CAM5), respectively; the average accuracy of no segment method is higher than that of no segment method on the Olympic data set and ucf101 data set. The results show that the introduction of atomic action concept is conducive to learning more discriminative complex action classifier. Compared with the no segment method, the motion recognition rate of the experimental design method increased by 11.50% (CAM1), 8.40% (CAM2), 13.60% (CAM3), 13.50% (CAM4), and 13.60% (CAM5), and the average accuracy rate increased by 2.80% (Olympic data set) and 3.50% (ucf101 data set). To sum up, the effect of




(a) Temporal localization results of atomic actions on composite datasets




(b) Hierarchical description of complex actions on Olympic dataset

Figure 7: Effect analysis of complex action description in teaching video.

the proposed method is better than that of linear SVM and no segments, because the proposed method takes the relationship between complex actions and atomic actions into account and establishes the corresponding relationship between atomic actions and video segments to show the temporal structure of atomic actions.

Figure 6(a) shows the comparison results of the recognition effects of the linear SVM method, the no segment method, and the proposed method on the Olympic data set. It can be seen from Figure 6(a) that the recognition effect of the proposed method is better than that of the linear SVM method and no segment method for all action categories except pole vault. Figure 6(b) shows the results of the comparison of the recognition effects of 13 categories of actions on the ucf101 data set by the linear SVM method, no segment method, and the experimental method. On the whole, the experimental method has better action recognition effect.

Figure 7(a) compares the prediction results of the proposed method and the real atomic action annotation on the same video segment on several videos of the composite data set. One of the colors corresponds to an atomic action, and the duration of atomic action is represented by color width. It can be seen that in most cases, the proposed method can achieve the goal of detecting atomic action in video and accurately locate the time sequence position of atomic action. Figure 7(b) shows an example of video description on the Olympic data set. Complex action categories are marked on the top of the time bar. The time bar is responsible for displaying the detected atomic actions. One color corresponds to one atomic action. Black means that the video segment is not associated with any atomic actions. Taking "high jump" as an example, it can be seen from Figure 7(b) that it can be divided into "run-up," "somersault," "landing," and other three atomic actions, and the relationship between atomic actions and video segments is roughly correct; most black video segments are static or only contain irrelevant actions; different complex actions can share a group of atomic actions. According to the detailed description of complex movements in the video, teachers or students can accurately learn the professional sports skills, sports posture, and power way in sports.

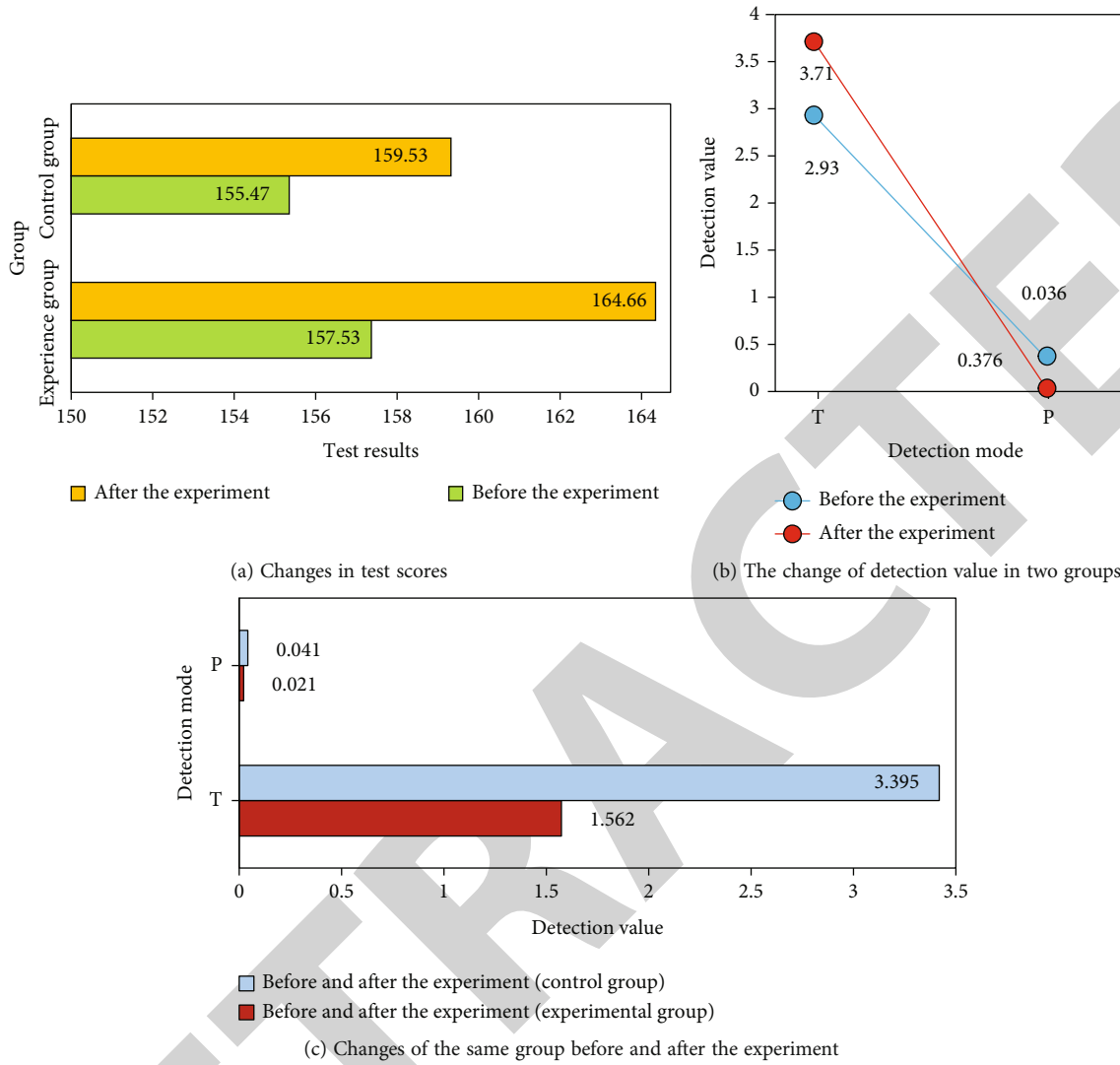



Figure 8: Before and after the experiment, the scores of the experimental group and the control group were compared.

*3.2. The Effect of Application in College Physical Education.* In order to test the application effect of sports video segmentation and recognition technology in College Physical Education Teaching under big data dynamic programming algorithm and understand the importance of traditional sports in College Physical Education Teaching under big data dynamic programming algorithm, a college physical education major student is selected as the experimental object. After a period of teaching experiment, taking Fosbury Flop as a test item, the test results of the experimental group were compared with those of the control group.

The experimental group introduced the traditional sports video teaching link of sports video action recognition technology based on big data dynamic programming algorithm, while the control group taught in the normal way (no video action recognition). It can be seen from Figure 8 that before the experiment, there was no significant difference between the two groups ($P = 0.376 > 0.05$); after the experiment, there was a significant difference between the two groups (experimental group and control group) ($P = 0.036 < 0.05$), and there was a significant difference between the two groups ($P = 0.021 < 0.05$). It shows that the introduction of traditional sports based on big data dynamic programming algorithm can improve students' sports performance to a certain extent.

Sports skills are the ability that students in sports colleges and departments must master and also the important foundation for students to engage in sports-related work in the future. Therefore, in the introduction of traditional sports based on big data dynamic programming algorithm to cultivate students, we should pay attention to the improvement of sports-related students' professional skills. In order to verify the influence of the introduction of traditional sports based on the video action recognition technology of big data dynamic programming algorithm on the students' skill level, the change of students' skill level before and after the experiment is taken.

As shown in Figure 9, before the beginning of the teaching experiment, there was no significant difference in the skill level between the experimental group and the control

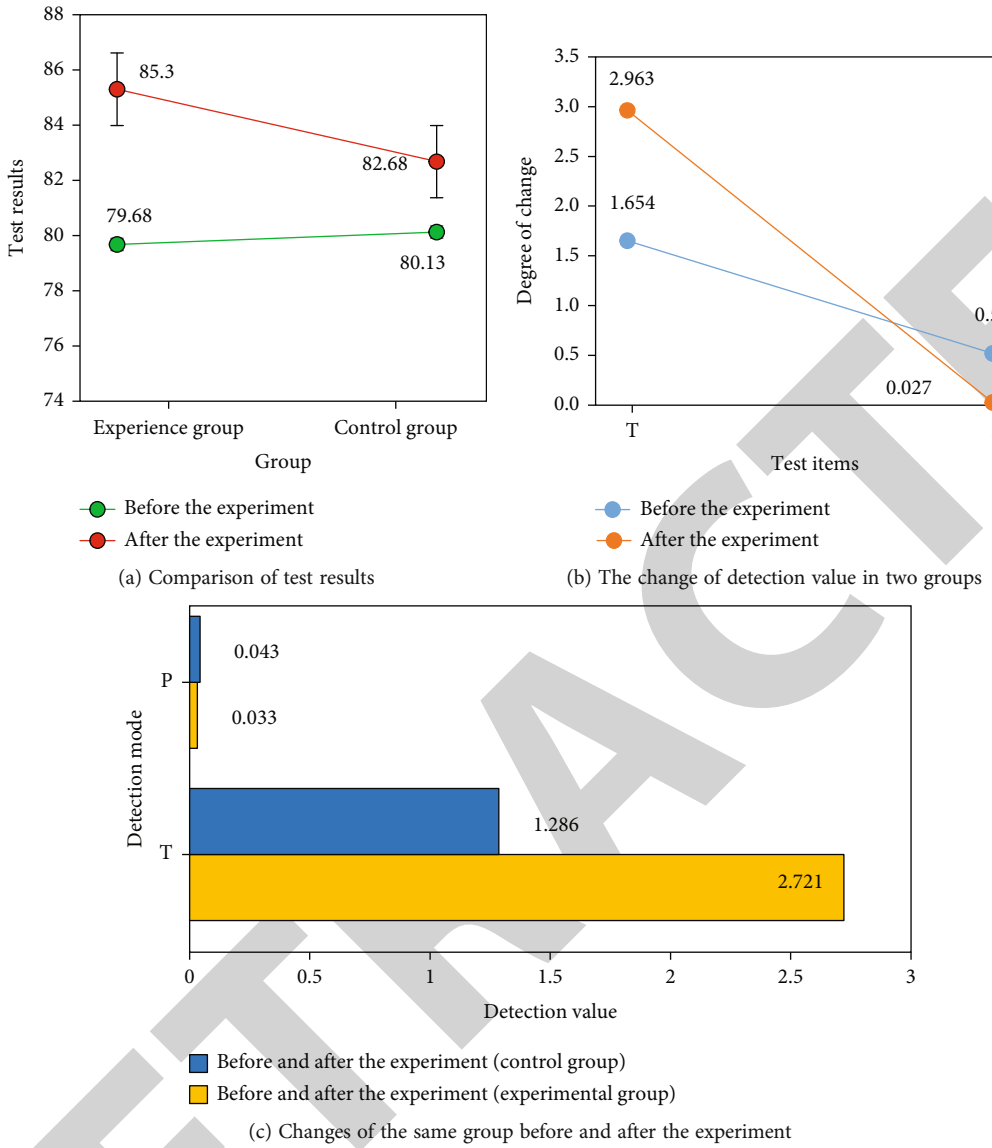



Figure 9: Comparison of technical evaluation between the experimental group and the control group before and after the experiment.

group ($P = 0.392 > 0.05$); after the end of the teaching experiment, there was a significant difference between the two groups ($P = 0.027 < 0.05$), there was a significant difference in the skill level of the students before and after the experimental group ($P = 0.033 < 0.05$), and there was a significant difference in the skill level of the students before and after the control group ($P = 043 < 0.05$), which shows that the progress of the experimental group is significantly stronger than that of the control group, indicating that the introduction of traditional sports based on big data dynamic programming algorithm video action recognition technology can effectively improve the students' skill level.

Figure 10 shows the comparison of teaching ability between the experimental group and the control group after the experiment. It can be seen from Figure 10 that after the experiment, under the guidance of the teaching mode of video action recognition technology based on big data dynamic programming algorithm, the five teaching abilities of the experimental group have been improved, and the improvement of teaching ability is significantly higher than that of the control group ($P < 0.05$). The improvement of students' teaching ability is conducive to learning. The degree of mastering the students' professional ability is improved, and the employment probability is increased.

## 4. Conclusion

Physical education is an important part of school education. In order to verify the importance of traditional sports based on big data dynamic programming algorithm into college physical education, the video action recognition technology under big data dynamic programming algorithm is designed to recognize the action of traditional sports teaching video, guide students to watch and learn, and compare the changes

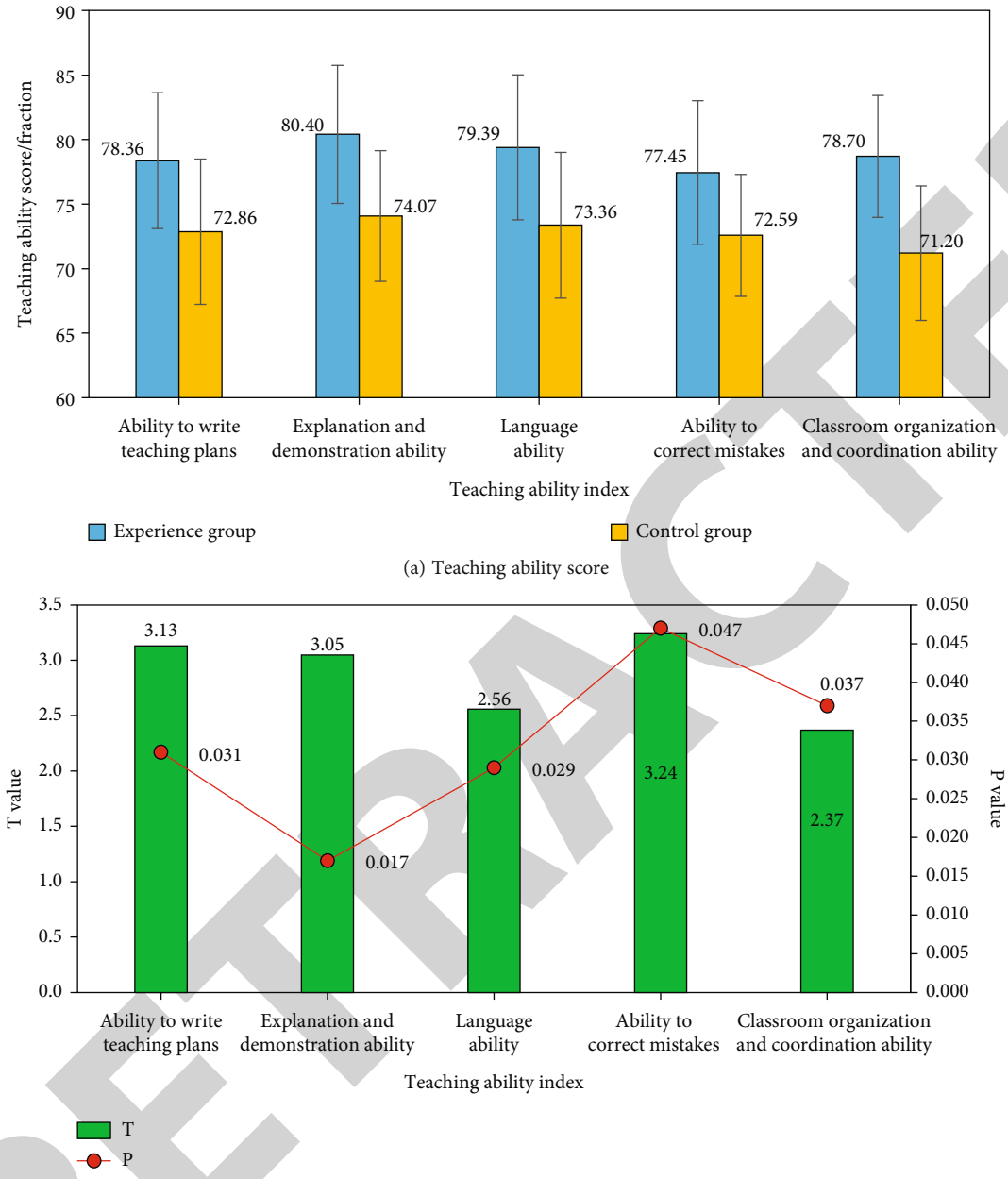



Figure 10: Comparison of teaching ability between the experimental group and control group after experiment.

of students' sports level. Compared with the no segment method, the motion recognition rate of the experimental design method was increased by 11.50% (CAM1), 8.40% (CAM2), 13.60% (CAM3), 13.50% (CAM4), and 13.60% (CAM5), and the average accuracy rate was increased by 2.80% (Olympic data set) and 3.50% (ucf101 data set), respectively, with better effect of motion recognition and accurate video description. There was a significant difference in the achievement of motor standard ($P = 0.021 < 0.05$), and there was a significant difference in the skill level of the experimental group ($P = 0.033 < 0.05$). To sum up, the introduction of the traditional sports teaching mode of video action recognition technology based on big data dynamic programming algorithm can effectively improve the teaching quality of physical education in colleges and universities and improve the level of students' professional sports skills. The experiment has achieved some results, but in the experiment, the initial segmentation of the video in the way of equal division easily causes the end of the previous atomic action combined with the beginning of the latter atomic action to be divided into the same video segment, resulting in wrong video segment annotation results. Therefore, a future research work is to improve the effectiveness of the initial video segmentation through the video segmentation method based on motion information.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no competing interest.

## Acknowledgments

The project is supported by the "Research on the Construction of College Students' Sports and Health Evaluation System, China (Grant No. 371202901405)."

## References

[1] A. Neuprez, J. F. Kaux, M. Locquet, C. Beaudart, and J. Y. Reginster, "The presence of erosive joints is a strong predictor of radiological progression in hand osteoarthritis: results of a 2-year prospective follow-up of the Liège Hand Osteoarthritis Cohort (LIHOC)," *Arthritis Research & Therapy*, vol. 23, no. 1, pp. 1–9, 2021.

[2] M. A. Lopez-Gordo, N. Kohlmorgen, C. Morillas, and F. Pelayo, "Performance prediction at single-action level to a first-person shooter video game," *Virtual Reality*, vol. 25, no. 3, pp. 681–693, 2021.

[3] F. Gurkan and B. Gunsel, "Integration of regularized _l_ 1 tracking and instance segmentation for video object tracking," *Neurocomputing*, vol. 423, pp. 284–300, 2021.

[4] U. Cosimo, C. Matteo, and A. Alessio, "Neuroanatomical substrates of action perception and understanding: an anatomic likelihood estimation meta-analysis of lesion-symptom mapping studies in brain injured patients," *Frontiers in Human Neuroscience*, vol. 8, p. 344, 2016.

[5] J. Liu, C. Wang, and Y. Liu, "A novel method for temporal action localization and recognition in untrimmed video based on time series segmentation," *IEEE Access*, vol. 7, pp. 135204–135209, 2019.

[6] M. Thangaraj and S. Monikavasagom, "A competent frame work for efficient object detection, tracking and classification," *Wireless Personal Communications*, vol. 107, no. 2, pp. 939–957, 2019.

[7] Y. Lyu, G. Vosselman, G. S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: a semantic segmentation dataset for UAV imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 165, pp. 108–119, 2020.

[8] J. Lian, K. Jia, Y. Li, Q. Zhang, Z. Zhang, and B. Zhang, "Area segmentation of images using watershed and anisotropic Gaussian kernels," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 04, p. 1954012, 2019.

[9] W. Huang, H. Yu, W. Zheng, and J. Zhang, "A framework to coordinate segmentation and recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8473–8480, 2019.

[10] H. Chen, X. Liu, J. Shi, and G. Zhao, "Temporal hierarchical dictionary guided decoding for online gesture segmentation and recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 9689–9702, 2020.

[11] F. Gao, J. Lin, H. Liu, and S. Lu, "A novel VBM framework of fiber recognition based on image segmentation and DCNN," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 963–973, 2020.

[12] A. F. M. Saifuddin Saif, M. A. Shahriar Khan, A. Mohammad Hadi, R. Prashad Karmoker, and J. Julian Gomes, "Aggressive action estimation: a comprehensive review on neural network based human segmentation and action recognition," *International Journal of Education and Management Engineering*, vol. 9, no. 1, pp. 9–19, 2019.

[13] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[14] C. M. Patil and Y. N. Sunitha, "A review of the multiple object detection, tracking and recognition in video surveillance systems using different approaches," *Restaurant Business*, vol. 118, no. 7, pp. 34–43, 2019.

[15] M. Ma and H. Song, "Effective moving object detection in H.264/AVC compressed domain for video surveillance," *Multimedia Tools and Applications*, vol. 78, no. 24, pp. 35195–35209, 2019.

[16] M. J. Hussain, S. H. Wasti, G. Huang, L. Wei, Y. Jiang, and Y. Tang, "An approach for measuring semantic similarity between Wikipedia concepts using multiple inheritances," *Information Processing & Management*, vol. 57, no. 3, p. 102188, 2020.

[17] S. Jain, K. R. Seeja, and R. Jindal, "A new methodology for computing semantic relatedness: modified latent semantic analysis by fuzzy formal concept analysis," *Procedia Computer Science*, vol. 167, pp. 1102–1109, 2020.

[18] R. L. Moseley and F. Pulvermüller, "What can autism teach us about the role of sensorimotor systems in higher cognition? New clues from studies on language, action semantics, and abstract emotional concept processing," *Cortex*, vol. 100, pp. 149–190, 2018.

[19] A. Tejero-de-Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, "Summarization of user-generated sports video by using deep action recognition features," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2000–2011, 2018.

[20] F. R. Dreyer and F. Pulvermüller, "Abstract semantics in the motor system? –an event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning," *Cortex*, vol. 100, pp. 52–70, 2017.

[21] X. Jiang, J. Yang, X. Tan, and H. Xi, "Observation-based optimization for POMDPs with continuous state, observation, and action spaces," *IEEE Transactions on Automatic Control*, vol. 64, no. 5, pp. 2045–2052, 2019.

[22] Y. Zhan, S. Dai, Q. Mao et al., "A video semantic analysis method based on kernel discriminative sparse representation and weighted KNN," in *IEEE international conference on Green Computing & Communications, IEEE & Internet of things*, pp. 1360–1372, 2015, IEEE.

[23] L. Y. Tarhan, C. E. Watson, and L. J. Buxbaum, "Shared and distinct neuroanatomic regions critical for tool-related action production and recognition: evidence from 131 left-hemisphere stroke patients," *Journal of Cognitive Neuroscience*, vol. 27, no. 12, pp. 2491–2511, 2015.

[24] A. Mubashar, K. Asghar, A. R. Javed et al., "Storage and proximity management for centralized personal health records using an IPFS-based optimization algorithm," *Journal of Circuits, Systems and Computers*, p. 2250010, 2021.

[25] C. Beaudry, R. Peteri, and L. Mascarilla, "An efficient and sparse approach for large scale human action recognition in

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3927773, 11 pages
https://doi.org/10.1155/2021/3927773

*Research Article*

# Design and Implementation of Rural Community Elderly Culture Platform Based on Real-Time Social Media Data Mining

**Yangang Zhou**[1] **and Xiao Hu**[2]

[1]*School of Fine Arts and Art Design of Kunming University, Kunming City, China 650214*
[2]*School of Arts, Wuhan Institute of Technology, Wuhan City, China 430205*

Correspondence should be addressed to Xiao Hu; 07091401@wit.edu.cn

Received 12 August 2021; Accepted 17 September 2021; Published 26 October 2021

Academic Editor: Rajesh Kaluri



In the current big data environment, science and technology not only provide a new governance model for rural community governance but also put forward higher requirements for rural community governance level. Under the background of rural revitalization, promoting the construction of rural community cultural service system is not only an important choice to realize the equalization of urban and rural basic public services but also an important way to protect the cultural rights and interests of rural residents. On the basis of analyzing the real-time data of social media, this paper studies the design and implementation method of rural community culture platform and then puts forward the strategy of community public culture informatization construction under the background of aging. From a global perspective, all countries have their own ways and means to invest in public cultural services. Especially from the perspective of countries with better development of public cultural services, multichannel funding sources are an important indicator of the quality of cultural undertakings. With the development of China's social economy, the rural endowment insurance system is becoming more and more perfect, and the basic living needs of the elderly are basically met.

## 1. Introduction

Based on the characteristics of social media digitization and interactivity, the rural community informatization construction mode is being reconstructed, and the new chapter of rural community governance with rural residents as the core and interactive dialogue as the characteristics is slowly opening [1]. The construction of rural community is a systematic project, covering infrastructure, economy, culture, and other dimensions. As an important part of the construction of new socialist countryside, the importance of informatization in rural community construction is self-evident [2]. The development and innovation of community cultural activities are of great significance for the city to build a harmonious society and carry forward the socialist spiritual civilization culture [3]. To strengthen the cultural construction of rural communities, with the support and joint participation of the government and all sectors of society, and in combination with the specific requirements of China's political, economic, social, and cultural development in the new era, efforts should be made to create a lifestyle based on meeting the material and spiritual needs of the elderly, which is to promote and practice the socialist core values in the new era [4]. It is an important guarantee to enhance the ideological and moral cultivation of citizens and promote the stability, harmony, civilization, and progress of the whole society [5]. Under the background of Rural Revitalization Strategy, it is of great practical significance and academic value for the construction of Chinese cultural undertakings to study the achievements, existing problems, causes and measures of the current construction of Chinese rural community cultural service system, explore the relationship between Chinese excellent traditional culture and public cultural service system, and study the spiritual connotation of Chinese traditional culture [6].

Although compared with urban areas, the Internet penetration rate in rural areas still needs to be improved, and the trend of rapid penetration of the Internet into rural areas

is irreversible. Especially in recent years, with the rapid development of mobile Internet, social media is quietly approaching rural residents, affecting all aspects of rural life [7]. All these have laid a good foundation for the social media era to open a new chapter of rural community informatization construction. Rural Revitalization Strategy is a major decision-making deployment of the 19th National Congress of the Communist Party of China, a general charter to guide rural work in the new period, and a driving force to realize rural modernization in an all-round way [8]. Under the background of rural revitalization, promoting the construction of rural community cultural service system is not only an important choice to realize the equalization of urban and rural basic public services but also an important path to protect the cultural rights and interests of rural residents [9]. The elderly is an important part of Chinese society, their living conditions are directly related to the level of social development, and their relationship directly affects the degree of social harmony [5]. With the development of China's social economy, the rural endowment insurance system is becoming more and more perfect, and the basic living needs of the elderly have been basically met. But today's elderly are not only satisfied with food and clothing, they are eager for rich and colorful cultural life and have strong spiritual and cultural needs [10]. To do a good job in cultural construction is an important work in the new era. As a complex and huge system engineering, the community culture construction under the aging trend needs the joint participation of all sectors of society to achieve the desired results [11].

Data mining technology is a kind of data processing technology, which is a process of extracting information and knowledge hidden in a large number of incomplete, noisy, fuzzy and random data, which people do not know in advance and are potentially useful. In the context of big data, promoting the informatization of rural communities is an important part of realizing rural "good governance." With the rapid development of the mobile Internet, social media is rapidly infiltrating the vast rural areas, not only integrating into the daily lives of rural residents but also providing a new carrier support for information dissemination in rural communities [12]. The cultural construction of rural communities is the basic project for the construction of spiritual civilization in our country, and it is also an important task for promoting and practicing the core values of socialism in the new era and building a harmonious, civilized, and stable community environment [13]. With the advancement of the rural revitalization strategy, the construction of public cultural services in rural communities in my country still faces many challenges. The insufficient supply of public cultural services has increasingly hindered the steady progress of the rural revitalization strategy. There is an urgent need to vigorously improve the construction of cultural services in rural communities. From the perspective of national strategy, rural culture, especially public culture, plays an increasingly important role in rural construction and development. It has become the key to rural modernization and the foundation of rural stability and harmony. How to create a stable, harmonious, and healthy social and cultural environment is an inevitable requirement for aging and social development, as well as an objective need to fully realize the goal of building socialism with Chinese characteristics [14]. Based on the real-time data analysis of social media, this paper studies the design and implementation method of rural community culture platform under the background of big data and then puts forward the strategy of community public culture informatization construction under the background of aging. The construction of rural community cultural system is helpful to promote the multidimensional coordinated development of rural areas and can promote the economic and social development of rural areas through the cooperation of multiple subjects.

This paper first discusses the relationship between social media big data and the construction of rural community information from three aspects: the construction requirements of rural community information, the business characteristics of knowledge governance, and the influence of population aging on the construction of rural community culture, then analyzes the application significance of big data in the construction of community cultural activities, and then puts forward the construction method of rural community elderly culture platform based on real-time social media data mining, in order to improve the rural community elderly culture and information construction.

## 2. The Connection between Social Media Big Data and Rural Community Informatization Construction

*2.1. Requirements for the Construction of Rural Community Information.* Social media, also known as social media, is the inevitable product of the development of network media to a certain stage. As a content production and exchange platform based on users' social relations, its basic characteristics are mainly manifested in two aspects, one is the combination of content production and socialization, and the other is that the protagonist of social media platform is users, not website operators [14]. Compared with urban communities, the informationization of rural community management started late, with few people skilled in using computers, lack of professional developers, and few applications of management information systems [15]. A small number of computers and networks are mainly used for simple document processing or low-level entertainment activities such as surfing the Internet and chatting, which are rarely combined with the daily management of the community. Community informatization construction provides a more reliable supporting platform and a good system environment for community service and governance and lays a foundation for the innovation of governance in a smart society. Governance in rural community information is a systematic project, which requires not only the organic division of labor and coordination among various departments but also the sharing and sharing of information [16]. From the current reality, rural community information has limited supply of governance resources and information

platform, limited space for social and market intervention, difficult to show the advantages of professional services of social organizations, difficult to extend the power of market mechanism to optimize resource allocation, and few opportunities for society and market to provide professional guidance for community governance, which is not conducive to giving full play to its role in resource collection, integration and development.

Generally speaking, rural community information governance breaks the barriers between vertical and horizontal governments, realizes deep cooperation between vertical departments and horizontal departments, breaks the pattern of separation of powers and responsibilities, information monopoly and exclusion, and builds government departments at all levels into an information network community by rebuilding the intranet office system and operation management system. The collaborative governance mechanism of rural community information governance is shown in Figure 1.

Information dissemination in the era of social media not only has the characteristics of "digitalization" and "interactivity" of general online media but also breaks the resource barrier of limited computers in the past. To some extent, as long as rural residents have a smart phone, they can catch up with the fast lane of rural community information construction in the new era. With the development of economy and society, the arrival of an aging society, and the advancement of urbanization, the security function of land is gradually weakened, and the traditional pension mode of relying on children and land is no longer enough to guarantee the elderly life of farmers. Carrying out cultural activities in rural communities can enrich the daily life of the elderly, improve their cultural and moral cultivation, change their inherent unhealthy lifestyle, be willing to communicate with each other, share the fun in life, and actively participate in community cultural entertainment and physical fitness activities [17]. There is a big gap between information governance and residents' needs. With the rapid transformation of rural society, the needs of community residents are increasingly diversified, and the demand for individualization and independence is increasing. However, the rural community information service platform focuses on community residents' convenience services, and localization and expansion services are limited. The construction of rural public service system is not only the concrete result of the overall promotion of rural civilization but also the value channel to meet the subjective needs of rural residents.

*2.2. Analysis of Knowledge Governance and Its Business Characteristics.* With the globalization of economy, common knowledge becomes more and more important as the foundation of global governance. However, the era of big data makes the original knowledge more fragmented and decentralized, which makes it very difficult for subjects to form common knowledge [18]. Common knowledge, as the cornerstone of social group decision-making, is also the premise of common action and common value in group decision-making, which greatly affects the judgment and behavior of cognitive subjects. In the process of modernization, informationization has increasingly become an important force to promote rural social change after industrialization, urbanization, and marketization. Information technology has extended into various fields of rural economy and society and promoted the leap forward development of rural economy and society. The diversification of knowledge governance subjects is embodied in the intersubjectivity of multisubjects in rural community public affairs governance, which refers to the relationship between two or more subjects. According to the reality of rural community economic and social development, the national rural strategic planning, the practical interests, and basic concerns of the majority of rural residents, a service function module with multiple functions, various forms and residents' acceptability can be constructed, which can not only ensure the party and state policies to go to the countryside quickly but also ensure the timely feedback of community residents' interests and demands, and realize the timely docking of supply and demand information and the effective supply of services.

*2.3. The Influence of Aging Population on the Construction of Rural Community Culture.* With the profound changes in the social and economic fields, great changes have taken place in the social life fields, which put forward newer and higher requirements for the cultural work, and also put forward new tasks for doing a good job in the cultural work for the elderly. Urban settlement has created wealth and opportunities for society [19]. But with it, the problem of urban population aging has become increasingly prominent. How to create a stable, harmonious, and healthy social and cultural environment, enrich the cultural life of the elderly and meet the spiritual and cultural needs of the elderly is the inevitable requirement of the age and social development of the elderly, and it is also the objective need to fully realize the grand goal of building socialism with Chinese characteristics. With the improvement of material living standards, the demand of the elderly for spiritual and cultural life is getting stronger and stronger. How to meet the growing spiritual and cultural needs of the elderly and enrich the spiritual life of the elderly is of great significance to the realization of family harmony, social harmony, and intergenerational harmony. Through the cultural construction of a certain cultural form and background, the elderly group can naturally form a whole with similar identity points at both psychological and formal levels, which has very important practical significance for the managers of the elderly group. Community is the primary way to solve social problems. Due to the general extension of the life span of the population, the life time of the whole effective labor has been continuously extended, and the overall labor cost in the social economy has decreased; that is, some elderly people have changed from the consumption population to the production population, making due contributions to the creation of social wealth. The analysis framework of cultural platform design in rural community public affairs under big data environment is shown in Figure 2.

Regional culture, as a soft constraint, also plays a vital role in the formation of consensus. The regional culture here is the established system, and the regional cognitive subjects

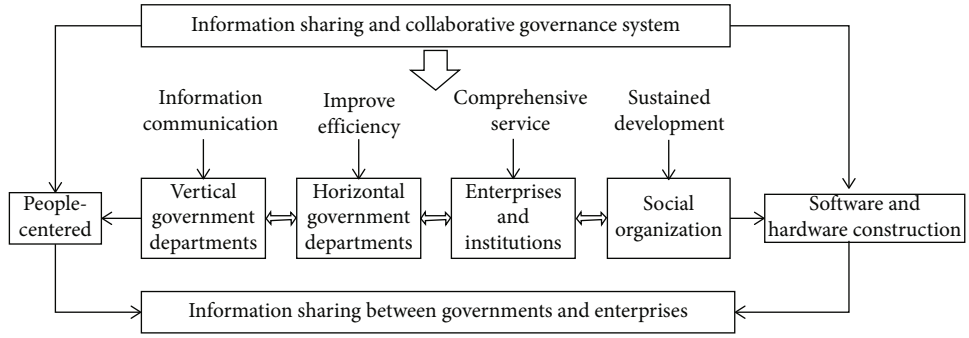



Figure 1: Collaborative governance mechanism of community informatization.

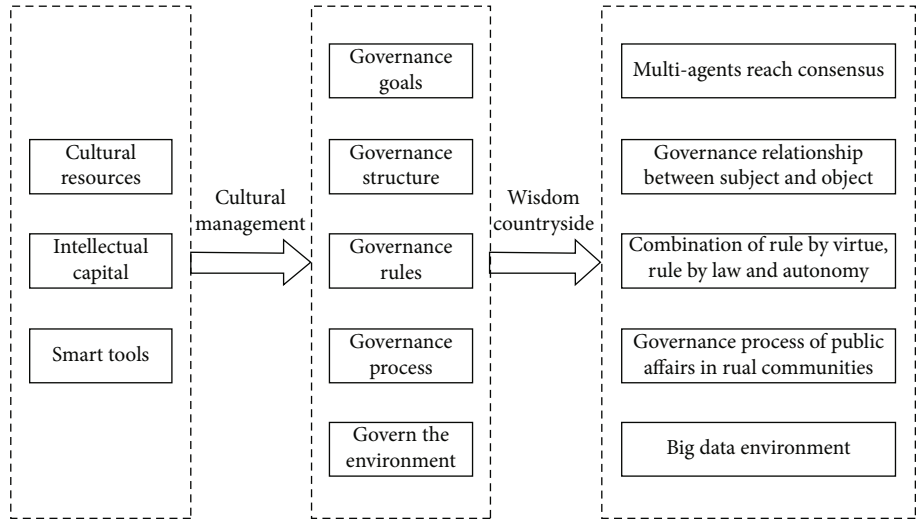



Figure 2: Cultural activity management analysis framework.

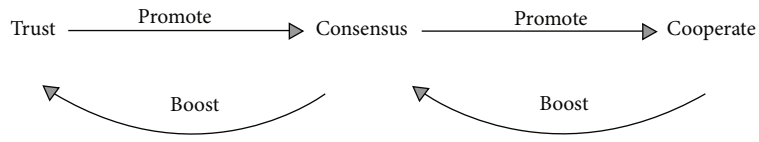



Figure 3: The relationship of trust, consensus, and cooperation.

have a strong resonance with the same regional culture, that is, trust; so, it is easier for the subjects to form consensus. The relationship among trust, consensus, and cooperation is shown in Figure 3.

Rural culture is an important continuation of national culture. After thousands of years of precipitation, it has become an important spiritual support for rural economic and social development. Rural culture has the characteristics of the times, which can closely fit with the development of the times and standardize the social order. From the perspective of farmers' needs, we should build a more diversified rural cultural service system and promote the coordinated development of rural community cultural service system with demand value. Actively carrying out cultural activities in rural communities and innovating new cultural service models are not only conducive to the construction of community culture, increasing the cohesion of community groups, but also reducing many burdens for solving the problem of social pension. On the one hand, the old-age culture has gradually become an indispensable part of community culture construction, and the construction of community old-age culture will meet the development requirements of social aging. On the other hand, the aging of the population puts forward higher requirements for the construction of socialist spiritual civilization. The construction of the rural community cultural service system is formed on the basis of rural culture. In the concrete construction process, it also provides public cultural products for the broad masses of rural residents, enriches the amateur cultural life of the broad masses of the people, and at the same time helps to unite the multisubject forces of rural revitalization.

## 3. The Role of Big Data in the Construction of Community Cultural Activities

*3.1. Promote Community Cultural Activities to Form Feedback Mechanism.* The construction of big data

environment can effectively understand the needs of multisubjects, collect the interests and needs of multirural community governance subjects for public affairs, and improve the accuracy of cultural activities management content. One of the important parts of the management of cultural activities in rural community public affairs is to collect the needs of various subjects and analyze their needs. Because of the characteristics of villagers' mobility, it is difficult for traditional methods to collect all the needs. Using big data, mobile Internet and other technologies can effectively collect the real needs of villagers. Big data technology can also actively stimulate residents' cultural needs and dynamically collect and analyze residents' cultural needs, so as to determine the specific forms of community cultural activities [20]. This kind of big data model has been used in the commercial field for a long time and is quite successful. We can learn from the successful model in the commercial field to build community culture and deal with residents' cultural needs in a timely and efficient manner. For rural communities, the objects that need to provide management of cultural activities may be villagers, enterprises, and public organizations, and the needs of multiple subjects are often different. If the knowledge services provided are uniform and not targeted, the user experience is often very low, relying on big data technology to provide personalized services, improve the utilization rate of knowledge, and improve the governance level and ability of rural communities by classifying governance objects and providing targeted and accurate services. The construction of big data environment can integrate the fragmentation knowledge of rural communities and improve the accuracy of cultural activity management content.

*3.2. Improve the Effective Utilization Rate of Community Cultural Activities Funds.* Big data provides data support for the decision-making behavior of public affairs governance in rural communities by collecting a large number of diverse data resources. And through big data means such as data mining, hidden information can be effectively mined to provide data support for scientific decision-making. The era of big data provides the possibility and feasibility for China's rural communities to move towards information-based governance and finally towards intelligent community governance. The data mining process is shown in Figure 4.

Through real-time social media data mining, residents' cultural needs can be accurately located, and an effective feedback mechanism can be formed, which is conducive to improving the utilization rate of cultural activities resources. It can also provide one-stop community cultural services for community residents, including activities reservation, venue reservation, community settlement, and other functions. Specific to each resident, the activities pushed on the home page may be different. Such accurate push will inevitably enhance the enthusiasm of residents to participate in community cultural activities. The construction of big data environment can effectively play the advantages of multiagents in rural community public affairs governance, achieve a community of interests, coordinate multiagent behavior, and realize the governance pattern of "smart countryside" with multiagent governance.

## 4. Construction of Rural Community Elderly Culture Platform Based on Real-Time Social Media Data Mining

*4.1. Strengthen the Mining and Integration of Resources.* In order to make the cultural construction achieve better results, we need to strengthen the mining of social resources on the basis of the existing cultural facilities for the elderly, gradually build and improve the cultural construction work, and build the service network of information consultation and help for the elderly. In the rational allocation of resources, we can take the form of individual donation, enterprise sponsorship, and government funding to gradually form a set of perfect cultural construction system. The rural community pension model is flexible, suitable for local conditions to develop pension service system, to meet the different needs of different groups. Different levels of economic development in different regions determine that there are differences in the elderly care services provided by rural communities, and the different living conditions and ideas of the elderly also determine that the elderly care needs are not the same [21]. Providing the necessary platform for the entertainment and self-cultivation of the elderly is the basic guarantee for the development of community elderly culture. It is also an effective way to meet the cultural consumption needs of the elderly and promote the development of community elderly culture. At the same time, it is also the most practical and specific work that community workers should carry out. Before infrastructure construction, we must first communicate with the villagers, understand the accurate needs and requests of the villagers in detail, and then formulate the relevant design and construction scheme. In the construction process, it is necessary to arrange the responsible staff to be present in the whole process, deal with the emergency in time, and ensure that the property of the rural people will not be lost in the construction process.

As shown in Table 1, some women working in the household farm and even some younger and healthy elderly people have clearly indicated that they are willing to join the elderly care service team and show a strong learning enthusiasm.

Under the environment of big data, individuals are both information publishers and information receivers. The development of mobile Internet and big data has strengthened the multiple attributes of cultural activity management, which makes it easy for those who are incapable and lack enthusiasm to participate in expressing their views, and enriches the universality of governance subjects, thus strengthening the multiple attributes of governance. The main body of cultural construction is shown in Figure 5.

The characteristics of the rural community pension model determine that it can formulate the pension service system and provide pension service according to local conditions according to the development of different regions and the living conditions of the elderly, so as to meet the different needs of the elderly. In addition, the providers and service contents of community aged care services are various. This diversified form of community aged care services is also convenient to meet the differentiated needs of the elderly and tailor different service contents for the elderly. In order to cope with the crisis caused by the current rural labor

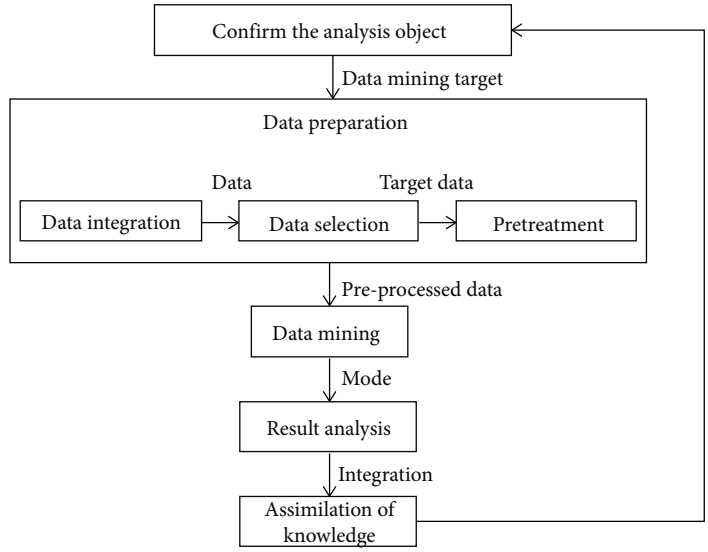



Figure 4: Data mining process.

Table 1: Investigation on farmers' wishes.

| Willingness survey | Willing to join | Unwilling to join | Uncertain |
|---|---|---|---|
| Number of people | 35 | 10 | 11 |
| Proportion | 62.5% | 17.9% | 19.6% |

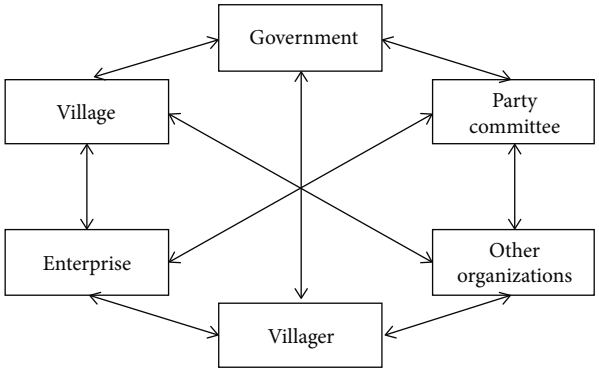



Figure 5: The main body of cultural construction.

outflow and promote the sustainable development of rural economy and society, it is necessary to increase the attractiveness of all kinds of talents including cultural talents and strengthen the construction of cultural service talents in rural communities. In the construction of resource sharing mechanism, government agencies at all levels should give full play to the macro guidance role and give strong support in policies. Community neighborhood committees and other institutions can use their own cultural facilities and resources to create better conditions for the cultural activities of the elderly. To develop the model of providing for the aged in rural communities, we only need to make use of the existing resources and appeal to all sectors of society to donate the equipment needed by the elderly.

Community residents can directly express their own interests through the bottom-up social conditions and public opinion feedback system and directly participate in the decision-making of related community affairs and participate in community public governance. The multilevel community resident two-way information transmission and participation governance system are shown in Figure 6.

Get all the items scored by community users $i$ and $j$, then calculate the similarity between them through different similarity measurement methods, and record it as $\mathrm{sim}(i, j)$. This article uses a modified cosine similarity calculation method:

$$\mathrm{sim}(i,j) = \frac{E_{c\in I_{i,j}}\left(R_{i,c}-\bar{R}\right)\left(R_{j,c}-\bar{R}_c\right)}{\sqrt{\sum_{c\in I_{i,j}}\left(R_{i,c}\left(R_{i,c}-\bar{R}_c\right)^2\times\sum_{c\in I_{i,j}}\left(R_{j,c}-\bar{R}_c\right)^2\right)}}. \quad (1)$$

$R_{i,c}$ is the rating of user $i$ on item $c$, and $\bar{R}_c$ is the average rating of item $c$. After calculating the similarity between users, for a user $u$, a set of "neighbors" arranged according to the similarity is generated, $N = \{U_1, U_2, \cdots, U_t\}$, $0 \le t \le m$, and $u$ does not belong to $N$. From $U_1$ to $U_t$, $\mathrm{sim}(u, U_i)$ $(1 \le i \le t)$ is arranged in descending order.

When providing public cultural service products, local grass-root governments should not only consider the universal public cultural service demand and the extension of regional focu, but also pay attention to meeting the cultural needs of different groups. Due to the long return period of public cultural service construction, it is limited by the financial sustainability of local governments. Therefore, it is necessary for relevant government departments to make efforts to help the poor with cultural precision, enhance the ability of local governments to effectively build public cultural services, and ensure the optimal allocation of funds and resources. The elderly living alone have many inconveniences and are prone to loneliness and depression, which indicates that the social psychological support of the elderly is also a problem that cannot be ignored in community

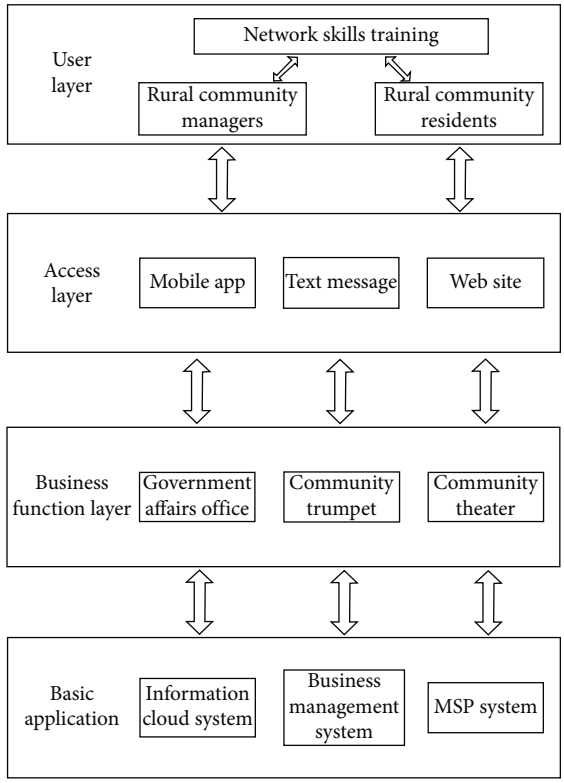



Figure 6: Two-way information transmission and governance system for community residents.

nursing services. The prevalence of elderly living alone in community was investigated, and the results are shown in Table 2.

Chronic diseases are the most common diseases that affect the health of the elderly, which may lead to a decline in the mobility of the elderly. The ranking and prevalence of chronic diseases of the elderly are shown in Table 3.

To construct rural community public cultural service with regional and national characteristics is to establish the key development direction according to the advantages of rural historical and cultural resources in different regions and to integrate urban cultural consumption and creative design elements into rural traditional culture to form a new brand of rural community public cultural service. Therefore, we can give full play to the cultural productivity in rural areas, make up for the imbalance between supply and demand of cultural service products in rural communities, and effectively play its cultural and economic functions while protecting traditional local culture. For rural residents, they are direct participants and beneficiaries of the old-age care model in rural communities. The most important thing to develop the old-age model in rural communities is to break the old-age concept of "raising children and preventing old age" in farmers' psychology for a long time and establish the old-age awareness of mutual assistance and mutual aid. Community workers should strengthen the guidance and management of cultural and sports activities for the elderly, take the community as the place and the residents as the main body, and carry out a series of interesting,

Table 2: The prevalence of elderly people living alone in the community.

| Number of disease types | Number of cases | Proportion (%) |
|---|---|---|
| 0 | 158 | 25.56 |
| 1 | 264 | 42.6 |
| 2 | 119 | 19.2 |
| 3 | 52 | 8.4 |
| More than three kinds | 27 | 4.4 |

Table 3: The order and prevalence of chronic diseases in the elderly.

| Disease name | Prevalence rate (%) |
|---|---|
| Hypertension | 30.05 |
| Diabetes | 7.76 |
| Ischemic heart disease | 7.54 |
| Cerebrovascular disease | 6.01 |
| Rheumatoid arthritis | 2.35 |
| Herniated disk | 1.68 |
| Cataract | 1.23 |
| Benign prostatic hyperplasia | 0.95 |

connotative, competitive, cooperative, and distinctive cultural and sports activities from time to time [22]. We should encourage more community residents volunteers to join the community cultural construction team, especially those who have artistic hobbies and stylistic expertise and are enthusiastic about community affairs, give full play to their leading role, organize the establishment of community cultural activity groups in various forms, and cultivate more elderly residents to become literary lovers and literary backbones, forming a linkage effect [23]. Cultural departments at all levels should actively carry out high-quality cultural activities in conjunction with aging departments, workers, youth, women, and other mass organizations, Federation of Literary and Art Circles associations and social and cultural organizations according to their respective work characteristics. We should encourage all sectors of society to care for and support cultural activities and vigorously advocate and organize volunteer teams to provide various services including spiritual comfort and cultural life for the elderly.

As shown in Table 4, 76% of rural residents think that they will rely on their children to support the elderly in the future, and they think that their old age life should be taken care of by their children.

A survey on whether farmers are willing to accept the old-age model in rural communities is shown in Table 5. The way of providing for the aged in rural communities can not only ensure that the elderly do not leave their familiar living environment as much as possible, so that they can live with their neighbors, friends and family, but also receive the aged care services provided by the community.

Under the current situation, a large part of rural community pension services need community members to help each other directly. That is to say, if residents and neighbors cannot form a good atmosphere of mutual help and mutual

Table 4: Choices of farmers' pension methods.

| Ways of providing for the aged | Family pension | New rural insurance | Other |
|---|---|---|---|
| Number of people | 41 | 8 | 11 |
| Proportion | 69.5% | 13.6 | 18.6 |

Table 5: Investigation on the willingness of the community pension model.

| Willingness survey | Receptibility | Unacceptable | Uncertainty |
|---|---|---|---|
| Number of people | 38 | 5 | 13 |
| Proportion | 67.9% | 8.9 | 23.2 |

support for the elderly, the development of rural community pension model cannot be discussed. Only when the rural people understand the national policy can they consciously exert their sense of ownership and actively exert their supervision. At the same time, rural people can learn the spirit of national policy documents through this platform, change their ideas, broaden their horizons, and improve their overall quality [24].

*4.2. Promote the Organic Combination of Cultural Construction and Cultural Industry for the Elderly.* To carry out cultural activities, we should start from the needs and aspirations of the elderly and combine the characteristics of different places. Relevant departments must adhere to the community-based, grassroot-oriented, and practical. For public cultural places such as exhibition halls, the elderly can be given appropriate preferential treatment or even free of charge. On the occasion of major traditional festivals every year, giving priority to arranging condolence performances for the elderly has gradually become a system and mobilizes the participation consciousness of the villagers. Villagers are not only the beneficiaries of rural community cultural services but also important participants [25]. Community-based old-age care is a kind of old-age care way based on family-based old-age care, taking the community as the platform, taking the relevant service organizations in the community as the leading force, effectively mobilizing and integrating various resources, and providing all-round and multilevel old-age care services for the elderly. The basic structure of the community pension system is shown in Figure 7.

In order to coordinate various actors with different interests, it is necessary to apply various governance methods, that is, to coordinate the needs and interests of multiple subjects through a series of formal and informal institutional arrangements to realize collective action. Traditional governance rules include bureaucratic governance rules, market governance rules, and network governance rules. Bureaucratic governance rules are based on power relations, and its governance structure is from top to bottom. In the current big data environment, human resources, intellectual capital, social capital, knowledge, and other resources are scattered and fragmented. It is difficult for a single subject

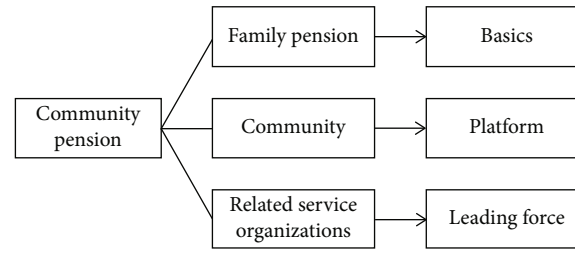



Figure 7: The basic structure of the community pension system.

to achieve its goal by relying on a certain resource. It is necessary to integrate the resources scattered in the hands of multiple subjects to achieve collective action. Market governance rules are based on the economic principle of interaction between consumers and product suppliers, which is the coordination of an exchange system constructed by a rational economic man through "invisible hand." Network governance rules are based on the basic principle of trust to allocate public goods resources. The advantages and disadvantages of bureaucratic governance, market governance, and network governance rules are shown in Table 6.

With the rapid development of society, rural people are paying more and more attention to personal health. It is necessary to strengthen the leadership of cultural work for the elderly, strengthen the research, guidance and management of cultural activities, and ensure the healthy and orderly development of cultural undertakings for the elderly. We should adhere to materialist dialectical thought and correct orientation, actively guide the elderly to carry out various healthy, beneficial and scientific cultural and fitness activities, and oppose various activities that are not conducive to the physical and mental health of the elderly and superstition and pseudoscience. By building a place for mass activities in rural areas, let the people have places and equipment for activities together, let the masses organize activities spontaneously, and enhance the cohesion in rural areas [22]. By strengthening physical exercise, the rural people can strengthen their physical fitness, relieve their fatigue, and benefit their physical and mental health. By organizing activities, the rural people's spiritual and cultural needs were met, and the socialist core values were also educated. Within the statutory authority, the relevant departments of local governments should formulate rules and regulations with legal effect and relevant standards of public cultural services according to the actual situation of the development of public cultural services in the region. The management process of the elderly culture of the two committees in rural community governance under the big data environment is shown in Figure 8.

Rural cultural management workers actively organize and guide rural people to carry out activities they are willing to carry out, and at the same time, embody socialist core values in the process of activities. Perfecting the multichannel funds for rural community cultural construction is an important link to ensure the steady advancement of rural

Table 6: Comparison of advantages and disadvantages of traditional governance rules in rural communities.

| Governance rules | Advantage | Disadvantage |
|---|---|---|
| Bureaucratic governance | Centralization and strong control | The system is rigid and inefficient |
| Market governance | Market regulates resource allocation and pays attention to efficiency | Market failures such as insufficient supply of public goods |
| Network governance | Closely linked | Prone to trust issues |

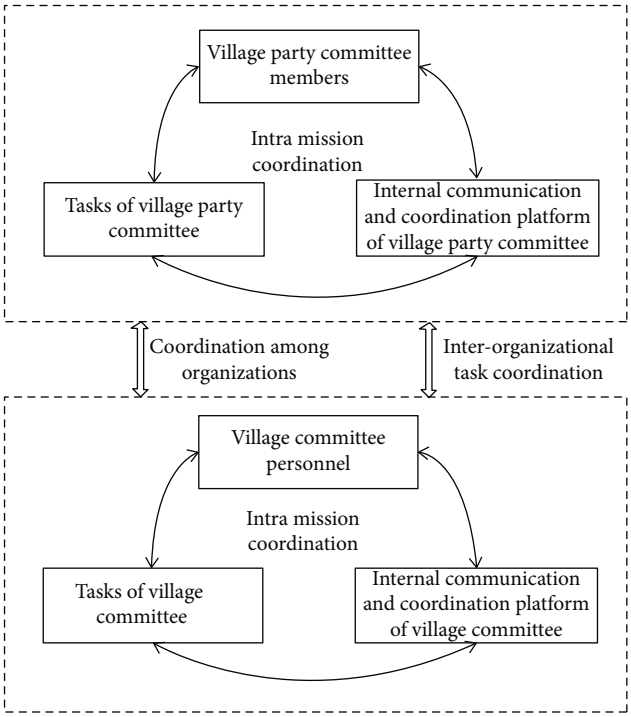



Figure 8: The management process of the elderly culture in the village committees.

cultural undertakings. The general situation and physiological functions of the respondents were analyzed by single factor, and the results of stepwise regression analysis are shown in Table 7.

Suppose the expected output is $z'_k$, and the global error between expected and actual output is defined as $L$:

$$L = \frac{1}{2}\sum_{k=1}^{m}\left(z_k - z'_k\right)^2. \tag{2}$$

Through the back propagation process, the error is expanded to the hidden layer as

$$L = \frac{1}{2}\sum_{k=1}^{m}\left[f(\lambda_k) - z'_k\right]^2 = \frac{1}{2}\sum_{k=1}^{m}\left[f\left(\sum_{k=1}^{t} w_{jk}y_j + b_k\right) - z'_k\right]^2. \tag{3}$$

Table 7: Stepwise regression analysis of general conditions and physiological functions.

| Variable | Partial regression coefficient | Standardized partial regression coefficient | Standard error |
|---|---|---|---|
| Self care ability | -14.37 | -0.32 | 2.51 |
| Perplexing problems | -2.87 | -0.15 | 0.78 |
| Medical expenses | 2.51 | -0.28 | 1.23 |
| Nursing needs | -6.79 | -0.16 | 2.59 |

Finally, the reverse transmission to the input layer is

$$\begin{aligned} L &= \frac{1}{2}\sum_{k=1}^{m}\left[f\left(\sum_{k=1}^{t} w_{jk}y_j + b_k\right) - z'_k\right]^2 \\ &= \frac{1}{2}\sum_{k=1}^{m}\left[f\left(\sum_{k=1}^{t} w_{jk}\left(\sum_{j=1}^{n} w_{ij}a_i + b_j\right) + b_k\right) - z'_k\right]^2. \end{aligned} \tag{4}$$

The network error is a function of the weights $w_{ij}$ and $w_{jk}$. Therefore, the error $E$ can be changed by changing the weight of the neuron; thus,

$$\Delta w_{ij} = -\varepsilon\frac{\partial L}{\partial w_{ij}}(i = 1\cdots m, j = 1\cdots n), \tag{5}$$

$$\Delta w_{jk} = -\varepsilon\frac{\partial L}{\partial w_{jk}}(j = 1\cdots n, k = 1\cdots t). \tag{6}$$

Among them, $\varepsilon$ represents the rate, and $\varepsilon$ (0, 1).

From the world point of view, countries have their own ways and means to invest in public cultural services, especially from the perspective of countries with better development of public cultural services, and multichannel funding sources are an important indicator of the quality of cultural undertakings. Rural cultural workers should work hard, make full use of the existing material conditions, realize the "sharing" of rural community cultural service system infrastructure, give full play to the role of infrastructure, meet the spiritual and cultural needs of rural people, and also carry out the education of socialist core values [26].

*4.3. Development Proposal.* Chinese rural people have lived in the environment of local traditional culture for generations, forming the rural life regulations and ethical management with Chinese traditional rural characteristics and maintaining the social stability of Chinese rural areas. The purpose of realizing the rural revitalization strategy is to meet the spiritual and cultural needs of the people in rural areas of China, and it is necessary to protect the diversity of Chinese regional culture. On the basis of understanding people's real needs through various investigation methods, the villagers should be guided to actively participate in the construction of public cultural services in rural communities, so as to stably promote the sustainable operation of the construction. Theoretically speaking, we should pay attention to the continuous study of relevant laws and regulations, mold our own cultural legal literacy, and firmly establish the legal concept of equalization of public cultural services. The construction of community culture can promote Chinese excellent traditional culture, and at the same time, it can continuously improve the moral cultivation, political consciousness, and cultural accomplishment of community residents and then enhance their cultural self-confidence and consciously practice the socialist core values. Only when the rural people have a good health can they live and work in peace and contentment. It is also part of the rural revitalization strategy to build a public activity place so that rural people can move freely.

As a basic work of revitalizing rural culture and promoting rural revitalization in China, while following the task requirements of implementing the rural revitalization strategy and adhering to the socialist core values as the guide, we should take meeting the public cultural needs of farmers as the starting point and strengthen the specific path of working coordination among governments at all levels and relevant departments [27]. As far as the activity places for the elderly are concerned, the existing village branches, rural school buildings, and even some abandoned houses can be used for activities in rural areas. In the process of infrastructure construction, make clear the responsibilities of each post, so as to ensure that the process is supervised, the actions are disciplined, and the results are fed back. At the same time, the acceptance of the project should realize the lifelong responsibility system, ensure that the public cultural service construction funds are used in the cutting edge, and form a high-pressure situation in the system. The community should actively carry out the construction of cultural and sports organizations for the elderly and cultivate a number of elderly cultural and sports backbone teams by relying on regional elderly activity centers and cultural centers to promote the development of grassroots elderly cultural and sports activities. Actively introducing market mechanism and realizing the economic value of old-age culture in the way of market mechanism operation can effectively solve the dilemma of lack of funds, and also help to provide a steady stream of manpower and material resources for the development of old-age culture.

## 5. Conclusions

With the increasing popularity of rural Internet, social media is helping to open a new chapter in the development of rural community information. Under the background of aging population, the construction of rural community culture is conducive to the formation of harmonious, harmonious, and peaceful community atmosphere. Based on the real-time data analysis of social media, this paper studies the design and implementation method of rural community culture platform under the background of big data and then puts forward the strategy of community public culture informatization construction under the background of aging. In a good community humanistic environment, neighbors watch each other, respect each other, and love each other, which effectively alleviates the widespread insecurity and depression of modern people and enables the elderly to live a truly quality old age. Rural villagers are beneficiaries and builders of the construction of the rural community cultural service system. Governments at all levels should adhere to the people-centered working principle, let rural villagers participate in the whole process of building the rural community cultural service system, and give full play to their subjective initiative.

The research on the management of cultural activities in rural community public affairs under the big data environment plays an important role in promoting the modernization of rural community governance level and governance capacity. In the future, the community should actively carry out the construction of cultural and sports organizations for the elderly, rely on regional activity centers and cultural centers for the elderly, cultivate a number of backbone teams of cultural and sports for the elderly, and promote the development of cultural and sports activities for the elderly at the grass-root level.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no competing interest.

## References

[1] M. Shade, J. Boron, N. Manley, K. Kupzyk, and C. Pullen, "Ease of use and usefulness of medication reminder apps among rural aging adults," *Journal of Community Health Nursing*, vol. 36, no. 3, pp. 105–114, 2019.

[2] B. Elizalde-San Miguel and V. Díaz-Gandasegui, "Aging in rural areas of Spain: the influence of demography on care strategies," *History of the Family*, vol. 21, no. 2, pp. 214–230, 2016.

[3] H. Yangyang, "Social reconstruction mode of left-behind elderly under the deconstruction of Rural Society in Aggregated Nursing Homes," *Journal of Sichuan University of Science & Technology (Social Science Edition)*, vol. 31, no. 4, pp. 58–67, 2016.

[4] L. Gongping and F. Fang, "Practice and research on health education for the elderly in township communities under the background of rural revitalization strategy," *China Rural Education*, vol. 287, no. 1, pp. 42–44, 2019.

[5] H. Wang, "The aging of rural population and the evolution of rural space," *Population Research*, vol. 43, no. 5, pp. 66–80, 2019.

[6] T. Guojiang, "Experiments and reflections on elderly education carried out by rural community education centers," *China Rural Education*, vol. 285, no. 23, pp. 50-51, 2018.

[7] G. Yaqiao, J. Ma, J. Wang, and Q. Yang, "The lack of humanistic feelings in the countryside and its reasons," *Housing and real estate*, vol. 492, no. 7, pp. 282–290, 2018.

[8] L. Liao and X. Gao, "Research progress and prospects of the impact of population aging on rural development," *Advances in Geographical Sciences*, vol. 37, no. 5, pp. 617–626, 2018.

[9] Z. Mengna, "Research on the design of suitable aging of rural environmental landscape under the background of aging," *Art Technology*, vol. 32, no. 10, pp. 183-184, 2019.

[10] H. Pan, "Opportunities, challenges and countermeasures for rural revitalization caused by the aging of migrant workers," *China economic and trade guide*, vol. 949, no. 20, pp. 57–79, 2019.

[11] F. Xiao, J. Tan, L. Zhaojun, and S. Jinyun, "The enlightenment of Taomi Village community construction on the activation of rural cultural heritage," *Urban Architecture*, vol. 16, no. 2, pp. 85–87, 2019.

[12] X. Deng and H. Jianyun, "Rural community planning methods and practices from the perspective of cultural guidance," *The Planner*, vol. 35, no. 23, pp. 40–46, 2019.

[13] P. Xiaowen and W. Junying, "The impact of rural population aging on rural revitalization strategy," *Cooperative Economy and Technology*, vol. 597, no. 22, pp. 13–15, 2018.

[14] H. Shuo, ""aging of farmers" under the strategy of rural revitalization and its countermeasures," *Shanxi Agricultural Economics*, vol. 270, no. 6, pp. 78-79, 2020.

[15] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big dat," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[16] Q. Wu, L. Tongtong, Q. Jianling, L. Yingchun, and Z. Runyun, "Research on rural community pension model under the background of rural revitalization strategy," *Anhui agricultural Science Bulletin*, vol. 399, no. 17, pp. 13–44, 2020.

[17] N. Deepa, Q.-V. Pham, D. C. Nguyen et al., "A survey on blockchain for big data: Approaches, opportunities, and future directions," 2020, http://arxiv.org/abs/2009.00858.

[18] W. Xu and W. Wang, "Research on the integration of rural hollow governance and community construction under the background of rural revitalization strategy," *Journal of Agricultural and Forestry Economic Management*, vol. 18, no. 3, pp. 416–423, 2019.

[19] Q. Feng and Z. Guanghua, "Thoughts on the development model of rural ecological communities under the background of beautiful rural construction," *Science & Technology Economic Guide*, vol. 639, no. 13, pp. 88-89, 2018.

[20] F. Qiong and Q. Wu, "The reconstruction of traditional etiquette culture in the rural revitalization strategy," *Chongqing Social Sciences*, vol. 287, no. 10, pp. 33–41, 2018.

[21] Y. Guo, Z. Yang, and H. Yue, "The spatiotemporal evolution of the aging of rural population in China and rural revitalization strategies," *Geographical Research*, vol. 38, no. 3, pp. 667–683, 2019.

[22] Y. Zaiting, "The changes, predicaments and choices of rural culture from the perspective of modernity," *Journal of Beijing Institute of Graphic Communication*, vol. 25, no. 8, pp. 51–55, 2017.

[23] C. Ye, X. Wang, C. Zheng, Y. Ji, and Y. Xie, "Research on the issue of elderly care for the elderly in empty nest in the new situation," *Rural Science & Technology*, vol. 252, no. 12, pp. 33-34, 2020.

[24] C. Ge and S. Shangeng, "Research on the suitable ageing reformation of rural living environment in urban suburbs," *Rural Science & Technology*, vol. 207, no. 3, pp. 16–18, 2019.

[25] Y. Hu, "The predicament and solution of rural cultural construction under the background of rural revitalization strategy," *Rural Economy and Technology*, vol. 30, no. 11, pp. 247-248, 2019.

[26] H. Wang, H. Lu, F. Zhang, A. Zhang, J. Guo, and J. Dong, "Pursuing a better eating life: a new concept of building shared restaurants in rural communities," *Society and Public Welfare*, vol. 2020, no. 1, pp. 58–61, 2020.

[27] P. Ying, "Discussion on rural revitalization strategy and intangible cultural heritage protection," *Shanghai Urban Management*, vol. 27, no. 4, pp. 8–13, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3920325, 13 pages
https://doi.org/10.1155/2021/3920325

*Research Article*

# Real-Time Twitter Trend Analysis Using Big Data Analytics and Machine Learning Techniques

**Anisha P. Rodrigues, Roshan Fernandes, Adarsh Bhandary, Asha C. Shenoy, Ashwanth Shetty, and M. Anisha**

*Department of Computer Science and Engineering, NMAM Institute of Technology, Nitte, Karkala, India*

Correspondence should be addressed to Roshan Fernandes; roshan_nmamit@nitte.edu.in

Received 28 July 2021; Accepted 4 October 2021; Published 25 October 2021

Academic Editor: Rajesh Kaluri



Twitter is a popular microblogging social media, using which its users can share useful information. Keeping a track of user postings and common hashtags allows us to understand what is happening around the world and what are people's opinions on it. As such, a Twitter trend analysis analyzes Twitter data and hashtags to determine what topics are being talked about the most on Twitter. Feature extraction and trend detection can be performed using machine learning algorithms. Big data tools and techniques are needed to extract relevant information from continuous steam of data originating from Twitter. The objectives of this research work are to analyze the relative popularity of different hashtags and which field has the maximum share of voice. Along with this, the common interests of the community can also be determined. Twitter trends plan an important role in the business field, marketing, politics, sports, and entertainment activities. The proposed work implemented the Twitter trend analysis using latent Dirichlet allocation, cosine similarity, K means clustering, and Jaccard similarity techniques and compared the results with Big Data Apache SPARK tool implementation. The LDA technique for trend analysis resulted in an accuracy of 74% and Jaccard with an accuracy of 83% for static data. The results proved that the real-time tweets are analyzed comparatively faster in the Big Data Apache SPARK tool than in the normal execution environment.

## 1. Introduction

Twitter is a popular social networking site where millions of people tweet every second about various topics related to society, politics, sports, entertainment, and many more. The standard syntax followed by Twitter users while tweeting involves hashtags, retweets, and user mentions. Hashtags are words or phrases which are prefixed with "#," and user mention means mentioning other people, companies, brands, or precisely other Twitter users in the tweet by using the "@" symbol at the beginning of their user name. There is a restriction of 140 characters on the length of any tweet which allows users to post tweets quickly. At the same time, users all across the globe can tweet about anything happening or their thoughts at any given time of the day. Tweets thus help people to understand how others feel about different ongoing events, government policies, sports tournaments, etc. Brands can analyze tweets to know people's sentiments towards their products. Government and politicians get an idea of how people are responding to the different policies, acts, and amendments. During elections, Twitter plays a vital role in campaigning too. For a given day or a span of days, any topic can be made trending by the repeated use of the same hashtag. Thus, Twitter trends play an important role in the process of decision-making by different organizations and companies. The main motivation for the Twitter trend analysis is to identify the recent trends happening across the world using big data machine learning techniques. This will help to analyze what has happened in the past and what may happen in the future. It helps to track customer trends and interests especially what customers like, what their behaviors are, and how this changes over the time.

In the proposed work, the tweets are collected using Twitter API and applied counting methods and different machine learning algorithms to identify trending topics on

Twitter. Twitter API provides a standard way to read and write Twitter data. This API provides a set of methods that can be used to communicate with the application. To process a huge volume of tweets instantaneously, we have used SPARK streaming. SPARK is a big data tool that can be effectively used to deal with a large volume of data in a short time. Hashtag counting and noun counting are the two basic methods that count the hashtags and nouns in tweets, respectively, to determine which particular word is trending. Topic modeling technique latent Dirichlet allocation (LDA) is used, which groups the tweets into clusters of topics based on keywords. Cosine similarity measures how similar two or more documents are and groups the tweets accordingly. K means clustering and Jaccard similarity also help us to classify tweets into clusters. By using SPARK streaming, we were able to identify real-time trends more quickly as compared to a normal execution environment. We have performed an analysis of the time taken to execute the programs on static data and real-time data collected using SPARK. We have also included analysis for May 2021 which shows us the output obtained using different techniques and helps us to conclude that all algorithms run efficiently and give accurate trends.

*1.1. Contributions of the Proposed Work.* By carefully analyzing many works of literature in the field of Twitter data analysis, we have concluded that the majority of researchers have contributed towards Twitter sentiment analysis than trend analysis. Few researchers who contributed to trend analysis have used LDA and clustering techniques using SPARK. The main contribution of the proposed work is to perform the Twitter trend analysis. This includes applying the various techniques for Twitter trend analysis and comparing the results using various evaluation parameters. The techniques used are hashtag counting, noun counting, cosine similarity, Jaccard similarity, LDA, and K-means. These techniques are applied to static Twitter data as well as real-time streaming data and compared the results. We obtained better results in terms of execution speed for real-time Twitter trend analysis using SPARK.

## 2. Related Work

To identify sentiments in tweets, lexicon-based methods and polarity multiplication have been used [1]. NLP techniques like tokenization, removal of stopwords, and stemming are used for preprocessing. The lexicon method is simpler and has lower accuracy compared to machine learning. Hence, machine learning techniques must be used for analyzing tweet sentiments and trends. Machine learning algorithms like Naïve Bayes, SVM, and KNN were used for sentiment analysis [2, 3]. Out of the three Naïve Bayes was found to achieve the highest accuracy, i.e., 80.9% followed by KNN with an accuracy of 75.58%. Latent Dirichlet allocation which is a topic modeling algorithm was used to analyze tweets and extract useful information from them [4]. Using LDA, a large number of tweets are processed as a collection of documents, where each document is associated with a collection of topics. Each topic is associated with a set of words, and each document has a different proportion of topics based on the frequency of words that appear in each topic. The same method was used by Negara et.al [5] to process a large number of tweets and divide them into 4 clusters, namely, economic, sports, military, and technology. LDA algorithm was found to have optimal performance for Sports tweets with an accuracy of 98% which is better than LSI topic modeling.

Shahreen et al. [6] have used the machine learning and neural network approach for the text analysis. SVM was used for text analysis, and weight optimizers like Limited-memory BFGS, Stochastic gradient descent, and Adam were used to attaining maximum accuracy using neural networks. They obtained an accuracy of 95.2% with SVM and 97.6% using a neural network. Hidayatullah et al. [7] performed topic modeling on a dataset obtained from the official Twitter account of traffic management center in Java to create a topic model regarding traffic information. Hasan et al. [8] have planned the analysis in two phases. In phase 1 after data acquisition, researchers have preprocessed the data stream using tokenization and stop word removal; then, they have clustered using improved fuzzy C-means clustering and adaptive particle swarm optimization. They have examined Twitter data streaming using an Apache SPARK engine. In phase 2, the data is preprocessed, and they have classified Higgs data using modified SVM, and Higgs data streaming is examined using an Apache SPARK engine. The computational analysis shows that it achieved better results compared to existing methods in terms of F-score, precision, ROC curve, and accuracy. Garg and Kaur [9] have explained the analysis of Twitter data using components of Cloudera distribution of Hadoop. The objective is to assign polarity to each tweet. Map reduce and Apache SPARK frameworks were used for sentiment analysis. The result shows that Apache SPARK is better than MapReduce. Saad and Yang [10] have performed sentiment analysis of Twitter data using ordinal regression. The preprocessed tweets are run using different machine learning algorithms. These algorithms reveal the polarity of tweets. The algorithms used were support vector regression, decision tree, random forest, and multinomial logistic regression of which decision tree showed the highest accuracy.

Hasan et al. [11] used machine learning techniques to perform sentiment analysis. Polarity calculation and sentiment analysis were performed using Text Blob, Sentiwordnet, and W-WSD and then classified using Naive Bias and SVM. It gives a comparison of techniques of sentiment analysis by applying supervised machine learning algorithms like Naïve Bayes and SVM. Huq et al. [12] have also performed sentiment analysis on Twitter data using machine learning algorithms. They have used SVM (support vector machine), and sentiment classification algorithm (SCA) was built using KNN (K Nearest Neighbour). The performance of both was compared, and SCA is found to be better than SVM. Jianqiang et al. [13] found that convolutional neural networks are better for sentiment classification of tweets. An RBF kernel SVM and LR exploiting unigram and bigram features (BoW) were also used. For twitter sentiment analysis, DCNN using pretrained word vectors was found to have

good performance. Ahmed and Rodríguez-Díaz [14] have performed sentiment analysis on online customer reviews. Here, text selection, text collection, text processing, sentiment analysis, and regression analysis techniques were used. This project analyzes the customer experience and helps to meet customer demands. Predeveloped lexicons were used to determine positive and negative signs as there is no dynamic element to guide feelings. Rathod and Barot [15] researched the same field to predict public opinion on ongoing events by analyzing tweet sentiments using machine learning classifiers like SVM, Naïve Bayes, logistic classifier, and KNN classifier. SVM was found to be the best classifier with the least mean square error for the classifications. Garg et al. [16] have identified the trending pattern in Twitter using SPARK. These patterns were obtained by collecting tweets on a real-time basis and identifying trending hashtags at the same time. It was implemented using a big data technology SPARK streaming. This helps companies to know about their brand awareness and customer needs. To handle a large number of tweets from Twitter on a real-time basis, SPARK framework has been used. Sentiment analysis and opinion mining of tweets have been done using the same [17]. Machine learning techniques can be extended to classify the fake reviews and fake news [18, 19]. The text classification is improved using the two-stage text feature selection algorithm [20]. Big data Hadoop framework is used to classify the product reviews based on aspects [21].

This research article proposed Twitter trend analysis using hashtag counting, noun counting, cosine similarity, Jaccard similarity, LDA, and K-means. These techniques are applied to static Twitter data as well as real-time streaming data and compared the results. The proposed work obtained better results in terms of execution speed for real-time Twitter trend analysis using SPARK tool.

The rest of the content is organized as follows. Section 3 discusses the proposed methodology. Section 4 gives the detailed results and analysis, and Section 5 highlights the conclusion and future scope in this research work.

## 3. Proposed Methodology

The proposed methodology includes the various steps, namely, collecting the static and real-time tweets from the Twitter and to perform the trend analysis. The proposed technique uses both static tweets and also real-time tweet trend analysis. Initially, the tweets need to be preprocessed for further analysis. Later, various machine learning techniques are applied on these static and real-time tweets to analyze the trends. Figure 1 depicts the proposed architecture for the real-time Twitter trend analysis.

This model is aimed at analyzing the trending topics in Twitter by using different approaches. Initially, the tweets are collected and preprocessed for further analysis. This preprocessed data is then analyzed using various methods like hashtag counting, noun counting, cosine similarity, Jaccard similarity, LDA, and K-means techniques. The performance of each algorithm is evaluated. The results are then analyzed to obtain the trending topic. We have also used SPARK framework to analyze real-time tweets. Using real-time streaming by SPARK, we have streamed tweets in real-time from Twitter and produced trending results faster. The various components involved in the proposed work are discussed in this section.

*3.1. Data Collection.* Tweets are collected using Twitter API. Tweets belonging to different domains like sports, health, economy, politics, and social, which were tweeted between January 15, 2021, and June 30, 2021, are collected. We have collected as many as 20,000 tweets. The dataset follows the JSON format. While streaming data through SPARK, we used a TCP socket as a data source to which tweets were written. SPARK will read and process the data from the socket.

*3.2. Data Preprocessing.* The collected tweets were stored locally in JSON file. This data is preprocessed by the following steps:

(1) Converting emoticons present in the tweets to text

(2) Removing hyperlinks (https/url) present in each tweet

(3) The tweets are made ready to be processed by removing punctuations and white spaces

(4) Removing stop-words

(5) *Performing Stemming*. Stemming is the process of removing suffixes in a word and retaining only the root word. For example, eating will become eat after stemming

(6) *Performing Lemmatization*. Lemmatization is similar to stemming where the output after the lemmatization process is called "lemma." In lemmatization, the reduced form of the words is found to be more meaningful when compared to the results of stemming

The preprocessed data is then fed as input to various algorithms to track the trending topics. The various techniques used were hashtag counting, noun counting, cosine similarity, Jaccard similarity, LDA, and K-means clustering.

*3.3. Hashtag Counting.* Hashtag counting is a primitive and simple method of predicting a trending topic. The collected dataset is subjected to counting, based on the number of times the hashtag appears in the dataset its trend value is set. Then, this value is used to get the top trends corresponding to the processed dataset. Here, the hashtag with the highest count is said to be a trending hashtag.

*3.4. Noun Counting.* In this method, the tweet contents are tagged with corresponding parts of speech. Tweet contents are categorized as nouns, verbs, adverbs, adjectives, and so on. Now, we detect the trend by counting the repeated nouns. The noun with the highest count is said to be trending.

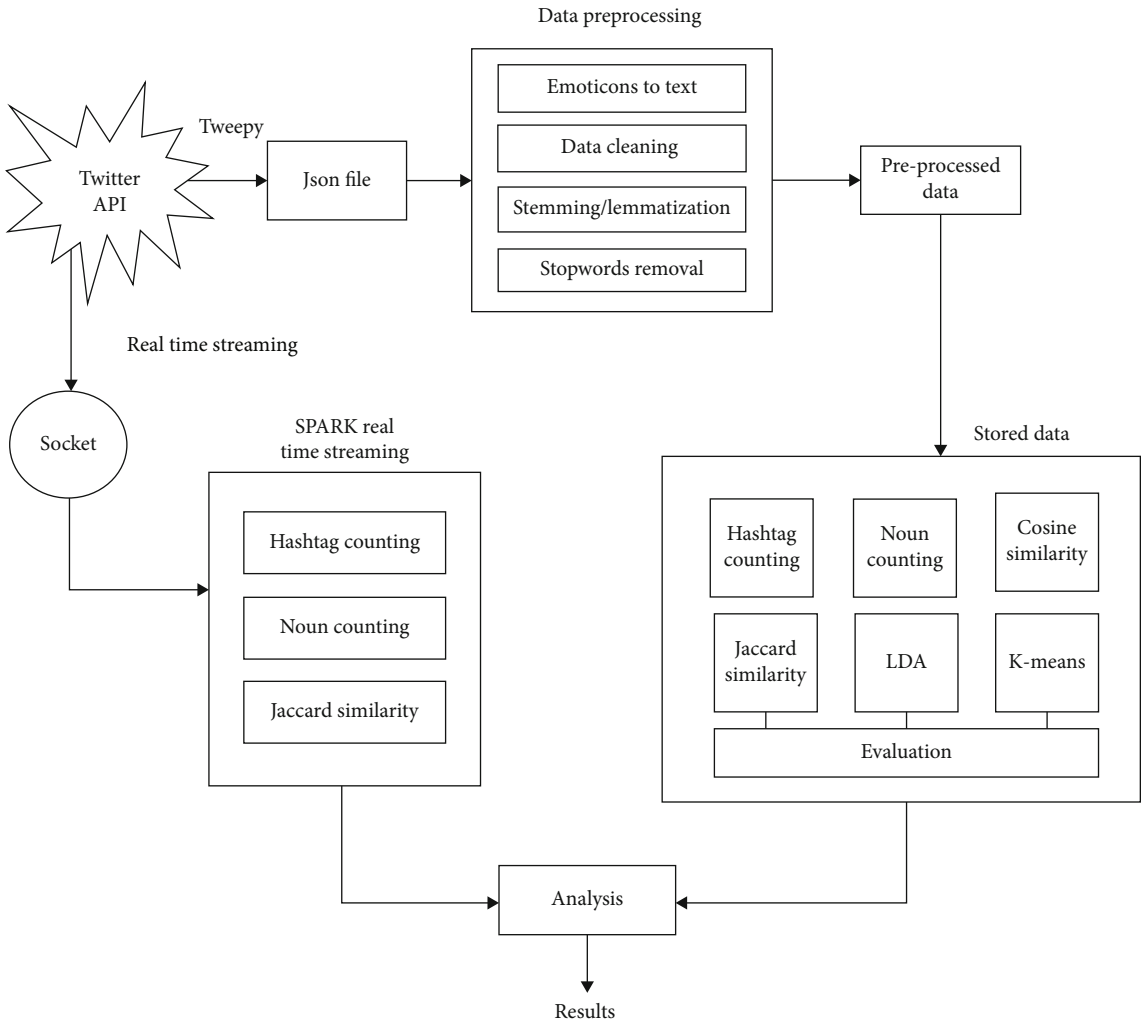



Figure 1: Proposed architecture for real-time tweet trend analysis.

*3.5. Clustering Using Latent Dirichlet Allocation (LDA).* Topic modeling techniques can be used to analyze Twitter trends based on the tweet text. The goal of this type of analysis is to find the different hidden topics in the dataset of tweets and then to determine the trending topic based on the number of tweets for each topic. LDA is one of the topic modeling algorithms specially designed for text data. This technique considers each document as a mixture of some of the topics that the algorithm produces as a final result. The topics are the probability distribution of the words that occur in the set of all the documents present in the dataset. For the Twitter trend analysis, the dataset can be considered as the set of documents where each document will be a tweet.

For example, consider the following three tweet texts:

(1) "What a champion. Simply the best. So calm, so sure in a run-chase. No big celebration, no theatrics, just a job finished. Superstar of the game."

(2) "IPL is postponed because of COVID. It is sad but safety is first."

(3) "Day 486 of lockdown, no effective vaccine rollout, restaurants are only doing takeaways and honestly this is all taking away my happiness."

The preprocessing of these tweet texts will give keywords as follows: ['champion', 'best', 'job', 'run-chase, 'celebration', 'superstar', 'game'], ['IPL', 'postpone', 'COVID', 'safety', 'first'] and, ['lockdown', 'vaccine', 'restaurant', 'takeaway', honest', 'happy'].

Each keyword array will be considered as a document, and LDA will try to find the hidden topics based on the probability distribution of keywords. We observe that the above tweet texts are related to sports and the COVID-19 pandemic. Initially, the algorithm will assign each word in the document to a random topic out of $n$ number of topics. As we already know theoretically, the above tweets consist of two topics; the algorithm may assign the first word that is "champion" for topic 2 (COVID-19). We know this assignment is wrong, but the algorithm will try to correct this in the future iteration based on two factors that are how often the topic occurs in the document and how often the word occurs in the topic. As there are not many COVID-19-

related terms in tweet 1 and the word "champion" will not occur many times in topic 2 (COVID-19), so the algorithm may assign the word "champion" to the new topic that is topic 1 (sports). With multiple such iterations, the algorithm will achieve stability in topic recognition and word distribution across the topics. Finally, each document can be represented as a mixture of determined topics; in the example, under consideration, tweet 1 is 100% topic 1, tweet 2 is 70% topic 1 and 30% topic 2, and tweet 3 is 100% topic 2. The number of topics and other tuning parameters can be altered to get better results in terms of clear topics.

*3.6. Trend Analysis Using Cosine Similarity.* Cosine similarity is a standard of measurement used to determine how much similar the records are regardless of their size. In terms of mathematics, it is a measurement of the cosine of the angle between two vectors plotted in multidimensional space. In this context, the two vectors are dictionaries (with the key being word and value being the count of that particular word) of those two documents. When we plot these two vectors in a multidimensional space, where each dimension corresponds to the keys in the dictionary (i.e., words in the document) and corresponding values represent how far is the point from that dimension, the cosine similarity calculates the angle between those two vectors, not the Euclidean distance. The cosine similarity metric is beneficial because even when two documents with the similar resemblance in word count but are far apart by the Euclidean distance because of the size (e.g., the word "baseball" occurred 100 times in the first document and 10 times in the second document), but they could have had a minor angle between them. The lesser the angle, the greater the similarity as we know the cosine of the angle increases as the angle decreases. Given two vectors $\vec{a}$ and $\vec{b}$, the angle between those two vectors is calculated by the equation (1) [22]:

$$\theta = \cos^{-1}\left(\left(\vec{a} \,.\, \vec{b}\right) / \left(\left|\left|\vec{a}\right|\right| * \left|\left|\vec{b}\right|\right|\right)\right) \cdots, \tag{1}$$

where

$$\vec{a} \,.\, \vec{b} = \sum_{1}^{n} (ai * bi). \tag{2}$$

On Twitter, hashtags are not mandatory for any tweet, so some tweets may not be having the hashtag attached. If we directly apply the hashtag counting algorithm for analysis on such data, the algorithm will simply ignore the tweets without hashtags thereby making the analysis inaccurate. Hence, in the proposed work, we divided the dataset into two sections as tweets with hashtags and tweets without any hashtags. Each tweet in the first section is stored as a document and labeled with the respective hashtag thereby creating a document for each available hashtag. Now for each tweet in the second section, we try to introduce the missing hashtag as one among the many available options in the stored document set based on the cosine similarity between the tweet under consideration and documents. This way we can make sure that each tweet will be attached with relatable hashtags and thereby considered by the hashtag counting technique.

*3.7. Trend Analysis Using Jaccard Similarity.* Jaccard similarity can be used to get the similarity coefficient of the tweet text and the predefined clusters and then can be classified based on the score obtained. Jaccard similarity algorithm works using the set intersection and union operations as shown in equation (3) [22].

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \cdots. \tag{3}$$

Consider a tweet text "Arranging ambulance, oxygen, and beds at the hospital during this COVID19 pandemic was not at all easy." Now after preprocessing the text, we get "Arranging ambulance oxygen bed hospital pandemic easy" as keywords. If we have two predefined clusters, namely, health and sports as "hospital ambulance doctors medicine COVID19 vaccine bed cough lungs pandemic oxygen pandemic" and "cricket match championship ipl football pandemic suspended umpire loss win," respectively, then we can represent all these as follows.

$$\begin{aligned} \text{Tweet}_{\text{text}} &= [0, 1, 2, 3, 4, 5, 6], \\ \text{Health}_{\text{related}_{\text{words}}} &= [4, 1, 7, 8, 9, 10, 3, 11, 12, 2, 5], \\ \text{Entertainment}_{\text{related}_{\text{words}}} &= [13, 14, 15, 16, 17, 5, 18, 19, 20, 21], \end{aligned} \tag{4}$$

where {'Arranging': 0, 'ambulance': 1, 'oxygen': 2, 'bed': 3, 'hospital': 4, 'pandemic': 5, 'easy': 6, 'doctors': 7, 'medicine': 8, 'COVID19': 9, 'vaccine': 10, 'cough': 11, 'lungs': 12, 'cricket': 13, 'match': 14, 'championship': 15, 'ipl': 16, 'football': 17, 'suspended': 18, 'umpire': 19, 'loss': 20, 'win': 21}.

For the given text intersection with health, $n$ sports will be {1, 2, 3, 4, 5} and {5}, respectively. Hence, the Jaccard coefficient for health and sports will be 5/18 and 1/17. As the score of similarity for the health cluster is higher, the tweet will be classified as health related to tweet.

*3.8. Trend Analysis Using K-Means Clustering.* Interests of Twitter users vary from user to user; some may tweet more about social events, some are much into politics, and some tweet more about sports. Twitter users' behavior or interests and in turn likes and dislikes can be analyzed based on the number of tweets they tweet on different various topics. By using the results of Jaccard similarity, we can cluster Twitter users into multiple categories with the help of K-means clustering. K-means algorithm attempts to cluster the given dataset into $k$ number of nonoverlapping groups, such that every data point in the dataset belongs to a unique cluster. Cluster formation is done such that maximum similarity is maintained within a cluster, and different clusters are as far as possible from each other. Euclidian distance is used to achieve the clustering goal. For example, if we consider some tweets for four Twitter accounts related to the health

and sports category as [1], [1, 2], [3, 4], and [4, 5], respectively, where [a,b] represents *a* number of health-related tweets and *b* number of sports-related tweets from a user. K-means algorithm follows below simple steps in a loop until it meets converging conditions.

(1) Find the coordinates of the centroid

(2) Calculate the distance of each object to the centroid

(3) Assign the objects to a cluster based on minimum distance

*3.9. Real-Time Streaming Using SPARK.* A good analysis needs a large amount of data. The more is the data, the much better will be the analysis. It is very important to cover a large volume of tweets across the globe on various topics from different people to get accurate trends while analyzing Twitter trends. Thankfully, Twitter provides all the support we need to get tweets for such analysis. But if we choose to write a program for collecting the tweets and then preprocess, store, and finally apply algorithms on the stored data to find out the trends, a lot of time and resources will be wasted when we can do the same task with the help of real-time streaming and SPARK. Thanks to Twitter again, which will support streaming the Twitter data. A TCP socket in a system will be used instead of a file to hold the incoming Twitter stream. If a SPARK session is connected to this same TCP socket, it will read incoming data as soon as it will be written to the socket. This powerful combination can be used to enhance the results of the algorithms that have mentioned earlier in the paper. The advantage here is that the model will not wait until we are done collecting the required amount of tweets. Every time Twitter data is written to the socket, SPARK will immediately start processing it. With the structured streaming support in SPARK, the result will get updated as the incoming data get processed. So even though SPARK produces results in batches, every batch will be having the result corresponding to the data streamed until that point in time.

In our analysis, we have used a TCP socket as a data source. Tweets are collected on real-time basis using Twitter API and tweepy writes it to this data source. Pspark processes tweets in batches. The program runs on SPARK for the specified time interval where in tweets are streamed in batches and output is also obtained in batches. Depending on what is being tweeted about the most at a given time, the numbers will keep changing in every batch. In this way, we were able to stream data continuously on a real-time basis. Compared to static data, larger number of tweets could be collected, and it also provides accurate trends at any given time of a day.

## 4. Results and Discussion

The experimental results of different techniques and algorithms used for trend analysis provide us insights on which method is best suited for real-time analysis and gives accurate results. The outputs of these techniques have been presented and analyzed in the form of graphs, and tables and a close match have been found between the different results obtained. We have performed trend analysis on static and dynamic data. For static analysis, data is collected beforehand and stored in a file. Basic counting methods and machine learning algorithms are applied to this stored data to identify the trends. In the case of dynamic analysis, data is streamed on a real-time basis, and analysis is performed at the same time by using SPARK structured streaming. This has allowed us to process a large number of tweets and obtain accurate trends. For the experimental analysis, we have used a sample dataset having 20000 static tweets. For real-time trend analysis, we have extracted live tweets.

*4.1. Static Twitter Data Analysis.* The first and the basic method we have used to predict what is being talked about a lot on Twitter is counting the hashtags. Hashtags being the key elements of tweets are used widely by people to express their opinions and as supporting elements of the tweet content. For the experimental analysis, we have used a sample dataset having 20000 tweets where in "#COVID19" has been found to be used the highest number of times, which is 87.

Hashtag counting does not consider the actual tweet content to predict the trend. To overcome this drawback, we have used the noun counting method which identifies the nouns in all the tweets and hence tells us which nouns have been used repeatedly. At first, we have used part-of-speech tagging to tag the words in the tweets with their corresponding part of speech. Next, the words tagged as "nouns" were collected and counted to find the noun that has been used frequently.

As seen in the result, the word "vaccine" is found to be used the highest number of times. The results obtained using the two counting techniques are related and strongly comply with the actual scenario too. Hence, we decided to use these two techniques for real-time analysis as well. Results of Twitter trends using hashtag counting and noun counting are shown in Figures 2 and 3, respectively.

A sample dataset with 6000 tweets has been used to get the results of the designed LDA model. The preprocessed dataset stored in a data frame was given as input to the model. The execution of the algorithm for a different number of topics ($k$) produced different results. To find an optimal number of topics for the given dataset, the coherence value has been calculated for each value of $k$. This measure helps us to figure out how coherent the topics are, in other words how well the recognized topics support each other. Figure 4 shows the coherence value for different $k$ values.

Choosing the right $k$ value is not straightforward always, and there is no such standard way to do that. Either we can manually try to tune the $k$ value based on the topic interpretation or we can consider the $k$ with a larger coherence value. In the above sample, $k = 3$ can be taken as an optimal number of topics and further can be improvised by modifying the other parameters such as alpha, beta, or even number of iterations. The following is the result of LDA after tuning $k$ values.

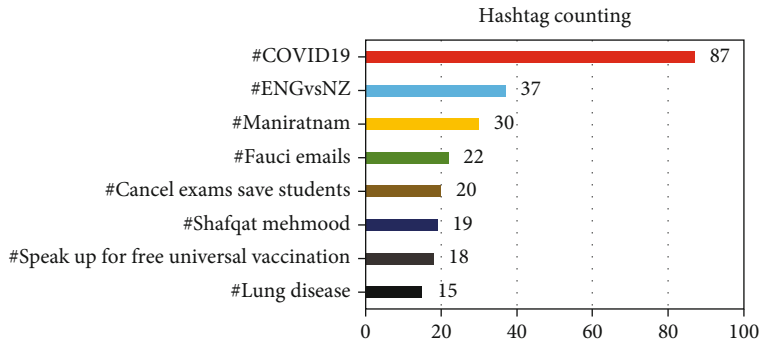



Figure 2: Twitter trends using hashtag counting.

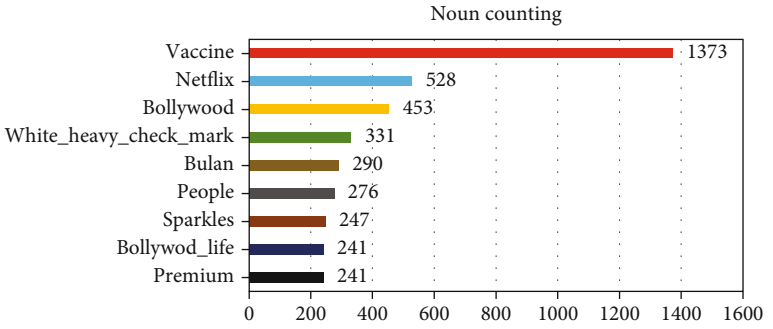



Figure 3: Twitter trends using noun counting.

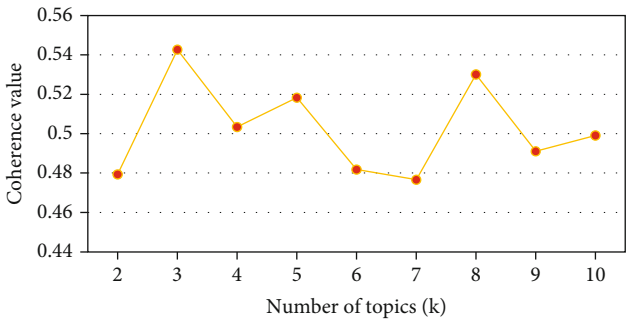



Figure 4: Number of topics vs. coherence value.

[(0,
'0.056∗"vaccine" +0.020∗"month" +0.016∗"covid" +0.010∗"netflix" +0.009∗"hold" +0.008∗"day" +0.008∗"solo" +0.007∗"php" +0.007∗"amp" +0.006∗"people"'),
(1,
'0.048∗"bollywood" +0.013∗"movie" +0.009∗"netflix" +0.006∗"amp" +0.006∗"film" +0.005∗"time" +0.005∗"actor" +0.005∗"good" +0.004∗"song" +0.003∗"sushant"'),
(2,
'0.039∗"netflix" +0.014∗"bulan" +0.012∗"premium" +0.012∗"jual" +0.010∗"spotify" +0.009∗"viu" +0.008∗"canva" +0.008∗"legal" +0.008∗"youtube" +0.007∗"garansi"')]

The output contains 3 topics with id topics 0, 1, and 2, a close look at these clustered topics can give some insights on what that topic represents. In the above case, we can say that topic 0 is the COVID-19 pandemic, topic 1 is Bollywood, and topic 2 is Netflix. Table 1 gives the topic assignment for each tweet in the dataset.

Table 1: Topic assignment for the tweets in the data set.

| Tweet number | Topic distribution |
|---|---|
| tweet2186 | 99% Bollywood |
| tweet1782 | 98% Netflix |
| tweet2640 | 99% Bollywood |
| tweet5518 | 99% COVID-19 |
| tweet1718 | 26% Bollywood, 73% Netflix |
| tweet5725 | 99% COVID-19 |
| tweet1051 | 99% Bollywood |
| tweet936 | 99% COVID-19 |
| tweet2522 | 99% Bollywood |
| tweet2679 | 77% COVID-19, 22% Netflix |

If we approximate the topic distribution by assigning the most probable topic in the distribution for each tweet, we get 2307 tweets related to COVID-19, 2100 tweets about Bollywood, and 1593 tweets on topic Netflix. Figure 5 shows the analysis.

Next, we used cosine similarity to group the tweets into documents based on different topics and then measured the cosine of the angle between the documents by considering them as vectors. First, it creates documents that have hashtags and then label that particular document with that hashtag. For the tweets which have no hashtag, it will

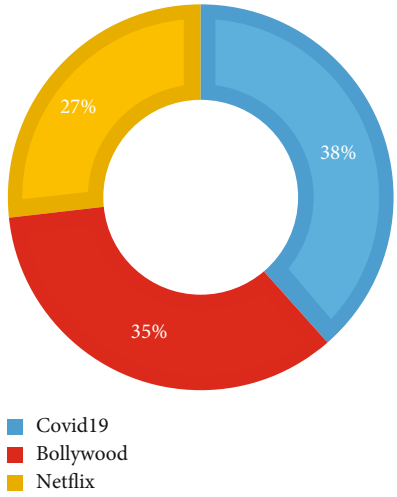



Figure 5: Trend analysis using LDA.

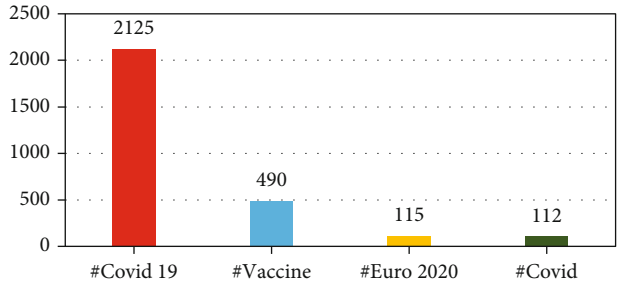



Figure 6: Top trends using cosine similarity.

calculate cosine similarity with TF-IDF to all the documents that were created earlier, and one with the highest similarity is labeled with the hashtag, and the count of that hashtag is also incremented. Here, #covid19 is trending in the generated documentation for the dataset that we have collected. This is depicted in Figure 6. When compared with Figure 2, the following graph in Figure 6 shows the huge rise in the hashtag counts after the usage of cosine similarity to generate hashtags for the tweets without any tags attached. The accuracy of this model in terms of introducing relatable hashtags at the missing value is calculated to 0.7397 which is approximately 74%.

The collected Twitter data has been classified by the model, designed using the Jaccard similarity classification algorithm. It shows that health-related tweets are more in number when compared to other categories with 60% of the collected tweets being health tweets. This is depicted in Figure 7.

The performance of the model was determined by the accuracy and Jaccard score. To find the accuracy, a dataset is prepared with tweets including their actual category which is assigned manually. Later, tweets in the dataset were given as input to the model, and the predictions done by the model are compared against the actual results stored in the dataset.

The confusion matrix for the above results is shown in Table 2.

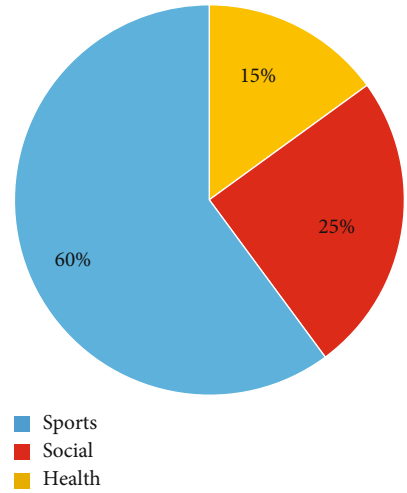



Figure 7: Trend analysis using Jaccard similarity.

Table 2: Confusion matrix for Jaccard similarity.

| *Expected\predicted* | Social | Health | Sports |
|---|---|---|---|
| Social | 112 | 20 | 33 |
| Health | 16 | 283 | 6 |
| Sports | 25 | 21 | 153 |

Based on the above results, the model accuracy is calculated to 0.8316 approximately 83%. Similarly, the Jaccard score will be 0.7117(average = "micro") and 0.68551 (average = "macro").

*4.2. Trend Analysis Using K-Means Clustering.* The model has been designed in such a way that it can group users into various categories based on the results of Jaccard similarity. For the sample data set, four categories were predefined, namely, economy, health, social, and culture. At first, each tweet in the dataset was classified using Jaccard similarity, and the number of tweets tweeted by each user id was calculated.

If we cluster users based on their interest in the social and economic sectors, we get the result that shows that a group of users show limited involvement in the social sector with less than 100 tweets and not much interest in the economic sector either. There is one more group of people who tweets on both sectors but show high interest in the social sector when compared to the economy. This is shown in Figure 8.

Silhouette outline can be adapted to fix the degree of separation among clusters. The optimal number of clusters will be decided based on the silhouette coefficient.

$$\text{Silhouette coefficient} = \frac{b^i - a^i}{\max\left(b^i, a^i\right)} \cdots, \tag{5}$$

where $a^i$ is the average length from all score points in the corresponding cluster and $b^i$ is the average length from all

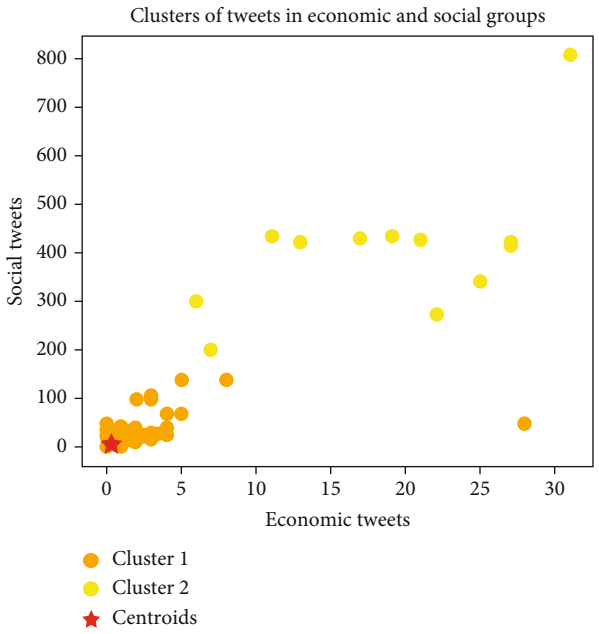



Figure 8: Results of clustering after applying on economic and social tweets.

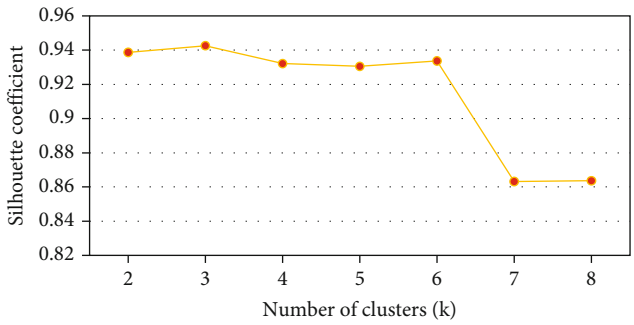



Figure 9: Number of clusters vs. silhouette coefficient.

score points in the nearest cluster. The coefficient can get values within the interval [-1, 1]. If it is 0, the unit is very close to the nearby clusters. If it is 1, the unit is notably apart from the nearby clusters. If it is -1, the unit is attached to the incorrect clusters. Hence, we look for the $k$ value with a higher coefficient value. For the dataset under consideration, we can say that $k = 2$ or $k = 3$ is not a bad choice which has a silhouette coefficient of 0.938 and 0.942, respectively. But for $k = 7$ or $k = 8$, we can observe that there is a decrease in the coefficient value. This is depicted in Figure 9.

An alternative method for Silhouette analysis is the elbow method. The elbow method helps in deciding a good match of $k$ value for a given dataset as per the sum of squared distance between data points and corresponding clusters' centroids. The optimal $k$ value is the spot where SSE forms an elbow and starts to flatten out. This is shown in Figure 10.

The above graph shows the formation of the elbow at $k = 2$ and $k = 3$. $K = 2$ is chosen as the optimum number for clusters of the given dataset as the SSE curve forms elbow at that point.

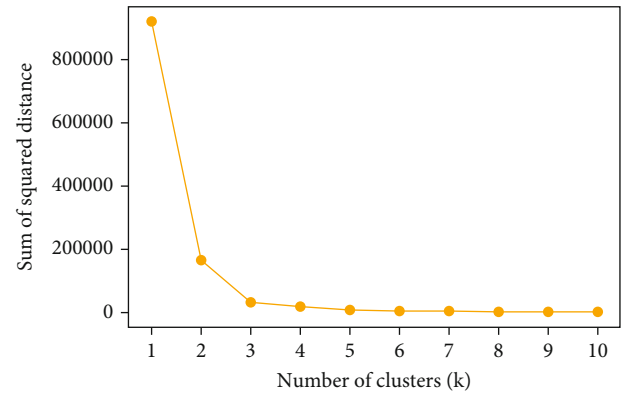



Figure 10: Finding the best value of $k$ using the elbow technique.

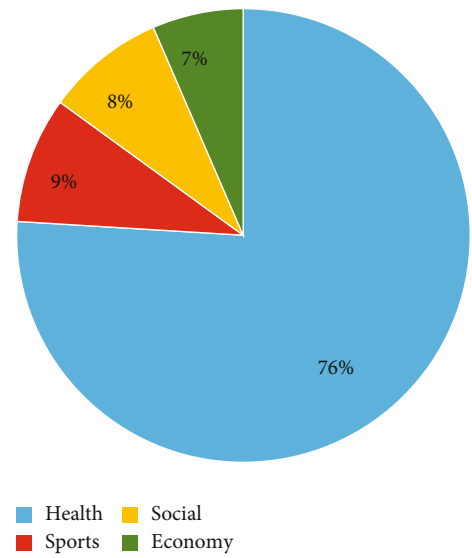



Figure 11: Real-time tweet trend distribution using Jaccard similarity.

*4.3. Real-Time Twitter Data Analysis Using SPARK.* To process a large number of tweets in a fast manner, we have used SPARK streaming. SPARK is a big data tool that enables fault tolerance parallelism in the data processing. The results obtained using SPARK were rightly matching with the real-time Twitter trends. We have applied hashtag counting to find popular hashtags, noun counting to obtain the most prominent words in the tweets, and Jaccard similarity to group the tweets into different categories like health, economy, sports, and social. We collected tweets for one month (May 2021) to analyze and compare the output of different techniques. By using Jaccard similarity, we were able to group the tweets into different categories and obtain a pie chart for the distribution shown in Figure 11.

Figures 12–14 depict the real-time sports, health, and social trends using the hashtag counting technique. Figures 15–17 depict the real-time sports, health, and social trends using the noun counting technique.

The above graph shows that for each separate category obtained in Jaccard similarity, similar words are found to be trending when both hashtag counting and noun counting

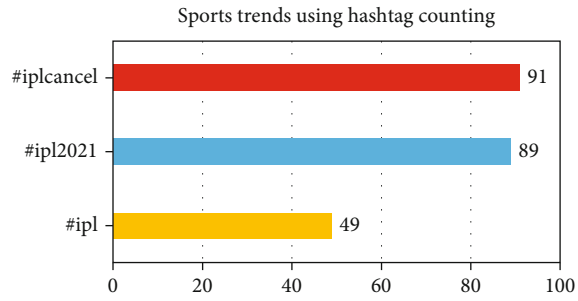



Figure 12: Real-time sports trends using hashtag counting.

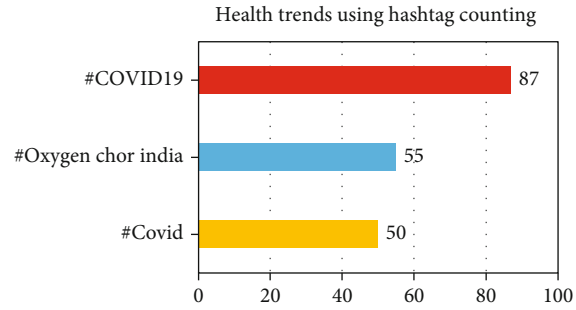



Figure 13: Real-time health trends using hashtag counting.

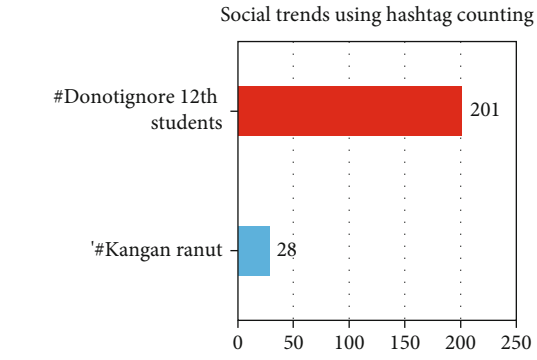



Figure 14: Real-time social trends using hashtag counting.

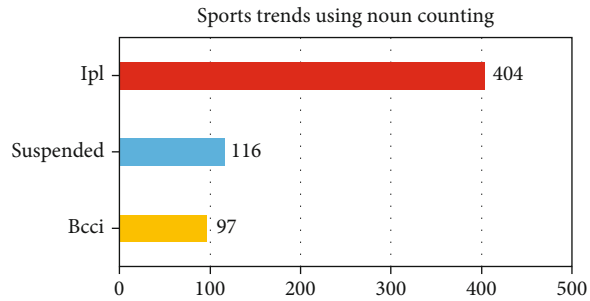



Figure 15: Real-time sports trends using noun counting.

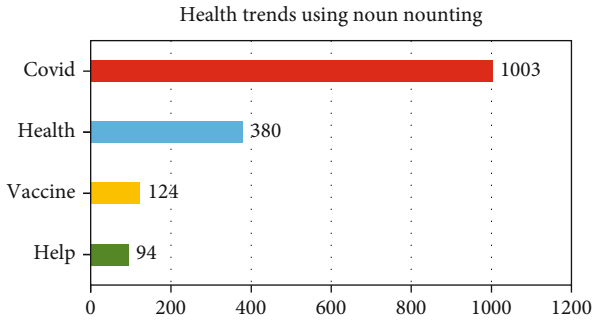



Figure 16: Real-time health trends using noun counting.

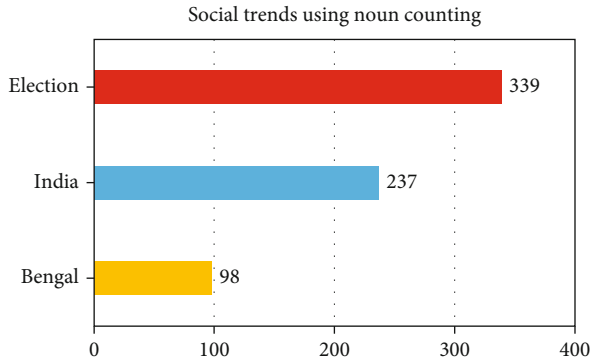



Figure 17: Real-time social trends using noun counting.

methods are applied. We can track the counts for certain handpicked nouns using noun counting technique for a particular interval of time. Figure 18 depicts the trend plot for few handpicked nouns based on the Twitter activities for 5 days of interval in June of 2021.

Similarly, we can even track the trends for few selected hashtags as well using hashtag counting techniques. Figure 19 depicts the trend chart of some of the popular hashtags for 3 days period in June of 2021.

Figure 20 shows the variation in the volume of real-time tweets related to health, economy, and social for five days session in June of 2021 using the Jaccard similarity method. Table 3 gives the comparison of Twitter trend analysis using SPARK and without using SPARK in terms of execution time required for the hashtag counting method (in seconds).

Figure 21 shows the execution time comparison between two cases with and without using SPARK for real-time and stored tweet trend analysis using hashtag counting technique, respectively.

In the graph shown in Figure 21, we can see that for a smaller number of tweets, SPARK is taking relatively higher time but as the number of tweets increases, we can see the difference and need for using SPARK. With real-time streaming and using SPARK as we can generate the results in batches, the response time will be much better compared to the program without using stream and SPARK. Table 4 gives the response time in seconds while applying different analysis techniques on the stored datasets without using SPARK tool.

Figure 22 depicts the comparative execution time in seconds, for the real-time trend analysis by streaming with

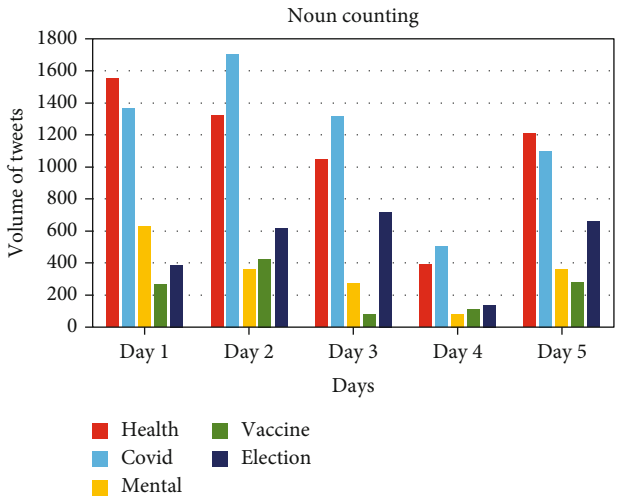



Figure 18: Real-time Twitter trend analysis for 5 days in June using noun counting.

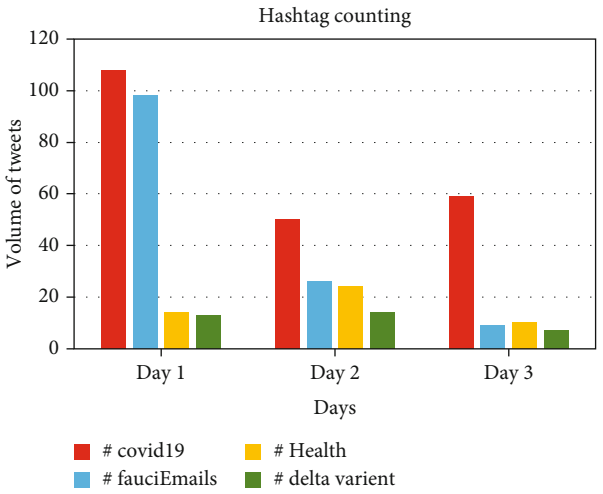



Figure 19: Real-time Twitter trend analysis for 3 days in June using hashtag counting.

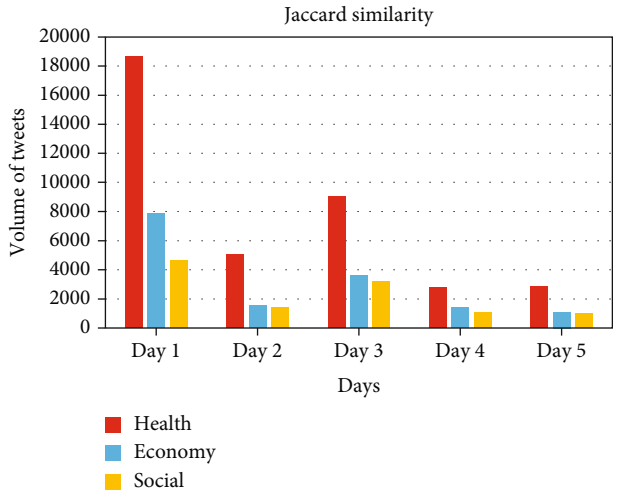



Figure 20: Real-time Twitter trend analysis for 5 days using Jaccard similarity.

Table 3: Comparison of execution time (in seconds) with and without using SPARK for hashtag counting technique.

| Number of tweets processed | Total execution time using SPARK | Total execution time without using SPARK |
|---|---|---|
| 457 | 51.49812531 | 28.36422133 |
| 1075 | 101.5530577 | 66.67746949 |
| 1509 | 151.5708518 | 432.3796902 |
| 1955 | 201.9629259 | 651.3878441 |
| 2316 | 231.7777178 | 806.4659975 |
| 2615 | 271.9080777 | 817.5746803 |
| 4814 | 302.536587 | 1656.805031 |

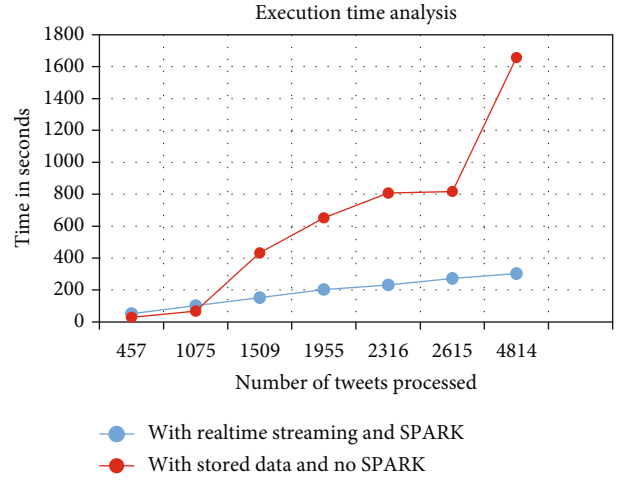



Figure 21: Execution time analysis.

Table 4: Response time (in seconds) without using SPARK.

| Number of tweets in the dataset | Response time in hashtag counting | Response time in noun counting | Response time in Jaccard similarity |
|---|---|---|---|
| 10 | 1.222835064 | 2.911193 | 2.761557 |
| 30 | 2.047077417 | 3.765497 | 3.606873 |
| 50 | 4.008028526 | 5.937976 | 5.719659 |
| 100 | 6.942538977 | 8.9464 | 8.678112 |
| 500 | 33.71384072 | 36.51038 | 35.83625 |
| 1000 | 66.13316321 | 70.19669 | 68.6782 |
| 1200 | 82.5823097 | 86.82858 | 85.56904 |

SPARK and hashtag counting, noun counting, and Jaccard similarity without using SPARK.

The graph in Figure 22 shows that SPARK will take around 40 seconds to start the response to the stream and produces the first batch, and then, it keeps on updating the result as the input stream keeps coming. The constant time shown here depends on the system specification on which we are running the program and also the speed of the internet connection to the system because of the streaming of tweets from Twitter. This response time can still be

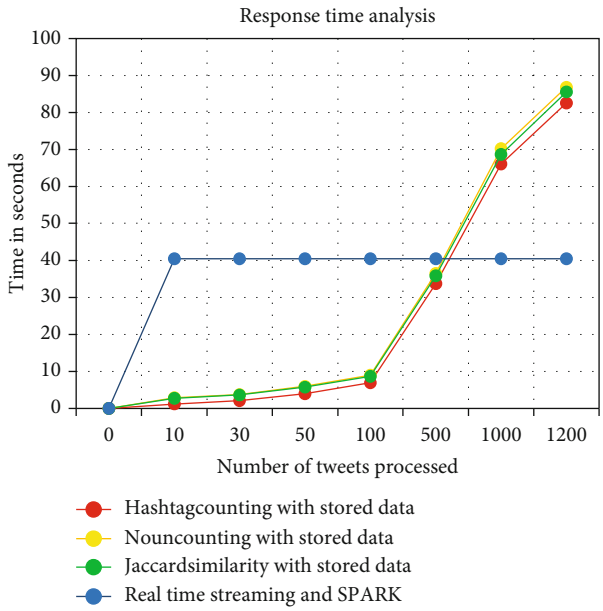



Figure 22: Comparative analysis of the execution time.

decreased using machines with powerful processors. Without using SPARK, the response will be the result itself hence have to wait until the program executes completely which is not preferable always as Twitter is a platform where so many people will tweet on so many topics in very little time.

*4.3.1. API Used.* We have used Twitter API to get access to the Twitter data, using Twitter API we can programmatically retrieve the data. In order to get access to the Twitter API, you will need to follow the following steps.

*Step 1.* Apply for twitter developer account and wait until we receive approval. Generally, Twitter provides two levels of access: one is "Standard," and another one is for "Academic research." The proposed work chosen academic research option.

*Step 2.* Once our account is approved, we will be able to generate or find the twitter API access credentials which are discussed below:

*API key*. This is basically a user name that allows you to make request to the Twitter to get access for the data.

*API key secret*. This is the password for your API key.

*Access token*. This token represents the associated Twitter account.

*Access token secret*. This also represents the associated Twitter account.

*Bearer token*. This token represents the application for which you are using the Twitter data.

Since we are building application in Python, its package manager pip provides a library called "tweepy" which is used to connect programmatically and get access to the Twitter API using the credentials that we will get in Step 2 and then download or stream data in real time.

## 5. Conclusion

Twitter is one of the major platforms with a large number of users worldwide. People, their interest, their opinion, likes, dislikes, events, sports tournaments, politics, movies, and the music everything are part of it. Analyzing such a rich data content platform and observing trends in it definitely will be beneficial. Analyzing Twitter trends helps to know what people are more interested in and thus helps business organizations or brands to improve their sales, political parties to understand people's emotions and needs, movie industries to get valid feedback for their performances, and much more. In this article, we have proposed some of the possible techniques that can be used to analyze Twitter trends from brute force counting techniques to topic modelling and machine learning clustering techniques. Choice of the technique depends on the purpose of analysis, the amount of data expected to be covered in the analysis model, and even the expected output formats. As everyone expects the model to be run faster and smoother, we also have proposed model development using real-time streaming and SPARK which is a big data analytics tool. The LDA technique for trend analysis resulted in an accuracy of 74% and Jaccard with an accuracy of 83% for static data. The results proved that the real-time tweets are analyzed comparatively faster in the Big Data Apache SPARK tool than in the normal execution environment.

## 6. Future Work

As future work, the proposed models can be modified to develop a trend analysis system that tracks the trends in a particular geographical location. This will help business organizations to target the right people in the right places to build their brand values. As the application and demand for Twitter trend analysis are always rising, the proposed techniques with few modifications can be used to fit most of the requirements.

## Data Availability

The JSON data used to support the findings of this study are included within the supplementary information file.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Supplementary Materials

The supplementary file consists of 20,000 tweets collected from Twitter for the experimental analysis. The file is in the text format, and the tweets are in the JSON format. The real-time analysis is performed on live streaming of tweets and hence not stored offline. This data set may be used by the researchers to further carryout more experiments and obtain better results. *(Supplementary Materials)*

## References


[1] M. Mashuri, "Sentiment analysis in twitter using lexicon based and polarity multiplication," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, pp. 365–368, IEEE, 2019.

[2] M. Wongkar and A. Angdresey, "Sentiment analysis using naive Bayes algorithm of the data crawler: Twitter," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, pp. 1–5, IEEE, 2019.

[3] R. Sharma, *Twitter Sentiment Analysis*, 2019, https://github.com/sharmaroshan/Twitter-Sentiment-Analysis.

[4] S. Yang and H. Zhang, "Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis," *International Journal of Computer and Information Engineering*, vol. 12, no. 7, pp. 525–529, 2018.

[5] E. S. Negara, D. Triadi, and R. Andryani, "Topic modelling Twitter data with latent Dirichlet allocation method," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pp. 386–390, IEEE, 2019.

[6] N. Shahreen, M. Subhani, and M. M. Rahman, "Suicidal trend analysis of Twitter using machine learning and neural network," in *2018 international conference on Bangla speech and language processing (ICBSLP)*, pp. 1–5, IEEE, 2018.

[7] A. F. Hidayatullah and M. R. Ma'arif, "Road traffic topic modeling on Twitter using latent Dirichlet allocation," in *2017 international conference on sustainable information engineering and technology (SIET)*, pp. 47–52, IEEE, 2017.

[8] R. A. Hasan, R. A. I. Alhayali, N. D. Zaki, and A. H. Ali, "An adaptive clustering and classification algorithm for twitter data streaming in Apache Spark," *Telkomnika*, vol. 17, no. 6, pp. 3086–3099, 2019.

[9] K. Garg and D. Kaur, "Sentiment analysis on Twitter data using Apache Hadoop and performance evaluation on Hadoop MapReduce and Apache Spark," *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, , pp. 233–238, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2019.

[10] S. E. Saad and J. Yang, "Twitter sentiment analysis based on ordinal regression," *IEEE Access*, vol. 7, pp. 163677–163685, 2019.

[11] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine learning-based sentiment analysis for twitter accounts," *Mathematical and Computational Applications*, vol. 23, no. 1, p. 11, 2018.

[12] M. R. Huq, A. Ali, and A. Rahman, "Sentiment analysis on Twitter data using KNN and SVM," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, pp. 19–25, 2017.

[13] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep convolution neural networks for Twitter sentiment analysis," *IEEE Access*, vol. 6, pp. 23253–23260, 2018.

[14] A. Z. Ahmed and M. Rodríguez-Díaz, "Significant labels in sentiment analysis of online customer reviews of airlines," *Sustainability*, vol. 12, no. 20, pp. 1–18, 2020.

[15] T. Rathod and M. Barot, "Trend analysis on Twitter for predicting public opinion on ongoing events," *International Journal of Computing Applications*, vol. 180, no. 26, pp. 13–17, 2018.

[16] P. Garg, R. Johari, H. Kumar, and R. Bhatia, "Trending pattern analysis of Twitter using spark streaming," in *International Conference on Application of Computing and Communication Technologies*, pp. 3–13, Springer, Singapore, 2018.

[17] N. D. Zaki, N. Y. Hashim, Y. M. Mohialden, M. A. Mohammed, T. Sutikno, and A. H. Ali, "A real-time big data sentiment analysis for Iraqi tweets using spark streaming," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 4, pp. 1411–1419, 2020.

[18] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems*, vol. 117, pp. 47–58, 2021.

[19] H. Khan, M. U. Asghar, M. Z. Asghar, G. Srivastava, P. K. R. Maddikunta, and T. R. Gadekallu, "Fake review classification using supervised machine learning," *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, , pp. 269–288, Springer International Publishing, 2021.

[20] G. SRIVASTAVA, P. K. R. MADDIKUNTA, and T. R. GADEKALLU, *A Two-Stage Text Feature Selection Algorithm for Improving Text Classification*, ACM Transactions on Asian and Low-Resource Language Information Processing, 2021.

[21] A. P. Rodrigues, N. N. Chiplunkar, and R. Fernandes, "Aspect-based classification of product reviews using Hadoop framework," *Cogent Engineering*, vol. 7, no. 1, p. 1810862, 2020.

[22] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *International conference on intelligent data engineering and automated learning*, pp. 611–618, Springer, Berlin, Heidelberg, 2013.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 1709363, 11 pages
https://doi.org/10.1155/2021/1709363

*Research Article*

# A Robust Optimization Modeling for Mine Supply Chain Planning under the Big Data

**Wenbo Liu**

*Department of Logistics Management, Liaoning Provincial College of Communications, Shenyang 110122, China*

Correspondence should be addressed to Wenbo Liu; wenbo-315@163.com

Received 12 August 2021; Accepted 27 September 2021; Published 23 October 2021

Academic Editor: Rajesh Kaluri



With the rapid development of information technology, large-scale data is collected and stored, which provides a huge amount of information for decision-making. This paper focuses on the planning of mine supply chain under the big data. The mine supply chain usually contains three stages, which is mining, processing, and ore product transportation. This paper tackles the difficulty of variable cut-off grade by establishing a robust optimization model. To solve the robust optimization model, the nonlinear constraints in the model were linearized first. Then, the specific parameter values were determined through the employment of the hypothesis test in statistics, and the robust optimization model was solved finally. The analysis results show that the robust optimization model can be stabilized when the parameters are subject to disturbance. Finally, sensitivity analysis experiments are carried out for several parameters in the model to find out the influence of each parameter on the model. This paper combines mine supply chain planning with big data, which not only improves the production and transportation efficiency of ore products, but also reduces related costs.

## 1. Introduction

The steel industry is an important basic industry sector, which is the material base of developing national economy and national defense construction. It consumes large amount of iron ores in daily production and thus tends to maintain a high level of inventory. Meanwhile, the mining companies, which provide the iron ores suffer from the uncertainty of ore grade in mining and the market fluctuations in transportation. Based on this, it is necessary to establish a long-term, stable, and safe mine supply chain.

The mine supply chain under big data involves many links and contains a large number of relevant data or parameters. The emergence of big data has an unmistakable impact on the amount and speed of data processed in supply chains [1–3]. In the mine supply chain, the raw ores are extracted from multiple mining locations first, and each of the location might have different ore grade. Then, the raw ores are mixed to meet the ore grade requirements and then transport to the concentrating mills for processing. Finally, the ore products are transported through railway or sea transportation from the concentrating mills to the logistics centers. The demands of the final users are met by the accurate services of the logistics centers. During the whole mine supply chain, the ore grade is a key factor involved in the decision-making. For each mining location, based on the estimation of the natural ore grade, the manager must determine a cut-off grade to optimize the quality and quantity of the mined ore. Generally speaking, the higher the cut-off grade, the higher the ore quality while the lower the ore quantity, since the iron ore blocks with a natural ore grade lower than the cut-off grade is dumped as waste. The concentrating mill has a minimum ore grade for the raw ore, and the decision of the cut-off grades in multiple ore locations will be subject to the mining of the iron ore, as well as the corresponding processing, inventory, and transportation largely. In this paper, the variable cut-off grade, which varies in different time periods is considered, for the purpose of promoting the flexibility of the whole ore supply chain. Most of the previous research attaches importance to the solution to a single component of the mine supply chain, e.g., mining, processing, production, warehousing, or transportation. It is of great significance to study the whole mine supply chain, which is composed of the entire process of production, processing,

inventory, and transportation. Mine supply chain planning is a data-driven process that focuses on developing plans to operate the supply chain efficiently and optimizing outcomes under given constraints [4]. Big data analysis tools such as optimization algorithms can help the supply chain to balance resource and to determine reasonable production planning capabilities, inventory levels, distribution capabilities, etc. [5, 6].

We followed the modeling methods of Liu et al. [7] and built a robust optimization model based on the important parameter of the cut-off grade. It was the logic of this paper that firstly determined the range of the parameter by the statistical analysis for the ore grade parameters through the employment of actual observation and exploration results, and then helped obtain the value of the relevant parameters by using hypothesis testing. Hypothesis testing with statistical significance is one of the most important methods of big data analysis. In addition, the objective function values obtained by the robust optimization model were compared with each other by carrying out numerical experiments, and then the verification for the stability and optimality of the robust optimization model was performed. As for robust optimization, it was able to address the problem of data uncertainty by guaranteeing the feasibility and optimality of the solution for the worst instances of the parameters. Through a number of numerical experiments, it was shown by the experimental results that the proposed robust optimization model was very stable under the circumstance of parameter perturbations.

This paper is aimed at coordinating mining, inventory, and transportation between all the aspects of the mine supply chain under the premise of considering relevant constraints and minimizing the total costs on the mining, blending, processing, inventory, and transportation of the supply chain. The main contribution of this paper is to integrate the key node enterprises in the mine supply chain, the optimization of its production and logistics systems, reduce or eliminate the phenomenon of unbalanced production, avoid the fluctuation of upstream and downstream, make the ore mining, processing, and transportation orderly, and realize the balanced development of supply and demand, so as to improve the overall benefit and efficiency of the mine supply chain. The research objects of this paper include monomer ore and other multimetal.

This paper is organized as follows: In Section 2, we review relevant literature. Section 3 describes our problem and constructs a robust optimization model. In Section 4, performance analysis and sensitivity analysis experiments are carried out for the robust optimization model. It is verified by the numerical experiments that the robust optimization model is able to meet the actual needs and obtain ideal results. Section 5 concludes the paper.

## 2. Literature Review

Ghiani et al. [8] defined the supply chain as "a complex logistics system in which raw materials are converted into finished products and then distributed to final users (consumers or companies)." The competitive advantage of supply chain management in various industries is realized though supply chain planning [9]. The supply chain planning involves many functional areas of procurement, production and distribution, and across strategic network planning, production planning and scheduling, purchasing and material requirement planning, and distribution and transport planning [6, 10]. Brunaud and Grossmann [11] put forward the statement that some researchers carried out relevant studies on the supply chain modeling and optimization issues in a variety of industries. Nishi et al. [12] proposed a framework for distributed optimization of supply chain planning and coordination approach. Steinrücke [13] studied the practical problems of the global aluminum for supply chain network. Members of the aluminum for supply chain network are scattered around the world. They constructed a novel type of mixed-integer decision-making model, which can coordinate the production quantities and times of all supply chain members, in order to minimize the production and transportation cost of the whole supply chain. Vintró et al. [14] and Söderholm et al. [15] focused their research on green supply chain, such as the issues related to the society, environment, health, and safety. Apart from that, Kusi-Sarpong et al. [16] fixed their attention on the studies carried out for a framework and evaluation of Ghana's mining green supply chain practices. Azapagic [17] devoted himself to the development of a framework for the sustainable development of the mining and mining industry, in which economic, environmental, social, and other comprehensive indicators for the mining industry stakeholders are included.

Big data analysis is highly relevant to supply chain planning [5]. Optimization techniques can provide fundamental support to demand planning, production planning, inventory plans, and logistics planning by improving planning accuracy and flexibility [18–20]. The big data is collected from a wide range of diversified sources with various perspectives and data formats (i.e., variety) [21]. The digitization of supply chains [22] for better tracking of supply chains further highlights the role of big data analytics. He et al. [23] carried out some researches in the fields of big data analysis, which used dimensionality reduction algorithm of big data to quickly extract valuable parts from a huge amount of data and improve computational efficiency.

The production and distribution logistics planning in the open pit mine is known for its own unique features of the mining industry. It is the uncertainty available in the quality requirements of ore iron concentrates that challenges the choice about which grade of ore to mine. Newman et al. [24] review several decades of literature, including mine design, long-term and short-term production scheduling, equipment selection, and dispatching. Chen and Wang [25] constructed a linear programming model for integrated production planning for Canadian steel making company. The production plan is considered as an integrated process, including raw material procurement, semi-finished product procurement, capacity allocation, and finished product production and distribution. The purpose of the model is to optimize production planning based on production costs, product throughput rates, customer demands, sales prices,

and facility capacities. Lagos et al. [26] consider a mining problem involving extraction and processing decisions under capacity constraints, and they solved the uncertainty of the ore grade.

The measure of dealing with uncertainty in supply chain optimization is usually completed by four common methods, stochastic programming (Azaron et al. [27]), fuzzy programming (Mitra et al. [28] and Lin et al. [29]), probabilistic programming (You and Grossmann [30]), and sensitivity analysis for instance. Robust optimization (Ben-Tal and Nemirovski [31]) is able to provide a framework to handle the uncertainty of parameters in the problems related to optimization within the category of a given set of bounded uncertainties, and the optimal solution can also be offered in an uncertain implementation. In addition, robust optimization can solve the model with uncertain parameters and realize the feasibility and optimality of the solution in the worst case of parameters. Bertsimas and Thiele [32] proposed a general robust methodology for the purpose of solving the problem of inventory with fixed costs and constraints of capacity for production/inventory. Gurnani et al. [33] focused on the analysis of a supply chain in assembly systems, where there was uncertainty in yields and demand. Zahiri et al. [34] proposed a robust model of demand-deterministic supply chain, which was composed of the centers for collection and distribution. Pishvaee et al. [35] fixed their attention on the study of a robust optimization model to manage uncertainty data for the problems existing in supply chain design. Vahdani et al. [36] proposed a reliable design model of supply chain in the circumstance of uncertainty. They present a dual-objective mathematical programming method, which can minimize the total cost and expected transportation cost under the failure of logistics network facilities. Additionally, Paydar et al. [37] proposed a MILP model for the used oil supply chain. Based on the uncertainty of oil collection, the robust optimization method was proposed by Mulvey et al. [38]. [32] Safaei et al. [39] proposed a mixed integer linear programming model, which can be used to optimize the paper and cardboard recycling network and solve demand uncertainty of the network by using the robust optimization. Jiao et al. [40] studied the design of sustainable closed-loop supply chain in a variety of uncertain circumstances. They also presented the data-driven approaches to generate robust closed-loop supply chain which can reduce uncertainty. It follows that most of the studies conducted previously put emphasis on the mechanism modeling and failed to set up the supply chain system from the data-driven perspective. Therefore, it is quite difficult to ensure the accuracy of the mechanism model, due to the subjective factors. In fact, data on relevant industries are easy to collect because it reflects the objective world.

## 3. Problem Description and Model

*3.1. Problem Description.* In recent years, with the continuous increase of global mining investment and production capacity, the overall supply capacity of ore has been significantly improved, which further promotes the continuous increase of global iron ore production and sea transport. The long-term stability of ore supply chain will attach great importance.

Macroscopically, the relevant nodes of the mine supply chain include mines, concentrating mills, logistics centers, and final users. The efficient operation of mine supply chain can realize the minimum total cost, including mining, processing, inventory, and transportation.

Different from general enterprises, mining enterprises extract ore resources from nature. Each ore body has its own characteristic; there are significant differences in mining technology, methods, and equipment. It is widely believed that ore grade is a very important factor in the process of mining. The production capacity of mine is affected by the grade distribution of ore body. The fluctuation of ore grade distributed in different orebodies will have a great influence on the output and quality of mined ore and then affect the subsequent processing. Therefore, the variable cut-off grade should be considered in the mining plan to allow for the flexibility of the mining stage under production capacity constraints. Due to the development of technology and industry, the quality of mineral requirements is more and more high, the need for mineral processing of raw ore, which is mineral processing. The mineral processing is the most important section in the whole production of mine. In addition, in order to meet the final users' demand for ore products in the supply chain, it is necessary to coordinate the mining and processing of ore. Therefore, the mine supply chain mentioned in this paper is different from the traditional single (multi) product and single (multi) period supply chain.

Figure 1 shows the framework of mine supply chain. Ore mined from multiple locations needs to be blended to meet the grade requirements of the concentrating mill. Then, the ore products from the concentrating mill will be transported to logistics centers in batches. According to demand, the logistics centers will store ore products of specific grades. Finally, the ore product will be transported from the logistics centers to the final users according to requirements.

The actual assumptions are as follows:

(1) Each mine has a number of ore extraction locations and can extract a variety of ore grade of raw ore, but the mining capacity of each mine is limited

(2) In each mine, raw ore from different locations is transported to the concentrating mill which belongs to mine for processing, and only one type of ore products is produced in the whole process

(3) Ore products of different grade are transported from the concentrating mills to logistics centers in batches, and ore products are then sent to final users in batches

In the next section, a robust optimization model is proposed.

*3.2. Model.* A robust optimization model is built based on the sets, parameters, and variables as follows.

*Sets and parameters*

$M$: set of concentrating mills corresponding to mines

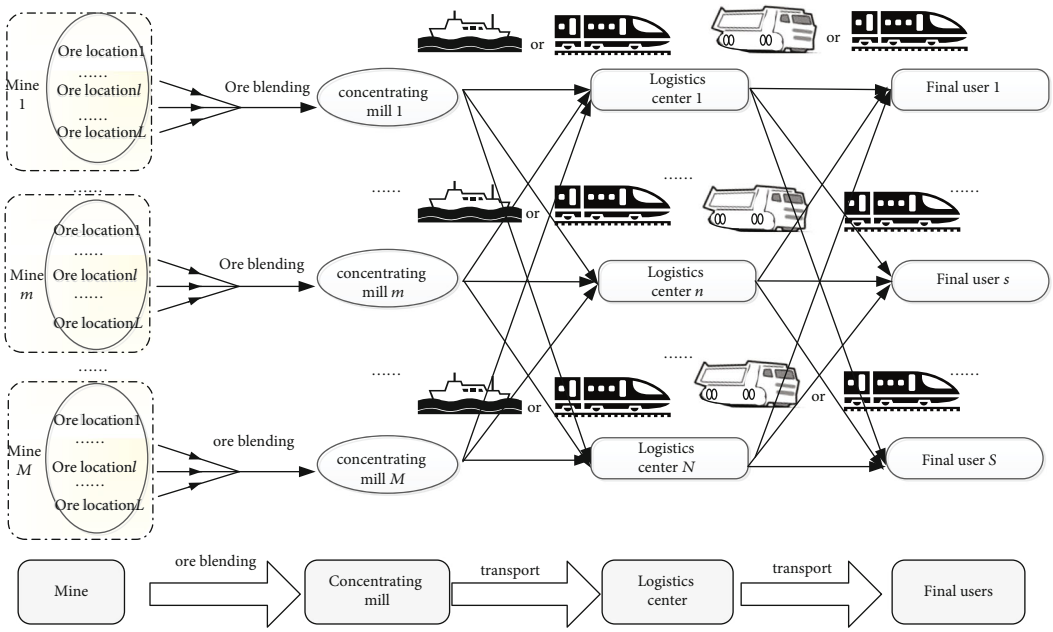



Figure 1: The framework of mine supply chain.

$N$: set of logistics centers

$S$: set of final users

$T$: set of time periods

$L_m$: number of mining sites in mine $m$ ($l = 1, 2, \cdots, |L_m|$)

$O_m$: number of optional cut-off grade of mine $m$ ($o = 1, 2, \cdots, |O_m|$)

$r_{ml}$: consumption resource associated with the extraction of site $l$ of mine $m$

$C_{mlo}$: unit exploitation cost of site $l$ of mine $m$ according the cut-off grade $o$

$P_m$: unit processing cost of concentrating mill of mine $m$

$\mathrm{IO}_m$: unit inventory cost of concentrating mill of mine $m$

$\mathrm{TR}_{mn}$: unit transport cost of ore product from concentrating mill of mine $m$ to logistics center $n$

$\mathrm{ID}_{mn}$: unit inventory cost at logistics center $n$ for the ore product from concentrating mill of mine $m$

$\mathrm{DC}_{ns}$: unit transportation cost of ore product from logistics center $n$ to final user $s$

$\mathrm{SD}_{mnt}$: start-up cost of ore products from concentrating mill of mine $m$ at logistics center $n$ in time period $t$

$\mathrm{ST}_{nst}$: start-up cost when ore products are transported from logistics center $n$ to final user $s$ in phase $t$

$U_{mt}$: maximum storage of ore products at concentrating mill of mine $m$ in time period $t$

$L_{mt}$: minimum storage of ore products at concentrating mill of mine $m$ in time period $t$

$H_{nt}^{\max}$: maximum storage of ore products from all concentrating mill at logistics center $n$ in time period $t$

$D_{mst}$: demand of final user $s$ for ore products of mine $m$ in time period $t$

$g_{mlot}$: the cut-off grade $o$, which is proportion of useful ore obtained from total mining, at the site $l$ of mine $m$ in time period $t$. In the mining process, for the reason of the error existing in exploration and the complex and variable geology, geomorphology, and grade of the mine, the actual value may deviate from the preestimated value, and the parameter $g_{mlot}$ can be random. Through statistical analysis by a large amount of actual mining data, it is observed that the parameter $\tilde{g}_{mlot}$ changes in the polyhedron (Khan and Asad [41]) and it can be expressed by the formula as follows: $\tilde{g}_{mlot} \in \Omega = \{g | Hg \le q\}$. Here, $H$ is uncertainty set matrix, and $q$ is parameter matrix

$g_m^{\min}$: minimum selected grade of the concentrating mill of mine $m$

$j_m$: productivity of ore products of concentrating mill of mine $m$

$\mathrm{SC}_{mlot}$: start-up cost of location $l$ of mine $m$ using the cut-off grade $o$ in time period $t$

$\mathrm{CA}_{mt}$: production ability of mine $m$ in time period $t$

$\mathrm{PA}_{mt}$: processing ability of concentrating mill of mine $m$ in time period $t$

$\mathrm{re}_{mlot}$: unit consumption resource of location $l$ of mine $m$ using the cut-off grade $o$ in time period $t$

$MC_{mnt}$: maximum transport ability of ore products of concentrating mill of mine $m$ to logistics center $n$ in time period $t$

$M$: large positive number

*Variables*

$X_{mlot}$: amount of ore mined from location $l$ of mine $m$ using the cut-off grade $o$ in period time $t$

$X_{mt}$: amount of ore blended in the concentrating mill of mine $m$ in time period $t$

$I_{mt}$: inventory of ore products processed in concentrating mill of mine $m$ in time period $t$

$X_{mnt}$: amount of ore products transported from concentrating mill of mine $m$ to logistics center $n$ in time period $t$

$I_{mnt}$: inventory of ore products of mine $m$ in logistics center $n$ in time period $t$

$X_{mnst}$: amount of the ore products of mine $m$ transported from logistics center $n$ to final user $s$ in time period $t$

$\lambda_{mlot}$: 1 if ore is mined from location $l$ of mine $m$ using the cut-off grade $o$ in time period $t$, 0 otherwise

$\omega_{mnt}$: 1 if ore products shipped from the concentrating mill of mine $m$ to logistics center $n$ in time period $t$, 0 otherwise

$\pi_{nst}$: 1 if the ore products shipped from logistics center $n$ to final user $s$ in time period $t$, 0 otherwise

*Formulation*

$$\begin{aligned}\text{Min } Z = & \sum_{m=1}^{M}\sum_{l=1}^{L_m}\sum_{o=1}^{O_m}\sum_{t=1}^{T}\text{SC}_{mlot}\lambda_{mlot} + \sum_{m=1}^{M}\sum_{l=1}^{L_m}\sum_{o=1}^{O_m}\sum_{t=1}^{T}C_{mlo}x_{mlot} \\ & + \sum_{m=1}^{M}\sum_{t=1}^{T}P_m X_{mt} + \sum_{m=1}^{M}\sum_{t=1}^{T}\text{IO}_m I_{mt} + \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=1}^{T}\text{SD}_{mnt}\omega_{mnt} \\ & + \sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{t=1}^{T}\text{ST}_{nst}\pi_{nst} + \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=1}^{T}\text{TR}_{mn}X_{mnt} \\ & + \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{t=1}^{T}\text{ID}_{mn}I_{mnt} + \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{t=1}^{T}\text{DC}_{ns}X_{mnst},\end{aligned} \tag{1}$$

subject to

$$\sum_{l\in L_m}\left(r_{ml}\sum_{o\in O_m}\lambda_{mlot} + \sum_{o\in O_m}\text{re}_{mlot}X_{mlot}\right) \le \text{CA}_{mt}, \forall m\in M, t\in T, \tag{2}$$

$$\sum_{l\in L_m}\sum_{o\in O_m}X_{mlot} \le \text{PA}_{mt}, \forall m\in M, t\in T, \tag{3}$$

$$\min_{\substack{\varepsilon^T H\ge x\\ \varepsilon\ge 0}}\sum_{l\in L_m}\sum_{o\in O_m}q\varepsilon_{mlot}^T \ge g_m^{\min}\sum_{l\in L_m}\sum_{o\in O_m}X_{mlot} \quad \forall m\in M, t\in T, \tag{4}$$

$$\sum_{l\in L_m}\sum_{o\in O_m}X_{mlot} \ge j_m X_{mt}, \forall m\in M, t\in T, \tag{5}$$

$$X_{mlot} \le M\lambda_{mlot}, \forall m\in M, l\in L_m, o\in O_m, t\in T, \tag{6}$$

$$\sum_{o\in O_m}\lambda_{mlot} \le 1, \forall m\in M, l\in L_m, t\in T, \tag{7}$$

$$X_{mt} + I_{m,t-1} - \sum_{n\in N}X_{mnt} = I_{mt}, \forall m\in M, t\in T, \tag{8}$$

$$X_{mnt} + I_{m,n,t-1} - \sum_{s\in S}X_{mnst} = I_{mnt}, \forall m\in M, n\in N, t\in T, \tag{9}$$

$$L_{mt} \le I_{mt} \le U_{mt}, \forall m\in M, t\in T, \tag{10}$$

$$\sum_{m\in M}I_{mnt} \le H_{nt}^{\max}, \forall n\in N, t\in T, \tag{11}$$

$$X_{mnt} \le MC_{mnt}\omega_{mnt}, \forall m\in M, n\in N, t\in T, \tag{12}$$

$$\sum_{m\in M}X_{mnst} \le M\pi_{nst}, \forall n\in N, s\in S, t\in T, \tag{13}$$

$$\sum_{n\in N}X_{mnst} \ge D_{mst}, \forall n\in N, s\in S, t\in T \tag{14}$$

$$X_{mlot}, X_{mt}, I_{mt}, X_{mnt}, I_{mnt}, X_{mnst} \ge 0, \forall m\in M, l\in L_m, o\in O_m, n\in N, s\in S, t\in T, \tag{15}$$

$$\lambda_{mlot}, \omega_{mnt}, \pi_{nst} \in \{0,1\}, \forall m\in M, l\in L_m, o\in O_m, n\in N, s\in S, t\in T. \tag{16}$$

Equation (1) presents the objective function that minimizes the total cost of production, processing, inventory, transportation, and other related.

Constraint (2) indicates the tonnage of ore removed does not exceed the production ability. Constraint (3) indicates the tonnage of ore blended does not exceed the processing ability.

Constraint (4) ensures that ore processed in the concentrating mill meets the minimum ore grade requirements. Using the dual gap of linear programming is zero, the constraints (17)–(19) can be substituted for the constraint (4).

$$g_m^{\min}\sum_{l\in L_m}\sum_{o\in O_m}X_{mlot} - \sum_{l\in L_m}\sum_{o\in O_m}q\varepsilon_{mlot}^T \le 0 \quad \forall m\in M, t\in T, \tag{17}$$

$$\varepsilon_{mlot}^T H \ge X_{mlot} \quad \forall m\in M, l\in L_m, o\in O_m, t\in T, \tag{18}$$

$$\varepsilon_{mlot} \ge 0 \quad \forall m\in M, l\in L_M, o\in O_m, t\in T. \tag{19}$$

Constraint (5) means to achieve the productive rate. Constraint (6) indicates logical constraints between variables related to mining. Constraint (7) requires that for each time period, only one ore cut-off grade can be selected. Constraints (8) and (9) represent the balance constraints. Constraints (10) and (11) ensure that the amount of ore products is required to be between the minimum and maximum inventory in each concentrating mill and logistics center. Constraints (12) and (13) represent the logical constraint. Constraint (14) means meeting the final user's demand for ore products in terms of time and quantity. Finally, constraints (15) and (16) enforce nonnegativity and integrality, as appropriate.

## 4. Performance Analysis of Model

The methods to deal with the problems of uncertain optimization mainly include analysis on sensitivity, fuzzy programming, stochastic programming, and robust optimization. The purpose of sensitivity analysis is to analyze the influence of uncertain parameter changes on the optimal solution. Sensitivity analysis can be used to study the stability of the optimal solution when the original data is inaccurate or has change, and it can also determine which parameters have a greater impact on the system or model. The robust optimization is derived from the traditional robust control theory, and it is regarded as a replacement of sensitivity analysis and stochastic programming. Robust optimization can limit the uncertain parameters within the disturbance range. The purpose is finding a solution that can be

Table 1: The objective function value results of the robust optimization model.

| Numerical example (M-N-S-T) | $L_m/O_m$ | Mean (T RMB) | Variance | Maximum (T RMB) | Minimum (T RMB) |
|---|---|---|---|---|---|
| 3-2-2-3 | I | 7.7698 | 0.0645 | 8.4266 | 7.3586 |
| | II | 6.5342 | 0.0831 | 7.7921 | 6.3082 |
| | III | 5.4109 | 0.1035 | 6.1293 | 5.4098 |
| | IV | 5.4125 | 0.0792 | 6.0424 | 5.3905 |
| | V | 6.8091 | 0.0592 | 8.0009 | 6.7298 |
| 4-3-3-5 | I | 3.2983 | 0.0628 | 3.8723 | 3.0799 |
| | II | 2.1532 | 0.0592 | 2.8901 | 2.0003 |
| | III | 2.4103 | 0.0425 | 2.9904 | 2.0212 |
| | IV | 2.1096 | 0.0691 | 2.7903 | 2.0109 |
| | V | 2.0271 | 0.0701 | 2.6312 | 2.0002 |
| 6-5-6-8 | I | 111.834 | 0.1463 | 145.342 | 101.093 |
| | II | 85.853 | 0.1596 | 128.093 | 75.334 |
| | III | 85.212 | 0.1394 | 130.984 | 76.987 |
| | IV | 99.035 | 0.0809 | 140.242 | 87.092 |
| | V | 90.352 | 0.1783 | 134.091 | 77.023 |
| 8-5-12-10 | I | 289.355 | 0.0532 | 289.731 | 246.985 |
| | II | 302.437 | 0.1903 | 380.921 | 251.243 |
| | III | 250.876 | 0.3094 | 359.248 | 235.042 |
| | IV | 274.902 | 0.0732 | 320.942 | 256.914 |
| | V | 289.351 | 0.3190 | 341.253 | 268.933 |

effectively resist the uncertainty and ensure the feasibility of the solution over the uncertain sets.

### 4.1. Robust Optimization Analysis

*4.1.1. Hypothesis Testing.* The parameter $\tilde{g}_{mlot}$ changes in the polyhedron as mentioned above and satisfies the following formula:

$$\tilde{g}_{mlot} \in \Omega = \{g | Hg \leq q\}. \tag{20}$$

Through the observation, derivation, and statistical analysis of multiple sets of uncertain parameters, the value in matrix $H$ is assumed to be -10, and in matrix $q$, it is assumed to be -3. In this section, hypothesis testing is used to verify whether the values of these two parameters are feasible.

The process is as follows:

(1) Propose reasonable original hypothesis ($H_0$) and alternative hypothesis ($H_1$) based on the problem

(2) Select suitable test statistics according to hypothetical characteristics

(3) Calculate the value of the test statistic from the sample observation at $H_0$

(4) For a given of significance level $\alpha$, check the table or calculate the critical value through the distribution of test statistics and then get rejection domain and acceptance domain of $H_0$

(5) The decision is to accept $H_0$ when the value of the test statistic falls into the accepted domain, and otherwise, reject $H_0$

Find the 20 samples from actual observations, due to the small sample size ($n < 30$), the $T$-test is used.

Establish two assumptions:

*Original hypothesis ($H_0$)*: $\mu < 0.3$

*Alternative hypothesis ($H_1$)*: $\mu \geq 0.3$

$\mu$ represents the average value of the ratio of the useful ore component quality to the total ore quality obtained by selecting different mining cut-off grade at different periods of the mine.

Based on the original hypothesis, probability is obtained of the sample mean or more extreme mean. If the probability is large, the original hypothesis $H_0$ is considered correct; otherwise, the original hypothesis is considered wrong, and the alternative hypothesis $H_1$ is accepted.

Since the sample is normal distribution, the number of samples is 20, and the statistic is the $t$-statistic, and the formula is as follows:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t(n-1). \tag{21}$$

$\bar{x}$ is the sample average, $\mu$ is the ensemble average, $S$ is the sample standard deviation, and $n$ is the sample size.

This statistic is the $t$ distribution of the degree of freedom is $(n-1)$.

Actual observations were made on the ratio of the useful ore component quality to the total ore mining quality

Table 2: The objective function value of different mining capacity and combination of the $L_m$ and $O_m$.

| Numerical example (M-N-S-T) | $L_m/O_m$ | $U^1$ (T RMB) | $U^2$ (T RMB) | $U^3$ (T RMB) | $U^4$ (T RMB) | $U^5$ (T RMB) | $U^6$ (T RMB) | $U^7$ (T RMB) |
|---|---|---|---|---|---|---|---|---|
| 3-2-2-3 | I | 7.6839 | 7.6923 | 7.7231 | 7.8423 | 7.8901 | 7.9021 | 7.9521 |
| | II | 6.3369 | 6.3588 | 6.3730 | 6.4484 | 6.7816 | 6.8315 | 6.9723 |
| | III | 5.3950 | 5.3950 | 5.3966 | 5.4326 | 5.5778 | 5.7789 | 5.8452 |
| | IV | 5.3776 | 5.3776 | 5.3951 | 5.4094 | 5.4739 | 5.5349 | 5.5892 |
| | V | 6.7086 | 6.7096 | 6.7452 | 6.9160 | 7.4193 | 7.3130 | 7.5295 |
| 4-3-3-5 | I | 3.2654 | 3.2832 | 3.2937 | 3.3597 | 3.1307 | 3.3642 | 3.3958 |
| | II | 2.1203 | 2.1203 | 2.1203 | 2.1216 | 2.1228 | 2.1394 | 2.1518 |
| | III | 2.3111 | 2.3117 | 2.3121 | 2.3137 | 2.3206 | 2.3356 | 2.3813 |
| | IV | 2.0817 | 2.0846 | 2.0891 | 2.0967 | 2.1044 | 2.2041 | 2.2402 |
| | V | 2.0198 | 2.0198 | 2.0208 | 2.0228 | 2.0303 | 2.0621 | 2.0772 |
| 6-5-6-8 | I | 109.147 | 109.589 | 110.184 | 110.870 | 111.021 | 111.068 | 112.098 |
| | II | 83.989 | 84.152 | 84.456 | 84.792 | 85.411 | 85.896 | 86.231 |
| | III | 83.895 | 84.008 | 84.229 | 84.399 | 84.712 | 85.023 | 85.932 |
| | IV | 95.637 | 95.724 | 95.913 | 96.056 | 96.305 | 96.832 | 97.235 |
| | V | 86.961 | 86.983 | 86.992 | 87.009 | 87.135 | 87.832 | 88.016 |
| 8-5-12-10 | I | 265.303 | 266.831 | 269.233 | 270.665 | 272.116 | 274.359 | 283.012 |
| | II | 291.983 | 292.580 | 293.404 | 293.878 | 294.387 | 295.346 | 296.063 |
| | III | 238.284 | 238.423 | 238.535 | 238.666 | 238.783 | 239.012 | 239.983 |
| | IV | 250.909 | 250.961 | 250.985 | 250.979 | 251.092 | 251.931 | 252.096 |
| | V | 276.022 | 276.049 | 276.077 | 276.092 | 276.115 | 276.903 | 277.012 |
| 10-5-12-10 | I | 340.973 | 341.329 | 341.772 | 342.006 | 342.366 | 343.001 | 343.893 |
| | II | 334.036 | 334.357 | 334.877 | 335.255 | 335.768 | 335.982 | 336.023 |
| | III | 284.802 | 284.851 | 284.857 | 284.878 | 285.912 | 286.012 | 288.022 |
| | IV | 294.162 | 294.181 | 294.261 | 294.231 | 294.257 | 294.719 | 294.975 |
| | V | 309.256 | 309.253 | 309.276 | 309.299 | 309.542 | 309.892 | 310.021 |
| 15-5-15-10 | I | 537.261 | 537.321 | 537.982 | 537.998 | 538.231 | 539.012 | 539.823 |
| | II | 432.421 | 432.498 | 432.512 | 432.832 | 432.987 | 433.053 | 433.875 |
| | III | 421.905 | 421.998 | 422.012 | 422.429 | 422.498 | 423.712 | 423.891 |
| | IV | 415.235 | 412.389 | 412.578 | 412.698 | 412.986 | 413.245 | 414.046 |
| | V | 409.357 | 409.398 | 409.406 | 409.502 | 409.584 | 410.302 | 410.356 |
| 25-5-15-10 | I | 589.599 | 589.904 | 590.052 | 590.146 | 590.250 | 590.457 | 590.602 |
| | II | 508.920 | 508.874 | 508.922 | 508.930 | 509.103 | 509.521 | 510.903 |
| | III | 509.103 | 509.428 | 509.754 | 509.981 | 509.995 | 510.325 | 510.932 |
| | IV | 488.837 | 488.835 | 488.879 | 488.914 | 488.941 | 488.982 | 489.041 |
| | V | 479.032 | 479.321 | 479.832 | 479.982 | 479.998 | 480.216 | 480.427 |
| 30-5-20-10 | I | 673.923 | 674.022 | 675.921 | 675.995 | 676.023 | 676.901 | 679.313 |
| | II | 665.321 | 665.492 | 665.792 | 666.912 | 666.998 | 670.321 | 679.352 |
| | III | 650.384 | 650.822 | 650.982 | 651.342 | 652.094 | 653.926 | 654.932 |
| | IV | 642.394 | 642.834 | 643.136 | 643.213 | 644.932 | 645.138 | 658.227 |
| | V | 639.356 | 639.722 | 640.201 | 640.835 | 642.930 | 631.159 | 638.315 |

obtained by selecting different cut-off grades at different sites at different periods, and 20 sample values were obtained, respectively:

$x_1, \cdots, {}_{20} = \{0.62, 0.27, 0.44, 0.50, 0.26, 0.21, 0.48, 0.69, 0.36, 0.28, 0.41, 0.65, 0.21, 0.24, 0.50, 0.38, 0.11, 0.45, 0.31, 0.25\}$.

After calculation, the following results can be obtained: $\bar{x} = \sum_{i=1}^{20} x_i/20 = 0.381$, $S = \sqrt{1/19\sum_{i=1}^{20}(x_i - \bar{x})^2} \approx 0.16$, and $t = 2.27$.

Taking the threshold is 0.025, according to the table of quantile of $t$ distribution, the $t_{19}(0.025) = 2.093$ can be

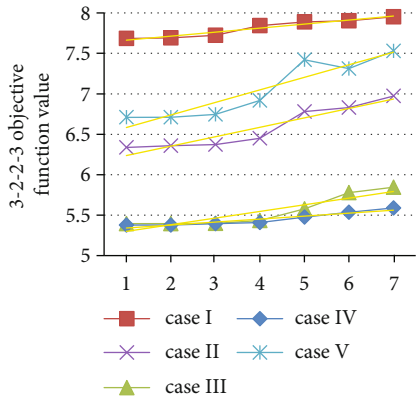



Figure 2: Sensitivity curve of the objective function value on the disturbance of mining capacity of example "3-2-2-3."

found. Since the absolute value of the $t$-statistic is 2.27 and falls within the "rejection domain," the original hypothesis is rejected and the alternative hypothesis is accepted. This shows that the sample average is significantly different from the overall average. So, it can be concluded that the average value grade of the useful ore component quality obtained by different cut-off grades at different periods at different sites to the total ore quality is greater than or equal to 0.3.

Through the above statistical hypothesis testing, it can be finally determined that the value in the matrix $H$ is -10, and the value in the matrix $q$ is -3.

*4.1.2. Solution of Robust Optimization Model.* After determining the value range of $H$ and $q$, the results of $H$ and $q$ are substituted into Equations (17)–(19), the robust optimization model is directly solved by mainstream optimization software ILOG CPLEX V12.6.1, and then, the objective function value was obtained.

The purpose of numerical experiments is to verify the optimality and stability of the robust optimization model, so only small-scale examples are tested.

For the robust optimization model, the objective function values are obtained by 20 times disturbances.

Table 1 lists the mean, variance, maximum, and minimum values of the objective function values of the robust optimization model under disturbed 20 times for 4 examples. A different numbers of mines, logistics centers, final users, and time periods are considered, denoted by "M-N-S-T," which are as follows: "3-2-2-3," "4-3-3-5," "6-5-6-8," and "8-5-12-10." Based on the combinations of mining sites ($L_m$) and ore cut-off grades ($O_m$), there are 5 cases in each group, i.e., case I: $L_m = \{3\}$ and $O_m = \{3\}$, case II: $L_m = \{5\}$ and $O_m = \{3\}$, case III: $L_m = \{10\}$ and $O_m = \{5\}$, case IV: $L_m = \{15\}$ and $O_m = \{5\}$, and case V: $L_m = \{20\}$ and $O_m = \{5\}$.

The variance was a measure of the dispersion of a set of data. The results show that the variance of these examples was small, which proved that the fluctuation of examples was small as well. This also verifies that the robust optimization model was quite stable. Even if the actual mining grade differed greatly from expected, it can ensure the stability of the entire supply chain and meet demand.

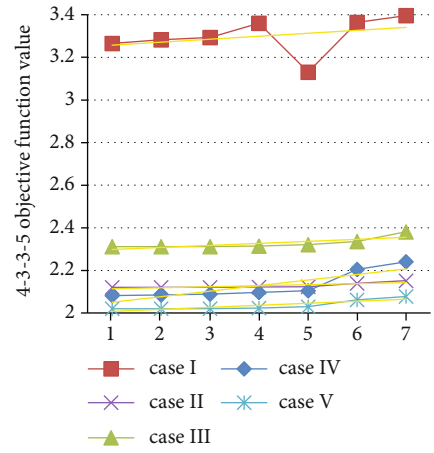



Figure 3: Sensitivity curve of the objective function value on the disturbance of mining capacity of example "4-3-3-5."

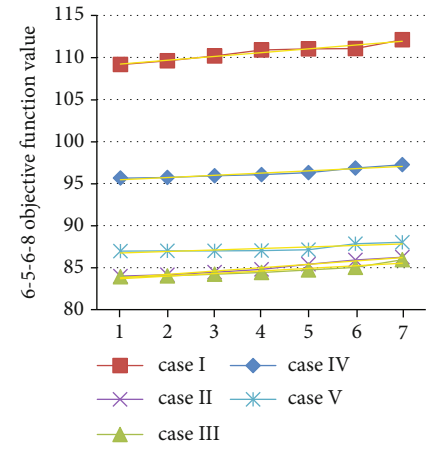



Figure 4: Sensitivity curve of the objective function value on the disturbance of mining capacity of example "6-5-6-8."

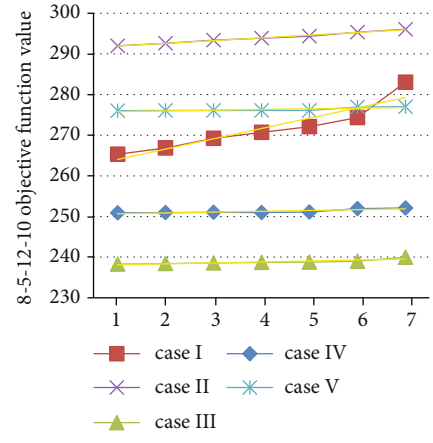



Figure 5: Sensitivity curve of the objective function value on the disturbance of mining capacity of example "8-5-12-10."

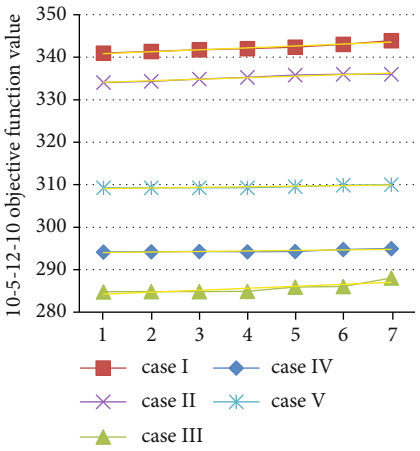



Figure 6: Sensitivity curve of the objective function value on the disturbance of mining capacity of example "10-5-12-10."

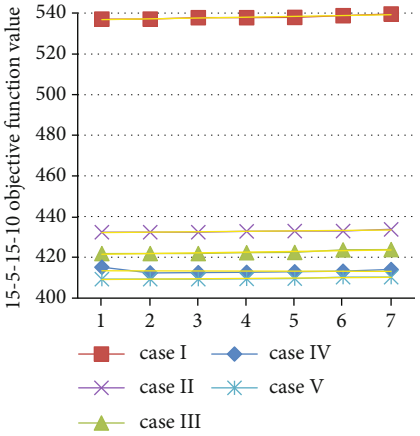



Figure 7: Sensitivity curve of the objective function value on the disturbance of mining capacity of example "15-5-15-10."

*4.2. Sensitivity Analysis.* The calculation examples in this section are similar to the previous section, and four groups are again added here, which are as follows: "10-5-12-10," "15-5-15-10," "25-5-15-10," and "30-5-20-10."

For each group of examples "M-N-S-T," the parameters of seven situations are set according to different mining capabilities under the corresponding mining site $L_m$ and the cut-off grade $O_m$, so $U^1$, $U^2$, $U^3$, $U^4$, $U^5$, $U^6$, and $U^7$ are shown, respectively. In these seven examples, other parameters remain unchanged, and only the mining capacity $CA_{mt}$ changes. Table 2 lists the best objective function values for different mining capacities and mining site $L_m$ and the cut-off grade $O_m$.

The objective function value is obtained based on five different mining sites and combination of schemes under seven different mining capacities. The sensitivity curve of the whole objective function value after the disturbance of the mining capacity is plotted, as shown in Figures 2–9. The yellow line among the five broken lines is the linear trend line in each figure, and it represents the variation trend of objective function value.

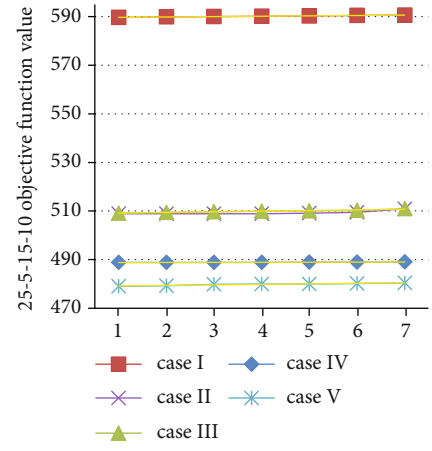



Figure 8: Sensitivity curve of the objective function value on the disturbance of mining capacity of example "25-5-15-10."

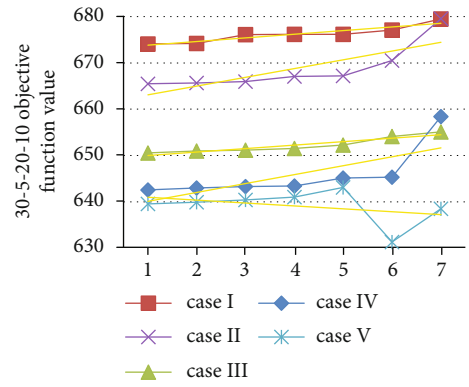



Figure 9: Sensitivity curve of the objective function value on the disturbance of mining capacity of example "30-5-20-10."

From Table 2 and sensitivity curves 1~8, the following conclusions can be drawn:

(1) In terms of vertical comparison, for each example, as the number of mining sites and the programs increases, the overall objective function value shows a downward trend, indicating that the increase of the mining site and multiple alternatives can help reduce the total cost

(2) In terms of horizontal comparison, as far as the overall trend is concerned, with the same number of mines, logistics centers, final users, and time periods and the same set of mining sites and schemes, as the mining capacity decreases, the total cost increases (in some cases, the target value first decreases and then increases), which shows that as long as resources are allocated and used reasonably, the total cost can be reduced

## 5. Conclusion

In this paper, a robust optimization model was established for mine supply chain under big data, which includes mines,

concentrating mills, logistics centers, and final users. The model not only considers the production details such as mining, grinding and separation, and ore blending, but also satisfies the intermediate links such as transportation and inventory and will ultimately meet the final users' requirements for ore product amount, ore grade, and time period. The model has universality and can be applied to different types of mines with different properties.

In this paper, the model is solved using the actual mine data from open pit mines, and the results have practical significance and value for mining enterprises. Through statistical analysis, it is obtained that the variable cut-off grade changes within a polyhedron. The analysis of the robust performance results shows that when the actual survey data deviates from the expected value, the robust optimization model built in this paper can be used to obtain the optimal solution, and even if the parameters are disturbed, the solution of the model is still stable. In conclusion, the robust optimization model proposed in this paper has stability and optimality.

Additionally, the sensitivity analysis was performed on the model, and the influence on the objective function value imposed by the parameter such as mining capacity, the change of the mining site, and the cut-off grade was obtained. Through the rational integration and allocation of resources, the production and logistics planning of open-pit mines can make more decisions that are reasonable.

In future research, we will continue to study in-depth changes in ore grades, hoping that there will be breakthroughs in big data analysis, and explore more static or dynamic influencing factors that may affect the entire mine production and logistics system, hoping to bring more profits to related enterprises in the mine supply chain.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The author declares no competing interest.

## Acknowledgments

This work was supported by the Scientific Research Project of Education Department of Liaoning Province (L2019639).

## References

[1] A. Gunasekaran, T. Papadopoulos, R. Dubey et al., "Big data and predictive analytics for supply chain and organizational performance," *Journal of Business Research*, vol. 70, pp. 308–317, 2017.

[2] A. McAfee and E. Brynjolfsson, "Big data: the management revolution," *Harvard Business Review*, vol. 90, no. 90, pp. 60–68, 2012.

[3] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management," *Journal of Business Logistics*, vol. 34, no. 2, pp. 77–84, 2013.

[4] Supply Chain Council, *Supply-Chain Operations Reference Model*, Revision 11.0, Supply Chain Council, 2012, http://www.supply-chain.org.

[5] M. Brinch, "Understanding the value of big data in supply chain management and its business processes," *International Journal of Operations & Production Management*, vol. 38, no. 7, pp. 1589–1614, 2018.

[6] H. Stadtler and C. Kilger, *Supply Chain Management and Advanced Planning*, Springer, Darmstadt, 3rd edition, 2005.

[7] W. Liu, D. Sun, and T. Xu, "Integrated production and distribution planning for the iron ore concentrate," *Mathematical Problems in Engineering*, vol. 2019, Article ID 7948349, 10 pages, 2019.

[8] G. Ghiani, G. Laporte, and R. Musmanno, *Introduction to Logistics Systems Planning and Control*, John Wiley & Sons, 2004.

[9] P. Jonsson and J. Holmström, "Future of supply chain planning: closing the gaps between practice and promise," *International Journal of Physical Distribution and Logistics Management*, vol. 46, no. 1, pp. 62–81, 2016.

[10] Y. Mauergauz, *Advanced Planning and Scheduling in Manufacturing and Supply Chains*, Springer, Moscow, 2016.

[11] B. Brunaud and I. E. Grossmann, "Perspectives in multilevel decision-making in the process industry," *Frontiers of Engineering Management*, vol. 4, no. 3, pp. 256–270, 2017.

[12] T. Nishi, R. Shinozaki, and M. Konishi, "An augmented Lagrangian approach for distributed supply chain planning for multiple companies," *IEEE Transactions on Automation Science and Engineering*, vol. 5, no. 2, pp. 259–274, 2008.

[13] M. Steinrücke, "An approach to integrate production-transportation planning and scheduling in an aluminium supply chain network," *International Journal of Production Research*, vol. 49, no. 21, pp. 6559–6583, 2011.

[14] C. Vintró, L. Sanmiquel, and M. Freijo, "Environmental sustainability in the mining sector: evidence from Catalan companies," *Journal of Cleaner Production*, vol. 84, pp. 155–163, 2014.

[15] K. Söderholm, P. Söderholm, H. Helenius et al., "Environmental regulation and competitiveness in the mining industry: permitting processes with special focus on Finland, Sweden and Russia," *Resources Policy*, vol. 43, pp. 130–142, 2015.

[16] S. Kusi-Sarpong, J. Sarkis, and X. Wang, "Assessing green supply chain practices in the Ghanaian mining industry: a framework and evaluation," *International Journal of Production Economics*, vol. 181, pp. 325–341, 2016.

[17] A. Azapagic, "Developing a framework for sustainable development indicators for the mining and minerals industry," *Journal of Cleaner Production*, vol. 12, no. 6, pp. 639–662, 2004.

[18] P. Russom, "Big data analytics," *TDWI best practices report*, vol. 19, The Data Warehousing Institute (TDWI), 2011.

[19] M. Seyedan and F. Mafakheri, "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities," *Journal of Big Data*, vol. 7, no. 1, 2020.

[20] G. Wang, A. Gunasekaran, E. W. T. Ngai, and T. Papadopoulos, "Big data analytics in logistics and supply chain management: certain investigations for research and applications," *International Journal of Production Economics*, vol. 176, pp. 98–110, 2016.

[21] R. G. J. Richey, T. R. Morgan, K. Lindsey-Hall, and F. G. Adams, "A global exploration of big data in the supply chain," *International Journal of Physical Distribution & Logistics Management*, vol. 46, no. 8, pp. 710–739, 2016.

[22] G. Büyüközkan and F. Göçer, "Digital supply chain: literature review and a proposed framework for future research," *Computers in Industry*, vol. 97, pp. 157–177, 2018.

[23] H. S. Yuan, L. Shan, and G. Chao, "Research and implementation of dimension reduction algorithm in big data analysis," *Artificial Intelligence and Security, 7th International Conference*, pp. 14–26, 2021.

[24] A. M. Newman, E. Rubio, R. Caro, A. Weintraub, and K. Eurek, "A review of operations research in mine planning," *Interfaces*, vol. 40, no. 3, pp. 222–245, 2010.

[25] M. Chen and W. Wang, "A linear programming model for integrated steel production and distribution planning," *International Journal of Operations & Production Management*, vol. 17, no. 6, pp. 592–610, 1997.

[26] G. Lagos, D. Espinoza, E. Moreno, and J. Amaya, "Robust planning for an open-pit mining problem under ore-grade uncertainty," *Electronic Notes in Discrete Mathematics*, vol. 37, pp. 15–20, 2011.

[27] A. Azaron, K. N. Brown, S. A. Tarim, and M. Modarres, "A multi-objective stochastic programming approach for supply chain design considering risk," *International Journal of Production Economics*, vol. 116, no. 1, pp. 129–138, 2008.

[28] K. Mitra, R. D. Gudi, S. C. Patwardhan, and G. Sardar, "Towards resilient supply chains: uncertainty analysis using fuzzy mathematical programming," *Chemical Engineering Research and Design*, vol. 87, no. 7, pp. 967–981, 2009.

[29] K. P. Lin, P. T. Chang, K. C. Hung, and P. F. Pai, "A simulation of vendor managed inventory dynamics using fuzzy arithmetic operations with genetic algorithms," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2571–2579, 2010.

[30] F. You and I. E. Grossmann, "Design of responsive supply chains under demand uncertainty," *Computers and Chemical Engineering*, vol. 32, no. 12, pp. 3090–3111, 2008.

[31] A. Ben-Tal and A. Nemirovski, "Selected topics in robust convex optimization," *Mathematical Programming*, vol. 112, no. 1, pp. 125–158, 2008.

[32] D. Bertsimas and A. Thiele, "A robust optimization approach to inventory theory," *Operations Research*, vol. 54, no. 1, pp. 150–168, 2006.

[33] H. Gurnani, R. Akella, and J. Lehoczky, "Supply management in assembly systems with random yield and random demand," *IIE Transactions*, vol. 32, no. 8, pp. 701–714, 2000.

[34] B. Zahiri, M. Mousazadeh, and A. Bozorgi-Amiri, "A robust stochastic programming approach for blood collection and distribution network design," *International Journal of Research in Industrial Engineering*, vol. 3, no. 2, p. 1, 2014.

[35] M. S. Pishvaee, M. Rabbani, and S. A. Torabi, "A robust optimization approach to closed-loop supply chain network design under uncertainty," *Applied Mathematical Modelling*, vol. 35, no. 2, pp. 637–649, 2011.

[36] B. Vahdani, R. Tavakkoli-Moghaddam, M. Modarres, and A. Baboli, "Reliable design of a forward/reverse logistics network under uncertainty: a robust-M/M/c queuing model," *Transportation Research Part E: Logistics and Transportation Review*, vol. 48, no. 6, pp. 1152–1168, 2012.

[37] M. M. Paydar, V. Babaveisi, and A. S. Safaei, "An engine oil closed-loop supply chain design considering collection risk," *Computers & Chemical Engineering*, vol. 104, pp. 38–55, 2017.

[38] J. M. Mulvey, R. J. Vanderbei, and S. A. Zenios, "Robust optimization of large-scale systems," *Operations Research*, vol. 43, no. 2, pp. 264–281, 1995.

[39] A. S. Safaei, A. Roozbeh, and M. M. Paydar, "A robust optimization model for the design of a cardboard closed-loop supply chain," *Journal of Cleaner Production*, vol. 166, pp. 1154–1168, 2017.

[40] Z. Jiao, L. Ran, Y. Zhang, Z. Li, and W. Zhang, "Data-driven approaches to integrated closed-loop sustainable supply chain design under multi-uncertainties," *Journal of Cleaner Production*, vol. 185, pp. 105–127, 2018.

[41] A. Khan and M. W. A. Asad, "A method for optimal cut-off grade policy in open pit mining operations under uncertain supply," *Resources Policy*, vol. 60, pp. 178–184, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 4637876, 19 pages
https://doi.org/10.1155/2021/4637876

*Research Article*

# Novel Stream Ciphering Algorithm for Big Data Images Using Zeckendorf Representation

**Liangshun Wu**[1] **and Hengjin Cai**[1,2]

[1]*School of Computer Science, Wuhan University, Wuhan 430079, China*
[2]*Zall Research Institute of Smart Commerce, Wuhan 430010, China*

Correspondence should be addressed to Hengjin Cai; hjcai@whu.edu.cn

Received 5 August 2021; Revised 30 August 2021; Accepted 9 September 2021; Published 21 October 2021

Academic Editor: Rajesh Kaluri



Big data is a term used for very large data sets. Digital equipment produces vast amounts of images every day; the need for image encryption is increasingly pronounced, for example, to safeguard the privacy of the patients' medical imaging data in cloud disk. There is an obvious contradiction between the security and privacy and the widespread use of big data. Nowadays, the most important engine to provide confidentiality is encryption. However, block ciphering is not suitable for the huge data in a real-time environment because of the strong correlation among pixels and high redundancy; stream ciphering is considered a lightweight solution for ciphering high-definition images (i.e., high data volume). For a stream cipher, since the encryption algorithm is deterministic, the only thing you can do is to make the key "look random." This article proves that the probability that the digit 1 appears in the midsection of a Zeckendorf representation is constant, which can be utilized to generate the pseudorandom numbers. Then, a novel stream cipher key generator (ZPKG) is proposed to encrypt high-definition images that need transferring. The experimental results show that the proposed stream ciphering method, with the keystream of which satisfies Golomb's randomness postulates, is faster than RC4 and LSFR with indistinguishable performance on hardware depletion, and the method is highly key sensitive and shows good resistance against noise attacks and statistical attacks.

## 1. Introduction

The development of digital sensor technology and storage device leads to the rapid expansion of the digital image library, and all kinds of digital equipment produce vast amounts of images every day. Though image compression reduces the bandwidth, transferring compressed images alone is still not secure. Thus, how to effectively and securely transfer these images has become a hot research direction in recent years. A variety of encryption algorithms have been investigated to image cryptosystems. Most of them are based on permutation and diffusion architecture [1]. The permutation process alters the location of image pixels, and the diffusion process changes the pixel values so that a small change in one pixel can spread to almost all pixels in the entire image [2]. These two procedures are independent. Modern cryptography includes symmetric encryption, asymmetric encryption, and hash function, among which symmetric encryption is divided into two types: block ciphers and stream ciphers. Block ciphers such as DES, AES, and IDEA, are not suitable for practical image encryption because of intrinsic features of some images such as mass data capacity, strong correlation among pixels, and high redundancy [3].

A stream cipher is a symmetric key encryption where the crypto keys used to encrypt the binary image is randomly changed so that the cipher image produced is mathematically impossible to break. Also, each bit of data is encrypted with each bit of key. The random keys are changed so that it will not allow any pattern to be repeated, giving a clue to the cracker to break the cipher image. The advantage of using stream cipher is that the execution speed is higher when compared to block ciphers and has lower hardware complexity. Unlike block ciphers, stream cipher will not produce the same ciphertext even for repetitive blocks of plaintext, since the keys are changed constantly for every bit of

plaintext. Basically, in stream ciphers, for simplicity, the manner you encrypt is by bitwise XOR, and if you intend to decrypt a ciphertext, you simply do XOR once more. The exclusive or (XOR or $\oplus$) operation, which is simple to implement on hardware, gives a ray of hope for fast image encryption.

However, if multiple data are encrypted with the same key, the attacker can decrypt the data without guessing the key. For example, suppose that two strings of plaintext data, $P_1$ and $P_2$, are encrypted using the same key, $K$. The ciphertexts are as follows: $E_1 = P_1 \oplus K$ and $E_2 = P_2 \oplus K$. Because $E_1 \oplus E_2 = P_1 \oplus K \oplus P_2 \oplus K = P_1 \oplus P_2 \oplus (K \oplus K) = P_1 \oplus P_2$, if XOR $P_2$ on both sides, then $E_1 \oplus E_2 \oplus P_2 = P_1 \oplus P_2 \oplus P_2 = P_1$. At this point, it is clear that the attacker recovered the plaintext without obtaining the key.

Therefore, in stream ciphering, the difficulty of cracking depends on the randomness and unpredictability of the keystream. Alternatively, a keystream generated by a specified generator should at least "look random." The motivation of this paper is to generate such pseudorandom keystreams to resist chosen plaintext attacks and statistical attacks.

This paper is organized as follows: Section 2 introduces preliminary knowledge. Section 3 reviews the related work. Section 4 elaborates on generating a pseudorandom keystream that satisfies Golomb's randomness postulates. Section 5 proves the randomness of the keystream theoretically. Section 6 does some experiments. Finally, Section 7 draws the conclusion.

## 2. Preliminaries

*2.1. Golomb's Randomness Postulates.* Golomb's randomness postulates [4] defines the requisite properties to be sufficiently random looking. Those properties are given as follows: the runs of 0's are called "gaps"; runs of 1's are called "blocks".

(1) In a cycle, the number of 1's differs from that of 0's by at most 1

(2) At least half the runs have length 1, at least one-fourth have length 2, at least one-eighth has length 3, and so forth. Moreover, for each of these lengths, there are (almost) equally many gaps and blocks. In other words, the number of any possible $n$-runs is approximately equal to $\Lambda/2^n$, where $\Lambda$ denotes the length of the keystream

(3) The autocorrelation function $\Gamma(\tau)$:

$$\Gamma(\tau) = \sum_{i=1}^{\Lambda} c_i c_{i+\tau} = \begin{cases} \dfrac{\Lambda}{2}, & \tau = 0, \\ \dfrac{\Lambda}{4}, & 0 < \tau < \Lambda \end{cases} \tag{1}$$

*2.2. Zeckendorf Representation.* It is known from Zeckendorf's theorem [5] that each nonnegative integer can be addressed as a sum of distinct Fibonacci numbers. For instance, 17 is the sum of the 7th, 4th, and 1st Fibonacci numbers, viz. $17 = F_7 + F_4 + F_1$. Every nonnegative integer $N$ admits a representation:

$$M = F_{m_1} + F_{m_2} + \cdots + F_{m_r}, \tag{2}$$

with $m_1 > m_2 > \cdots \geq m_r$, and as usual, $F_0 = 0$, $F_1 = 1$, and $F_{n+2} = F_{n+1} + F_n$ for all $n \geq 0$. We call this a Zeckendorf representation or $F$-addend representation of $N$. It is convenient to write this representation as a word/sequence $\{\varepsilon_{L-1}, \varepsilon_{L-2}, \cdots, \varepsilon_0\}$ of length $L$ with each $\varepsilon_i$ oscillating over the alphabet $\{0, 1\}$, where 1 indicates the respective Fibonacci addend appears in the sum, and 0 otherwise. Let

$$S_L(N) = \{\varepsilon_{L-1}, \varepsilon_{L-2}, \cdots, \varepsilon_0\}, \tag{3}$$

be the aforementioned Zeckendorf representation, for instance, $17 = F_7 + F_4 + F_1$; then,

$$S_9(17) = \{0, 1, 0, 0, 1, 0, 0, 1, 0\}. \tag{4}$$

If imposing the additional requirement that consecutive 1 are not allowed (viz. $m_i \geq m_{i+1} + 2$) and $\varepsilon_1$ cannot be '1' ($F_1 = F_2 = 1$, provided only $F_2 = 1$ admissible), then we obtain the canonical version of the definition [6–9]. Such a "canonical Zeckendorf representation" always exists and is unique [5].

## 3. Related Work

*3.1. Stream Ciphers and PRNG.* Stream cipher is a symmetric key cryptography in which the key is randomly altered in a way that the cipher image created is mathematically impossible to break. The benefit of using stream cipher is that when it is compared to block ciphers, the execution speed is higher and has less hardware complexity. Unlike block ciphers, stream ciphers, even for repetitive blocks of plain text, will not generate the same ciphertext, since the keys are changed constantly for every element of plaintext [10].

Image encryption using some of the existing standard stream cipher methods such as RC4 and Vernam cipher methods have drawbacks. The RC4 algorithm is vulnerable to analytic attacks of the state table. In every 256 keys, there can be a weak key [11]. These keys are identified by cryptanalysis that is able to find circumstances under which one of more generated bytes are strongly correlated with a few bytes of the key [12]. Also, the same sequence of keys is repeated which would enable the hacker to break the ciphertext. Also, the first three words of the secret key can be found, and by iteration, each word of the key used in RC4 can be obtained. The Vernam cipher considered a perfect cipher is a type of one-time pad cipher. The drawback in this method is the need for the unlimited number of keys and the distribution of large number of random keys becomes a problem [13]. Recently, the use of a chaotic system in

cryptography to encrypt images has emerged for its random characteristics [14].

As the core component of stream ciphers, the generation of random numbers is essential. There are two basic types of generators used to produce random sequences: random number generators (RNGs) and pseudorandom number generators (PRNGs). For cryptographic applications, both of these generator types produce a stream of zeros and ones that may be divided into substreams or blocks of random numbers. A random bit sequence could be interpreted as the result of the flips of an unbiased "fair" coin with sides that are labeled "0" and "1." Obviously, the use of unbiased coins for cryptographic purposes is impractical. An RNG considers a nondeterministic source (i.e., the entropy source). The source typically consists of some physical quantity, such as the noise in an electrical circuit, the timing of user processes (e.g., keystrokes or mouse movements), or the quantum effects in a semiconductor. The outputs of an RNG may be used directly as a random number or may be fed into a PRNG. However, producing high-quality random numbers may be so time-consuming. Inputs to PRNGs are called seeds. The outputs of a PRNG are typically deterministic functions of the seed, which is the origin of the term "pseudorandom." Ironically, pseudorandom numbers often appear to be more random than RNGs, because a series of transformations can eliminate statistical autocorrelations between input and output [10].

*3.2. Applications of Zeckendorf Representation.* Zeckendorf representation works well in certain situations. For example, Leroy et al. [15] count the number of distinct (scattered) subwords occurring in a given word. More precisely, it considers the generalization of the Pascal triangle to the binomial coefficients of words and the Zeckendorf representation counting the number of positive entries on each row. Epifanio et al. [16] proved that Zeckendorf representation has deep connections with the Sturmian graph, and Bernat [17] connected Zeckendorf representation with continued fractions.

In addition, the research of stream ciphers that resorts to Zeckendorf representation has long been done. For example, feedback with carry shift registers (FCSRs) plays a vital role in the hardware design of stream ciphers besides LFSRs. Galois representation is often considered the first choice for FCSRs, howbeit, recently, a new representation that generalizes both Galois and Zeckendorf representations for FCSR automata was presented [18]. It is immune to previous attacks and can dramatically improve internal diffusion. Later, Lin [19] further improved the aforementioned FCSR circuit.

Similarly, the U-Quark hash function with FCSRs of Zeckendorf representation [20]. Fish (Fibonacci shrinking), a fast software stream cipher, was proposed to achieve solid performance simulated on an Intel 486 processor [21]. Nevertheless, these researches do not focus on the generation of keystream of the stream cipher at the software level but the hardware level. Our research is suggested adding a small stone to the wall of the application of Zeckendorf representation. Recently, several studies apply Zeckendorf representation in blockchain and big data encryption [22, 23].

## 4. The Proposed Method

*4.1. Probability Structure of Zeckendorf Representation.* Suppose there exist $m$ ones in a canonical Zeckendorf representation, denoted as $S_{L,m}$, it is known from [15] that the quantity of $S_{L,m}$ is given by

$$N_{L,m} = \begin{cases} \binom{L-m+1}{m}, & 0 \le m \le \left\lceil \frac{L+1}{2} \right\rceil, \\ 0, & m > \left\lceil \frac{L+1}{2} \right\rceil, \end{cases} \tag{5}$$

where $\lceil \cdot \rceil$ is the ceiling function.

Let $N_{L,m}(k)$ be the number of representations that own 1 in the $k$th position. Filipponi and Wolfowicz [24] proves the fact that

$$N_{L,m}(k) = N_{L,m}(L-k+1), \tag{6}$$

$$N_{L,m}(k) = \sum_{i=0}^{k-1} (-1)^i \binom{L-m-i}{m-1-i}, \tag{7}$$

where $1 \le k \le \lceil (L+1)/2 \rceil$, or

$$N_{L,m}(k) = \frac{1-(-1)^k}{2} \binom{L-m-k+1}{m-k} + \sum_{i=0}^{\lceil (k-2)/2 \rceil} \binom{L-m-2i-1}{m-2i-1}, \tag{8}$$

resting on the assumption [25]:

$$\binom{a}{-|b|} = 0. \tag{9}$$

It can be easily inferred from (7) that the addend disappears for $i > m+1$. And, from (6), the following are found.

**Theorem 1.** *$N_{L,m}(k) = N_{L,m}(m)$, and it is constant for any $m < k < L-m+1$.*

*Proof.* By using (7), we compute

$$\begin{aligned} N_{L,0}(k) &= 0 \\ N_{L,1}(k) &= 1, \\ N_{L,2}(k) &= \begin{cases} L-2, k \in \{1, L\}, \\ L-3, k \notin \{1, L\}, \end{cases} \end{aligned} \tag{10}$$

and using (8), we get

$$\begin{aligned} N_{L,m}(m) &= \frac{1-(-1)^m}{2} \binom{L-m-m+1}{m-m} \\ &\quad + \sum_{i=0}^{\lceil (m-2)/2 \rceil} \binom{L-m-2i-1}{m-2i-1}, \end{aligned} \tag{11}$$

then

$$N_{2m-1,m}(m) = \begin{cases} 1, 2\nmid m, \\ 0, 2\mid m, \end{cases}$$
$$N_{2m,m}(m) = \left\lceil \frac{m+1}{2} \right\rceil, \tag{12}$$
$$N_{2m+1,m}(m) = \begin{cases} \frac{(m+1)^2}{4}, 2\nmid m, \\ \frac{m(m+2)}{4}, 2\mid m. \end{cases}$$

□

Therefore, the probability that the $k$th position of a Zeckendorf representation containing $m$ 1 s locates the digit 1 is

$$\Pr_{L,m}(k) = \frac{N_{L,m}(k)}{N_{L,m}}. \tag{13}$$

In a similar fashion of (10) and (12), we have

$$\Pr_{L,m}(1) = \Pr_{L,m}(L) = \frac{m}{L-m-1},$$
$$\Pr_{L,m}(2) = \Pr_{L,m}(L-1) = \frac{m(L-2m+1)}{(L-m-1)(L-m)},$$
$$\Pr_{2m-1,m}(m) = \begin{cases} 1, 2\nmid m, \\ 0, 2\mid m, \end{cases}$$
$$N_{2m,m} = \begin{cases} \frac{1}{2}, & 2\nmid m, \\ \frac{m}{2m+2}, & 2\mid m, \end{cases} \tag{14}$$
$$N_{2m+1,m} = \begin{cases} \frac{m+1}{2m+4}, & 2\nmid m, \\ \frac{m}{2m+2}, & 2\mid m. \end{cases}$$

Let $L = 30$, $m = 9$, then, the value of $\Pr_{L,m}(k)$ is shown as Figure 1, where the straight line in the midsection means that the probability is constant.

*4.2. Pseudorandom Keystream Generation.* The following procedures could generate a reasonably satisfying keystream $C$:

Suppose that both the sender and the receiver know a pair of keys $(e_1, e_2)$, each of which consists of three integers:

$$\begin{cases} e_1 = (a_n, c_n, \mu_n), \\ e_2 = (a_m, c_m, \mu_m), \end{cases} \tag{15}$$

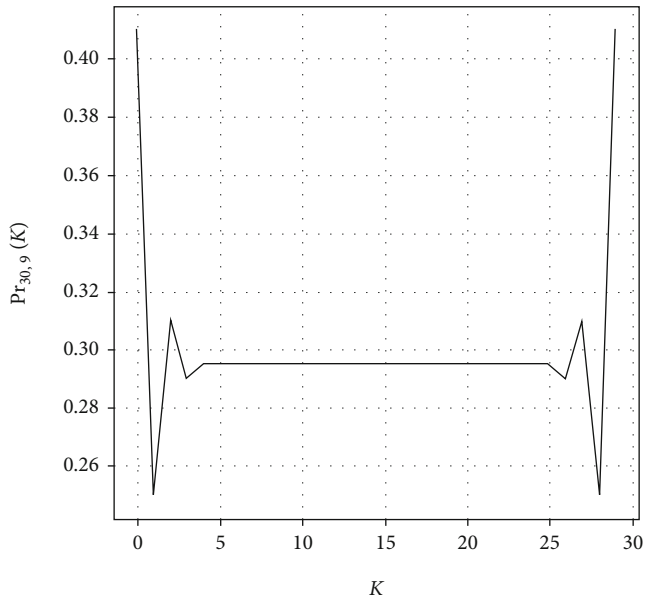



Figure 1: The diagram of $\Pr_{30,9}(k)$.

where $\mu_n$ and $\mu_m$ are primes of the same order of magnitude. There exist pseudorandom integral sequences

$$\begin{cases} \mathbb{N} = (N_1, N_2, \cdots, N_H), \\ \mathbb{M} = (M_1, M_2, \cdots, M_H), \end{cases} \tag{16}$$

with starting values $(N_0, M_0)$ satisfying

$$\begin{cases} N_0 > 0, \quad a_n, c_n < \mu_n, \\ M_0 > 0, \quad a_m, c_m < \mu_m, \end{cases} \tag{17}$$

and each item of $\mathbb{N}$ and $\mathbb{M}$ is obtained by the algorithm described in [26] that

$$\begin{cases} N_{h+1} = (a_n N_h + c_n) \mod \mu_n, & 0 \le h \le H-1, \\ M_{h+1} = (a_m N_h + c_m) \mod \mu_m, & 0 \le h \le H-1, \end{cases} \tag{18}$$

where $H$ is decided by the message length. The integers $N_i$, $M_i$ $(i = 1, 2, \cdots, H)$ are converted into canonical Zeckendorf representations:

$$\begin{cases} u_L(N_i) = (n_1, n_2, \cdots, n_L), \\ u_L(M_i) = (m_1, m_2, \cdots, m_L). \end{cases} \tag{19}$$

Then, we carry out the bitwise logical addition (OR) on their midsection (where the probability is constant) in this way acquiring $C_i$ of length $t$:

$$\mathbf{C}_i = \{c_1, c_2, \cdots, c_t\}, \tag{20}$$

**Require:** A pair of key, $e_1 = (a_n, c_n, \mu_n)$, $e_2 = (a_m, c_m, \mu_m)$; message length, $H$;
**Ensure:** Keystream, $C = \{C_1, C_2, \cdots, C_H\}$;
//initialize;
$N_0 \longleftarrow Random()$;
$M_0 \longleftarrow Random()$;
**for** $h = 1$ to $H$ **do**
/**********
Generate integral sequences
**********/
$N_{h+1} \longleftarrow (a_n * N_h + c_n) \bmod \mu_n$;
$M_{h+1} \longleftarrow (a_m * M_h + c_m) \bmod \mu_m$;
/**********
Convert into Zeckendorf representations
**********/
$S_L(N_h) \longleftarrow \text{Encode}(N_h)$;
$S_L(M_h) \longleftarrow \text{Encode}(M_h)$;
/**********
Intercept midsection
**********/
**for** $u = S_L(N_h), S_L(M_h)$ **do**
//Count the number of 1 in sequence;
$m \longleftarrow u.\text{Count}(1)$;
//Determine the middle-section;
$L_{\text{start}} \longleftarrow m$;
$L_{\text{end}} \longleftarrow L - m + 1$;
//Intercept;
$u \longleftarrow u[L_{\text{start}}, L_{\text{end}}]$;
**end for**
/**********
Bitwise OR
**********/
$s \longleftarrow S_L(N_h)$ OR $S_L(M_h)$
/**********
Take random piece (of length $t$)
**********/
$U_L = (5(L+2) - 8 - [5(L+2)^2 + 4]^{1/2}/10) + 1$;
$t = L - 2U_L + 2$;
$C_h = \{c_1, c_2, \cdots, c_t\} \longleftarrow \text{RandomSelect}(s)$
**end for**
**Return** $C = \{C_1, C_2, \cdots, C_H\}$;

Algorithm 1: Pseudorandom keystream generation algorithm.

where $t$ is given by (see next section for the value of $U_L$)

$$t = L - 2U_L + 2,$$
$$c_j = n_k + m_k, j = 1, 2, \cdots, t\,; k = U_L + j - 1. \tag{21}$$

By juxtaposing $C_i$, we finally obtain the keystream sequence $C$ of length $Ht$:

$$\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_H\}. \tag{22}$$

The sequence $C$ is exactly what we need. Just for the sake of convenience narration, we refer to this kind of pseudorandom keystream generation algorithm as "ZPKG" hereinafter, and the pseudocode is shown as Algorithm 1.

## 5. Randomness Analysis

Let $F_n$ be the greatest Fibonacci number no greater than $N_i$, then the length of the shortest Zeckendorf representation $S_L(N_i)$ will be $L = n - 1$. It can be proved that

$$L = \left\lfloor \log_{\Phi} \sqrt{5} \left( N_i + \frac{1}{2} \right) \right\rfloor - 1, \tag{23}$$

where $\Phi = (1 + \sqrt{5})/2$ denotes the golden ratio. $U_L$ of 1's in $S_L(N_i)$ is most likely to be [15]:

$$U_L = \frac{5(L+2) - 8 - \left[5(L+2)^2 + 4\right]^{1/2}}{10} + 1. \tag{24}$$

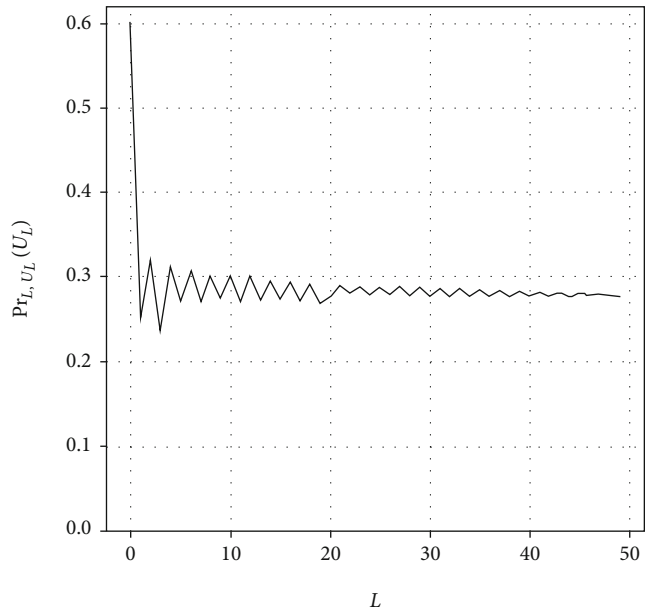



Figure 2: The diagram of $\Pr_{L,U_L}(U_L)$.

Table 1: The probability of the $n$-runs that appear in $C$ ($n = 1, 2, 3, 4$).

| $n = 1$ | | $n = 2$ | | $n = 3$ | | $n = 4$ | |
|---|---|---|---|---|---|---|---|
| Runs | Probability | Runs | Probability | Runs | Probability | Runs | Probability |
| (0) | 0.525 | (00) | 0.201 | (000) | 0.078 | (0000) | 0.030 |
| (1) | 0.475 | (01) | 0.322 | (001) | 0.124 | (0001) | 0.047 |
| | | (10) | 0.322 | (010) | 0.225 | (0010) | 0.089 |
| | | (11) | 0.155 | (011) | 0.096 | (0011) | 0.036 |
| | | | | (100) | 0.124 | (0100) | 0.086 |
| | | | | (101) | 0.197 | (0101) | 0.138 |
| | | | | (110) | 0.096 | (0110) | 0.059 |
| | | | | (111) | 0.060 | (0111) | 0.037 |
| | | | | | | (1000) | 0.047 |
| | | | | | | (1001) | 0.076 |
| | | | | | | (1010) | 0.136 |
| | | | | | | (1011) | 0.060 |
| | | | | | | (1100) | 0.038 |
| | | | | | | (1101) | 0.058 |
| | | | | | | (1110) | 0.037 |
| | | | | | | (1111) | 0.023 |

The probability that a digit '1' lies in the $k$th position in the midsection of $S_L(N_i)$ is

$$p(1) = \Pr_{L,U_L}(U_L) = \frac{N_{L,U_L}(U_L)}{N_{L,U_L}}. \tag{25}$$

Similarly, it holds for $S_L(M_i)$.

It draws from [15] that as $L$ approaches infinity, even in this case, $L > 25$ would be enough, $p(1)$ is expected to approach the limit of $1/(\Phi + 2) \approx 0.2764$ (see Figure 2)

Then, the probability that a '0' lies in the $k$th position in both $S_L(N_i)$ and $S_L(M_i)$ is readily given as below:

$$\Pr(0) = p^2(0) \approx \frac{\Phi^2}{5} \approx 0.524, \tag{26}$$

where

$$p(0) = 1 - p(1) \approx \frac{(\Phi + 1)}{(\Phi + 2)} \approx 0.724. \tag{27}$$

Table 2: NIST-800-22 statistical testing result of ZPKG algorithm.

| Test item | Params 1 | Params 2 | Params 3 | Params 4 | Result |
|---|---|---|---|---|---|
| Approximate entropy | 0.026853 | 0.013829 | 0.068205 | 0.034937 | Pass |
| Block frequency | 0.058378 | 0.870831 | 0.724584 | 0.297646 | Pass |
| Cumulative sums | 0.459642 | 0.069717 | 0.963210 | 0.328997 | Pass |
| FFT | 0.358795 | 0.919848 | 0.081236 | 0.713570 | Pass |
| Frequency | 0.435391 | 0.447255 | 0.888660 | 0.193601 | Pass |
| Linear complexity | 0.186537 | 0.203633 | 0.569565 | 0.232544 | Pass |
| Longest run | 0.359643 | 0.087189 | 0.789913 | 0.250387 | Pass |
| Nonoverlapping template | 0.348045 | 0.680967 | 0.106169 | 0.068529 | Pass |
| Overlapping template | 0.512834 | 0.063236 | 0.020689 | 0.490518 | Pass |
| Random excursions | 0.319514 | 0.181174 | 0.524622 | 0.304589 | Pass |
| Random excursion variant | 0.579380 | 0.177934 | 0.108254 | 0.659874 | Pass |
| Rank | 0.949536 | 0.648387 | 0.862457 | 0.648387 | Pass |
| Runs | 0.340097 | 0.086469 | 0.041369 | 0.027231 | Pass |
| Serial test-1 | 0.407933 | 0.213432 | 0.648688 | 0.814738 | Pass |
| Serial test-2 | 0.462490 | 0.880617 | 0.584615 | 0.512974 | Pass |
| Maurer's universal | 0.026152 | 0.538143 | 0.142680 | 0.600293 | Pass |

In this way, $\Pr(1)$ is given by

$$\Pr(1) = 1 - \Pr(0) = 1 - p^2(0) \approx 0.476. \tag{28}$$

From (26) and (28), it follows that, for $L > 25$, Golomb's first postulate is enough fulfilled.

Golomb's second postulate does not seem, by all accounts, to be so all around fulfilled. In this paper, we are going to assess the probabilities $\Pr(00)$, $\Pr(01)$, $\Pr(10)$, and $\Pr(11)$ of any conceivable pair in $C$.

First, we think about the probabilities $\Pr(00)$, $\Pr(01)$, and $\Pr(10)$ and see that '1' fundamentally preexists or is followed by '0'—the fact that there is no pair (11) at all that is blamed. Therefore, we have

$$p(01) = p(10) = p(1) \approx \frac{1}{(\Phi + 2)}, \tag{29}$$

$$p(00) = 1 - (p(01) + p(10)) = 1 - 2p(1) \approx \frac{\Phi}{(\Phi + 2)}. \tag{30}$$

We can apply some bitwise logical additions (OR or +) to get each pair of $C$:

$$(00) = (00) + (00), \tag{31}$$

(1) $(01) = (01) + (00)$ or $(00) + (10)$ or $(01) + (10)$

(2) $(10) = (10) + (00)$ or $(00) + (10)$ or $(10) + (10)$

(3) $(11) = (10) + (01)$ or $(01) + (10)$

Table 3: Hardware resource depletion comparison.

| | ZPKG | RC4 | LFSR |
|---|---|---|---|
| Logic elements | 1110 | 375 | 205 |
| Registers | 303 | 80 | 34 |
| Pins | 37 | 104 | 21 |
| RAM | 21B | 0B | 20B |
| PLLs | 32 | 30 | 0 |

Next, from (29), (30), and 1-4, we have

$$\begin{aligned} \Pr(00) &= p^2(00) = \frac{1}{5} = 0.2, \\ \Pr(01) = \Pr(10) &= 2p(00)p(01) + p^2(01) \approx \frac{\Phi}{5} \approx 0.324, \\ \Pr(11) &= 2p^2(01) \approx \frac{2}{5\Phi^2} \approx 0.152. \end{aligned} \tag{32}$$

## 6. Experiment

### 6.1. Randomness Test

*6.1.1. Golomb's Postulate Testing.* Given the initial values $e_1 = \{a_n = 29, c_n = 4{,}712{,}321{,}103, \mu_n = 4{,}500{,}000{,}013\}$, $e_2 = \{a_m = 31, c_m = 5{,}666{,}778{,}007, \mu_m = 5{,}127{,}312{,}451\}$, $N_0 = 5{,}123{,}123{,}007$, $M_0 = 4{,}901{,}976{,}445$, $H = 100$. Put them into the formula, we have $L = 47$, $U_l = 13$, and $t = 23$. Therefore, $\Lambda = Ht = 23{,}000$. The probability of $n$-runs is obtained by enumerating the occasions they appear in $C$ and dividing $\Lambda$. Table 1 shows the results. The results of $n = 1$ and $n = 2$ prove that (26), (28), and (32) hold

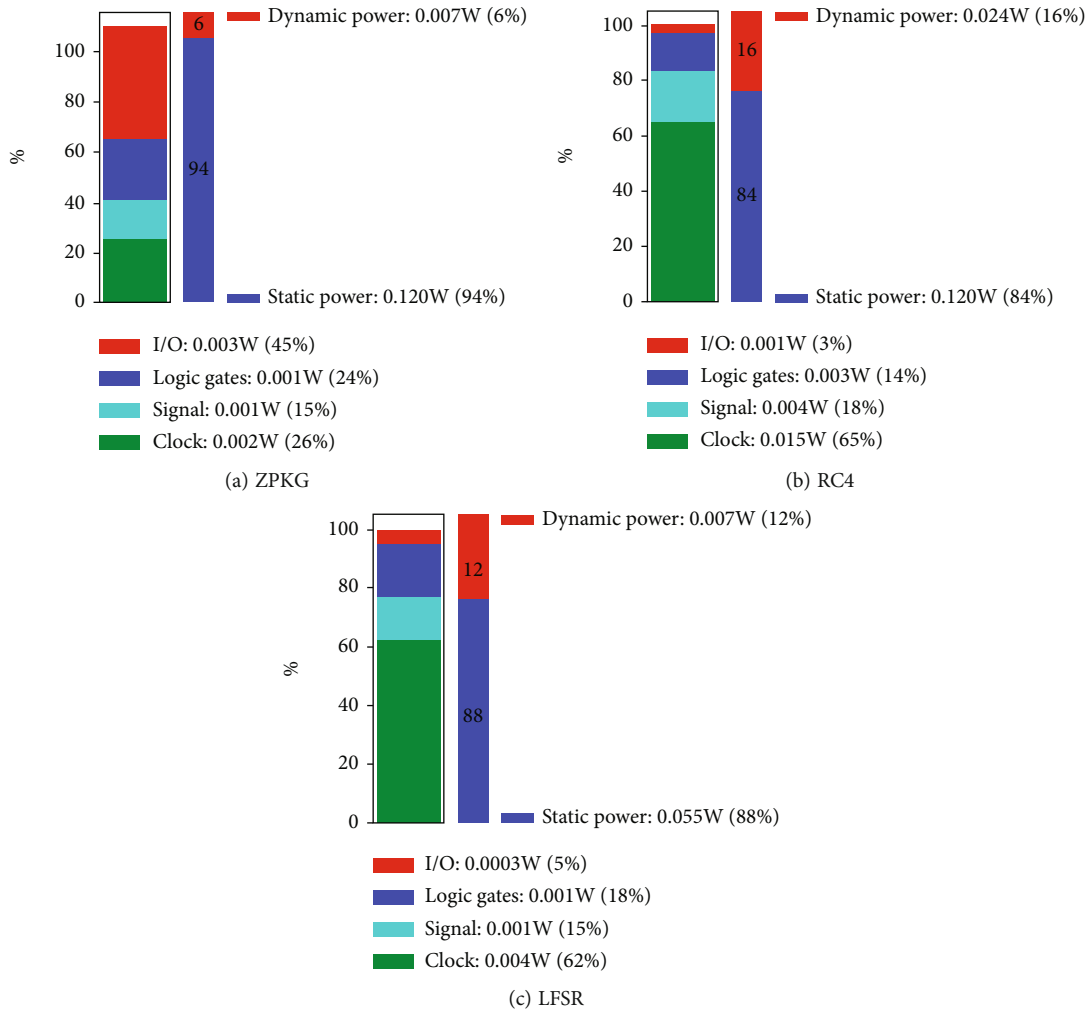



Figure 3: The power dissipation distribution.

for relatively large $L$ ($L = 47$); in other words, the experimental estimations are near theoretical calculation when $L \longrightarrow +\infty$.

We extract a segment of length $\Lambda' = 1{,}000$ randomly from $C$ and then compute the estimation of $\Gamma(\tau) = \sum_{i=1}^{\Lambda} c_i$ $c_{i+\tau}$ for $\tau = 1, 2, \cdots, \Lambda' - 1$, and unquestionably:

$$\begin{aligned} &\Gamma(0) = 0.47\Lambda', \\ &\Gamma(1) = \Gamma\left(\Lambda' - 1\right) = 0.163\Lambda', \\ &0.2\Lambda' \le \Gamma(\tau) \le 0.247\Lambda' \left(\overline{\Gamma}(\tau) = 0.223\Lambda'\right), \\ &\quad 2 \le \tau \le \Lambda' - 2, \end{aligned} \tag{33}$$

where $(\overline{\Gamma}(\tau) = (1/\Lambda')\sum_{\tau=0}^{\Lambda-1}\Gamma(\tau))$.

A few more cases were acquired by alternating the parameters $N_0$, $M_0$, $e_1$, and $e_2$ rendered insignificant differences from the preceding cases, which prove that the ZPKG algorithm satisfies the Golomb randomness postulates.

Table 4: The power dissipation comparison.

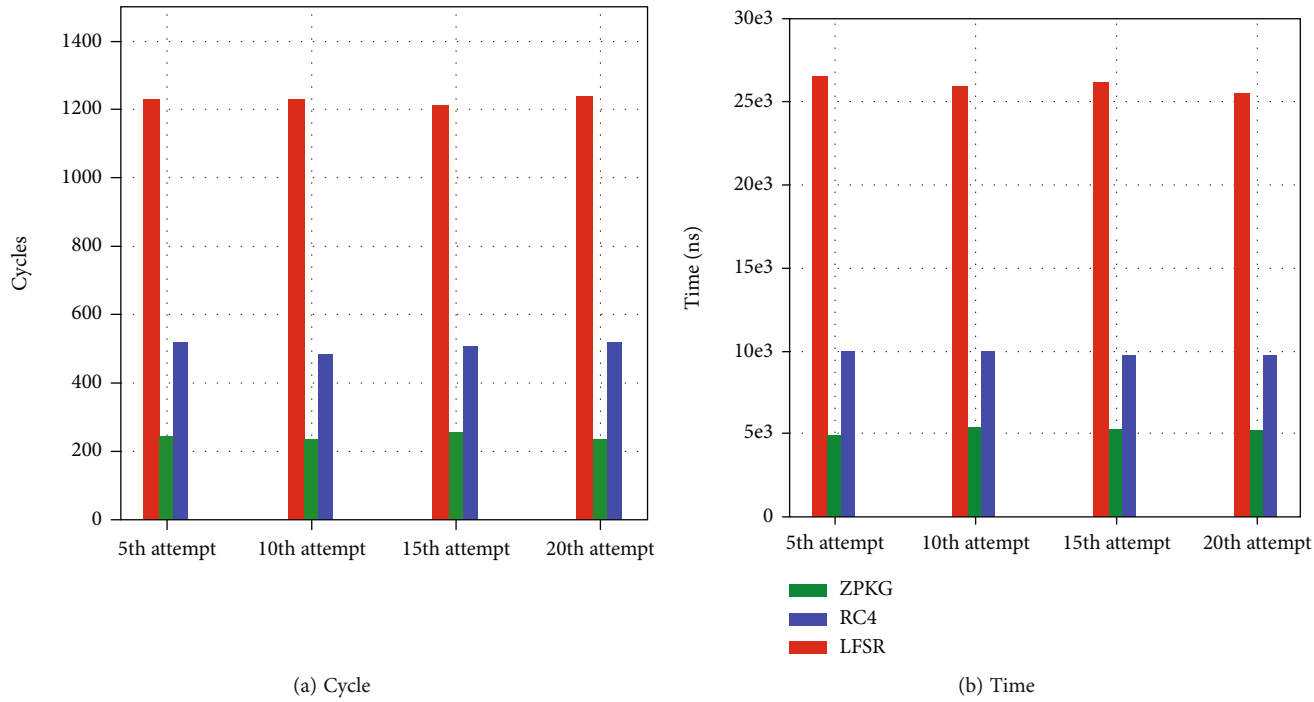
| | ZPKG | RC4 | LFSR |
|---|---|---|---|
| On-chip power | 0.127 W | 0.144 W | 0.062 W |
| TJ * | 26.5 °C | 26.7 °C | 26.5 °C |
| Thermal margin | 58.5 °C | 58.3 °C | 26.2 °C |
| Off-chip power | 0 W | 0 W | 0 W |

*TJ: junction temperature.

Table 5: The key generation time.

| | ZPKG | RC4 | LFSR |
|---|---|---|---|
| Cycles | 233 | 490 | 1275 |
| Time (ns) | 4670 | 9790 | 25500 |

*6.1.2. NIST-800-22 Statistical Testing.* NIST-800-22 [27] is a statistical test suite for random and pseudorandom number generators for cryptographic applications. This test standard was enacted by the Information Technology Laboratory




(a) Cycle (b) Time

Figure 4: Statistics of key generation time and clock cycle of repeated attempts.

(ITL) at the National Institute of Standards and Technology (NIST). The test suite describes 16 statistical tests, including the longest run test, cumulative sums, and the linear complexity test, which are useful in detecting deviations of a binary sequence from randomness. The $p$ value summarizes the strength of the evidence against the null hypothesis in each statistical test. If $p$ value $\geq \alpha$ (level of significance), then the null hypothesis is accepted; i.e., the sequence appears to be random. If $p$ value $< \alpha$, the null hypothesis is rejected; i.e., the sequence appears nonrandom. Typically, $\alpha$ is chosen in the range $[0.001,0.01]$. Common values of $\alpha$ in cryptography are about 0.01 based on the NIST-800-22 test standard.

We configured four sets of initial parameters of $N_0$, $M_0$, $e_1$, and $e_2$.The experimental results of the NIST-800-22 test are shown in Table 2. Table 2 shows that the generated sequences of four sets of parameters passed all the tests. These sequences show good randomness and meet the requirements of the stream cipher.

*6.2. Performance Evaluation.* $m$-sequence based on linear feedback shift register (LFSR) is a widely used keystream generator for its long period, good statistical characteristics, easy to be analyzed by algebraic methods, and adapted for hardware implementation. Another type is word-based stream ciphers, for example, RC4 [28, 29]. RC4 has a variable key length and is based on the word-driven operation using random permutations. Unlike LFSR, RC4 works better with software implementation. RC4 consists of two parts: PRGA algorithm, which is for a pseudorandom number generator, and KSA algorithm, which is for key generation. RC4 is extensively used in the secure sockets protocol/transport layer security (SSL/TLS) and WEP protocols, part of the IEEE802.11 wireless LAN standard.

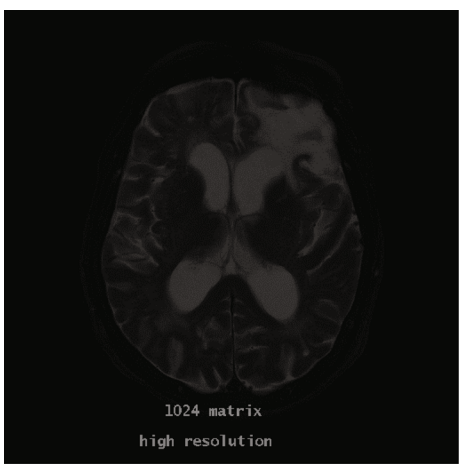
In this section, the ZPKG algorithm, RC4 algorithm, and LFSR algorithm are successfully applied to the encryption of more than 50 images of CVG-UGR test image set, including gray image, biometric image, medical image, and magnetic resonance image (MRI). The experimental simulation platform is FPGA, and the simulation software is ModelSim SE-64 10.4.

*6.2.1. Hardware Depletion.* As shown in Table 3, the ZPKG circuit employs significantly more logic gates and registers than RC4 and LFSR; it occupies the same RAM as LFSR but is slightly higher than RC4. However, the I/O pins of the ZPKG circuit are only 1/3 of RC4, slightly more than that of LFSR. The number of PLLs is similar to that of RC4 but higher than LFSR. This shows that the ZPKG and RC4 circuits have different priorities regarding the disposal of hardware resources; ZPKG and RC4 occupied more hardware resources than LFSR. From a power distribution perspective, they are all primarily based on static power. ZPKG has the highest I/O power other than dynamic power (see Figure 3 and Table 4). However, as shown in Table 3, the number of ZPKG I/O pins is considerably lower than RC4, implying that ZPKG requires frequent I/O operations. In addition, the clock power consumption of RC4 is much higher than that of ZPKG and LFSR, which indicates that RC4 requires more clock cycles, which suggests that the generation speed of a pseudorandom keystream of RC4 is the slowest.

*6.2.2. Key Generation Speed.* Under crystal vibration frequency of 50 MHz, the ZPKG circuit spends 4670 ns to generates a 64-bit pseudorandom keystream, while that of the RC4 is 9790 ns, and that of the LFSR is 25500 ns (see Table 5). In other words,

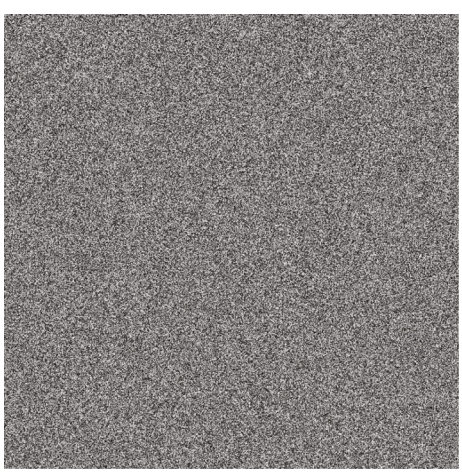
(a)

(b)




(c)

(d)



(e)

(f)

Figure 5: Continued.


(g) (h)

(i) (j)

(k) (l)

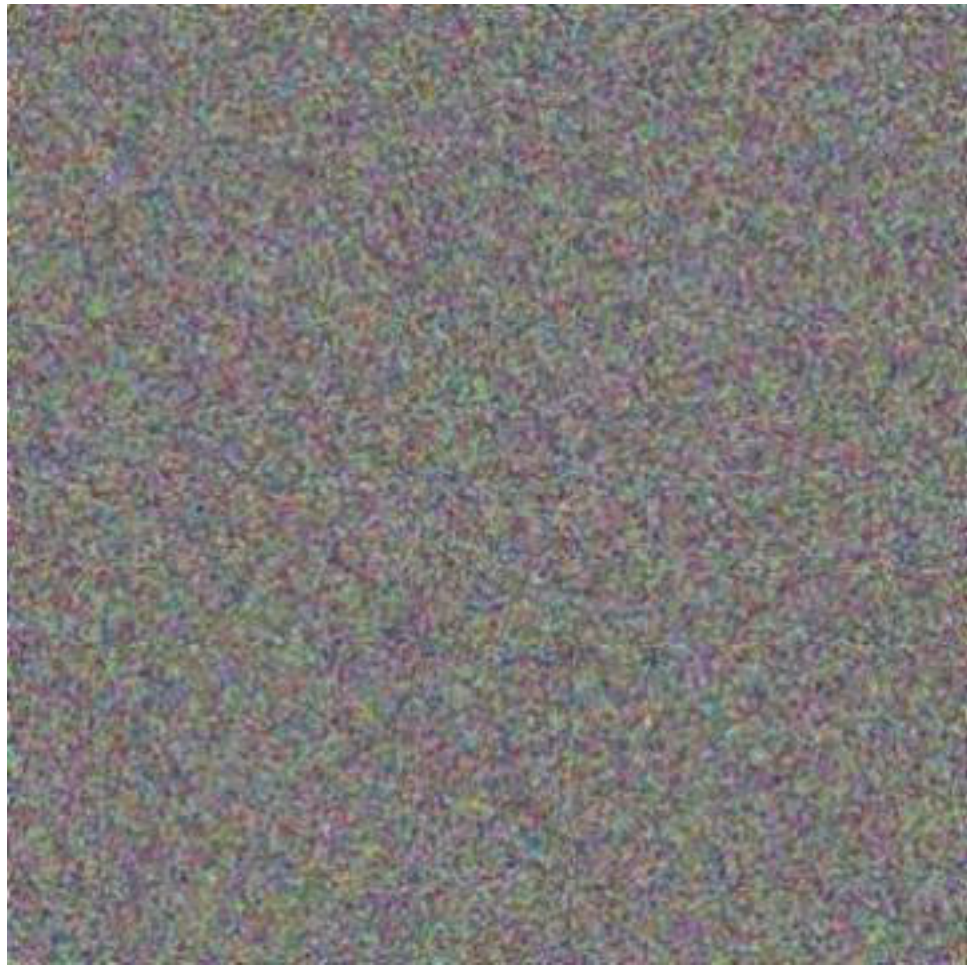
Figure 5: Grayscale images encryption and decryption using ZPKG algorithm. (a–d) Original. (e–h) Encrypted images. (i–l) Decrypted images (SSIM = 1).

(a)

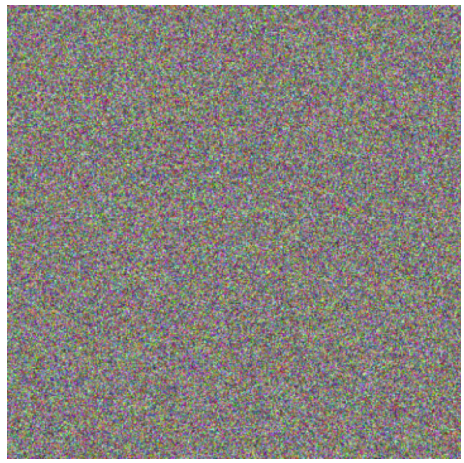
(b)

(c)

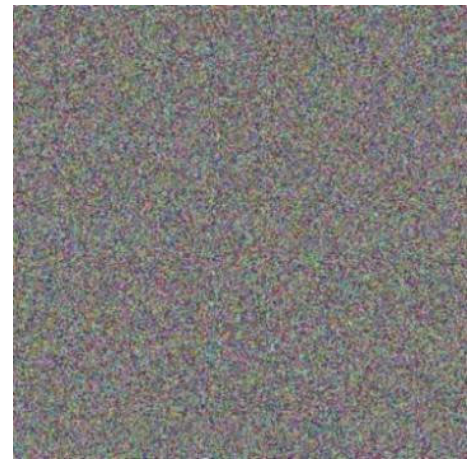
Figure 6: Continued.



(d)



(e)



(f)

Figure 6: Color image encryption using ZPKG algorithm. (a–c) Original. (d–f) Encrypted images.

ZPKG is approximately one time faster than RC4; both ZPKG and RC4 are much slower than LFSR.

Twenty simulations were completed, each generating a pseudorandom 64-bit key. Figure 4 presents the statistics, suggesting that the results are stable with the simulations.

*6.3. Security Analysis.* Security is important not only for the encrypted objects but also for the encryption algorithms themselves.

In what follows, we discuss some security issues of the ZPKG algorithm, such as scrambling effect, statistical

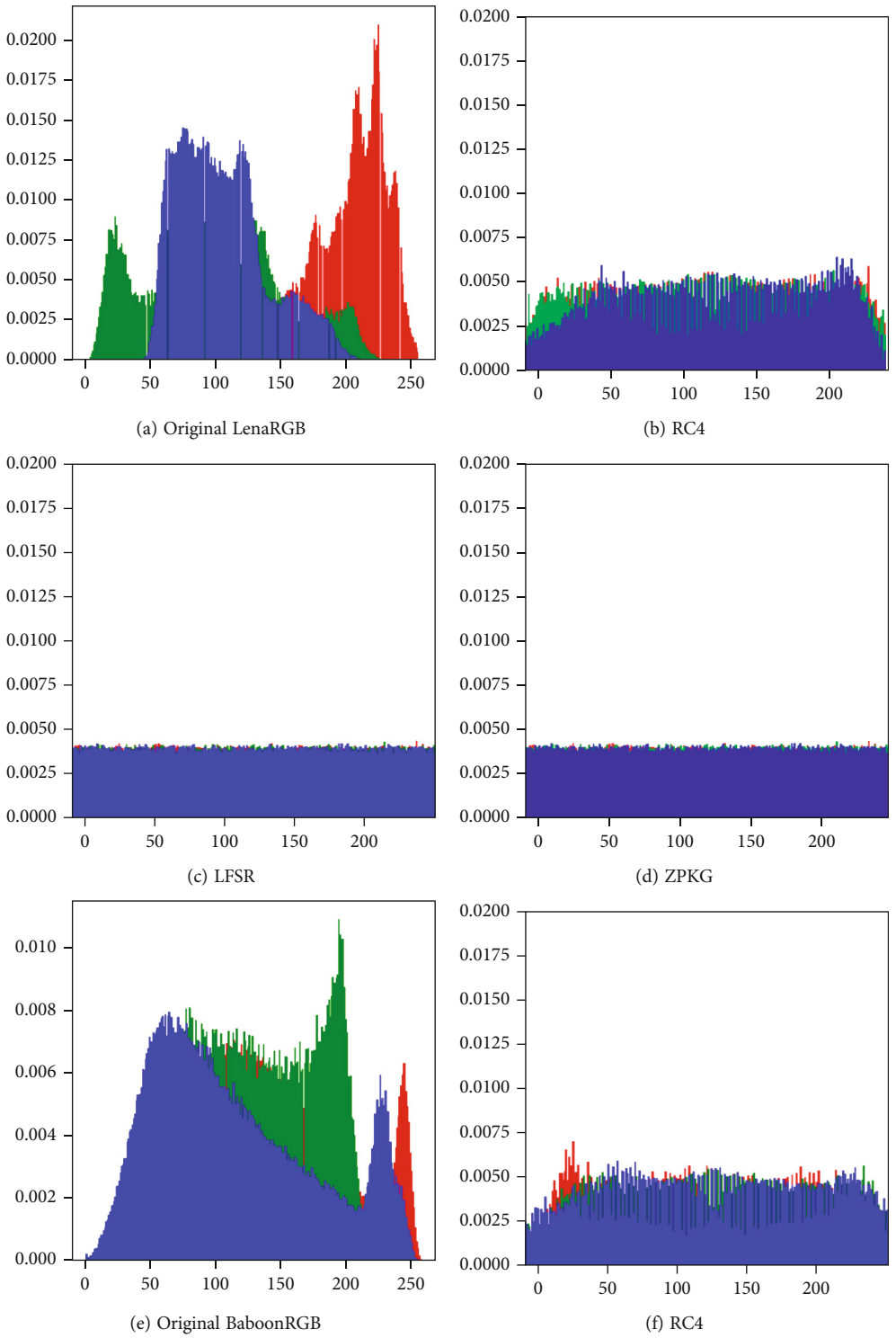



Figure 7: Continued.

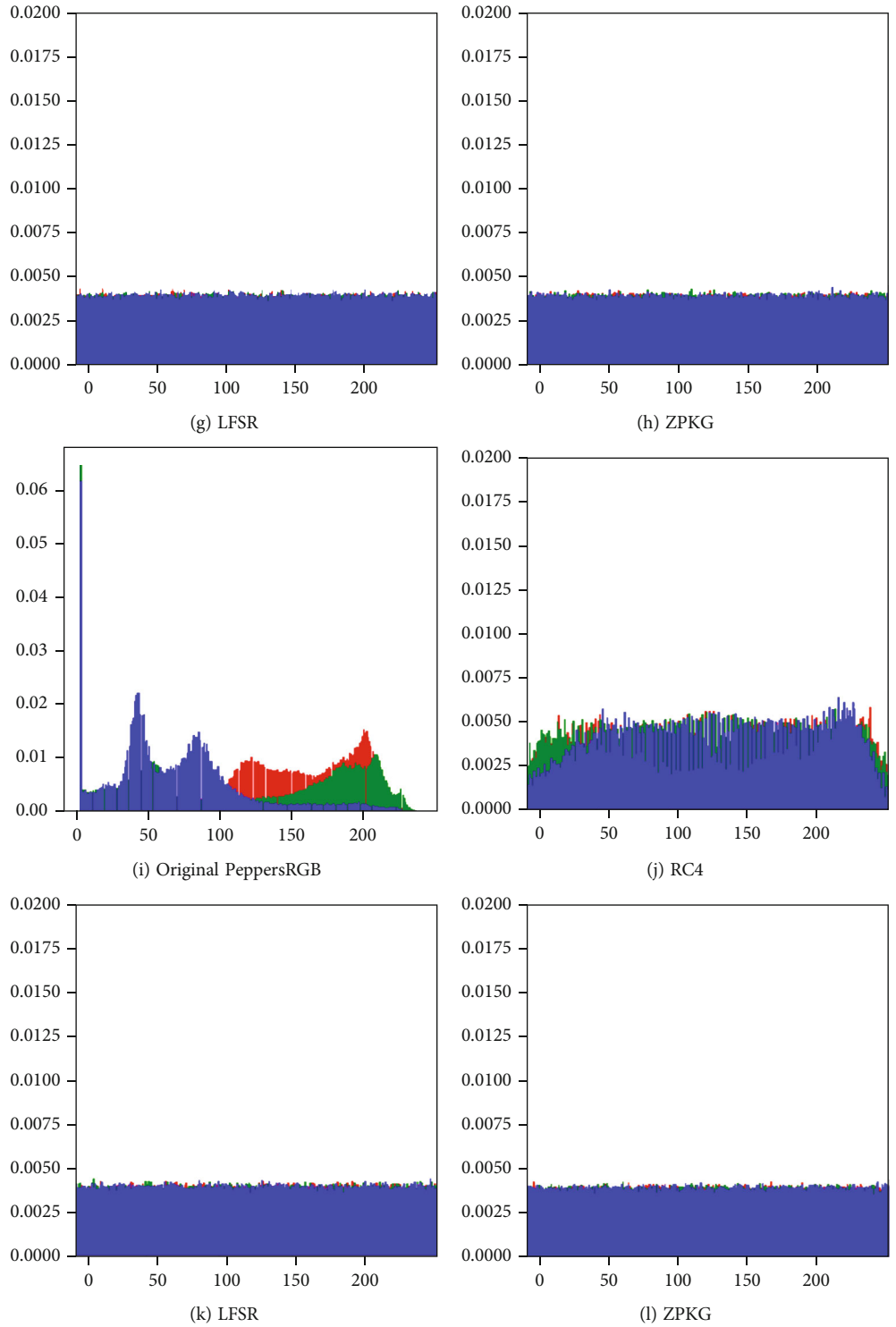
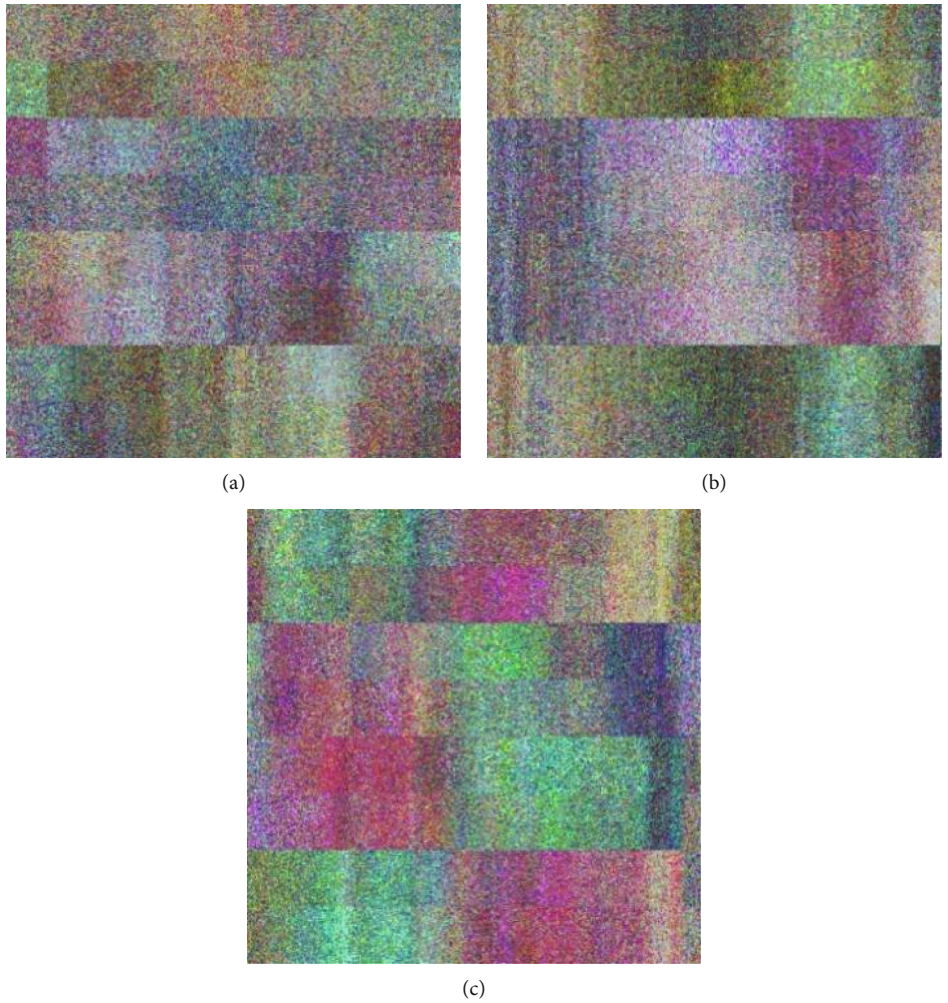



Figure 7: Color image histograms before and after encrypting with ZPKG, RC4, and LFSR.

histogram analysis, key sensitivity testing, robustness, and noise attacks.

*6.3.1. Scrambling Effect.* Figure 5 shows the results of ZPKG algorithm encrypting and decrypting different types of gray-scale images. The encrypted images are visually close to noise images. The structural similarity (SSIM) index is a quantitative assessment method for measuring the similarity between two images [29]; it reflects whether the original images are completely reconstructed or not. A value 1 of



(a) (b) (c)

Figure 8: Color image decryption using ZPKG with 1-bit flipped key. (a) Decrypt LenaRGB. (b) Decrypt BaboonRGB. (c) Decrypt PeppersRGB.

the SSIM index indicates that two measured images are identical. SSIM is computed as

$$\text{SSIM} = \frac{\text{Cov}(X, Y)}{\sqrt{\bar{X}\bar{Y}}}, \tag{34}$$

where

$$\begin{aligned} \text{Cov}(X, Y) &= N \sum_{i=1}^{N} x_i, y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i, \\ \bar{X} &= N \sum_{i=1}^{N} x_i^2 - \left( \sum_{i=1}^{N} x_i \right)^2, \\ Y &= N \sum_{i=1}^{N} y_i^2 - \left( \sum_{i=1}^{N} y_i \right)^2, \end{aligned} \tag{35}$$

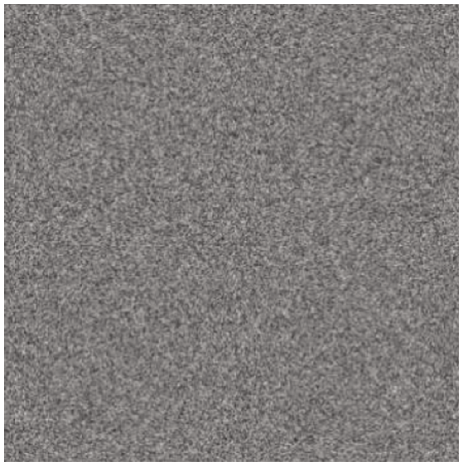
where $x_i$ represents the gray value of the $i$th pixel of the first image, and $y_i$ represents the gray value of the $i$th pixel of the second image.

Figures 5(i)–5(l) show the grayscale images decrypted by ZPKG, and its SSIM index equals to 1, which proves the correctness of the ZPKG algorithm.

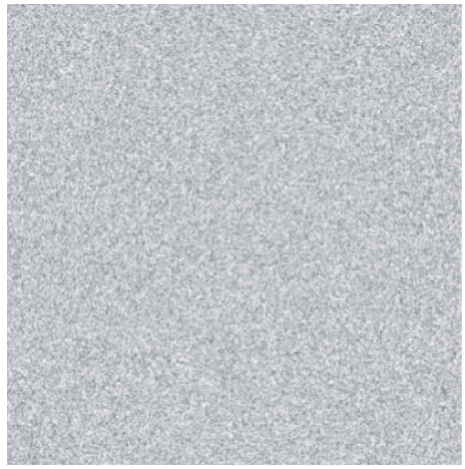
The 3D images such as color images contain several 2D data matrices called 2D components. Color images, for example, contain three color channels called R, G, and B. Each color channel is a 2D component. In this manner, the 3D images can be considered the combination of several 2D images. The 3D image encryption can be accomplished by encrypting all its 2D components one by one. Figure 6 shows three examples of the ZPKG algorithm encrypting color images. The encrypted images (Figures 6(d)–6(f)) look like noise images visually.

*6.3.2. Histogram Analysis.* An image histogram is a graphic representation of the pixel intensity distribution of an image. To overcome statistic attacks, the encrypted image should have a histogram with random behavior and uniform distribution.

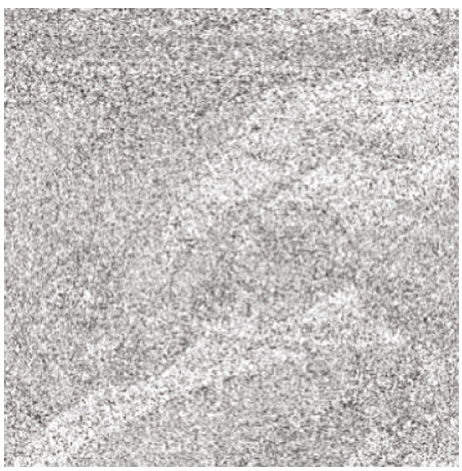
The encrypted image histograms using the ZPKG algorithm, LFSR algorithm, and RC4 algorithm are completely different from the original image (see Figure 7). Nevertheless,



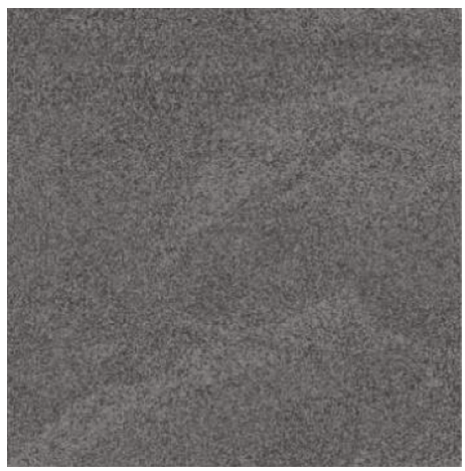
(a) ZPKG encrypt



(b) LFSR encrypt



(c) RC4-KDA encrypt

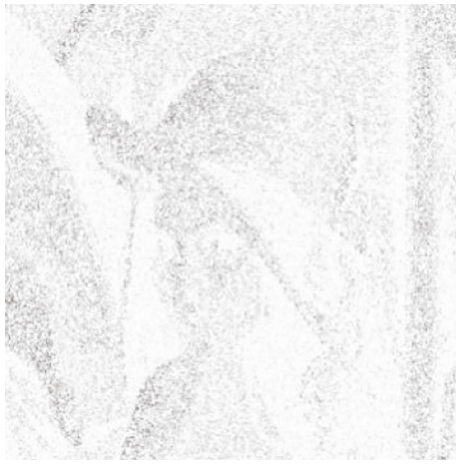


(d) RC4-XOR encrypt

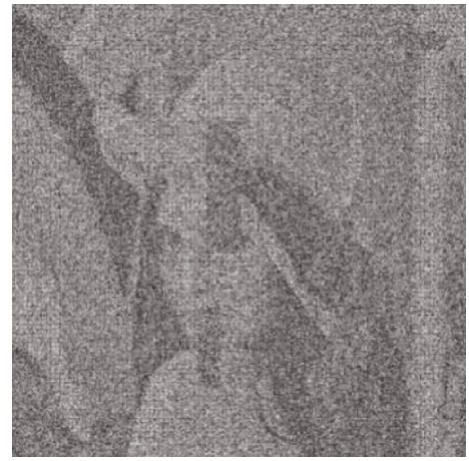
(e) ZPKG decrypt, SSIM = 0.9021

(f) LFSR decrypt, SSIM = 0.5068

Figure 9: Continued.



(g) RC4-KDA decrypt, SSIM = 0.7037



(h) RC4-XOR decrypt, SSIM = 0.4548

Figure 9: Encrypting and decrypting Lena image with noise overlap added.

the encrypted image of RC4 still retains some visual information of the original image, and the intensity distribution of the corresponding histogram is uneven. By contrast, the encrypted images generated by the LFSR and ZPKG algorithms follow a nearly uniform distribution, indicating that the ZPKG algorithm has better performance against statistical attacks than RC4.

*6.3.3. Key Sensitivity.* We flip one bit of keystream, then decrypt images with flipped key. We notice that the decoded picture is in a state of chaos (see Figure 8) and deviates from the original image. The ZPKG algorithm shows good key sensitivity.

*6.3.4. Robustness.* The communication and networking channels are generally in the presence of different types of noise. To test the robustness of the ZPKG algorithm against noise attacks, the salt and pepper noise with density 0.05 is added to the encrypted images. We then try to reconstruct the original image from these noised encrypted images. The SSIM index is used to quantitatively evaluate the similarity between the reconstructed images and the original images. The results are shown in Figure 9. The SSIM index of ZPKG (0.9021) is higher than that of RC4 (0.7037 and 0.4548) and LFSR (0.5068), so we can say ZPKG is more robust when facing noise attacks.

## 7. Conclusion

Stream ciphers cannot really work without the keystream randomness. We proved that the probability of occurrence of the number 1 in the middle part of Zeckendorf coding is constant, which can generate pseudorandom numbers. The pseudorandom numbers generated by the proposed algorithm satisfy the Golomb randomness hypothesis. Experimental results show that our method is three times faster than the RC4 and LSFR algorithms, has no significant difference in hardware occupation, and has high key sensitivity and good resistance to noise attacks and statistical attacks. There is no doubt that our research has its limitations, for example, the lack of theoretical analysis of the characteristics of cryptography, and the lack of comparison with recent research such as chaotic stream ciphering algorithms. We will spend more time devoting on theoretical research on Zeckendorf representation in the future. Furthermore, more experimental investigations are needed to evaluate the performance of ZPKG by comparing it with other stream ciphers such as RC4A, VMPC, and Spritz.

## Data Availability

The CVG-UGR test images data used to support the findings of this study have been deposited in the repository (https://ccia.ugr.es/cvg/dbimagenes/).

## Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

## Acknowledgments

This work is sponsored by the National Natural Science Foundation of China under grant number 61832014.

## References

[1] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.

[2] S. M. Seyedzadeh, B. Norouzi, and S. Mirzakuchaki, "RGB color image encryption based on Choquet fuzzy integral," *Journal of Systems and Software*, vol. 97, pp. 128–139, 2014.

[3] X. Wu, D. Wang, J. Kurths, and H. Kan, "A novel lossless color image encryption scheme using 2D DWT and 6D hyperchaotic system," *Information Sciences*, vol. 349-350, pp. 137–153, 2016.

[4] S. W. Golomb, *Shift Register Sequences*, Holden-Day Inc., San Francisco, 1967.

[5] E. Zeckendorf, "A generalized Fibonacci numeration," *Fibonacci Quarterly*, vol. 10, no. 4, pp. 365–372, 1972.

[6] C. G. Lekkerkerker, "Voorstelling van natuurlijke getallen door een som van getallen van fibonacci," in *Stichting Mathematisch Centrum*, pp. 190–195, Zuivere Wiskunde, 1951.

[7] M. Deza, "On minimal numbers of terms in representation of natural numbers as a sum of Fibonacci numbers," *Fibonacci Quarterly*, vol. 15, no. 3, pp. 237-238, 1977.

[8] M. Edson and L. Q. Zamboni, "On representations of positive integers in the Fibonacci base," *Theoretical Computer Science*, vol. 326, no. 1-3, pp. 241–260, 2004.

[9] E. P. Davlet'yarova, A. A. Zhukova, and A. V. Shutov, "Geometrization of the Fibonacci numeration system, with applications to number theory," *St Petersburg Mathematical Journal*, vol. 25, no. 6, pp. 893–907, 2014.

[10] N. K. Sreelaja and G. A. Vijayalakshmi Pai, "Stream cipher for binary image encryption using ant colony optimization based key generation," *Applied Soft Computing*, vol. 12, no. 9, pp. 2879–2895, 2012.

[11] "RC4 encryption algorithm," http://www.vocal.com.

[12] A. S. Mantin, "Weaknesses in the key scheduling algorithm of RC4," in *Selected Areas in Cryptography. SAC 2001*, S. Vaudenay and A. M. Youssef, Eds., vol. 2259 of Lecture Notes in Computer Science, pp. 1–24, Springer, Berlin, Heidelberg, 2001.

[13] Chung-Ping Wu and C. C. J. Kuo, "Design of integrated multimedia compression and encryption systems," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 828–839, 2005.

[14] K. Khan, "Chaotic cryptography and its applications in telecommunication systems," *Telecommunication Systems*, vol. 52, no. 2, pp. 513-514, 2013.

[15] J. Leroy, M. Rigo, and M. Stipulanti, "Counting the number of non-zero coefficients in rows of generalized Pascal triangles," *Discrete Mathematics*, vol. 340, no. 5, pp. 862–881, 2017.

[16] C. Epifanio, C. Frougny, A. Gabriele, F. Mignosi, and J. Shallit, "Sturmian graphs and integer representations over numeration systems," *Discrete Applied Mathematics*, vol. 160, no. 4-5, pp. 536–547, 2012.

[17] J. Bernat, "Continued fractions and numeration in the Fibonacci base," *Discrete Mathematics*, vol. 306, no. 22, pp. 2828–2850, 2006.

[18] T. P. Berger, M. Minier, and B. Pousse, "Software oriented stream ciphers based upon FCSRs in diversified mode," in *Progress in Cryptology - INDOCRYPT 2009. INDOCRYPT 2009*, B. Roy and N. Sendrier, Eds., vol. 5922 of Lecture Notes in Computer Science, pp. 119–135, Springer, Berlin, Heidelber, 2009.

[19] Z. Lin, "The transformation from the Galois NLFSR to the Fibonacci configuration," in *2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies*, pp. 335–339, Xi'an, China, 2013.

[20] S. S. Mansouri and E. Dubrova, "An improved hardware implementation of the Quark hash function," in *Radio Frequency Identification. RFIDSec 2013*, M. Hutter and J. M. Schmidt, Eds., vol. 8262 of Lecture Notes in Computer Science, pp. 113–127, Springer, Berlin, Heidelberg, 2013.

[21] U. Blocher and M. Dichtl, "Fish: a fast software stream cipher," in *Fast Software Encryption. FSE 1993*, R. Anderson, Ed., vol. 809 of Lecture Notes in Computer Science, pp. 41–44, Springer, Berlin, Heidelberg, 1993.

[22] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 99, pp. 54776–54788, 2020.

[23] N. Deepa, Q. V. Pham, D. C. Nguyen et al., "A Survey on Blockchain for Big Data: Approaches, Opportunities, and Future Directions," 2020, https://arxiv.org/abs/2009.00858.

[24] P. Filipponi and W. Wolfowicz, "A statistical property of non-adjacent ones binary sequences," *Note Recensioni Notizie*, vol. 36, no. 3, pp. 103–106, 1987.

[25] T. Mansour, "Combinatorial identities and inverse binomial coefficients," *Advances in Applied Mathematics*, vol. 28, no. 2, pp. 196–202, 2002.

[26] A. Doğanaksoy, F. Sulak, M. Uğuz, O. Şeker, and Z. Akcengiz, "New statistical randomness tests based on length of runs," *Mathematical Problems in Engineering*, vol. 2015, 14 pages, 2015.

[27] A. Rukhin, J. Soto, J. Nechvatal et al., *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*, 2010.

[28] A. Popov, *RFC 7465: Prohibiting RC4 Cipher Suites*, IETF, 2015, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.682.1589&rep=rep1&type=pdf.

[29] S. L. Miao, J. Y. Zuo, and Y. F. Song, "Design and achieve of FPGA-based RC4 encryption algorithm," *Computer Measurement and Control*, vol. 26, no. 2, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 1999154, 12 pages
https://doi.org/10.1155/2021/1999154

*Research Article*

# Parallel Differential Evolutionary Particle Filtering Algorithm Based on the CUDA Unfolding Cycle

**Kaijie Huang and Jie Cao**

*Lanzhou University of Technology, Lanzhou, China*

Correspondence should be addressed to Kaijie Huang; h18893471259@gmail.com

Received 13 August 2021; Revised 31 August 2021; Accepted 9 September 2021; Published 15 October 2021

Academic Editor: Rajesh Kaluri



Aiming at the problem of low statute efficiency of prefix sum execution during the execution of the parallel differential evolutionary particle filtering algorithm, a filtering algorithm based on the CUDA unfolding cyclic prefix sum is proposed to remove the thread differentiation and thread idleness existing in the parallel prefix sum by unfolding the cyclic method and unfolding the thread bundle method, optimize the cycle, and improve the prefix sum execution efficiency. By introducing the parallel strategy, the differential evolutionary particle filtering algorithm is implemented in parallel and executed on the GPU side using the improved prefix sum computation during the algorithm update. Through big data analysis, the results show that this parallel differential evolutionary particle filtering algorithm with the improved prefix sum statute can effectively improve differential evolutionary particle filtering for nonlinear system states and real-time performance in heterogeneous parallel processing systems.

## 1. Introduction

Particle filtering is a sequential Monte Carlo method that employs particles to approximate the posterior probability density distribution. In [1], the multi-intelligent coevolution mechanism is introduced into particle filtering, and the resampling process is realized by the competition, crossover, mutation, and self-learning among particles, which effectively solve the problem of particle degradation and particle scarcity. Literature [2] compared the filtering accuracy of particle filtering under different search strategies, and the accuracy of the differential evolutionary particle filtering algorithm was improved, but the computational complexity was increased. To address the computational complexity problem, literature [3–5] proposed a GPU-based particle filtering parallel algorithm, which effectively combines the traditional particle filtering algorithm with GPU to make full use of the performance of GPU parallel computing and accelerate the computational speed of the particle filtering algorithm. Literature [6, 7] proposed a GPU-based parallel optimization design and implementation of particle filtering to improve the computational speed of the tracking algorithm. Literature [8–10] designed and implemented a parallel particle swarm optimization algorithm based on CUDA, which uses a large number of GPU threads to accelerate the convergence speed of the whole particle swarm. Parallel statute algorithms are used in the abovementioned literature for parallel particle filtering algorithms to simplify thread operations. Prefixes and algorithms are an important primitive for parallel algorithm programming and are utilized as basic modules for many different algorithms. Compared to serial algorithms, CUDA-based parallel algorithms execute single instruction multithreaded commands, which can perform more operations and improve the efficiency of algorithm execution. However, due to the execution mode and memory access mode of the prefix sum algorithm [11–13], the execution process is prone to thread division and memory access conflict phenomena, which cannot effectively utilize the hardware resources of GPU. Prefix summation contains a large number of repetitive operations, which are simple but inefficient. Segmented prefix summation avoids thread repetition but suffers from serious memory access problems, making the utilization of GPU hardware resources low. Literature [14] introduces additional instructions and

demonstrates their application in the construction of efficient parallel algorithm primitives, such as prefix sums and segmented binary prefix sums. In literature [15], researchers used parallel segmented prefixes to construct data processing and optimize them to improve the overall performance of the algorithm. In literature [16], researchers used GPUs and the practical parallel particle swarm well to solve the problem of singular facility locations, demonstrating that particle swarm optimization is a flexible optimization technique. In literature [17], several tree data structures are studied for the prefix sum problem, providing a variety of practical solutions, all of which obtain a good speedup factor.

To address the problem of thread differentiation in the execution of the differential evolutionary particle filtering parallel algorithm, based on CUDA architecture, this paper proposes a differential evolutionary particle filtering algorithm based on unfolding cyclic prefixes and optimization to remove thread differentiation and reduce the lag caused by judgment and branch prediction, which makes the particle filtering algorithm gradually improve the computational performance.

## 2. Differential Evolutionary Particle Filtering Algorithm

Differential evolutionary algorithm (DE) is a stochastic parallel direct search algorithm, whose basic idea is to start from a certain randomly generated initial population, iterate continuously according to certain operation rules, and according to the fitness value of each individual, keep the good individuals and eliminate the inferior ones, and guide the search process to approach the optimal solution. The algorithm has the advantages of simple structure, easy implementation, no need for gradient information, fewer parameters, etc., and has a variety of different search strategies.

The calculation process of the DE-PF algorithm in this paper is as follows.

*Step 1.* For the initialization step, sampling is performed at time $k = 0$. The resulting $N$ particles $\{x_0^i\}_{i=1}^N$ are used as initial samples, and the distribution of the initial samples is $x_0^i \sim p(x_0)$. All particles have the same initial weights $w_0^i = 1/N$. Repeat iterations for $T = 1, 2, 3, \cdots, N$.

*Step 2.* For the prediction step, set$k = k + 1$, sample particle $\{x_k^i\}_{i=1}^N$ at the current moment through the state transfer model, and calculate the current measure $\{y_k^i\}_{i=1}^N$.

*Step 3.* The weights are calculated and normalized, and after receiving the measurements in Step 2, each particle needs to update the weights according to the likelihood function $p(y_T|x_T^i)$:

$$w_t^i = w_{t-1}^i \cdot p\left(y_t \middle| x_t^i\right). \tag{1}$$

The normalization process makes the sum of the particle weights equal to one, and the normalization process is expressed as

$$w_t^i = \frac{w_t^i}{\sum_N w_t^i}. \tag{2}$$

*Step 4.* For differential evolutionary resampling, we have the following:

(1) $g = 1$. The initial particle of evolution $\{x_k^{g,i}\}_{i=1}^N = \{x_k^i\}_{i=1}^N$

(2) The variation operation is performed on the particle set $\{x_k^{g,i}\}_{i=1}^N$, and then the crossover operation is performed to obtain the candidate particle set $\{\tilde{x}_k^{g,i}\}_{i=1}^N$

(3) The fitness value of the candidate particle set $\{\tilde{x}_k^{g,i}\}_{i=1}^N$ is calculated, and the selection operation is performed, and the resulting particle set is$\{x_k^{g+1,i}\}_{i=1}^N$

(4) If $g < G_{\max}$ and $\sigma > \sigma_{\min}$, then set $g = g + 1$; turn to Step 2; otherwise, go to the next step

*Step 5.* Arrange the particles in descending order.

*Step 6.* Count the number of times each particle is copied, except for its own.

*Step 7.* Calculate the weighted sum of the weights in Step 4, except for its own.

*Step 8.* Eliminate the small particles.

*Step 9.* For the state output step, the optimized set of particles is used as a sample of equal weights $\{x_k^i, w_k^i = N^{-1}\}_{i=1}^N$:

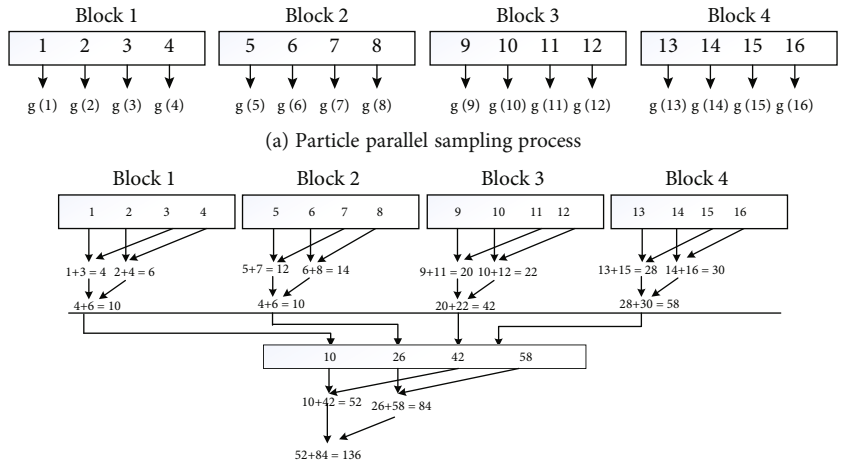
$$\text{Calculated state estimates}: x_k = \sum_{i=1}^N w_k^i, x_k^i. \tag{3}$$

## 3. Improved Parallel Prefix Sum

The parallel algorithm needs to calculate the cumulative distribution function (CDF) of the particles when performing the computation, which is a simple continuous prefix and operation described as follows:

$$y[n] = y[n-1] + x[n], \tag{4}$$

where$n = 0, 1, N-1, y[-1] = 0$, and$N$is the size of the data$y[n]$. The sequential computation is very straightforward and makes parallelization difficult due to the dependencies between output data. For small prefix sum problems, only one thread block is used and recursive multiplication is used to solve the problem. However, parallel particle filtering requires a longer computation of the prefix sum problem when the number of particles $N = 16$, and Figure 1 expresses the same operation on different particles, i.e., the parallel way of prefix summation.




(a) Particle parallel sampling process

(b) Parallel execution process of particle prefix summation

Figure 1: Parallel prefix sum based on unfolding loop improvement.

The parallel prefix sum can be understood as the parallelization of the process of summing all the numbers in an array. In general, the idea of parallelization is based on the binary statute of "trees," as shown in Figures 2 and 3. The implementation of parallel prefix summation can be divided into two types:

*(1) Direct Prefix Sum*. Elements are paired with their direct neighbors to find the sum

*(2) Interleaved Prefix Sums*. Elements are paired according to a given span

Based on the problem of idle threads in the parallel operation of the interleaved prefix sum algorithm, this paper proposes a spread-loop prefix sum method to reduce idle threads and improve the efficiency of prefix sum execution.

By assessing the interleaved prefix sum method, the initial value of stride is half of blockDim.x. When (tid < stride) and then executing subsequent instructions, it means that half of the threads in the first iteration are idle, which wastes GPU computing resources and targets a new problem: idle threads. The performance of the parallel algorithm can still be improved if all of them can be utilized, which is also pending the next step to be optimized and improved.

Expanding loops is a technique that is intended to optimize loops by reducing the frequency of branch occurrences and loop maintenance instructions. In a loop expansion, the body of the loop is written multiple times in the code, rather than just writing the body of the loop once and then using another loop to execute it repeatedly. Any closed loop can have its number of iterations reduced or removed altogether. The number of copies of the loop body is referred to as the loop expansion factor, and the number of iterations becomes the singular number of iterations divided by the loop expansion factor. In sequential arrays, loop expansion is the most efficient way to improve performance when the number of iterations of the loop is known before the loop is performed. Assuming a thread block length of 1024, the threads

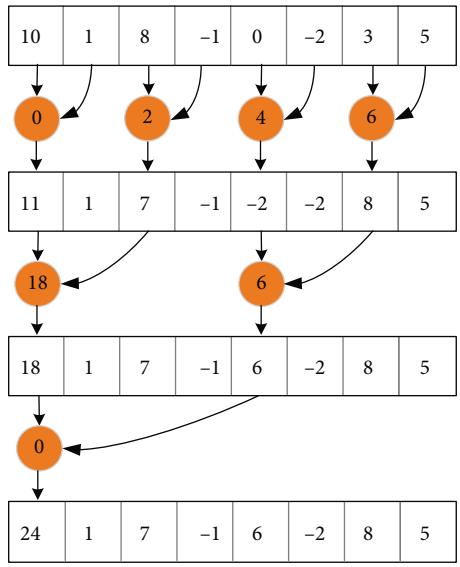



Figure 2: Direct prefix sum.

involved in the computation of the statute iterations at 512, 256, 128, and 64 are distributed in different thread bundles (since each warp can only have 32 threads executing simultaneously); then, there is an order of precedence in the SM execution of these thread bundles, so each step of the statute iteration needs to be synchronized within the block. Only when the statute iterates to 32, 16, 8, 4, and 2, the thread bundle execution they are in is not associated with other thread bundles and no interblock synchronization is needed, while there is implicit synchronization after each instruction in the process of thread bundles in SM, so the intrabundle synchronization problem can be solved, making the global array corresponding to the threads get updated in time without affecting the execution of the next instruction.

In the preceding prefix and computation, each thread block was responsible for one corresponding data block.

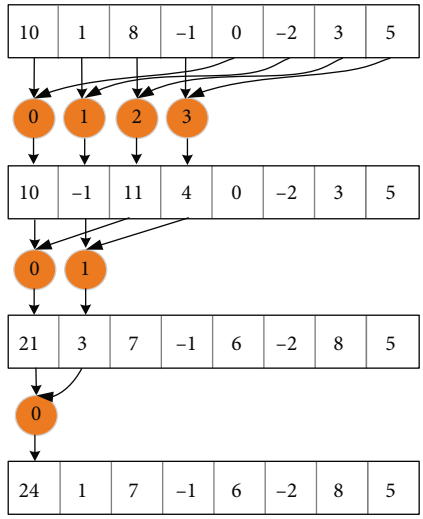



Figure 3: Interleaved prefix sum.

Table 1: Interleaving prefix and expanding loop (expanding factor is 2).

```
if (index + blockDim.x < N)d_data[index] + = d_data
[index + blockDim.x];
__syncthreads();
for (int strize = blockDim.x/2 ; strize > 0 ; strize>> = 1)
{if (tid < strize)
data[tid] + = data[tid + strize];
__syncthreads();
```

Now, each thread block is responsible for prefixing and calculating two data blocks, thus eliminating instruction consumption and increasing the scheduling of more independent instructions to improve performance. The following is a schematic diagram of the prefix sum with expansion factors of 2 and 4. There are three scales of expansion, 2, 4, and 8, where a block computes 2 blocks, 4 blocks, and 8 blocks of data, respectively, adds the adjacent data blocks to the data block corresponding to the current thread block, and then sums them, listed as Tables 1–3.

The parallel prefix and method algorithm strength is low, so the bottleneck in the system may be due to the scheduling instructions. The solution is to expand the for loop. __syncthreads is used for intrablock synchronization. In the statute kernel function, it is used to ensure that all threads in each round have written their local results to global memory before the thread moves to the next round. During the statute, the number of active threads decreases, and when there are less than 32 active threads, we will have only one warp. In a single warp, the execution of instructions follows the SIMD (single instruction multiple data) pattern; i.e., when there are less than 32 active threads, there is no need for synchronization control, and each instruction is followed by an implicit intrabundle synchronization process after each instruction. Therefore, it is necessary to solve the problem of loop control and thread synchronization when

Table 2: Interleaved prefix and loop expansion with a factor of 4.

```
if (index + 3 ∗ blockDim.x < N)
{int a = d_data[index];
int a1 = d_data[index + blockDim.x];
int a2 = d_data[index + 2 ∗ blockDim.x];
int a3 = d_data[index + 3 ∗ blockDim.x];
d_data[index] = (a + a1 + a2 + a3);
```

Table 3: Interleaved prefix and thread bundle expansion.

```
if (tid < 32)
{volatile int ∗ vmen = data;
vmen[tid] + = vmen[tid + 32];
vmen[tid] + = vmen[tid + 16];
vmen[tid] + = vmen[tid + 8];
vmen[tid] + = vmen[tid + 4];
vmen[tid] + = vmen[tid + 2];
vmen[tid] + = vmen[tid + 1];
```

Table 4: Interleaved prefixes and fully expanded.

```
if (index + 7 ∗ blockDim.x < N)
int a = d_data[index];
int a1 = d_data[index + blockDim.x];
int a2 = d_data[index + 2 ∗ blockDim.x];
int a3 = d_data[index + 3 ∗ blockDim.x];
int a4 = d_data[index + 4 ∗ blockDim.x];
int a5 = d_data[index + 5 ∗ blockDim.x];
int a6 = d_data[index + 6 ∗ blockDim.x];
int a7 = d_data[index + 7 ∗ blockDim.x];
d_data[index] = (a + a1 + a2 + a3 + a4 + a5 + a6 + a7);}
```

there is only one thread bundle. Based on this, the thread bundle expansion method with interleaved prefix sum is proposed.

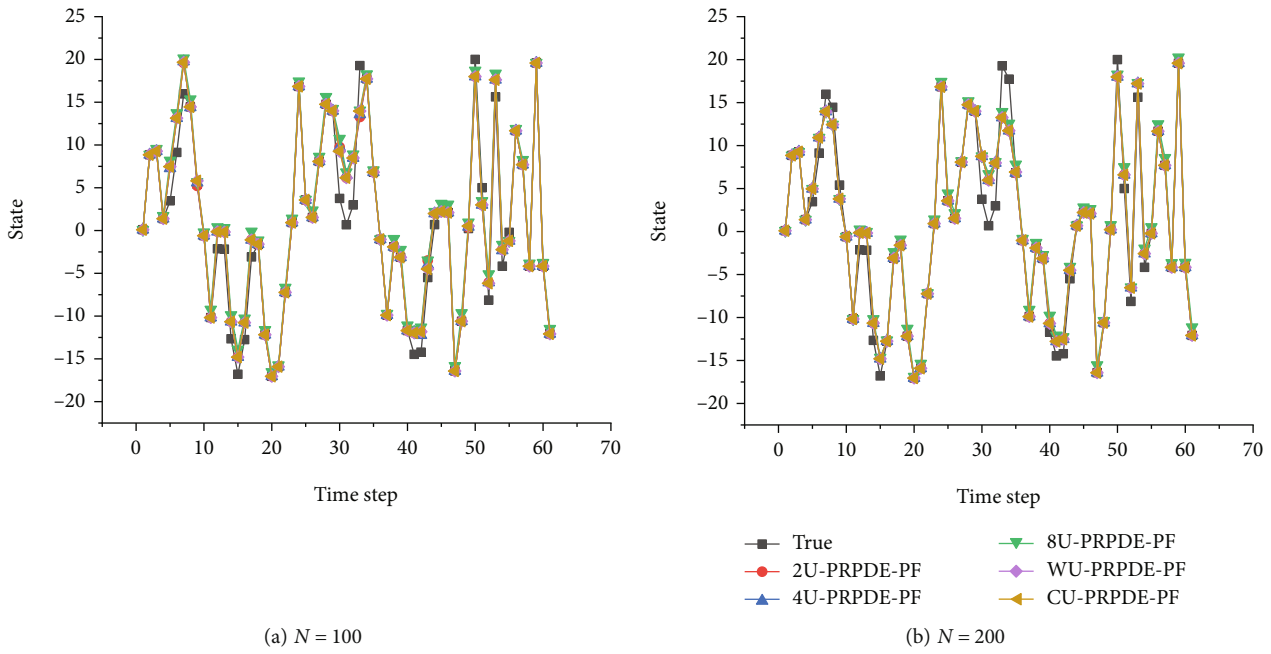
Through the previous experimental analysis, the iterative loop below 32 threads is unfolded. In fact, because of the length limit of the thread block (generally 1024), the number of loops is determined, so the loop can be fully unfolded, i.e., 1024, 512, 256, 128, and 64, and calculated, and the only thing that needs to be noted is that each calculation should be synchronized afterwards. Table 4 shows the pseudocode for a fully expanded loop.

## 4. Experiment and Performance Analysis

In order to verify the basic performance of the parallel algorithm with the improved prefix sum, the performance of the algorithm is simulated using a typical one-dimensional nonlinear system model and compared with the parallel prefix sum based on the unfolding cycle, parallel prefix sum based on the thread unfolding cycle, and parallel prefix sum based on the full unfolding filtering algorithm. The experimental platform includes the Win10 64-bit system, Visual Studio 2013 programming software, and CUDA9.2-based programming framework, where the GPU

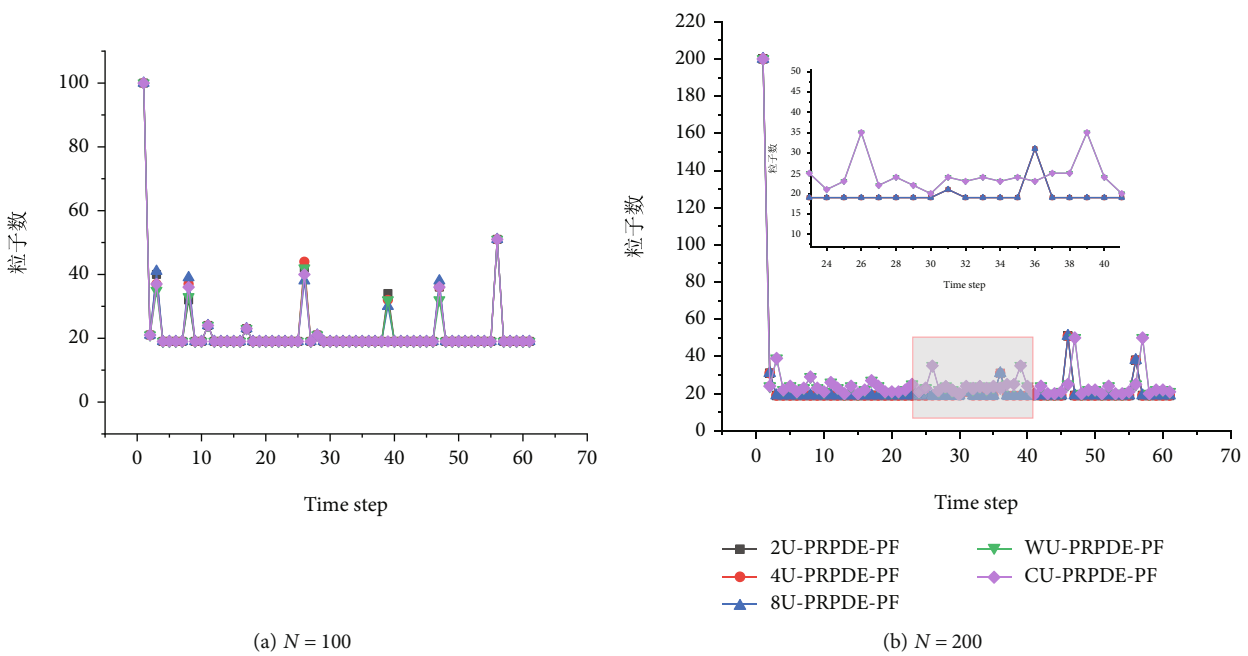
Table 5: The detailed parameters of the experimental platform.

| GPU | | CPU | |
|---|---|---|---|
| GTX1080Ti | | Intel® Core™ i5-4460 | |
| Stream processor unit | 3584 | CPU | Intel® i5-4460 |
| Video memory | 11 GB | Core number | 4 |
| Clock frequency | 1582 MHz | Memory | 8 GB |
| Memory bit width | 352 bits | Clock frequency | 3.2 GHz |



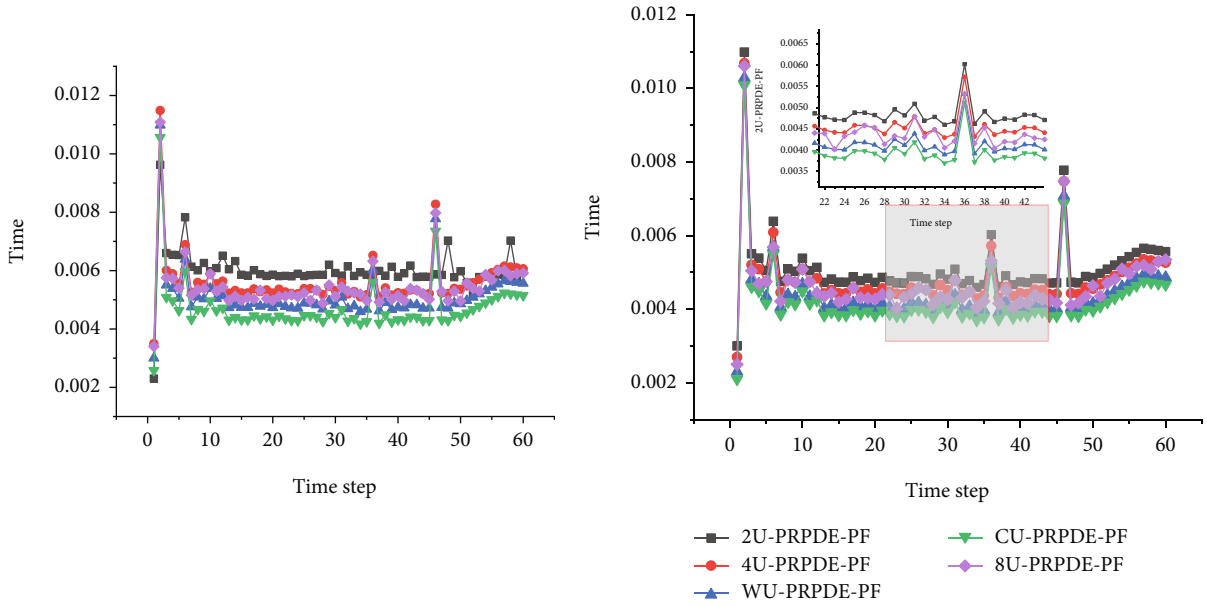

(a) $N = 100$ (b) $N = 200$

Figure 4: State estimation results of five improved algorithms.




(a) $N = 100$ (b) $N = 200$

Figure 5: Particle number curves of the five improved algorithms.




(a) $N = 100$ (b) $N = 200$

Figure 6: Time curves calculated by five algorithms.

Table 6: Comparison of filter calculation time for $N = 200$.

| | 2U-PRPDE-PF | 4U-PRPDE-PF | 8U-PRPDE-PF | WU-PRPDE-PF | CU-PRPDE-PF |
|---|---|---|---|---|---|
| Time (s) | 0.005012 | 0.004815 | 0.00462 | 0.004465 | 0.00425 |

is GTX1080Ti and the CPU is i5-4460. Detailed parameters are listed in Table 5.

The one-dimensional nonlinear system model is as follows:

$$\begin{cases} x_k = 1 + \sin(0.04\pi k) + 0.5x_{k\text{-}1} + u_{k\text{-}1}, \\ y_k = \begin{cases} 0.2x_k^2 + v_k & (1 \le k \le 30), \\ 0.5x_k\text{-}2 + v_k & (30 < k \le T), \end{cases} \end{cases} \tag{5}$$

System noise of the model is $u_{k-1} \sim \Gamma(3, 2)$, total observation time is$T = 60$, crossover probability is$CR = 0.6$of the evolutionary algorithm, and the maximum evolution time is iteration number$G_{\max} = 10$. In this paper, we use the parallel prefix and expansion factors 2, 4, and 8 (2U-PRPDE-PF, 4U-PRPDE-PF, and 8U-PRPDE-PF) based on the expansion loop. In this paper, the comparison experiments are conducted among the three algorithms, i.e., PF, 8U-PRPDE-PF, warp unrolling PRPDE-PF, and complete unrolling PRPDE-PF.

*4.1. Experimental Analysis of Root Mean Square Error.* The algorithm is simulated $R_{MC} = 200$ times by independent Monte Carlo, and the root mean square error of the time is defined as follows:

$$L_k^{\text{RMSE}} = \sqrt{\left(\frac{1}{R_{MC}}\right) \sum_{j=1}^{R_{MC}} \left(x_{k,j} - \bar{x}_{k,j}\right)^2}. \tag{6}$$

$x_{k,j}$ and $\bar{x}_{k,j}$ denote the actual and predicted states at the moment $k$ in the $j$th simulation, respectively. The measurement noise $v_k \sim N(0, 0.001)$. Figure 4 gives a comparison of the momentary root mean square error of the five algorithms for the two settings of the particle number $N = 100$ and $N = 200$.

The performance of the algorithm state estimation is basically the same. It can be seen from Figure 4 that the mean square error of the five improved algorithms, 2U-PRPDE-PF, 4U-PRPDE-PF, 8U-PRPDE-PF, WU-PRPDE-PF, and CU-PRPDE-PF, under the same experimental conditions of the particle number, is reduced relative to the IIPRPDE-PF algorithm, and all of them can guarantee the state estimation ability with the accuracy of the algorithm improved to some extent, indicating that the improved methods improve the state tracking performance of the filtering algorithm to some extent.

*4.2. Particle Distribution and Calculation Time Experiments.* Figures 5 and 6 show the curves of particle number variation and computation time of the improved prefix sum algorithm based on the unfolding cycle for 60 simulation moments. The comparison of the simulation curves in Figure 5 shows that the particle numbers of 2U-PRPDE-PF, 4U-PRPDE-PF, 8U-PRPDE-PF, WU-PRPDE-PF, and CU-PRPDE-PF decrease gradually and adjust the numbers adaptively with time. In Figure 6, at the time, it can be seen that the time of CU-PRPDE-PF is lower than that of the other filters due to performing full unfolding, fully improving the recursive

Table 7: Computation time of the five parallel algorithms.

| $N$ | Computation time (s) | | | | |
|---|---|---|---|---|---|
| | CRPF | Block CRPF | 2U-PRPDE-PF | 4U-PRPDE-PF | Optimized block CRPF |
| 1024 | 0.15862 | 0.1125 | 0.03751 (2.99x) | 0.03564 (3.15x) | 0.1461 |
| 2048 | 0.21301 | 0.1676 | 0.05 (3.35x) | 0.048 (3.49x) | 0.1747 |
| 3200 | 0.26709 | 0.21225 | 0.0625 (3.396x) | 0.06 (3.53x) | 0.24 |
| 4096 | 0.32074 | 0.256 | 0.075 (3.41x) | 0.0692 (3.69x) | 0.291 |
| 6400 | 0.44983 | 0.3672 | 0.096 (3.825x) | 0.0783 (4.689x) | 0.4076 |

Table 8: Six algorithms' running schedule. Unit: ms.

| $N$ | IIPRPDE-PF | 2U-PRPDE-PF | 4U-PRPDE-PF | 8U-PRPDE-PF | WU-PRPDE-PF | CU-PRPDE-PF |
|---|---|---|---|---|---|---|
| $2^{10}$ | 39.2 | 37.51 | 36.35 | 35.64 | 33.56 | 33 |
| $2^{11}$ | 51.7 | 50.592 | 49.011 | 48 | 44.2 | 43 |
| $2^{12}$ | 76.05 | 75.888 | 73.5165 | 69.2 | 65 | 64 |
| $2^{13}$ | 152.35 | 139.536 | 135.1755 | 136.5 | 130.21 | 128 |
| $2^{14}$ | 250.6 | 241.536 | 233.988 | 225.82 | 214.18 | 210 |
| $2^{15}$ | 493.3 | 471.648 | 456.909 | 448.2 | 421.62 | 414 |
| $2^{16}$ | 972 | 930.24 | 901.17 | 883.6 | 830.77 | 816 |
| $2^{17}$ | 1927.85 | 1861.296 | 1803.1305 | 1748.59 | 1647.73 | 1620 |
| $2^{18}$ | 3753.85 | 3607.536 | 3494.8005 | 3410.6 | 3208.42 | 3154 |
| $2^{19}$ | 7526.55 | 7215.888 | 6990.3915 | 6842 | 6432.94 | 6324.8 |
| $2^{20}$ | 15031 | 14394.24 | 13944.42 | 13662 | 12847 | 12631 |

Table 9: Acceleration ratios of the five algorithms relative to IIPRPDE-PF.

| $N$ | 2U-PRPDE-PF | 4U-PRPDE-PF | 8U-PRPDE-PF | WU-PRPDE-PF | CU-PRPDE-PF |
|---|---|---|---|---|---|
| $2^{10}$ | 1.04505 | 1.0784 | 1.09969 | 1.16806 | 1.18788 |
| $2^{11}$ | 1.0219 | 1.05487 | 1.07708 | 1.16968 | 1.20233 |
| $2^{12}$ | 1.00213 | 1.03446 | 1.09899 | 1.17 | 1.18828 |
| $2^{13}$ | 1.09183 | 1.12705 | 1.11612 | 1.17003 | 1.19023 |
| $2^{14}$ | 1.03753 | 1.071 | 1.10973 | 1.17004 | 1.19333 |
| $2^{15}$ | 1.04591 | 1.07965 | 1.10062 | 1.17001 | 1.19155 |
| $2^{16}$ | 1.04489 | 1.0786 | 1.10005 | 1.17 | 1.19118 |
| $2^{17}$ | 1.03576 | 1.06917 | 1.10252 | 1.17 | 1.19003 |
| $2^{18}$ | 1.04056 | 1.07412 | 1.10064 | 1.17 | 1.19019 |
| $2^{19}$ | 1.04305 | 1.0767 | 1.10005 | 1.17 | 1.19001 |
| $2^{20}$ | 1.02042 | 1.07792 | 1.1002 | 1.17 | 1.19 |

loop, increasing the prefix and execution efficiency, i.e., increasing the execution rate of resampling and computation time consumption, while the time of 2U-PRPDE-PF, 4U-PRPDE-PF, 8U-PRPDE-PF, and WU-PRPDE-PF is smaller than that of IIPRPDE-PF with a decreasing trend.

After improving the recursive loop in resampling, the particles are reduced adaptively, and the computation time of the parallel filtering algorithm after all five unfolded loops is relatively reduced and smaller than IIPRPDE-PF. After unfolding the recursive loop within resampling, the overall complexity of the algorithm increases, and the time required for recursive sampling to update the number of particles for calculation in real time is not enough to offset the time saved by the reduction of particles when the number of particles is small, and this situation disappears at the time when the computation time of the parallel differential evolutionary

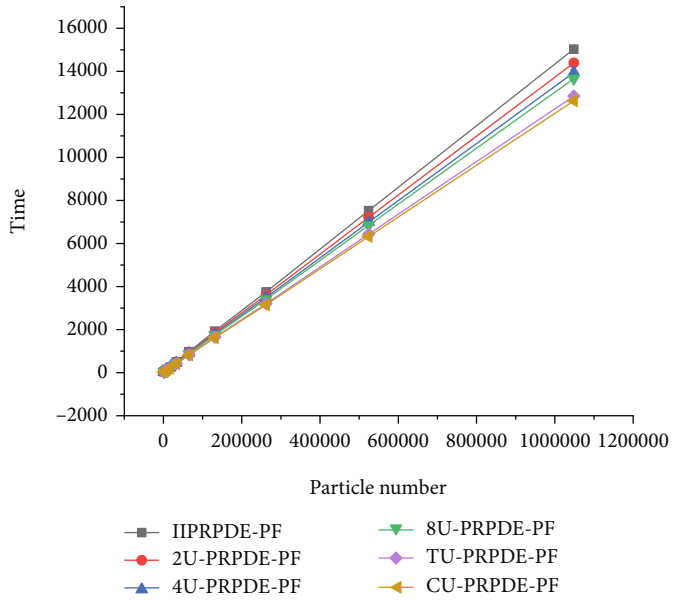



Figure 7: Schedule of the three algorithm runs.

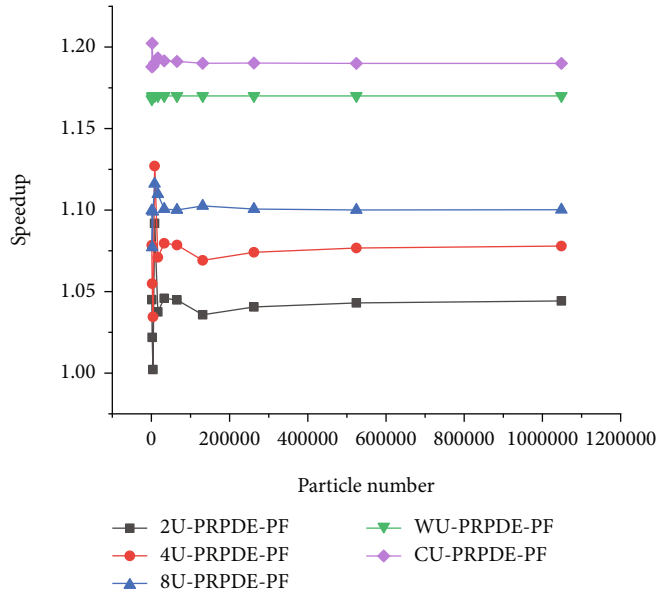



Figure 8: Acceleration ratios of five filtering algorithms with circular unfolding relative to IIPRPDE-PF.

Table 10: The parameters of different GPUs.

| GPU | CUDA cores | Base frequency | Video memory | Memory bit width |
|---|---|---|---|---|
| GTX1080Ti | 3584 | 1.58 GHz | 11 GB | 352 bits |
| GTX960 | 1024 | 1127 MHz | 2 GB | 128 bits |
| GTX950 | 768 | 1024 MHz | 2 GB | 128 bits |
| GTX750Ti | 640 | 1020 MHz | 2 GB | 128 bits |

particle filter for all unfolding loops is smaller than that of the corresponding parallel differential evolutionary particle filter, which also indicates that the PRPDE-PF sampling of unfolding loops improves the computation time more significantly. The filter computation time shown in Table 6 is obtained by 60 independent Monte Carlo experiments and taking the average of the running time of each filter for each experiment, and it can be seen that CU-PRPDE-PF requires

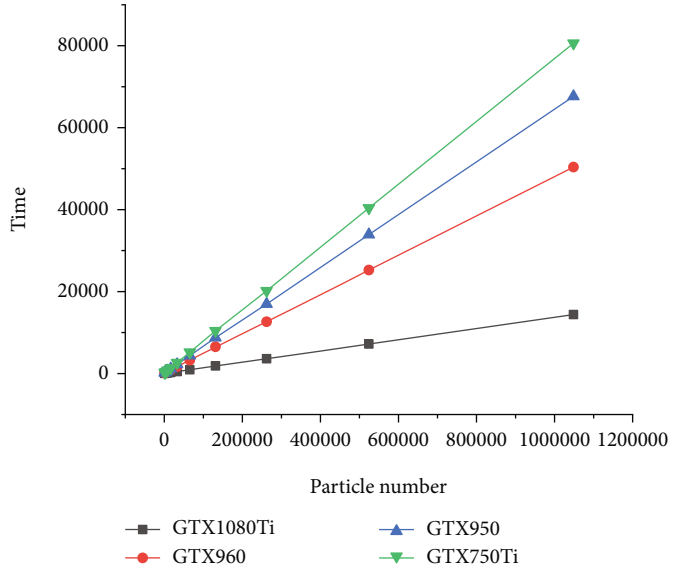



Figure 9: 2U-PRPDE-PF algorithm under different GPUs.

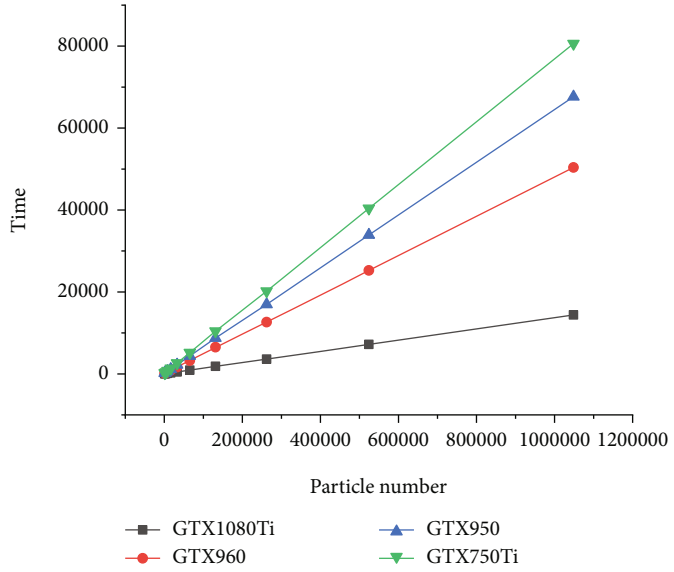



Figure 10: 4U-PRPDE-PF algorithm in different GPUs.

the least computation time. Combining the performance indicators of computational accuracy and computation time of each filter, the fully unfolded loop filter algorithm CU-PRPDE-PF has the least computation time and is the best performance among the five improved filtering algorithms in this paper.

Also, compared with the three smart optimized parallel particle filtering algorithms in the article of Wang et al. [18], the computation time of the improved algorithms (2U-PRPDE-PF, 4U-PRPDE-PF) and the block parallel particle smart optimized particle filtering algorithm in this paper are shown in Table 7, respectively. It can be seen from the table that among the three intelligent optimized parallel algorithms, the block parallel particle filtering algorithm block parallel CRPF has the best performance, followed by the optimized block parallel; the optimization part increases the complexity of the algorithm; and the computational performance decreases compared to the block parallel algorithm. The algorithms proposed in this paper, 2U-PRPDE-PF and 4U-PRPDE-PF algorithms, are compared with the block parallel algorithm, respectively. From Table 7, it is concluded that the improved 2U-PRPDE-PF algorithm in this paper has stronger computational performance than the block parallel CRPF, and a 3.82x acceleration ratio is obtained as the number of particles grows, and the 4U-PRPDE-PF algorithm obtains a speedup ratio of 4.689x as the number of particles increases asymptotically, so the algorithm proposed in this paper has improved

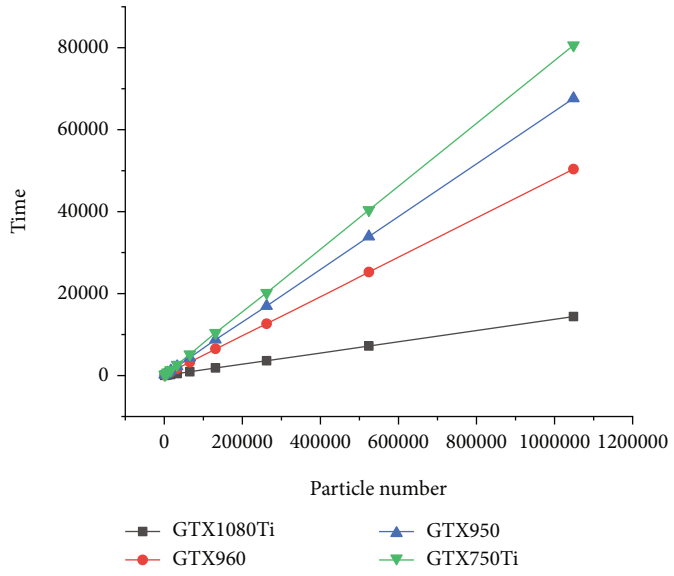



Figure 11: 8U-PRPDE-PF algorithm in different GPUs.

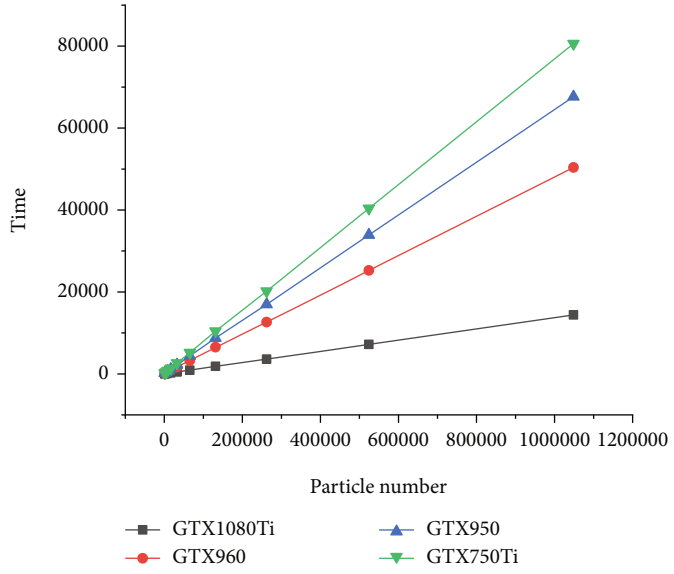



Figure 12: TU-PRPDE-PF algorithm in different GPUs.

performance and can obtain a good speedup ratio compared to the block parallel algorithm.

Comparison of the runs of the five parallel differential evolutionary particle filtering algorithms 2U-PRPDE-PF, 4U-PRPDE-PF, 8U-PRPDE-PF, WU-PRPDE-PF, and CU-PRPDE-PF was based on CUDA cyclic unfolding with improved prefixes and postimprovement in the same GPU case. Tables 8 and 9 show the running schedules and speedup ratios of the five improved algorithms, respectively, and Figures 7 and 8 correspond to Tables 8 and 9, respectively, where the speedup ratio is defined as the value obtained by dividing the running time of the original algorithm by the running time of the improved algorithm under the same particle count condition. Figure 8 shows the values obtained by dividing the operation time of the original algorithm IIPRPDE-PF by the operation times of 2U-PRPDE-PF, 4U-PRPDE-PF, 8U-PRPDE-PF, WU-PRPDE-PF, and CU-PRPDE-PF, respectively. The acceleration ratio of CU-PRPDE-PF is the largest and remains around 1.19 as the number of particles increases, while the acceleration ratios of the other four types of 2U-PRPDE-PF, 4U-PRPDE-PF, 8U-PRPDE-PF, and WU-PRPDE-PF eventually remain at a certain value as the number of particles increases. Under GTX1080Ti, the number of particles is 1024; after the direct unfolding with unfolding factors of 2, 4, and 8, from 39.2 ms to 35.64 ms, it can be seen that the direct cyclic unfolding has a very big impact on the efficiency; this is not only because of saving the extra thread block running but also because the

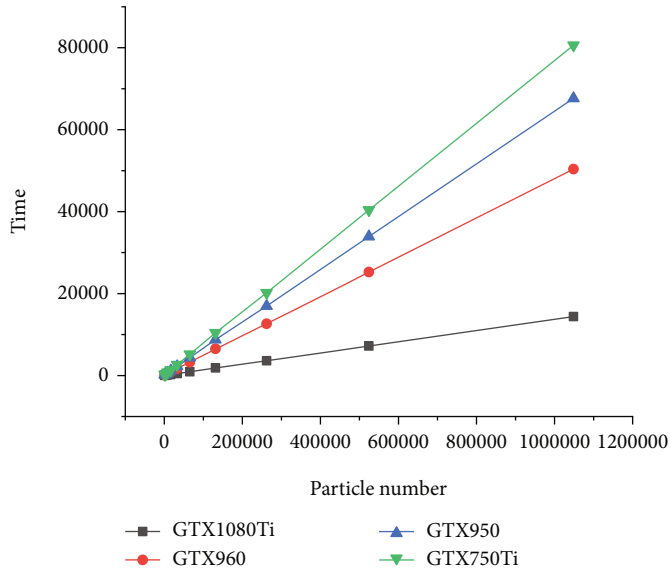



Figure 13: CU-PRPDE-PF algorithm in different GPUs.

improvement has more independent memory loading, and storage operations can produce better performance, with better hidden latency. The algorithm in this paper has been improved incrementally to obtain an overall performance improvement of up to 1.19 times.

*4.3. Real-Time Performance of Algorithms on Different GPUs.* Experimental simulations are performed with the above five improved algorithms based on different GPU conditions. The whole experimental platform includes the Win10 system, Visual Studio 2013 programming software, and CUDA9.2-based programming framework with i5-4460 CPU, listed as Table 10, running the algorithms on four different GPUs with the number of particles from $2^{10}$to $2^{20}$.

The performance experiments of the five improved algorithms in this paper are done based on the same CPU and different GPU conditions. According to the analysis in Figures 9–13, compared with the IIPRPDE-PF algorithm, the five improved algorithms of this paper based on IIPRPDE-PF for cyclic unfolding, 2U-PRPDE-PF, 4U-PRPDE-PF, 8U-PRPDE-PF, WU-PRPDE-PF, and CU-PRPDE-PF, exhibit approximately the same growth rate for different GPUs. It is discussed that the acceleration ratio of the algorithm under different GPU conditions is basically proportional to the computational power of the GPU itself, and the performance of the algorithm is optimal under the experimental environment of GPU GTX1080Ti. In this paper, the performance improvement of GPU computation is limited to the improved prefix and problem. Based on the direct segmentation prefix and the improved differential evolutionary particle filtering algorithm, the overall performance improvement speed of CU-PRPDE-PF is up to 19% relative to the IIPRPDE-PF algorithm, and the performance improvement factor of CU-PRPDE-PF can reach up to 1.45 compared with that of the original PDE-PR algorithm. The main reason for the limited performance improvement of the improved algorithm is just not simply parallel on the GPU and requires complex operations or even contains quite a few logical judgments. However, some performance gains can be achieved by prefixing and incremental improvements.

## 5. Conclusion

In this paper, we propose a CUDA unfolding loop-based state estimation method for differential evolutionary particle filtering to address the problem of inefficient parallel differential evolutionary particle filtering with parallel execution threads and improve the execution efficiency of the prefix sum by unfolding the prefix sum method with an unfolding loop and a thread bundle. The proposed method uses the segmented prefixes after the unfolding loop and the improved resampling and the latest moment of observation to update the proposed distribution of the optimized particle filter in real time and adaptively adjusts the number of particles to be sampled for the particle filter to a smaller number using differential evolutionary resampling. In addition, for the execution of the particle filtering algorithm, the prefix and execution have the problem of inefficient thread execution, and the GPU does not have the branch prediction capability, at every branch it performs, so the algorithm removes the thread bundle differentiation and thread idleness existing in the parallel prefix by unfolding the loop and unfolding the thread bundle method, eliminating the lag caused by the failure of judgment and branch prediction, further improving the overall computational performance. The current CUDA compiler cannot do this optimization for us and requires artificially unfolding the loop within the kernel function, which can greatly improve the kernel performance. The purpose of unfolding the loop in CUDA is twofold: to reduce instruction consumption and to increase the performance by adding more independent scheduling instructions to reduce fragmentation. Simulation results show that the parallel differential evolutionary particle filtering algorithm with this unfolding loop can effectively improve intelligent

optimal particle filtering for nonlinear system states and real-time performance. Finally, experimental simulations show that the algorithm with the improved prefix sum can achieve the best speedup factor of 1.19 relative to the IIPRPPDE-PF algorithm and 1.48 relative to the PDE-PF algorithm under GTX1080Ti, and the experimental data show that the overall performance of the algorithm under different GPUs is proportional to the GPU. The experimental data show that the overall performance of the algorithm under different GPUs is proportional to the GPU computational power, which indicates that the improved algorithm in this paper has universal applicability.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] Y. U. Chunchao, Z. Yang, Z. Xia, X. Yuan, and M. Yan, "Heterogeneous source image alignment based on mutual information and particle swarm algorithm using GPU parallel architecture," *Infrared Technology*, vol. 38, no. 11, pp. 938–946, 2016.

[2] H. W. Li, J. Wang, and H. T. Su, "Improved particle filter based on differential evolution," *Electronics Letters*, vol. 47, no. 19, pp. 1078-1079, 2011.

[3] S. W. X. J. C. Jiazhong and Y. Shengsheng, "GPU-based parallel algorithm for particle filtering," *Journal of Huazhong University of Science and Technology (Natural Science Edition)*, vol. 5, pp. 63–66, 2011.

[4] S. K. Das, C. Mazumdar, and K. Banerjee, "GPU accelerated novel particle filtering method," *Computing*, vol. 96, no. 8, pp. 749–773, 2014.

[5] L. M. Murray, A. Lee, and P. E. Jacob, "Parallel resampling in the particle filter," *Journal of Computational & Graphical Statistics*, vol. 25, no. 3, pp. 789–805, 2016.

[6] V. P. Jilkov and J. Wu, "Efficient GPU-accelerated implementation of particle and particle flow filters for target tracking," *Journal of Advances in Information Fusion*, vol. 10, no. 1, pp. 73–88, 2015.

[7] W. Liu, Z. Meng, and D. Xue, "A multi-feature fusion video target tracking algorithm based on CUDA and particle filtering," *Computer System Applications*, vol. 22, no. 11, pp. 123–128, 2013.

[8] S. Lalwani, H. Sharma, S. C. Satapathy, K. Deep, and J. C. Bansal, "A survey on parallel particle swarm optimization algorithms," *Arabian Journal for Science and Engineering*, vol. 44, no. 4, pp. 2899–2923, 2019.

[9] F. Bourennani, "Cooperative asynchronous parallel particle swarm optimization for large dimensional problems," *International Journal of Applied Metaheuristic Computing*, vol. 10, no. 3, pp. 19–38, 2019.

[10] X. Lin and Y. Wu, "Parameters identification of photovoltaic models using niche-based particle swarm optimization in parallel computing architecture," *Energy*, vol. 196, p. 117054, 2020.

[11] S. S. C. B. Fraser, *Computing Included and Excluded Sums Using Parallel Prefix*, Doctoral dissertation, Massachusetts Institute of Technology, 2020.

[12] H. Tokura, T. Fujita, K. Nakano, Y. Ito, and J. L. Bordim, "Almost optimal column-wise prefix-sum computation on the GPU," *The Journal of Supercomputing*, vol. 74, no. 4, pp. 1510–1521, 2018.

[13] G. Thakur, H. Sohal, and S. Jain, "A novel parallel prefix adder for optimized Radix-2 FFT processor," *Multidimensional Systems and Signal Processing*, vol. 3, 2021.

[14] M. Harris and M. Garland, "Optimizing parallel prefix operations for the Fermi architecture," in *GPU Computing Gems Jade Edition*, pp. 29–38, Morgan Kaufmann, 2012.

[15] A. Pîrjan, "Optimization solutions for the segmented sum algorithmic function," *Wseas Us*, 2013.

[16] E. Wynters, "Parallel particle swarm optimization can solve many optimization problems quickly on GPUs," *Journal of Computing Sciences in Colleges*, vol. 33, no. 6, pp. 114–123, 2018.

[17] "Analysis of dimensionality reduction techniques on big data, a novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU".

[18] J. Wang, J. Cao, W. Li, P. Yu, and K. Huang, "A novel parallel accelerated CRPF algorithm," *Applied Intelligence*, vol. 50, no. 3, pp. 849–859, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8802458, 12 pages
https://doi.org/10.1155/2021/8802458

*Research Article*

# Research on Digital Application of Lighting Design in Public Space Based on Cloud Computing and Data Mining

**Yan Huang and Yongfeng Zhang**

*Sichuan Fine Arts Institute, Chongqing 400053, China*

Correspondence should be addressed to Yan Huang; 2019024@scfai.edu.cn

Received 6 August 2021; Accepted 6 September 2021; Published 15 October 2021

Academic Editor: Thippa Reddy G



Light comes along with everything in the world, which makes the world full of vigor and vitality. The appearance and development of urban landscape art lighting have added a surprise to human beings and have also increased their desire and pursuit for life. This paper describes the characteristics and functions of intelligent lighting control system in landscape public space based on cloud computing and big data, and how to use remote measurement and control technology. The system is composed of intelligent sensing layer, network transport layer, and big data processing application layer, which can collect the location and running status information of urban lighting equipment in real time. The application mode, management mode, and service mode of traditional lighting were changed through a remote platform, and the sustainable development mode of energy saving and environmental protection in the future was realized. The importance of humanized lighting design in landscape public space and the scientificalness of humanized lighting design in landscape public space lighting design are further clarified. This paper hopes to have certain reference value and guiding significance in landscape public space lighting design.

## 1. Introduction

With the rapid development of artificial lighting technology, human beings are presented with a colorful light and magical world. Artificial light source is more important in human life, and the application of artificial light source in architectural space has become an important topic in contemporary architectural design. With the construction of cultural buildings and the improvement of public aesthetic quality, higher requirements are put forward for the lighting of cultural buildings [1]. At present, the construction of smart cities in many developed cities abroad started earlier, the concept of smart cities is more mature, and it has made great progress and been widely recognized in the specific implementation and construction [2]. An excellent landscape public space lighting design is not only for people's lighting needs and functional design but also for people to show a comfortable and beautiful landscape public environment through a complete lighting design scheme, and to increase the enjoyment and interest of urban public space environment.

With the improvement of white light technology, its application in the field of general lighting shows great potential. In order to take the lead in the field of semiconductor lighting, countries all over the world rush to launch their own semiconductor lighting plans [3–4]. How to provide visitors with a better visual experience, how to make visitors deeply understand the designer's design intention by using light, and how to study light by means of display have become important issues faced by designers. At present, the lighting perception of landscape public space in China is monotonous, and the old color is the general feeling of visitors [5]. The designers of landscape public space lighting still stay at the stage of technical indicators, ignoring the visual quality of landscape public space lighting. The difference between Chinese and foreign landscape public space lighting is not only reflected in lighting equipment and

funds, but more importantly, there is a big distance between ideas and technologies.

The most prominent application of information age development in smart city construction is the application of information technology in urban infrastructure construction [6]. We need to evaluate and predict the effect of landscape public space lighting in the early stage with the help of digital technology, so as to check the working performance more conveniently and intuitively. The application of digital technology in architectural space lighting design will also have far-reaching significance. In order to strengthen the research on lighting and lighting system of landscape public space, this paper hopes to make a deeper and more comprehensive research on lighting and lighting system of landscape public space from practice to theory and summarize some methods and theories of lighting and lighting design of landscape public space. It is very gratifying to provide some useful basic information and professional knowledge for the practitioners of lighting design in landscape public space.

## 2. Research Purpose and Innovation

With the continuous improvement of urbanization level, the urban lighting system is expanding, especially in the current energy shortage environment, lighting energy saving has become a common concern of the society. How to save energy and improve the lighting design level of public space is an urgent problem to be solved. Based on this topic, this paper puts forward the research of digital application of public space lighting design based on cloud computing and data mining.

The innovation of this research is to build an intelligent lighting control system for landscape public space based on cloud computing and big data. The system is composed of intelligent sensing layer, network transport layer, and big data processing application layer, which can collect the location and running status information of urban lighting equipment in real time, thereby realizing the digital design of public space lighting.

## 3. Research Status at Home and Abroad

*3.1. Overseas.* The lighting design of landscape public space in foreign countries was earlier. For example, landscape lighting, a relatively advanced lighting method, appeared in Louis XIV cities from 1643 to 1715. Although the development of this kind of landscape lighting is far behind the future lighting means, it provides a strong reference for the future lighting. Entering the 20th century, thanks to the advancement of the second scientific and technological revolution, the lighting industry has also made considerable progress [7]. The design of residential landscape lighting environment is not only to illuminate the environment in the traditional sense but also a derivative and recreation of the existing landscape art. If the design is unreasonable, it will often lead to excessive lighting or abuse of lighting, which will lead to problems such as too tacky and dazzling lighting, unobtrusive scenery, lack of stereoscopic impression, and even more serious light pollution [8]. Alla et al. [9] did not use a lot of light to beautify the plant landscape by using LED light source but made the plant landscape more beautiful than before. Sybilski et al. [10] focused on creating artistic conception of design. Through the use of light, the artistic conception of landscape lighting can be created, and the light and shadow of plant lighting can be handled skillfully to achieve a more poetic atmosphere. The top of the exhibition hall of London Art Museum is designed with diamond hollowing out, and visitors can look up to the blue sky when standing in the middle. The efforts made by foreign designers in saving energy, preventing glare, providing a clear and comfortable angle, avoiding vision disorder or causing fatigue, etc. are worth learning and learning from our designers.

*3.2. Domestic.* Data show that there are some problems in the lighting of landscape public space in China. For example, most of the light sources are lamps for general use in the market, the light source shielding effect is not very good, the light reflection phenomenon is serious, and the display lighting lacks professional design. In contemporary residential landscape design, the beautification function of lighting art is obvious to all. Reasonable lighting design not only meets people's basic needs functionally but also enables people to enjoy visually and makes the residential environment achieve ideal artistic conception [11]. Wong [12] brings a good visual impression by harmonizing with the landscape of the big environment, which makes people have a strong sense of belonging to the living environment and admire the beautiful environment. Wei [13] thinks "Only landscape plant lighting with outstanding style and characteristics can be regarded as excellent works." According to the understanding of plant landscape design style, it puts forward "the logical analysis diagram of plant lighting design thinking." Yanguo et al. [14], combined with the design lighting design project, measured the internal light environment of Tibetan architecture and combined the quantitative analysis of traditional light environment with modern lighting design strategy, so that the final design results not only met the functional illumination requirements of landscape public space but also realized the higher requirements of expressing Tibetan traditional light environment.

*3.3. Lighting Design Level of Landscape Public Space.* The lighting design of landscape public space belongs to a part of landscape lighting in terms of large content. Usually, the design of landscape lighting is actually the realization of functional lighting design. In residential areas, if the plant lighting design is to achieve the light environment effect with functionality, comfort, and artistry, it is necessary to consider and understand the level of lighting design.

*3.4. Lighting Planning of Landscape Public Space.* The description of plant landscape by light is actually a way to express design ideas by using three-dimensional light forms. For living space, the expression of light form should be rigorous. The planning of landscape public space lighting is a detailed technical analysis of living space, and the analysis should be sufficient. According to different plant growth and morphological characteristics, the corresponding plant

lighting methods should be determined to meet the needs of lighting design [15], have a relatively correct grasp of the overall space, outline the lighting layout, and then take the overall space as the basic point, gradually and orderly refine to different functional areas, and refine the plant landscape in the functional areas. Only in this way can we ensure the orderly and unified lighting effect at night, thus forming a good visual impression.

The area is a whole composed of function and environmental landscape. Defining the lighting area is a clear understanding of the design function of humanized lighting and the content elements of the whole environment. In the design and understanding of lighting nodes, we should also pay attention to the fact that lighting nodes are the main landscape center in lighting design. As the key and important content of regional analysis, lighting nodes should have a clear design method for the design of lighting nodes. Lighting structure analysis refers to the design and organization of various elements in the lighting environment. Through the design and organization, the lighting elements are coordinated and unified, and then the lighting space architecture in the lighting space is optimized.

*3.5. Lighting Design of Landscape Public Space.* Lighting design is an extension of lighting planning, which is the early stage planning and pattern division of lighting schemes. Lighting design is aimed at the content established by lighting planning and carries out corresponding analysis and refinement. In a sense, it is also the lighting design scheme that meets the needs of humanized design to the utmost extent. Everyday lighting effect is commonly used, and it is also the most common and common description of landscape public space lighting in residential lighting environment most of the time.

The lighting design of landscape public space focuses on the landscape center and main landscape axis of residential area, and the lighting design of landscape center and main landscape axis of residential area are often the most intuitive and explicit expression of the general characteristics of plant landscape in residential area [16–17], mainly through the contrast of light in illumination. In the lighting design, in order to create different lighting features for the plant landscape of the large living environment, the use of contrast technique is the most effective means to highlight the lighting effect [18]. For example, in order to form an affiliation and set the contrast between subject and individual, the landscape center is the main body of lighting design, and other landscape lighting is the subsidiary part of this main body. In this way, the differences between subject and individual in landscape environment can be effectively displayed through lighting design.

*3.6. Program of Humanized Lighting Design for Plant Landscape in Residential Area.* The design of landscape public space lighting scheme is a combination of various activities. In the specific scheme design, it should be carried out in turn according to reasonable steps and procedures to complete the design task more scientifically [19–20]. After receiving the design task, it is necessary to comprehensively, systematically, and carefully analyze the design task book and various materials provided by the design entrusting party, so as to better carry out the next design work and have more exact requirements and design specifications for the design content.

## 4. Research Technique

To sum up, it is possible to establish intelligent lighting monitoring and management system based on the Internet of Things technology, such as big data processing, cloud computing, platform management, and mobile application mode. Therefore, building a lighting monitoring and management system based on the combination of Internet of Things, big data, and cloud services has become one of the inevitable measures for urban ecological and economic transformation. In this chapter, Apriori association rule algorithm based on cloud computing platform and spatial distribution descriptive analysis method are used to design landscape public space lighting digitally. The method flow is shown in Figure 1.

*4.1. Algorithm Design of Apriori Association Rules Based on Cloud Computing Platform.* Based on the idea of dividing distributed database, the database $D$ is divided into $D_1, D_2, \cdots D_m$, among $D_1 \cup D_2 \cup \cdots \cup D_m = D, D_1 \cap D_2 \cap \cdots \cap D_m = \phi$; after division, each data block is sent to the corresponding server node, and each data node processes one block, so that multiple nodes can work in parallel.

Apriori algorithm solves the frequent $k$-item sets for each node and obtains the local frequent item sets contained in the corresponding data. Apriori uses the iterative method layer by layer, that is, searching $(k+1)$-item sets with $k$-item sets. $L_k - 1$ is used as a connection to generate a set $C_k$ of candidate $k$-item sets and get $k$-item sets $L_k$. This step is completed by each task independently. Finally, the global end performs the threshold operation of $L_k$ obtained by $C_k$. Finally, the global control end uses the following formula [21]:

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)}. \quad (1)$$

The confidence of association rules is calculated, where $\text{support\_count}(A \cup B)$ is the number of things containing item sets $(A \cup B)$, and its value is the support count of frequent item sets obtained by the global controller. $\text{support\_count}(A)$ is the number of things containing item set $A$, and its value should be the accumulation of the number of things containing $A$ in each server node.

*4.2. Descriptive Analysis Method of Spatial Distribution.* Descriptive analysis of spatial distribution is a global description of the spatial distribution characteristics of a group of discrete points from the macro level, based on the spatial position and nonspatial attributes of the discrete points, using specific spatial statistical indicators to find a certain rule [22–23]. Statistical indicators of descriptive analysis of spatial distribution are mainly divided into two categories: concentrated trend statistical indicators and discrete trend statistical indicators.

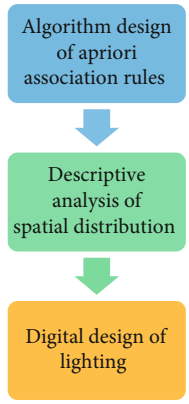



Figure 1: Method flow chart.

The concentrated trend statistical index mainly refers to various center positions of a set of point sets, among which the mean center index in spatial distribution description analysis refers to the average value of $x$ coordinate value and $y$ coordinate value of a set of discrete points, which is the center position or average position of the set of discrete point sets. In some cases, this index is called the center of gravity, which indicates the distribution state of discrete points.

We assume that there are $n$ cases of urban management problems in the study area in a certain time period, in which the plane coordinate point of the $i$th case can be expressed as $(x_i, y_i)$; then, the mean center position $(x, y)$ in this area in this time period is the arithmetic mean value of the coordinates of $n$ cases [24], and its definition is expressed by the following formula [25]:

$$(x, y) = \left( \sum_{i=1}^{n} x_i/n, \sum_{i=1}^{n} y_i/n \right), \tag{2}$$

where $n$ represents the total number of case points in this area in this time period.

Among them, the standard distance refers to the deviation of the actual occurrence position of the case from the central position, and its expression is shown in the following formula:

$$d = \left( \sum_{i=1}^{n} (x_i - x)^2 + \left( \sum_{i=1}^{n} (y_i - y) \right) \right) /n. \tag{3}$$

The standard distance $sd$ actually reflects the average deviation of each point position relative to the center position of the spatial mean. In the actual research process, we can compare and analyze the average distance of the same type of cases in different case types or different time periods, so as to investigate the dispersion degree of the spatial distribution of relevant cases relative to the mean center position.

The standard deviation ellipse index is described by the center position, ellipse major and minor axes, and ellipse orientation. Among them, the center position can be any center index that gathers trend statistics, such as mean center or median center. In the research, the mean center is used for aggregation trend statistics, and the standard deviation of $(x, y)$ coordinate value is calculated first, as shown in the following formula:

$$\begin{aligned} sdx &= \left( \sum_{i=1}^{n} (x_i - x)^2 \right) /n, \\ sdy &= \left( \sum_{i=1}^{n} (y_i - y)^2 \right) /n. \end{aligned} \tag{4}$$

Then, the orientation of the standard deviation ellipse is calculated, which is an angle rotated relative to the true north direction, so that the sum of the distances from all the points in the point set to the major and minor axes of the ellipse is the shortest.

Rotation angle is $\tan\theta = (A + B)/C$, where $A, B$, and $C$ are shown in the following formula:

$$\begin{aligned} A &= \left( \sum_{i=1}^{n} sdx_i^2 - \sum_{i=1}^{n} sdy_i^2 \right) = \left( \sum_{i=1}^{n} sdx_i^2 - \sum_{i=1}^{n} sdy_i^2 \right)^2 + 4 \left( \sum_{i=1}^{n} sdx_i sdy_i \right)^2, \\ C &= 2 \sum_{i=1}^{n} sdx_i sdy_i. \end{aligned} \tag{5}$$

Finally, the standard deviation along the $x$ axis and $y$ axis is calculated, as shown in the following formula:

$$\begin{aligned} \sigma_x &= \left( \sum_{i=1}^{n} (sdx_i \cos\theta - sdy_i \sin\theta)^2 \right) /n, \\ \sigma_y &= \left( \sum_{i=1}^{n} (sdx_i \sin\theta - sdy_i \cos\theta)^2 \right) /n. \end{aligned} \tag{6}$$

To sum up, by using the statistics such as the mean center of concentrated trend indicators, standard distance, and standard deviation ellipse of discrete trend indicators, we can quantitatively analyze the data of digital urban management cases from a macro perspective and preliminarily grasp the overall spatial distribution pattern of cases.

### 4.3. Digital Design of Public Space Lighting

*4.3.1. Selection of Light Source and Lighting Fixture.* In the practical process of outdoor lighting setting, we use a lot of lights for artistic treatment of night scenes in buildings, which will inevitably form some dazzling lights. If these dazzling lights are not properly handled, a new kind of pollution may occur. If dazzling lights affect motor vehicle drivers, potential safety hazards may arise. Therefore, we should choose the lamp position, projection angle, and projection direction reasonably. It is the first principle to determine the lamp position according to the lamp efficiency. We also require the lamp position on the facade to be concealed, so as not to destroy the day-to-day effect of the building. In

addition, the hidden light source gives people a sense of trance and has special effects.

All kinds of lighting sources and lamps have become the most basic and core carrier of landscape public space lighting, and they are also one of the contents that we need to pay special attention to in the process of landscape lighting design. Therefore, when designing space lighting, we should fully consider the system attribute and importance of each element, pay attention to the influence degree between lighting elements, and coordinate the relationship between each element. The landscape of residential area is Chinese or Western, natural or regular, and the lighting design should show tranquility or style, elegance or magnificence. Different types of residential area landscape need different types of waterscape lighting design to cooperate with it. When designing the interior lighting environment of the civic center, we should start from the public's psychology, avoid the high-brightness and complex changing space, tap the spiritual needs of the citizens, extract the unique space atmosphere of each space, and cooperate with the changes of architectural space and indoor materials and colors to create a comfortable and harmonious indoor lighting environment.

MapReduce is a new parallel programming system invented by Google Inc. in recent years. It puts parallelization, fault tolerance, data distribution, load balancing, etc. in one database, and all data operations of the system are summarized into two steps: Map stage and Reduce stage. Map function and Reduce function only need to be defined in all operation processing job programs submitted by programmers to MapReduce, and the MapReduce system can automatically initialize the job into multiple same Map tasks and Reduce tasks, read different input data blocks, and call the Map functions and Reduce functions for processing according to information such as the size of input data and the configuration of the job.

The working mode of MapReduce is as follows, Map is responsible for decomposing tasks and Reduce is responsible for merging decomposed tasks. The work flow of MapReduce is shown in Figure 2.

By determining the brightness contrast, the lighting relationship between the plant itself and its environment is made clear, that is, the rationality of the selection and arrangement of lamps and light sources is made clear. In terms of actual lighting effect, the visual feeling produced by reflection to our eyes is also different. In order to verify the rationality of this inference, the author conducted an experiment with lighting software (Figure 3).

With the emergence of light projection technology, LCD liquid crystal projector, optical-processing DLP digital projector, and CRT cathode ray tube projector are developed. The projection technology of LCD projector is more mature. The chip of DLP digital light processing projector completes the visual digital information display technology. There is rapid development of DLP digital light processing projector, through the lighting design activities to meet the needs of people's visual viewing, while taking into account people's psychological perception of the lighting environment. Expression of local characteristics, highlights the natural and cultural characteristics of the regional scope, highlights

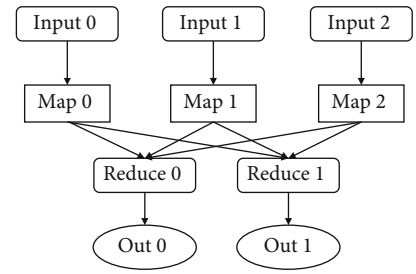



Figure 2: Work flow of MapReduce.

the landscape content and specific landscape intention, and gives artistic conception to the humanistic atmosphere. With lighting design as the medium, we can better coordinate the relationship between natural environment and humanistic environment and create a more humane lighting atmosphere.

*4.3.2. Spatial Characteristics Analysis and Atmosphere Refinement.* Landscape lighting environment design is an important means to decorate and beautify the environment and create artistic atmosphere. In order to decorate the residential landscape, increase the spatial level, and render the environmental atmosphere, it is very important to use decorative lighting and decorative lamps. The same lighting effect can be embodied in different technical ways. For example, the spectrum with the same intensity that people feel may be composed of light with different wavelengths. Contour lighting, floodlight lighting, and interior transparent lighting all belong to the landscape lighting modes of modern urban waterfront space. In the design process of landscape lighting scheme, it is necessary to take the factors such as lighting source, lighting environment, and lighting function zoning as reference and select and apply these lighting modes reasonably.

The GFS (Google File System) is composed of a Master and a large number of block servers. Master stores all metadata of the file system, including namespace, access control, file block information, and file block location information. The files in GFS are divided into 6 MB blocks for storage, as shown in Figure 4. After obtaining the write authorization from Master, the client transmits the data to all the data copies, and only after all the data copies receive the modified data does the client send out the write request control signal. After all data copies update the data, the master copy sends a write operation completion control signal to the client.

In order to achieve the core design concept of maximum humanization, the layout and features of landscape elements should be followed, and the functionality of lighting design should be met; the road system of the park is analyzed in landscape design (Figure 5), hoping to cover the content requirements of humanized lighting design more comprehensively.

The function of light and color is to render the space environment and reflect the theme characteristics of public places. When designing light and color, the influence of light and color on surrounding plants should be fully considered to avoid interference with the growth of plants. There are three kinds of light shapes: point, line, and plane. Point light source can be used for local illumination to attract people's

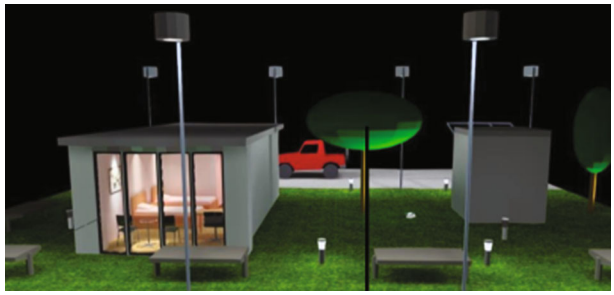


Figure 3: Lighting environment.

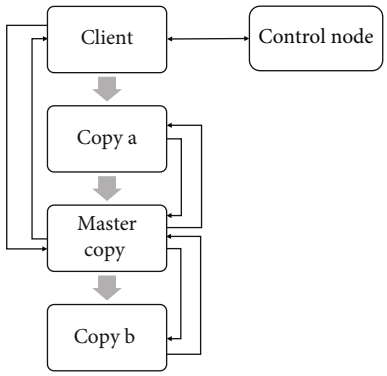



Figure 4: The write control signal of GFS is separated from the data stream.

attention. Among them, the garden lamp not only plays a role in ensuring lighting but also pays great attention to its shape, material, color, proportion, and scale. The garden lamp has become an indispensable ornament in residential landscape. In the lighting design, obviously, we should emphasize this sense of depth and, at the same time, let the public enter the building from the urban space. Emotions are gradually moved to control the different dimming modes of the panel, adjust the brightness and illumination of light and shadow, and create a three-dimensional and layered sense with abundant light, so that the landscape public space can show the best side.

*4.3.3. Design of Intelligent Lighting Control System for Landscape Public Space Based on Big Data.* Combined with big data technology, this project puts forward the architecture of intelligent lighting control system in landscape public space. By transforming the existing street lamps, a variety of sensors, controllers, and communication modules are used to build a smart city information perception network with high intelligence. This project collects the location and running status information of urban lighting equipment in real time, which can not only provide all-round street lamp and operation and maintenance of information technology service for the city but also provide personalized decision support for road sections under different environmental parameters by using big data platform and also provide effective support for urban intelligent transportation and intelligent security.

This project focuses on the intelligent lighting management of landscape public space based on big data technology and proposes a framework of intelligent lighting control system of landscape public space including multiple sensors, central controller, and communication module. On this basis, a big data platform for urban public lighting management is constructed. The system is divided into intelligence perception layer, network transport layer, and big data processing application layer. The system structure diagram is shown in Figure 6.

Intelligent sensing layer is the sensory organ of urban lighting intelligent energy-saving system, which captures all kinds of information in the area covered by lighting lines in real time through various intelligent sensors. The collected information includes various information of lighting equipment, such as current, voltage, and fault information. Lighting use related information, such as the flow of people passing through per unit time and the number of vehicles. Environmental parameters include illuminance, temperature, humidity, air pollutant monitoring, and noise monitoring.

The application layer of big data processing is the "brain" of urban lighting system, and the intelligent lighting control strategy reflects the intelligence degree of the whole system. Comprehensive utilization of all kinds of collected data information and related information of environmental parameters can realize big data analysis, and at the same time, it can assist to realize real-time monitoring of street lamp working conditions and provide functions such as alarm of street lamp fault type and notification of fault location. The functional structure of the system is shown in Figure 7.

Big data processing center includes five modules: data source, data collection, data storage, data processing, and data display. All kinds of data collected from the intelligent sensing layer are transmitted to the big data processing center as data sources through the network layer. Through the analysis and processing of the data, the intelligent lighting control schemes in different environments are obtained, which makes the lighting control scheme break the traditional unified control mode and analyzes the characteristic lighting control schemes according to the different situations of the actual lighting scene, thus improving the energy-saving efficiency. The structural diagram of the big data processing center is shown in Figure 8.

*4.3.4. Lighting Control and Energy Saving.* The lighting control is mainly based on the automatic lighting controller of the branch circuit, supplemented by manual control. Intelligent lighting controller is used to switch the power supply with different functions in each partition, which can be used for overall control, partition control, and timed scene control. The green courtyard is divided into active area and quiet area, and the active area is designed with high illumination according to the requirements of the lighting code. Lighting design should be people-oriented, starting from people's physiological and psychological needs, select the light source and color suitable for garden scenic spots, and determine the reasonable light consumption and illumination, which can not only fully show the night viewing of

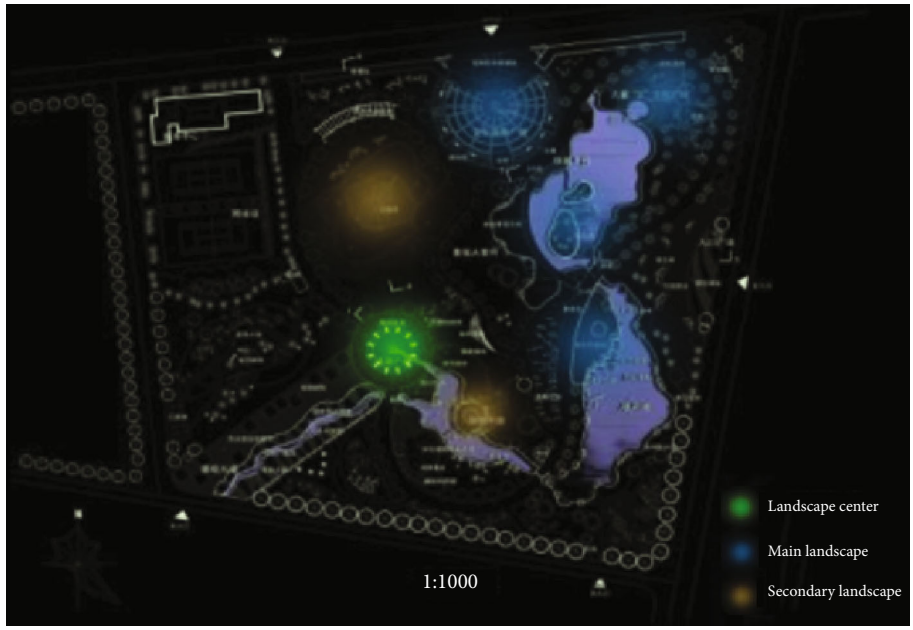



Figure 5: Landscape structure analysis diagram.

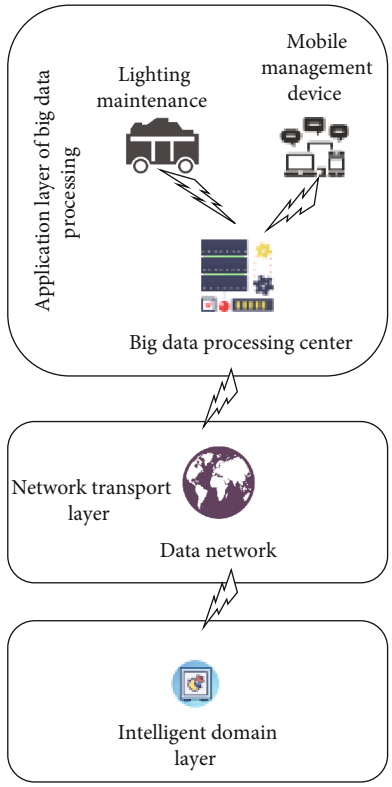



Figure 6: System structure diagram.

residential landscape and the interaction with residents but also avoid designing excessive night lighting to meet the effect of landscape as the key decorative lighting of nodes.

Sphere adopts the flow processing calculation mode [26]. In the stream processing mode, each element of the input array is independently processed by the same processing function using multiple computing units. Assume that a sector data set consists of one or more physical files. In Sphere, a data stream is actually an abstraction, which represents either a data set or a part of a data set.

The data processing process of Sphere is shown in Figure 9.

Figure 8 shows the specific process of Sphere processing data segments in a stream. Generally speaking, the number of data segments is larger than the number of SPEs. Sphere provides a simple mechanism to deal with the load balancing problem. Because the slower SPEs will process fewer data segments, they will be allocated one more data segment to the faster SPEs accordingly. Each SPE obtains a data segment from the stream as an input and generates a segment from the stream as an output. In turn, these output segments can be used as inputs for other Sphere processing processes.

According to the requirements of illumination index, the number of lamps and lanterns and the power of selected lamps and lanterns should be reasonably configured. Energy-saving and environment-friendly lamps were chosen. The selection of lamps and lanterns should comply with the national standard for lamps and lanterns (GB7000), power saving should be considered, and timing control should be carried out (timing control in different seasons in summer and winter) [27]. Different from the lighting modes of commercial streets and buildings, there is less flow of people at night in residential areas, and many lighting fixtures are directly accessible by human hands. Flooding lighting has a large demand for electric energy, and the only effective way is to control the demand for landscape lighting by illumination value [28]. High-pole lighting can be adopted, so that the light intensity can be distributed to a farther space as far as possible. Diffuse lamps and lampshades made of diffuse materials can be used to determine

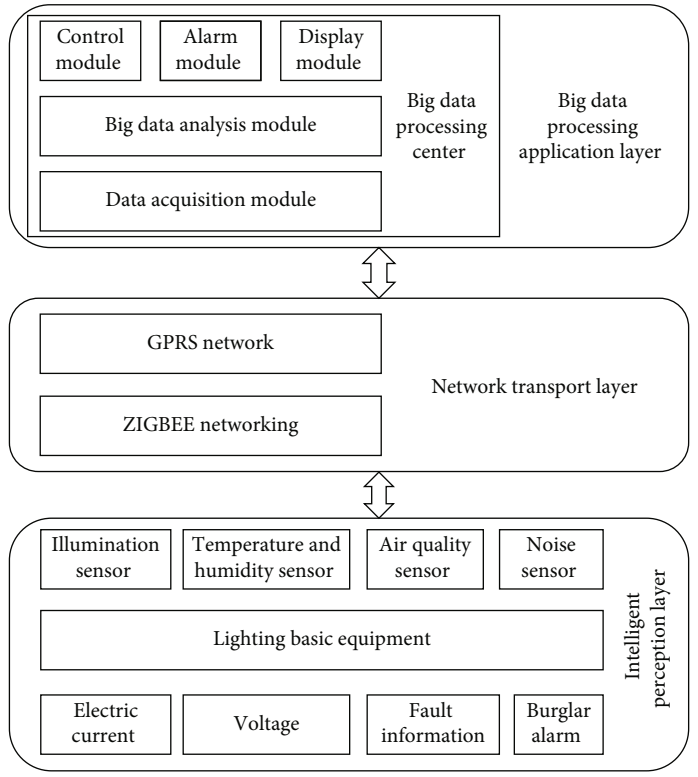



Figure 7: System functional structure diagram.

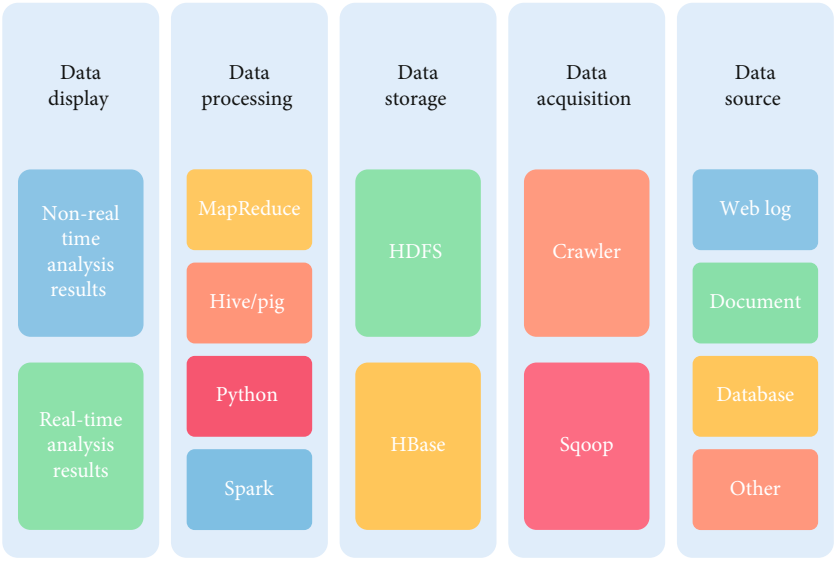



Figure 8: Structure diagram of big data processing center.

the appropriate lamp spacing and meet the requirements of uniform lighting.

Considering the safety of electrical equipment, the metal shell of lighting fixture should be equipotentially connected with the metal part of the building. Metal bases of lamps and lanterns and PE lines from terminals to lamp holders (cross sections with power lines, etc.) shall be reliably connected with PE lines of loop cables. A set of repeated grounding shall be made at the entrance of lighting distribution power supply of control box, and the grounding resistance shall not exceed 4 ohms. Considering personal safety, TN grounding system is adopted in the design. The grounding

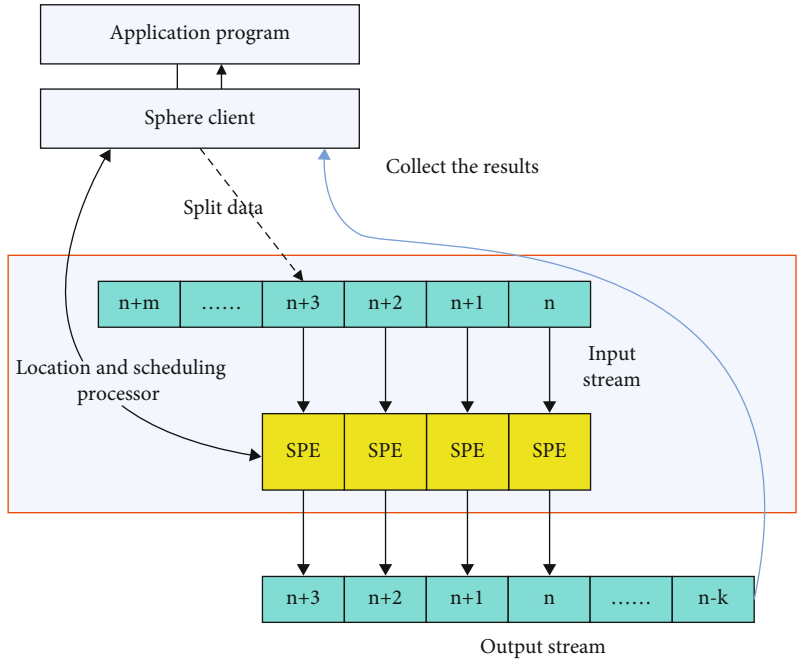



Figure 9: Sphere stream data processing process diagram.

resistance of PE line should be controlled within a reasonable range, and the resistance should be no more than 4 ohms.

## 5. Analysis and Discussion

Self-made data are used to collect the experimental information, in which the training sample set contains 1000 samples and 100 categories. The test data sets are 300 M and 580 M, respectively, containing tens of millions of test samples. The dimension of the algorithm is open (the highest test dimension in the experiment is 20 dimensions), but in order to make the data most representative, this experiment uses 10-dimensional samples.

Through a series of experiments on OptimDM in cluster environment, this paper collected the data of the relationship between the number of computing nodes and the running time of OptimDM algorithm in the edge cluster environment of Graph G, as shown in Table 1. The number of edges in the table is millions, and the number of nodes is 10, 20, and 30, respectively. The unit of running time is hours.

It can be seen from Figure 10 that there is a linear relationship between the number of edges in Graph G and the running time of the algorithm, which is consistent with the analysis of the time complexity of the algorithm in section 4. It can be seen that using the OptimDM algorithm, with the increase of the number of edges in the graph, the running time of the algorithm is within an acceptable range, and the performance of the algorithm remains stable all the time. By comparing the experimental results with the experimental results based on the traditional breadth-first search algorithm, it can be found that the optimal algorithm has better time performance when processing large-scale graph data, which is obviously improved compared with the traditional single-node solution.

Table 1: The relationship between the number of edges of Graph G, the number of nodes in cluster and the running time of OptimDM.

| The number of edges of G | 10 | 20 | 30 |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 100 | 0.36 | 0.62 | 0.21 |
| 205 | 2.14 | 1.78 | 0.88 |
| 330 | 3.38 | 2.66 | 1.28 |
| 418 | 4.24 | 3.19 | 1.96 |

Then, the influence of the number of computing nodes on the running time of OptimDM algorithm is verified under the premise that the scale of the graph data is unchanged. The experimental data are from Table 1. When processing the graph data with $992 \times 10^6$ edges, the relationship between the number of computing nodes in the cluster and the running time of OptimDM algorithm is shown in Figure 11.

It can be seen from Figure 7 that increasing the number of computing nodes in the cluster can greatly shorten the running time of OptimDM algorithm and improve the algorithm performance on the premise that the number of edges in the graph is unchanged. Experimental results directly prove that the performance of the OptimDM algorithm is scalable and depends on the size of cluster. That is, the OptimDM algorithm can make full use of distributed computing resources and reduce the time cost of algorithm operation. Comparing this experiment with the traditional single-node environment experiment, we can find that the algorithm based on cloud computing distributed environment has advantages in performance scalability. As long as computing devices are added, the computing power can be

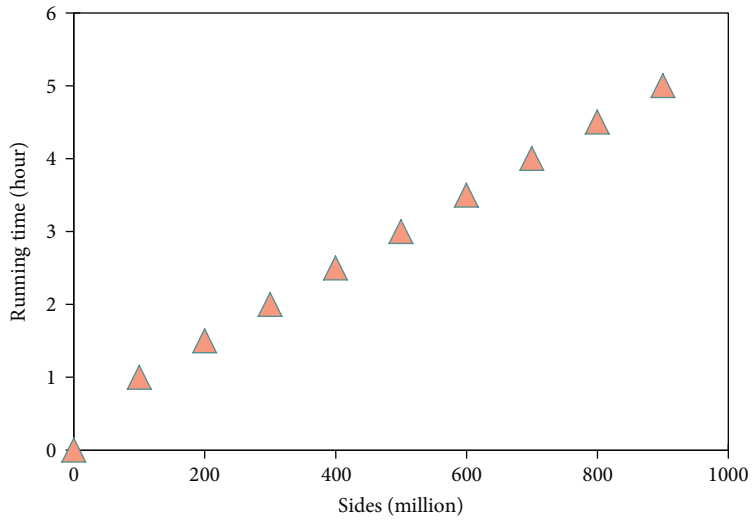



Figure 10: Relationship between the number of edges and running time in the cluster of nodes.

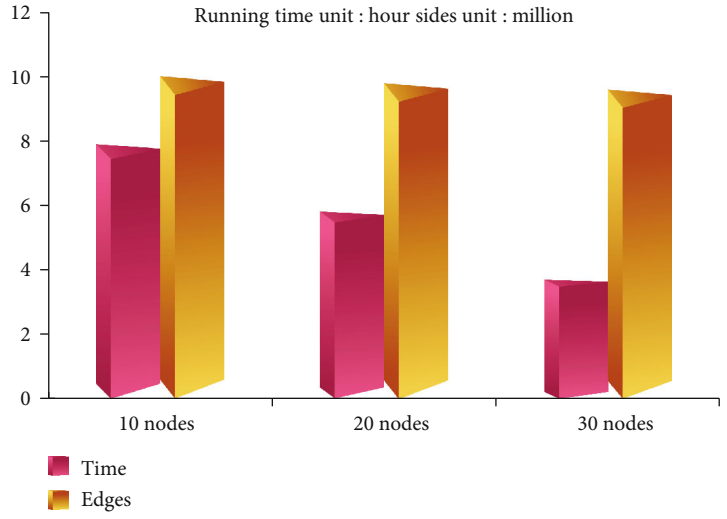



Figure 11: Relationship between cluster size and running time when dealing with the graph with fixed scale.

continuously enhanced. Therefore, this experiment also indirectly proves the powerful computing power brought by cloud computing cluster.

In order to detect the influence of the number of nodes in the Hadoop cluster on cluster performance, this experiment will kill off different numbers of data nodes in the cluster, and let the Hadoop clusters with different numbers of nodes process the same batch of input data (100,000 web pages and 10,000,000 web browsing records) and then analyze the performance changes between clusters with different numbers of machines, so as to obtain the performance acceleration of Chinese hotspot extraction algorithm MapReduce on the Hadoop cluster with the number of machines changing. In this experiment, the relative acceleration ratio coefficient is taken as an important measure, which is defined as follows: relative acceleration ratio = single data node cluster running time/multi – data node cluster running time.

After starting the Hadoop cluster, a certain number of data nodes is killed every time, then the MapReduce Chinese hotspot extraction algorithm is run, the running time of each cluster is recorded, and finally the acceleration ratio of each cluster experiment is calculated according to the definition of the above relative acceleration ratio formula, and we get Figure 12.

The data in Figure 12 shows that after Map/Reduce, the algorithm has received a good speedup effect when running on the Hadoop cluster. With the increase of data nodes in the cluster, the running speed of the cluster shows a nearly linear growth trend.

In structured data, entities appear in the form of tables, and the relationships among various entities are embodied by keys and constraints. The structure of structured data is similar to that of XML to a certain extent. The work of structure mapping is to map the tree graph of structured data

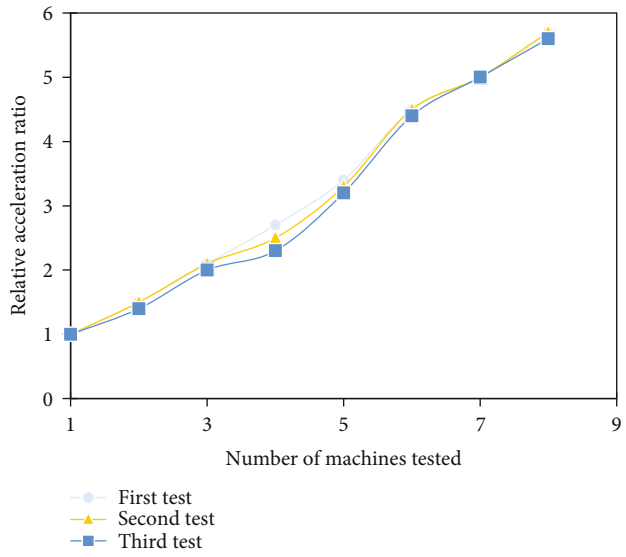



Figure 12: Hadoop cluster performance acceleration chart.

with edges as labels to the tag tree of XML with nodes as labels. However, the mapping of this structure does not contain semantic information in structured data. Once the domain structure is determined, the corresponding relationship between structured data and XML data can be determined, and the mapping from structured data to XML data can be automatically realized by the program through the field/domain interval comparison table.

Unstructured data is difficult to retrieve because of its various formats and large amount of data, and it lacks good organization strategy and mechanism when it is stored. If the unstructured data storage management system only provides simple data storage and maintenance functions, it is only a container for data, and it is impossible to mine valuable information in data on this basis. When adding or modifying the path field to the data table, the query system should find the source file. Only when the source file exists can the path and other information be recorded or modified in the data table. At the same time, the monitoring system updates the monitoring list and brings the newly added or modified data file into (or re-into) the monitoring range. Similarly, to delete a source file, we first remove the monitoring and then realize the consistent deletion of the source file and attribute data through the monitoring system.

## 6. Conclusion

Urban landscape public space lighting is a comprehensive landscape image composed of the city's natural landscape, humanistic elements, and lighting language. With the rapid development of science and technology, people's aesthetic ability is improving day by day, and the pace of social development is accelerating. People are no longer satisfied with the traditional municipal lighting. In the current cloud computing environment, the architecture of data mining platform is becoming more and more perfect, and its functions are becoming more and more powerful and diversified. In this paper, in the design and implementation of landscape public space lighting system, wireless sensor technology, embedded technology, and cloud computing framework are comprehensively used to form a series of complete systems in hardware and software. The wireless sensor technology is used to collect different parameter data of lamps, the embedded technology is used to forward different parameter data, and the server node is used to analyze and filter the data. Finally, the large-scale website is built to display the effect and realize data processing and real-time monitoring. Through practical exploration, a set of operation modes are condensed, and the perceptual design concept is combined with the rational operation mode, which enriches the lighting design method of landscape public space.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no competing interest.

## References

[1] W. Qipeng, "Urban road lighting design," *Urban Construction Theory Research*, vol. 10, pp. 284-285, 2017.

[2] L. Chang, W. Lei, L. Yanling, and H. Rui, "Discussion on visual communication art lighting design of highway tunnel," *Modern Tunnel Technology*, vol. 57, 2020.

[3] Y. Zhang, M. Zhang, S. Liu, F. Pan, N. Zou, and R. Zhang, "Research on classroom lighting design based on spectrum technology[J]," *Management and Technology of Small and Medium-sized Enterprises*, vol. 549, no. 08, pp. 188–197+201, 2018.

[4] Z. Wang, "Analysis of lighting design and energy-saving countermeasures of municipal roads," *Economic Vision*, vol. 2, pp. 162–162, 2017.

[5] M. T. Adams, R. O. Cleveland, and R. A. Roy, "Modeling-based design and assessment of an acousto-optic guided high-intensity focused ultrasound system," *Journal of Biomedical Optics*, vol. 22, no. 1, 2017.

[6] L. Zhang, E. Simo-Serra, Y. Ji, and C. Liu, "Generating digital painting lighting effects via RGB-space geometry," *ACM Transactions on Graphics*, vol. 39, no. 2, pp. 1–13, 2020.

[7] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems*, vol. 117, no. 6, pp. 47–58, 2021.

[8] K. Alla, "Lighting design for better health and well being[J]," *Nature*, vol. 354, no. 237, pp. 529–533, 2019.

[9] M. Alazab, S. Khan, S. Krishnan, Q. V. Pham, M. P. K. Reddy, and T. R. Gadekallu, "A multidirectional LSTM model for predicting the stability of a smart grid," *IEEE Access*, vol. 8, pp. 85454–85463, 2020.

[10] M. Parimala, R. M. Swarna Priya, M. Praveen Kumar Reddy, C. Lal Chowdhary, R. Kumar Poluru, and S. Khan, "Spatiotemporal-based sentiment analysis on tweets for risk assessment of

event using deep learning approach," *Software: Practice and Experience*, vol. 51, no. 3, pp. 550–570, 2021.

[11] H. Xianglin and W. Ji, "Research on lighting design of modern commercial space," *Lighting and Lighting*, vol. 41, no. 1, pp. 36–40, 2017.

[12] I. L. Wong, "A review of daylighting design and implementation in buildings," *Renewable & Sustainable Energy Reviews*, vol. 74, pp. 959–968, 2017.

[13] A. K. Bashir, S. Khan, B. Prabadevi et al., "Comparative analysis of machine learning algorithms for prediction of smart grid stability," *International Transactions on Electrical Energy Systems*, vol. 31, no. 9, 2021.

[14] F. Fathy, Y. Mansour, H. Sabry, M. Refat, and A. Wagdy, "Conceptual framework for daylighting and facade design in museums and exhibition spaces," *Solar Energy*, vol. 204, pp. 673–682, 2020.

[15] H. Yanguo and S. Fenghua, "Application of genetic ant colony algorithm in lighting design of highway tunnel," *Highway Engineering*, vol. 43, no. 4, pp. 39–43, 2018, 91.

[16] I. Acosta, C. Varela, J. F. Molina, J. Navarro, and J. J. Sendra, "Energy efficiency and lighting design in courtyards and atriums: a predictive method for daylight factors," *Applied Energy*, vol. 211, pp. 1216–1228, 2018.

[17] C. Jiang, K. T. Chau, Y. Y. Leung, C. Liu, C. H. T. Lee, and W. Han, "Design and analysis of wireless ballastless fluorescent lighting," *IEEE Transactions on Industrial Electronics*, vol. 66, 2019.

[18] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[19] N. Deepa, Q.-V. Pham, D. C. Nguyen et al., "A survey on blockchain for big data: approaches, opportunities, and future directions," http://arxiv.org/abs/2009.00858v2.

[20] G. Parise, M. Allegri, L. Parise, R. Pennacchia, F. Regoli, and G. Vasselli, "Topology of continuous availability for LED lighting systems," *IEEE Transactions on Industry Applications*, vol. 55, no. 6, pp. 5659–5665, 2019.

[21] X. Zhiyong, "Preliminary study on lighting design of commercial space," *Building Materials and Decoration*, vol. 4, pp. 84-85, 2017.

[22] K. Konis, "A circadian design assist tool to evaluate daylight access in buildings for human biological lighting needs," *Solar Energy*, vol. 191, no. Oct., pp. 449–458, 2019.

[23] L. Chao, "On energy-saving methods in architectural electrical lighting design," *Housing and Real Estate*, vol. 490, no. 5, pp. 64-65, 2018.

[24] S. Hu, "Research on data mining platform architecture and its key technologies based on cloud computing," *Wireless Internet Technology*, vol. 17, pp. 125-126, 2020.

[25] G. Wenwen, "Application of color perception in urban night lighting design," *Housing and real estate*, vol. 65, 2020.

[26] J. Yang and P. Liu, "Research on innovative design thinking based on data mining technology," *Design*, vol. 33, no. 3, pp. 76-77, 2020.

[27] S. Farooq, A. Ahmed, and M. A. Kamal, "Assessment of lighting design of restaurants with reference to its aesthetics and function," *Civil Engineering and Architecture*, vol. 8, no. 4, pp. 714–720, 2020.

[28] V. Zheltov, "Spatial angle distribution of brightness in lighting design of lighting scene [J]," *Automation and Modeling in Design and Management*, vol. 2020, no. 3, pp. 4–11, 2020.

[29] H. Gao, Z. Wang, D. Zhu, C. Zhang, and N. Zou, "Research on the influence of lighting mode and CCT on the lighting design of art museum based on subjective experiment," *AIP Advances*, vol. 10, no. 12, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 4923804, 11 pages
https://doi.org/10.1155/2021/4923804

*Research Article*

# Big Data Energy Consumption Monitoring Technology of Obese Individuals Based on MEMS Sensor

**Yongjun Zhao**,[1,2] **Juan Zhao**,[3] **Liang Ding**,[1,2] **and Congcong Xie**[4]

[1]*Department of Physical Education, Lvliang University, Lvliang, Shanxi, China*
[2]*Institute of Fitness and Rehabilitation, Lvliang University, Lvliang City, Shanxi Province, China*
[3]*Rehabilitation Department, Taiyuan Peace Hospital, Taiyuan, Shanxi, China*
[4]*Research Institute of Family Planning of Hebei Province, Key Laboratory for Family Planning and Birth Health of the National Health, Shijiazhuang, Hebei, China*

Correspondence should be addressed to Yongjun Zhao; 37018@llu.edu.cn

Received 12 August 2021; Revised 16 September 2021; Accepted 21 September 2021; Published 14 October 2021

Academic Editor: Rajesh Kaluri



The application of micro electro mechanical system (MEMS) is more and more extensive, involving military, medical, communication and other major fields. The progress of science and technology has brought cross era changes to human beings, but also brought troubles to human beings. Because machines can replace most people, which leads to a significant reduction in human exercise, many people have the symptoms of obesity. Therefore, how to effectively detect human exercise energy consumption is of great significance to improve obesity symptoms. The energy consumption detector takes stm32f103zet6 as the core processor and uses the inertial sensor mpu6050 to build a MEMS sensor system to monitor the daily motion state and gait of human body in real time. In the design of the big data algorithm, the adaptive peak detection and step, decision tree two-level classification of motion recognition big data algorithm are organically integrated, and then combined with the acceleration vector value of the motion energy detection big data algorithm, to process the collected motion data, including the acceleration signal, gyroscope and other data processing, and finally complete the feature extraction, get the final recognition and detection results. Through the data reference, we can know that the system can recognize different human motion states. Among them, it has 95% accuracy in the motion recognition of sitting, standing, walking, running, going up and down stairs and lying back, which is basically the same as the top detectors on the market. In the energy consumption detection, it also has 95% accuracy, which proves the correctness of the experimental big data algorithm design, and also improves the accuracy It is proved that the system has good performance and high practicability, and can provide a new idea for obese individual motion detection.

## 1. Introduction

With the continuous progress of MEMS technology, the big data motion monitoring industry has developed rapidly in recent years. At the same time, in recent years, the construction of health monitoring system in various medical and health industries and enterprises has accelerated [1]. As an information content, gait data is becoming more and more popular. In the future, the motion state data will use the motion energy consumption classification big data algorithm to achieve specific target tasks, such as daily motion or motion energy consumption monitoring. With the application of these technologies in daily life, it is very important to accurately classify the existing motion state capture and recognition models [2]. Waegli et al. proposed an improved coarse alignment method and quaternion estimation method to calculate the initial direction. Simulation and practical experiments show that the two methods are stable in any initial direction for large data sensors with MEMS Imus error characteristics [3]. Bottenfield et al. used MEMS IMU to obtain angle, acceleration, velocity and displacement data for the simulation of eye response and head impact test. The simulated collision test was carried out by 3D printing skull model and manikin [4]. Sang et al. proposed a method

to locate the pulse source using MEMS microphone and delay and beamforming. The time and position of the ball in three-dimensional space are determined by high-speed infrared scanning method. The experimental results show that the sound based ball motion estimation has a larger range of motion than the camera based method [5]. Nguyen et al. proposed a MEMS based pressure sensor, which can simultaneously measure blood pulse wave and respiratory rate with only one sensor element. The tube of the sensor device contacts the angular artery of the object and the area above the nasal cavity, resulting in changes in pressure in the tube caused by the subject's pulse wave and breathing. Experiments show that it is feasible to extract pulse wave and respiratory related information by using the low-frequency and high-frequency components of sensor signals [6]. Park et al. uses MEMS sensor information of wearable devices to identify relevant basic hand movements. The acceleration and gyroscope collected by MEMS are used to process the sensor information, and the distance between the data sensor and the sensor is calculated according to the stored information. The experimental results show that compared with the traditional research, the action recognition method can distinguish many actions, and the average recognition rate is 97.1% [7]. Tu et al. proposed a big data algorithm for gait recognition based on MEMS acceleration sensor. Local key points are used to generate sparse gait feature location templates, and template fusion is used to transform sparse gait periodic features effectively. Finally, the nearest neighbor big data algorithm and voting mechanism are used to identify gait features. The experimental results show that the recognition rate is 98.67% and the authentication rate is 99.89% [8].

Morozov et al. proposed a MEMS accelerometer model with two movable beam elements located between two fixed electrodes. The longitudinal inertial force changes the spectrum characteristics of the system and can be used as the output signal of big data sensor. The results show that the sensitivity of the big data sensor based on mode positioning is higher than that based on natural frequency drift [9]. Using data from synchronous sensors, Francesco et al. can give a comprehensive overview of user mobility. MEMS system can measure pole angle, arm cycle frequency and synchronization, and thrust applied on the ground. In addition, the data from the GPS module gives the environment image of the active session in terms of distance, slope and ground type [10]. Sheng et al. designed an adaptive attitude measurement big data algorithm based on MEMS gyroscope and accelerometer. The big data algorithm adopts extended Kalman filter to realize data fusion. At the same time, the Allan variance is used to estimate the dynamic noise of MEMS Gyro, and the big data algorithm with forgetting factor and limited memory is added. The experimental results show that the combination of the two can achieve high-precision attitude measurement, verify that the big data algorithm has good dynamic noise suppression ability, and improve the adaptability of the system to environmental changes [11]. Wang et al. proposed an adaptive tracking Kalman filter for UAV MEMS navigation. The filter transforms the strong inertial navigation system, and uses the optimal adaptive factor technology to overcome the influence of noise uncertainty and motion model error. For small UAV applications, large or small initial attitude error can be ensured without changing the model [12]. Tao et al. proposed a gait authentication method based on MEMS inertial sensors. They are fixed on smart shoes, collect motion signals and send them to the server. Then, gait parameters such as step length, gait frequency, attitude phase, swing phase and pitch angle are calculated as the characteristics of personal recognition. A new big data probabilistic neural network classification mechanism is proposed as the only mechanism to identify different users. Experimental results show that the method is effective. The average classification rates of 22 people reached 85.3% and 85.7%, respectively [13]. Reddy et al. studied two important dimensionality reduction techniques, linear discriminant analysis (LDA) and principal component analysis (PCA), on four popular machine learning (ML) algorithms [14]. Resnik et al. demonstrated the process of this study by using specially designed MEMS acceleration sensors for measurement and subsequent evaluation on a bridge in Armenia. These findings confirm the great potential of this method in monitoring support structures [15]. The fan s team used PDMS and polyimide MEMS flexible strain/pressure sensors. The device structure of MEMS can integrate three electrical sensors that output digital signals into the strain and pressure detection area. The results show that the engineering data sensors with different bending sensitivity can be applied to the control of robot arm [16].

The innovative contribution of this paper is to use the inertial sensor mpu6050 to build a MEMS sensor system to monitor the daily motion state and gait of human body in real time, and introduce the dynamic threshold detection method for gait recognition. Aiming at the shortcomings of the system gyroscope, the adaptive dynamic threshold is used to improve it.

## 2. Motion State Feature Extraction Technology and Energy Consumption Monitoring Big Data Algorithm Design

*2.1. Analysis of Gait Detection Methods.* Walking is a very effective aerobic exercise for obese individuals. At present, there are mainly two methods to detect the number of steps, one is dynamic threshold detection. This method mainly calculates the number of experimental steps through the detection of dynamic threshold and dynamic accuracy. In a complete step cycle, three directions of acceleration will be generated. In the measurement, three directions are defined as X, Y and Z. Through the monitoring of acceleration in these three directions, the acceleration threshold in each direction is extracted. At the same time, the sampling frequency is set to 50 Hz to adjust the dynamic accuracy. Finally, the average value of the maximum threshold and the minimum threshold in the same direction is calculated as the dynamic threshold. The calculation method of the dynamic threshold is shown in Figure 1.

In Figure 1, displacement register and dynamic threshold play a key role in accurate recognition. The former is composed of new and old sampling value registers, which are used to store new and old sampling data, respectively.

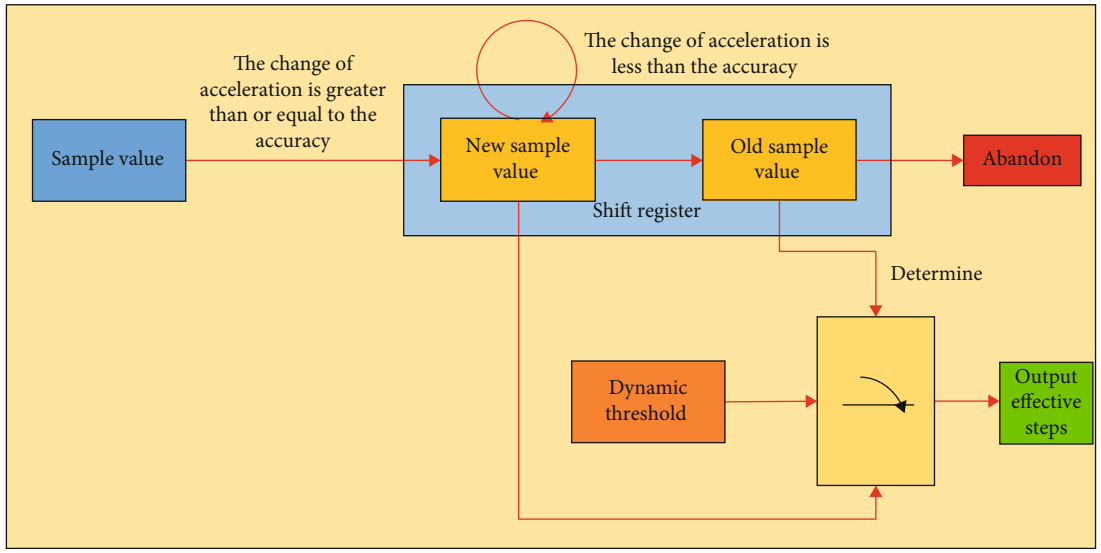
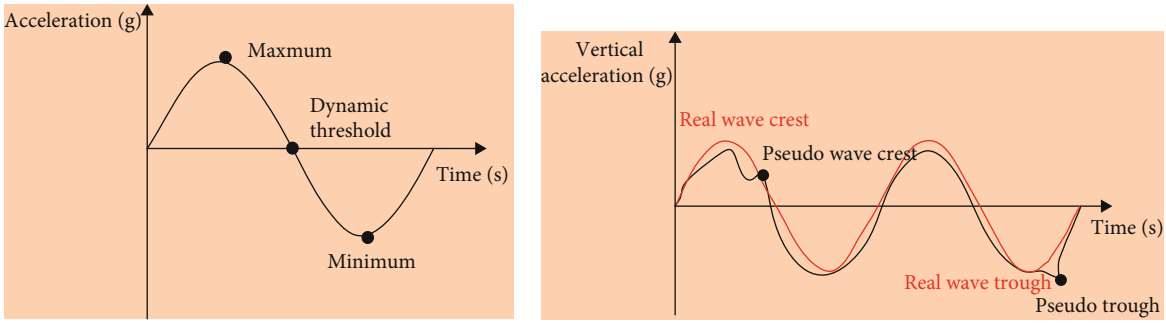



Figure 1: Schematic diagram of dynamic threshold big data algorithm.




(a) Dynamic threshold diagram

(b) Periodic sinusoidal waveform of vertical acceleration

Figure 2: Periodic sinusoidal waveform of vertical acceleration.

In the judgment of acceleration, the acceleration that exceeds or equals to the dynamic accuracy is input into the new sampling value register. On the contrary, if the acceleration is ignored, the data in the new sampling value register remains unchanged. Figure 2(a) shows the change of dynamic threshold in motion. It can be seen that the acceleration curve in a complete step will pass through the dynamic threshold from top to bottom, but this method is not accurate in vibration recognition, and the recognition accuracy is low. In view of this defect, the second method can be used to optimize and improve the first method.

In the process of walking, with the constant change of the center of gravity, the change of the acceleration signal is similar to the sine wave change curve, and the most significant change is the $z$-axis acceleration signal. As shown in Figure 2(b), the dotted line represents the ideal sine waveform, and the solid line represents the vertical acceleration waveform generated during walking. According to the different slopes before and after the peak value of the sine wave, the sampling values before and after the difference are made in turn, and the positive and negative slopes are used to locate the turning point of the slope, so as to judge the peak value. In addition, in the process of walking, people will have two different movement states, namely normal and abnormal, and the corresponding movement states are walking and running. In view of these two states, different threshold parameters need to be set for segmented recognition. The big data algorithm uses the acceleration vector to replace the single axis acceleration signal, which can effectively avoid the problem of low adaptability of the sensor attitude when the human body is in different movements to some extent, and can effectively avoid the confusion of the initial big data algorithm in dealing with different states of movement, which greatly improves the accuracy of the pedometer. The flow chart of the improved big data algorithm is shown in Figure 3.

As shown in Figure 3, the frequency range of people in normal walking state is almost in the range [1,2.5], while the frequency range of people in fast motion state is in the range [1.5,5]. In view of the increase of step rate during running, appropriately increasing the sampling frequency can maximize the complete acquisition of gait information. To sum up, set the sampling frequency to 50 Hz. At the same time, two threshold parameters and domain window length are set for different motion states to detect the peak. The process can be divided into four parts. The first part is to calculate the acceleration vector value generated in the movement and filter the invalid data. The purpose is to detect

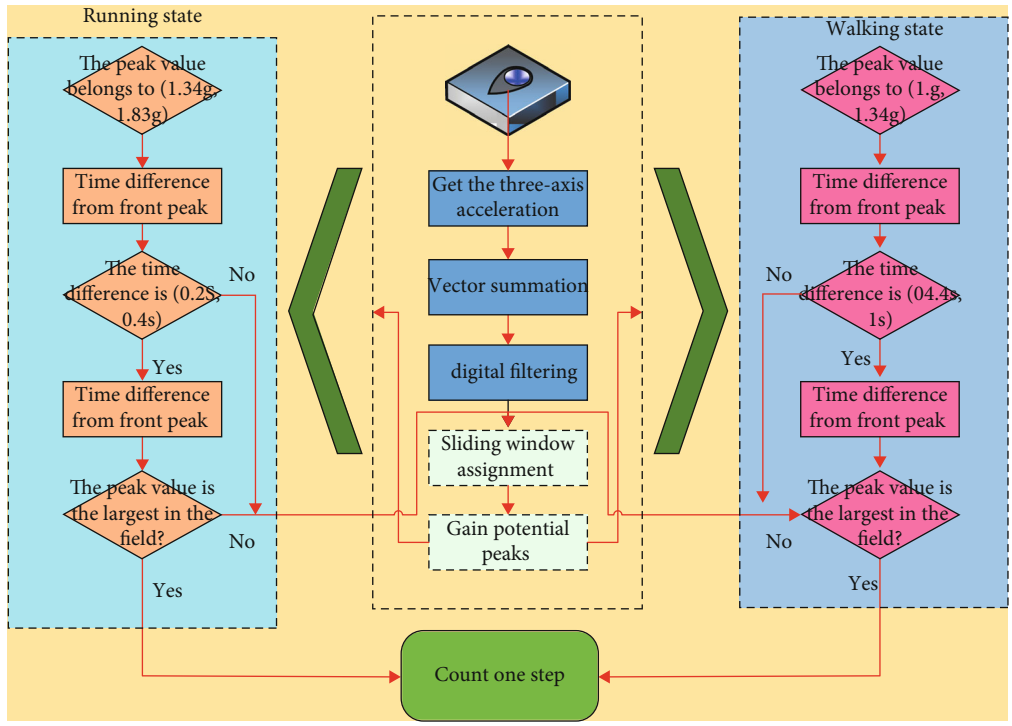



Figure 3: Adaptive peak detection method.

the change of the overall acceleration and reduce the influence of the sensor posture. The second part is to determine the gait. In this process, a 20 length sliding window is selected, and the maximum value is selected as the potential peak value The acceleration threshold in walking state is between 1 g-1.34 g, and that in running state is between 1.34 g-1.83 g. These two intervals are the important basis for judging gait. The third part sets the time threshold discrimination range of normal state and abnormal state, the former range is [0.4 s, 1 s] and the latter range is [0.2S, 0.4 s]. The calculation here is based on the difference between the peak time, and the difference between the previous peak and the potential peak is calculated, and the value will be used as the basis for the second judgment; the last part is the step judgment. Here, we need to set the corresponding size of the field window for different motion states, and set the window length of the normal field and the abnormal field to be 10 and 5, respectively, and calculate the difference according to the peak sequence And the maximum difference is selected for the next judgment. If the peak is within the set range, it is determined that the monitoring object has completed one step walking, otherwise, the data remains unchanged.

*2.2. Two Level Classified Action Recognition Big Data Algorithm Based on Decision Tree.* Based on MEMS sensor technology, a two-layer classified motion recognition big data algorithm with decision tree as the core is proposed to recognize and capture common human actions. Firstly, the acceleration signals in X, Y and Z directions collected by the acceleration sensor are vector calculated to obtain the acceleration vector value VM. The energy consumption is calculated according to the formula w = FS. When the acceleration vector value changes, the time is integrated according to formula (1), and the motion energy consumption of the body in the time range is calculated.

$$E = umg\int_{VM1}^{VM2} dVM\int_{t1}^{t2} tdt \tag{1}$$

In formula (1), $E$ and $u$ represent the energy consumption and parameters of human body movement; $VM$, $VM1$ and $VM2$, respectively, represent the acceleration vector value, the initial acceleration and the end acceleration before the movement; $t1$ represents the starting time, $t2$ represents the termination time; $mg$ represents the weight of the subject. After consulting a large amount of data, it can be seen that when the independent operation and the velocimeter measure the attitude angle, generally speaking, in the static state, The accuracy of the system to obtain attitude angle by three directions is high, but there are some defects in dynamic performance, so gyro drift easily. Therefore, data fusion technology is needed to make up for this defect. When the object is in a static state, it can refer to formula (2).

$$G2 = A_x^2 + A_y^2 + A_z^2 \tag{2}$$

$A_x$, $A_y$, $A_z$ is the acceleration of three axes, which needs to be normalized. When the monitoring object is at rest, the vector value in $z$-axis direction is equal to 1. Common data fusion methods such as weighted average fusion method, although the method is simple and practical, the accuracy deviation is large. Neural network fusion method has the ability of self-

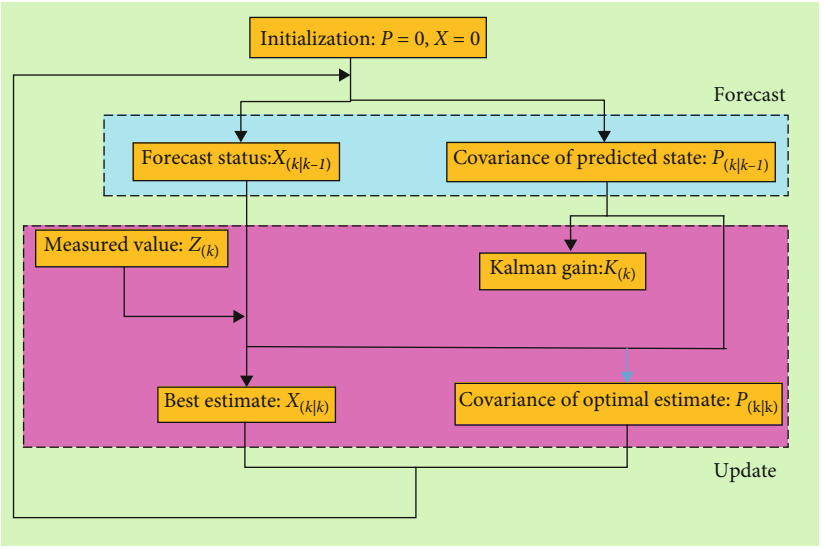



Figure 4: Schematic diagram of Kalman filter.

learning and good nonlinearity, but the effect is still unsatisfactory in the application of parameter optimization and structural model, because of its complex operation and high limitations. Compared with the former two methods, Kalman filtering method can combine the advantages of the two methods, the big data algorithm is simple and efficient, has high robustness and strong adaptability, and can better fit the linear filtering characteristics. The principle of Kalman filter is to evaluate the system state before the unit time stage based on the state space model, find the optimal state estimation of the model in the time period, and then evaluate the actual measured value and the estimation of the system, and finally get the optimal value. The principle is shown in Figure 4.

As can be seen from Figure 4, Kalman filter operation can be roughly divided into two parts, namely prediction and update correction. In the prediction stage, the first step is to calculate the prediction state $X_{(k|k-1)}$ of the system at $k$, and get the matrix $P_{(k|k-1)}$, $P_{(k|k-1)}$ can accurately reflect the trust degree of the current estimated state. The larger the value is, the smaller the trust degree is. $P_{(k|k-1)}$ is also defined as the error covariance matrix. The calculation formulas of the two are shown in formulas (3) and (4).

$$X_{(k|k-1)} = AX_{(k-1|k-1)} + BU_{(k)} \tag{3}$$

$$P_{(k|k-1)} = AP_{(k-1|k-1)}A^T + Q_k \tag{4}$$

Among them, $A$, $B$, $U_k$, $P_{(k|k-1)}$, $P_{(k-1|k-1)}$, $Q_{(k)}$ represents the matrix coefficient, the control quantity of current state, the covariance of $X_{(k|k-1)}$ and $X_{(k-1|k-1)}$, and the covariance of *k-value* estimation process. On the basis of formula (3), the angle measurement model equation is established for prior estimation, and the established equation is shown in formula (5).

$$Angle+ = (Gyro - Q\_bias) * dt \tag{5}$$

Among them, *angle* and *gyro*, respectively, represent the optimal estimation of angle at a certain time and the time before that time, and the angle design measurement value of gyroscope. *Q-bias* is the zero drift value of the gyroscope. Generally speaking, the *Q-bias* of the system is the same every time. Then the system state estimation matrix is obtained by combining formula (3) and formula (4), and the result is shown in formula (6).

$$\begin{bmatrix} Angle \\ Q\_bias \end{bmatrix} = \begin{bmatrix} 1 & -dt \\ 0 & 1 \end{bmatrix} \begin{bmatrix} Angle \\ Q\_bias \end{bmatrix} + \begin{bmatrix} dt \\ 0 \end{bmatrix} Gyro \tag{6}$$

The error covariance of the system is derived from formula (4), as shown in formula (7).

$$Q_k = \begin{bmatrix} \operatorname{cov}(Q\_angle, Q\_angle) & \operatorname{cov}(Q\_bias, Q\_angle) \\ \operatorname{cov}(Q\_angle, Q\_bias) & \operatorname{cov}(Q\_bias, Q\_bias) \end{bmatrix} \tag{7}$$

Where *Q-angle* is the covariance of gyro noise, because gyro drift noise and angle noise are two independent factors, the covariance between them is equal to 0. The second part of Kalman filter is the update phase. In this process, $K_{(k)}$ is a two-dimensional variable, corresponding to two Kalman gains, *Angle* and *Q-bias*. The observation variable is the angle value obtained from the acceleration signal, and the measured value *Accel* is the measured value $Z_{(k)}$, and the formula (8) is obtained.

$$\begin{cases} K_{(k)} = P_{(k|k-1)}H^T\left(HP_{(k|k-1)}H^T + R\right)^{-1} \\ K_{(k)} = \begin{bmatrix} K\_0 \\ K\_1 \end{bmatrix} \end{cases} \tag{8}$$

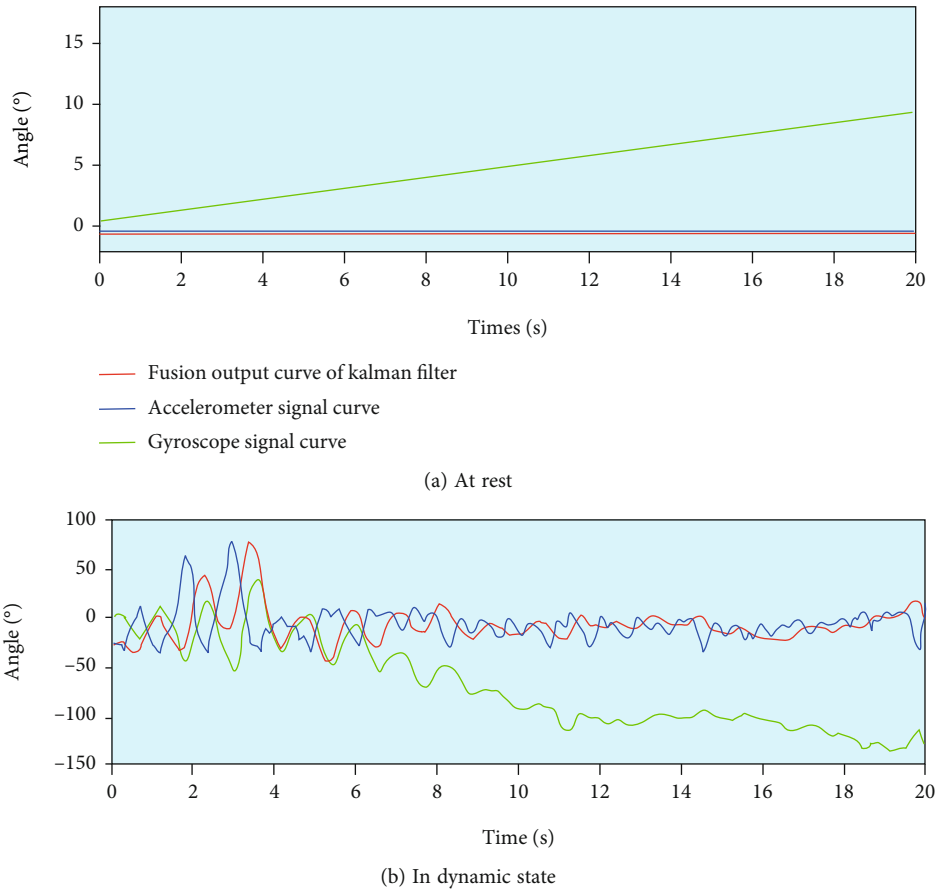



Figure 5: Effect of Kalman filter.

Among them, $R$, $K\_0$ and $K\_1$ represent the angle measurement noise, the Kalman gain of angle and the Kalman gain of gyroscope, respectively. The measurement equation in Kalman formula can be directly quoted here. The basic equation of Kalman is shown in formula (9).

$$Z_{(k)} = HX_{(k)} + V_{(k)} \tag{9}$$

Since $Q\_bias$ and $Accel$ have no relationship, it is concluded that,

$$H = [1 \quad 0] \tag{10}$$

$$P \times H^T = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \times [1 \quad 0]^T = [a \quad c] \tag{11}$$

$$H \times P \times H^T = [1 \quad 0] \times \begin{bmatrix} a & b \\ c & d \end{bmatrix} \times [1 \quad 0]^T = a \tag{12}$$

$$X_{(k|k)} = X_{(k|k-1)} + K_{(k)}\left(Z_{(k)} - HX_{(k|k-1)}\right) \tag{13}$$

Finally, the error covariance is updated, as shown in formula (14).

$$\begin{cases} P_{(k|k)} = \left(I - K_{(k)}H\right)P_{(k|k-1)} \\ I = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \end{cases} \tag{14}$$

Finally, formula (15) is obtained from formula (8):

$$P_{(k|k)} = P_{(k|k-1)} - K_{(k)}\left(HP_{(k|k-1)}\right) \tag{15}$$

Among the above formulas, formula (3) (4) (8) (13), and (14) is the basic equation of Kalman filter, and the effect diagram of Kalman filter is shown in Figure 5.

Finally, through the conversion, the complete Kalman filter is obtained. Figure 5 (a) shows the waveform of mpu6050 when the sensor is stationary. It can be seen that the attitude angle of the gyroscope has a serious deviation with time. At this time, the X and Y axes of the sensor are parallel to the placement plane, and the Z axis is the direction of gravity. Therefore, from the overall force situation, the sensor is only affected by the acceleration of gravity. In this state, the attitude angle of the sensor is 0 in theory. The main reason

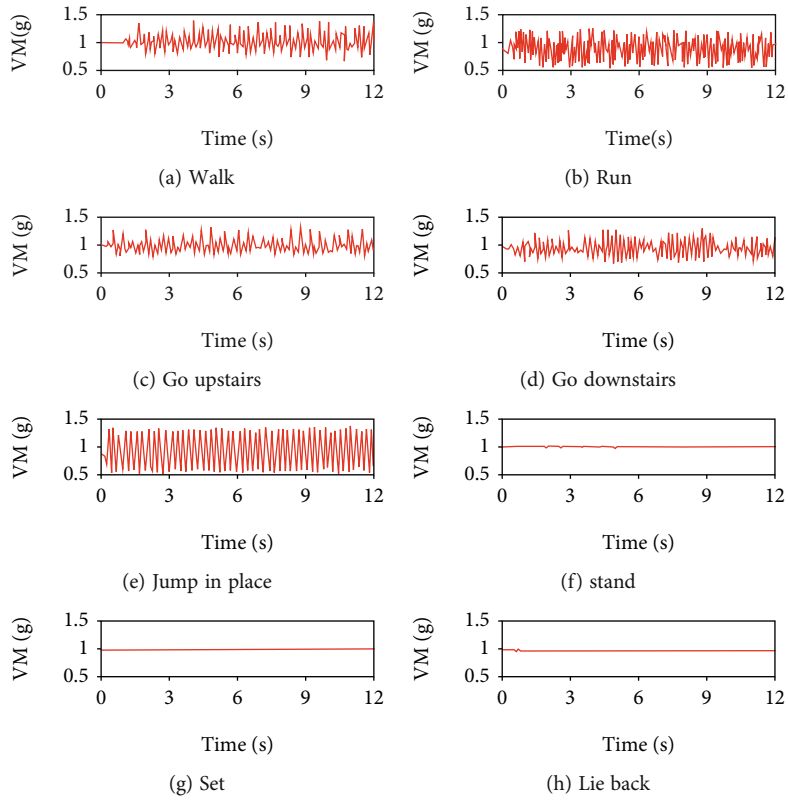



Figure 6: Variation curve of acceleration vector VM under different states.

of drift is the integral error of gyroscope signal. After Kalman filtering, the drift is corrected. Figure 5(b) shows the waveform generated by mpu6050 sensor after different degrees of vibration. It can be seen that the signal generated by gyroscope is calculated correctly in the period of 0-20, and it begins to drift after the time value exceeds 20, and the dynamic adaptability decreases seriously. After Kalman filter correction, the signal is very smooth, without bias. So it can be concluded that Kalman filter can accurately calculate the low attitude of the monitoring object in any state. On the one hand, it makes up for the defects of gyroscope, on the other hand, it provides a guarantee for the stable acquisition of the dynamic acceleration signal of the system, which greatly increases the accuracy of the system and improves the robustness of the system.

## 3. Analysis of Big Data Energy Consumption Monitoring Technology

*3.1. Performance Analysis of Action Recognition.* The wearing position of monitoring equipment has a great impact on data collection. In order to maximize the accuracy of measurement, the wearing position is set as the waist of the monitoring object. The reason is that compared with other parts, such as wrist, ankle and so on, the detection point is closest to the position of the center of gravity of the human body and produces less interference, which can more accurately reflect the real-time state of the human body in motion. The subjects were 10 males with an average age of 24 years and an average height of $171 \pm 5$ cm. The subjects completed the following movements according to the requirements: normal walking; 10 km/h running; up and down 10 floors, jumping in place and three static states, which were upright, sitting flat and lying back, respectively. Each movement was repeated five times. Through the data recorded by the monitoring instrument, the change of acceleration vector value VM is obtained. The experimental results are shown in Figure 6.

It can be seen from Figure 6 that the actions under different states are reflected by the change of the acceleration vector value VM. The acceleration vector values VM in Figures 6(b) and 6(e) are, respectively, distributed after 1.34 g-1.83 g and 1.83 g, and the boundaries of the two ranges are relatively obvious, so the two actions can be better identified. For other motion states, as shown in Figures 6(a), 6(c) and 6(d), the change curves of acceleration vector VM of normal walking, going upstairs and going downstairs are relatively close, and the distribution range is in the interval [1 g, 1.5 g], so the similarity degree is small, and it is unable to make an effective distinction. For the three static states, the fluctuation is also small, and the range value is about 1 g, so the equipment can not effectively distinguish. In view of this situation, the experimenters are arranged to test again according to the sequence of actions: slow walking, going upstairs, going downstairs, and static actions of standing up, sitting flat and lying back. Each group of actions is performed five times, and the change curve of angle is detected. The test results are shown in Figure 7.

It can be seen from the changes of the four attitude angles in Figure 7 that the angle detection based on the

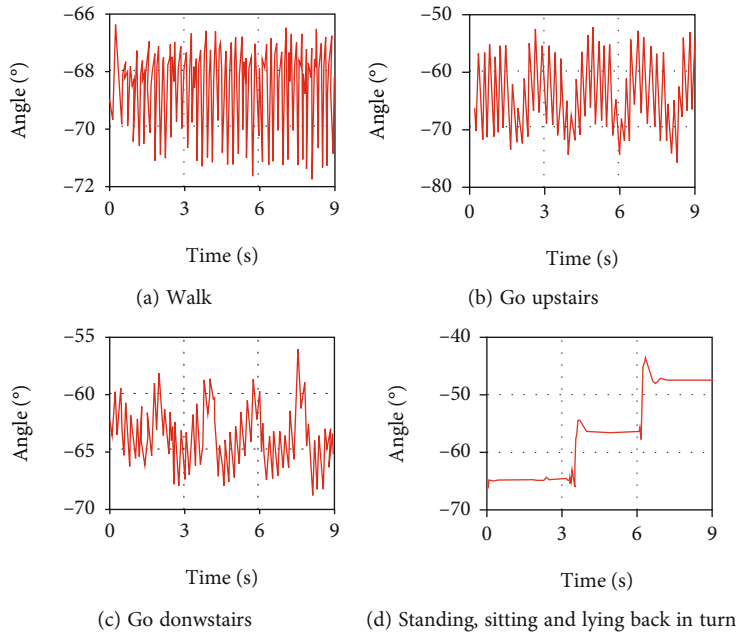



Figure 7: Angle change curve under different actions.

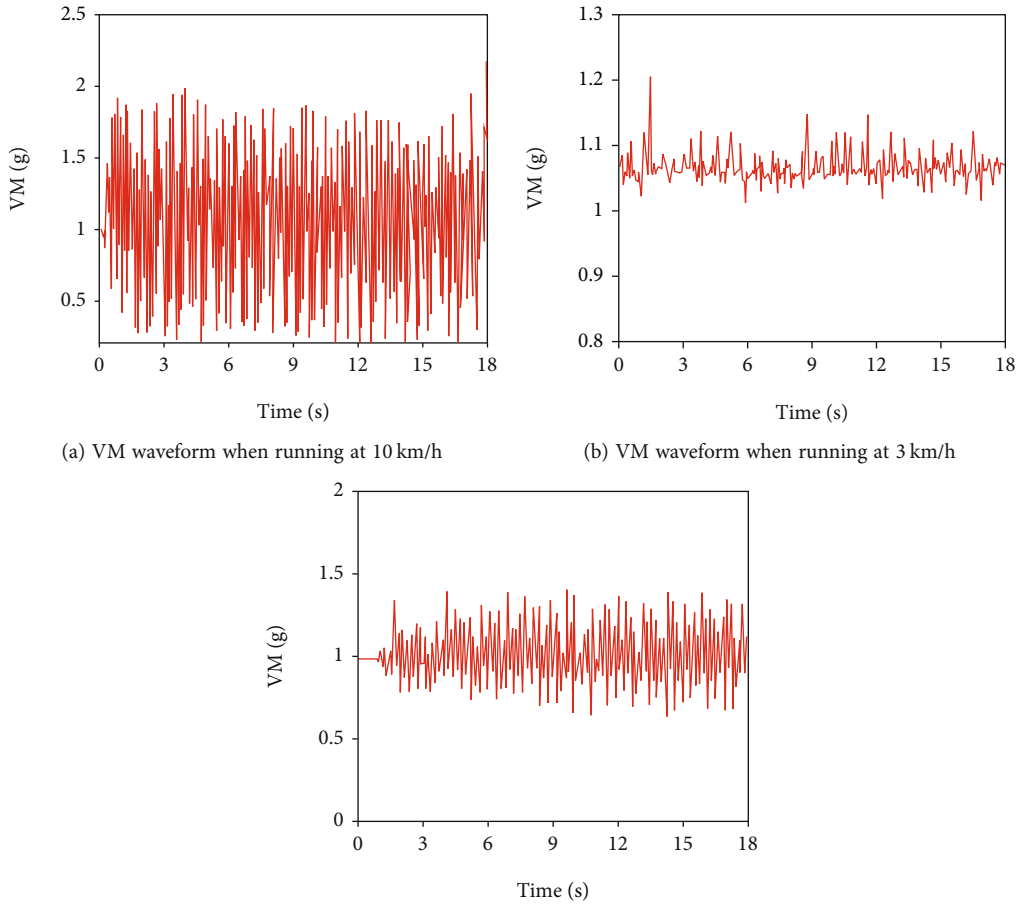
vector value VM detection can effectively distinguish walking, going upstairs and going downstairs. From the change trend of the curve, we can know that when the monitoring object is in different states of walking, going upstairs and going downstairs, the peak range of the attitude inclination signal is in [-69, -66.8], [-60, -50], [-66.5, -60], respectively. When the three movements are carried out in turn, it is obvious that the wave crests are segmented, and this segmented performance is just the key to identify different movements. The peak ranges of standing, sitting and lying back are, respectively, distributed in [-70, -65], [-60, -55], [-50, -40]. For this, we only need to set the peak of posture inclination angle of different actions as the action judgment threshold, that is, we can effectively identify different states.

*3.2. Performance Analysis of Motion Energy Consumption Detection.* In this paper, the energy consumption detector takes stm32f103zet6 as the core processor and uses the inertial sensor mpu6050 to build a MEMS sensor system to monitor the daily motion state and gait of human body in real time. The commonly used method of human motion energy consumption detection for wearable devices is double label method, because this method has very high accuracy and anti-interference ability. Although this method is good, it has high cost and low practicability. The energy consumption of treadmill in gymnasium is compared, which is simple and economical. The treadmill used for comparison is solid f63 plus in the United States, with the maximum speed of 20 km/h. After many times of verification, the deviation of running energy consumption detection is ±3%, which fully meets the experimental requirements. The object of the experiment is a 70 kg male, and the wearing position of the device remains unchanged. In order to ensure the accuracy of the big data algorithm to the greatest extent, the sampling window period is set to t = 1 s, and the refresh frequency is 50 Hz, that is, 50 samples are taken in 1 second. The experimental results are shown in Figure 8. As can be seen from Figure 8(a), when the detected object runs at different speeds, the serial port data records the vector change VM value at T = 0 and T = 1, and the gravity acceleration g is 9.8 m/s. After formula calculation, when the test object runs at the speed of 10 km/h, the energy consumption of human movement is 9.6 kcal/min, that is, 0.16 kcal/s, and the calculated u = 0.0018. The experimental results are shown in Figures 8(b) and 8(c).

It can be seen that when the experimenter moves at the speed of 3 km/h and 5 km/h, respectively, the corresponding energy consumption is calculated to be 0.017 kcal/s and 0.06 kcal/s, respectively. At this time, the energy consumption value on the treadmill is 0.017 kcal/s and 0.058 kcal/s, and the accuracy rate is 96%, which proves that the reliability and accuracy of the U value obtained from the test meet the requirements.

*3.3. Accuracy Analysis.* Through the comparison of the accuracy with several common sports wristbands on the market, including XM Sports Wristband, HW Sports Wristband and PG watch, three aspects of experimental accuracy were compared, including step test, action recognition and motion energy consumption detection. The experimental environment is corridor, the ambient temperature is 26°C, and the relative humidity is 26%. The experimental results are shown in Figure 9.

As can be seen from the experimental results in Figure 9, the average accuracy of common motion detectors on the market, such as XM bracelet, PG watch and HW bracelet, is above 90%, which can meet the detection requirements, with the average accuracy of 94% ± 1.2%, 96% ± 1.5% and 97% ± 0.6%, respectively. The average accuracy of this design system is 96.3 ± 0.8%, which is the closest to the top




(a) VM waveform when running at 10 km/h

(b) VM waveform when running at 3 km/h

(c) VM waveform when running at 5 km/h

Figure 8: VM waveform when running at different speeds.

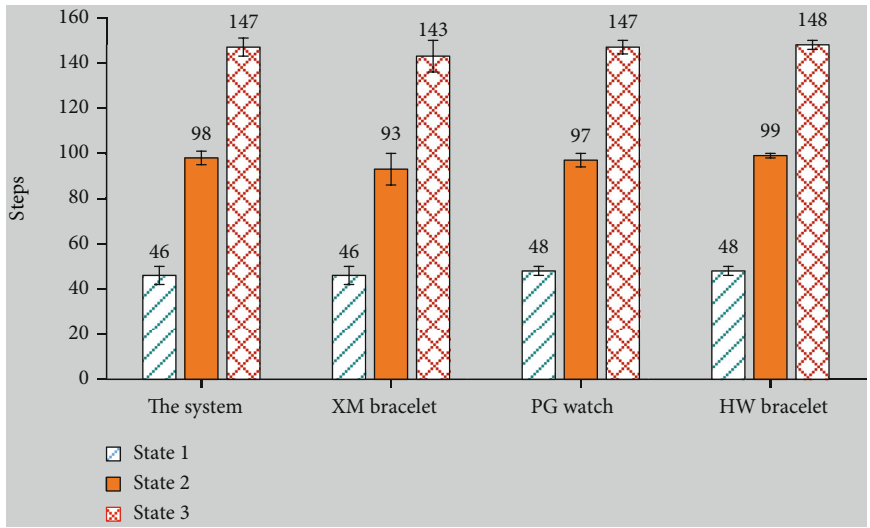



Figure 9: Comparison results of conventional accuracy.

detection system on the market. It can maintain high accuracy in the step detection, which proves that the big data algorithm is accurate. Next, the accuracy of different action recognition is tested, and the experimental results are shown in Figure 10.

Because the common motion detectors on the market have no action recognition function, as can be seen from Figure 10, in the recognition of walking, running, going upstairs, going downstairs, standing, sitting and lying back, it can basically guarantee more than 90% accuracy, with an

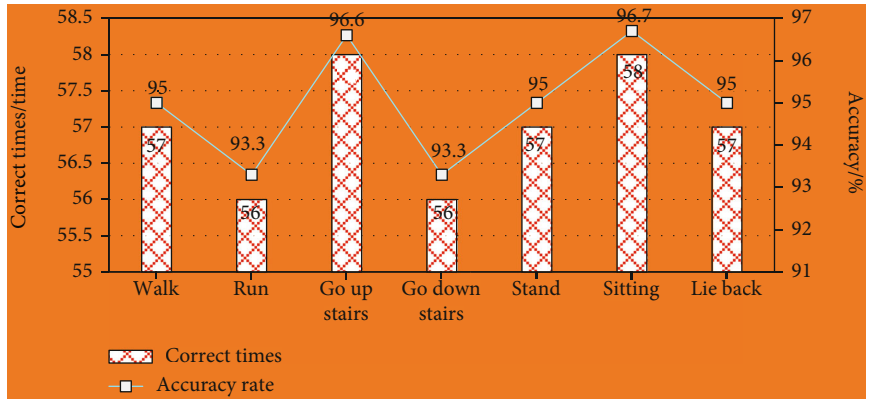
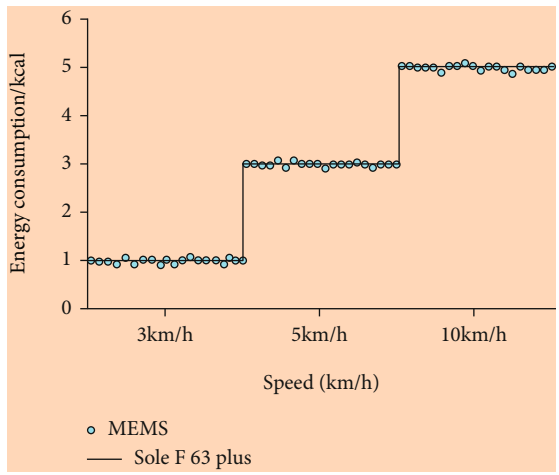
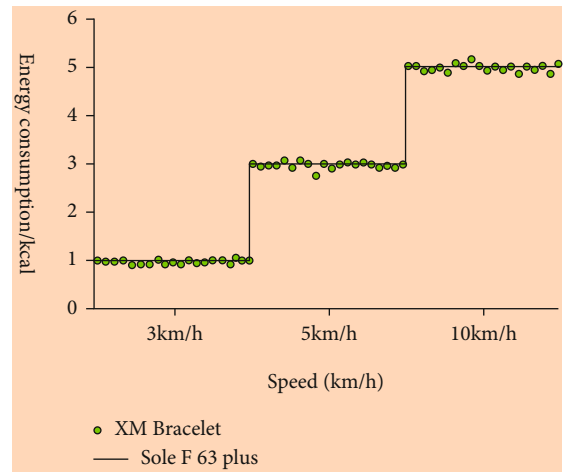


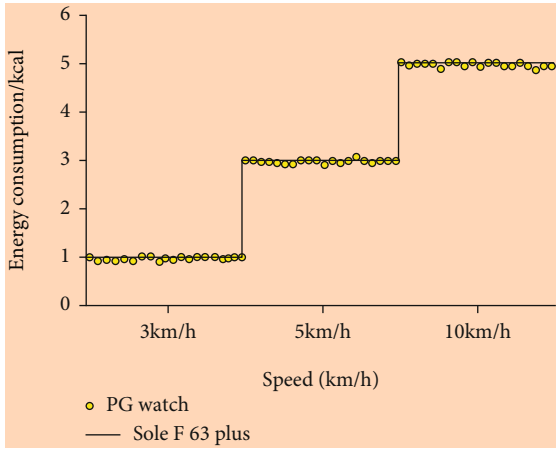

Figure 10: Comparison of action recognition accuracy.



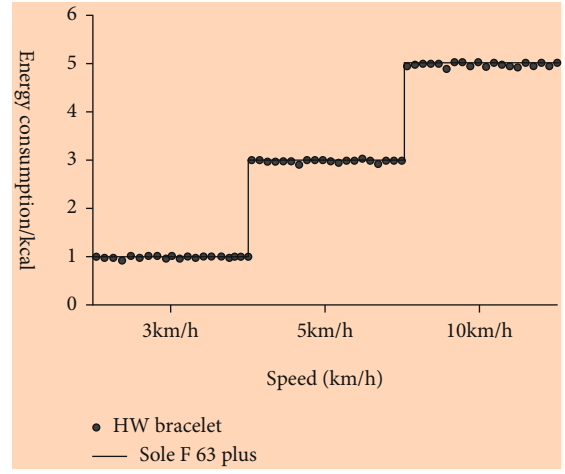

(a) MEMS accuracy




(b) XM Bracelet accuracy




(c) PG Watch accuracy




(d) HW Bracelet accuracy

Figure 11: Comparison of motion energy consumption accuracy.

average accuracy of 95%, which proves that the accuracy of the system is high and the big data algorithm is accurate. Finally, the energy consumption detection accuracy of the four big data algorithms is analyzed, and the results are shown in Figure 11.

It can be seen from Figure 11 that the detection results of four kinds of sports energy consumption are densely distributed near the control results of solid running, and the dispersion of XM bracelet is large, which reflects that the detection accuracy of the detector is low. Through calculation, the

average accuracy of XM bracelet, PG watch and HW bracelet is 87%, 95% and 96%, respectively, while the average accuracy of this system is 95%, which can basically meet the top technical requirements of PG and HW manufacturers. The results show that the system can fully meet the accuracy requirements when measuring the energy consumption of human movement, and can provide a new idea for the movement detection of obese individuals.

## 4. Conclusion

In the obese individual motion state monitoring technology, the inertial sensor mpu6050 is used to build MEMS sensor system to monitor the human daily motion state and steps in real time, and the dynamic threshold detection method is introduced to recognize the gait. In view of the shortcomings of the system gyroscope, adaptive dynamic threshold is used to improve it. In the design of the big data algorithm, the adaptive peak detection and step, decision tree two-level classification action recognition big data algorithm are organically integrated, and then combined with the acceleration vector value of the cloud top energy detection big data algorithm, to process the collected motion data, including the acceleration signal, gyroscope and other data processing, and finally complete the feature extraction, get the final recognition and detection results. Through data reference, we can know that the system can recognize different human motion states. Compared with the motion monitor with better performance on the market, the performance of this system is close to it, and to a certain extent, it is more practical. However, the data research in the big data algorithm design in this paper is not deep enough. Therefore, in the future research, the experiment needs further statistics and more data content.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no competing interest.

## Acknowledgments

The study was supported by "Scientific and Technologial Innovation Programs of Higher Education Institutions in Shanxi (Grant No. 2019L0945)".

## References

[1] L. Zhang, "Research and Design of A Motion Sensor Based on MEMS," *IOP Conference Series: Earth and Environmental Science*, vol. 170, no. 2, pp. 12-13, 2017.

[2] V. Camomilla, E. Bergamini, S. Fantozzi, and G. Vannozzi, "Trends Supporting the In-Field Use of Wearable Inertial Sensors for Sport Performance Evaluation: A Systematic Review," *Sensors*, vol. 18, no. 3, p. 873, 2018.

[3] A. Waegli and J. Skaloud, "Optimization of two GPS/MEMS-IMU integration strategies with application to sports," *GPS Solutions*, vol. 13, no. 4, pp. 315–326, 2009.

[4] B. Bottenfield, M. Yuan, and M. L. Adams, "Instrumentation for sensing passive eye response due to head impact via MEMS IMUs," *Advancing Microelectronics*, vol. 46, no. 2, pp. 16–19, 2019.

[5] S. W. Seo, M. Kim, and Y. Kim, "Optical and acoustic sensor-based 3D ball motion estimation for ball sport simulators," *Sensors*, vol. 18, no. 5, p. 1323, 2018.

[6] T. V. Nguyen and M. Ichiki, "MEMS-Based Sensor for Simultaneous Measurement of Pulse Wave and Respiration Rate," *Sensors*, vol. 19, no. 22, p. 4942, 2019.

[7] I. Park, S. Y. Lee, and J. J. Ko, "Hand gesture recognition with correlation using MEMS sensor," *The Journal of Korean Institute of Communications and Information Sciences*, vol. 42, no. 11, pp. 2139–2147, 2017.

[8] B. B. Tu, L. H. Gu, R. Y. Chuai, and H. Xu, "Gait recognition based on MEMS acceleration sensor," *Journal of China Inertial Technology*, vol. 25, no. 3, pp. 304–308, 2017.

[9] N. F. Morozov, D. A. Indeitsev, V. S. Igumnova et al., "A novel model of a mode-localized MEMS accelerometer," *Doklady Physics*, vol. 65, no. 10, pp. 371–375, 2020.

[10] F. Mocera, G. Aquilino, and A. Somà, "NordicWalking performance analysis with an integrated monitoring system," *Sensors*, vol. 18, no. 5, p. 1505, 2018.

[11] J. H. Sheng, Z. A. Zhang, and B. N. Xing, "Adaptive attitude measurement big data algorithm based on MEMS gyroscope and accelerometer," *Journal of Test Technology*, vol. 32, no. 4, pp. 277–284, 2018.

[12] D. Wang, H. Lv, and J. Wu, "In-flight initial alignment for small UAV MEMS-based navigation via adaptive unscented Kalman filtering approach," *Aerospace Science and Technology*, vol. 61, pp. 73–84, 2017.

[13] S. Tao, X. Zhang, H. Cai, Z. Lv, C. Hu, and H. Xie, "Gait based biometric personal authentication by using MEMS inertial sensors," *Journal of Ambient Intelligence & Humanized Computing*, vol. 9, no. 5, pp. 1705–1712, 2018.

[14] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[15] B. Resnik and A. Sargsyan, "Anwendung von MEMS-Beschleunigungs-sensoren im Rahmen von Bauwerksuberwachungen anhand eines typischen Beispiels," *Allgemeine Vermessungs Nachrichten*, vol. 126, no. 6-7, pp. 163–172, 2019.

[16] S. Fan, L. Meng, L. Dan, W. Zheng, and X. Wang, "Polymer microelectromechanical system-integrated flexible sensors for wearable technologies," *Sensors Journal IEEE*, vol. 19, no. 2, pp. 443–450, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 4556157, 11 pages
https://doi.org/10.1155/2021/4556157

*Research Article*

# A Novel SAR Image Target Recognition Algorithm under Big Data Analysis

**Xiang Chen, Xing Wang, You Chen, and Haihan Wang**

*Aeronautics Engineering College, Air Force Engineering University, Xi'an, China*

Correspondence should be addressed to Xiang Chen; 18811506679@163.com

Received 11 August 2021; Accepted 17 September 2021; Published 13 October 2021

Academic Editor: Thippa Reddy G



Synthetic aperture radar (SAR) image target recognition technology is aimed at automatically determining the presence or absence of target information from the input SAR image and improving the efficiency and accuracy of SAR image interpretation. Based on big data analysis, dirty data is removed, clean data is returned, and standardized processing of SAR image data is realized. At the same time, by establishing a statistical model of coherent speckles, the convolutional autoencoder is used to denoise the SAR image. Finally, the network model modified by softmax cross-entropy loss and Fisher loss is used for automatic target recognition. Based on the MSTAR data set, two scene graphs containing the target synthesized by the background image and the target slice are used for experiments. Several comparative experiments have verified the effectiveness of the classification and recognition model in this paper.

## 1. Introduction

In recent years, SAR image automatic target recognition technology (SARATR) has been widely used [1] and has formed a fixed three-level flow process: detection, identification, and classification [2]. The detection module is mainly based on the detection algorithm to obtain slices containing the SAR image target [3]; the discrimination module eliminates the false alarm value for the target slice [4]; the classification module selects the best decision mechanism to judge the category. With the rapid development of deep learning [5] in machine vision, its models can independently learn the internal laws of data from massive sample data and have strong feature expression ability. For the classification and recognition task of natural images, researchers have proposed some very typical deep learning network models, such as AlexNet [6], VGG [7], Googlenet [8], ResNet [9], and DenseNet [10]. These network models have obtained excellent results in the large-scale visual recognition challenge.

In order to accelerate the realization of efficient classification and recognition of deep learning in SAR targets, many scholars at home and abroad have carried out a series of studies on it. Housseini et al. [11] proposed to learn the convolution kernel and bias based on the convolutional autoencoder (CAE). This method guarantees a high recognition rate; the recognition speed is approximately 27 times that of the CNN architecture alone. Wang et al. [12] studied the influence of coherent speckles in SAR images on CNN for SAR target recognition. On this basis, they proposed a bipolar coupled CNN structure. They firstly used the denoising subnetwork to denoise and then learned the residual speckle characteristics and target information through the classification subnetwork. This structure can improve the noise robustness of the network. Wagner [13] combined convolutional neural network and support vector machine to classify 10 types of targets in the MSTAR database and finally got a recognition rate of 98.6%. Chen et al. [14] took VGG as the reference object and designed a 5-layer CNN network (A-ConvNet) with a large convolution kernel and full convolution mapping, which was cut from the target slice of the MSTAR data set and used as a training sample. The recognition rate reaches 99% in the standard environment and 96% in the 17°-30° extended environment. Zelnio et al. [15] used an 8-layer CNN structure to classify 10 types

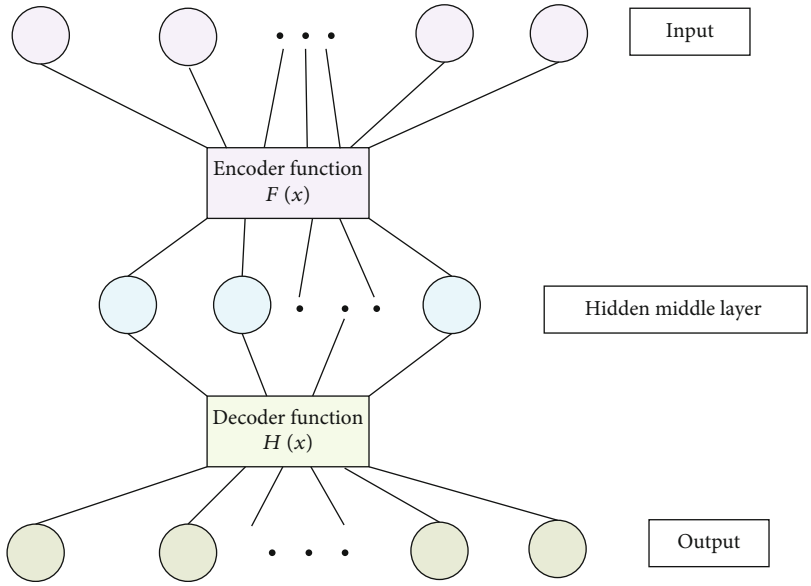



Figure 1: principle structure of CAE.

of vehicle targets, and its network structure design was similar to AlexNet, with a final recognition rate of 92.3%.

In practical application scenes, the accuracy of SAR image recognition is easily affected by noise interference and multiscale target transformation. In order to obtain effective battlefield situation awareness information, it is of great significance to study advanced SARATR technology. In this paper, a novel model framework based on deep learning is proposed to study the automatic target recognition technology of SAR image. Firstly, preprocess SAR image data through big data analysis methods. At the same time, by establishing a statistical model of coherent speckles, the convolutional autoencoder is used to denoise the SAR image. Finally, the network model modified by softmax cross-entropy loss and Fisher loss is used for automatic target recognition. Based on the MSTAR data set, experimental results show that the proposed method has better recognition performance than other traditional algorithms.

## 2. SAR Image Target Recognition Method

*2.1. SAR Image Data Preprocessing.* The big data analysis method is used to preprocess the SAR image data. The standardization process [16, 17] includes three main steps:

(1) *Image Data Analysis*. Firstly, analyze the SAR image data sample, obtain the attributes and various values of each basic data, and analyze the data quality based on the big data analysis according to the correctness and consistency of the SAR image data.

(2) Define workflow and conversion rules. For dirty data, build standardized data processing steps and conversion rules based on the quality and scale of the data. For data with pattern-level problems, it is necessary to specify a matching and query language to generate its processing code

(3) Reflow clean data. First, back up the SAR image data to be processed, and implement standardized processing on the SAR image data through the final workflow and conversion rules. Store the processed SAR image data in the original data source, and delete the original data entry

*2.2. SAR Image Coherent Speckle Noise Statistical Model.* Due to the coherent interference of radar waves reflected by many basic scatterers, the SAR image itself is affected by coherent speckle noise, which will reduce the spatial resolution of the image, blur the edge and texture characteristics of the image [18], and greatly affect the interpretation of the SAR image work [19]. Therefore, modeling and suppressing coherent speckle noise in SAR images is an important part of target recognition.

Coherent speckle noise modeling is mainly based on the assumption of fully developed coherent speckle, which can usually be regarded as a random multiplicative noise:

$$Y = X.I. \tag{1}$$

Among them, $Y$ represents the actual observed image intensity (including coherent speckle noise); $X$ is the noise-free image intensity (the actual situation does not exist); $I$ represents the random multiplicative coherent speckle noise intensity, which is statistically independent of $X$.

When the multiview coherent speckle noise satisfies the gamma distribution, the corresponding probability density function [20] is as follows:

$$P_I(I) = \frac{L^L I^{L-1} e^{-LI}}{\Gamma(L)}. \tag{2}$$

Among them, $L$ is the number of sights; $I$ is the intensity of coherent speckle noise.

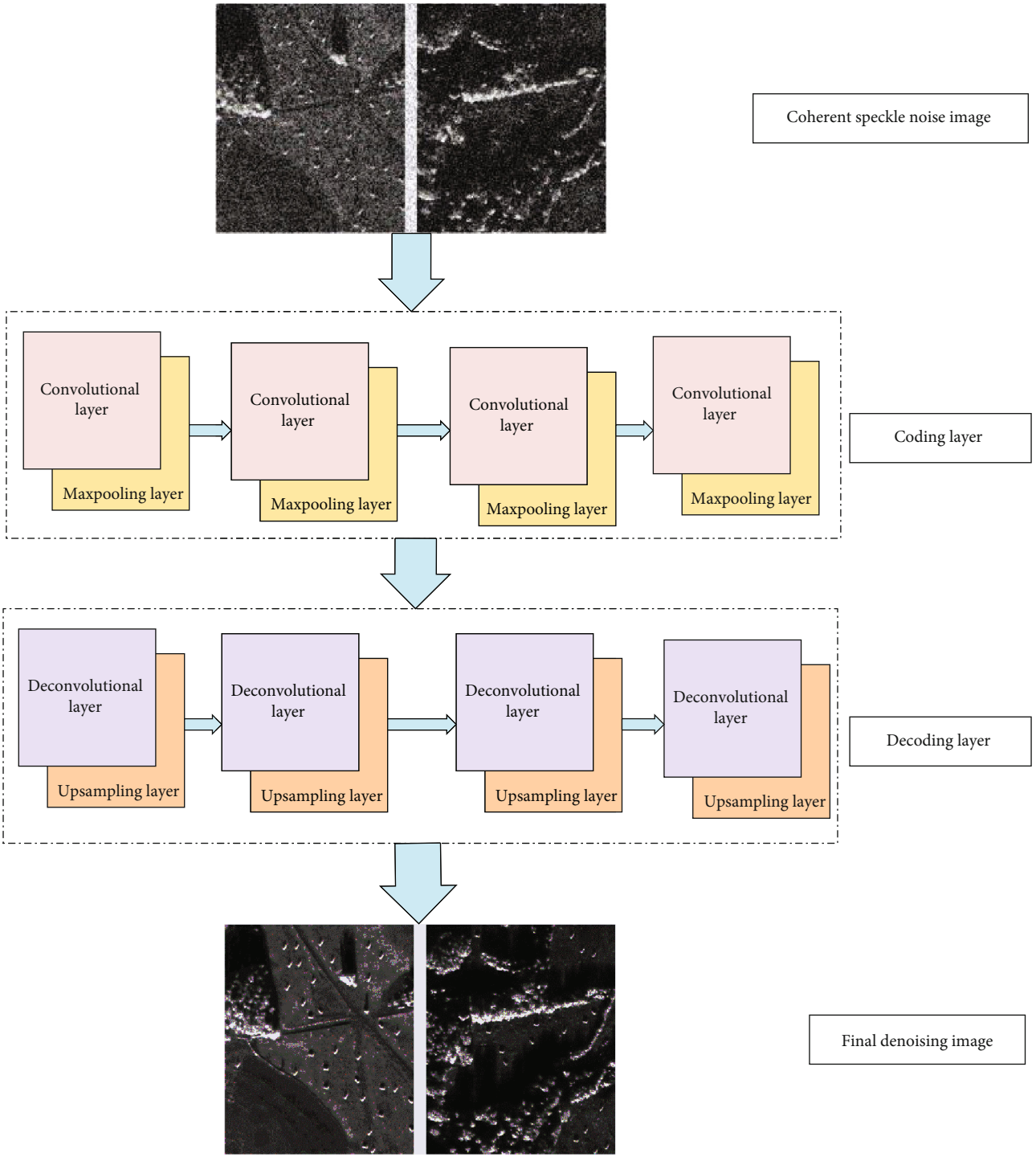



Figure 2: Flow chart and network structure of speckle suppression algorithm.

## 3. SAR Target Recognition

This research proposes a SAR image target classification feature learning framework based on CNN and trains the network model through classification tasks. The trained network performs feature extraction on the region of interest in the SAR image and finally inputs the feature information into the classifier to achieve target classification.

*3.1. SAR Image Speckle Suppression Based on CAE Network.* Convolutional autoencoder (CAE) [21–23] replaces the fully connected neurons in the autoencoder with convolutional neurons. The principle structure of CAE is shown in Figure 1. CAE consists of two parts: encoder and decoder. An encoder is used to extract image features, and a decoder is used to generate images. In the coding process, feature information of the image is saved and redundant information is

deleted [24]. As the capability of feature extraction is enhanced with the deepening of network layers, when deep CAE network is applied to image tasks, the layers of encoder and decoder are usually deepened at the same time to extract, encode, and decode input features.

After reflowing clean data through the big data analysis method, according to the SAR image coherent speckle noise statistical model [25] in Section 2.2, the coherent speckle noise of the SAR image data needs to be processed [19].

The coherent speckle suppression network used in this article combines the structural characteristics of CAE, as shown in Figure 2. First, the noise suppression network needs to be trained with the coherent speckle statistical model. The noise suppression network includes an encoding layer and a decoding layer. The encoding layer contains four identical convolutional layers and pooling layers. The compression factor is 16 after the entire encoding process. The corresponding decoding layer also includes four deconvolution layers and the upper layer combination structure. The decoding layer is finally enlarged by 16 times, and the original image is restored.

As SAR image coherent spot generally unable to get the noise level, in order to make the network generalization performance, this study adopts mixed noise simulation coherent spot data of network training. The cost function of fine-tuning optimization is as follows:

$$L_{\mathrm{CoAE}}\left(x, \widehat{x}\right) = \lambda \sum_{i=1}^{n} \left| x_i - \widehat{x}_i \right| + (1-\lambda) \sum_{i=1}^{n} (x_i - x_i)^2, \quad (3)$$

where $\lambda$ is the super parameter that controls the mixing loss, $x$ represents the image with noise, $\widehat{x}$ represents the estimated image without noise, and $L_{\mathrm{CoAE}}$ represents the reconstruction loss of the entire network. The weights are fine-tuned through the softmax classification layer, which not only effectively avoids network overfitting, but also enables the model to better adapt to the SAR image classification and denoising task.

*3.2. Target Feature Extraction.* In order to perform target recognition, first, need to extract relevant features to distinguish between different targets and then obtain a list of regions of interest (ROI). This paper uses the TP-CFAR algorithm to extract target feature information to provide support for the training of the classifier. The main process of the algorithm is shown in Figure 3.

*3.2.1. DP-CFAR Detection Algorithm.* Double parameter constant false alarm rate (DP-CFAR) is a kind of constant false alarm probability algorithm. The DP-CFAR algorithm [26, 27] refers to the use of mathematical statistics theory to estimate the parameters of the detection model while keeping the target false alarm rate unchanged, which not only reduces the calculation amount of the algorithm, but also adaptively adjusts the threshold in complex background environment.

Figure 4 shows the main steps of the DP-CFAR algorithm. When the DP-CFAR algorithm is used to detect the target image, sliding window method is usually used to traverse the detection and estimate the target scale [22]. In the actual environment, the target detection background environment of SAR images is more complicated, and it is necessary to design a DP-CFAR detection algorithm that adaptively adjusts the threshold. This paper uses the sliding window method to traverse the pixels to detect the background and the target. The size of the sliding window is estimated according to the estimated target scale acquisition. The detection algorithm first calculates the clutter statistical characteristics of the background area and establishes the background clutter distribution probability density function.

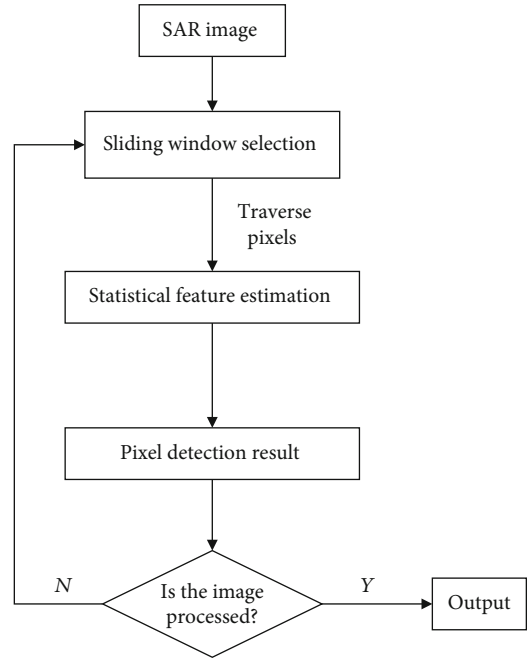



Figure 3: Detection algorithm flow.

Assume that the statistical distribution of clutter pixel values in the background window is Gaussian. Thus, the false alarm rate of two-parameter CFAR can be calculated as follows:

$$P_{\mathrm{fa}} = \int_{VT}^{+\infty} P(x|H_0)dx = \int_{VT}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{(x_i-\mu)^2/2\sigma^2} dx, \quad (4)$$

where $\sigma$ is the standard deviation of the Gaussian distribution, $\mu$ is the mean, and $x_i$ is the value of pixels in the clutter background.

Let $t = x - \mu/\sigma^2$, then the above equation can be written as follows:

$$P_{\mathrm{fa}} = \int_{t}^{\infty} p(t|w_0)dx. \quad (5)$$

When the false alarm rate of dual-parameter CFAR is obtained, according to the formula $\Phi(\alpha) = 1 - P_{\mathrm{fa}}$, the threshold factor $\alpha$ can be calculated. The threshold corresponding to two-parameter CFAR is $V_T$.

$$V_T = \mu + \alpha\sigma. \quad (6)$$

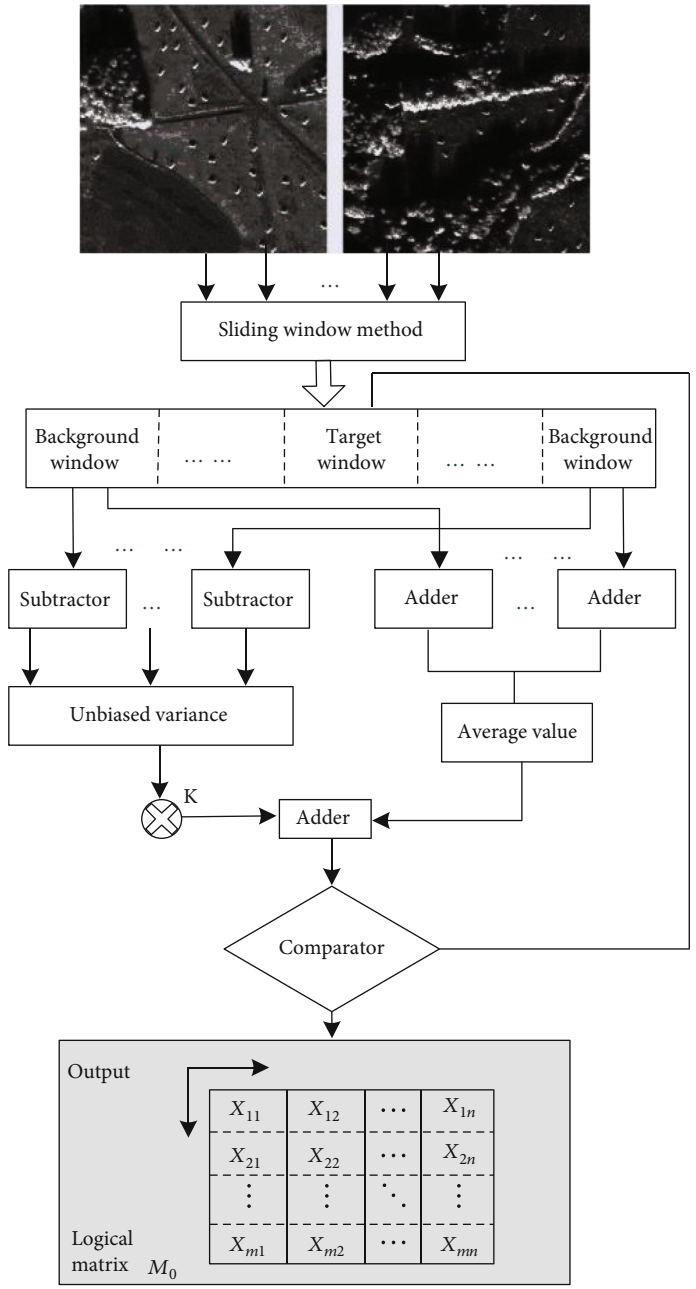



Figure 4: Detection algorithm flow.

Formula (6) also satisfies the condition of constant false alarm rate. The estimated values of Gaussian distribution statistics $\mu$ and $\sigma$ are as follows:

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$
$$\widehat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(xi - \mu\wedge)^2}. \tag{7}$$

In the formula, $\widehat{\mu}$ represents the mean estimation; $\widehat{\sigma}$ represents the standard deviation estimate of the Gaussian distribution. The final discriminant formula of the DP-CFAR detector is as follows:

$$V_T = \frac{x_i - \widehat{\mu}}{\widehat{\sigma}} \begin{cases} \geq \alpha, & x_i \text{ is target,} \\ < \alpha, & x_i \text{ is background.} \end{cases} \tag{8}$$

The sliding window method is used to adaptively estimate the threshold of target detection, and finally, the classification of the pixels to be detected is judged. The judgment standards are as follows:

When $P(x|w_0) > P(t|w_0)$, $x$ is background; when $P(x|w_0) \leq P(t|w_0)$, $x$ is target.

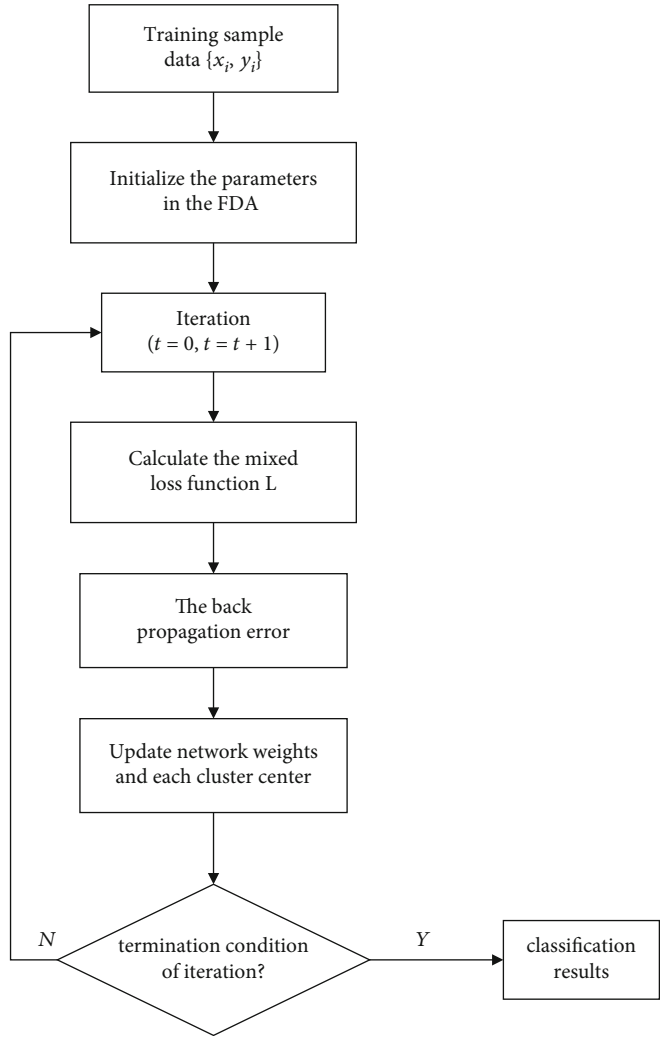



Figure 5: Classification network optimization with FDA.

After detecting the image, the DP-CFAR algorithm outputs a logical matrix with the same size as the image to be detected. The pixel value of the image mask is 0 for background and 1 for target. In order to remove the isolated false alarm points, the calculated logic matrix and mask image are used for processing. In this paper, the image mask is first corroded to cut off the small connections between adjacent targets in the mask; then, expansion is performed to make up for the small holes of the target in the image mask; finally, the prior knowledge is used to estimate the size of the connected components where the target may exist, and the small connected components are removed.

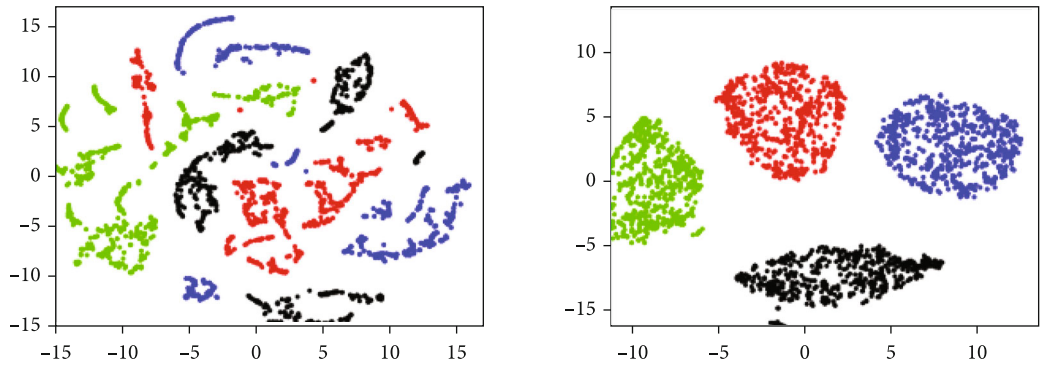
*3.2.2. Generate ROI Based on NPA Algorithm*

(1) *Neighbor Aggregation*. In order to avoid the phenomenon of "increased batches," the influence of noise may cause pixels with pixel values below the DP-CFAR detection threshold to exist in the target area, resulting in a target being detected as several small targets. In this paper, the nearest neighbor point pixel aggregation (NPA) algorithm is used for target aggregation. First, define the set of neighborhood points:

$$N_i = \left\{ S | D(i,j) \le \sqrt{M^2 + P^2} \right\}, \tag{9}$$

where $S$ is the set of spatial neighborhoods of pixel points $(x_i, y_i)$. $D(i,j)$ is the Euclidean distance between pixel points $(x_i, y_i)$ and $(x_j, y_j)$. According to the prior information of the target, $M$ is the number of pixels occupied by the length of the target; $P$ is the number of pixels occupied by the target width, and the number of pixels occupied by the largest scale of the target is usually taken.

(2) This paper uses NPA algorithm to achieve small target aggregation. The removal of false alarm targets mainly uses the target area generated by the aggregation and estimates the target size based on prior knowledge. When the pixel exceeds the set threshold, the false target is removed. Finally, according to the position of the center point of the clustering target area, take




(a) Softmax cross-entropy loss corrected classification results (b) Softmax loss and Fisher loss corrected classification results

Figure 6: Classification results of two correction methods.

the center point as the center point of the ROI, cut out a fixed size region from the image to be detected, and output the coordinate information of the ROI

*3.3. SAR Image Automatic Target Classification and Recognition.* Fishier discriminant analysis (FDA) is a commonly used subspace analysis method. The main idea of the algorithm is that the constraint of calculating the projection axis is to minimize the intraclass divergence and maximize the intersample divergence. Generally speaking, the CNN network is optimized by softmax cross-entropy loss. When the training samples are limited or the diversity is insufficient, the network is easy to overfit, resulting in poor model generalization ability. The intraclass divergence matrix and the interclass divergence matrix of FDA are integrated into the target equation of network optimization, which is used to fine-tune the weight of each network to improve the compactness and class compactness of the feature space in the case of limited training samples, as shown in Figure 5. The separability improves the generalization performance of the entire network. If the loss function of the classification network training only has the softmax classification loss,

$$L_s = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} O\left\{h_i^k = c\right\}\left(w_j^T x^i - \sum_{j=1}^{k} w_j^T x^i\right),$$
$$O\left\{h_i^k = c\right\} = \begin{cases} 1 & h_i^k = c \text{ is true,} \\ 0 & h_i^k = c \text{ is false.} \end{cases} \tag{10}$$

Define intraclass divergence matrix $F_w$ and interclass divergence matrix $F_b$:

$$F_w = \sum_{k=1}^{C}\sum_{i=1}^{Nc}(\mu_i - \bar{\mu}_k)(\mu_i - \bar{\mu}_k)^T,$$
$$F_b = \sum_{k=1}^{C} N_c\left(\vec{\mu} - \bar{\mu}_c\right)\left(\vec{\mu} - \bar{\mu}_c\right)^T, \tag{11}$$

where $C$ represents the number of CNN training categories, $N_c$ is the total number of training samples belonging to the $c$-th category, $\mu_i$ represents the $i$-th training sample, and $\vec{\mu}$ is the mean vector of all training samples: $\vec{\mu} = \sum_{i=1}^{N} u_i/N$. $\bar{\mu}_c$ is the mean vector of training samples belonging to the $c$ type. Then, the FDA algorithm finds the best projection matrix $F_{\text{FDA}}$:

$$F_{\text{FDA}} = \max_w \left[\frac{Tr\left(W^T F_b W\right)}{Tr\left(W^T F_w W\right)}\right]. \tag{12}$$

According to the FDA algorithm, the Fisher loss constraining the interclass and intraclass distance $L_F$ can be obtained:

$$L_F = \frac{1}{cN}\sum_{c=1}^{C}\sum_{i=1}^{Nc} O\left\{h_i^k = c\right\}.h_i^k - \bar{h}_{i2}^{k2} + \frac{\beta}{\bar{h}_i^k - \bar{h}_2^{k2}},$$
$$O\left\{h_i^k = c\right\} = \begin{cases} 1 & h_i^k = c \text{ is true,} \\ 0 & h_i^k = c \text{ is false,} \end{cases} \tag{13}$$
$$L = L_S + \lambda L_F.$$

Table 1: Number of MSTAR data sets.

| Category | Quantity | | | |
|---|---|---|---|---|
| | 15° | 17° | 30° | 45° |
| BMP2 | 587 | 698 | 0 | 0 |
| BTR-70 | 196 | 233 | 0 | 0 |
| T-72 | 582 | 691 | 0 | 0 |
| 2S1 | 274 | 299 | 288 | 303 |
| BRDM-2 | 274 | 298 | 420 | 423 |
| D7 | 274 | 299 | 0 | 0 |
| T-62 | 273 | 299 | 0 | 0 |
| ZIL-131 | 274 | 299 | 0 | 0 |
| ZSU-23-4 | 274 | 299 | 406 | 422 |
| BTR-60 | 195 | 256 | 0 | 0 |
| Total | 3203 | 3671 | 1114 | 1148 |

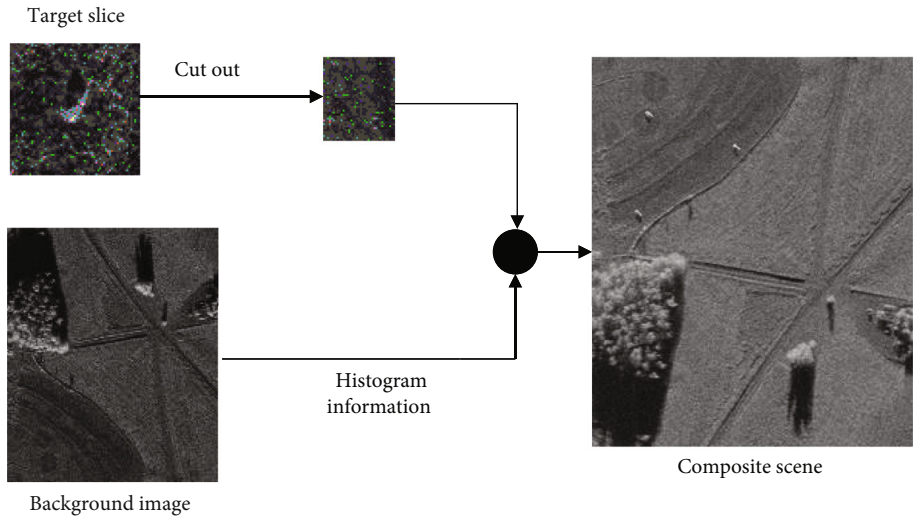



Figure 7: Synthesis process of scene image.

In the formula, $h_i^K$ represents the output of the corresponding feature layer (the first layer before softmax) of the sample $x_i$; $\bar{h}_i^K$ represents the average output of class $C$ samples at the corresponding feature layer; $\bar{h}^K$ represents the average output of all samples in the corresponding feature layer.

Figure 6(a), using the same data set, shows the network model training structure corrected by softmax cross-entropy loss, and Figure 6(b) shows the network model training structure corrected by softmax cross-entropy loss and Fisher loss. The experimental results show that the clustering effect of the same kind of samples in Figure 6(b) is better, and the center spacing of different types of samples is larger, while the sample clustering effect in Figure 6(a) is poor.

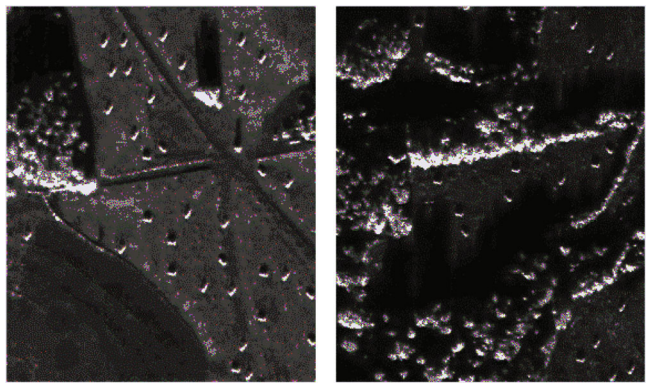


Figure 8: Composite scene diagram.

## 4. Experimental Results and Analysis

In order to evaluate the effectiveness of the proposed SAR image target recognition framework, the experimental part of this paper mainly uses the proposed recognition system to recognize the synthetic SAR image of the target scene to be detected and analyzes the experimental results.

*4.1. Experimental Data.* The experimental data in this section is the MSTAR data set, as shown in Table 1. Due to the high cost of acquiring scene images with a large number of targets, only SAR image slices of 10 types of targets and some background images of the same area are provided in the MSTAR data set. The background image in the MSTAR data set has an imaging elevation angle of 15°, with a total of 100 scene images.

Firstly, preprocess SAR image data through big data analysis methods. Because the background image and the background in the target slice can be regarded as homogeneous regions and both are informed by the same radar, it is recommended to synthesize the scene image containing the target using the target slice and the background in the MSTAR data set.

In this experiment, two scenes to be detected were synthesized, and the synthesis process is shown in Figure 7.

*4.2. Experimental Results and Analysis*

*4.2.1. Speckle Inhibition.* According to the scene synthesis method in Figure 7, two different scene graphs were synthesized in this experiment, as shown in Figures 8(a) and 8(b), and the image size of the two scene graphs is $1474 \times 1784$.

Figure 9 shows the denoising results of synthetic scenes 1 and 2 at different noise levels. It can be seen from Figure 9(a) that with the increase of $L$, the noise level also increases. When $L = 1$, the image is in single-view mode, the image is most destroyed, and the fine texture part of the image (the road in the middle of scene 1) is completely covered by the coherent spots. Judging from the speckle suppression images in Figure 9(b), the restored images of the model under different noise levels are roughly the same, and the speckle suppression effect in homogeneous areas is more obvious (the woods below scene 2).

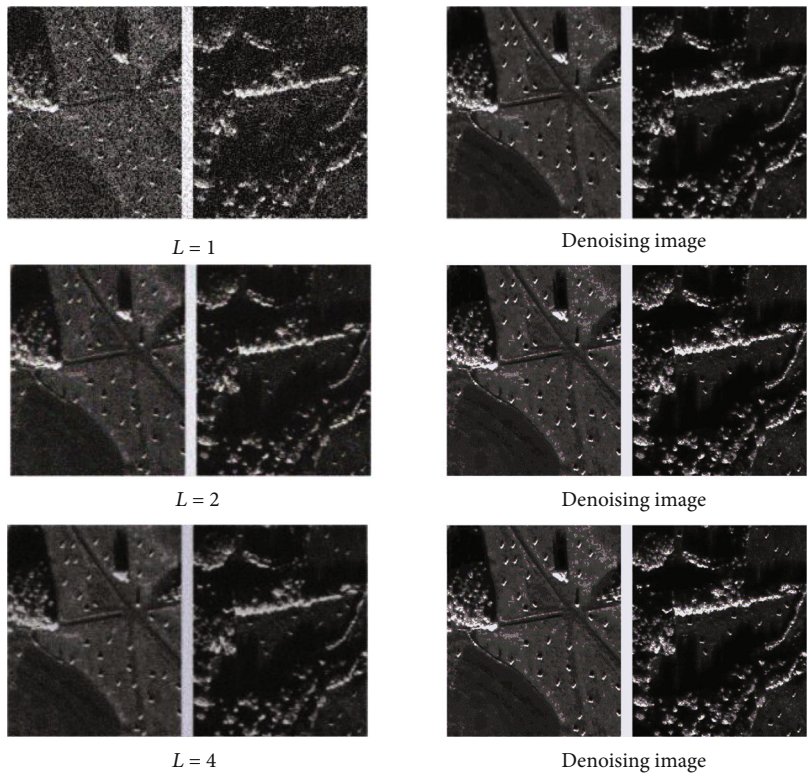



Figure 9: Results of speckle suppression.

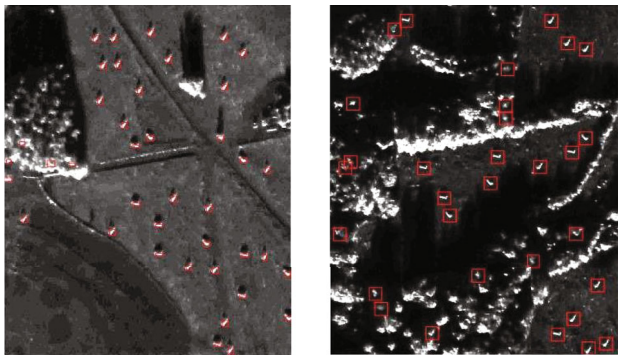



Figure 10: Target recognition results of two scenes.

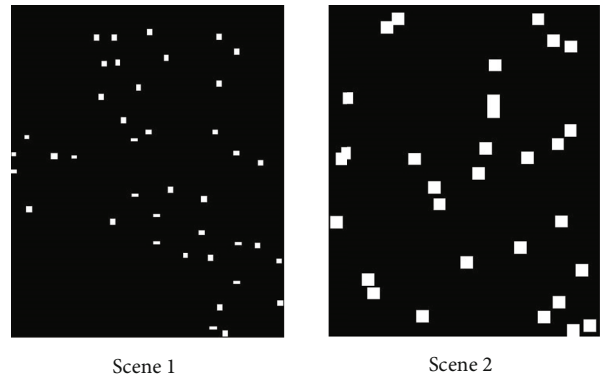
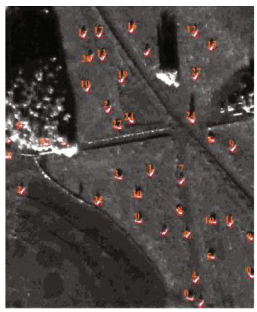
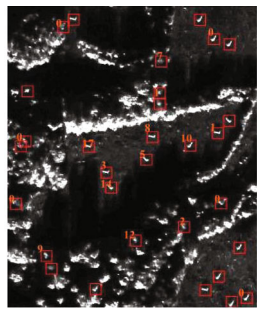
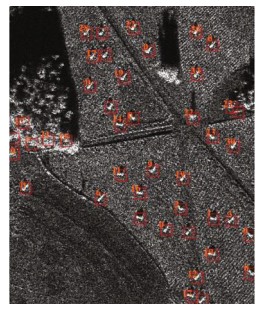
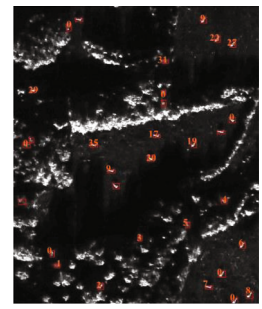



Figure 11: Image mask corresponding to two scenes.

*4.2.2. Target Feature Learning and Detection Results.* Figure 10 shows the results of target detection in scene 1 and scene 2. The experimental results show that a total of 42 target areas are detected in scene 1, including 7 false alarm targets, and there is no missed detection. In scene 2, a total of 35 targets were detected, including 12 false alarm targets, no missed detection, almost all false alarm targets appeared in the bushes, because the gray value of the bushes is relatively large, and the proportion of bushes in the whole scene is relatively high, and it causes more false alarms.

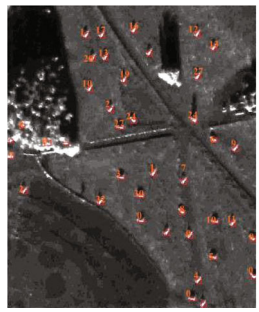
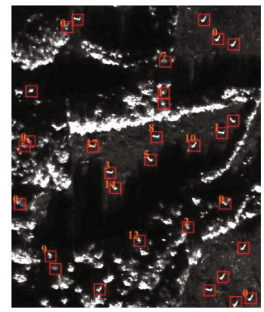
Figure 11 shows the corresponding image masks output by the two scenes of Figure 10 in the target detection module. It can be seen that the shape of the region of interest at this time is square, and there is overlap between the targets, but in general, different categories are distinguished.

*4.2.3. Result of Target Recognition.* According to the target feature learning and detection results, a threshold is set for the classifier to identify false alarms. The threshold is 0.5, that is, when the probability of the current category is greater than 0.5, it is judged as a true category; otherwise, it is a false alarm. The target recognition result using the algorithm in this paper is shown in Figure 12(c). The experimental results show that the number of false alarms in scene 1 is 0, and all categories are successfully recognized. The number of false alarms in scene 2 is 1, and only the bush in the lower left corner is identified as target 6, and the recognition effect is good.

Scene 1 Scene 2 Scene 1 Scene 2

(a) CA-CFAR algorithm (b) Original CFAR algorithm

Scene 1 Scene 2

(c) Proposed algorithm

Figure 12: Identification results.

In order to verify the effectiveness of the algorithm in this paper, this study analyzes the recognition effect of the CA-CFAR algorithm and the traditional CFAR algorithm, and the experimental results are shown in Figures 12(a) and 12(b). In scene 1, CA-CFAR and the algorithm in this paper performed very well, no false positives and false negatives occurred, and the recognition effect was good. However, the traditional CFAR algorithm has false alarms due to noise interference. In scene 2, the performance of the three algorithms is not perfect. However, the false positive rate of the algorithm proposed in this paper is lower than the other two algorithms, the number of false positives is 2 times, and there is no false negative number, and the effect is better.

## 5. Summary and Prospect

This paper has carried out in-depth research on the problem of automatic target recognition in SAR images and designed a SAR image classification feature extraction network framework based on deep convolutional self-encoding CNN, which can use the CFAR algorithm to detect a series of regions of interest at the target location, and at the same time, effective classification features can be automatically learned from training data through classification tasks. To solve the problem of poor network overfitting and generalization performance caused by insufficient data diversity, a loss function was designed based on the Fisher criterion, and the weight of the network was fine-adjusted by the loss function, so that the distance between different categories of samples in the feature space of the network mapping was more discrete and the same samples were more clustered. The trained network model can be used as an effective feature extraction for classification.

With the continuous development of deep learning technology, there are many mature end-to-end optimization recognition systems in the field of ordinary image automatic target recognition. How to combine these algorithms with SAR image automatic target recognition is a promising research direction.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] C. Xu, C. Yin, D. Wang, and W. Han, "Fast ship detection combining visual saliency and a cascade CNN in SAR images," *IET Radar, Sonar & Navigation*, vol. 14, no. 12, pp. 1879–1887, 2020.

[2] X. Xue, "Research on SAR image processing technology," Science and Technology Press, 2017.

[3] Z. Liu, B. Qu, and J. Guo, "Target recognition of SAR images using fused deep feature by multiset canonical correlations analysis," *Optik*, vol. 220, article 165156, 2020.

[4] Z. Qiang, *Analysis of SAR Image Automatic Target Recognition Technology*, Information Technology, 2018.

[5] T. R. Gadekallu, D. S. Rajput, M. Reddy et al., "A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU," *Journal of Real-Time Image Processing*, vol. 18, no. 4, pp. 1383–1396, 2021.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, http://arxiv.org/abs/1409.1556.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, NV, USA, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, Honolulu, HI, USA, 2017.

[11] A. El Housseini, A. Toumi, and A. Khenchaf, "Deep learning for target recognition from SAR images," in *2017 seminar on detection systems architectures and technologies (DAT)*, pp. 1–5, Algiers, Algeria, 2017.

[12] J. Wang, T. Zheng, P. Lei, and X. Bai, "Ground target classification in noisy SAR images using convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 4180–4192, 2018.

[13] S. Wagner, "Combination of convolutional feature extraction and support vector machines for radar ATR," in *17th international conference on information Fusion (FUSION)*, pp. 1–6, Salamanca, Spain, 2014.

[14] S. Chen, H. Wang, F. Xu, and Y. Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4806–4817, 2016.

[15] E. Zelnio, F. D. Garber, and D. Morgan, "Deep convolutional neural networks for ATR from SAR imagery," in *Algorithms for synthetic aperture radar imagery XXII*, Baltimore, Maryland, United States, 2015.

[16] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 99, pp. 1–1, 2020.

[17] D. Geng, "SAR image ship detection data mining method under big data analysis," *Ship Science and Technology*, vol. 42, no. 22, pp. 61–63, 2020.

[18] Z. U. Rehman, M. S. Zia, G. R. Bojja, M. Yaqub, F. Jinchao, and K. Arshid, "Texture based localization of a brain tumor from MR-images by using a machine learning approach," *Medical Hypotheses*, vol. 141, article 109705, 2020.

[19] H. Shen, C. Zhou, J. Li, and Q. Yuan, "SAR image despeckling employing a recursive deep CNN prior," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 273–286, 2021.

[20] G. Reddy and N. Neelu, "Hybrid firefly-bat optimized fuzzy artificial neural network based classifier for diabetes diagnosis," *International Journal of Intelligent Engineering and Systems*, vol. 10, no. 4, pp. 18–27, 2017.

[21] K. O. Oyebade and D. Kamil, "Pattern recognition: invariance learning in convolutional auto encoder network," *International Journal of Image*, vol. 8, no. 3, pp. 19–27, 2016.

[22] M. R. Ayaluri, K. Sudheer Reddy, S. R. Konda, and S. R. Chidirala, "Efficient steganalysis using convolutional auto encoder network to ensure original image quality," *PeerJ Computer Science*, vol. 7, p. e356, 2021.

[23] G. R. Bojja, M. Ofori, J. Liu, and L. S. Ambati, *Early Public Outlook on the Coronavirus Disease (COVID-19): A Social Media Study*, AMCIS 2020 Proceedings, 2020.

[24] X. Y. Wu, Y. C. Jiang, Z. Lv, and H. Kuang, "Target recognition of SAR image based on improved convolutional auto-encoding network," in *2019 6th Asia-Pacific conference on synthetic aperture radar (APSAR)*, pp. 1–5, Xiamen, China, 2019.

[25] F. Khaldi, F. Soltani, and M. Baadeche, "Fuzzy CFAR detectors for MIMO radars in homogeneous and non-homogeneous Pareto clutter," *Journal of Communications Technology and Electronics*, vol. 66, no. 1, pp. 62–69, 2021.

[26] M. Sahal, Z. A. Said, R. Y. Putra, R. E. A. Kadir, and A. A. Firmansyah, "Comparison of CFAR methods on multiple targets in sea clutter using SPX-radar-simulator," in *2020 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pp. 260–265, Surabaya, Indonesia, 2020.

[27] A. Abu and R. Diamant, "CFAR detection algorithm for objects in sonar images," *IET Radar, Sonar & Navigation*, vol. 14, no. 11, pp. 1757–1766, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3373535, 12 pages
https://doi.org/10.1155/2021/3373535

*Research Article*

# Computer-Aided Teaching System Based on Data Mining

**Yonghua Tang**,[1] **Qiang Fan**,[2] **and Peng Liu**[1]

[1]*LuXun Academy of Fine Arts, Dalian 116600, China*
[2]*Criminal Investigation Police University of China, Shenyang 110035, China*

Correspondence should be addressed to Peng Liu; liupengpromise@lumei.edu.cn

Received 16 July 2021; Revised 9 September 2021; Accepted 21 September 2021; Published 11 October 2021

Academic Editor: Rajesh Kaluri



The traditional teaching model cannot adapt to the teaching needs of the era of smart teaching. Based on this, this paper combines data mining technology to carry out teaching reforms, constructs a computer-aided system based on data mining, and constructs teaching system functions based on actual conditions. The constructed system can carry out multisubject teaching. Moreover, this paper uses a data mining system to mine teaching resources and uses spectral clustering methods to integrate multiple teaching resources to improve the practicability of data mining algorithms. In addition, this paper combines digital technology to deal with teaching resources. Finally, after building the system, this paper designs experiments to verify the performance of the system. From the research results, it can be seen that the system constructed in this paper has certain teaching and practical effects, and it can be applied to a larger teaching scope in subsequent research.

## 1. Introduction

With the continuous development and popularization of information technology, comprehensive informatization has become the inevitable development direction of this era. This is especially true in the field of education, and digital education makes teaching content more reliable and has a uniform standard [1]. From the recording and broadcasting online school in the late 1990s to the higher education resource sharing platform, "Love Course," which provides complete video resources, to the "China University MOOC" online open course class that integrates teaching, Q&A, testing, and homework teaching, to the "NetEase Cloud Classroom" that can realize real-time interactive online live teaching [2], modern education is closely related to the information age, and building a teaching platform based on information technology is the direction of reform of teaching informatization in major colleges and universities [3].

Teaching decisions are the process of analyzing, judging, exploring, and choosing teaching implementation plans in order to achieve teaching objectives. Teaching is a controlled dynamic complex system, to achieve effective control of system motion, requires timely confirmation of various elements and its interaction relationships in the system, and makes decisions according to the corresponding teaching principles. Data is an important basis for teaching decisions. The smart teaching platform provides data on the status of the learner but is currently only an evaluation method for students. In traditional teaching activities, students are often collected in the forms of students or exams due to the collection and analysis of students' data, and the teaching behavior of teachers has a great impact. With today's scientific and technical assistance, the extraction of teaching data is no longer a problem, but the use of teaching data is used in teaching evaluation, teaching management, statistical attendance, etc., and the teaching of teachers is not much. In summary, in the environment of education big data, how to help teachers give full play to the teaching process and how to use the data in the smart teaching platform to optimize the teaching decisions are the core issues of the research in education.

This paper applies data mining technology to computer-assisted teaching to build a data mining-based smart teaching system and improves the traditional teaching mode to improve the teaching effect.

## 2. Related Work

Regarding the characteristics and value of smart classrooms, literature [4] believed that mobile terminal-based interaction can help improve children's developmental learning and social skills. Literature [5] believed that with the help of wireless smart devices, learners' participation in classroom learning can be improved. Literature [6] believed that the teaching terminal based on the smart classroom can clarify the geographical location and learning progress of the students, determine the current teaching activities of students, recommend learning resources according to the needs of students, and support effective real-time collaboration and resource sharing between teachers and students, as well as students and students. Literature [7] believed that a smart classroom is a new type of classroom, and the fundamental goal is to cultivate students' wisdom. The smart classroom focuses not only on the knowledge level of the learner and the scores obtained in the exam but also on stimulating the potential of students and focusing on cultivating students' wisdom. Literature [8] believed that the cultivation of wisdom should exist throughout the entire classroom teaching process and use experience and accumulated thinking experience to enhance wisdom so as to achieve the ultimate goal of using wisdom to solve problems. Literature [9] reshaped and upgraded the flipped classroom and proposed a breakthrough from a flipped classroom to a smart classroom in terms of resource quality improvement and teaching method optimization.

As a probe into the teaching mode of the theoretical level in a wisdom classroom, Literature [10] believes that learners can personalize and autonomously learn from their own rhythm in wisdom class. Document [11] is a research on the autonomous constructive processing and treatment of learners in the smart classroom. Literature [12] believed that wisdom learning is in a contextual environment, providing students with a wide range of learning resources for students, and promotes new learning paradigms of education development. In order to achieve smart learning, literature [13] designed a smart learning system model including cloud computing, learning analysis technology, and mixed reality, reflecting the three major characteristics in the process of education, interactive, personalized custom, and independent control. Document [14] shows that on the basis of studying the intelligence learning connotation, the conceptual framework of wisdom learning is built, and four wisdom learning models are designed. Literature [15] believed that the technical characteristics of the wisdom class have designed a learning model based on the wisdom class, and the application research of the learning model is designed. Literature [16] gave a "three-section ten-step" structural model of the wisdom class through the comparison with traditional classroom teaching processes. As a research on wisdom classroom teaching practice, Literature [17] reshaped and upgraded the wisdom classroom learning environment from hardware and software, thereby solving the failure of teachers and students in LMS (Learning Management Services). Document [18] shows that based on the ITLA (Integrated Teaching and Learning Assistance) system, it studied important factors that determine the effective development of the wisdom class. Literature [19] is supported by the HiTeach Interactive System, compared to traditional classrooms and wisdom classroom teaching, exploring the positive significance of teaching in teaching. About wisdom classroom teaching evaluation, Literature [20] shows the response to the teaching strategy of the wisdom classroom teaching under the network learning space while presenting the teaching strategy of the teacher's teaching behavior and the student's learning behavior for teaching evaluation.

## 3. Spectral Clustering-Aided Teaching Algorithm

Spectral clustering is a very popular research field in cluster analysis. Its main idea is to obtain a graph cut through the feature decomposition of the graph (Laplacian matrix). It is a clustering method based on a graph cut. This research mainly introduces the spectral graph theory, the graph partition method, the spectral clustering algorithm of spectral clustering, and the problems existing in spectral clustering at present.

We first give a set $X = \{x_1, x_2, \cdots, x_n\}$, $x_i \in R^l$, which contains $n$ data points. We assume that each data sample is regarded as a vertex $V$ in the graph, and the edge $E$ between the vertices is assigned a weight value $W$ according to the similarity between the samples so that an undirected weighted graph $G = (V, E)$ based on the sample similarity is obtained. It can be defined with the following formula [21]:

$$w_{ij} = \exp\left(-\frac{d^2(x_i, x_j)}{2\delta^2}\right). \tag{1}$$

Among them, $w_{ij}$ is the similarity value between points $x_i$ and $x_j$, and $d(x_i, x_j)$ is the Euclidean distance between points $x_i$ and $x_j$, and $\delta$ is the scale parameter, which controls the speed at which the similarity value $w_{ij}$ decays with the Euclidean distance $d(x_i, x_j)$.

The clustering problem can be expressed as a cut problem on the graph. The result of the cut is as far as possible to minimize the similarity between the two subgraphs and maximize the internal similarity. The quality of the clustering results is directly related to the quality of the cut criteria. As shown in Figure 1, through a cut strategy, H can be divided into one category, and all other points can be divided into another category. Alternatively, another classification strategy can be used to classify H, A, B, and C into one category and D, E, G, and F into another category. Obviously, the second cut strategy can divide undirected graphs into two more balanced categories. Common cut criteria include the minimum cut criterion, the normalized cut criterion, the ratio cut criterion, and the average cut criterion.

(1) Minimum cut set criterion

Now, we first consider the simplest case: two-way cut. The undirected weighted graph $G$ is divided into two subsets $A$ and $B$ that do not want to intersect, where $A \cap B = \varnothing$, $A$

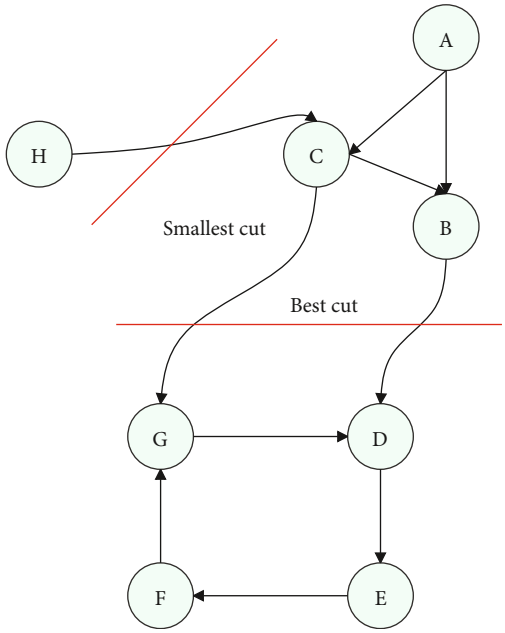



Figure 1: Undirected graph partition of spectral clustering.

$\cup B = V$. The easiest way is to take the method of minimizing the cut to produce two disjoint subsets. Cutting is defined as [22]

$$\text{cut}(A, B) = \sum_{p \in A, q \in B} w(p, q). \tag{2}$$

$w(p, q)$ is the similarity between point $p$ and point $q$. When the data point is in Euclidean space $R^d$, a reasonable default candidate similarity function is the Gaussian similarity function, which is defined as follows:

$$w_{ij} = \exp\left(\frac{-\|x_i - x_j\|}{2\delta^2}\right). \tag{3}$$

We used a clustering algorithm based on the graph theory to solve the segmentation problem, and we obtained very good segmentation results in the experiment. But at the same time, the minimum cut criterion is easy to produce unbalanced results: one type contains most data points, while the other type contains only a few data points or even only one data point. This is because the minimum cut does not consider the size of the cluster. In order to solve this problem, we proposed the normative cut set criterion. The experimental results show that this cutting method can obtain relatively balanced clustering results.

(2) Normative cut set criterion (Ncut)

The normative cut set criterion focuses on the data in the global scope, not only the local solution of the dataset. The Ncut algorithm has two-way partition and multiway partition. The two-way cut only uses the eigenvector corresponding to the second smallest eigenvalue of the Laplacian matrix for segmentation. However, the multipath cut uses $k$ eigenvectors starting from the eigenvector corresponding to the second smallest eigenvalue to cluster together, where $k$ is a predetermined constant. For the two-partition problem of cutting $V$ into two regions, $A$ and $B$, the objective function of the Ncut's optimized graph partition is shown in the following formula:

$$\text{Ncut}(A, B) = -\frac{\text{cut}(A, B)}{\text{assoc}(A, V)} + \frac{\text{cut}(A, B)}{\text{assoc}(B, V)}. \tag{4}$$

Among them,

$$\text{cut}(A, B) = \sum_{p \in A, q \in B} W(p, q), \tag{5}$$

is the minimum cut set shown above, and

$$\text{assoc}(A, V) = \sum_{p \in A, q \in B} W(p, q), \tag{6}$$

is the total number of connections from all data points in $A$ to all data points in graph $G$. From the criterion function, it can be seen that the normalized cut set criterion considers the separation state of the two subsets, so it can avoid unbalanced clustering results.

At the same time, Shi and Malik proposed another standard to measure the total normalized association criterion for a given divided area. Its formula is as follows:

$$\text{Nassoc}(A, B) = \frac{\text{assoc}(A, A)}{\text{assoc}(A, V)} + \frac{\text{assoc}(B, B)}{\text{assoc}(B, V)}. \tag{7}$$

Among them, $\text{assoc}(A, A)$ and $\text{assoc}(B, B)$ are the total weights of the edges connecting the nodes in $A$ and $B$, respectively. By inference, the relationship between these two division standards is $\text{Ncut}(A, B) = 2 - \text{Nassoc}(A, B)$. Obviously, in order to achieve a better clustering effect, it is necessary to minimize the noncorrelated function and maximize the correlation function at the same time.

The above criteria are all two-way division methods for undirected graphs. We proposed a method to divide the graph into $k$ subgraphs. The objective function of the $k$-dimensional normalized cut set criterion can be written as follows:

$$\text{MNcut}(A_1, \cdots, A_k) = \sum_{i=1}^{k} \frac{\text{cut}\left(A_i, \overline{A_i}\right)}{\text{vol}(A_i)}. \tag{8}$$

Among them, $\text{vol}(A) = \sum_{i \in A} d_i$ is the number of data in dataset $A$. $A_k$ is the complement of dataset $A$. When $k = 2$, multichannel division is equal to two-channel division.

(3) Average cut set criterion (Avcut)

The objective function of the average cut set criterion is as follows:

$$\text{Avcut} = \frac{\text{cut}(A, B)}{|A|} + \frac{\text{cut}(A, B)}{|B|}. \tag{9}$$

It can be seen from the formulas of the normative cut set criterion and the average cut set criterion that these two formulas both express the relationship between the boundary loss and the correlation of the divided regions in the undirected weighted graph $G$ in the form of the sum of ratios. This shows that the objective function of the minimum scale cut set criterion and the normalized cut set criterion can be cut more accurately. Their common disadvantage is that it is easy to segment very small subgraphs containing only a few points, and they are undersegmented. It can be seen from the experimental results in the literature that the cut result of the normative cut set criterion is better than that of the average cut set criterion.

(4) Ratio cut set criterion (Rcut)

The formula of the ratio cut set criterion function is as follows:

$$\text{Rcut} = \frac{\text{cut}(A, B)}{\min(|A|, |B|)}. \tag{10}$$

Among them, $|A|$ and $|B|$ represent the number of vertices of subgraphs $A$ and $B$, respectively.

The advantage of this criterion is that when minimizing the criterion function, only the minimum similarity between classes needs to be considered, which reduces the possibility of oversegmentation, but the disadvantage is that the operating efficiency is too low.

(5) Minimum-maximum cut set criterion (Mcut)

The min-max cut criterion maximizes $\text{assoc}(A, A)$ and $\text{assoc}(B, B)$ while minimizing $\text{cut}(A, B)$. The objective function of this criterion is as follows:

$$\text{Mcut} = \frac{\text{cut}(A, B)}{\text{assoc}(A, A)} + \frac{\text{cut}(A, B)}{\text{assoc}(B, B)}. \tag{11}$$

Among them, $\text{assoc}(A, A)$ and $\text{assoc}(B, B)$ are the total weights of the edges connecting the nodes in $A$ and $B$, respectively. By minimizing the criterion function, undersegmentation or only segmentation of smaller subgraphs containing a few vertices can be avoided. Therefore, by minimizing the minimum-maximum cut set criterion function, a more balanced cut set can be obtained, but the realization speed is relatively slow. Both the minimum-maximum cut set criterion and the normative cut set criterion can satisfy the clustering principle that the similarity within a class is small, but the similarity between classes is large. The difference is that when the overlap between classes is too large, the cut effect of the normative cut set criterion is not as good as the minimum-maximum cut set criterion.

The adjacency matrix (denoted as $W$ or $A$) is also called the similarity matrix, and the Laplacian matrix (denoted as $L$) is a common representation of graphs. The similarity matrix of the weighted graph uses real numbers to reflect the different relationships between the vertices. The elements in this matrix can be expressed by the following formula:

$$w_{ij} = \exp\left(-\frac{d^2(x_i, x_j)}{2\delta^2}\right), \quad i \neq j. \tag{12}$$

Among them, $x_i$ is the $i$-th sample point in the dataset, and $d(x_i, x_j)$ is the distance between the sample point $x_i$ and the sample point $x_j$. The distance can use any form of distance function, and the more commonly used is the Euclidean distance $\|x_i - x_j\|$. In the formula, $\delta$ is the nuclear radius, which is a parameter that needs to be given in advance. In the clustering algorithm, the attenuation rate $w_{ij}$ is constrained by the parameter $\delta$, so a proper value of $\delta$ must be given to improve the clustering accuracy of the algorithm. The row vectors in $w_{ij}$ represent the distribution of the dataset, and they are usually distributed on the hypersphere of the $k$-dimensional space. Researchers usually use degrees to represent the distribution of the dataset around the point, and the diagonal matrix composed of all the degree values as diagonal elements is the degree matrix, which is usually represented by $D$:

$$D_{ij} = \sum_{i=1}^{n} w_{ij}. \tag{13}$$

Among them, $n$ is the number of sample points.

The Laplacian matrix is $L = D - W$, where $D$ is the degree matrix and $W$ is the similarity matrix. Most spectral clustering algorithms cut graphs based on the spectrum of the Laplacian matrix. Laplacian matrices can be divided into two types: nonnormalized Laplacian matrix $L$ and normalized Laplacian matrix. The normalized Laplacian matrix includes a symmetric form (denoted as $L_s$) and a random walk form (denoted as $L_r$).

There are many spectral clustering algorithms. The difference lies in how to choose the object function and how to construct the affinity matrix of the graph, but the basic framework is the same.

*Step 1.* According to the given dataset, the algorithm constructs a graph matrix, and there are different methods for different situations.

*Step 2.* The algorithm solves the first $k$ eigenvectors of the matrix and constructs the eigenspace $R^k$.

Typical spectral clustering algorithms such as the Shi and Marik algorithm; Kan R, Vempala S, and Vetta A algorithm; Ng, Jordan, and Weiss algorithm; link algorithm; and Markov random walk algorithm have achieved the expected application effect.

(1) SM algorithm

The normative cut set criterion is a very popular technique in spectral clustering, and it has achieved good results

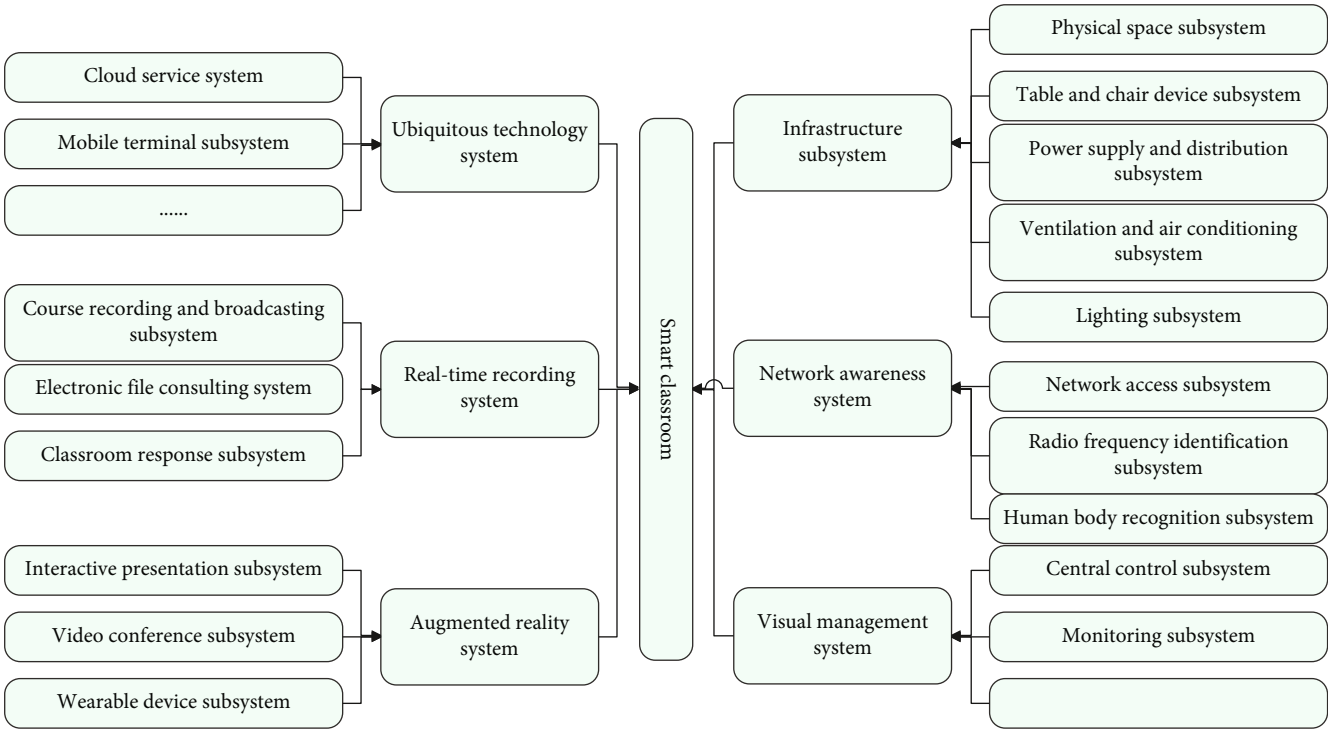



Figure 2: Smart classroom "I-SMART" model.

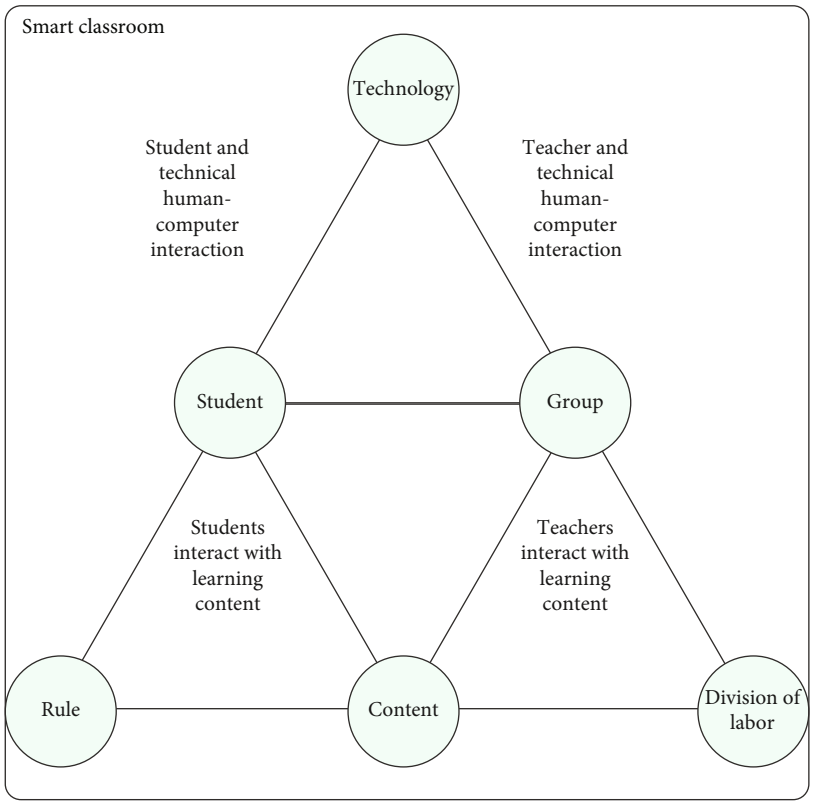



Figure 3: Interactive teaching model in a smart classroom environment.

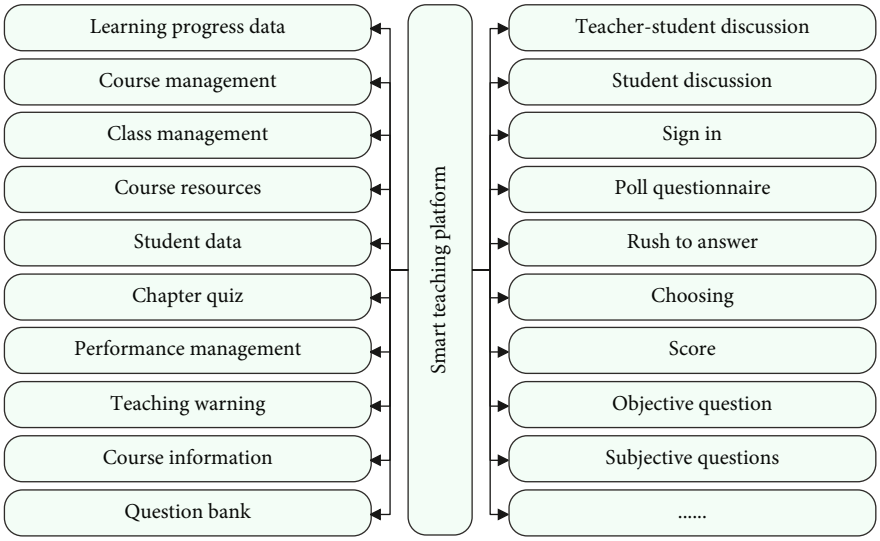



Figure 4: Data in the smart teaching platform.

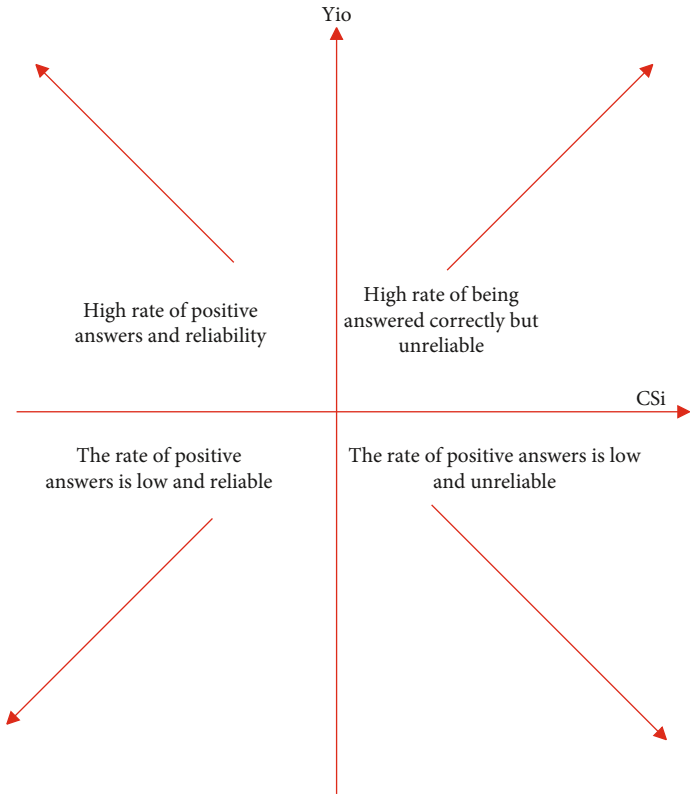



Figure 5: Yio-CSi model.

in the application of image segmentation. The normative cut set criterion can be described as follows:

$$\begin{cases} \min \quad \mathrm{Ncut}(A,B) = \min \dfrac{x^T(D-W)x}{x^TDx} \\ \text{s.t.} \quad x^T\mathrm{We} = x^TDe = 0. \end{cases} \tag{14}$$

This is an NP-hard problem. Fortunately, the problem can be solved by relaxing the discrete constraint of $x$. We assume that $x$ is a real value, the objective function can be solved by the Rayleigh quotient, and this problem can be transformed into the second smallest eigenvalue of the solving formula $(D-W)x = \lambda Dx$. The SM algorithm can be described as follows.

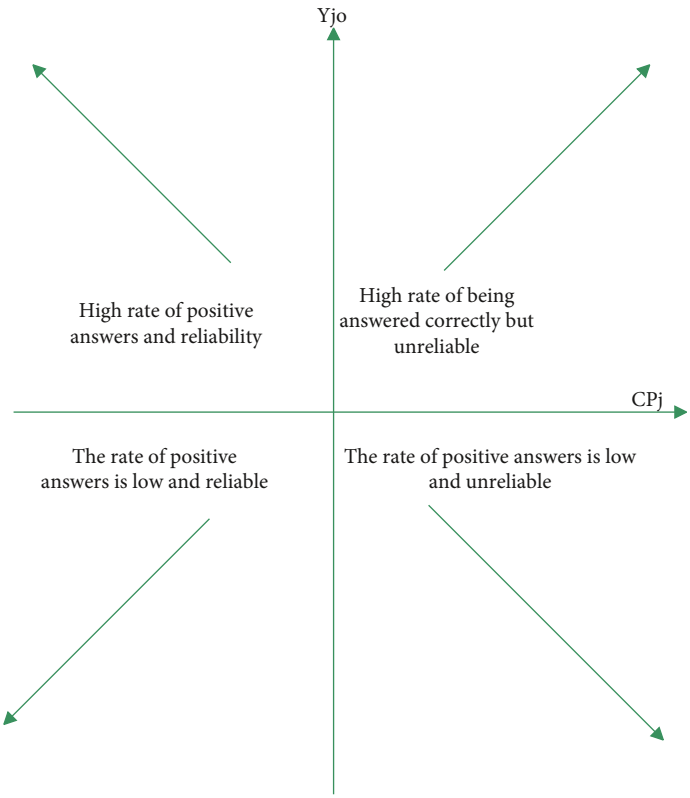
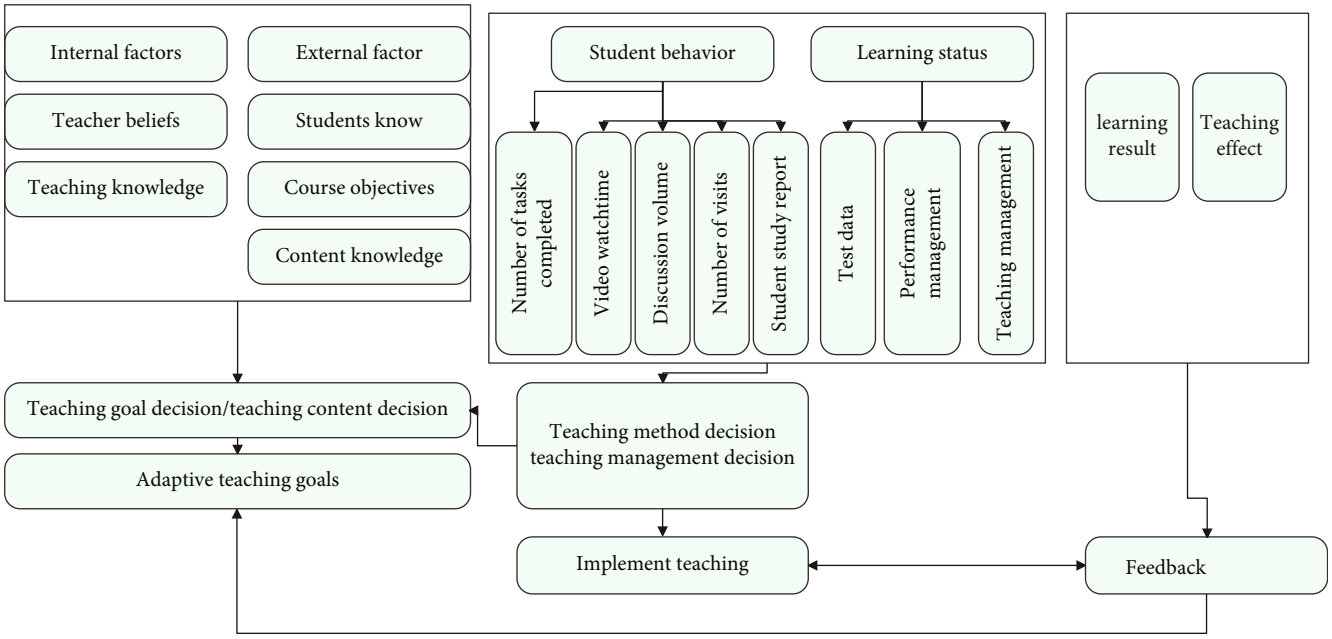



Figure 6: Yjo-CPj model.




- Teaching planning stage
- Interactive teaching stage
- Teaching evaluation stage

Figure 7: Teaching decision model based on teaching data.

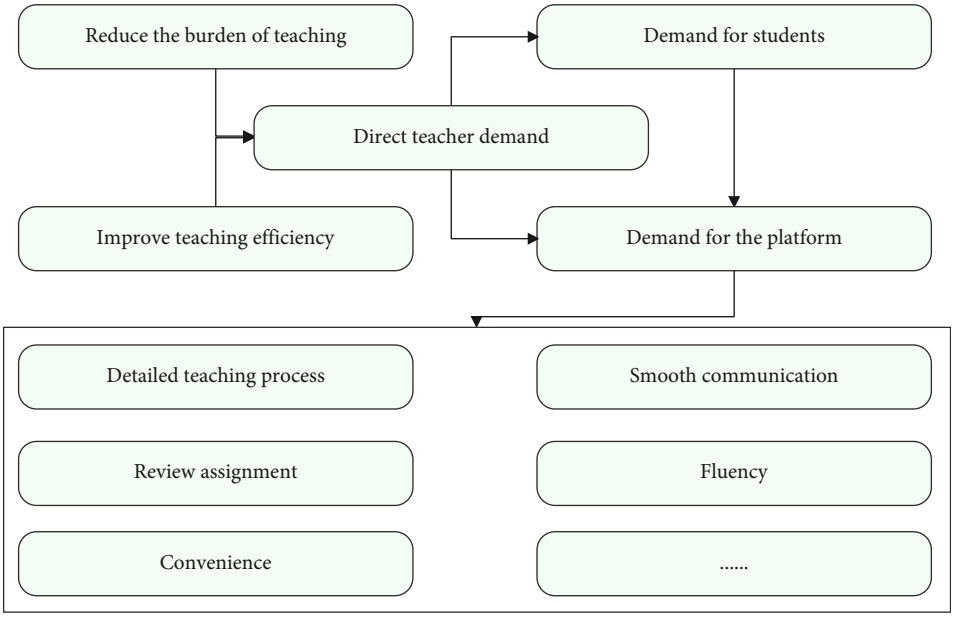



Figure 8: Teacher-platform-student needs analysis.

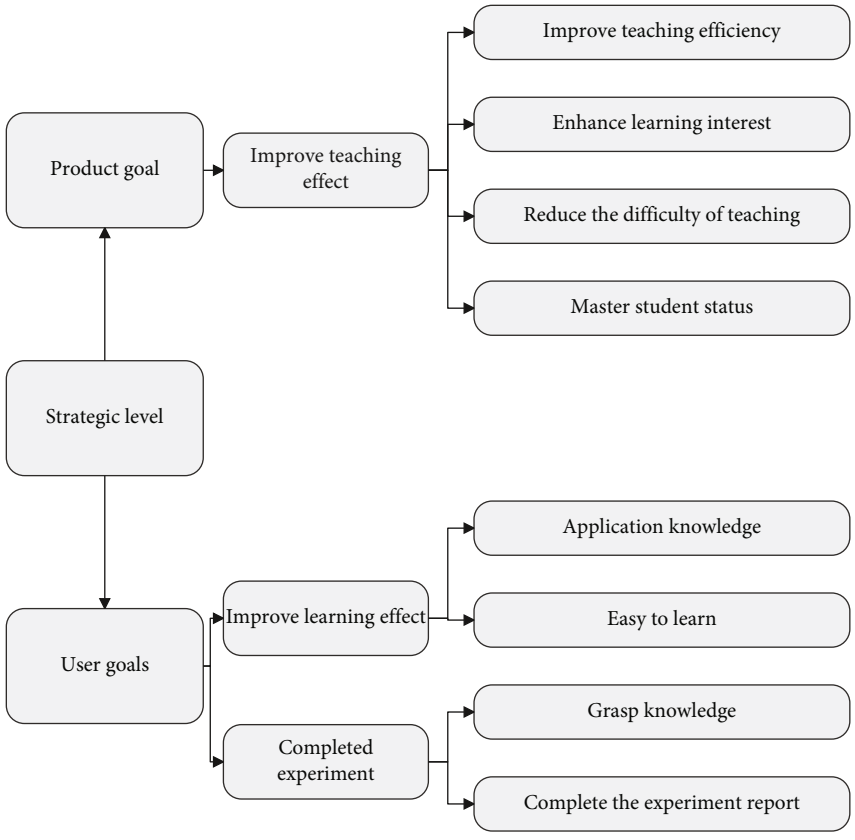



Figure 9: The strategic layer of the smart teaching platform.

*Step 1.* The algorithm first constructs the similarity matrix $W \in R^{N\times N}$ and calculates the Laplacian matrix according to the formula $L = D - W$.

*Step 2.* The algorithm calculates the first $k$-generated eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_k$ and obtains the corresponding eigenvector $v_1, v_2, \cdots, v_k$ according to the formula $(D - W)x = \lambda Dx$.

*Step 3.* The algorithm uses the eigenvectors calculated in the second step to construct a matrix $Q \in R^{N\times K}$.

Table 1: Statistical table of the evaluation of the mining effect of teaching resources.

| No. | Resource mining | No. | Resource mining | No. | Resource mining |
|---|---|---|---|---|---|
| 1 | 92.9 | 21 | 91.4 | 41 | 84.1 |
| 2 | 77.4 | 22 | 90.6 | 42 | 77.7 |
| 3 | 82.6 | 23 | 81.0 | 43 | 91.2 |
| 4 | 89.6 | 24 | 84.1 | 44 | 79.3 |
| 5 | 83.8 | 25 | 78.4 | 45 | 92.1 |
| 6 | 91.6 | 26 | 90.1 | 46 | 91.4 |
| 7 | 82.1 | 27 | 81.3 | 47 | 87.8 |
| 8 | 83.7 | 28 | 83.3 | 48 | 79.0 |
| 9 | 90.5 | 29 | 92.6 | 49 | 78.3 |
| 10 | 92.6 | 30 | 83.8 | 50 | 92.0 |
| 11 | 90.2 | 31 | 84.0 | 51 | 84.1 |
| 12 | 88.3 | 32 | 92.8 | 52 | 90.5 |
| 13 | 91.5 | 33 | 81.1 | 53 | 92.4 |
| 14 | 89.5 | 34 | 90.2 | 54 | 78.5 |
| 15 | 86.7 | 35 | 92.1 | 55 | 77.1 |
| 16 | 92.7 | 36 | 80.0 | 56 | 81.3 |
| 17 | 80.1 | 37 | 85.8 | 57 | 88.1 |
| 18 | 92.6 | 38 | 82.1 | 58 | 88.9 |
| 19 | 90.0 | 39 | 85.2 | 59 | 81.6 |
| 20 | 87.2 | 40 | 82.9 | 60 | 91.6 |

*Step 4.* The algorithm uses the second smallest eigenvalue and the corresponding Fiedler vector to find the cutting point to cut the graph through the Fiedler vector so that Ncut is the smallest. In the Fiedler vector, the larger than this point is divided into one category, and the smaller than this point is divided into another category. The normalized adjacency matrix defined by the algorithm is as follows:

$$N = D^{-1/2}WD^{-1/2} : N(i,j) = \frac{W(i,j)}{\sqrt{D(i,i)D(j,j)}}. \quad (15)$$

*Step 5.* The algorithm uses the $k$-means algorithm to cluster the matrix $O$ into $k$ categories $C_1, C_2, \cdots, C_k$.

The computational complexity of solving the eigenvalue problem of all eigenvectors is $O(n^3)$, where $n$ is the number of input sample sets.

(2) NJW algorithm

The above SM algorithm only uses Fiedler vectors, but the NJW algorithm uses k feature vectors at the same time. Because when calculating the k-way partition, using more feature vectors will achieve better results. The NJW algorithm can be described as follows:

*Input*: $N$ data points $\{x_i\}_{x=1}^{N}$

*Output*: cluster $A_1, A_2, \cdots, A_k$, where $A_i = \{j \mid r_j \in c_i\}$

(3) Markov random walk algorithm

*Step 1.* The algorithm constructs the similarity matrix $W \in R^{N\times N}$, the diagonal matrix $D$, and the Laplacian matrix $L = D^{-1/2}WD^{-1/2}$ through the formula $w_{ij} = \exp(-d^2(x_i, x_j)/2\delta^2)$, $i \neq j$.

*Step 2.* The algorithm calculates the first $k$ eigenvectors of the Laplacian matrix $v_1, v_2, \cdots, v_k$.

*Step 3.* The algorithm uses the eigenvector $v_1, v_2, \cdots, v_k$ calculated in the second step to construct a matrix $Q \in R^{N\times K}$.

*Step 4.* The algorithm forms a normalized matrix $M$ by normalizing the rows to norm 1, where $m_{ij} = x_{ij}/\sqrt{\sum_k x_{ik}^2}$.

*Step 5.* The algorithm makes $r_i \in R^K$ a vector corresponding to the $i$-th row of the normalized matrix $M$.

*Step 6.* The algorithm uses typical clustering algorithms such as the $k$-means algorithm to cluster the abovementioned matrix into $k$ classes. By repeatedly executing the NJW algorithm, the scale parameter $\delta$, which measures the similarity between sample points, can be obtained, but this increases the time used by the algorithm.

Another point of view in spectral clustering is the Markov random walk algorithm. It uses a probability model to obtain the spectrum method, and the random walk on the spectrum is considered to be a random jump from one node to another node. Spectral clustering can be understood as looking for such a graph partitioning; that is, the random walk stays in the same class for a long time and rarely stays in another class. From this point of view, random walk and graph partitioning have the same idea.

The random transition matrix $P = D^{-1}W$ can be obtained from the normalized similarity matrix $W$, so the sum of each row of the matrix is 1. Among them, $P_{ij}$ is the probability from point $v_i$ to point $v_j$. The Markov random walk algorithm is basically the same as the NJW algorithm. The Markov random walk algorithm has achieved good results and can automatically determine the clustering value.

In addition to the above SC algorithm, there are many other SC algorithms. Perona and Freeman proposed the PF algorithm. Scott et al. proposed the SLH algorithm, and WISS combined the SLH algorithm and the SM algorithm to propose a new algorithm. Ding et al. proposed a new division criterion Mcut. In addition, Meila et al. proposed a new spectral clustering algorithm called the MS algorithm under the framework of the Markov random walk.

*(1) The Choice of the Laplacian Matrix.* In the spectral clustering algorithm, it is very important to construct the similarity graph matrix. After the number of data samples is given, the most commonly used formula for constructing similarity graphs and selecting the Laplacian matrix is as

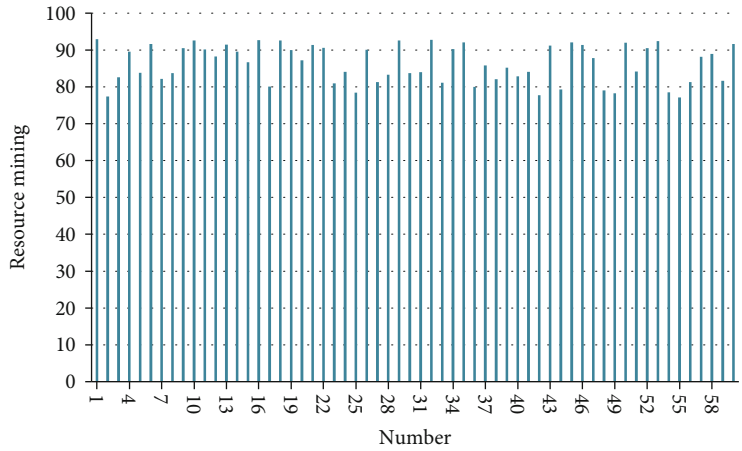



Figure 10: Statistical diagram of the evaluation of the mining effect of teaching resources.

follows:

$$w_{ij} = \exp\left(\frac{-\|x_i - x_j\|}{2\delta^2}\right). \quad (16)$$

In this paper, some commonly used Laplacian construct methods have been collected, but how to select the Laplacian matrix is uncertain.

*(2) Selection of Parameters*. The spectrum clustering is uncertain when constructing similar matrices, and the determination of 8 is often necessary to obtain the uncertainty of clustering results according to the experience of the researcher and multiple attempts. So the choice of parameters is an important research direction.

*(3) Determination of the Number of Clusters*. The choice of clustering directly affects the cluster results. Now, the current spectrum clustering research does not give a strategy for the number of choices of clusters, which is also a more important research direction in cluster research.

*(4) Differential Problems with Uneven Distribution of Density Distribution*. The existing spectrum clustering method is still unable to obtain a good clustering effect for the density distribution.

## 4. Computer-Aided Teaching System Based on Data Mining

After the SMART conceptual model was put forward, this paper designs the "I-SMART" model from the aspect of hardware and software configuration, which provides framework support for the construction of smart classrooms (Figure 2).

Under the guidance of the activity theory, based on the above analysis of teaching elements and types of teaching interaction, an interactive teaching model in a smart classroom environment is proposed (Figure 3).

Table 2: Statistical table of the evaluation of the teaching effect.

| No. | Teaching effect | No. | Teaching effect | No. | Teaching effect |
|---|---|---|---|---|---|
| 1 | 80.4 | 21 | 92.4 | 41 | 76.3 |
| 2 | 80.6 | 22 | 72.9 | 42 | 90.9 |
| 3 | 87.3 | 23 | 94.8 | 43 | 82.2 |
| 4 | 76.3 | 24 | 89.1 | 44 | 91.6 |
| 5 | 87.7 | 25 | 76.9 | 45 | 77.3 |
| 6 | 74.8 | 26 | 90.8 | 46 | 75.2 |
| 7 | 83.3 | 27 | 79.2 | 47 | 85.0 |
| 8 | 73.5 | 28 | 78.6 | 48 | 86.5 |
| 9 | 84.3 | 29 | 91.7 | 49 | 87.4 |
| 10 | 90.2 | 30 | 81.2 | 50 | 93.7 |
| 11 | 87.9 | 31 | 85.0 | 51 | 83.8 |
| 12 | 78.2 | 32 | 93.1 | 52 | 80.0 |
| 13 | 88.5 | 33 | 74.8 | 53 | 89.9 |
| 14 | 72.5 | 34 | 89.3 | 54 | 93.7 |
| 15 | 79.6 | 35 | 90.8 | 55 | 90.1 |
| 16 | 86.8 | 36 | 88.1 | 56 | 90.3 |
| 17 | 73.9 | 37 | 82.8 | 57 | 93.4 |
| 18 | 93.1 | 38 | 83.4 | 58 | 94.6 |
| 19 | 80.6 | 39 | 95.0 | 59 | 86.5 |
| 20 | 85.2 | 40 | 80.0 | 60 | 80.9 |

There are nine modules in the teaching platform: homepage, statistics, classroom activities, homework, examination, discussion, information, notification, and management. Users can directly access teaching data. The various types of data that can be read by the teacher after being automatically collected by the platform and analyzed by the model are summarized in Figure 4.

Through the correlation analysis of multiple factors in the S-P table, it can provide a clear direction for teaching decision-making. S-line tomography is used to analyze the types of students. Using the Yio-CSi (Student Positive Answer Rate-Student Attention Coefficient) model (Figure 5), we can find the students with the best and stable grades, the students with the worst and stable grades, and the students with unstable

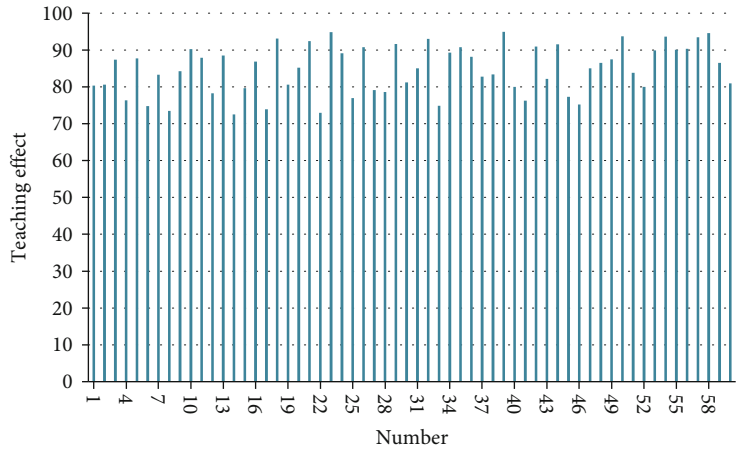



Figure 11: Statistical diagram of the evaluation of the teaching effect.

grades that cannot be found in the general statistical description. Using the Yjo-CPj (Question Positive Answer Rate-Question Attention Coefficient) model (Figure 6), it is possible to find knowledge points that students understand and are reliable, knowledge points that students seem to understand but are unreliable, knowledge points that are not understood by students, and knowledge points that are not understood but have accidental factors. Based on this, we make scientific decisions on the selection and application of teaching media.

This research is based on the smart teaching platform environment and attaches importance to the value of the data captured by the platform in the process of monitoring student behavior and status. Therefore, it is considered to incorporate the data that can be used for decision-making into the decision model to form a teaching decision model based on the data of the smart teaching platform, as shown in Figure 7.

The users of the education platform are not only students but also teachers, so we have to think deeply about teachers' needs for online education platforms. We can further analyze the teacher's demand for the ergonomic experimental teaching platform in two aspects. Among them, one is the demand between the platform and the teacher, and the other is the demand between the teacher and the student, as shown in Figure 8.

The strategic layer of the platform is shown in Figure 9. User goals refer to what kind of needs the users hope to meet through the product. The main users of the platform are students. For students, the most important goal of this platform is to help students improve their learning effects. On the one hand, it shows that students can learn and apply knowledge faster through this platform, and on the other hand, it shows that students can quickly learn how to use this platform, that is, the ease of learning of the platform.

## 5. Performance Test of the Computer-Aided Teaching System Based on Data Mining

The above constructs a computer-aided teaching system based on data mining. After constructing the system, the performance of the system is verified, and the system operation performance and system teaching effect are mainly studied. The research on its operating performance is mainly the effect of teaching resource mining. This paper obtains effective resources from the massive network teaching resources through the simulation system, tests multiple sets of statistical data, and evaluates the collected data through expert evaluation methods. The results are shown in Table 1 and Figure 10.

From the above research, we can see that the system constructed in this paper can effectively tap the required teaching resources. After that, this paper evaluates the teaching effect of this system through a small-scale teaching test, and the statistical results are shown in Table 2 and Figure 11.

It can be seen from the above experiments that the system constructed in this paper has certain teaching and practical effects, so it can expand the scope of teaching in follow-up research and conduct practical research from the perspective of multiple subjects and multiple audiences.

## 6. Conclusion

The smart education system based on data mining services is a typical current smart education. It realizes the intelligence, visualization, and high efficiency of classroom teaching, and its application conforms to the concept of building a strong education nation and is more in line with the specific requirements of educational modernization. With the continuous progress of cloud computing, artificial intelligence, big data, and other technical means, the service quality of the smart education system has also been upgraded, which provides favorable conditions for the safe and effective operation of the system. This paper combines data mining technology to build a computer-aided teaching system and builds the function of the teaching system based on the actual situation. The constructed system can carry out multi-subject teaching. Moreover, this paper uses a data mining system to mine teaching resources and combines digital technology to process teaching resources. After constructing the system, this paper designs experiments to verify the performance of the system. From the research results, it can be

seen that the system constructed in this paper has certain teaching and practical effects.

## Data Availability

The labeled datasets used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare no competing interests.

## Acknowledgments

This study is sponsored by the following: (1) Research and Practice of Mixed Teaching Mode of Criminal Technology Courses Based on OBE Concept—take Speech Recognition and Appraisal Course as an example (No. 5-3 1017501)—and (2) application research of IPv6-based video transmission and control technology in teaching platform—Scientific research project (No. NGII20180124).

## References

[1] R. Nunes Linhares, C. Mário Guimarães Alcântara, E. Ávila Gonçalves, F. Ramos, and M. José Loureiro, "Teaching evaluation by teachers from Brazil and Portugal: a comparative analysis," *American Journal of Educational Research*, vol. 5, no. 5, pp. 546–551, 2017.

[2] N. Huang, "Analysis and design of university teaching evaluation system based on JSP platform," *International Journal of Education & Management Engineering*, vol. 7, no. 3, pp. 43–50, 2017.

[3] J. A. Moreno-Murcia, Y. Silveira Torregrosa, and N. Belando Pedreño, "Questionnaire evaluating teaching competencies in the university environment. Evaluation of teaching competencies in the university," *Naer Journal of New Approaches in Educational Research*, vol. 4, no. 1, pp. 54–61, 2015.

[4] S. Liu and P. Chen, "Research on fuzzy comprehensive evaluation in practice teaching assessment of computer majors," *International Journal of Modern Education & Computer Science*, vol. 7, no. 11, pp. 12–19, 2015.

[5] L. Zhou, H. Li, and K. Sun, "Teaching performance evaluation by means of a hierarchical multifactorial evaluation model based on type-2 fuzzy sets," *Applied Intelligence*, vol. 46, no. 1, pp. 34–44, 2017.

[6] J. Porozovs, L. Liepniece, and D. Voita, "Evaluation of the teaching methods used in secondary school biology lessons," *Journal of Pedagogy and Psychology "Signum Temporis"*, vol. 7, no. 1, pp. 60–66, 2015.

[7] M. A. Oliveros, A. García, and B. Valdez, "Evaluation of a teaching sequence regarding science, technology and society values in higher education," *Creative Education*, vol. 6, no. 16, pp. 1768–1775, 2015.

[8] M. S. Cerón and F. M. del Sagrario Corte Cruz, "The evaluation of teaching: some consequences for Latin America," *Revista Mexicana De Investigacion Educativa*, vol. 20, no. 67, pp. 1233–1253, 2015.

[9] L. Liu, J. Feng, Q. Pei et al., "Blockchain-enabled secure data sharing scheme in mobile-edge computing: an asynchronous advantage actor–critic learning approach," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2342–2353, 2021.

[10] K. Angell, E. Tewell, and Long Island University, Brooklyn, "Teaching and un-teaching source evaluation: questioning authority in information literacy instruction," *Communications in Information Literacy*, vol. 11, no. 1, pp. 95–121, 2017.

[11] M. Brkovic and P. Chiles, "'Spector - the sustainability inspector': participatory teaching, learning and evaluation game for architects, architecture students and pupils," *Facta Universitatis*, vol. 14, no. 1, pp. 1–20, 2016.

[12] T. H. Reisenwitz, "Student evaluation of teaching: an investigation of nonresponse bias in an online context," *Journal of Marketing Education*, vol. 38, no. 4, pp. 139–144, 2015.

[13] Y. Jiang and Y. Wang, "Evaluation of teaching quality of public physical education in colleges based on the fuzzy evaluation theory," *Journal of Computational and Theoretical Nanoence*, vol. 13, no. 12, pp. 9848–9851, 2016.

[14] W. Shang, J. Chen, H. Bi, Y. C. Sui, Y. Chen, and H. Yu, "Impacts of COVID-19 pandemic on user behaviors and environmental benefits of bike sharing: a big-data analysis," *Applied Energy*, vol. 285, article 116429, 2021.

[15] F. Garofalo, P. Mota-Moya, A. Munday, and S. Romy, "Total extraperitoneal hernia repair: residency teaching program and outcome evaluation," *World Journal of Surgery*, vol. 41, no. 1, pp. 100–105, 2017.

[16] G. Gong and S. Liu, "Consideration of evaluation of teaching at colleges," *Open Journal of Social Sciences*, vol. 4, no. 7, pp. 82–84, 2016.

[17] H. Zhao, "College physics teaching model design and evaluation research of students' seriousness," *Open Cybernetics & Systemics Journal*, vol. 9, no. 1, pp. 2017–2020, 2015.

[18] N. D. Tran, "Reconceptualisation of approaches to teaching evaluation in higher education, issues in educational research," *Issues in Educational Research*, vol. 25, no. 1, pp. 50–61, 2015.

[19] P. Wu, S. P. Low, J. Y. Liu, J. Pienaar, and B. Xia, "Critical success factors in distance learning construction programs at Central Queensland University: students' perspective," *Journal of Professional Issues in Engineering Education and Practice*, vol. 141, no. 1, article 05014003, 2015.

[20] E. A. Willis, A. N. Szabo-Reed, L. T. Ptomey et al., "Distance learning strategies for weight management utilizing social media: a comparison of phone conference call versus social media platform. Rationale and design for a randomized study," *Contemporary Clinical Trials*, vol. 47, no. 6, pp. 282–288, 2016.

[21] H. J. Ye, D. C. Zhan, and Y. Jiang, "Fast generalization rates for distance metric learning," *Machine Learning*, vol. 108, no. 2, pp. 267–295, 2019.

[22] Y. Luo, Y. Wen, T. Liu, and D. Tao, "Transferring knowledge fragments for learning distance metric from a heterogeneous domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 1013–1026, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3938409, 13 pages
https://doi.org/10.1155/2021/3938409

*Research Article*

# Better Effectiveness of Multi-Integrated Neural Networks: Take Stock Big Data as an Example

**HangLin Lu and XiuYun Peng**

*Zhengzhou University, Zhengzhou, China*

Correspondence should be addressed to HangLin Lu; luhanglinde@163.com

HangLin Lu and XiuYun Peng contributed equally to this work.

Received 9 August 2021; Revised 3 September 2021; Accepted 6 September 2021; Published 6 October 2021

Academic Editor: Thippa Reddy G



With the development of big data, in the financial market, the stock price prediction has many research directions from the perspective of big data. The classical time series prediction model cannot adapt to the high-latitude information of stock data in the era of big data. The development of deep learning provides a new idea for high-latitude stock data prediction. Four neural network models and three integrated learning models form different strategy sets, and the opening price of the next timestamp is predicted by backtracking information over the past 15 days with the characteristics of 12 indexes of the stock. The experimental results show that the prediction effect of the integration model based on the average weight policy and stacking policy is better than that of the single neural network, and the integration model based on stacking policy is expected to have the highest prediction accuracy and the minimum expected error. The accuracy was 80.2%, and the mean square error was 0.024. Compared with the single model, the accuracy is increased by 2%~7%, and the error is reduced by 0.01~0.03. The innovation of this article lies in the traditional machine learning thinking is applied to deep learning, as an individual with a variety of neural network to study, through the integration of learning strategies, fusion for the integration model, the experimental results show that the effect of the integrated model is better than that of a single model, to improve the robustness and accuracy of the model; the performance of the integrated model is more stable. For the utilization of big data resources, the integrated model of neural network has better prediction effect.

## 1. Introduction

In recent years, with the maturity and development of financial market, the theory of stock forecast is increasingly diversified. Early stock index predictions were based on market theories, such as Osborne's walking theory in 1959, which held that stocks could not predict the Brownian motion in physics [1]. Fama, winner of the Nobel Prize in Economics in 1970, puts forward the efficient market hypothesis, which believes that the trend of stock prices can be predicted under the condition of sufficient information [2].

In the beginning, the prediction of stock-related indexes also adopted statistical traditional regression model analysis, such as autoregressive conditional heteroscedasticity model (ARCH), autoregressive conditional heteroscedasticity model (ARIMA), and GARCH model [3].

With the development of machine learning and the improvement of the ability of machine learning in time series prediction, the relationship between machine learning and the prediction of financial-related indicators is getting closer and closer. Whether it is deep learning or basic machine learning model, it can constantly improve its performance during the training process. With the development of artificial intelligence and the improvement of computer performance in recent years, machine learning has been widely applied in the financial industry. In 1999, Allen and Karjalainen applied the genetic algorithm to the historical data of American stocks and deduced the trading rules from it [4].

Endemic learning has also been used in stock prediction. In 2003, Kim compared support vector machine and neural network to try the effect of support vector machine in stock index prediction process. In 2016, Khaidem used an integrated learning model of random forests to predict stock returns and reduce investment risk. Huang and Chen used support vector machine (SVM) to test the prediction of the model on the stock price of the Bank of China [5].

However, for stock index prediction, more scholars rely on a single model. Cui and Li used the GARCH model and BP network to carry out stock price prediction experiments, and BP network was better than the traditional statistical model [3]. However, the generalization ability of these networks is not strong, and it is easy to overfit the data of the training set, which makes the prediction effect of the test set worse. For time series, RNN, LSTM, and other cyclic neural networks have obvious advantages. Wang et al. compared the effect of RNN and LSTM in stock price prediction [6]. With the improvement of computer computing speed, some neural network algorithms have been more verified and applied. Yenidoğan et al. [7] used the LSTM model to predict the stock fluctuations of the CSI 300 Index and found that its prediction effect was better than the classical time series analysis method.

In 2017, Nelson et al. used LSTM neural network to build a stock index prediction model and predicted the rise and fall with the help of historical data. Meanwhile, by comparing the model with other machine learning methods, they found that deep learning could better extract data information and make more accurate predictions with stronger robustness. [8] immediately some integrated learning models such as forest. With the development of deep learning be replaced gradually, but the application of the integration strategy of integrated study on deep learning network can also make the performance of the model for promotion, Xie et al. than using LSTM neural network model, such as building-integrated learning model, in the stock of quantitative trading experiments achieved better effect [9].

The time series data used in this paper consists of 12 indicators, and each time, the data of the previous 15 trading days is backtracked to predict the opening price of the next time step. MLP, RNN, LSTM, and GRU were set as four basic models, and three combination strategies were adopted to form three integrated models. The first 70% data were taken as the training set and the last 30% as the prediction set. The training set data were formed into multiple batches of data by the self-help sampling method. In the training process, Adam was used as the optimizer of the deep neural network to predict the opening price of the next trading day and record the accuracy and error of the forecast rise and fall. We found that for the utilization of big data resources, the integrated model of neural network has better prediction effect.

## 2. Problem Raising and Theoretical Analysis

In the training setting of the deep learning model, in order to prevent overfitting on the test set data, the training samples of neural network are often randomly selected as training batches in the training set. Because of this randomness, the effect of the model in the process of training convergence is constantly fluctuating, and this fluctuation still exists in the training after model training convergence. This leads to a problem: the actual performance of the final model is unstable, and its effect has random error.

This error is reflected in the data of the verification set, which is actually the error caused by the different sensitivity of the model to different samples. In the test set, the fluctuation can be understood as the lack of feature learning of the model to the training data.

Then, different deep learning cell structures are determined, which determines the attributes of the corresponding network, which makes a model may have a better effect on a certain type of samples in the data set, but the learning effect on other types of samples is not ideal. That is, different network models may have different adaptive learning abilities to different samples. How to make the network adapt to the training of more types of samples as much as possible? We think of integrating neural networks with different structures to obtain a more robust model.

On the other hand, from the perspective of statistics, the fusion of prediction results of multiple models is similar to multiple sampling to take the mean, which can reduce the random error of prediction results and improve the stability of model prediction effect.

In conclusion, in order to reduce the randomness of this model training effect, capture the characteristics of the test set data as much as possible to improve the model performance. Here, we try to propose an integrated neural network model. Through the training and integration of deep learning networks with different structures, a more stable model can be obtained.

## 3. Individual Learner

Individual learner is one of the basic structures of the ensemble learning model, which can also be called the basic model. Individual learners also have their own learning prediction ability. In this paper, according to the characteristics of time series, we select 4 kinds of neural network models as individual learners, which are multilayer perceptron (MLP), recurring neural network (RNN), long-short-term neural network (LSTM), and gated recurring neural network (GRU).

*3.1. MLP.* Multilayer perceptron is one of the most classical feedforward artificial neural network models.

The model has strong nonlinear fitting and generalization ability, and the weight and bias of each neuron are adjusted continuously with the help of the error back propagation algorithm, so as to reduce the error in the training set. However, the generalization ability of this model is insufficient, and it is easy to appear the phenomenon of overfitting for the data of the training set. Chen et al. proposed the problem of insufficient generalization ability of MLP [10].

*3.2. RNN.* Cyclic neural network, also known as recursive neural network, was proposed by Schuster and Paliwal [11]

in 1997. It is a kind of neural network built for sequential data and can fully reflect the correlation of data at different time nodes [12]. Cyclic neural networks have some advantages in learning the nonlinear characteristics of sequences because of their memory in time. Recurring neural network has many applications in natural language processing [13], time series prediction, and other fields.

$$O_t = g(V \cdot S_t),$$
$$S_t = f(U \cdot X_t + W \cdot S_{t-1}), \quad (1)$$

where $x_t$ represents the input value, $s$ represents the value of the hidden layer, $U$ represents the weight matrix from the input layer to the hidden layer, $V$ represents the weight matrix from the hidden layer to the output layer, and $o$ represents the output value. As can be seen from the figure, the weight matrix of the hidden layer of the cyclic neural network depends not only on the current input $x$ but also on the value $s$ of the hidden layer on the last timestamp.

*3.3. LSTM.* Long and short-term memory artificial neural network (LSTM) is a chain form designed to solve the long-term data dependence existing in the recursive neural network (RNN) [14], proposed by Hochreiter and Schmidhuber [13]. As a special cyclic neural network, LSTM is formed by repeating module chains. LSTM also has this chain-like structure, of which the most important basic structure is the cell. Each cell has a specific gate structure to realize selective information transfer. Through the information transfer of LSTM gate structure (forgetting gate, input gate, update gate, and output gate), each cell state can be updated according to the last output and the current input. The specific structure is shown in Figure 1.

*3.4. GRU.* Gated circulation unit is a variant of LSTM proposed by Chung et al., whose special structure can solve the phenomenon of gradient dispersion in the training process of standard RNN [15]. The GRU controls the input and memory of information through two gate structures, a reset gate and an update gate. The reset gate determines the combination of the new input information with the information previously memorized by the GRU cell, while the update gate is used to save the memorized information from the previous timestamp to the information retained by the current timestamp. This gate control structure can better preserve the information in the long-term time series and will not forget or erase the effective information because of the longer time series.

The basic structure of GRU is shown in Figure 2.

## 4. Ensemble Learner

*4.1. Basic Theory.* Ensemble learning is a learning mode that constructs multiple individual learners and integrates them to achieve related classification or fitting tasks. Its basic structure is a single individual learner. Ensemble learning combines basic models through a combination strategy to achieve an integrated model that exceeds the effect of individual learners and improve the robustness of the model.

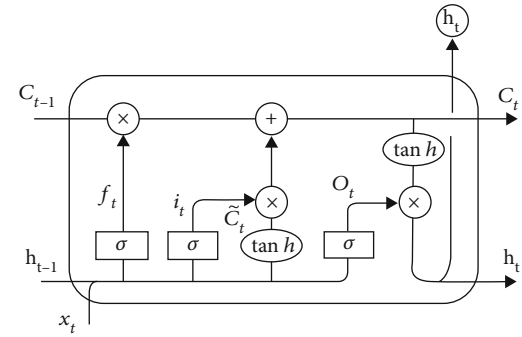



Figure 1: LSTM structure.

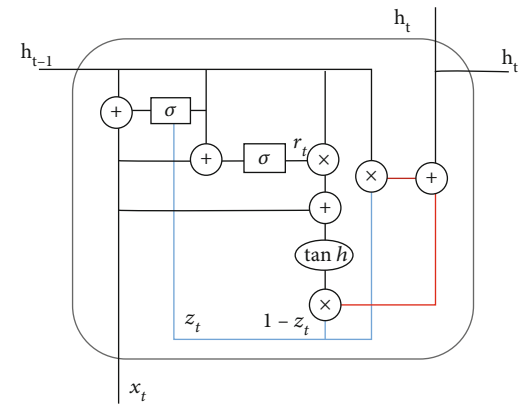



Figure 2: GRU structure.

Common ensemble learning models are mostly built based on weak learners, such as random forest and boosting. In order to obtain a good integrated model, individual learners should be good but different, that is to say, individual learners have better performance, but at the same time, different learners have differences in principle or architecture. In this paper, we will use the model with strong learning ability as the basic learner. Multiple neural networks with different structures are used as individual learners, and MLP, RNN, LSTM, and GRU are selected as individual learners according to the characteristics of stock index prediction time series. Each individual learner itself has strong learning ability. Different types of individual learners are included in the integration, so the integration model constituted is "heterogeneous," and each basic model has a parallel relationship. The selection and structure setting of such an individual model can improve the robustness of the integration model in principle.

*4.2. Integration Strategy*

*4.2.1. Average Weight Method.* The combination strategy of the average weight method is a commonly used learning strategy for numerical regressions in ensemble learning. The method is to average the output of several individual learners to get the final predicted output.

The final forecast results are as follows:

$$H(x) = \frac{1}{T}\sum_{1}^{T} \text{model}_i(x). \quad (2)$$

When combining the average weight strategy in this paper, four models were selected, that is, the combining layer multiplied the output of each model by 0.25 and then summed up the predicted value. After the predicted value was compared with the target value to solve the mean square error (MSE), the error was transmitted in reverse and fed back to the four individual learners. In the error feedforward process, each weight of the binding layer is always maintained at 0.25, which does not change with the error feedforward. Model settings are shown in Figure 3.

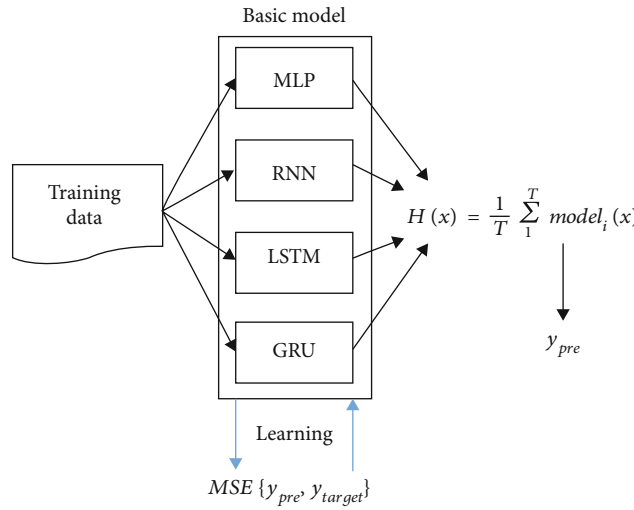



Figure 3: Average weight method.

*4.2.2. Stacking Method.* The combination strategy of stacking is to regression integration of the output of each model through one or more metalearners. The whole training set is used to train the basic model, and the metamotor trains the predicted value of the basic model as the characteristic.

In the integration model of this paper, the basic model for different learning algorithms, stacking is heterogeneous integration. The model settings are shown in Figure 4.

The algorithm pseudocode is as follows:

*4.2.3. Global Learning Method.* The combination strategy of the global learning method refers to the connection between the basic model and the secondary learners, as shown in the figure. Then, the whole training set data is used to train the model. During the training process, the error feedback not only changes the internal network parameters of each basic model but also changes the internal weight of the metalearner. Model settings are shown in Figure 5.

## 5. Experimental

In this paper, four kinds of neural networks are integrated, and three kinds of integration learning strategies are adopted for model fusion. The whole process is shown in the figure. In the model, the first stage is data preprocessing, including removing missing values, data segmentation, and data normalization and then dividing the training set and the test set. In the second stage, the training model is constantly evaluated for the performance of the model in the training set and test set to determine the number of training iterations. Finally, the network is used to predict the results of the stock index. Step settings are shown in Figure 6.

The specific steps are as follows:

(1) Missing values were eliminated, minimum-maximum normalization method was used to normalize each stock index in its dimension to values within the range of [0,1], and then, the historical data were divided into two groups: training data set and test data set

(2) Batch the data of the training set. At the beginning of each training round, the order of the data of the training set is disrupted and 64 data are extracted as a batch

(3) Initialize the model network, including the weight and bias of parameters on each layer, determine the stochastic gradient descent algorithm as Adam, and set the maximum number of iterations as 100. The fluctuation data of error and accuracy were recorded after the model was stabilized

(4) Conduct training on different basic models and integrated strategy models

(5) Set the number of iteration training: 100. Evaluate the suitability of the model. If the model is suitable, save the model and record the mean square error and accuracy at the same time; if not, continue the error feedforward training

(6) Import the test set data into the model to confirm the optimal one for index prediction. Ensure that complete predictions are made after optimal test set predictions are made

(7) Evaluate the prediction accuracy of each model through five performance indicators

The experiment was conducted on a PC (CPU: AMD Ryzen 5 3500U,8 Gbps RAM). The development environment is Python 3.8, and Spyder is running on a Windows 10 operating system. The model is implemented using PyTorch.

*5.1. Data Settings.* In this study, we select Shenzhen stock: Ping An Bank (stock code: 000001), Ping An Bank Co., Ltd., is a national joint-stock commercial bank headquartered in Shenzhen (Shenzhen Stock Exchange: 000001). Its predecessor, Shenzhen Development Bank, is a national joint-stock bank publicly listed in the mainland of China. The experimental data set consists of daily historical data for the past 10 years as of January 4, 2010, BBB 0 and

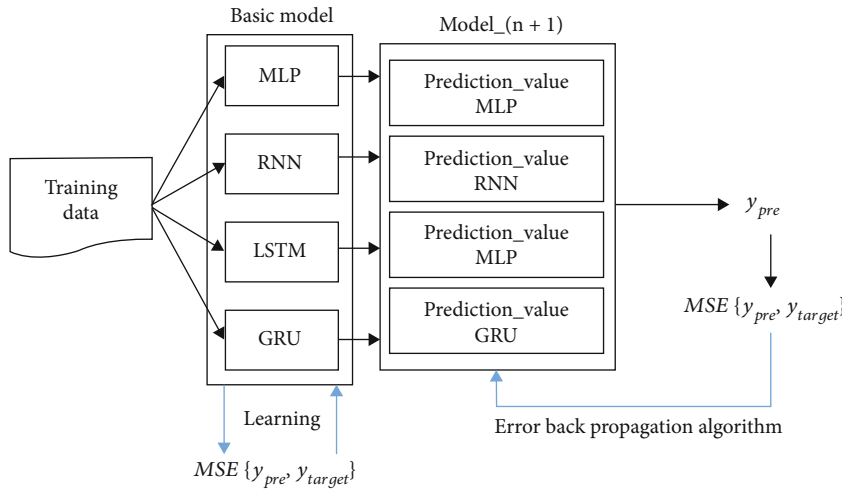



Figure 4: Stacking method.

1: Input: training data $D = \{x_i, y_i\}_{i=1}^{m}$
2: Output: ensemble classifier $H$
3: *Step1: learn base-level classifier*
4: **for** $t = 1$ to $T$ **do**
5: learn $h_t$ based on $D$
6: **end for**
7: *Step2: construct new data set predictions*
8: **for** $i = 1$ to $m$ **do**
9: $D_h = \{x_i, y_i\}$, where $x_i' = \{h_1(x_i), \cdots, h_T(x_i)\}$
10: **end for**
11: *Step3: learn a meta-classifier*
12: learn $H$ based on $D_h$
13: return $H$

Algorithm 1: Stacking.

December 31, 2019. All data are from the Oriental Fortune Market Center, the data as shown in the figure.

As shown in Table 1, each timestamp contains 12 characteristics, which are closing price, maximum price, minimum price, opening price, previous closing price, up/down amount, up/down amount, turnover rate, volume, transaction amount, total market value, and circulating market value.

*5.2. Data Preprocessing.* Data preprocessing is a crucial step in data analysis and model training. High-quality data leads to better models and predictions. First, a small number of missing values in the data set were eliminated, and then, each index was normalized to between. The normalization method is as follows:

$$X = \frac{X_t - X_{\min}}{X_{\max} - X_{\min}}, \tag{3}$$

where $X_t$ is the value $f$ a feature on the timestamp $t$, $X_{\max}$ is the maximum value on this feature dimension, and $X_{\min}$ is the minimum value on this feature dimension.

After data preprocessing, the original data set is divided into two independent sets at a fixed ratio. Among them, the first 70% of daily historical data is taken as the training set, and the remaining 30% of data is taken as the test set. As shown in the figure, the blue part is the training set data, and the orange part is the test set data, as shown in Figure 7.

The normalized data distribution of the training set and the test set is shown in the figure. As shown in Figure 8, it can be seen that the distribution domain of the training set is larger than that of the test set, that is, the training set and the test set of the data are properly divided.

*5.3. Performance Index Evaluation.* There are many ways to measure the effect of prediction. In order to properly evaluate the prediction ability of various models, the following five indicators are used in the experiment to measure the accuracy of the model: mean square error (MSE), mean absolute error (MAE), MAPE, and $R$ squared.

$$\begin{aligned}
\mathrm{MSE} &= \frac{1}{N}\sum_{t=1}^{N}\left(y_{\mathrm{test}}(t) - y\wedge_{\mathrm{pre}}(t)\right)^2,\\
\mathrm{MAE} &= \frac{1}{N}\sum_{i=1}^{N}|y_{\mathrm{test}}(t) - \widehat{y}_{\mathrm{test}}(t)|,\\
\mathrm{MAPE} &= \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_{\mathrm{test}}(t) - \widehat{y}_{\mathrm{test}}(t)}{y_{\mathrm{test}}(t)}\right|,\\
R^2 &= 1 - \frac{\mathrm{MSE}(y_{\mathrm{test}}(t) - \widehat{y}_{\mathrm{test}}(t))}{\mathrm{Var}(y_{\mathrm{test}}(t))},
\end{aligned} \tag{4}$$

where $N$ represents the number of samples of this group of data, while $y_{\mathrm{test}}$ and $\widehat{y}_{\mathrm{test}}$ represent the real and predicted values of the test set data, respectively.

Considering that the stock index pays more attention to the trend of rise and fall, the accuracy rate is introduced at the same time. The definition is if the index value of stock $T + 1$ is greater than that of stock $T$, and the predicted value of stock is also greater than that of stock $T$, it is considered

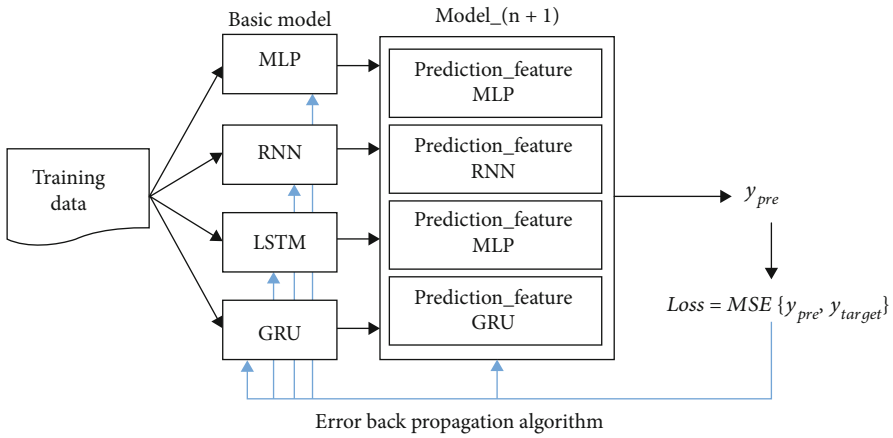



Figure 5: Global learning method.

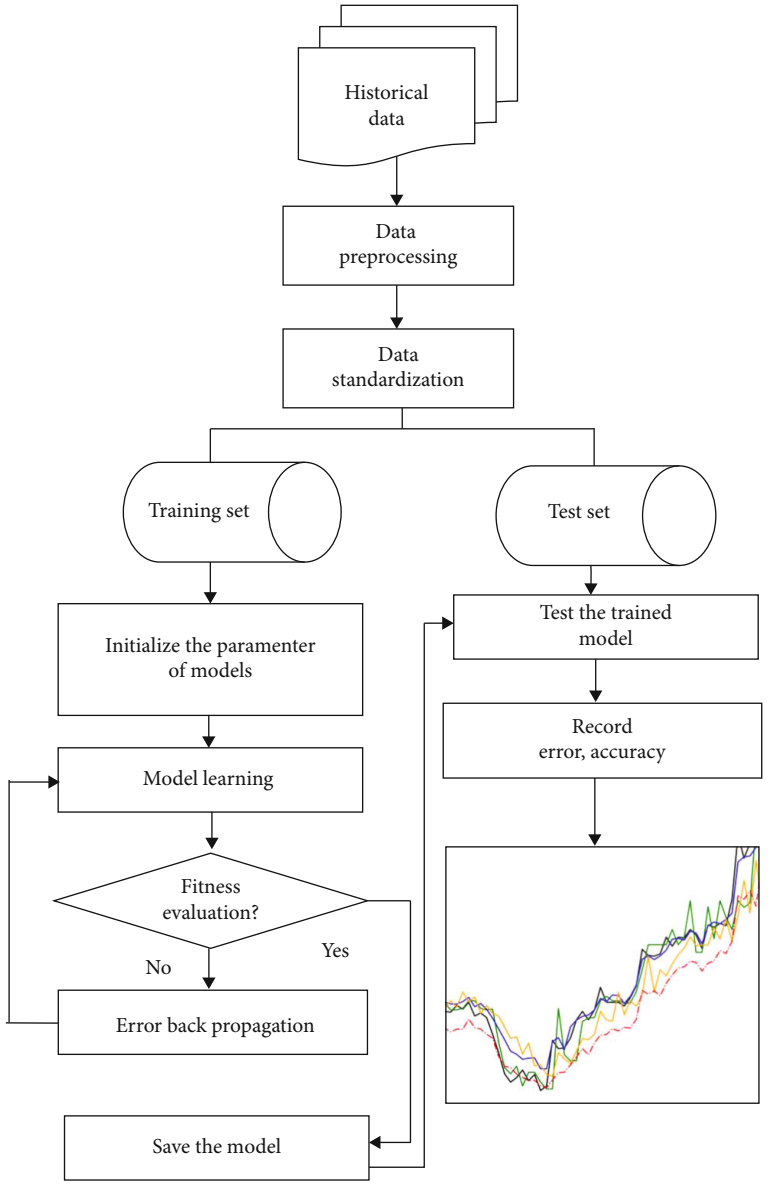



Figure 6: The experimental process.

Table 1: Stock data format.

| Date | Close | Highest | ··· | Change (%) | Vol (share) | Vol (RMB) | EUR (RMB) | FAMC (RMB) |
|---|---|---|---|---|---|---|---|---|
| 2010/1/4 | 23.71 | 24.58 | ··· | -2.7082 | 24192276 | 580249472 | 73629834497 | 69330922520 |
| 2010/1/5 | 23.3 | 23.9 | ··· | -1.7292 | 55649982 | 1293476939 | 72356606655 | 68132032674 |
| 2010/1/6 | 22.9 | 23.25 | ··· | -1.7167 | 41214313 | 944453697 | 71114433150 | 66962384045 |
| 2010/1/7 | 22.65 | 23.05 | ··· | -1.0917 | 35533685 | 804166316 | 70338074709 | 66231353651 |
| 2010/1/8 | 22.6 | 22.75 | ··· | -0.2208 | 28854306 | 650667405 | 70182803021 | 66085147572 |

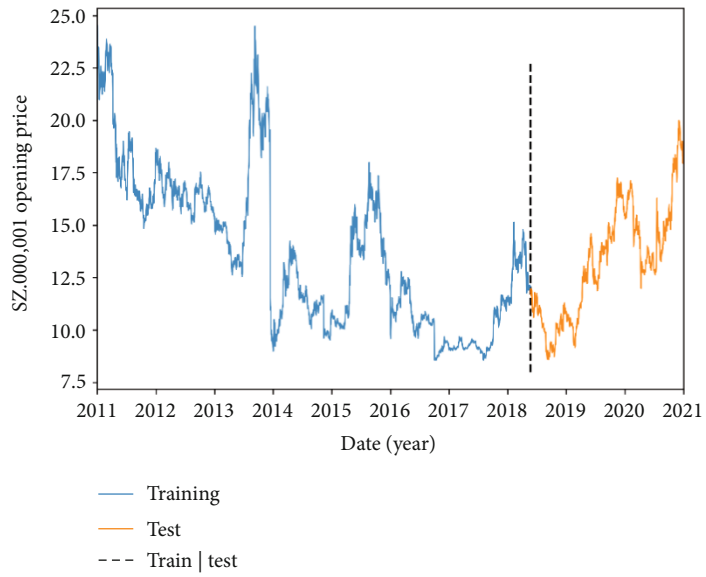



Figure 7: Training set test set partition.

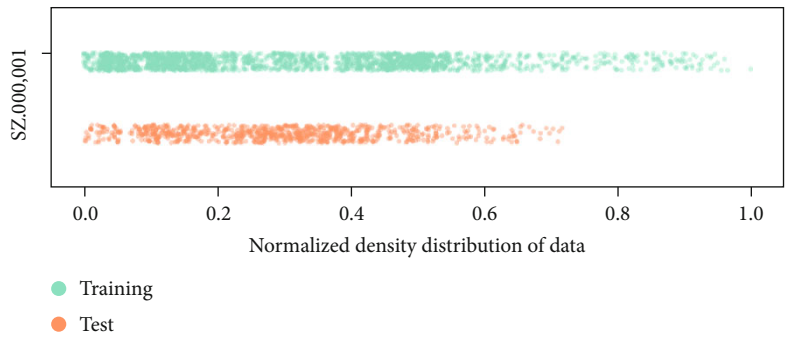



Figure 8: Data distribution of training set and test set.

that the forecast is correct. The reverse is the opposite. Namely, if the successful prediction of stock index trend is up or down, it is considered that the prediction is successful.

The formula is expressed as

$$\text{Accuracy} = \frac{n_{\text{right trend}}}{N_{\text{total}}}. \tag{5}$$

*5.4. Parameter Setting of the Model.* The setting of model parameters often affects the amount of information that can be recorded by the model, thus affecting the effect of

Table 2: Parameter settings.

| Model | Parameter settings |
|---|---|
| MLP | Unit number = 32<br>Layer number = 2 |
| RNN | Batch_size = 64<br>Epoch = 100 |
| LSTM | Parameter = Adam<br>Learing rate = 5e − 4 |
| GRU | |

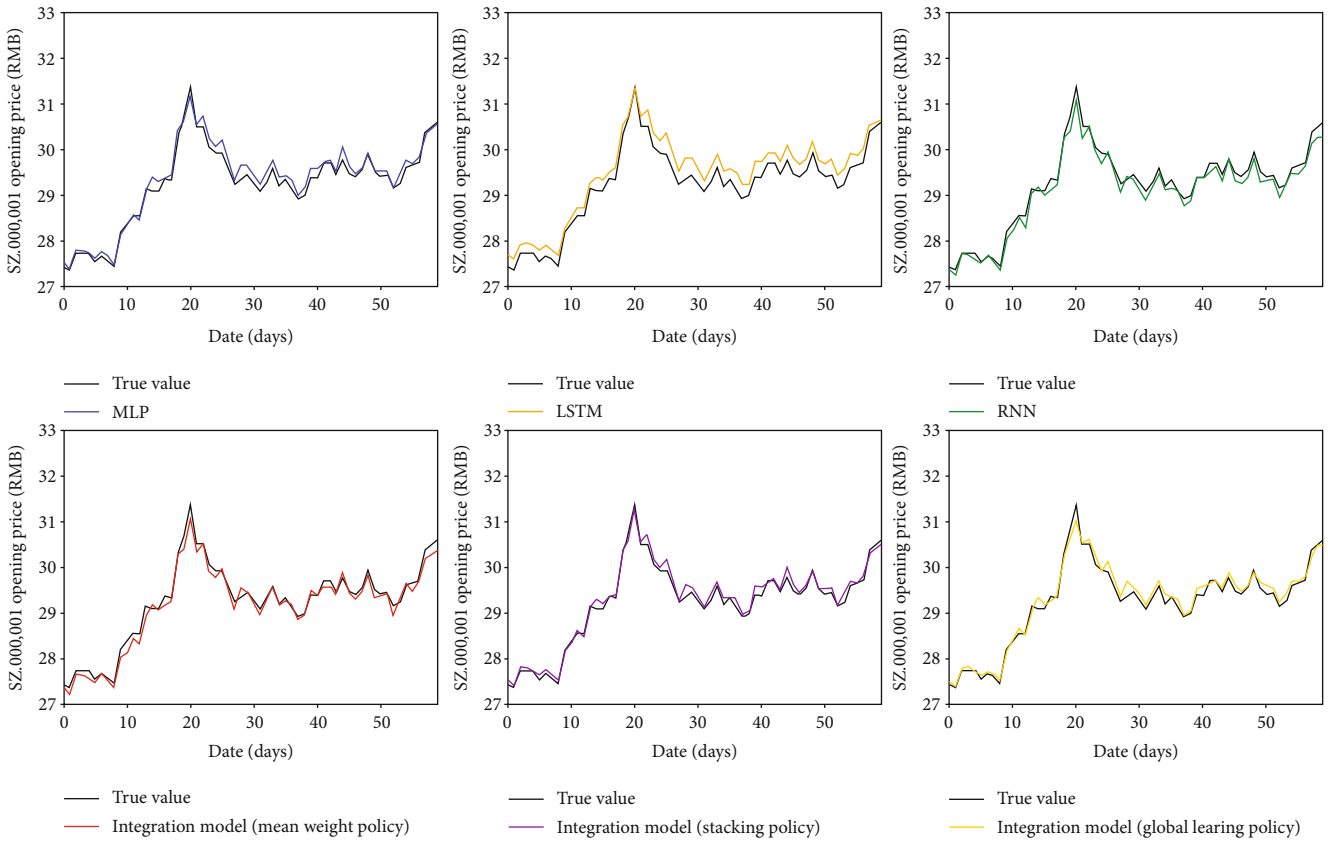



Figure 9: The prediction of several models.

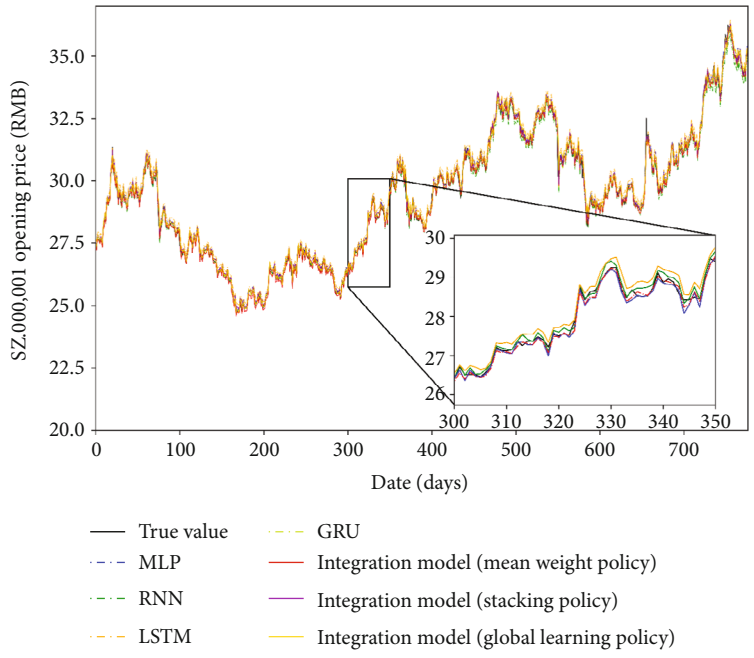



Figure 10: Complete data set prediction.

the model. In order to fairly compare the effects of different models, the parameters of the basic model will be set the same here. The model parameters are set as Table 2.

*5.5. Forecast of Indicators.* The whole data contains 2560 records, each timestamp records the closing price, the high, the low, the opening price, and other 12 characteristic

Table 3: Evaluation of prediction results.

| Model | Accuracy | MAE | MAPE | MSE | R_2 |
|---|---|---|---|---|---|
| MLP | 0.758301 | 0.151200 | 0.011381 | 0.041578 | 0.993498 |
| RNN | 0.785069 | 0.133305 | 0.010259 | 0.032996 | 0.994840 |
| LSTM | 0.778255 | 0.143340 | 0.010028 | 0.042996 | 0.996720 |
| GRU | 0.776190 | 0.129919 | 0.009749 | 0.031959 | 0.995003 |
| MEAN | 0.788030 | 0.129757 | 0.009710 | 0.032805 | 0.994870 |
| STACK | 0.802240 | 0.118611 | 0.003991 | 0.028048 | 0.995695 |
| LEARN | 0.755701 | 0.154355 | 0.005160 | 0.043889 | 0.993264 |

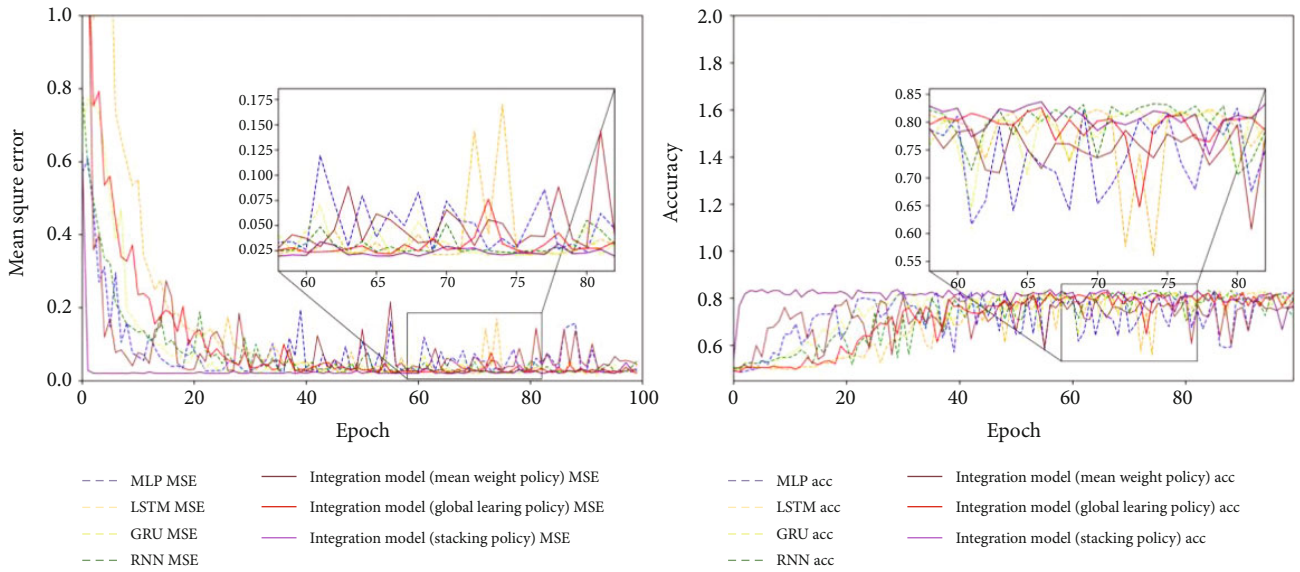



Figure 11: Models train batches with fluctuating stability and error conditions.

indicators. The first 1792 records were taken as the training set, 64 records were taken as a batch, and 28 batches were formed. The last 777 records were used as test set data. In order to comprehensively evaluate the prediction performance of the model, the prediction results are evaluated by the four performance indicators mentioned above.

Backtracking days refer to the data of the previous days used as the feature in the forecast, and the opening price of the next day is predicted by the feature data of the previous $N$ days.

In order to evaluate the performance of various models, based on previous studies, we set the number of days for backtracking data to be 15.

## 6. Result Evaluation

*6.1. Prediction Effect.* Figure 9 shows the prediction of the opening price index of the test set during 60 days by the models under the three basic models and the three integration strategies. It is obvious from the figure that the curves of the integration strategy using the average weight method and the stacking strategy are the closest to the real value, as shown in Figure 9.

In order to reflect the prediction effect of the model more comprehensively, the prediction effect diagram of the model on the complete test set is drawn, as shown in Figure 10. It can be seen clearly from the enlarged subgraph in the figure that the integrated model of the average weight method and the stacking strategy are very close to the predicted value.

In order to evaluate the prediction accuracy of each model, the prediction accuracy of the model is evaluated through several evaluation indexes described above. The model of the iteration training batch (the last 50 batches) after the training results are stable is evaluated, and the average value of the evaluation indicators of each iteration is taken as the expected value of each indicator. The final results are shown in Table 3.

It can be seen from the table that among the single training models MLP, RNN, LSTM, and GRU, the evaluation results of RNN, LSTM, and GRU cycle networks are better than that of MLP network. The accuracy of RNN in predicting the rise and fall is higher, reaching 78%, and the R_2 of LSTM is higher, which indicates that the overall prediction fitting of LSTM is better. The MAE, MAPE, and MSE of GRU are smaller, which means the overall error of GRU prediction is smaller. From the analysis of the prediction and

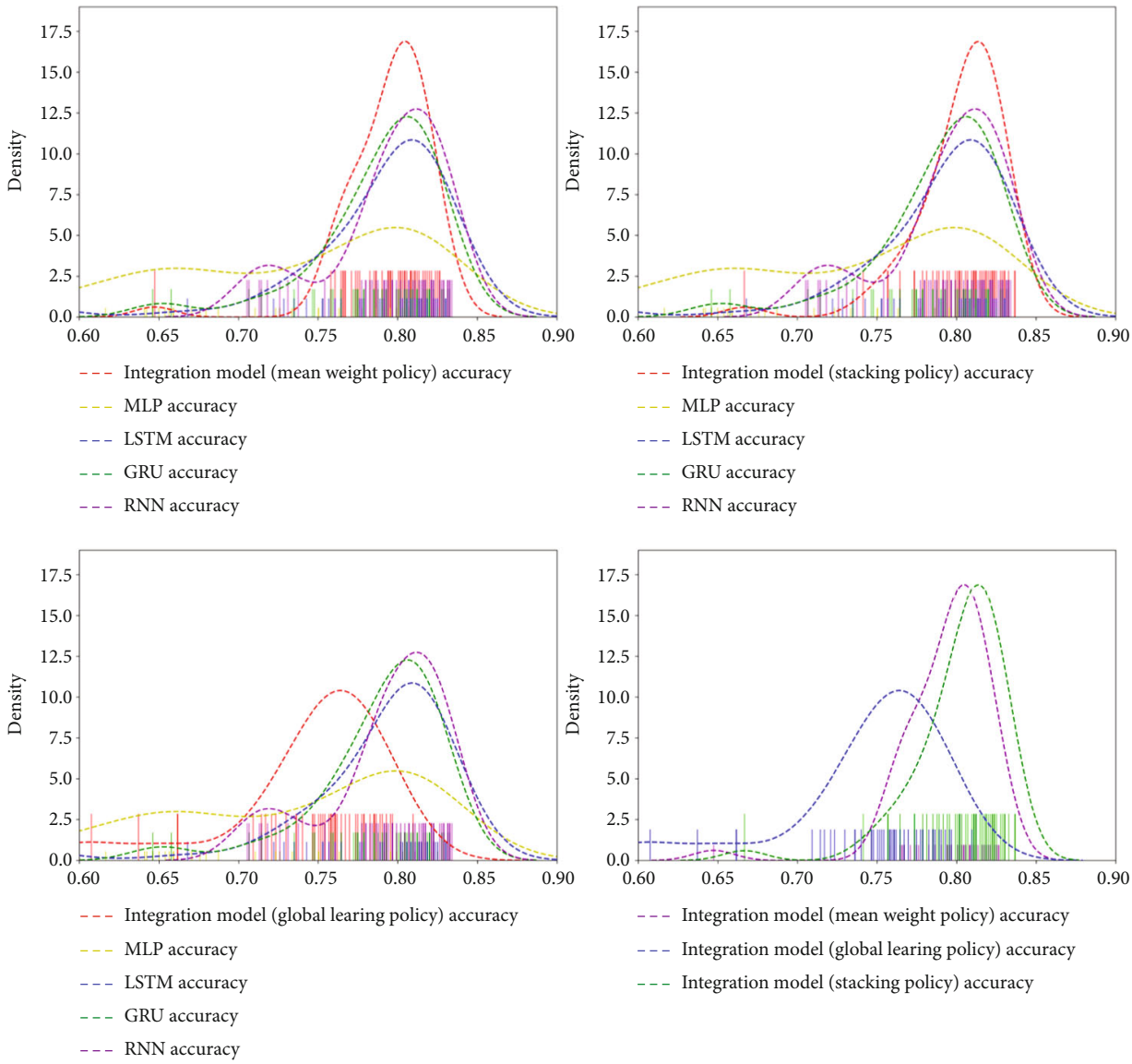



Figure 12: Prediction accuracy distribution after stable model training.

evaluation of a single model, it can be seen that the sensitivity of different single models to data has different characteristics.

In the integration model, the prediction effects under the three integration strategies are also different. Using the integrated model of average weight strategy and stack strategy, the prediction effect is better. The accuracy is 78.8% and 80.2%, respectively. Meanwhile, the overall prediction error of the integrated model of stacking strategy is smaller, and its MAPE is as low as 0.003991, and R_2 is as low as 0.9956. But on the other hand, the prediction effect of the integrated model of global learning strategy is not very good, with an accuracy of 75.5% and a MSE of 0.0438, which is the largest error among the tested models, and the prediction effect is even worse than that of the single model.

*6.2. The Stability and Validity of the Model.* In the process of experiment, we set the maximum training session as 100 times, but in the process of model training, it is often difficult to get the optimal model.

In most cases, it fluctuates at a certain level after the model training is stable. Whether the model parameters can be stabilized around the optimal parameters is an important reference to evaluate the stability of the model effect, as shown in Figure 11.

The figure reflects the convergence process and stability of each model after convergence. It can be seen from the figure that each model has basically converged after 30 rounds of training, but each model has fluctuations. In the MSE curve, the value is small and stable: the integrated model of the average weight method integration strategy and the stacking strategy, and the MSE is stable at [0.025,0.035]. In the accuracy figure, the more stable accuracy is also the integrated model of the average weight method integration strategy and the stacking stack strategy, and its accuracy is stable at [0.75,0.83].

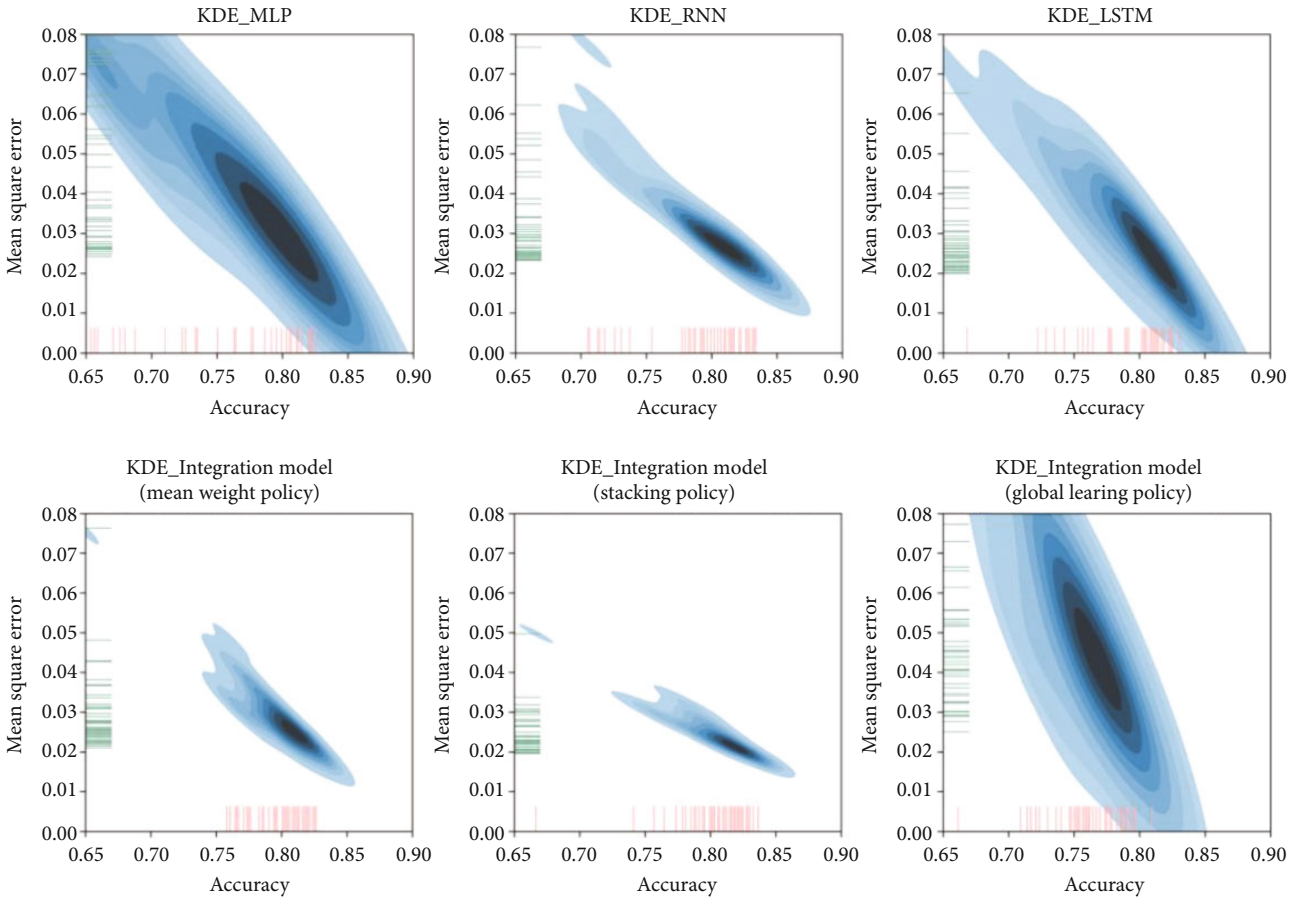



Figure 13: Two-dimensional distribution of prediction accuracy and error after stable model training.

In order to more accurately evaluate the stability of model training, we recorded the changes of MSE and accuracy of different models during the training process. Meanwhile, in the training round (the last 50 rounds) after confirming the stable fluctuation of the model, we drew the kernel density estimation curves of MSE and accuracy of each model. It reflects the effectiveness and stability of individual learner and three kinds of integrated learning models.

As can be seen from Figure 12, the density peak position of the integration model of the integration strategy of the average weight method and the integration model of the stacking strategy is close to the density peak position of the basic model, while the density of the integration model of the integration strategy of the average weight method and the integration model of the stacking strategy is more concentrated at the peak. That is to say, the accuracy error of these two integrated models is smaller, and the possibility of getting a better model is greater. However, the accuracy of the integrated model of global learning strategy is worse than that of each basic model. The expected accuracy of the integrated model of global learning strategy is worse than that of the single model, and the accuracy of the corresponding density peak is even lower than that of any single model.

According to the fourth picture in the figure, it can be seen that among the three kinds of integration strategies, the density peak of accuracy of the integration model of the integration strategy of the average weight method and the integration model of the stacking strategy is similar, that is, the error of stability is smaller in the training process and the effect is better. Relatively speaking, the density peak of the integration model of stacking strategy is more to the right, and the accuracy of the corresponding peak is higher than that of the integration model of the average weight method integration strategy. This means that the integrated model for the stacking policy works better.

As shown in Figure 13, for the training round (the last 50 rounds) after the stable fluctuation of the model, a two-dimensional kernel density estimation diagram of its accuracy and mean square error was drawn, as shown in the figure. The larger the area of blue shadow in the figure is, the more dispersed the density is, which means that the accuracy and mean square error of the model fluctuate more and become more unstable during the training process. As can be seen from the diagram, the use of the average weight strategy integration model and integrated model of the stacking strategy of the shadow area is small, intensity bigger, which means that the model is more stable in the process of training, and maximize the density of these two models (i.e., the deepest shadow color) for greater accuracy, the corresponding smaller mean square error (MSE). This means that the expected performance of these two models is better than that of the other models.

Table 4: Expectation and standard deviation of models, accuracy, and error.

| Model | Accuracy | | Mean square error | |
|---|---|---|---|---|
| | Mean | Standard deviation | Mean | Standard deviation |
| MLP | 0.73756 | 0.075862 | 0.05546 | 0.036732 |
| RNN | 0.79318 | 0.036592 | 0.03173 | 0.011244 |
| LSTM | 0.78512 | 0.054487 | 0.03355 | 0.026722 |
| GRU | 0.78798 | 0.039606 | 0.02939 | 0.011710 |
| Integration model (mean) | 0.79393 | 0.027936 | 0.02919 | 0.008925 |
| Integration model (learning) | 0.74497 | 0.049791 | 0.05828 | 0.037212 |
| **Integration model (stacking)** | **0.80224** | **0.028740** | **0.02441** | **0.005195** |

In order to concretely evaluate the error fluctuation, the expectation and standard deviation of the mean square error and accuracy of the training round (the last 50 rounds) after the stable fluctuation of the model were calculated. The calculation formula is as follows:

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i,$$
$$\sigma = \sqrt{D(X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2}. \tag{6}$$

In Table 4, the expected values of model accuracy and model mean square error are shown, in which the best effect is shown in bold. The integrated model with stacking strategy has the highest expected accuracy and the lowest standard deviation of accuracy. Meanwhile, the expected mean square error and the standard deviation of the mean square error are the least. The above indicates that this model expects the best effect, and its effect is the most effective and stable.

## 7. Conclusion

In this paper, four deep learning network models are constructed and three strategies are adopted to combine them into three integrated networks. Through the training of the model, the performance of the model is reflected by the data of the test set. The accuracy and error of the model were recorded during the training process. Through the experiment, we found the following:

(1) To take the average weight strategy composed of the integrated model and to take the stacking strategy composed of the integrated model, the effect of these two models is better than the effect of a single neural network. By recording the fluctuation of accuracy and error during the training of each model, we analyzed the stability of the model. It is found that the value of the peak accuracy density of the ensemble model composed of the average weight strategy and the ensemble model composed of the stacking strategy is higher than that of the basic model and the ensemble model adopting the global learning strategy. The density peak of the integration model adopts the first two integration strategies simultaneously. It is higher than the peak density of single strategy and global learning strategy. This means that the first two models, performance and stability, are expected to be better than other models. At the same time, these two models have stronger robustness. By comparing the integration model formed by the two combined strategies, we find that the integration model with stacking policy has the highest expected accuracy. At the same time, the expected error of the model is minimum, which means that the model takes into account the stability while optimizing the effect

(2) The integrated model composed of global learning strategy has the worst effect among all models, and its performance and stability are even inferior to that of the single base model. Through the analysis of the weight layer of this network, it is found that in the process of training, this strategy not only promotes the training of each single basic model in the process of error back propagation but also affects the weight allocation of each basic model by the metalearner. This leads to a complex game relationship between individual learner and metalearner, that is, when an individual learner with poor performance is optimized, the metalearner reduces the weight of the output of the individual learner. In this case, the better performance of the learner is constantly changing, which makes the metalearner unable to confirm an optimal weight strategy. In this case, the larger fluctuation of the whole model during the training process can also be explained theoretically

(3) Disadvantages of the model. Through the above experiments, we find that a more stable model can be obtained by integrating a variety of neural network models. However, the improvement of this stability requires a large time cost. For example, for a model integrated with four individual learners, four models need to be trained separately. This increases the time cost by four times, but the performance improvement is small

Future research will try to use the overlay strategy module integration model built in this paper. This multimodal

model will be applied to more fields. We will forecast the indicators of financial products such as foreign exchange, commodities, bonds, and futures. We plan to use an integration strategy that combines several different deep learning networks to build a better, more stable model. With the help of these integration strategies, the application of traditional machine learning thinking to the field of deep learning has great inspiration for the construction of multipattern deep learning models. With the development of the era of big data, the advantages of this new integration approach will be gradual.

## Data Availability

The source of the data has been declared and is readily available on the network.

## Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

## Authors' Contributions

Hanglin Lu and Xiuyun Peng contributed equally to this work.

## Acknowledgments

National Innovation Training Program for College Students. (Item No.202110459137) Website: http://gjcxcy.bjtu.edu.cn/Index.aspx. The paper is funded by the "national innovation training program for college". The fund aims to encourage college students to innovate and apply for it through the project plan.

## References

[1] M. F. M. Osborne, "Brownian motion in the stock market," *Operations Research*, vol. 7, no. 2, pp. 145–173, 1959.

[2] B. Chairman and E. F. Fama, "Efficient capital markets: a review of theory and empirical work†," *The Journal of Finance*, vol. 25, no. 2, pp. 418–420, 1970.

[3] C. Jianfu and L. Xingxu, "Stock price forecasting: a comparison between GARCH model and BP neural network model," *Statistics and Decision Making*, vol. 6, pp. 21-22, 2004.

[4] F. Allen and R. Karjalainen, "Using genetic algorithms to find technical trading rules (revision of 20-93)," Wharton School Rodney L. White Center for Financial Research, 1999.

[5] H. Tongxing and C. Fangfang, *Journal of Chongqing University of Technology (Natural Science)*, vol. 30, no. 2, 2016.

[6] W. Jun, Z. Peng, and Y. Shuai, "Comparison of SEQ2SEQ RNN and LSTM models based on stock forecasting," *Time Financial*, vol. 35, pp. 381-382+392, 2018.

[7] I. Yenidoğan, A. Çayir, O. Kozan, T. Dağ, and Ç. Arslan, "Bitcoin forecasting using ARIMA and PROPHET," in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pp. 621–624, Sarajevo, Bosnia and Herzegovina, 2018.

[8] D. Nelson, A. Pereira, and R. Oliveira, "Stock market's price movement prediction with LSTM neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, USA, 2017.

[9] Q. Xie and X. X. Cheng Gengguo, "Research on stock forecasting model based on neural network enrollment learning," *Computer Engineering and Applications*, vol. 55, no. 8, pp. 238–243, 2019.

[10] X. W. Chen, W. Y. Zhu, X. M. Qian et al., "Estimation of surface layer optical turbulence using artificial neural network," *Acta Optica Sinica*, vol. 40, no. 24, article 2401002, 2020.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning (Vol. 1)*, MIT Press, Cambridge, 2016.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction usingthe ARIMA model," in *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pp. 106–112, Cambridge, UK, 2014.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, https://arxiv.org/abs/1412.3555.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 1628996, 15 pages
https://doi.org/10.1155/2021/1628996

*Research Article*

# Electronic Guidance Cane for Users Having Partial Vision Loss Disability

**Asad Khan**,[1] **Muhammad Awais Ashraf**,[2] **Muhammad Awais Javeed**,[2] **Muhammad Shahzad Sarfraz**,[3] **Asad Ullah**,[4] **and Muhammad Mehran Arshad Khan**[5,6]

[1]*School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China*
[2]*School of Information Engineering, Chang'an University, Xi'an 710064, China*
[3]*Department of Computer Science, FAST-National University of Computer and Emerging Sciences, Faisalabad, Pakistan*
[4]*Department of Mathematical Sciences, Karakoram International University (KIU), Gilgit-Baltistan, Pakistan*
[5]*Department of Examinations, GC University Faisalabad, Pakistan*
[6]*School of Computer Science and Technology, Chongqing University, Chongqing 400044, China*

Correspondence should be addressed to Asad Khan; asad@gzhu.edu.cn
and Muhammad Mehran Arshad Khan; dr.mehranarshad@gcuf.edu.pk

Received 5 July 2021; Revised 3 August 2021; Accepted 27 August 2021; Published 5 October 2021

Academic Editor: Mamoun Alazab



Vision is, no doubt, one of the most important and precious gifts to humans; however, there exists a fraction of visually impaired ones who cannot see properly. These visually impaired disabled people face many challenges in their lives—like performing routine activities, e.g., shopping and walking. Additionally, they also need to travel to known and unknown places for different necessities, and hence, they require an attendant. Most of the time, affording an attendant is not easier and inexpensive, especially when almost 2.5% of the population of Pakistan is visually impaired. There exist some ways of helping these physically impaired people, for example, devices with a navigation system with speech output; however, these are either less accurate, costly, or heavier. Additionally, none of them have shown perfect results in both indoor and outdoor activities. Additionally, the problems become even more severe when the subject/the people are partially deaf as well. In this paper, we present a proof of concept of an embedded prototype which not only navigates but also detects the hurdles and gives alerts—using speech alarm output and/or vibration for the partially deaf—along the way. The designed embedded system includes a cane, a microcontroller, Global System for Mobile Communication (GSM), Global Positioning System (GPS) module, Arduino, a speech output module speaker, Light-Dependent Resistor (LDR), and ultrasonic sensors for hurdle detection with voice and vibrational feedback. Using our developed system, physically impaired people can reach their destination safely and independently.

## 1. Introduction

Vision is, undoubtedly, one of the most important senses for humans as humans get 83% of information from the environment via this sense. Unfortunately, as per this study [1], carried out by the World Health Organization (WHO) in 2011, 285 million people are visually impaired, 39 million are completely blind, and 246 million had weak vision/sight. In Pakistan alone, ~2 million people are suffering from blindness/sightlessness and visual impairment [2]. Vision loss disability or sightlessness is a phenomenon of lacking the vision perceptions, due to physiological or neurological mental factors, having a visual acuity of just 1-2/10 with both eyes open (less than or equal to 30 degrees [3]). This fraction of people faces many challenges, like performing simple daily routine activities, and hence always remains dependent on others [4–6]. This is not just the loss of the individual but also for the country, as they are unable to play their role in the growth of their country's economy [7]. Using a "white cane" simple stick (introduced by James Biggas of Bristol in 1921 [8]), as an indication, was the only way out for a visually impaired person to move around. The US congress proclaimed

October 15 the "White Cane Safety Day," to show love, respect, and sympathy to the visually impaired people.

In the 1960s, the research on assistive technologies—associated with data transmission [9, 10], navigation, and orientation aids, to confer accurate help to the visually handicapped person—was initiated [11–14]. For example, the first approach to observe the hurdle (coming in the way of an impaired person) was the thought of a vector field bar graph. The technique, somehow, solved the indoor navigational activities; however, it was less helpful for the outdoor activities [14]. Hence, the developed system was considered overcomplicated and less accurate and needed a mobile terminal to send and receive the information [15]. Later, the use of a microcontroller in a renowned system, namely, "NAVEBELT and Guidance," consisted of a series of devices (placed at the belt), to observe the hurdles [16]. Afterward, Yuan provided a device for varying sensing and surrounding discovery, intending to assist a visually impaired person. The device measures active triangulation and observes the environmental feature exploitation using the Kalman filter tracking [17, 18]. Another proposed solution takes the help of a camera and a transmitter to transmit the real-time video to the server—to be monitored and assisted by another person. This system also proved not accurate especially in outdoor applications and could not help the person much [19]. Bolgiano and Meeks [20] proposed a visual aid system based on the GPS, GSM, and ultrasonic sensors to detect the hurdles and to obtain the information related to user location [21].

Nowadays, different kinds of those canes are introduced, e.g., smart cane [22] and optical maser cane [23]; however, these tools have many constraints: unnecessary longer length of the cane and limitations in recognizing obstacles, making it difficult for the person to access public places. Recently, some other solutions are also proposed like GPS devices for landmark identification (near-infrared (IR) lightweight or radio frequencies), supersonic obstacle detectors (sonar, UltraCane, Miniguide, Palm-sonar, Ultra-Body-Guard, and iSONIC cane), and optical devices (the laser long cane) [24–29]. However, these devices have shown to be less useful in crowded environments especially in outdoor environments, mainly because of the multiple reflections. Alternatively, several techniques have been developed to revisit the quality of life of visually impaired people by introducing smart devices with built-in signal processing and sensing technology. These referred to electronic travel aid (ETA) devices which facilitate the blind to maneuver freely in an atmosphere dynamically changing in real time. As per the literature, ETAs are broadly classified into 2 major types: sonar input systems (laser signal, infrared signals, or supersonic signals) [30, 31] and camera input systems (consisting principally of a mini-CCD camera and fuzzy system) [32–35]. Bat K sonar, sensible cane, smart vision, and guide cane to observe obstacles in front of the person by transmitting and receiving the mirrored wave normally used supersonic sensors or optical maser sensors [36–39]. In response to detected obstacles to warn the person, it produces either associate audio or vibration. Systems like voice [40], Sound View [32], SVETA [41], and CASBLIP [42] use a single camera or stereo video cameras mounted on a wearable device to capture pictures. These captured pictures are resized and processed to regenerate the corresponding speech, audio, musical sounds, or vibrations. In such systems, the frequency of warning sound signals is related to the orientation of pixels. Some advanced systems use Global Positioning System (GPS) integration with the main system. It is also noteworthy that GPS receivers are beneficial for understanding the present location of the subject and close landmarks. Some solutions are already accessible within the market such as UltraCane [43], iSONIC [44], and Teletact [45, 46]. These products facilitate the visually impaired ones by grouping the data collected through sensors and then transmitting the recommendations through vibration or sound messages to the user. ETAs provide a warning by using sensing modality or/and tactile signals once an obstacle is estimated within the range and recommend the user to avoid it.

Blind aid and security systems have already been prepared with many other solutions. However, none of them is capable of completely deterring the needs of the visually impaired person. In summary, the main contributions of the proposed work are as follows:

(i) We provide an accurate and usable proof-of-concept electronic stick to make the user's life easier

(ii) Our approach used supersonic sensing technology to detect the hurdle and generates an audio alarm/-vibration to get the user's attention

(iii) Our prototype uses the LDR (Light-Dependent Resistor or photoresistor) to detect the darkness around the user and generates a darkness alert and lights the stick in parallel

(iv) Additionally, our system also detects water and smoke as well. Our solution uses GSM and GPS for the detection of the location of a user, and in case of an emergency, a user can press an emergency button to automatically generate and distribute a SMS to the desired person

The rest of this paper is structured as follows: next, we will discuss in Section 2 a concept of the indoor guide system based on partial vision loss which is presented with a detailed description of the design and further development of electronic apparatus used in the proposed electronic guidance cane, based on outdoor detection of objects, and direction indication-simulation system based only on the detection of objects. In addition, our experimental results for a lab-scale prototype are provided, and at last Section 3 concludes the paper.

## 2. Materials and Methods

Presently, visually impaired people normally use a white cane as a tool for directing them when they move or walk for their daily routine work. Here, in Figure 1, a smart electric cane prototype design was developed as a tool that can serve as an electric guidance cane for visually impaired

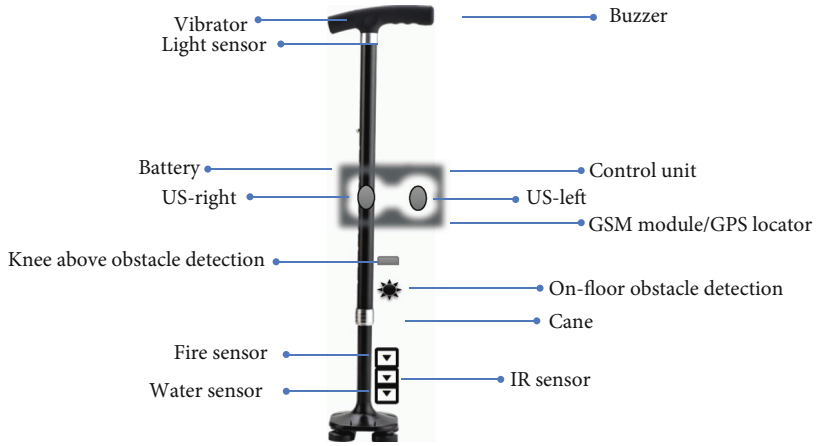



Figure 1: Prototype of the smart electronic stick for visually impaired humans.

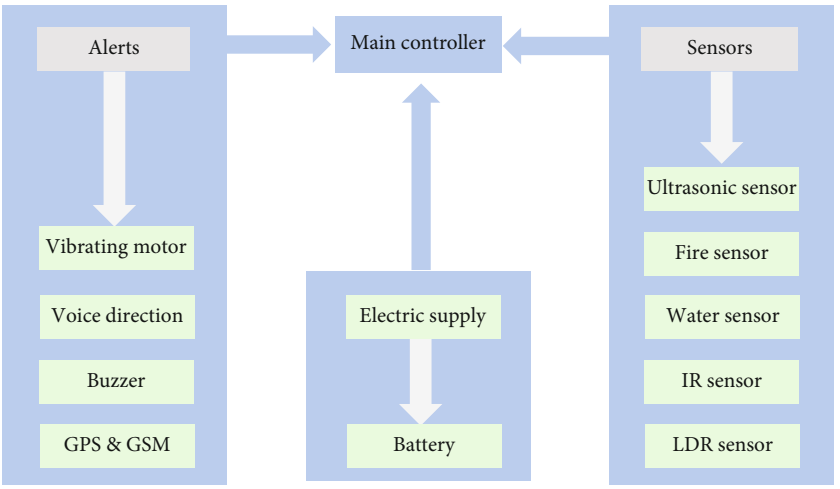



Figure 2: The block diagram of the smart stick of visually impaired human.

persons being a more efficient and helpful technique than the conventional one. The prototype explains different parts of the proposed system, and an ultrasonic sensor is used on the top side and bottom right and left for obstacle avoidance. The vibrator is fixed on the top of a stick which vibrates when an obstacle is encountered which helps in alarming the visually impaired person and allows that person to change their path. On the bottom of the cane, an IR sensor is used for pit and staircase detection to guide someone in that direction where there are fewer obstacles. To avoid sleeping in a water area, a water sensor is attached to the initial side of the guided system, and for fire protection, a fire sensor is placed on the initial bottom; a light sensor is useful at night which alerts the people in the surrounding area that a visually impaired person is walking and allows space so that the person can walk easily.

For further discussion and simplicity, Figure 2 illustrates the overall working mechanism and the features included in the proposed designed electric guidance cane concept. As shown within the figure, the Arduino Uno microcontroller is the heart of the proposed system as all the individual units are controlled and interfaced with the heart of the machine. The supersonic detector and the GPS unit act because the input to the microcontroller provides the obstacle data and user data severally when walking and using this embedded system. The LCD, GSM unit, and buzzer unit essentially are the outputs results from the microcontroller. The buzzer unit gives alerts via voice when the user encounters an obstacle in front of a specified distance. In case of emergency and any failure, the GSM unit provides message service to the expensive ones of the user. IR and LDR represent the obstacle avoidance sensor as a heat-sensitive sensor and light-dependent resistance, respectively. LDR and IR first sense the intensity value of an object in front of the user and send it to a microcontroller. The rest of the sensors including the water sensor work in the same procedure. After receiving the data, the microcontroller converts it into different discrete values from initial step 0–1023 and checks whether the received value is above the threshold level (a limit value that is set independently by the visually impaired person from the range of discrete values: 0–1023) or not; it will then be considered as there is no object in front of an electric guidance cane, and the buzzer alarming will remain off; if the received value is below the threshold level, the

Table 1: Specification of electronic components used to design the proposed system.

| Components | Specifications |
|---|---|
| Arduino Mega2560 | Digital I/O pins:54; clock speed: 16 MHz; operating voltage: 5 V |
| Ultrasonic sensors | Min-max range 3-400 cm; current consumption 15 mA |
| GPS module | Position accuracy: 1 m; update rate:18 Hz; sensitivity: -167 dBm |
| GSM module | Baud rate: 9600; supply voltage: 12 V DC; data bits: 8 |
| Voltage regulator | Min-max input voltage is 7 V-25 V; operating current (IQ) is 5 mA |
| Rf module | Operating voltage: 3 V-12 V; frequency: 433 MHz; current: 5.5 mA. |
| Vibratory motor | Maximum operating torque kg cm: ST-10 is 21.0 |
| Switches | Input output voltage:5 V-120 V |
| Capacitor | Max energy density: 2.6 W/kg; rated voltage: 48 V |
| Diode | Voltage of around 0.2 to 0.3 V |
| SD card module shield | Chipset: AMS 1117-3.3 V; English manual/spec: yes |

microcontroller will consider it as an object or hurdle in front of the guidance cane. During outdoor activities, if the value of IR and LDR is high and detects an object in front of a visually impaired person, then the vibrator will be on, or if the ultrasonic sensor value is high and identifies an object, then beep will be generated, voice coming from the speaker will tell the user to move right or left, and the light on the stick will also shine. For the practical decomposition of the merchandise, the structure of the proposed architecture is convenient and efficient enough to cover all aspects of the observant because it demonstrates the presence of every module ranging from its arrival until the service is served to the desired user and it describes that the modules are interlinked and intercepted with one another and work along to attain the required goal.

*2.1. Electronic Components.* Multiple electronic materials and apparatus are used for building electronic guidance cane circuits. Our proposed circuit designs contain these materials and apparatus that are described in Table 1.

*2.1.1. Arduino Mega2560.* The Arduino Mega2560 microcontroller board (licensed under a Creative Commons Attribution-Share-Alike 2.5) is based on the ATmega328 series controllers, and the schematic of the Arduino Mega consists of three blocks: the voltage regulator, the ATmega2560 with supporting circuitry, and the header. The default input is Arduino (Atmega), so they do not need to be explicitly declared as inputs with pin Mode () when we are using them as inputs. This means that it takes very little current and energy consumption to move the input pin from one state to another. Arduino Mega based on ATmega328 15, the main features of Arduino Mega (ATmega2560 running at 16 MHz with an external resonator (0.5% tolerance), also USB connection off the board it which supports auto-reset,5 V regulator which weighs covers <=2 grams the current performance DC input 5 V up to 12 V, Onboard power and status LEDs.

*2.1.2. Ultrasonic Sensors.* The sensor consists of three parts: transmitter, receiver, and control circuit. For robotics and hardware embedded systems, ultrasonic sensors are relatively simple and easy to interface. We have used the sonar sensor HC-SR04 in our project to detect hurdles on the left and right and in front of the stick and on the footstep. Sensors calculate the distance between the hurdle and receiver of the sensor. To get a desired obstacle localization effect, it is not enough to just improve the performance of the sonar sensor; filtering the measured data is additionally necessary. In this paper, the sequential sonar returned data is processed through the dynamic filtering method using the orientation and therefore the trajectory information of the cane control box. It can sense hurdle in a range of 2-80 cm. We can set the range of sensor according to our desire. The working principle of sonar needs a high-level signal for at least of 10 $\mu$s using an IO trigger. It sends pulses of 40 kHz, and the pulse back signal is scanned. Sonar has main parts called transducers who send pulses of eight 40 kHz and sense returned pulses or echo: an ultrasonic sensor emits ultrasonic wave through a transmitter which spread in the air when it hit any hurdle and then returns to the ultrasonic sensor receiver and a counter stops counting time when it receives reflected wave. For the basic purpose of choosing, we have used three ultrasonic sensors on the left and right and in front of the stick to detect any hurdle accurately. Ultrasonic sensors are very useful and cost-effective. There is no effect of sunlight on the performance of these sensors. We can use it for both indoor and outdoor systems. It has a range of 2-400 cm. We have a fixed range at 45 cm which means if there is a hurdle in 45 cm, then there would be an alert for a blind person in a form of audio voice. In Table 2, we discuss the specifications of ultrasonic distance sensor-HC-SR04 embedded with an electric cane control box using an Arduino microcontroller.

In the control box, we use four HC-SR04 sensors, which have the main functionality to detect obstacles in front of the cane within the mentioned distance in the table and guide a disabled person about obstacles and hurdles indoor and outdoor and further demonstrate a person to move left or right via signals sent to a microcontroller and attached speaker pronounce voice command, as shown in Figure 3.

Table 2 demonstrates the values of specific working voltage, measurement range, I/O pins needed, operating current, and dimensions.

*2.1.3. GPS Module.* In 1960, U.S. Air Force invented a satellite-based navigation system. The tracking sensitivity of GPS is -165 dBm. The working principle of GPS is that it can connect with 24 satellites which continuously encircle the earth once in 12-hour duration and provide very useful

Table 2: Ultrasonic distance sensor-HC-SR04 specifications for a microcontroller electric cane control box.

| Sensor | Ultrasonic distance sensor | HC-SR04 |
|---|---|---|
| Working voltage | 3.3 V/5 V compatible<br>Wide voltage level: 3.2 V–5.2 V | 5 V |
| Measurement range | 3 cm–350 cm | 10 cm–440 cm |
| I/O pins needed | 3 | 4 |
| Operating current | 8 mA | 15 mA |
| Dimensions | 50 mm × 25 mm × 16 mm | 45 mm × 20 mm × 15 mm |
| Ease of pairing with Raspberry Pi & Arduino | Easy and direct connection | Voltage conversion circuit required |

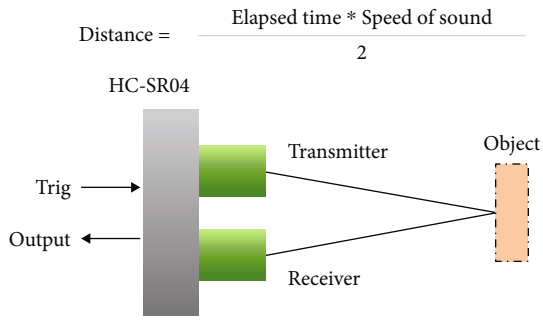



Figure 3

information regarding time, velocity, and position on earth. We can identify the position of any object by measuring the distance of that object from the satellite. The distance from satellite is

$$d = cxT, \tag{1}$$

where $c$ is the speed of light ($3 \times 10^8$ m/sec) and $T$ is the time period of satellite.

GPS needs three satellites for a 2D position, but for accuracy it needs four satellites for providing the location of any object on earth. When someone turned on the GPS module, it firstly downloads orbital information of all satellites and saves information in its memory. Then, the receiver calculates the distance between a satellite and itself by multiplying the velocity of the transmitted signal which is the speed of light $3 \times 10^8$ m/sec$^2$ by the time they receive the transmitted signal from a satellite. Along with the measure of travel time by the receiver, it should also know the exact position of the satellite for accuracy. GPS transmits a signal on two different carrier frequencies: one is L1 which is 1575.42 MHz and the other is L2 that is 1227.60 MHz which is more precise and used for military purposes. In our paper on the purpose of GPS, we are using the (GY-NEO6MV2) GPS module. These apparatus provide the ability to find the path and exact location of a visually impaired person. It may also use GPS for path planning. This GPS receiver with better performance having 20 million correlations helps in finding satellite location as soon as possible when it is turned on. It has six engines and a member of the NEO-6 series of GPS receivers. The power supply range is from a 3 V to 5 V LED signal indicator and 3 configuration pins with a 1-time pulse which supports DGFS (SBAS, WAAS, EGNOS, and MSAS) at a maximum altitude of 50,000 m with 500 m/s velocities. The performance of GPS in both cold and warm climate starts to be delivered after 27 sec.

*2.1.4. GSM Module.* GSM was developed in 1970 by Bell Laboratories Inc. This device has great importance in communication to transmit and receive data by using mobiles across the world. It had made communication easy since its development. It works at different frequency bands in which the band is used depending on the application. GSM handles multiple access at a time that is why today's communication system has become better than past years' systems. Its works between frequencies 850 MHz-00 MHz and 1800 MHz-1900 MHz, and data rates from 64 kbps to 120 Mbps can be conceded by this system. Time division multiple access overlapping of one signal to another signal was common in an older communication system. To avoid this major error in a communication system, engineers introduced a new method that is time division multiple access. In this method, some specific time slots were assigned to every user that helped a lot to solve many problems regarding overlapping, but one thing is important that frequency remained the same for all users. The use of the GSM module is that because of its better spectrum competence, international roaming, support for new facilities, and real-time clock with alarm management, phone calls are secure by using encryption and short message service (SMS). Most secure telecommunications currently accessible as security strategies are standardized for it. This module can be used in several applications in GPRS mode remote data logging, transaction terminals, weather stations, security applications, and supply chain management.

*2.1.5. Voltage Regulator.* In this prototype, a voltage regulator, LM 7805, is used. Voltage regulators are used for regulating voltages to our desired circuit voltage requirement, as shown in Figure 4. It will protect our circuit whenever there is excess voltage or current level beyond our circuit limit. It provides a constant voltage at its output terminal. It is a member of 78xx linear voltage regulator ICs, and xx indicates that it will provide a constant voltage level. The voltage regulator LM 7805 ratings described are its input voltage range from 7 V to 35 V, current IO = 1 A, and output voltage $V_{MIN} = 4.8$ V and $V_{MAX} = 5.2$.

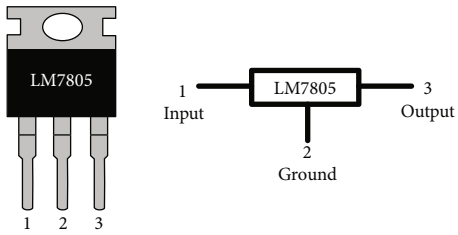



Figure 4

*2.1.6. RF Module.* The RF module wireless system design has two overriding constraints: which operates over a certain limit and specific distance and transfers a specific quantity of convergence results within a data rate. The RF modules are very small in dimension and have a wide operating voltage range; it operationally works between 3 V and 12 V. The working techniques of RF modules are RF transmitter and receiver modules. When transmitting logic 0, the transmitter draws no power while the carrier frequency thus consumes significantly less power in battery operation. From the transmitter, the data is usually sent serially which is received by the tuned receiver. Duly interfaced to two microcontrollers for data transfer, the RF transmitter and receiver are working at frequency 433 MHz, receiver current supply 3.5 mA, sensitivity 105Dbm, operating voltage 5 V. low power consumption with a transmitter frequency range of 433.92 MHz, and transmitter supply voltage of 3 V~6 V with output power of 4~12 Dbm.

*2.1.7. Vibratory Motor.* The vibratory motor is using a vibrator in an electric stick which will activate when there is water in the path of a blind person. It is connected to the bottom of the stick through two terminals when there is water current which will flow through terminals, and the vibrator is turned on and alerts a blind person.

*2.1.8. Switches.* An electric stick had one switch to control power consumption of the battery to get maximum time for the functioning of a blind person stick. The switch is used for the main power to turn on or off. The working usage is that it installed this switch just to save battery consumption.

*2.1.9. Capacitor.* Capacitors are used to filter out frequencies. It acts as a high-pass filter allowing high frequencies to pass and blocking the DC signal as well as a low-pass filter allowing low frequencies to pass and blocking the AC signal. We connect the capacitor in parallel to our circuit instead of series to block the AC signal and get only the DC signal.

*2.1.10. Diode.* A semiconductor diode is a device used for the unidirectional flow of current. It blocks current in the reverse direction. It has two terminals: anode (positive) and cathode (negative). A diode is a PN junction. It works when it is forward biased and behaves like a short circuit. When it is reverse biased, it behaves like an ideal open circuit.

*2.1.11. SD Card Module Shield.* In a smart electric stick to store the voice for four ultrasonic sensors, an SD card shield is installed with Arduino Mega. It is very useful to store different voice files for right, left, and front hurdle at the footstep, fire, water, etc. The SD module can be used by likely different microcontrollers for saving data rates like Arduino, AVR, PIC, and ARM. The module shield follows technical parameters which are operating voltage of 5 V, SPI interface, and dimensions of $20 \times 28$ mm.

*2.2. Indoor and Outdoor Prototype of the Electric Guidance Cane.* As we see in Figure 5, it shows the embedded hardware circuit design of an electric guidance cane based on hurdle detection with multiple hardware items using Arduino Uno, with the sonar sensor HC-SR04. In this methodology, the beep alarm and audio signal will turn to a high state only with the detection of an object in front of the user; otherwise, there is no action and the audio will remain off at that time.

In this scenario, a photocell LDR sensor, LEDs, LCD screen, resistors (220 ohms), micro-SD socket (SD card), ultrasonic sensor (HC SR04), speaker (SPKR 1), rotatory potentiometer (large) (R3), round pushbutton (S1), vibration motor (ROB 08-449), Adafruit Ultimate GPs, GSM (sim 800 L), water level sensor, flame sensor, and single Arduino Uno Mega (2560-ReV3) were used. One leg of the LDR sensor is attached to the microcontroller analog PIN A3, and another leg to a 5 V pin, and the same is done with a resistor which is attached to the GND port of the microcontroller. Besides, the threshold value for the sonar sensor (HC SR04) was adjusted to 5 from the discrete values (0–1023) for understanding whether it is a hurdle or not. On the other side, right and left pins of the sonar sensor pin VCC are connected to a speaker and potentiometer, and trig and echo pin is related to pin D34 and D36 Arduino Uno as well as GNDs. The front sensor (HC SR040) VCC was connected the same as left and right pins, and trig and echo were connected to pins D13 and D14 Arduino Uno and GNDs. Here, Arduino Mega ADK analog pins A1 and A0 are connected to the LDR R1 side and water sensor sport, and other to 3v3 and GND. The other ground of battery's negative also put against to the battery's positive. Furthermore, the D52 (SCK) and D50 (MISO) of Arduino were connected to the GSM port sim_rxd, sim, and txd, as we can see in Figure 5. The Arduino pins D28, D30, and D32 are related to the LDR positive side and D26 with the flame sensor port Ky-026. The other considerable connections are Txl D18 to GPS pin rx, D16 to SD card pin (U RSV), D2-D5 with LCD PIN DB7, and DB6-DB5 and DB4. The rest of the Arduino D8 and D12 with push-button leg 1 is connected to the LCD cathode side.

At the last part of the circuit design, GND negative part of all the obstacle avoidance sensors was connected to the GND port and all VCC (input voltage) to the Arduino 5 V pin. Initially, the sonar HC-Sr04 has a distance set from 2 cm to 1 m (by default) at the start; if there was a hurdle detected, the vibrator vibrates and will generate a beep through the speaker and the voice (turn left and right and forward) will be heard by the blind person.

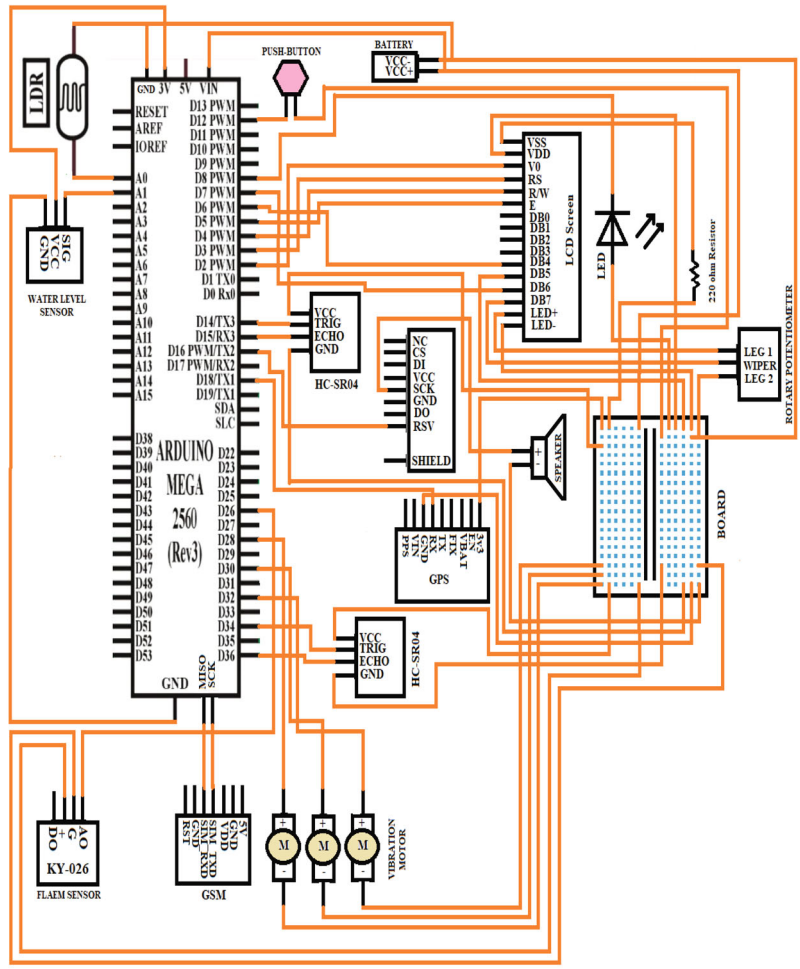



Figure 5: Circuit design of the indoor and outdoor prototype of the smart stick of visually impaired human.

*2.2.1. Indoor Prototype of the Electric Guidance Cane.* Concerning indoor cane usage patterns, a user follows it to fight the steer in its center where they are walking in order that any open proportional font window, door, or split unit mechanical device and house objects might not hurt them. They will swipe the electric guidance cane on the ground rather than the left-right direction technique if the floor surface is okay enough. It is quite difficult to measure the huge area square for the navigation, which does not need personal or infrastructure help. On stairs, they use the pencil vogue technique. They follow the railing of the steps or wall with manus. When returning down, they use a guidance cane to feel and observe the depth and breadth of stairs by a sound alarm and swiping cane. For escalator use in a mall in the door, the bumped start line allows them to comprehend the initial line. To get out of it on the finish, they either get information by sloping down reeling at the top or place a guidance cane on a further next step in order that they shall recognize once it touches the ground level area. One subject aforesaid that he uses to face on heel once approaching towards finish; therefore, he could safely find himself on escalators. For carry usage and other altitude purposes, they use a guidance cane to grasp once the carry arrived if it does not ring. Then, they use a cane to gauge the height of a lift's floor from the building floor, step in, and face within the center of the carry.

*(1) Results and Discussion*. In the beginning, Figure 6 shows the flow of the electric guidance cane for every step of the blind walking. It conjointly shows the sensors and actuator's work and also the control method done by Arduino Uno.

The flowchart of the obstacle detector using the supersonic sensing element is shown in Figure 6, which has 2 components: the initial half part deals with the obstacle detection and the other second half deals with distance measuring and alerting the users counting on the distance of the near obstacle to avoid collision. Counting on the gap of the obstacle from the person, four zones of the area unit formed: far zone, near zone, shut zone, and zone. If the next detected object is at ≥2 meters or more, then it comes underneath the way (safe) zone. If the next item object is found at 1 meter or more, then it comes underneath some close to the zone; if the item object is found at 100 cm or more, then it comes underneath the shut zone; and if the item is detected at 100 cm, then it comes underneath the zone. A voice instruction in conjunction with vibration and a buzzer alert voice is

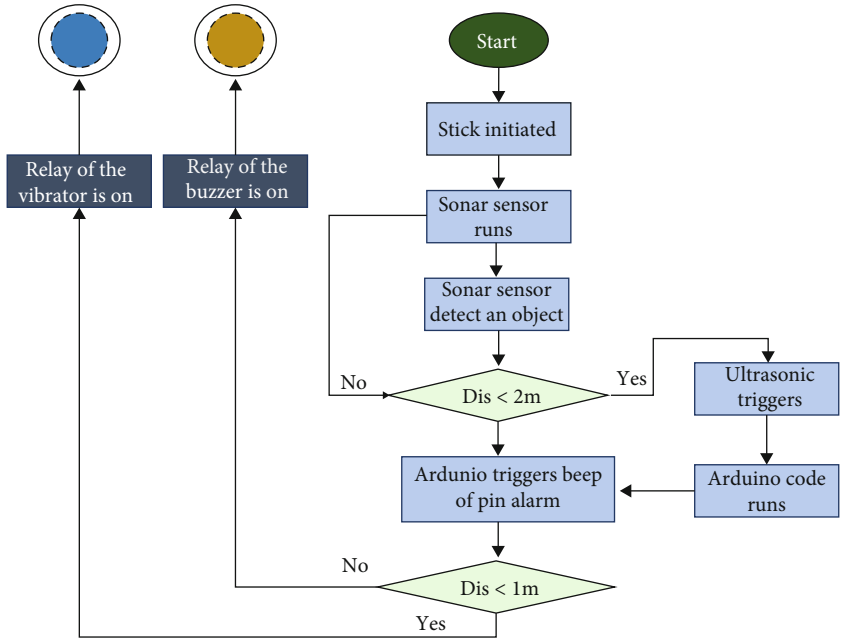



Figure 6: The flow diagram of the indoor prototype of the smart stick of visually impaired human.

processed to a user at each zone to alarm him/her and let individuals around that visually handicapped person to assist. For usage, a disabled person can control it very easily which is our main point to give more priority for partially disabled person and old age humans. When a disabled person wants to use it, he/she just needs to press "Button," then all features and functions will be activated like a microcontroller and sensors with GPS location. He uses his cane to move on road—indoor or outdoor, the cane will guide the user about their path and obstacles via voice (turn right, left, etc.) and send the location to person guardian; also, it can be used to track a disabled person in case of emergency.

Figure 7 shows the final demonstration of the proposed smart aid electronic stick that detects the obstacles and object's movement indoor using Arduino Uno. Figure 7(a) represents a lab-scale prototype of an electronic stick. Figure 7(b) shows the movement of the forward base because the sensed object intensity value of the sonar sensor was less than the threshold value and there was no obstacle detected by any of the obstacle avoidance sensors, so as a result, no beeping alarm turns on. The formal design of the proposed model and methodology can be seen in Figures 7(c) and 7(d), with the motive that only those obstacles will be detected in an indoor area, and the rest of the module including the voice alarm, vibrator, and beep will keep maintaining their state and start working in this condition. As an example, as discussed in Figure 7(c), the sonar sensor detects the initial hurdle and provides a voice assistant to handle the obstacles. Therefore, if the sonar sensor observes the sense intensity value or a value above the threshold, it suggests solutions to handle the obstacles. Moreover, when the object moves to the first obstacle, it indicates to turn right; that condition becomes true. Similarly, when an object moves to the second obstacle, it indicates to turn left (Figure 7(d)). The demonstrated results proposed the efficiency of the idea and give immediate validation for the RF module, vibrator, speaker, sonar sensor (HC SR04), and proposed model. These types of applications can be implemented and installed in the indoor area, parking area, hotels, schools' shopping malls, college's boundary, inside sectors, hospitals, and malls and homes to detect objects and give a straight path to the blind person.

*2.2.2. Outdoor Prototype of a Smart Stick of Visually Impaired Human.* As per our vision, the central idea of this work is to create such an innovation for our current visually impaired persons and an embedded system so that the results of location accuracy and direction can be improved. As presented in Figure 8, in outside use, all of the themes use left and right sound of the cane technique on a flat surface path. The reason of this can be that all of them use a cane with a vacant tip, not with a roller ball that will be swiped on the surface. In an outdoor setting, when a visually impaired person walks with angular deviation, three speech types will be on: turn "right, turn "left," or "forward," dependent on sensor detection varying frequency on the pedestrian's adherence to the way.

*(1) Results and Discussion.* The practicality of the supersonic sensor device remained the same as shown in Figure 6; this could be an additional display in Figure 8. During this, a combination of GPS and GS-modem technologies would possibly offer additional aid for the visually impaired persons.

Initially, whenever there is an emergency, the blind individuals have to be compelled to press the trigger button that activates the GPS and GSM. GPS identifies the placement of the visually handicapped person in real time which is sent to GSM within the sort of coordinates. An alert message is sent together with the precise location of the visually

(a)

(b) (c) (d)

Figure 7: Result diagrams of the indoor area of the smart aid electronic stick to move "forward," "turn right," and "turn left" at object detection. (a) In the real-time simulation, the proposed system finds the path by the situation of the surface and the place of the object on it and helps decide to which direction to move. (b) In the representation, they check the obstacles if there is any and then change their path which is dictated by speaker voice. (c) When an object comes in front of an obstacle avoidance sensor, the first set dictates a command to turn left while the other remaining dictations are neglected. (d) When motion is in front of the second obstacle sensor, only the part of the second set of the visually impaired person dictates a command to turn right, and all the other positions are neglected.

handicapped person to the receiver. For additional aid, supersonic detectors with voice recognition also are accustomed to finding an obstacle and active torch sensors. Therefore, this stick will not be misused by others except for the approved users. Figure 9 shows the design results of the outdoor prototype of an electric guidance cane for users having a vision loss disability using the Arduino Mega microcontroller. In this way, Figure 9(a) represents a visually impaired person from outside of the city where the location can be traced, and a user can also follow the path which he decided at the initial start level when he starts a journey. On the other hand, in Figures 9(b) and 9(c), the person turns right and left on the surface sonar sensor placed on the electric cane because there was a motion that was detected by the first IR obstacle avoidance sensor. At the same time, the object moved forward from the second avoidance sensor, then the third obstacle avoidance sensor detected its motion, and a voice alarm coming from the speaker helps the person to choose which side is better. Meanwhile, when a fire was detected by the flame sensor placed on the cane, the vibrator automatically operates to alert the person who is visually disabled; also, the water sensor senses the road water or depth of the road if there is any, as seen in Figures 9(b) and 9(c). In Table 3, the functionality of old systems differs from the new purposed embedded system of the electric guidance cane for users having vision loss problem. In the old system,

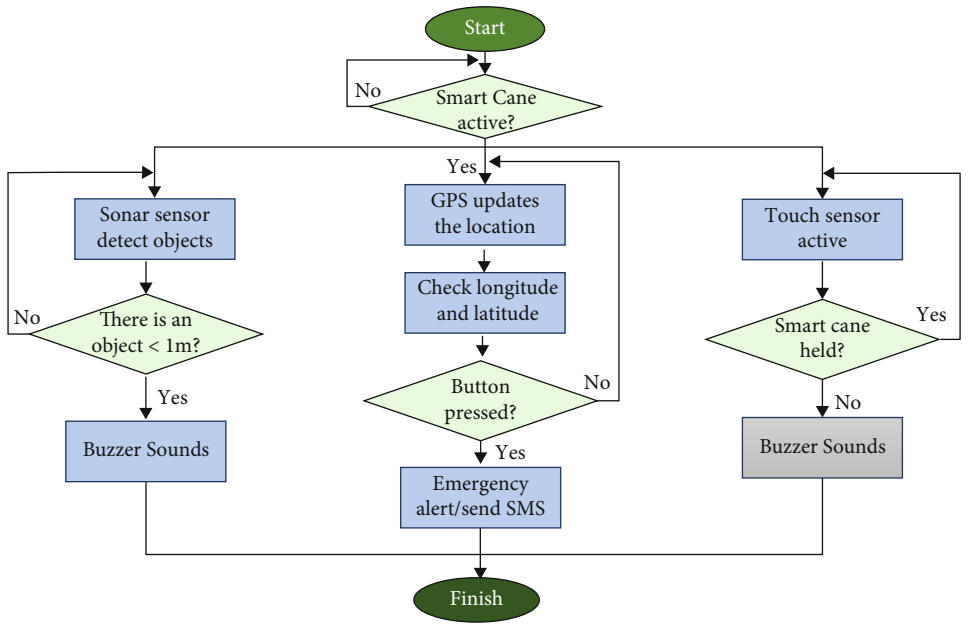



Figure 8: Presented flow diagram of the outdoor prototype of the smart stick of visually impaired human.

(a) (b) (c)

Figure 9: Result status diagrams of further enhanced work with an outdoor prototype of the electric guidance cane of visually impaired human. (a) During an outdoor real-time simulation, working on location, GPS and GSM modules were employed to check the longitude and latitude. (b) Display hurdle in front of the first sonar obstacle avoidance sonar sensor: turn right. (c) Motion in front of the second sonar obstacle avoidance sonar sensor: turn left. Water, road depth, and flame are detected in front of the stick, so a vibrator vibrates and the speaker generates a relevant direction command.

a basically single microcontroller was used with single distance and infrared sensor to guide the user in only prescribed one way direction. On other hand, some used only an obstacle sensor or flame sensor and different engineers used GPS for location. At the end, we study and concluded that the overall solution of this problem is to minimize the

Table 3: Comparison between old systems and the proposed system.

| Functionality | Old systems | Proposed system |
|---|---|---|
| Based on microcontroller-Arduino | Yes | Yes |
| Operates with the sonar and infrared module | Yes | Yes |
| Flame and water intelligence and sensing capability with hurdle detection generating vibration and sound to divert the person to the right direction where there is no obstacle | No<br>No | Yes<br>Yes |
| Locate the location of the blind person and give alert to the guardian (longitude & latitude capability) | No | Yes |

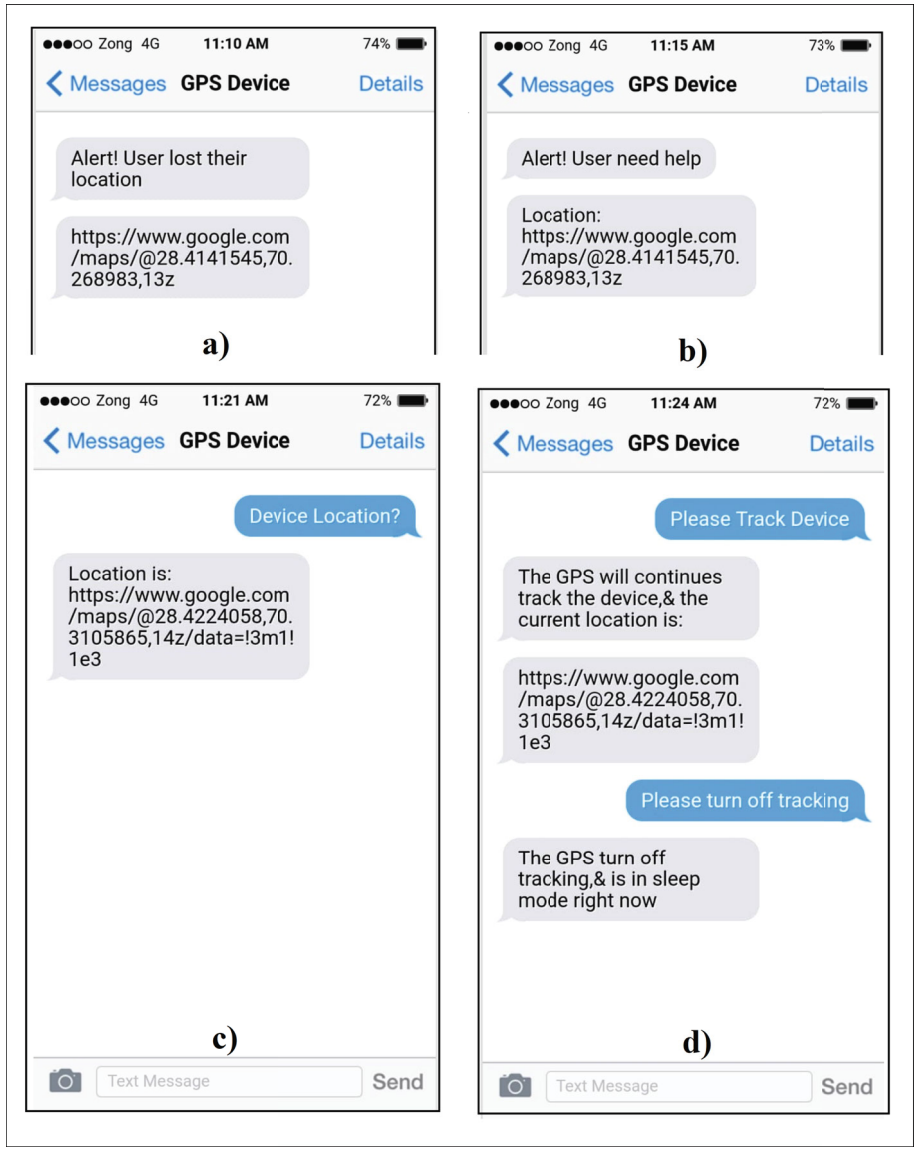
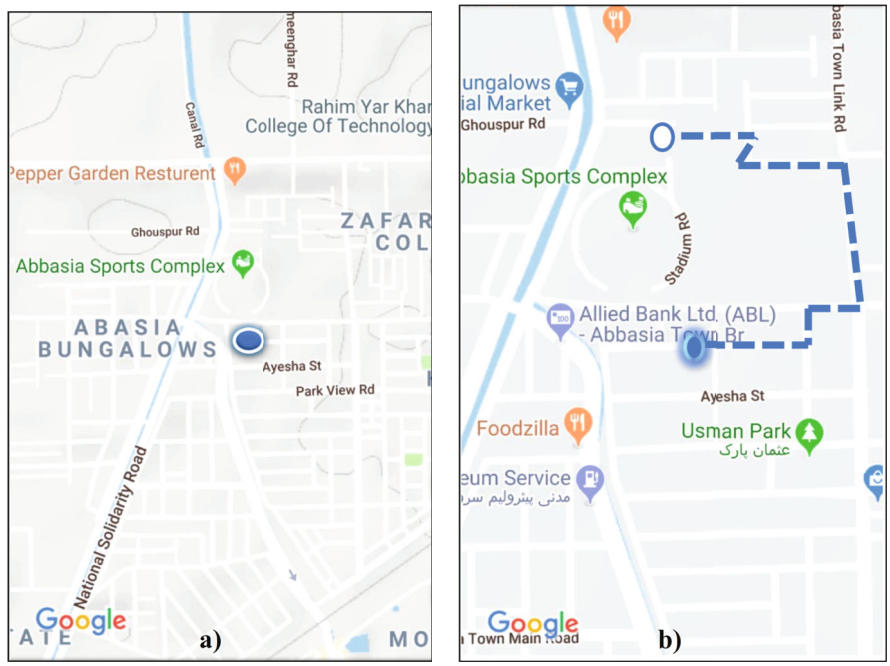



Figure 10: Alarming and alert notification when a disabled person (a) lost their path and (b) required help, when (c) the device location is requested by a guardian and responsible person, and when (d) GPS (information) tracking queries and live data are asked by a family member or guardian.

cost and increase the usability of the device with lots of other features. In this study, the authors have designed a low-cost system. Therefore, the system has various productive functions. First, it detects obstacles in multiple directions and informs the user through voice and beep alert, detecting also flame and ground depth. Second, the system also detects water existence and movement. Third, the system is also beneficial as it sends and receives the real-time location. At the end, if the visually impaired lost their direction, then he/she alarms other persons nearby by pressing a button on the cane, hinting them that he/she lost his/her path, and a message is also sent to the guardian via longitude




(a) (b)

Figure 11: Device information: (a) current location and (b) device path information tracking.

and latitude as we discuss in Figures 10(a) and 10(b) with a real-time location of that person.

*2.3. User Path Finding.* We designed the proposed embedded system and tested its functionality and results with a partially visually impaired person holding it, as featured in Figure 10. The alarm SMSs transmitted from the system to mobile phones of the disabled person's family and guardian are shown in Figures 10(a) and 10(b) for generating alert and request to need help, respectively. The SMSs with the embedded system device's current location and tracking path and status requested by the guardians are featured in Figures 10(c) and 10(d), respectively. Finally, a web link is generated, and the current location of the electric guidance cane in Google Maps and the available tracking device are shown in Figures 11(a) and 11(b), respectively.

In our country, Pakistani visually impaired persons (aged between 17 and 60 years) who used the prototype, as shown in Figures 8 and 9, respectively, reported clear benefits from the ability to alarm and get connected with their family and guardians. Moreover, the data analyzed to check multiple users how to use this guidance cane and reported that the guidance system improved their life style by giving them additional confidence and life easiness.

We used the microcontroller Arduino-based electric guidance cane with GPS and GSM wireless connectivity, in which the visually impaired person could automatically control the directions based on some sensors which the provide the data in the form of direction, buzzer, and alert via voice, to avoid any hurdles which come in a path or way. In some cases, we required wireless connectivity to make the prototype more scalable to find the current location of a new or existing visually impaired person and online access to the information managed by the microcontroller boards to develop a user-friendly interface, and a message is sent to the required guardian. In this regard, the current location longitude and latitude can be used for access from the Internet, and Google Maps would provide easy-to-use access to the location of visually impaired humans. Meanwhile, to store previous location record history and secure the acquired information of the prototype for further study and analysis, different kinds of searching and dictation algorithms could be used in our proposed designs; in fact, a microcontroller is easy to integrate with several sensors. Moreover, the reported systems are presented as lab-scale prototypes; similarly, the proposed embedded systems can be managed for real-scale facilities by applying other technologies. For the question of wireless connectivity, the GSM and GPS are used as shown in Figure 5. The system can be replaced with different sensor types, or a combination of several types, for better results. For example, one can choose a long-range infrared sensor with a maximum range of 1.5 m or a combination of laser diodes and photoresistors.

The feedback of the test on visually impaired persons about the proposed embedded system after participating in the practical experiment in a university of the system was positive. Based on the results and outputs, this kind of system would aid their navigation. The study also collected usability suggestions for further development of the system and real-life use.

## 3. Conclusions

In this paper, we conclude, finalized, and developed an electric guidance cane for the safety, protection, and convenience of visually impaired persons. It guides the person

using voice and by vibration alert. Therefore, it is useful for people that are visually impaired as well as for those who are hearing impaired. It is used for navigation and hurdle detection with the help of three ultrasonic sensors installed in three different directions, i.e., front, right, and left, and for the footstep of the stick. Using the GSM and GPS, the user can send its location to the concerned person; we have developed a smart algorithm with compact hardware design in a small box which comprises of the GPS module, GSM module, Arduino Mega, and water sensor at the bottom and flame sensor. The battery and other components are inside the stick body. This is an efficient design through which visually impaired people can go anywhere without any problem and can easily avoid hitting any hurdle coming in their path. We have installed a GSM module for locating blind people anywhere on earth. There is a specific mobile number that is saved in coding; whenever a specific message is sent by this number to a blind stick GSM number, the location of the blind person would be sent to that specific number. We have used Arduino to make our hardware compact and to avoid making the stick heavy. It also increases the efficiency of the system. This stick is easy to carry by a blind person without any problem. In the future, we will convert a rechargeable battery to solar and road fraction recharge. In this way, the battery life will increase and interlink with the AI camera which would be able to detect obstacles.

## Data Availability

No data were used to support the findings of the study.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

## Authors' Contributions

AK, MAA, and MAJ devised the methodology and acquired funding. MSS, MMAK, and AU carried out the formal analysis and data curation. AK and MSS wrote the original draft, reviewed the writing, and edited the manuscript. AK and AU proofread the manuscript before its final submission. AK, MAA, and MAJ contributed equally to this work.

## Acknowledgments

This work was supported by the Guangzhou Government Project under grant no. 62104301.

## References

[1] R. P. A. Bourne, S. R. Flaxman, T. Braithwaite, M. V. Cicinelli, A. Das, and Vision Loss Expert Group, "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis," *Lancet Global Health*, vol. 5, no. 9, pp. e888–e897, 2017.

[2] T. R. Fricke, N. Tahhan, S. Resnikoff et al., "Global prevalence of presbyopia and vision impairment from uncorrected presbyopia: systematic review, meta-analysis, and modelling," *Ophthalmology*, vol. 125, no. 10, pp. 1492–1499, 2018.

[3] M. Varghese, S. S. Manohar, K. Rodrigues, V. Kodkani, and S. Pendse, "The smart guide cane: an enhanced walking cane for assisting the visually challenged," in *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, Mumbai, India, 2015.

[4] A. Shaha, S. Rewari, and S. Gunasekharan, "SWSVIP-smart walking stick for the visually impaired people using low latency communication," in *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, pp. 1–5, Mumbai, 2018.

[5] M. P. Agrawal and A. R. Gupta, "Smart stick for the blind and visually impaired people," in *2018 Second international conference on inventive communication and computational Technologies (ICICCT)*, pp. 542–545, Coimbatore, 2018.

[6] A. S. Al-Fahoum, H. B. Al-Hmoud, and A. A. Al-Fraihat, "A smart infrared microcontroller-based blind guidance system," *Active and Passive Electronic Components*, vol. 2013, 7 pages, 2013.

[7] T. Miura, Y. Ebihara, M. Sakajiri, and T. Ifukube, "Utilization of auditory perceptions of sounds and silent objects for orientation and mobility by visually-impaired people," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1080–1087, Anchorage, AK, 2011.

[8] H. Marion and J. Michael, *Assistive Technology for Visually Impaired and Blind People*, Springer Science & Business Media, 2010.

[9] V. Tiponut, D. Ianchis, Z. Haraszy, and I. Bogdanov, "Work directions and new results in electronic travel aids for blind and visually impaired people," *WSEAS Transactions on Systems*, vol. 9, pp. 1086–1097, 2010.

[10] V. Tiponut, S. Popescu, I. Bogdanov, and C. Caleanu, "Obstacles detection system for visually impaired guidance," in *12th WSEAS International Conference on SYSTEMS*, Heraklion, Greece, 2008.

[11] I. Ulrich and J. Borenstein, "Local obstacle avoidance with look-ahead verification," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, vol. 3, pp. 2505–2511, San Francisco, CA, USA, 2000.

[12] Y. Seki and T. Sato, "A training system of orientation and mobility for blind people using acoustic virtual reality," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 19, no. 1, pp. 95–104, 2011.

[13] E. S. Narayanan, D. D. Gokul, B. P. Nithin, and P. Vidhyasagar, "IoT based smart walking cane for typhlotic with voice assistance," in *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*, pp. 1–6, Coimbatore, 2016.

[14] D. Yuan and R. Manduchi, "A tool for range sensing and environment discovery for the blind," in *2004 Conference on computer vision and pattern recognition workshop*, pp. 39–39, Washington, DC, USA, 2004.

[15] D. Yuan and R. Manduchi, "Dynamic environment exploration using a virtual white cane," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 243–249, San Diego, CA, USA, 2005.

[16] P. Baranski, M. Polanczyk, and P. Strumillo, "A remote guidance system for the blind," in *The 12th IEEE International*

*Conference on e-Health Networking, Applications and Services*, pp. 386–390, Lyon, 2010.

[17] T. Miura, K. Ueda, S. Ino, T. Muraoka, and T. Ifukube, "Object's width and distance distinguished by the blind using auditory sense while they are walking," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3859, 2008.

[18] R. Boldu, D. J. C. Matthies, H. Zhang, and S. Nanayakkara, "AiSee: an assistive wearable device to support visually impaired grocery shoppers," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, pp. 1–25, 2020.

[19] A. A. Tahat, "A wireless ranging system for the blind long-cane utilizing a smart-phone," in *2009 10th International Conference on Telecommunications*, pp. 111–117, Zagreb, 2009.

[20] D. Bolgiano and E. Meeks, "A laser cane for the blind," *IEEE Journal of Quantum Electronics*, vol. 3, no. 6, pp. 268–268, 1967.

[21] S. Shoval, I. Ulrich, and J. Borenstein, "Robotics-based obstacle-avoidance systems for the blind and visually impaired - Navbelt and the guidecane," *IEEE Robotics and Automation Magazine*, vol. 10, no. 1, pp. 9–20, 2003.

[22] L. Shangguan, Z. Yang, A. X. Liu, Z. Zhou, and Y. Liu, "STPP: spatial-temporal phase profiling-based method for relative RFID tag localization," *IEEE/ACM Transactions on Networking*, vol. 25, no. 1, pp. 596–609, 2017.

[23] A. Raina and D. Bansal, "Poster: smart-phones as active sensing platform for road safety solutions," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services Companion*, pp. 74–74, Singapore, 2016.

[24] A. Zvironas, M. Gudauskis, and D. Plikynas, "Indoor electronic traveling aids for visually impaired: systemic review," in *2019 International conference on computational science and computational intelligence (CSCI)*, pp. 936–942, Las Vegas, NV, USA, 2019.

[25] N. A. Giudice and G. E. Legge, "Blind navigation and the role of technology," in *Engineering Handbook of Smart Technology for Aging, Disability, and Independence*pp. 479–500, John Wiley & Sons.

[26] F. Xiao, Q. Miao, X. Xie, L. Sun, and R. Wang, "Indoor anti-collision alarm system based on wearable Internet of things for smart healthcare," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 53–59, 2018.

[27] M. F. Saaid, I. Ismail, and M. Z. H. Noor, "Radio frequency identification walking stick (RFIWS): a device for the blind," in *2009 5th International Colloquium on Signal Processing & Its Applications*, pp. 250–253, Kuala Lumpur, 2009.

[28] B. Andò, S. Baglio, V. Marletta, and A. Valastro, "A haptic solution to assist visually impaired in mobility tasks," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 5, pp. 641–646, 2015.

[29] S. S. Reddy, G. Sathyaprabha, and P. V. V. P. Rao, "Voice based guidance and location indications system for the blind using GSM," *GPS and Optical Device Indicator*, vol. 3, pp. 822–832, 2014.

[30] E. Milios, B. Kapralos, A. Kopinska, and S. Stergiopoulos, "Sonification of range information for 3-D space perception," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 4, pp. 416–421, 2003.

[31] X. Xing, R. Zhou, and L. Yang, "The current status of development of pedestrian autonomous navigation technology," in *2019 26th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS)*, pp. 1–3, Saint Petersburg, Russia, 2019.

[32] P. B. L. Meijer, "An experimental system for auditory image representations," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 2, pp. 112–121, 1992.

[33] M. Olfati, W. Yuan, A. Khan, and S. H. Nasseri, "A new approach to solve fuzzy data envelopment analysis model based on uncertainty," *IEEE Access*, vol. 8, pp. 167300–167307, 2020.

[34] M. Ahmad, A. Khan, A. Khan et al., "Spatial prior fuzziness pool-based interactive classification of hyperspectral images," *Remote Sensing*, vol. 11, no. 9, article 1136, 2019.

[35] R. Chen, Z. Tian, H. Liu, F. Zhao, S. Zhang, and H. Liu, "Construction of a voice driven life assistant system for visually impaired people," in *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 87–92, Chengdu, 2018.

[36] Bay Advanced Technologies, "BAT "K" sonar: ultrasonic sensing device for the blind," 2012, http://www.batforblind.co.nz.

[37] A. Garg, M. Balakrishnan, K. Paul et al., "Cane mounted knee-above obstacle detection and warning system for the visually impaired," in *Proceedings of the 11th International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pp. 143–151, New York, NY, 2007.

[38] D. T. Hartong, F. F. Jorritsma, J. J. Neve, B. J. Melis-Dankers, and A. C. Kooijman, "Improved mobility and independence of night-blind people using night-vision goggles," *Investigative Ophthalmology & Visual Science*, vol. 45, no. 6, pp. 1725–1731, 2004.

[39] A. Ullah, M. Shaheen, A. Khan, M. Khan, and K. Iqbal, "Evaluation of topology-dependent growth rate equations of three-dimensional grains using realistic microstructure simulations," *Materials Research Express*, vol. 6, no. 2, article 026523, 2018.

[40] I. Ulrich and I. Borenstein, "Correspondence the guide cane — applying mobile robot technologies to assist the visually impaired," *IEEE Transactions on Systems Man and Cybernetics-Part A Systems and Humans*, vol. 31, no. 2, pp. 131–136, 2001.

[41] M. Nie, J. Ren, Z. Li et al., "SoundView: an auditory guidance system based on environment understanding for the visually impaired people," in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine (EMBC'09)*, pp. 7240–7243, Minneapolis, MN, USA, September 2009.

[42] K. W. Lin, T. K. Lau, C. M. Cheuk, and Y. Liu, "A wearable stereo vision system for visually impaired," in *2012 IEEE International Conference on Mechatronics and Automation*, pp. 1423–1428, Chengdu, 2012.

[43] L. Dunai, D. Ismael, L. Lengua, I. Tortajada, and S. Brusola, "Obstacle detectors for visually impaired people," in *2014 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM)*, Bran, Romania, 2014.

[44] M. A. Rahman, M. S. Sadi, M. M. Islam, and P. Saha, "Design and development of navigation guide for visually impaired people," in *2019 IEEE international conference on biomedical engineering, computer and information Technology for Health (BECITHCON)*, pp. 89–92, Dhaka, Bangladesh, 2019.

[45] A. Riazi, F. Riazi, R. Yoosfi, and F. Bahmeei, "Outdoor difficulties experienced by a group of visually impaired Iranian people," *Journal of current ophthalmology*, vol. 28, no. 2, pp. 85–90, 2016.

[46] A. Kumar, R. Patra, M. Manjunatha, J. Mukhopadhyay, and A. K. Majumdar, "An electronic travel aid for navigation of visually impaired persons," in *2011 Third International Conference on Communication Systems and Networks (COMSNETS 2011)*, pp. 1–5, Bangalore, 2011.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8626890, 17 pages
https://doi.org/10.1155/2021/8626890

*Research Article*

# Research on Campus Education Information System Based on Internet of Things and Artificial Intelligence Decision

**Hetiao Hong**

*Hangzhou Normal University, Hangzhou Zhejiang 311121, China*

Correspondence should be addressed to Hetiao Hong; hht@hznu.edu.cn

Received 27 July 2021; Revised 9 September 2021; Accepted 15 September 2021; Published 4 October 2021

Academic Editor: Rajesh Kaluri



Because of the different reasons between regions, the distribution of educational resources is also different, the development of each school is unbalanced, and the degree of campus education informationization is different. The complex functional structure not only does not facilitate teachers and students but also leads to many problems: the prevention and prevention of campus life safety. It is difficult to keep and use multiple cards owned by one person. Software and education platform cannot be seamlessly connected, and there are various barriers between data and data and people and data. The lack of learning materials leads to the inequality of information. There are no good feedback and solution between teachers and students. It is difficult to manage accurately with a large number of people. This study will be based on the Internet and artificial intelligence technology, to explore how to study a large (or super large), concise and efficient, and excellent performance of campus education information system; this system can meet the teachers and students no matter what year, month, and day of a large number of visits. For some problems in the process of building the system, actively optimize and refine them. After functional testing and analysis of the system, the experimental results show that the interface function of the new system is stable, the usability test is better than the feedback experience of the original system, the response time is reduced by 21.6% on average, and the overall power consumption of the system is reduced by about 1.43% on average.

## 1. Introduction

With the rapid development of the times and the impact of new technologies, all walks of life have changed their faces, and the education industry is no exception. Especially in the past two years, the global epidemic has seriously affected the academic process. Primary schools, junior high schools, senior high schools, and higher education institutions should build a platform integrating learning, living, monitoring, safety, and other functions for teachers and students to use and facilitate command and arrangement in case of emergencies. The Internet of Things and artificial intelligence technology are introduced into the campus, integrated with educational information, and a variety of rich application scenarios are designed and applied to meet the daily study and life of teachers and students, accelerate the completion of the modern process of educational informationization, keep up with the pace of the times, and build an intelligent campus. The literature [1] realizes the virtual scene simulation and multimedia information interaction design of multimedia information platform, intelligent classroom, library interactive platform, and student information management system on intelligent multimedia information processing terminal. The literature [2] discusses the introduction of Internet of Things technology into intelligent management and the combination of management concepts and methods with artificial intelligence technology, which has an impact on life. On the basis of analyzing the current situation of medical device management, the literature [3] incorporates the innovative achievements of information technology such as electronic signature technology, big data analysis technology, Internet of Things technology, artificial intelligence, and cloud services into the daily work of medical device management. The literature [4] introduces the domestic artificial intelligence popularization education resources and the exploratory work of artificial intelligence technology popularization on this platform. The literature [5] systematically introduces the research status of industry

standard system related to Internet of Things and introduces the general standard classification method, hierarchical architecture, conceptual model, and system table in UPIoT application layer. The literature [6], based on campus security management, is aimed at establishing an intelligent mobile information system in colleges and universities, which adopts Windows media player along RTP/RTSP protocol. The literature [7] analyzes the application of Internet of Things in the construction of intelligent campus in colleges and universities in detail. The literature [8] compares the campus information systems of three British universities and obtains the functions that should be possessed on an ideal CWIS. The literature [9] introduces the application methods and significance of big data and cloud computing technology in intelligent campus. The literature [10] introduces an intelligent sensor network and applies the technology of intelligent sensor network to the research of campus environmental management information system model. The literature [11] introduces the architecture design of big data platform of BI application system. This building has outstanding advantages in cost, scalability, and scalability, and the realization of BI analysis function plays an important role in the strategic decision-making of colleges and universities. The literature [12] eliminates the "isolated island" in university information system and provides effective data basis for teachers and students' work, study, and daily life. The literature [13] proposes a new algorithm and user support system based on users' personal data and multiattribute preferences and proposes an effective lecture allocation method. The document [14] was based on VRML and JavaScript, the client programming, design, and implementation of online virtual campus system literature. The literature [15] takes the design of virtual campus card system under microservice architecture as the goal and designs the virtual campus card system. As an important educational link in cultivating innovative talents, under the background of the development of Internet and artificial intelligence, many colleges and universities are constantly applying new technological achievements, which must be reformed to meet the requirements of cultivating innovative talents facing the information society, establish and own their own campus information systems, change the basic environment of traditional education and teaching, and improve the overall level and application of education and teaching. The development of information society also provides environment and conditions for this reform; under the background of Internet and artificial intelligence, the construction of educational informationization has played a role in promoting the teaching, scientific research, and student training in colleges and universities, which will effectively promote the informationization and modernization of higher education, which is also the requirement and inevitable result of educational informationization. The method proposed in this paper runs well in the system, but in the whole campus system, the system expands more. In this way, the performance of the whole system will decline, and the adoption of intelligent systems requires better equipment requirements and network environment. The key research work in the future needs to improve the system response time and server throughput, especially in the context of wireless networks, using multiprotocol mechanism. When different network structures are accessed, the system is continuously accessed to improve the application effect of the whole system.

## 2. Theoretical Basis of System Development

*2.1. Internet of Things Technology.* Internet of Things [16, 17] is a new technology in the 21st century, which has a good development trend in China and has gradually become a very important technology in the information age. It requires sensors to collect and process information, and at the same time, it exchanges information through the network, and the real and virtual worlds have completed a great leap. The Internet of Things involves a wide range and has many application levels.

*2.2. AI Artificial Intelligence Technology.* Artificial intelligence [18], also known as AI, has never stopped the research on intelligent machines, hoping to create advanced machines similar to human intelligence and ability and even put forward the theory of what should happen when intelligent robots surpass human development. In addition, people have created a series of books and audio-visual works related to artificial intelligence to tell their expectations and worries about artificial intelligence. But so far, the development of artificial intelligence is not ahead of schedule, there are many defects and drawbacks, and the core technology needs further research.

*2.3. Network Upgrade Technology.* So far, 4G network basically covers the whole campus. However, due to the limitation of 4G technology, with the passage of time, the network gradually becomes stuck, and sometimes, even the network is disconnected. Therefore, it is necessary to update the network. In this network upgrade, the latest 5G network technology will be used to upgrade the bandwidth to 5G, and the specificity of the campus network will be strengthened to prevent unknown access and attacks from outside, and various authentication methods [19] will be adopted to facilitate integration with the new system.

*2.4. ETL Technology.* In the ETL technology, which is mature, we can understand it as the process of building a building. The actual application process is shown in Figure 1.

ETL technology, that is, data warehouse technology, extracts, cleans, transforms, and loads all kinds of data into data warehouse. We can also use a metaphor to understand it as the process of building a building. Figure 1 shows the practical application process, in which the functions of ETL are data extraction, cleaning and transformation, data summary, data federation, data synchronization, data migration, and data distribution.

*2.5. Information Security Technology.* Information security is a very important issue. Criminals often attack computer software or hardware in various ways all over the world: stealing network information, causing serious troubles to countless people, spam calls and spam sale information, human flesh search, personal pictures and audio leakage,

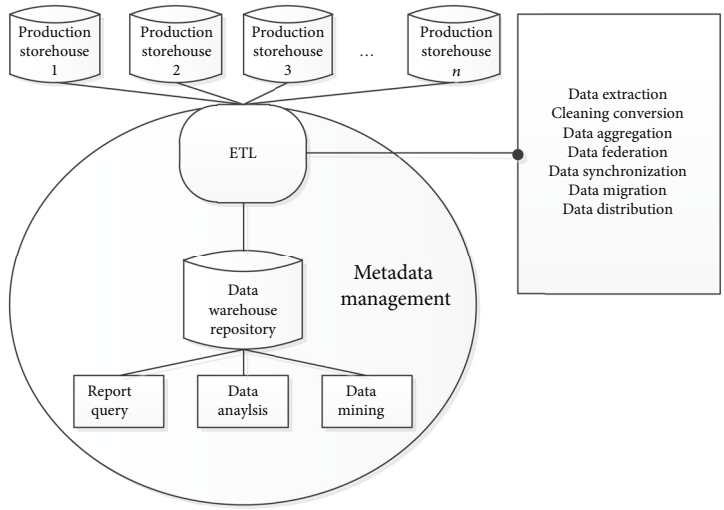



Figure 1: Practical application process of ETL.

Trojan virus, fraudulent links, and gambling propaganda. Our system needs to prevent malicious attacks, tampering, destruction, and rejection of individuals or computers that have not been accessed through formal channels, so as to ensure the confidentiality and security of information. In order to ensure information security, the system needs technical personnel to maintain and supervise the system frequently and often make up for system loopholes. For this system, we will mainly use five major information security technologies: firewall technology, information encryption technology, identity authentication technology, security protocol, and intrusion detection system. Figure 2 shows the classification diagram of information security technology.

As shown in Figure 3, the firewall operation diagram.

*2.6. Apriori Association Algorithm.* The Apriori algorithm uses a progressive search method to find the relation formation conditions in the database, which is composed of connection and interruption processes. We set the minimum support ($s$) and the minimum confidence ($C$).

(1) Find all frequent itemsets, and the frequency is greater than or equal to the minimum support ($s$)

(2) All frequent itemsets can be found by its frequent itemset

(3) The association rules should satisfy the minimum support ($s$) and the minimum confidence ($C$)

According to the Apriori algorithm, the purpose of campus education information system is to dig out hidden information that is usually ignored by people and find that teaching level and service need to be improved and improved from a new perspective. In Figure 4, we discuss the time consumption of minimum support. The greater the minimum support, the less time consumption. In Figure 5, with other contents fixed, the larger the amount of data found, the less time it takes.

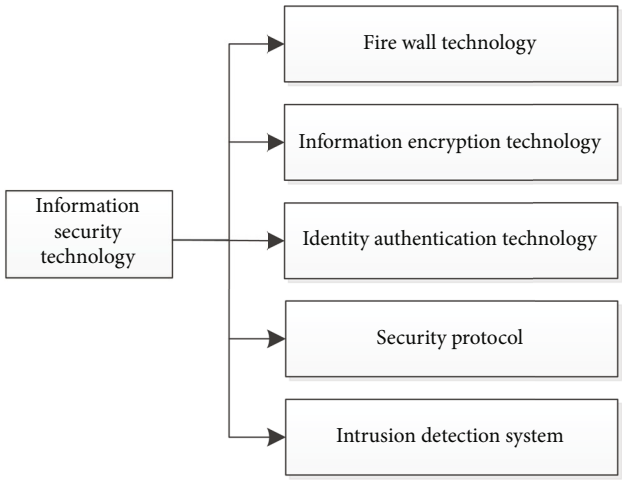



Figure 2: Classification of information security technology.

(1) Support:

$$(A \longrightarrow B) = P(A \cup B), \tag{1}$$

$$S(X, Y) = P(X, Y) = \frac{\text{num(xy)}}{\text{num(all samples)}}. \tag{2}$$

(2) Confidence:

$$(A \longrightarrow B) = P(B|A), \tag{3}$$

$$C(xY) = P(x|Y) = \frac{P(xy)}{P(y)}. \tag{4}$$

*2.7. Decision Tree Algorithm.* The decision tree algorithm classifies information through a series of rules.

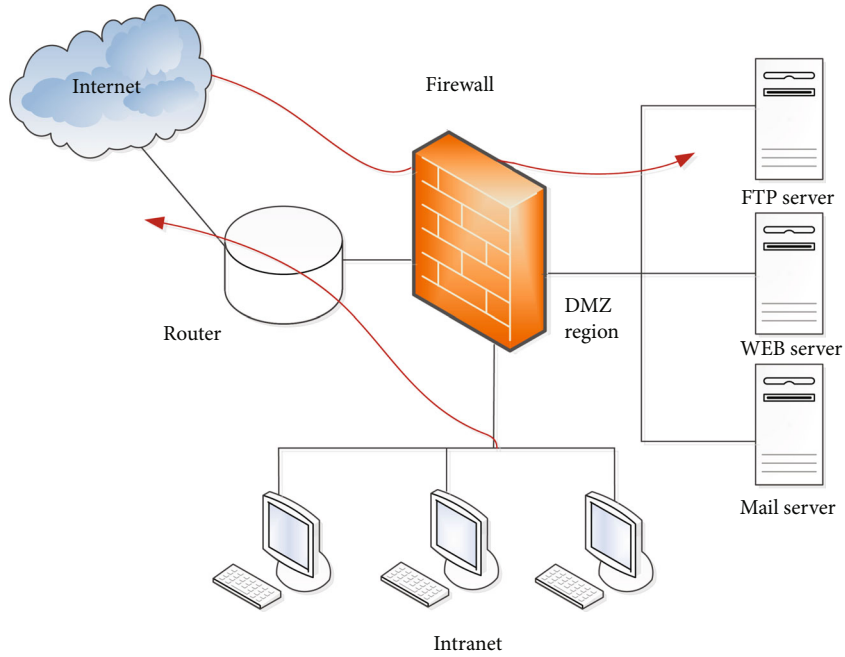



Figure 3: Schematic diagram of firewall operation.

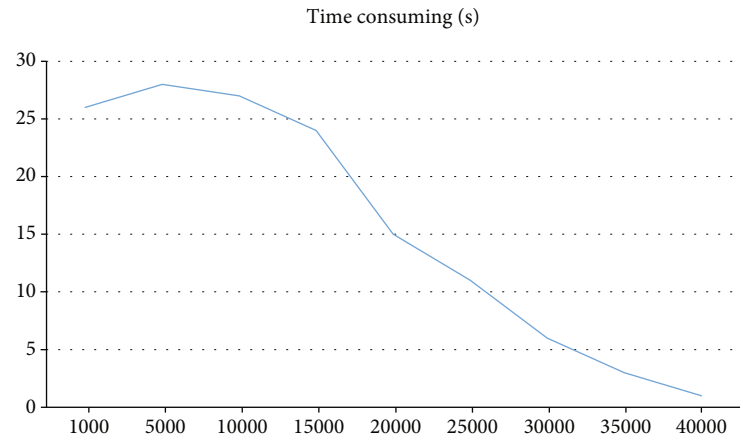



Figure 4: Time consumption of minimum support ($s$).

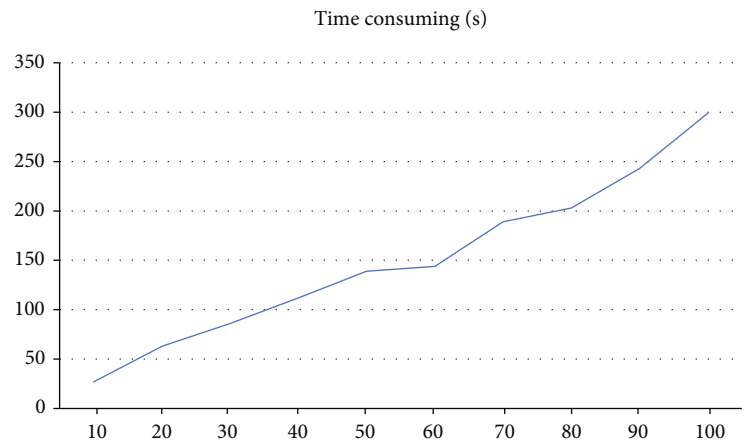



Figure 5: Time consumption of data volume ($d$).

*2.7.1. ID3 Algorithm*

(i) Information entropy:

$$P(x = x_i) = P_i, \quad i = 1, 2, \cdots, n,$$
$$H(x) = -\sum_{i=1}^{n} \log_2 p_i. \tag{5}$$

(ii) Conditional entropy:

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y \mid X = x_i), \quad p_i = P(x = x_i). \tag{6}$$

(iii) Information gain:

$$g(D, A) = H(D) - H(D|A). \tag{7}$$

*2.7.2. C4.5 Decision Tree*

(i) C4.5 which is an optimization decision tree defined as follows:

$$\$\$\text{SplitInfo}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * \log_2\left(\frac{|D_j|}{|D|}\right)\$\$. \tag{8}$$

(ii) Definition of information gain rate:

$$\$\$\text{GainRatio}(A) = \frac{G_{a_i n}(A)}{\text{Spl}\#\text{tInf}_0(a)}\$\$. \tag{9}$$

*2.7.3. CART Decision Tree.* Carnegie Mellon University decision tree is divided into classification tree and regression tree.

(i) Gini index:

$$\$\$G(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_k' = 1 - \sum_{k=1}^{|y|} p_k^2\$\$. \tag{10}$$

*2.7.4. Continuous Value Processing*

(i) For continuous attribute $a$:

$$Ta = \left\{ \frac{a^{\#} + a^{\#+1}}{2} \middle| 1 \le i \le n-1 \right\}. \tag{11}$$

*2.7.5. CART Regression Tree*

(i) Predictive regression continuous data:

$$D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots (x_n, y_n)\}. \tag{12}$$

(ii) Selecting variable $j$ and segmentation point $s$:

$$\min_{j,8}\left[\min_{C_1}\sum_{x_i}(y_i - C_1)^2 + \min_{C_2}\sum_{x\varepsilon}(y_i - C_2)^2\right] \tag{13}$$

(iii) Selecting variable and segmentation point $(j, s)$:

$$R_1(j, 8) = \left\{x|x^{(j)} \le 8\right\}, R_2\left(\dot{J}, 8\right) = \left\{x|x^{(j)} > 8\right\}\hat{c}_m = \frac{1}{N_m}\sum_{x_i} y_i, \quad x \in R_m, m = 1, 2. \tag{14}$$

(iv) Dividing the input space into m regions to generate a decision tree:

$$f(x) = \sum_{m=1}^{M} \widehat{C}_m I(x \in R_m). \tag{15}$$

(v) Solving the optimal output value with square error:

$$\sum_{x_i \in R_m} (y_i - f(x_i))^2. \tag{16}$$

## 3. System Architecture Design

The system is designed for data management, offline application, and online service.

(1) Unified data management refers to the platform for summarizing various data on campus, which controls the flow and storage of data; most people do not have permission to view or use this part of data

(2) The offline application part of the system mainly refers to the use of infrastructure on campus, including campus security services and life convenience services, including subsystems such as payment

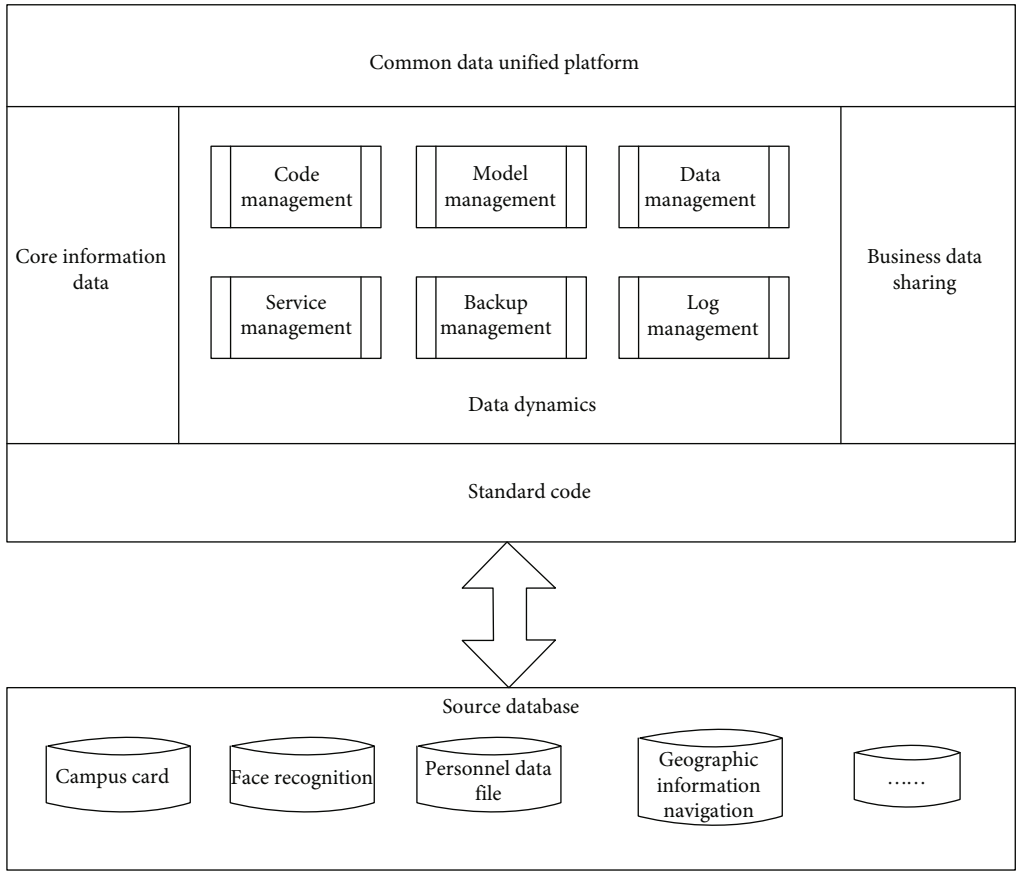



Figure 6: Architecture diagram of common data unified platform.

subsystem, face recognition subsystem for entering and leaving the door, classroom, and laboratory use subsystem

(3) Online services mainly refer to virtual campus services, which are mainly used through APP applications and WEB interface

Subsystem has an independent logical structure, but in the use of some subsystem functions, there will also be some other subsystem functions; this time we do not have to manually access to part of the subsystem permissions involved; the system will automatically judge and release the part of the data call and use of functions.

Offline application refers to the application of Internet of Things technology and the use of physical equipment on campus. Online service means that there is no entity at all, and all functions do not depend on physical devices on campus but are mainly used through APP applications and web interface. It can be distinguished according to whether physical equipment is needed.

### 3.1. Data Management

*3.1.1. Common Data Unified Platform.* The original intention of designing this unified public data platform is to process, clean, and integrate the shunted data of various subsystems and various basic original data, which need standardized and unified management, so as to effectively avoid the confusion, complexity and failure of data. In the common data unified platform, the data sources are processed and cleaned by the platform and finally diverted into the relevant information data of each subsystem after standardized and unified treatment. There is a total score relationship between data source and information data. It is also convenient to record and classify data when using it. Figure 6 shows the architecture diagram of the platform.

*3.1.2. Portrait of Student Behavior.* A campus has many students, so it is difficult to manage. Therefore, we should use certain tools and methods to collect and sort out students' abnormal data and analyze some abnormal behaviors and violations of students. The system cannot illegally collect students' personal information. Except for some technicians, ordinary people can only apply to view the analysis results. As shown in Figure 7, it is the related data analysis process.

As shown in Figure 8, it is the platform E-R diagram.

*3.1.3. Keeping of Personnel Data Files.* Data files record all aspects of a person. The preservation of archives is a very confidential job, which should be highly confidential. In general, except in exceptional circumstances, no one (especially the person himself) is allowed to view and read these materials. The system will automatically update the personnel file content and, if there is an error, will have a high authority to

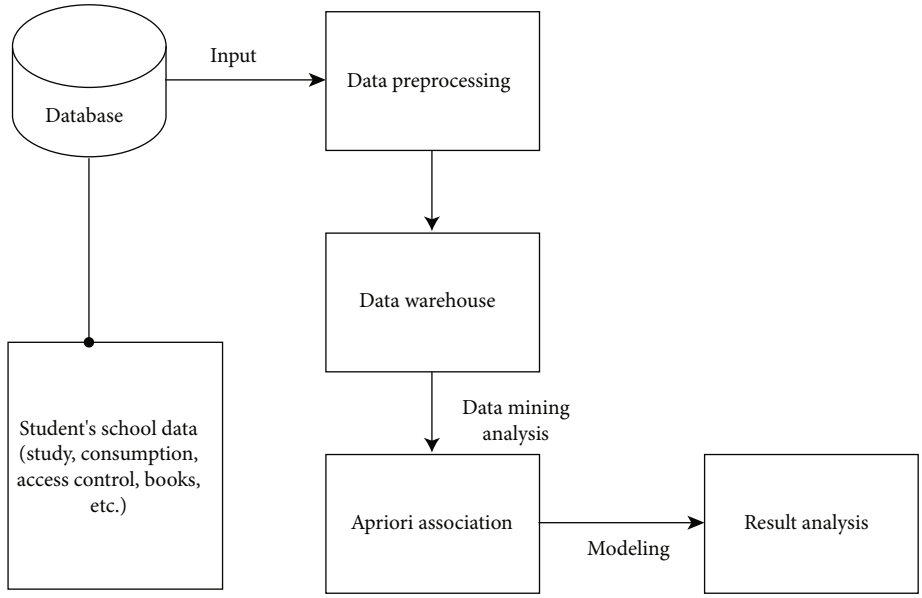



Figure 7: Student behavior portrait-data analysis process.

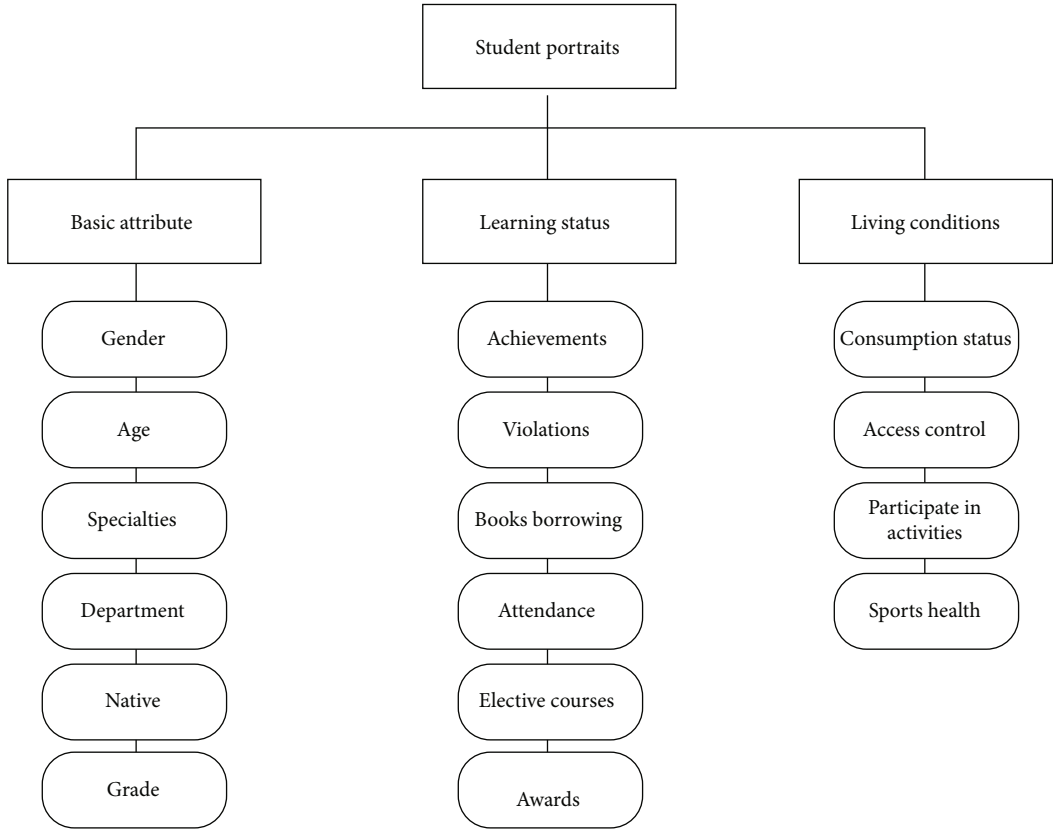



Figure 8: Platform E-R diagram.

deal with the personnel. Finally, this subsystem will make a safe backup to avoid any missing or missing files. As shown in Figure 9, it is the structure diagram of the subsystem.

*3.2. Campus Security Services.* Whether it is primary school, junior high school, high school, or university, campus safety is the key issue that has been emphasized again and again. In recent years, due to the development of network technology and the change of the times, data security and payment antifraud security have gradually become more and more important.

*3.2.1. Data Security.* It is far from enough to protect the security of data only by virtue of the security system of the

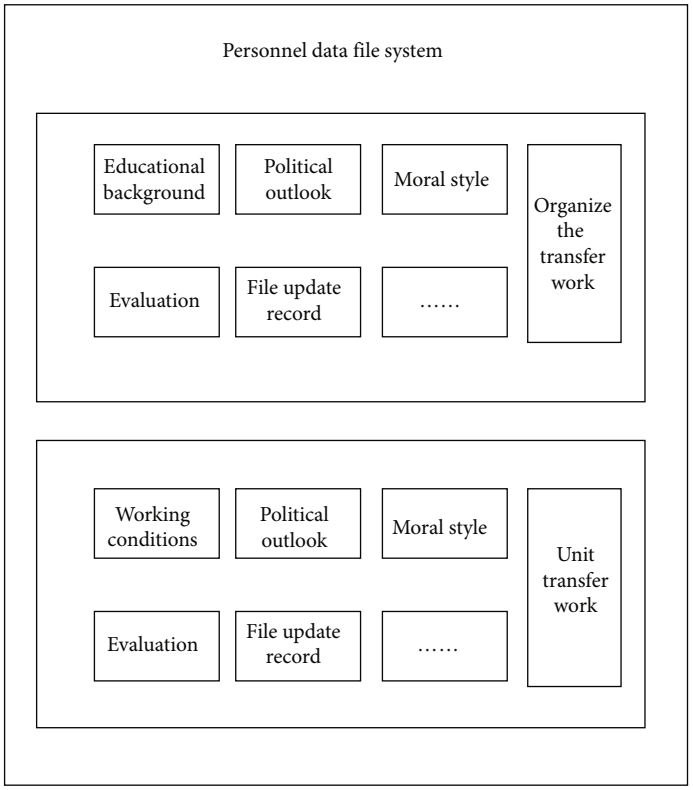



Figure 9: Personnel data file system.

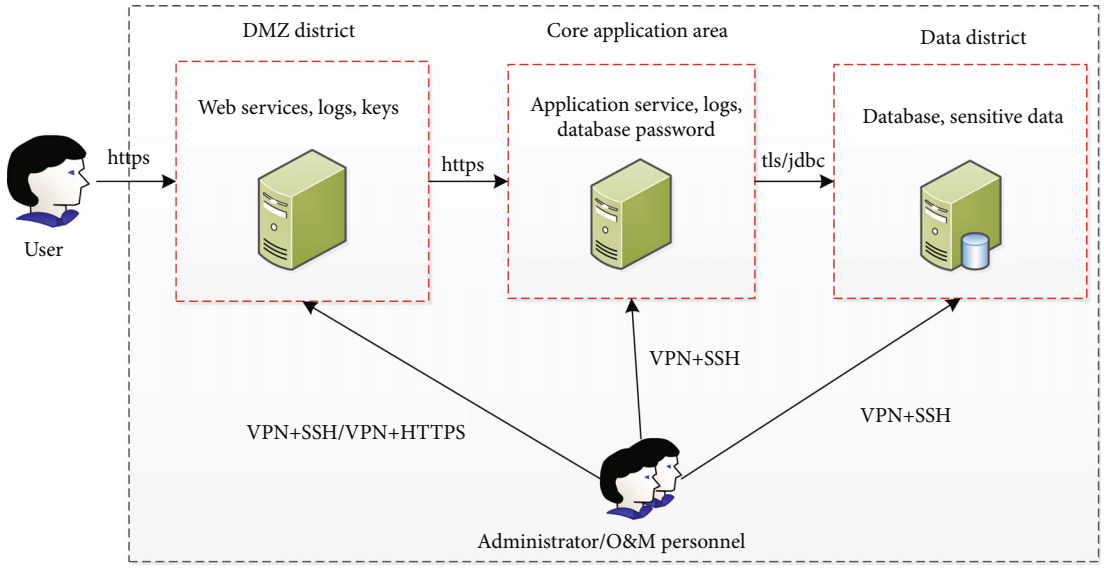



Figure 10: Security architecture diagram.

system itself, so we will specially design a security framework. Many criminals in the world make viruses (Trojan virus, Hack virus, Script virus, etc.) to steal information and data for illegal activities. The most terrible thing about viruses is that they infect computers through various channels and cause serious damage, which can lead to the destruction of data files and the paralysis of the network. We must focus on data security issues, maintain and update the system for a long time, make up for loopholes, and prevent the system from being attacked or crashing. The security architecture is shown in Figure 10.

*3.2.2. Payment Security.* The popularity of online shopping promotes the change of payment methods. Due to the development of information technology and communication industry in China, besides cash payment, credit card, or bank card payment, mobile phone scan code payment or two-dimensional code payment has become a mainstream

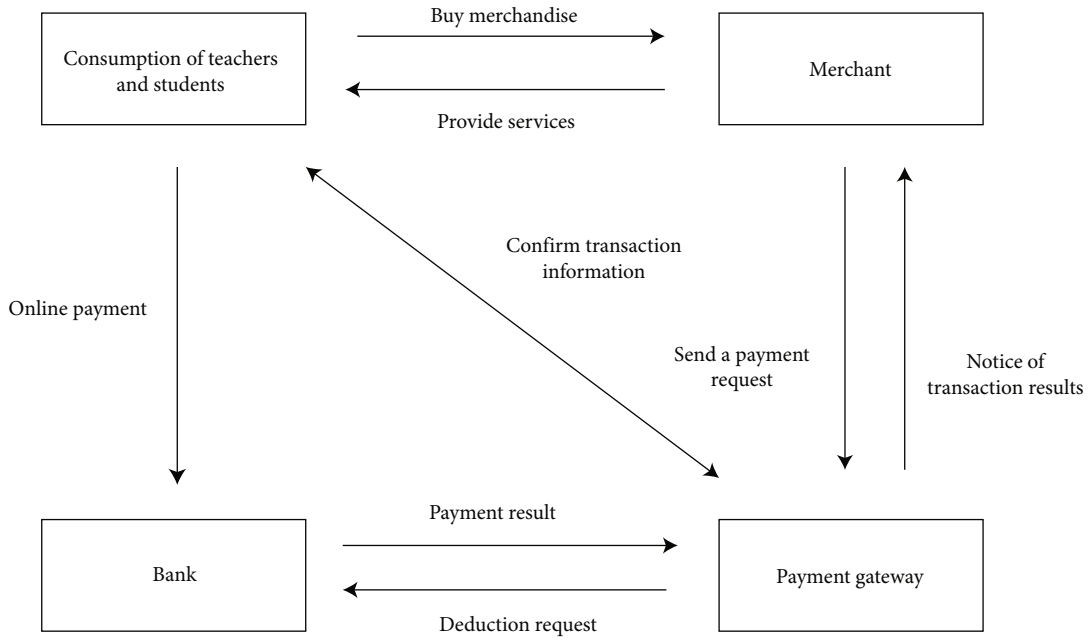



Figure 11: Electronic payment process.

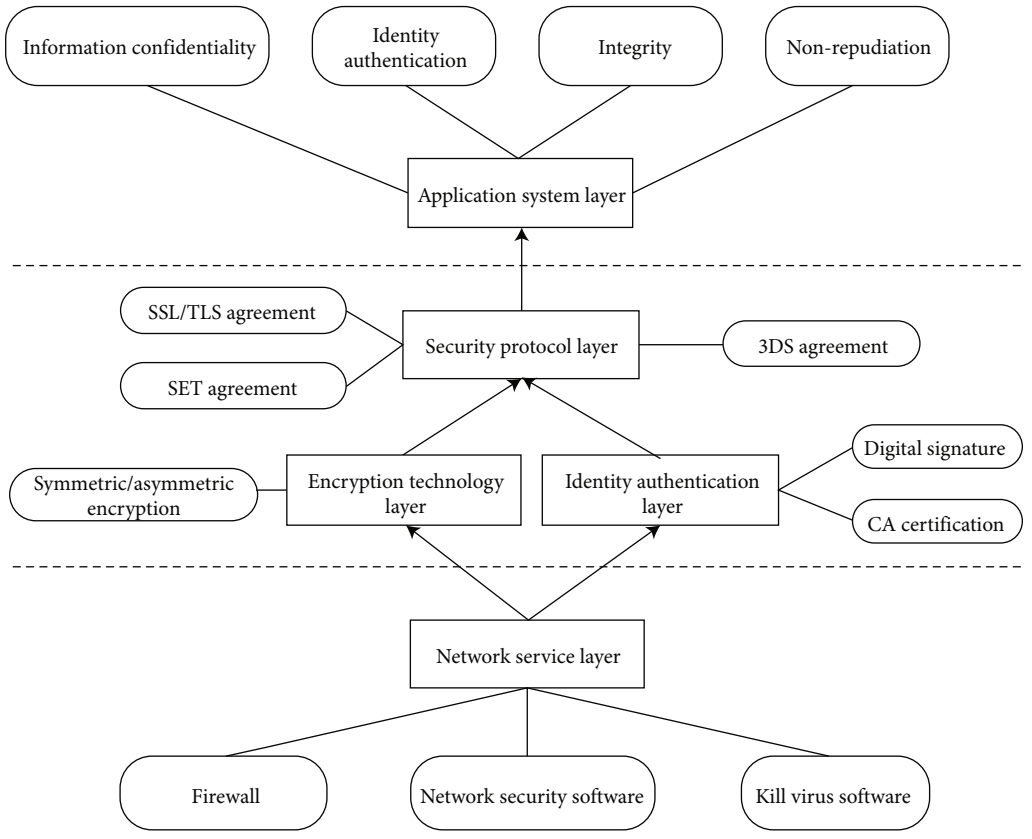



Figure 12: Electronic payment security technology.

trend. But the popularity of online payment also makes criminals eager to move. They have devised all kinds of network traps, which are difficult to capture telecom frauds and difficult to guard against malicious applications. Many people were cheated out of a lot of money because they did not pay attention for a while. Teachers and students on campus are also highly deceived places. As shown in Figure 11, it is an electronic payment flow chart.

As shown in Figure 12, it is a design drawing about electronic payment security technology.

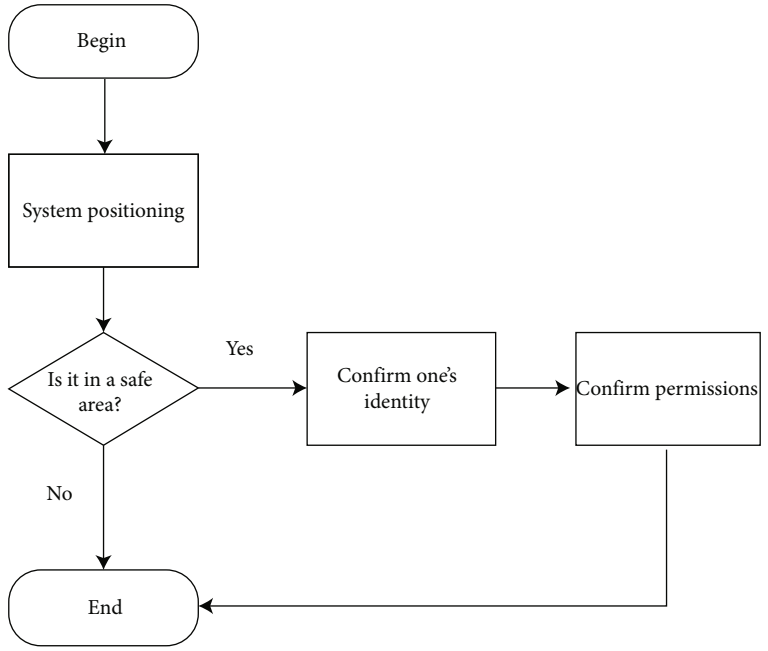



Figure 13: Positioning system.

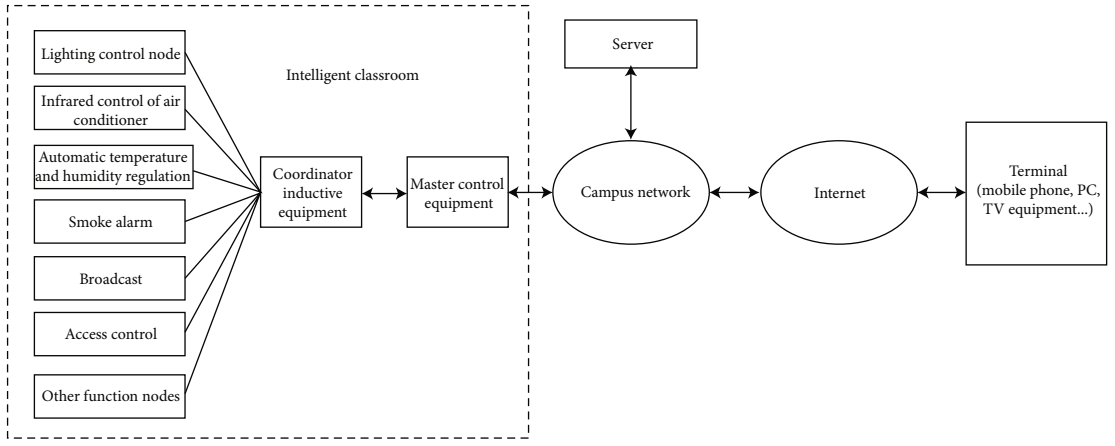



Figure 14: Functional structure diagram of intelligent classroom.

*3.2.3. Positioning System Usage.* In order to ensure the safety of school resources and the safety of teachers and students when they are mobilized, the system will use the school positioning system (generally placed on the virtual campus APP and various school infrastructures) to identify whether personnel are in the safe area, and only after confirming their identities and authorities can various resources be called to ensure personal safety. Figure 13 shows a flowchart of the positioning system.

### *3.3. Convenience Services*

*3.3.1. Intelligent Control of Classroom Equipment.* Old-fashioned classroom conditions are average, There are deficiencies in many aspects, such as air circulation, difficult debugging of projection equipment, too old-fashioned computers, and serious water and electricity consumption. After the classroom is updated, a new intelligent integrated classroom will be built, which can easily meet the requirements of teachers and students, and will be more conducive to teachers and students to carry out various teaching activities and even provide distance teaching services for students. The schematic diagram is shown in Figure 14.

*3.3.2. Intelligent Control of Laboratory Equipment.* Laboratory is a very important part in scientific research, teaching, and research. Experimental equipment and utensils should be more accurate and strict, and there should be a safety protection system to ensure the safety and integrity of experimental equipment. Special personnel should regularly check and maintain the cleanliness of the laboratory and check the laboratory situation. The structure is shown in Figure 15.

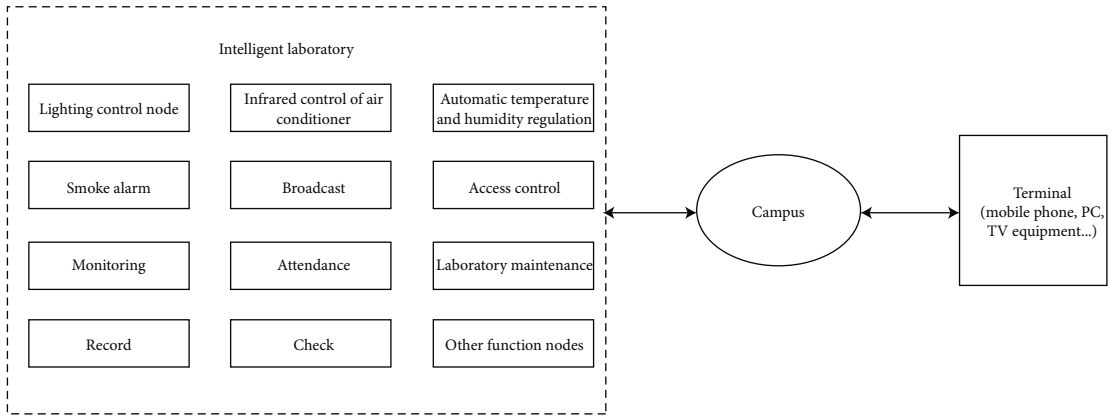



Figure 15: Functional structure diagram of intelligent laboratory.

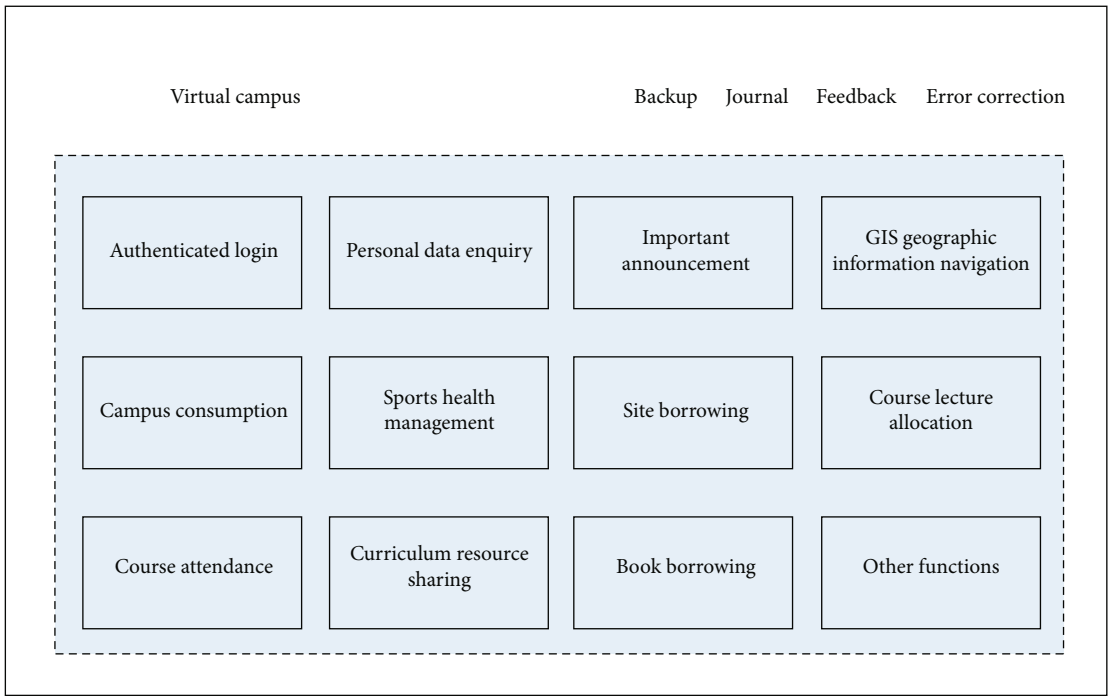



Figure 16: Virtual campus.

Table 1: Server requirements.

| Hardware platform type | PC server |
|---|---|
| CPU | 6G |
| Memory | 8G |
| Hard disk | 60G |
| Operating system | Windows or Linux |
| Database management system | Oracle |
| Description | According to the requirements of data exchange |

Table 2: Development environment.

| Category | Describe |
|---|---|
| Operating system | Win10 |
| IDE | 2021.1.1 |
| JAVA | JDK15 |
| Database | Oracle\MySQL |
| Application server | Tomcat8.0 |

*3.3.3. Application of Face Recognition.* The traditional identity authentication method, which is manually checked and approved or card-type, consumes manpower and material resources and is prone to flaws, which causes hidden dangers to campus security. Therefore, using artificial intelligence technology and Internet of Things technology, we introduce system equipment with face recognition application and place it in school gates, dormitories, laboratories, study rooms, libraries, and other access control places that need identity verification. Teachers and students only need to

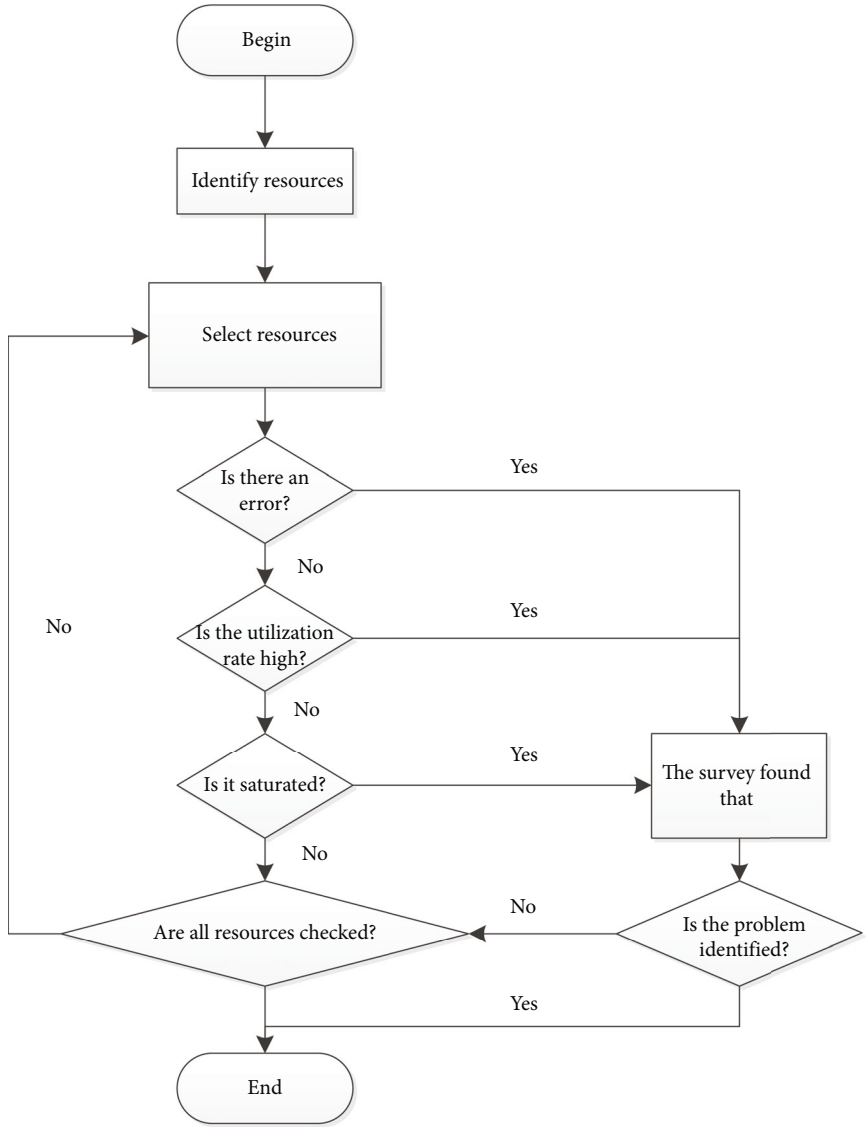



Figure 17: USE method.

Table 3: UI testing.

| Number | Test content | Test results |
|---|---|---|
| 1 | Home navigation position | Normal |
| 2 | Navigation bar content layout interface layout | Normal |
| 3 | Interface layout | Normal |
| 4 | Text display | Normal |
| 5 | Font size | Normal |
| 6 | Garbled code | None |
| 7 | Hyperlink | Normal |
| 8 | Color style | Accord with |
| 9 | Shortcut key | Normal |
| 10 | Options button | Normal |
| 11 | Text box, dialog box | Normal |
| 12 | Clarity | Clear |

complete the step of "brushing their faces," and they can use the corresponding venues and equipment resources according to their respective authorities. In this way, the inconvenience caused by numerous certificates is perfectly solved, and the work problems such as manpower, material resources, and time cost are greatly reduced. For outsiders, the system will give corresponding registration or alarm, which greatly improves the security of campus.

*3.3.4. Virtual Campus Services.* Virtual campus services include personal data inquiry, important announcement, GIS geographic information navigation, campus consumption, sports health management, venue borrowing, course lecture distribution, course attendance, course resource sharing, book borrowing, and other services. Teachers and students can enjoy convenient services only by logging in to APP or website on PC or mobile phone and authenticating their identities. The specific system interface of virtual campus service is shown in Figure 16.

Table 4: Compatibility testing.

| Number | Test case | Test results |
|---|---|---|
| 1 | It can be installed and enabled normally on different platforms, without card machine or flashback phenomenon. | Pass |
| 2 | Login, page browsing, search, comment, and other interfaces have no deformation, occlusion, uncoordinated size, and other problems and can be scaled and displayed at different resolutions. | Pass |
| 3 | Verify that interactive controls such as text boxes and keys in the interface can click and respond normally. | Pass |
| 4 | Verify that the controls in the interface can load network content, and the icons and texts of a single table item/list item have no distortion and occlusion. | Pass |
| 5 | Verify that the font and resolution in the interface are scaled to a certain extent. | Pass |

Table 5: Response time of original system.

| Concurrent request traffic | 100 | 1000 | 1900 | 2800 | 3700 | 4600 |
|---|---|---|---|---|---|---|
| Average response time per request (ms) of the original system | 25 | 100 | 168 | 228 | 307 | 411 |

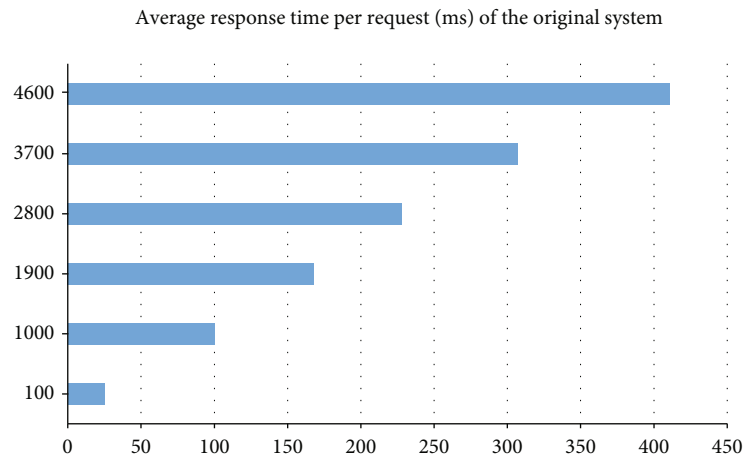



Figure 18: Response time.

## 4. System Optimization Research

The performance optimization of the system is not simply to detect one by one or to wait for feedback from users when they find problems. We should first automatically monitor performance problems and analyze them according to data. This is a long-term work, which requires continuous optimization research.

*4.1. Caching.* Caching technology can help us to develop high-performance and high-availability applications, save resources, and achieve optimization. We will cache data and computation results, as well as domain name resolution and resource objects themselves, which have been accessed, and they are likely to be used again in the future. This can reduce the waste or consumption of system resources and network resources to a certain extent.

*4.2. Inertia.* In order to save system resources, the system delays the calculation to some extent. When necessary, eliminate those parts that do not need to be calculated, and carry out the most precise calculation. This mode of operation can make the system have a "gap" to rest. Instead of running at high speed for a long time, it will cause irreversible damage to the life and performance of the system and cause a lot of troubles to maintenance personnel.

*4.3. Code Quality.* Excellent code is often efficient, usually with comments and code; code efficiency is very good, saving a lot of unnecessary work and trouble. When doing some optimization work, we should carefully modify our code unless we have to, because a little change can cause serious impact. The quality of the code and the logic of the code are important. Therefore, it is proposed that we should not optimize too early or overoptimize.

## 5. System Function Test and Analysis

Using professional testing system tools to test the campus education information system, according to the experimental results, the experimental analysis is given to determine the future improvement scheme and optimize the performance and function of the detailed system.

Table 6: Response time of new system.

| Concurrent request traffic | 100 | 1000 | 1900 | 2800 | 3700 | 4600 |
|---|---|---|---|---|---|---|
| Average response time per request for new system (ms) | 10 | 57 | 123 | 196 | 249 | 336 |

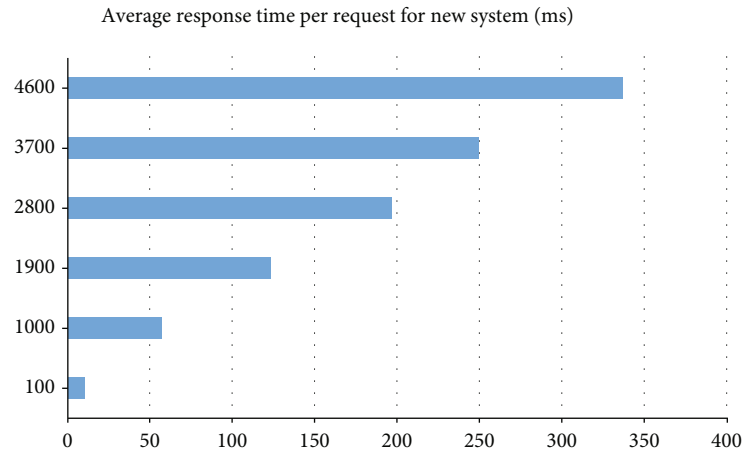



Figure 19: Response time.

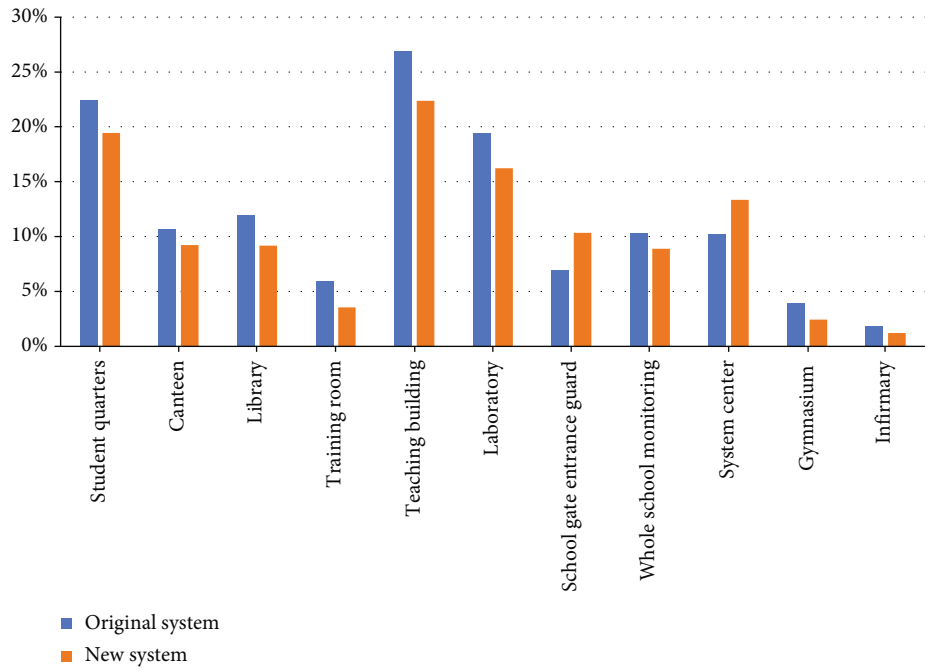



Figure 20: Comparison of regional power consumption between the original system and the new system.

*5.1. Running Environment.* The environment in which the system runs is shown in Tables 1 and 2.

*5.2. USE Method for Performance Analysis.* Check the utilization, saturation, and errors of all resources, as shown in Figure 17.

*5.3. UI Testing.* Through the Google Chrome browser manual operation identification detection, the main test platform operation is interface rationality. The test results are shown in Table 3.

*5.4. Compatibility Testing.* Test the compatibility of the system with software and hardware, as shown in Table 4.

*5.5. System Performance Comparison.* We will simulate two systems, one is the system before the transformation, and the other is the newly designed campus education information system. Because it is a simulation system, we consider many factors, such as cost, time, and experimental conditions. According to past experience, we reduce the system to an equal scale, which is convenient for the experiment.

Table 7: Data of the original system and the new system.

| Test point | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 | T11 | T12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Original system | 4 | 5 | 5 | 3 | 6 | 3 | 2 | 5 | 1 | 5 | 6 | 4 |
| New system | 5 | 6 | 6 | 5 | 5 | 7 | 5 | 5 | 7 | 4 | 5 | 6 |

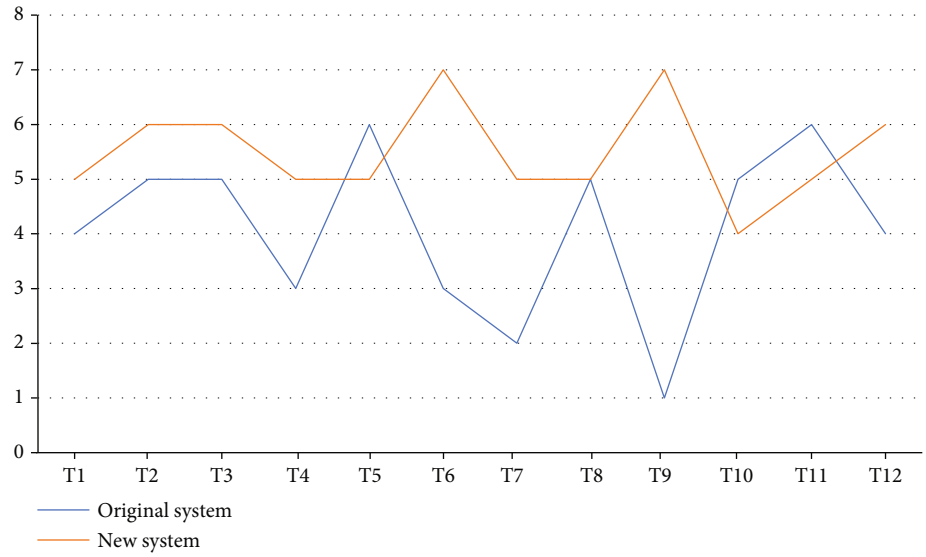



Figure 21: Data comparison diagram.

*5.5.1. System Response Time.* Response time refers to the time taken to execute a request or task. Response time means whether the performance of the system meets the requirements and whether it can bring smooth use effect. We set up six groups with visits of 100, 1000, 1900, 2800, 3700, and 4600.

(1) The data of the original system are shown in Table 5.

The average response time per request of the original system is shown in Figure 18.

(2) Data for the new system are shown in Table 6.

The average response time per request for the new system is shown in Figure 19.

(3) The following is the analysis of experimental results:

Comparing the data of the original system with the new system, it is not difficult to find that after adopting the Internet of Things and artificial intelligence technology, the average response time of our system for each request is obviously reduced, which is 21.6% less than that of the original system, and the working efficiency of the system is greatly improved.

*5.5.2. Area System Power Consumption.* In response to the national call to save resources, the power consumed by the system is also an important index to test the system performance.

(1) The regional data of the original system and the new system are shown in Figure 20.

(2) The following is the analysis of experimental results:

We can clearly see from the figure that most areas of the new system consume less power than the original system, and the overall power consumption decreases by 1.43% on average. Only the school access control and system center have different increases in power consumption.

*5.5.3. Usability Testing.* There are 12 test points in this part of the test, which test whether the software, hardware, and network conditions meet the requirements and the users' evaluation feelings. The test part is the user's feeling after using the system for a long time (the time is set to 30 days): 1 stands for very bad, 2 stands for very bad, 3 stands for bad, 4 stands for average, 5 stands for good, 6 stands for excellent, and 7 stands for perfect.

(1) Data comparison between the original system and the new system is shown in Table 7.

The comparison diagram between the original system and the new system is shown in Figure 21.

(2) The following is the analysis of the test results:

As can be seen from the chart, users' evaluation of the original system is very bad, showing extreme and fluctuating values. However, the evaluation of the new system is stable,

Table 8: Effect of minimum support ($s$) on performance.

| Minsup | 1000 | 5000 | 10000 | 15000 | 20000 | 25000 | 30000 | 35000 | 40000 |
|---|---|---|---|---|---|---|---|---|---|
| Time consumed (s) | 26 | 28 | 27 | 24 | 15 | 11 | 6 | 3 | 1 |

Table 9: Impact of data volume ($d$) on performance.

| Num/10000 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Time consumed (s) | 27 | 63 | 86 | 112 | 139 | 144 | 189 | 203 | 243 | 299 |

the use effect is very good, and the fluctuation value is very small.

*5.6. Apriori Algorithm Analysis.* The experimental setting data ($D$) is 100000 pieces, and the minimum confidence ($C$) is fixed at 0.75 without changing.

(1) Fixed data volume ($D$) changes minimum support ($s$):

The data are shown in Table 8.

The time consumption of minimum support is shown in Figure 4.

(2) Fixed minimum support ($s$) changes the amount of data ($D$):

The data are shown in Table 9.

The amount of data consumes time, as shown in Figure 5.

(3) The following is the analysis of experimental results:

As shown in the figure, the more data is, the more time it takes. When the amount of data is constant, the time decreases gradually with the increase of the minimum support.

## 6. Conclusion

To sum up, the algorithms and technologies used in this research have obvious progress and superiority. After the network upgrade, compared with the old and backward systems used in previous schools, the current new system has stronger update iteration, stronger intelligent optimization, and stronger user experience. However, it should be noted that our testing of the system is only in the initial stage. If we want to put it into the market formally, there are still more details to be improved and optimized in the process of putting it into use. With the passage of time and the update of technology, there is still many work that needs to be adjusted to be studied by later generations.

## Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The author declares that there are no conflicts of interest regarding this work.

## References

[1] X. U. Qing, "Design of intelligent campus multimedia interactive system based on internet of things technology," in *Proceedings of 2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS 2019) Volume I.Ed*, pp. 237–240, Conference Publishing Services, 2019.

[2] X. Lv and M. Li, "Application and research of the intelligent management system based on internet of things technology in the era of big data," *Mobile Information Systems*, vol. 2021, 6 pages, 2021.

[3] C. Yu and Y. Hang, "Application of artificial intelligence robots in precise management of medical equipment in the context of the internet of things," in *Proceedings of 2019 4th International Industrial Informatics and Computer Engineering Conference(IIICEC 2019)*, pp. 228–234, Ed. Francis Academic Press,UK, 2019.

[4] S. Zhao, "The development of artificial intelligence education resources under the background of the internet of things," in *Proceedings of the 32nd China Conference on Control and Decision-making, .Ed*, pp. 375–380, 2020.

[5] J. Zhang, Y. Ye, C. Hu, and B. Li, "Architecture design and demand analysis on application layer of standard system for ubiquitous power internet of things," *Global Energy Interconnection*, vol. 4, no. 3, pp. 304–314, 2021.

[6] L.-S. Chen, "Design and implementation of intelligent mobile information system for campus safety management," in *The Third (2008) Annual Meeting of Chinese Management–Proceedings of Marketing Sub-meeting.Ed*, pp. 132–139, 2008.

[7] Z. Zhu, "Research on the construction of efficient and intelligent campus based on the internet of things," in *Proceedings of 2020 International Conference on Artificial Intelligence and Communication Technology(AICT 2020).Ed*, pp. 287–290, Clausius Scientific Press, 2020.

[8] L. Rothnie, "Campus wide information system development at three UK universities," *Vine*, vol. 23, no. 4, pp. 18–30, 1993.

[9] Y. Chen and Y. Wei, "On the application of big data and cloud computing in the smart campus," in *Proceedings of 2019 4th International Industrial Informatics and Computer Engineering Conference (IIICEC 2019).Ed*, pp. 223–227, Francis Academic Press, UK, 2019.

[10] Q. Yin, "Design of campus management information system based on intelligent sensor network," *Journal of Guangdong Polytechnic Normal University*, vol. 33, no. 9, pp. 8–10, 2012.

[11] N. ZHANG, "A campus big-data platform architecture for data mining and business intelligence in education institutes," in *Proceedings of 2016 6th International Conference on Machinery,Materials, Environment, Biotechnology and Computer(MMEBC 2016).Ed*, pp. 313–319, 2016.

[12] S. Li and M. Li, "Research and practice of constructing "5A model intelligent campus in all aspects" in higher vocational colleges under the background of education informatization 2.0," in *Proceedings of 2019 3rd International Conference on Education,Management Science and Economics(ICEMSE 2019).Ed*, pp. 484–487, Atlantis Press, 2019.

[13] T. Matsuo and T. Fujimoto, "A new lecture allocation support system based on users' multiple preferences in campus information systems," *International Journal of Computational Intelligence & Applications*, vol. 6, no. 2, pp. 245–256, 2006.

[14] X. Cai, K. Zang, J. Li, and College of Information Engineering North China University of Technology Beijing, 100144, China, "Design and realization of online virtual campus system," in *Proceedings of 2010 Third International Conference on Education Technology and Training .Ed. Xueli Zhou*, pp. 159–161, Institute of Electrical and Electronics Engineers, Inc, 2010.

[15] P. Heping and Z. Jiao, "Design of virtual campus card system based on micro-service architecture," in *Proceedings of the Seventh International Conference on Computing and Information Science.Ed*, pp. 470–477, DEStech Publications, 2019.

[16] K. Yu, Z. Guo, and Y. Shen, "Secure artificial intelligence of things for implicit group recommendations," *IEEE Internet of Things Journal*, 2021.

[17] W. Wang, N. Kumar, J. Chen et al., "Realizing the potential of the internet of things for smart tourism with 5G and AI," *IEEE Network*, vol. 34, no. 6, pp. 295–301, 2020.

[18] K. Yu, M. Arifuzzaman, Z. Wen, D. Zhang, and T. Sato, "A key management scheme for secure communications of information centric advanced metering infrastructure in smart grid," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 8, pp. 2072–2085, 2015.

[19] W. Wang, F. Xia, H. Nie et al., "Vehicle trajectory clustering based on dynamic representation learning of internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 6, pp. 3567–3576, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 7147940, 10 pages
https://doi.org/10.1155/2021/7147940

*Research Article*

# A Metaheuristic Approach to Secure Multimedia Big Data for IoT-Based Smart City Applications

**Harsimranjit Singh Gill**,[1] **Tarandip Singh**,[2] **Baldeep Kaur**,[2] **Gurjot Singh Gaba**,[3] **Mehedi Masud**,[4] **and Mohammed Baz**[5]

[1]*Department of Electronics and Communication Engineering, Guru Nanak Dev Engineering College, India*
[2]*Department of Electronics Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, India*
[3]*School of Computer Science, Mohammed VI Polytechnic University, Ben Guerir 43150, Morocco*
[4]*Department of Computer Science, College of Computers and Information Technology, Taif University, P. O. Box 11099, Taif 21944, Saudi Arabia*
[5]*Department of Computer Engineering, College of Computer and Information Technology, Taif University, P. O. Box 11099, Taif 21994, Saudi Arabia*

Correspondence should be addressed to Mehedi Masud; mmasud@tu.edu.sa

Received 18 May 2021; Revised 16 June 2021; Accepted 15 September 2021; Published 4 October 2021

Academic Editor: Celestine Iwendi



Media streaming falls into the category of Big Data. Regardless of the video duration, an enormous amount of information is encoded in accordance with standardized algorithms of videos. In the transmission of videos, the intended recipient is allowed to receive a copy of the broadcasted video; however, the adversary also has access to it which poses a serious concern to the data confidentiality and availability. In this paper, a cryptographic algorithm, Advanced Encryption Standard, is used to conceal the information from malicious intruders. However, in order to utilize fewer system resources, video information is compressed before its encryption. Various compression algorithms such as Discrete Cosine Transform, Integer Wavelet transforms, and Huffman coding are employed to reduce the enormous size of videos. moving picture expert group is a standard employed in video broadcasting, and it constitutes of different frame types, viz., I, B, and P frames. Later, two frame types carry similar information as of foremost type. Even I frame is to be processed and compressed with the abovementioned schemes to discard any redundant information from it. However, I frame embraces an abundance of new information; thus, encryption of this frame is sufficient enough to safeguard the whole video. The introduction of various compression algorithms can further increase the encryption time of one frame. The performance parameters such as PSNR and compression ratio are examined to further analyze the proposed model's effectiveness. Therefore, the presented approach has superiority over the other schemes when the speed of encryption and processing of data are taken into consideration. After the reversal of the complete system, we have observed no major impact on the quality of the deciphered video. Simulation results ensure that the presented architecture is an efficient method for enciphering the video information.

## 1. Introduction

A city becomes smart when the physical objects are transformed into cyberphysical objects. The transformation facilitates real-time monitoring, manages resources, optimizes smart city operations, and improves citizens' quality of life. Few applications of a smart city include garbage van route optimization, automatic irrigation, wearable health network, and smart energy meters. Internet of Things (IoT) is the technology behind this revolution; it associates a sensor integrated into a communication unit with a physical object. Consequently, the cyberphysical object can then be accessed from anywhere using internet access. One of the primary applications of smart cities is monitoring the roads and rushy areas through closed-circuit television (CCTV) cameras for preventing crimes. There are plenty of CCTVs installed in

the smart city, resulting in enormous multimedia big data [1]. The big multimedia data generated by IoT nodes in the smart city contains sensitive information preventing tampering and other cyberthreats [2].

Information security is vital in communication and even while storing multimedia information [3]. The one way to protect the information is by blocking unauthorized access, but such a method is not very secure and reliable [4]. Another method is to encrypt the information in the gibberish form, so an end-user cannot decode it until the encryption method is known. Image and video encryption have various applications like multimedia messaging, military purposes, and internet communication like video calling, video conferencing, and satellite TV broadcasting [5, 6]. There are various encryption methods; AES is one of the most secure methods on which no possible attack is confirmed to date. In traditional approaches, there are two techniques of encryption, first is encrypting the whole data, and in another technique, entire data is compressed first by compression method, and then, it is encrypted, but these techniques take a lot of time and decrease the processing speed [7].

Currently, numerous compression and encryption techniques are proposed. However, these days' encryption techniques attract attention to joint compression and partial encryption techniques for secure video transmission. There are various methods for image and video encryption; for example, Alattar encrypts intracoded macroblocks of all frames of the MPEG video, which reduces the processing time and increases speed over full encryption of video. Another method called encrypting the header information of predicted macroblock and encrypting the whole data in all I-macroblocks is presented [8]. The method of encryption presented here constitutes three sections, i.e., the motion vector difference, intraprediction modes, and the signed bits of the texture data. Only the selected domain is secured according to the scalability types [9]. For encryption of video, the author proposed a different technique based on one-dimensional chaotic map in the DCT domain that uses multiple operations such as scrambling and encryption of I frame and three chaotic maps. In the whole process, five keys have been incorporated, which were not easy to find, and the I frame changes can make it complex [10, 11].

A novel encryption scheme that exploits partial information as an input used a secure encryption algorithm to encipher a part of compressed information through an orthogonal search algorithm. DCT and some other coding like quantization and arithmetic have been used for image compression, and then, the resultant information is encrypted by RSA algorithm [12, 13]. In the encryption techniques for the secure transmission of MPEG video bitstreams, another method was used, which comprised various encrypted I frames and header information of every predicted frame [14]. The encryption method has been presented where only the AC and DC coefficients of the I frame were encrypted. Both coefficients of I frame, AC coefficients of the P frame, and motion vector difference were encrypted [15]. Another approach based on the hash encryption model was demonstrated in [16]. In this approach, intraprediction, the difference of motion vector, and coefficients of quantization were encrypted. A novel key generation process was constructed using a hash function. In [17], Cheng and Li proposed a partial encryption method in which only a part of compressed data was encrypted. The presented scheme of partial encryption technique was later applied in numerous image and video compression algorithms. The encryption and integrated multimedia compression technique were illustrated in [18] based on modified entropy coders with multiple statistical models and selective encryption models.

Firstly, the limitations of selective encryption using cryptanalysis were explored and then processed the information through the selective encryption model. A similar approach based on multiple statistical models has been presented in which entropy coders were used to designing an encryption cipher. Using this technique, multiple encryption schemes were designed which incorporate the Huffman coder and the QM coder [19]. An unlike approach on text file was compressed and encrypted using chaotically mutated Huffman trees. Many Huffman tables were used to encode that text message. With the use of large keyspace, this technique provides robust security to the Brute-force attack. Another scheme employed lossless compression and contourlet transform before the encryption of image's most significant part. This method promised an increase in cipher image security [20, 21].

In [22], Setyaningsih and Wardoyo proposed a dissimilar technique comprised of compression and encryption technique, in which shared encryption occurs between the low- and high-frequency components. These coefficients and initial keys and total pixel values were used as an input to the hash function. The hash function value was used for encryption of the high-frequency components. For the joint compression and encryption of medical images, another author proposed a technique where the image was compressed by Discrete Wavelet Transform (DWT) and then encrypted by Advanced Encryption Algorithm. This scheme was designed to increase protection along with security [23]. A similar technique of joint image compression and encryption using the properties of integer wavelet transform (IWT) and SPHIT was presented where multiple methods were exploited such as hyperchaotic system, secure hash algorithm, nonlinear inverse operation, and plain text-based keystream to improve the security [24].

To enhance the compression ratio, Song et al. [25] presented a system that employs the intrinsic features of input images along with entropy encoding for the encryption process. SHA-256 has also been used to build a secure, chaotic cryptosystem that is resistant to certain common attacks. Another approach based on 3D chaotic maps was presented to decorate the adjacent pixels of an image after successfully implementing the arithmetic compression algorithm. This technique was developed for transfer images over a network for real-time application [26]. To encrypt [27, 28] the large data files and reduce execution time, the authors proposed the most secure and effective grid-based encryption technique. Here, the image is divided into grids and encrypted by AES algorithm [29]. A unique model has been shown

for the encryption of surveillance videos [30]. Numerous methods for image compression such as DWT, DCT, and Huffman encoding compression algorithm were presented. The medical image was compressed by using these methods [31]. Another technique of compression, IWT, was presented. The lifting process was used, which compressed the image by dividing the odd and even coefficients and then generating four subband images [32]. To achieve compression and scramble the pixels data of an image based on set partitioning in hierarchical trees (SPIHT) was suggested by Xiang et al. in [33]. The presented scheme can provide better resistance for different attacks compared to the original SPIHT technique.

For real-time applications [34–36], a new encryption method was proposed. It constitutes three sections: motion vector difference, intraprediction mode, and residual data. The encryption was executed by Network Abstraction Layer and distinguished the enhancement layer spatial scalability and temporal scalability [37]. In the field of compression and encryption, a new method of encryption with the scan pattern was proposed. This technique was based on scan methodology, which creates many scanning paths and space-filling curves. Firstly, lossy compression was applied on the difference of adjacent frames, and then, encryption was performed on compressed frame differences [38]. Another approach illustrates the usage of wavelength division multiplexed systems for end-to-end distribution of compressed video [39]. Moreover, speech signals can be transported between multiple entities of a network, keeping end-to-end encryption into consideration, using chaotic and cryptographic algorithms. The various chaotic maps have been employed to scramble the speech information, and semantic encryption techniques were used to encipher the information [40].

Based on the literature review, it is found that a number of encryption techniques have been applied to the video information, but no scheme has explored the possibility of joint compression technique and encryption technique. Therefore, this work focuses on combining both schemes, and we have tested it on a multimedia file. The proposed model is employed to broadcast video between two ends. Various techniques are explored to provide compression, such as IWT, DCT, Huffman coding, and encryption involved in AES algorithm. For video compression, the information of three frames, which includes I, B, and P frames, has been used. Among three frames, I frame contains the most information of the video. On the other side, P and B frames contain only a small portion of image information. The proposed model includes the following steps: the information of the I frame is extracted first from a MPEG video. In the second step, an IWT is applied to extracted I frame. After that, an image is divided into different subbands such as LL, HL, LH, and HH, respectively. The LL subband is the closest guesstimate of an original image. In the third step, DCT is applied to the LL band, and the resultant image is divided into one DC and various AC coefficients. During encryption, the DC coefficient is partially encrypted using the AES-128 bit, and the rest of AC coefficients are compressed by Huffman coding. AES and Huffman coding output is concatenated in the last step, and a cipher image is obtained.

The rest of the paper is organized as follows. A review of IWT, DCT, Huffman encoding, and AES is given in Section 2. In Section 3, the proposed approach is presented. Simulation results and discussion are presented in Section 4. Finally, we summarize the paper and present a conclusion in Section 5.

## 2. Preliminaries

*2.1. Integer Wavelet Transform (IWT).* When the image is decomposed, it is divided into different groups. The approximated content of an image is further divided into four subbands. The IWT provides a better result of compression prior to which approximate contents of the image are decomposed. It is a form of DWT and has many advantages of DWT, but it also has some functions that DWT cannot perform. It uses round-off values rather than floating-point values. Forward and reverse scheme is shown in Figures 1 and 2. Forward and reverse lifting scheme (LS) is used to perform simple shifting and adding operations. LS is used to divide the odd and even coefficients. This scheme is performed by three steps, i.e., split, predict, and update.

(i) Split: input image or signal is divided into even and odd coefficients

(ii) Predict: combining even values from predicted odd samples and then subtracting it from calculated odd samples to generate prediction error

(iii) Update: add the computed predicted error to update the entire even samples

Forward LS is used to compress the image and reverse LS for the reconstruction of the signal. Every transform by this scheme can be inverted [24, 32].

*2.2. Discrete Cosine Transform (DCT).* DCT is usually employed in almost all types of multimedia compression schemes. Likewise, in Discrete Fourier Transform, DCT converts a sequence of data or information from spatial-domain to frequency-domain. As DFT is based on complex numbers, DCT uses real numbers. The sequence generated by DCT is the addition of cosine functions that waver at various frequencies decorrelates the image information into different frequency bands. When calculating DCT of an image, the values that are in the high-frequency bands are near to zero, and then, compression occurs after quantization. Initially, the RGB image is translated into the $Y_{Cb}C_r$ color space. After that, each color space is converted into a number of $8 \times 8$ blocks which are again converted into DCT domain by using the 2D-DCT formula.

*2.3. Huffman Coding.* Huffman coding is a type of lossless data compression algorithm. It is the form of statistical coding which is used to reduce the input information bits and gives the strings of symbols. It assigns the dynamic length codes to the input characters. The length of that allocated codes depends on the occurrence of input characters. The

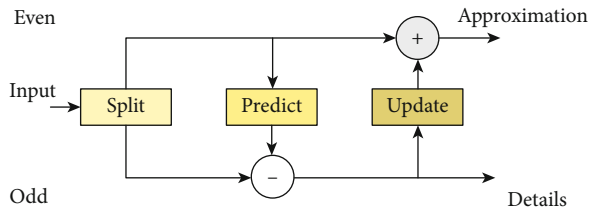



Figure 1: Forward lifting scheme.

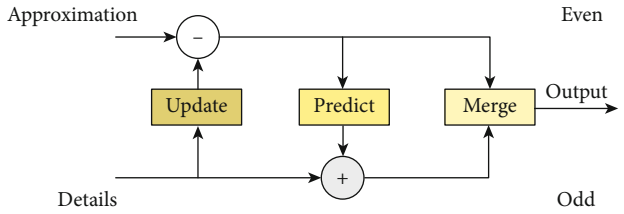



Figure 2: Reverse lifting scheme.

most recurring character is translated to a shorter code, and the character having the opposite frequency gets the longest code. The length of the codeword is not variable. It can reconstruct the original image or data [31].

*2.4. Advanced Encryption Standard.* AES is an encryption algorithm that is used to encrypt an image over the network. The AES was announced by the National Institute of Standards and Technology in the year 2001. AES falls under the category of an asymmetrical block cipher and was designed and implemented in both software and hardware. The block length varies from 128 bits to 512 bits and has a similar range of key length. Depending upon the size of the block length and key length is to be fixed with a similar size; thus, the number of rounds is selected, which range from 10 to 14 rounds, i.e., 16-byte key to 32-byte key. Each round is designed to perform four similar steps: permutation, arithmetic operations, byte substitution over a finite field, and XOR operation with a key. For the calculation of arithmetic operations, the modular reduction method can be used in Galois fields of mathematics. In AES, representation of each element is done as

$$A(x) = a_7x^7 + \cdots + a_1x + a_0, \tag{1}$$

where $a_i \in GF(2) = \{0, 1\}$. In addition to representation, each polynomial of AES is represented using the following notation of vector

$$A = [a_7, a_6, a_5, a_4, a_3, a_2, a_1, a_0]. \tag{2}$$

Modulo reduction plays a vital role in arithmetic operations, and the default irreducible polynomial is given as

$$P(x) = x^8 + x^4 + x^3 + x + 1. \tag{3}$$

This algorithm is applied to the DC coefficients of an image extracted by the compression algorithm and then transferred over the public network in the presented work. The original image can be reconstructed at the receiver side by applying a decryption algorithm on the cipher image. The length of all keys of the AES algorithm is sufficient to protect classified information up to the secret level. Thus, this algorithm gives better security and data confidentiality [14–18]. AES requires small space and low memory for the implementation of both encryption and decryption. An unlike modification in AES algorithm through primitive operations has been shown to mitigate low diffusion rate at the initial stage [41–43].

## 3. Proposed Approach

This section has provided a detailed description of enciphering and deciphering of I frames extracted from MPEG video.

*3.1. Compression and Encryption Approach.* The architecture of joint image compression and encryption is illustrated in Figure 3. There are various steps to perform joint image compression and encryption.

*Step 1.* Firstly, an I frame is selected from a MPEG video which contains more image information. I frame is an intra-coded completely specified picture and has a large amount of image information selected for compression.

*Step 2.* In the second step, the single-level decomposition of IWT is performed on I frame. IWT is based on the subband coding and lifting scheme. After transformation, four subbands LL, HL, LH, and HH are extracted. The LL subband is the closest estimation of the original image, HL subband signifies the detail about verticals, LH subband denotes the detail about the horizontal edge, and HH subband represents the detail about diagonal. Therefore, the LL subband is compressed because it has greater image information.

*Step 3.* In the third step, DCT is performed on LL subband according to the mathematical formula given as:

$$V_{pq} = \delta_p\delta_q \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} U_{mn} \cos\frac{\pi(2m+1)p}{2M} \cos\frac{\pi(2n+1)q}{2N}, 1 \le p \le M-1, /1 \le q \le N-1, \tag{4}$$

where $V_{pq}$ is known as DCT coefficients of $U$. The variables $\delta_p$ and $\delta_q$ are calculated as

$$\delta_p = \begin{cases} \frac{1}{\sqrt{M}}, & p = 0, \\ \sqrt{2/M}, & 1 \le p \le M-1, \end{cases} \qquad \delta_q = \begin{cases} \frac{1}{\sqrt{N}}, & q = 0, \\ \sqrt{2/N}, & 1 \le q \le N-1. \end{cases} \tag{5}$$

DCT is primarily used in the various types of multimedia compression schemes. It gives a finite sequence of data in

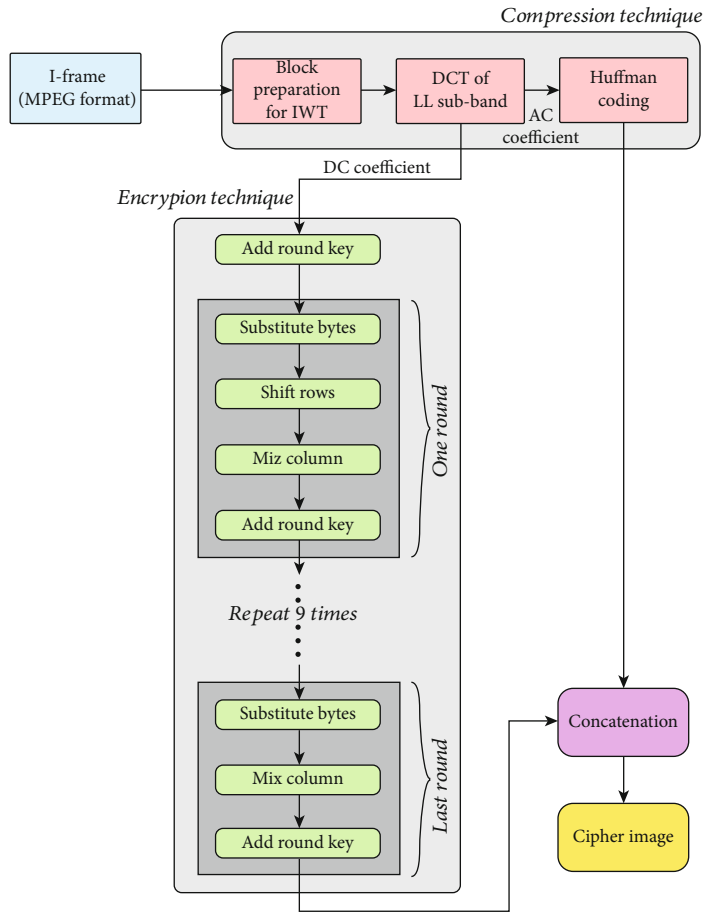



Figure 3: The basic building block of joint image compression and encryption.

terms of a cosine function. It alters an input image to the frequency domain from the spatial domain. When DCT is applied to the LL band, an image is obtained with one DC constant, and others are AC values. The DC coefficient is a low-frequency component with a huge value, and AC constants are high-frequency components close to zero.

*Step 4.* In this step, partial encryption is performed. The DC coefficient, which occurs after DCT compression, is partially encrypted by using the AES-128 bit. It performs 10 rounds on input data for the purpose of more confusion and diffusion. More rounds mean more security against the cryptanalysis attack. The detailed algorithm used for enciphering DC coefficient is presented in Algorithm 1.

*Step 5.* There are numerous tools used for compression purposes like Huffman coding, run-length encoding, entropy encoding, and arithmetic encoding. In this scheme, Huffman encoding is used, which is lossless data compression algorithm. The rest of AC coefficients obtained from DCT is further compressed by using Huffman encoding. It reduces the information bits to fewer bits, and the compressed image is obtained.

*Step 6.* In the final step, concatenation involves the output of the AES-128 bit and Huffman encoder. As a result, the cipher image is obtained, which is totally different from the original image. Moreover, if someone can access the AC coefficients of data, even then, the adversary may not be able to decrypt the image, owing to the robustness of the encryption model.

To address the issue of processing the data, we have incorporated the compression algorithms before preparing and presenting the data for encryption. In this way, the proposed architecture can be used to save a lot of system resources while still concealing the information and getting a plausible result from the presented system. In comparison to approaches discussed in [10, 15], the presented model shown in Figure 3 has superiority in terms of the time taken to process one frame of data. Since the data has already been compressed tremendously before the application of encryption, thereby, this approach is considered higher performance in terms of speed.

*3.2. Extraction and Decryption Approach.* To retrieve I frame from the cipher text or encrypted image, all the blocks presented in Figure 3 can be reversed. Primarily, the received data is segregated into two blocks, the former block is given to the Huffman decoder, and the latter block is fed for AES decryption. The Huffman decoder is used to retrieve actual AC coefficients for I-DCT; however, AES decryption is employed to recover the values of DC coefficients of

***Data:*** An input of DC coefficient calculated after application of forward DCT
***Result:*** Ciphertext
**1** for *key_expansion*
**2** $W = [W_0 W_1 \cdots W_s]$where $W \in K_{AB}$ and $W_s = W_3/W_5/W_7$
**3** *initialize* $k_r$
**4** for $i = s + 1$ to 43
**5** $temp = sbox(W_{i-1} \lll 8, 8)$
**6** $g = temp \bigoplus rcon(k_r + 1)$
**7** $W_i = W_{i-s-1} \bigoplus g$
**8** $W_{i+1} = W_{i-s} \bigoplus W_i$
**9** $W_{i+2} = W_{i-s+1} \bigoplus W_{i+1}$
**10** $i = i + s + 1$
**11** *increment* $k_r$
**12** *update* $W$ and ***go*** to *line* 3
**13** ***while*** *encryption*
**14** ***prepare*** *DCT data* in *blocks* of 128 *bits* in 4 *X* 4 *matrix*
**15** $intialphase = block_{1\_input} \bigoplus block_{first\_subkey}$
**16** for *round* 9/*downto* 1
**17** $bytesubs = sbox(intialphase)$
**18** for *shiftrow*
**19** *circular* − *shift row* 1/2/3/4 *right with* 0/3/2/1 *bytes*
**20** for *mixcolumn*
**21** for *each_row* and *each_column*
**22** $mcol = constants * shiftrow$**where***constants* is *a* 4 *X* 4 *matrix*
**23** $addrk = mcol \bigoplus block_{round\_subkey}$
**24** for *lastround*
**25** *repeat line* 17 to 19
**26** out = *line* 25 ⊕ $block_{last\_subkey}$
**27** *Ciphertext* = out
**28** ***go*** to *line* 14

Algorithm 1: Algorithm used for enciphering the data.

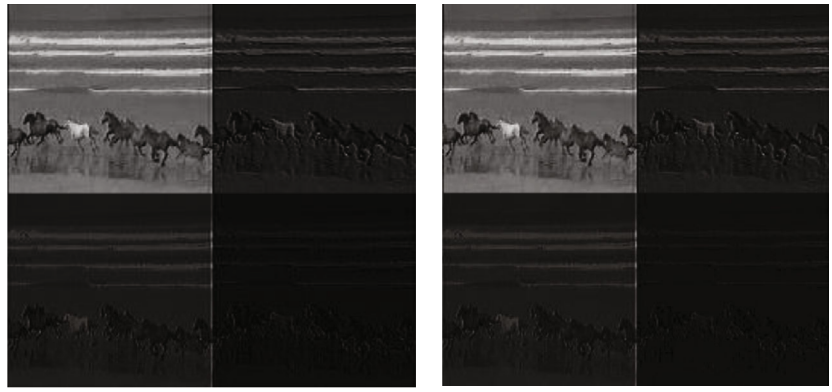
different macroblocks of the image. AES decryption process is an exact replica of AES encryption model, but only a reversal of the key schedule. This means DC coefficients are evaluated through the similar and symmetric key used for enciphering purposes. The evaluated coefficients are then processed through the inverse DCT stage and then given to inverse IWT block to construct the I frame of the transmitted video. To conclude, it is assessed that the extraction process of the I frame includes similar but reversal blocks and functions employed at the sender side.

## 4. Results and Discussions

The simulation of the proposed technique is performed on MATLAB with an Intel-core-5 i5 processor and 1 TB memory. The video is played in MATLAB video reader from where I, B, and P frames are extracted from MPEG video. Among these frames, a random I frame is selected from a video (as shown in Figure 4).

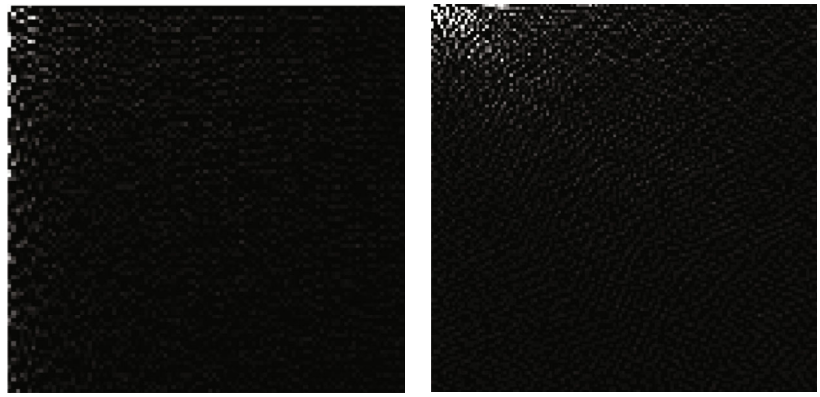
After selecting I frame, IWT is applied to transform the image, which divides the image into four subbands shown in Figure 5. Forward and reverse lifting scheme is also applied in which simple shifting and adding operations are performed. The data can be recovered by reverse lifting scheme without any loss. LS is used to divide the odd and even coefficients. Three stages perform the presented scheme: split, predict, and update. The split function divides the input image or signal into even and odd coefficients. Predict function forecasts the odd sample as a linear mixture of

Figure 4: Original I frame extracted from a MPEG video.



(a) (b)

Figure 5: Image after subband coding. (a) Operation is performed on $200 \times 200$. (b) Operation is performed on $256 \times 256$.



(a) (b)

Figure 6: DCT on LL subband. (a) Operation is performed on $200 \times 200$. (b) Operation is performed on $256 \times 256$.

even portions. Formerly, it is subtracted from the odd portion to generate a prediction error. Update function updates the even portion by totaling them to the previously generated value, i.e., prediction error. The same operation is performed on $200 \times 200$ (represented in Figure 5(a)) and $256 \times 256$ images (shown in Figure 5(b)).

The first band after decomposition is the LL band. LL band is the approximation of the original image, which contains more information. So, it is selected for compression purposes. After IWT, DCT is applied on the LL subband to compress the image. The image obtained after DCT is shown in Figure 6. When DCT is performed on an image, it divides into DC and AC coefficients. The DC coefficients are low-frequency components and have the highest value, and the AC coefficients in the high-frequency bands have a value close to zero, and then, compression occurs after quantization.

The sequence generated by DCT is the summation of cosine functions that oscillate at various frequencies or we can say that it decorrelates the image information into multiple frequency bands. Firstly, the color image of RGB pattern is converted into the $YC_b$ $C_r$ color space. After that, individual color space is separated into several $8 \times 8$ blocks, which are again converted into DCT domain by using the 2D-DCT formula in Equation (4). The white portion of the image is DC coefficient, and the other black portion of the image is AC coefficient. The operation is performed on $200 \times 200$ is shown in Figure 6(a) and on $256 \times 256$ pixel image is shown in Figure 6(b). After DCT, DC coefficients have the highest value selected for encryption purposes. To encrypt the DC coefficient through network's highest secure symmetric encryption algorithm, Advanced Encryption Algorithm is used. In the presented scheme, 128-bit AES algorithm is used. Here, the algorithm constitutes the following parameters: size of the key is 128 bits, block size of plain text is 128 bits, and ciphertext block size is similar to the size of plain text.

Some tests have been carried out to evaluate the effectiveness of the proposed system. The I frame extracted in Figure 4 is presented for the computation of the Average Peak Signal to Noise Ratio (PSNR) and compression ratio. Both of these parameters are observed to measure the quality of compressed, encrypted, and transmitted frames. The higher value of PSNR depicts higher fidelity or better reconstruction of the transmitted image. PSNR values in the range of 25-30 dB represent the decent quality of the reconstructed image; however, higher values indicate better visual quality. To calculate PSNR of an individual frame, the mean square error is observed beforehand, and then, the logarithmic value of PSNR is measured. The mathematical expressions

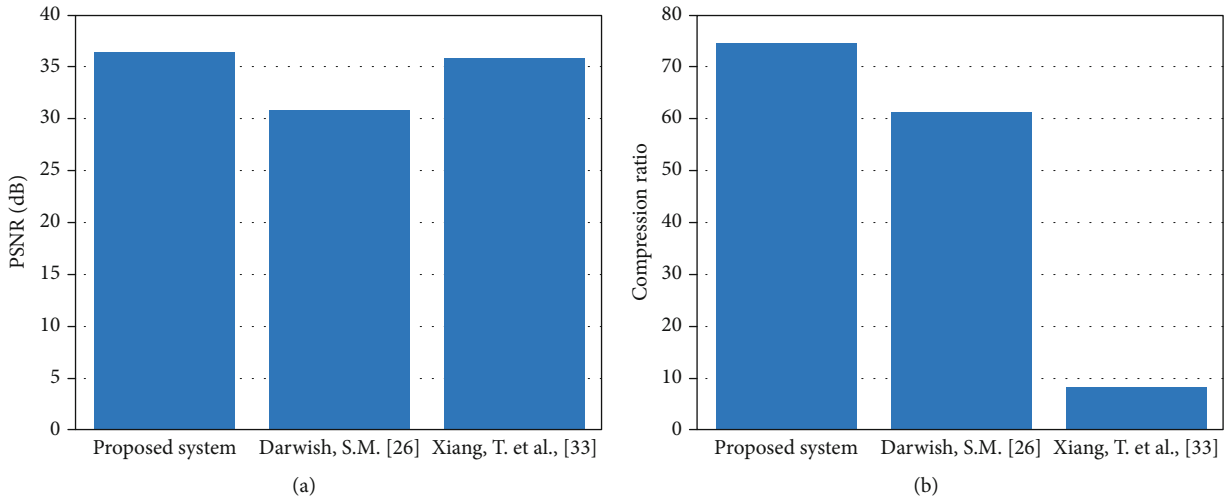



Figure 7: Performance analysis. (a) Calculated PSNR for $256 \times 256$ image. (b) Computed compression ratio for different techniques.

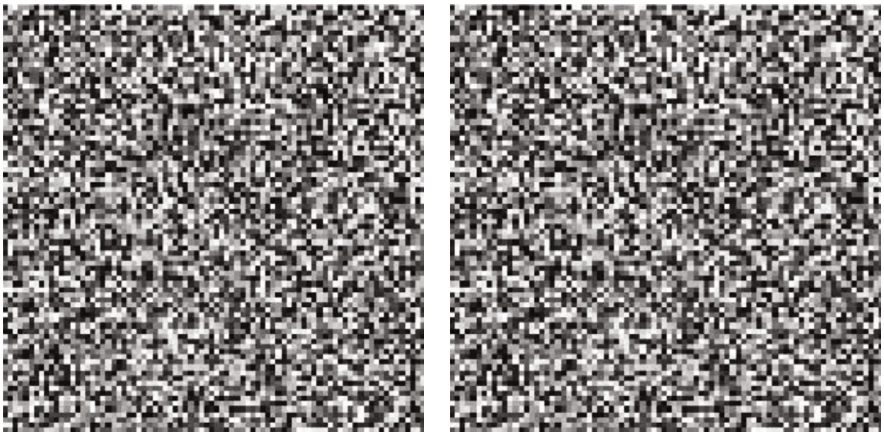



Figure 8: Cipher image. (a) Operation performed on $200 \times 200$. (b) Operation is performed on $256 \times 256$.

to evaluate PSNR and compression ratio are expressed as

$$\text{PSNR} = 10 \log_{10}\left(\frac{255^2}{\text{MSE}}\right),$$
$$\text{Compression ratio} = \frac{\text{Size of original I} - \text{frame}}{\text{Size of encoded frame}}. \quad (6)$$

Both of the mentioned parameters are shown in Figures 7(a) and 7(b). The proposed technique is compared with models presented by Darwish [26] and Xiang et al. [33]. It is clear from Figure 7(a) that PSNR value of each of the models is higher than 30 dB, which means the perceived frames at the receiver end are of high quality. Moreover, the calculated percentage of the compression ratio of individual I frame is close to 75% in comparison to 65% and 8% of techniques presented in [8, 26], respectively. After the analysis of compression ratio, it is examined that each pixel of I frame is represented by 0.107 bit as compared to 0.123 bit and 1 bit, respectively. Thus, the observations made demonstrate the effectiveness and efficiency of the proposed system.

Excluding the last round in every case, all other rounds are identical. Every round of operation includes one substitution step, a row-wise permutation step, a column-wise mixing step, and the addition of the round key. DC coefficients after DCT are encrypted. The rest of the AC coefficients are further compressed by Huffman encoding. It is a type of lossless data compression algorithm. It assigns the variable-length codes to the input characters that are AC coefficients. The most frequent character gets the smallest code, and the character which is least frequent gets the largest code. When Huffman coding is applied to the AC coefficients, it reduces the bits into fewer bits and gives output. After Huffman encoding, concatenation combines the output of Huffman encoder and AES, which results in cipher image shown in Figure 8. The same operation is performed on $200 \times 200$ shown in Figure 8(a) and on $256 \times 256$ pixel image shown in Figure 8(b). It is a scrambled form of the original image; it can only be decrypt by a person who has a secret key called a decryption key.

## 5. Conclusions and Future Directions

A joint image compression and encryption scheme for video broadcasting is proposed. The proposed technique includes two key operations: extracting I frame and encrypting that frame. I frame is selected from a MPEG video for the purpose of compression. IWT is performed on I frame, and images are divided into four subbands, then DCT is applied on LL band. After that, compression image is divided into DC and AC coefficients. After that, partial encryption is performed on the DC coefficient. AC coefficients are compressed further with Huffman coding. Finally, the compressed AC coefficients and encrypted data concatenate to form a cipher image. The simulation results and evaluated PSNR and compression ratio values show that the presented technique is efficient and gives proper security. The encryption process does not modify the compressed data and does not change the quality of the video. The results show that the frame encryption method is secure, and the proposed scheme fits the multimedia system and Internet communication or secret communication. The prospect of this research work includes the incorporation of Artificial Intelligence and Machine Learning to secure big multimedia data from cyberabuses.

## Data Availability

The data that support the findings of this study are available upon request.

## Conflicts of Interest

The authors declare that they have no competing interests.

## Acknowledgments

The authors would like to thank the Taif University Researchers Supporting Project number (TURSP-2020/239), Taif University, Taif, Saudi Arabia, for the support.

## References

[1] C. Iwendi, S. Ponnan, R. Munirathinam, K. Srinivasan, and C.-Y. Chang, "An efficient and unique TF/IDF algorithmic model-based data analysis for handling applications with big data streaming," *Electronics*, vol. 8, no. 11, p. 1331, 2019.

[2] M. Mohit, L. K. Saraswat, C. Iwendi, and J. H. Anajemba, "A neuro-fuzzy approach for intrusion detection in energy efficient sensor routing," in *4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*, pp. 1–5, Ghaziabad, India, 2019.

[3] A. Rehman, S. U. Rehman, M. Khan, M. Alazab, and G. Thippa Reddy, "CANintelliIDS: detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-based GRU," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 1456–1466, 2021.

[4] N. Deepa, Q.-V. Pham, D. C. Nguyen et al., "A survey on blockchain for big data: approaches, opportunities, and future directions," 2020, https://arxiv.org/abs/2009.00858.

[5] C. Iwendi, S. Khan, J. H. Anajemba, M. Mittal, M. Alenezi, and M. Alazab, "The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems," *Sensors*, vol. 20, no. 9, p. 2559, 2020.

[6] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[7] C. Iwendi, P. K. R. Maddikunta, T. R. Gadekallu, K. Lakshmanna, A. K. Bashir, and M. J. Piran, *A Metaheuristic Optimization Approach for Energy Efficiency in the IoT Networks*, Software: Practice and Experience, 2020.

[8] A. M. Alattar and G. I. Al-Regib, "Evaluation of selective encryption techniques for secure transmission of MPEG compressed bit streams," in *1999 IEEE International Symposium on Circuits and Systems (ISCAS)*, Orlando, FL, USA, 1999.

[9] S. W. Park and S. U. Shin, "Efficient selective encryption scheme for the H.264/scalable video coding (SVC)," in *International Conference on Networked Computing and Advanced Information Management*, Gyeongju, South Korea, 2008.

[10] S. Yang and S. Sun, *A Video Encryption Method Based on Chaotic Maps in DCT Domain*, Progress in Natural Science, China, 2008.

[11] S. A. Aliesawi, D. S. Alani, and A. M. Awad, "Secure image transmission over wireless network," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 2758–2764, 2018.

[12] A. A. Alhijaj and M. K. Hussein, "Stereo images encryption by OSA & RSA algorithms," *Journal of Physics: Conference Series*, vol. 1279, article 012045, 2019.

[13] S. SerElkhetm and S. Heshmat, "A survey study on joint image compression - encryption methods," in *International Conference on Innovative Trends in Communication and Computer Engineering*, Aswan, Egypt, 2020.

[14] A. M. Alattar and G. I. Al-Regib, "Improved selective encryption techniques for secure transmission of MPEG video bit streams," in *IEEE International Conference on Image Processing*, Kobe, Japan, 1999.

[15] F. Chiaraluce, L. Ciccarelli, E. Gambi, P. Pierleoni, and M. Reginelli, "A new chaotic algorithm for video encryption," *IEEE Transactions on Consumer Electronics*, vol. 48, no. 4, 2003.

[16] X. Wang, N. Zheng, and L. Tian, "Hash key-based video encryption scheme for H.264/AVC," *Signal Processing: Image Communication*, vol. 25, no. 6, pp. 427–437, 2010.

[17] H. Cheng and X. Li, "Partial encryption of compressed images and videos," *IEEE Transactions on Signal Processing*, vol. 48, no. 8, pp. 2439–2451, 2000.

[18] Chung-Ping Wu and C.-C. J. Kuo, "Design of integrated multimedia compression and encryption systems," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 828–839, 2005.

[19] H. Hermassi, R. Rhouma, and S. Belghith, "Joint compression and encryption using chaotically mutated Huffman trees," *Communications in Nonlinear Science and Numerical Simulation*, vol. 15, no. 10, pp. 2987–2999, 2010.

[20] S. Qiu, Y. Cui, and X. Meng, "A Data Encryption and Fast Transmission Algorithm Based on Surveillance Video," *Wireless Communications and Mobile Computing*, vol. 2020, no. - Article ID 8842412, p. 12, 2020.

[21] A. T. Hashim and B. D. Jalil, "Color image encryption based on chaotic shit keying with lossless compression," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 6, p. 5736, 2020.

[22] E. Setyaningsih and R. Wardoyo, "Review of image compression and encryption techniques," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 2, 2017.

[23] S. Ajili, M. A. Hajjaji, and A. Mtibaa, "Hybrid SVD-DWT watermarking technique using AES algorithm for medical image safe transfer," in *International Conference on Sciences and Techniques of Automatic Control and Computer Engineering*, Monastir, Tunisia, 2015.

[24] M. Zhang and X. Tong, "Joint image encryption and compression scheme based on IWT and SPIHT," *Optics and Lasers in Engineering*, vol. 90, pp. 254–274, 2017.

[25] Y. Song, Z. Zhu, W. Zhang, L. Guo, X. Yang, and H. Yu, "Joint image compression–encryption scheme using entropy coding and compressive sensing," *NNonlinear Dynamics*, vol. 95, no. 3, pp. 2235–2261, 2019.

[26] S. M. Darwish, "A modified image selective encryption-compression technique based on 3D chaotic maps and arithmetic coding," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 19229–19252, 2019.

[27] M. Masud, G. S. Gaba, K. Choudhary, M. S. Hossain, M. F. Alhamid, and G. Muhammad, "Lightweight and anonymity-preserving user authentication scheme for IoT-based healthcare," *IEEE Internet of Things Journal*, 2021, https://ieeexplore.ieee.org/document/9430932.

[28] G. S. Gaba, G. Kumar, H. Monga, T.-H. Kim, M. Liyanage, and P. Kumar, "Robust and lightweight key exchange (LKE) protocol for industry 4.0," *IEEE Access*, vol. 8, pp. 132808–132824, 2020.

[29] N. B. Rad and H. Shah-Hosseini, "Grid-based cryptography with AES algorithm," in *International Conference on Computer and Electrical Engineering*, Phuket, Thailand, 2008.

[30] X. Cao, M. Ma, X. Guo, L. Du, and D. Lin, "A new encryption scheme for surveillance videos," *Frontiers of Computer Science*, vol. 9, no. 5, pp. 765–777, 2015.

[31] T. Kumar and R. Kumar, "Medical image compression using hybrid techniques of DWT, DCT and Huffman coding," *International Journal of Innovative research in Electrical, Electronics, Instrumentation and Control Engineering*, vol. 3, no. 3, pp. 54–60, 2015.

[32] T. Mukherjee, B. Y. V. N. R. Swamy, and M. V. L. Bhavani, "Robust image compression using integer wavelet transform exploiting lifting scheme," *International Journal of Engineering Trends and Technology*, vol. 7, no. 5, pp. 217–220, 2014.

[33] T. Xiang, J. Qu, and D. Xiao, "Joint SPIHT compression and selective encryption," *Applied Soft Computing*, vol. 21, pp. 159–170, 2014.

[34] M. Masud, M. Alazab, K. Choudhary, and G. S. Gaba, "3P-SAKE: privacy-preserving and physically secured authenticated key establishment protocol for wireless industrial networks," *Computer Communications*, vol. 175, pp. 82–90, 2021.

[35] G. S. Gaba, G. Kumar, H. Monga, T.-H. Kim, and P. Kumar, "Robust and lightweight mutual authentication scheme in distributed smart environments," *IEEE Access*, vol. 8, pp. 69722–69733, 2020.

[36] M. Masud, G. S. Gaba, K. Choudhary, R. Alroobaea, and M. S. Hossain, "A robust and lightweight secure access scheme for cloud based E-healthcare services," *Peer-to-peer Networking and Applications*, vol. 14, no. 5, pp. 3043–3057, 2021.

[37] L. M. Varlakshmi, G. F. Sudha, and G. Jaikishan, "An efficient scalable video encryption scheme for real time applications," *Procedia Engineering*, vol. 30, pp. 852–860, 2012.

[38] S. S. Maniccam and N. G. Bourbakis, "Image and video encryption using SCAN patterns," *Journal of Pattern Recognition*, vol. 37, no. 4, pp. 725–737, 2004.

[39] Y.-T. Chang, Y.-C. Lin, and W.-H. Wang, "Intelligent shuffling cryptography with dynamic AWG/switch matrix for video transmission in WDM-PON network," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 2, pp. 1009–1022, 2019.

[40] G. Kaur, K. Singh, and H. S. Gill, "Chaos-based joint speech encryption scheme using SHA-1," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 10927–10947, 2021.

[41] E. M. de Los Reyes, A. M. Sison, and R. Medina, "Modified AES cipher round and key schedule," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 7, no. 1, 2019.

[42] M. Masud, G. S. Gaba, S. Alqahtani et al., "A lightweight and robust secure key establishment protocol for internet of medical things in COVID-19 patients care," *IEEE Internet of Things Journal*, 2021.

[43] S. Ibrahim, H. Alhumyani, M. Masud et al., "Framework for efficient medical image encryption using dynamic S-boxes and chaotic maps," *IEEE Access*, vol. 8, pp. 160433–160449, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9070182, 12 pages
https://doi.org/10.1155/2021/9070182

*Research Article*

# Recognition Method of Tunnel Lining Defects Based on Deep Learning

**Anfu Zhu, Shuaihao Chen, Fangfang Lu, Congxiao Ma, and Fengrui Zhang**

*North China University of Water Resources and Electric Power, Zhengzhou, China*

Correspondence should be addressed to Fengrui Zhang; fanbishunzhuisi@163.com

Received 9 June 2021; Accepted 10 August 2021; Published 30 September 2021

Academic Editor: Rajesh Kaluri



The defect identification of tunnel lining is a task with a lot of tasks and time-consuming work, and currently, it mainly relies on manual operation. This paper takes the ground-penetrating radar image of the internal defects of the lining as the research object, and chooses the popular VGG16, ResNet34 convolutional neural network (CNN) to build the automatic recognition model for comparative study, and proposes an improved ResNet34 defect-recognition model. In this paper, SGD and Adam training algorithms are used to update network parameters, and the PyTorch depth framework is used to train the network. The test results show that the ResNet34 network has faster convergence speed, higher accuracy rate, and shorter training time than the VGG16 network. The ResNet34 network using the Adam algorithm can achieve 99.08% accuracy. The improved ResNet34 network can achieve an accuracy of 99.25%, and at the same, reduce the parameter amount by 4.22% compared with the ResNet34 network, which can better identify defects in the lining. The research in this paper shows that the deep learning method can provide new ideas for the identification of tunnel lining defects.

## 1. Introduction

The number of tunnels in our country is increasing. Due to the influence of construction conditions and environment, the lining structure of tunnels will inevitably have defects during construction and operation. Common defects include imperfect lining, voids, cracks, and water seepage. If these defects cannot be detected in time, they will damage the structure of the tunnel and even cause major safety accidents. In order to discover these defects in time and ensure the safety of the tunnel structure, we need a faster and more accurate method to regularly check the quality of the tunnel. In the early days, tunnel detection was carried out by manual methods such as human eye observation and borehole survey methods. The human eye observation method is a common method for detecting defects on the tunnel surface, but this method has many unsafe factors for the inspectors. And the detection is slow, easy to be affected by environmental influences and personal factors [1]. With the development of detection technology, Ai et al. proposed a fast and effective method for obtaining the profile of subway tunnels based on photogrammetry. From the results of field application, the proposed method shows that the collection speed exceeds 5 km/h. It can scan a cross-sectional profile in one second, but it can only detect tunnel surface defects [2]. The borehole survey method can check the lining thickness, lining strength, lining cavity, and other lining internal conditions, but this method is a kind of damage detection, which will destroy the overall structure and waterproof performance of the tunnel, and the operation is complicated. Ultrasonic method and rebound method [3] are nondestructive testing methods that can detect the strength and thickness of concrete, but this method has certain limitations and is easily affected by concrete materials and the environment. The insufficiency of the acoustic wave method has prompted the development of ground-penetrating radar (GPR) technology. The ground-penetrating radar method [4] is a method of detecting the characteristics of the medium by emitting high-frequency electromagnetic waves. This detection method has no damage to the detected object, and the signal processing and image reconstruction technology is also widely used in the ground-penetrating radar images, so the

ground-penetrating radar technology is chosen for many nondestructive testing work. The research of this article also takes the radar image of the lining disease as the research object.

With the increase of the number of tunnels, the task of inspectors is getting heavier and heavier [5], and automatic tunnel identification technology is becoming more and more urgent. In terms of automatic identification of tunnel images, Dong et al. proposed an automatic identification method of tunnel lining cracks based on the local characteristics of the image and realized automatic identification of cracks by designing a clever cross-shaped template [6]. MOOn et al. can easily visualize concrete cracks by using image processing techniques such as improved subtraction, filtering, and segmentation and finally use BP network for classification. The recognition methods used in the above research all require manual production or extraction of features and rely on manual parameter adjustment, which is inefficient.

In recent years, electronic information technology based on optimization has been greatly developed [7–10]. Among them, deep learning technology is widely used in image recognition and speech recognition. This technology has also performed well in other scientific research fields [11]. These excellent performances are mainly attributed to convolutional neural networks. The network [12] can automatically complete the feature extraction in the process of model training. Compared with machine learning algorithms, it does not need to manually make and extract features. Song et al. [13] proposed an objective and fast tunnel crack recognition algorithm based on computer vision semantic segmentation, which achieved accurate segmentation of the location of tunnel cracks, thereby saving a lot of labor and financial resources in the railway sector and improving work efficiency. Decor et al. [14] use deep convolutional networks to train samples and establish a disease image classification system. The recognition rate of the network model can reach more than 95%. Xue et al. [15] proposed an automatic calculation method for the leakage area based on the tunnel surface image data set. After constructing the leakage segmentation model based on deep learning, three optimization measures were adopted: data enhancement, migration learning, and cascade strategy. It is very helpful to improve the segmentation performance of the original model.

Research on the application of deep learning technology to tunnel quality inspection [16] has gradually increased in recent years, but most of the studies are on the surface of tunnels or highways, and there are relatively few quality inspections on the interior of tunnel linings [17]. During the construction and operation of the tunnel, various defects will appear inside the lining. The safety of the tunnel lining is related to the safety of the entire tunnel. In this paper, the radar image of tunnel lining defects is used as the target object, and the deep learning method is used to identify the defects of tunnel lining. Deep learning methods are more intelligent than traditional recognition methods and do not need to manually design image features. This paper uses the excellent VGG16 and ResNet34 recognition networks for comparative research and proposes an improved ResNet network based on the ResNet34 network. The improved network has higher accuracy than the first two networks and uses fewer model parameters than the ResNet34 network. The second section of this article mainly introduces the data set and data preprocessing. The research data mainly uses incompact and hollow images, and the data is enhanced through methods such as cropping, mirror flipping, and adding noise. Section 3 gives an overview of the convolutional neural network and analyzes the structure and function of each network layer of the classic convolutional neural network. At the same time, the main deep learning classification network of VGG16 and ResNet and the improved network proposed in this paper are also studied. Section 4 mainly verifies the recognition effect of different network models through simulation experiments. The experimental results show the effectiveness of the deep learning method in the identification of tunnel lining defects, and it has a high accuracy rate. Moreover, the improved network proposed in this paper has a better effect in the identification of tunnel defects and can be used as a reference method for later tunnel lining inspection and maintenance.

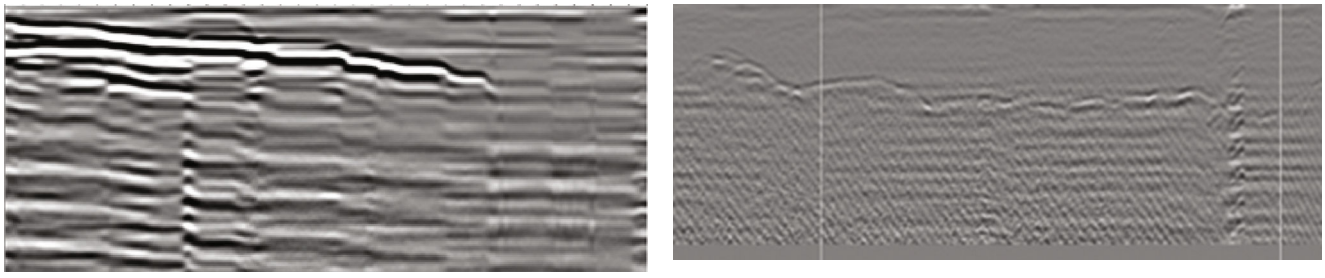
## 2. Establishment of Image Data Set

Deep learning and big data play a more and more important role in the field of artificial intelligence and have been widely used [18–20]. Data samples are the most important and indispensable part of deep learning. Network models need to learn the characteristic information in data samples to continuously evolve their models. At the same time, sufficient high-quality data can improve the robustness of the network model to the data and can also avoid network overfitting [21].

*2.1. Data Source.* The experimental data comes from a railway bureau in Xi'an. The experimental data are images obtained by scanning the interior of the tunnel lining with ground-penetrating radar equipment. In this experiment, two common disease images of voids and inconsistencies in the lining are selected as the research objects. Figure 1 shows two images randomly selected from the data. Figure 1(a) is a radar image with void defects, and Figure 1(b) is a radar image with imperfect defects. Images without defects are usually dense images with weak signal amplitude or even no interface signal. For imperfect defect images, the in-phase axis of the strong reflection signal at the lining interface is a diffracted arc, which is discontinuous and relatively scattered. For void defects, the lining interface has a strong reflection signal, the three-oscillation phase is obvious, and there is still a strong reflection interface signal in the lower part, and the two sets of signals have a large time path difference.

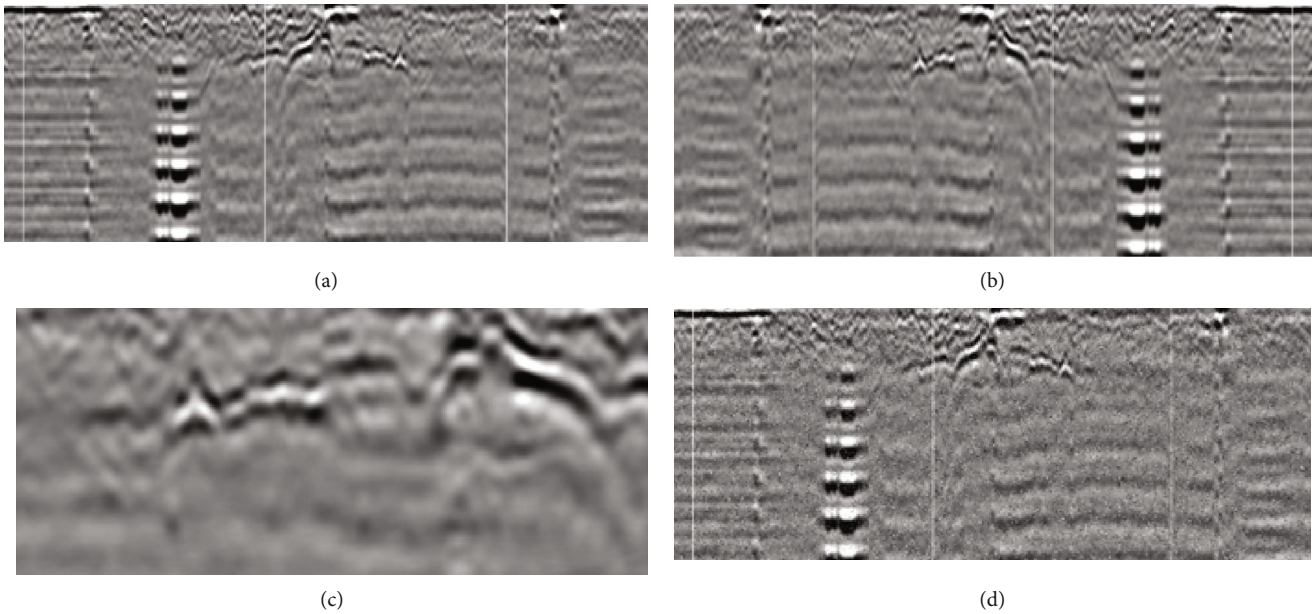
*2.2. Data Preprocessing*

(1) Data enhancement. Training deep networks in deep learning requires a lot of data. This experiment uses data enhancement methods such as cropping, mirror flipping, and noise addition to amplify data samples, as shown in Figure 2. Geometric transformation methods such as cropping and mirror flipping can



(a) (b)

Figure 1: Scanned image of ground-penetrating radar: (a) a radar image with void defects; (b) a radar image with imperfect defects.



(a) (b)

(c) (d)

Figure 2: Data enhancement methods: (a) original image; (b) mirror flip; (c) crop; (d) salt and pepper noise.

increase the space complexity of the disease. Adding salt and pepper noise to the sample can improve the robustness of the model. Using data augmentation methods to process sample data can allow the model to learn more and higher-level features of the data and improve the generalization ability of the model.

The disease image after data expansion has 1186 data samples, and the data samples are divided according to a certain proportion. Finally, the number of data of training samples and test samples is randomly selected to be 950 and 236, respectively. The classification of the data set is shown in Table 1.

(2) Batch normalization. In order to improve the training speed, the training method of deep neural networks is usually carried out in batch training. This method will cause different batches of data in the output layer of each layer of the network to have a different data distribution, which will cause the following training of a layer of network to become difficult, which makes the convergence speed of the model slow. In order to solve the problem of inconsistent input data distribution, Ding et al. [22] proposed a method of batch normalization. If a certain layer of network batch input data is $B = \{x_{1\cdots m}\}$, $x_1, x_2 \cdots x_m$ is the $m$ data in a batch, $\gamma, \beta$ are the parameter to be learned, and the output is $y_i = \{BN_{\gamma\beta}(x_i)\}$; then, the batch normalization can be performed by

Table 1: Classification and size of the data set.

| Category | Label | Training samples | Test samples | Total |
|---|---|---|---|---|
| Not dense | 0 | 470 | 117 | 587 |
| Escape | 1 | 480 | 119 | 599 |

$$\begin{cases} \mu_B \longleftarrow \dfrac{1}{m}\sum_{i=1}^{m} x_i, \\ \sigma_B^2 \longleftarrow \dfrac{1}{m}\sum_{i=1}^{m} (x_{i-}\mu_B)^2, \\ \widehat{x}_i \longleftarrow \dfrac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}, \\ y_i \longleftarrow \gamma\widehat{x}_i + \beta \equiv BN_{\gamma\beta}(x_i). \end{cases} \tag{1}$$

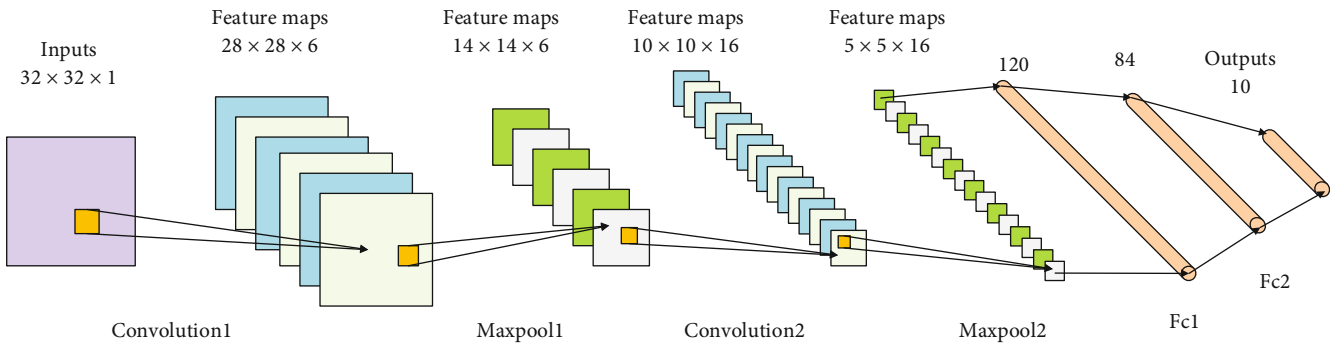



Figure 3: Convolutional neural network (CNN). The convolutional neural network is mainly composed of an input layer, two convolutional layers, two pooling layers, two fully connected layers, and an output layer.

In formula (1), $\mu_B$ and $\sigma_B^2$ are the values obtained by the network during the forward propagation process, representing the mean and variance of each dimension, respectively. The value of $\varepsilon$ in the formula is relatively small to prevent zero in the denominator. The two parameters $\gamma$ and $\beta$ are the parameters learned in the backpropagation process to adjust the variance and mean. After batch normalization, the convergence speed and accuracy of the model can be further improved.

## 3. Classification Method of Tunnel Lining Defects

*3.1. Convolutional Neural Network.* In the image recognition and target detection competitions, many network models are built on the basis of CNN [23] and have shown good performance. The CNN network contains multiple links. The main operations are carried out by convolution and pooling. The pooling layer is generally located behind the convolutional layer. The depth of the network can be achieved by superimposing the number of convolutional and pooling layers, as shown in Figure 3. Shown is a model diagram of a convolutional neural network.

Input layer: the input layer is the data entry. Generally, a series of simple preprocessing will be performed on the data before it is "feed" to the network through the input layer. The image is cropped to a suitable size, or the image is changed from multichannel to single-channel.

Convolutional layer: the constituent unit of the convolutional layer is the feature surface. The feature surface is obtained by a convolution operation at the convolution input level from a weight matrix, which is also called a convolution kernel. The role of the convolutional layer is also different, the low level is used to learn low-level features, and the high level is used to learn high-level features [24]. Wani et al. [25] also mentioned in their report (1) increasing the depth of the CNN network helps to improve the accuracy of the model and (2) increasing the number of convolutional layers is better than fully connected layers. Figure 4 shows an example of the convolution operation performed by the convolution layer. In the convolution operation, the depth and width of the input feature map can be adjusted by changing the size and number of the convolution kernel. In the example in Figure 4, the stride is 1, the padding is 0, the size of the input data is $4 \times 4$, the size of the filter is $3 \times 3$, and the size of the output data is $2 \times 2$. Both the input data and the output data have shapes. For the shape of the output data, it can be calculated according to formula (2): assuming that the padding is $P$, the stride is $S$, the size of the input data is $H\,x\,W$, the size of the filter is $FH\,x\,FW$, and the size of the output data is denoted as $H_1\,x\,W_1$:

$$\begin{cases} H_1 = \dfrac{H + 2P - FH}{S} + 1, \\ W_1 = \dfrac{H + 2P - FW}{S} + 1. \end{cases} \tag{2}$$

Pooling layer: the pooling layer plays the role of secondary extraction of features. The pooling layer is also called the downsampling layer. After the pooling operation, the resolution of the feature surface will be reduced. In other words, the operation will reduce the number of neurons on the characteristic surface but does not change its depth. There are many methods of pooling. Maximum pooling and average pooling are two of the commonly used methods. Other methods include hybrid pooling and spatial pyramid methods. Boureau et al. compared and analyzed the two commonly used methods of maximum pooling and average pooling. They believe that when the classification layer selects a linear classifier, the performance of the maximum pooling method is better than the average pooling method. Figure 5 shows an example of the maximum pooling operation. There are no parameters to be learned in the pooling layer, and the number of channels does not change after the pooling operation. The data shape after pooling can be calculated by referring to formula (2).

Fully connected layer: the fully connected layer is a tiled structure and is located at the end of the network model diagram. It can integrate the local feature information learned by the previous network layer to facilitate subsequent classification tasks. Because softmax has the advantages of high accuracy and small amount of calculation, it is often used as the preferred function of classification networks. The loss function usually adopts the cross-entropy loss function.

Other layers: (1) incentive layer. Whether it is used in the convolutional layer or the fully connected layer, the excitation layer can map the output of the previous network to

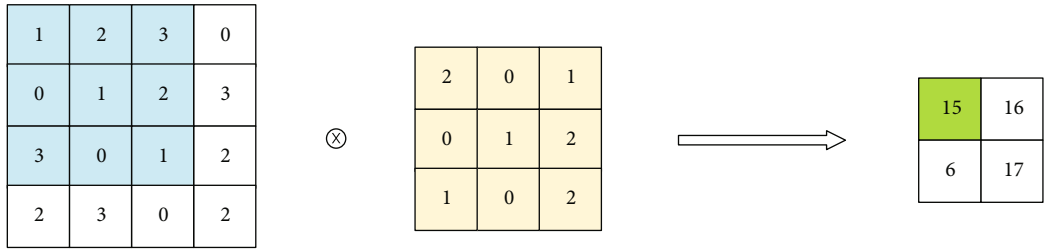



Figure 4: Example of convolution operation: in this example, "⊗" is used to indicate convolution operation. The input data size is $4 \times 4$, the size of the convolution kernel is $3 \times 3$, and the output result is obtained by convolution by the input data kernel convolution kernel.

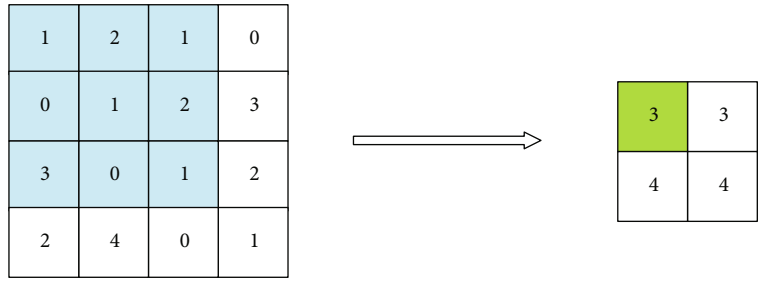



Figure 5: Pooling operation example. In this convolution example, the size of the input data is $4 \times 4$, the size of the pooling kernel is $3 \times 3$, and the size of the output data is $2 \times 2$.

Table 2: Four commonly used activation functions and their mathematical expressions in CNN.

| Function | Mathematical formula |
|---|---|
| Sigmoid | $h(x) = \dfrac{1}{1+e^{-x}}$ |
| Tanh | $h(x) = \dfrac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$ |
| ReLU | $h(x) = \max(0, x)$ |
| Leaky ReLU | $h(x) = \begin{cases} \alpha x, (x \le 0) \\ x, (x > 0) \end{cases}$ |

the nonlinear region, thereby enhancing the expressive power of the neural network. Early activation functions mostly used sigmoid. Due to the high computational cost of sigmoid, the gradient disappeared easily. Nowadays, ReLU is used more frequently. ReLU function as a nonlinear excitation function can control the gradient explosion and gradient disappearance phenomenon in the training process. Table 2 shows the commonly used activation functions.

(2) Random inactivation layer (dropout): random inactivation is a regularization method often used in fully connected layers. This technique is to randomly inactivate neurons at a certain rate. Reducing the interaction between neurons in the hidden layer can make the model more generalized and make the neurons learn more robust characteristics, and it is also an effective measure to avoid network overfitting.

### 3.2. VGG16, ResNet34 Network Model

(1) VGG16 network model. The VGG model achieved good results in the ILSVRC competition that year. VGG16 is one of the VGG models. Because of its excellent performance, it is used by many people who study image classification. Figure 6 shows the network architecture diagram of VGG16.

The VGG16 network deepens the depth of the network by superimposing the convolutional layer and the pooling layer and uses a $3 \times 3$ filter in all its layers, which can extract smaller features in the picture. The receptive field obtained by two $3 \times 3$ filters can replace the receptive field of a $5 \times 5$ filter and can reduce the number of parameters. The network model increases the receptive field by stacking multiple convolutional layers. The number of layers stacked in the VGG16 network has two and three layers. The model contains 5 pooling layers, using a $2 \times 2$ filter to perform maximum pooling, and the step size is 2. Since the categories in this study are two types, in order to reduce the number of parameters and accelerate training, the first two layers in the fully connected layer of the network are changed to 2048 channels, and the third layer is changed to 2 channels.

(2) The ResNet network refreshes the history of convolutional neural network depth, with a depth of up to 152 layers, which solves the problems of difficulty in training deep CNN models and network model degradation. ResNet34 is one type of ResNet network, and the structure is shown in Figure 7(a). This network is different from the previous deep network, not simply by superimposing the convolutional layer and the pooling layer, but introduces a residual block, the module structure diagram is shown in Figure 8(a).

The neural network extracts more features by continuously deepening the degree of the network, which is considered to be a way to improve the accuracy. However, as the

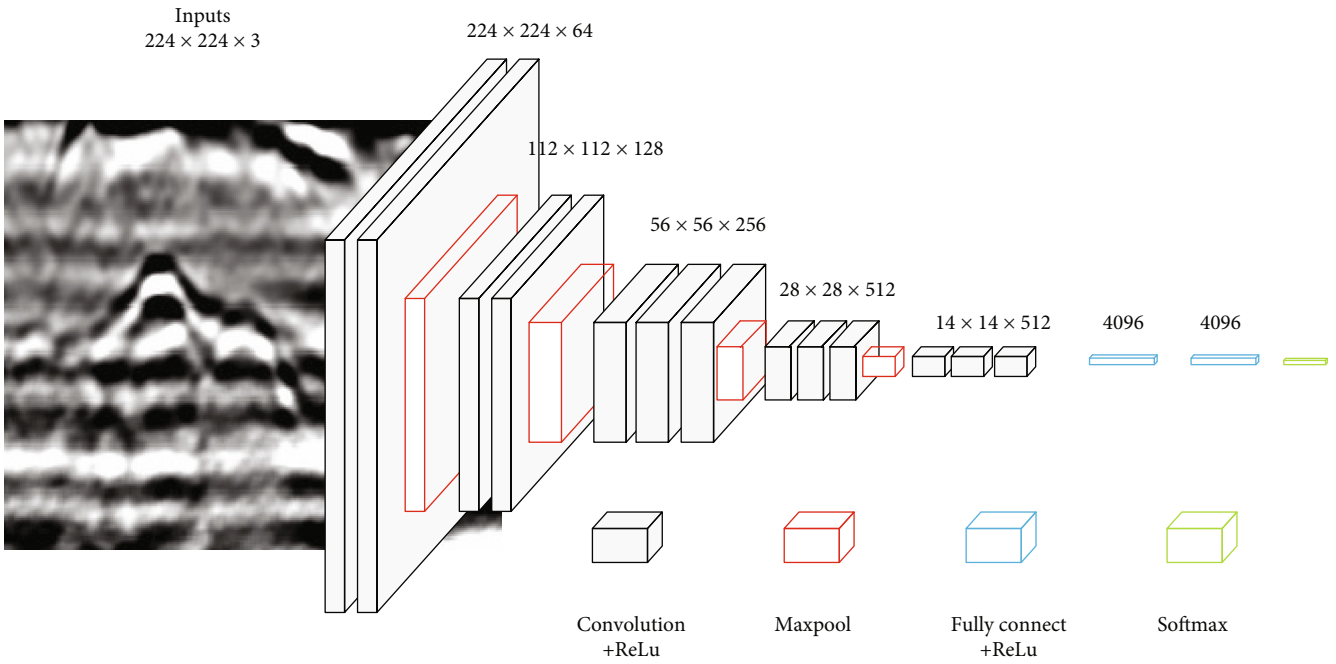



Figure 6: VGG16 network model. There are 16 layers in the network, and the squares in different colors in the figure represent the network layers with different functions.

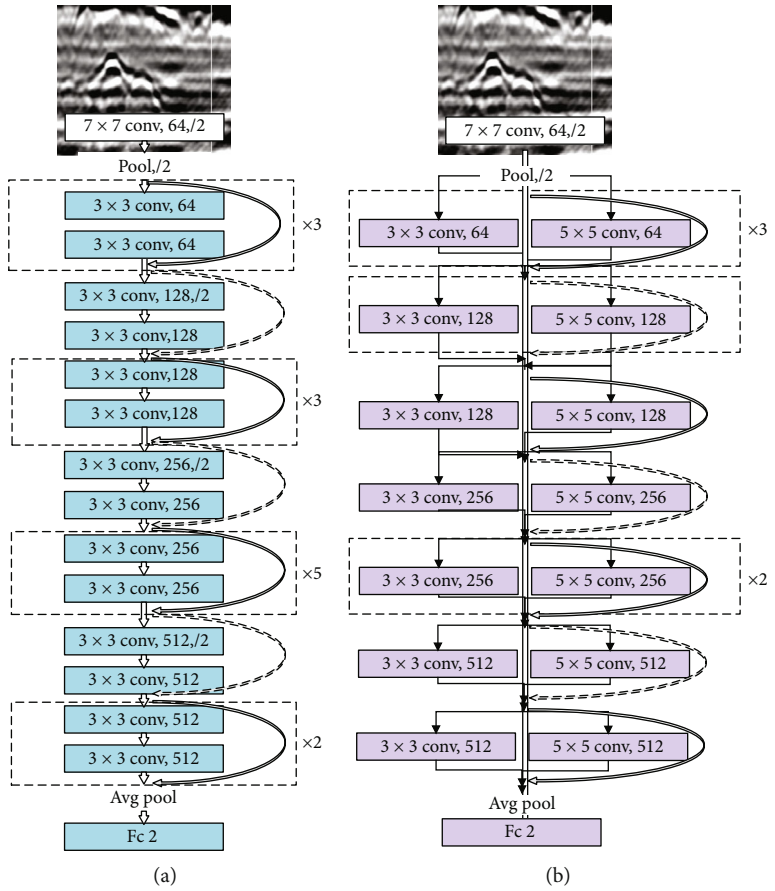



Figure 7: Network model. (a) ResNet34 network, except for the head and tail, the size of the convolution kernel of the network model is all $3 \times 3$. The 34-layer network is obtained by superimposing similar modules. (b) Improved ResNet34 network, which uses $5 \times 5$ convolution kernel in addition to $3 \times 3$ convolution kernel.

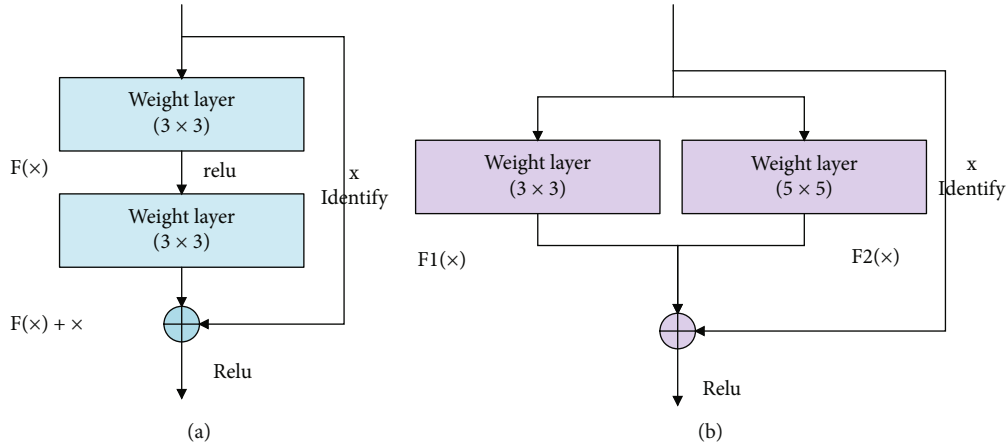



Figure 8: Residual module. (a) The original residual module, which consists of two layers of networks, the output of the first layer of network is $F(x)$, and the output of the second layer of network is $F(x) + x$. (b) The improved residual module, the network output of this module is $F1(x) + F2(x) + x$.

network deepens, the gradient disappears or the gradient explodes during the backpropagation of the network and the network is a degradation phenomenon; these factors cause the network to deepen and the accuracy cannot be improved or may even decrease.

Gradient disappearance or gradient explosion can be processed by data preprocessing and weight initialization, and network degradation can be solved by residual module. Formally use $H(x)$ to represent the desired mapping and use stacked nonlinear layers to fit the mapping $F(x) = H(x) - x$; then, the desired mapping $H(x)$ is converted to $H(x) = F(x) + x$. Assuming that optimizing the residual mapping is easier than optimizing the original mapping, then in extreme cases, pushing the residual to zero is simpler than approximating one mapping to another. Because the network has excellent performance in classification tasks, this paper chooses ResNet34 as one of the test networks to build a classification model of tunnel lining defect.

*3.3. Improved Convolutional Neural Network Model.* In the ResNet34 network, in addition to the $7 \times 7$ convolution kernel used in the initial convolution part, the other convolution kernels use $3 \times 3$ convolution kernels, and the network also continuously extracts image features by stacking similar residual modules. Based on a large number of experiments, this article improves the ResNet34 network for the image of tunnel lining defects. There are two main areas of improvement as follows:

(1) The residual module has been optimized and improved. In the improved residual module, a $5 \times 5$ size convolution kernel has been added. The $5 \times 5$ convolution operation is paralleled with the $3 \times 3$ size convolution operation, which not only increases the width of the network but also improves the adaptability to different scale defect images. The improved residual module is shown in Figure 8(b)

(2) The number of layers of the network layer has been improved. On the basis of improving residual module, the convolution operation of different network layers has been modified to reduce the parameter complexity of the model. The improved model framework is shown in Figure 7(b)

## 4. Experiments

*4.1. Training Algorithm and Hyperparameter Settings.* The training algorithm of the network model uses two algorithms, SGD and Adam, and verifies the performance of the two algorithms on ResNet34, VGG16, and the improved ResNet34 convolutional neural network.

(1) SGD and Adam algorithm. The SGD algorithm is what we often call the stochastic gradient descent method. We can understand the idea of the algorithm like this: if we want to solve the minimum value of a loss function $J(w)$, we need to give $w$ a random initial value and then follow a certain strategy that keeps changing the value of $w$ so that the function $J(w)$ keeps approaching the minimum value. Later, many update algorithms are improved by this algorithm. In the process of parameter update, the minibatch method is usually used to update the parameters. The loss of multiple samples is used as the total loss. The gradient descent method makes the network parameters along the direction of the fastest descending speed updated, so as to achieve the goal of continuously reducing the loss function $J(w)$. The update strategy of the parameter $w$ is as in

$$W \longleftarrow W - \eta \frac{\partial L}{\partial W}. \tag{3}$$

In the update strategy (formula (3)), $W$ is the weight parameter of the network, $L$ is the loss function of the network, and $\eta$ is the learning rate (usually 0.001). Through this formula, the parameters can be updated in the direction of the fastest decline.

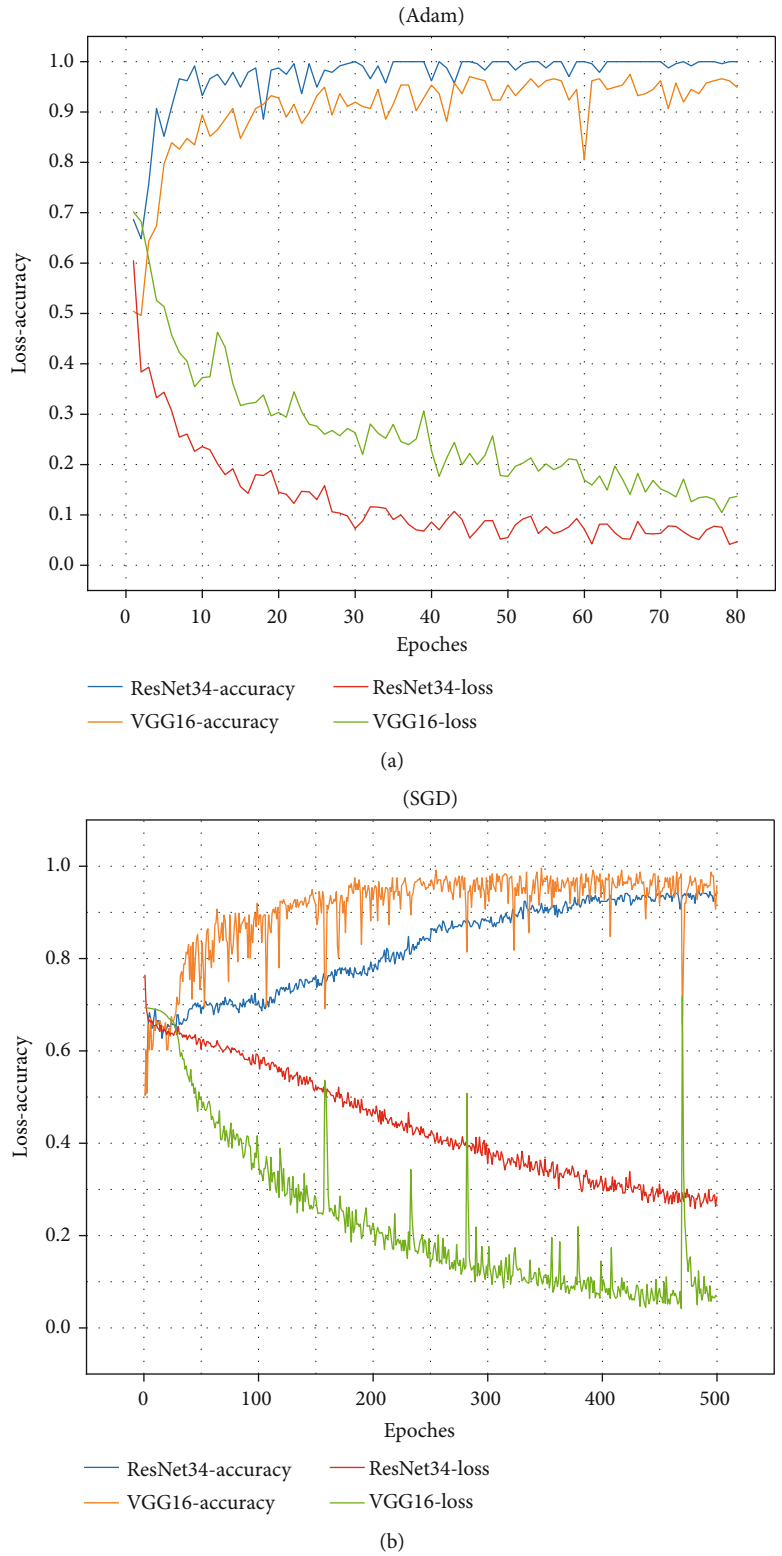



Figure 9: The loss and accuracy graphs of VGG16 and ResNet34: (a) network loss and accuracy curve diagram under the Adam algorithm; (b) graph of network loss and accuracy under the SGD algorithm.

The Adam algorithm has many similarities with the SGD algorithm, and its main purpose is to find a suitable parameter $w$ according to a certain algorithm strategy, so that $J(w)$ can be minimized. Different from the stochastic gradient descent method, the Adam algorithm automatically adjusts the learning rate in the process of updating the network parameters. The SGD learning rate is manually set. In the process of network learning, the learning rate will not

Table 3: Training results.

| Network | Algorithm | Parameter settings | Time (min) | Accuracy (%) |
|---|---|---|---|---|
| VGG16 | SGD | Momentum = 0 | 120.33 | 94.82 |
| VGG16 | Adam | $\beta_1 = 0.9$, $\beta_2 = 0.999$ | 27.09 | 94.93 |
| ResNet34 | SGD | Momentum = 0 | 83.15 | 93.58 |
| ResNet34 | Adam | $\beta_1 = 0.9$, $\beta_2 = 0.999$ | 12.35 | 99.08 |
| Improved ResNet | SGD | Momentum = 0 | 87.02 | 95.42 |
| Improved ResNet | Adam | $\beta_1 = 0.9$, $\beta_2 = 0.999$ | 13.07 | 99.25 |

change automatically. The size of the Adam learning rate is closely related to the first-order moment estimation and the second-order estimation of the gradient. This algorithm has high computational efficiency and combines the advantages of AdaGrad and RMSProp algorithms. The update strategy of the Adam algorithm is as shown in

$$\begin{cases} m \longleftarrow \beta_1 * m + (1 - \beta_1) * \dfrac{\partial L}{\partial W}, \\ v \longleftarrow \beta_2 * v + (1 - \beta_2) * \left(\dfrac{\partial L}{\partial W}\right)^2, \\ w \longleftarrow w - \eta \dfrac{m}{\sqrt{v}}. \end{cases} \tag{4}$$

In formula (4) $m$ and $v$ are the first and second moments of the gradient. $L$, $w$, and $\eta$ are the loss function, network parameters, and learning rate, respectively. $\beta_1$ and $\beta_2$ are the attenuation coefficients of the first and second moments, respectively. The two-parameter values are usually 0.9 and 0.999 by default.

(2) Hyperparameter setting. SGD: the number of training rounds (epoch) is 500, the batch size is 32, and the initial learning rate is set to 0.001. The Adam algorithm parameters are set to the following: the number of training rounds (epoch) is 80, the batch size is 32, and the initial learning rate is set to 0.001. Based on a large number of experiments, the dropout value of the VGG16 fully connected layer is set to 0.5. The activation function uses the ReLU function, and the loss function uses the cross-entropy function.

*4.2. Experiment Configuration.* The experiment was carried out on a computer with the ubuntu18.04 operating system. The computer used was Intel Xeon E5-2678 v3, 32 G random access memory, and equipped with an NVIDIA GeForce RTX 2080 (8 G graphics memory). This experiment is to train the network on the GPU [26], using the PyTorch deep learning framework. PyTorch is a deep learning framework developed by the Facebook artificial intelligence laboratory. Its interface is very easy to use, and the speed of the model is also excellent. In the process of reading the data, the transform module in PyTorch is used to cut the data, and at the same time, the data is standardized. Use dataset to load image data. After reading the image data, use the parallel computing capability of GPU to train the network by batch training. The last classification layer of the network structure uses the softmax classifier and uses the cross-entropy function as the loss function to evaluate the gap between the predicted value and the true value.

*4.3. Result Analysis*

(1) Comparative analysis of the experimental results of VGG16 and ResNet34. The training results are shown in Figure 9 and Table 3. Whether from Figure 9(a) or Figure 9(b), it can be seen that during the training process, the network loss and the test accuracy rate as a whole show a downward and upward trend, respectively. It shows that the network is always learning. Comparing Figure 9(a) and Figure 9(b), we can see that the Adam algorithm performs better than the SGD algorithm, has a higher accuracy rate, and can avoid the problem of large fluctuations in the loss function during the parameter update process. The Adam algorithm can automatically adjust the learning rate of the algorithm. Based on the relevant information in the figure and table, we can conclude that the ResNet34 network converges faster than the VGG16 network, has higher accuracy, and trains the network faster. The ResNet34 network with the Adam training algorithm can achieve 99.08% recognition. Therefore, in this experiment, the ResNet34 network can be considered as an alternative to constructing the tunnel lining defect network.

(2) Comparative analysis of ResNet34 and improved ResNet34 network. The experimental results of the two networks are shown in Figure 10 and Table 3. As can be seen from Figure 10(a), the recognition effect of the two networks is very good when the Adam algorithm is used. It can be seen from Table 3 that the improved network has a 0.17% improvement in recognition rate, and it is the network with the highest recognition rate among all the networks listed in Table 3, which can reach a recognition rate of 99.25%. The convergence speed of the two networks on the SGD algorithm is not as fast as the Adam algorithm, but the improved Resnet34

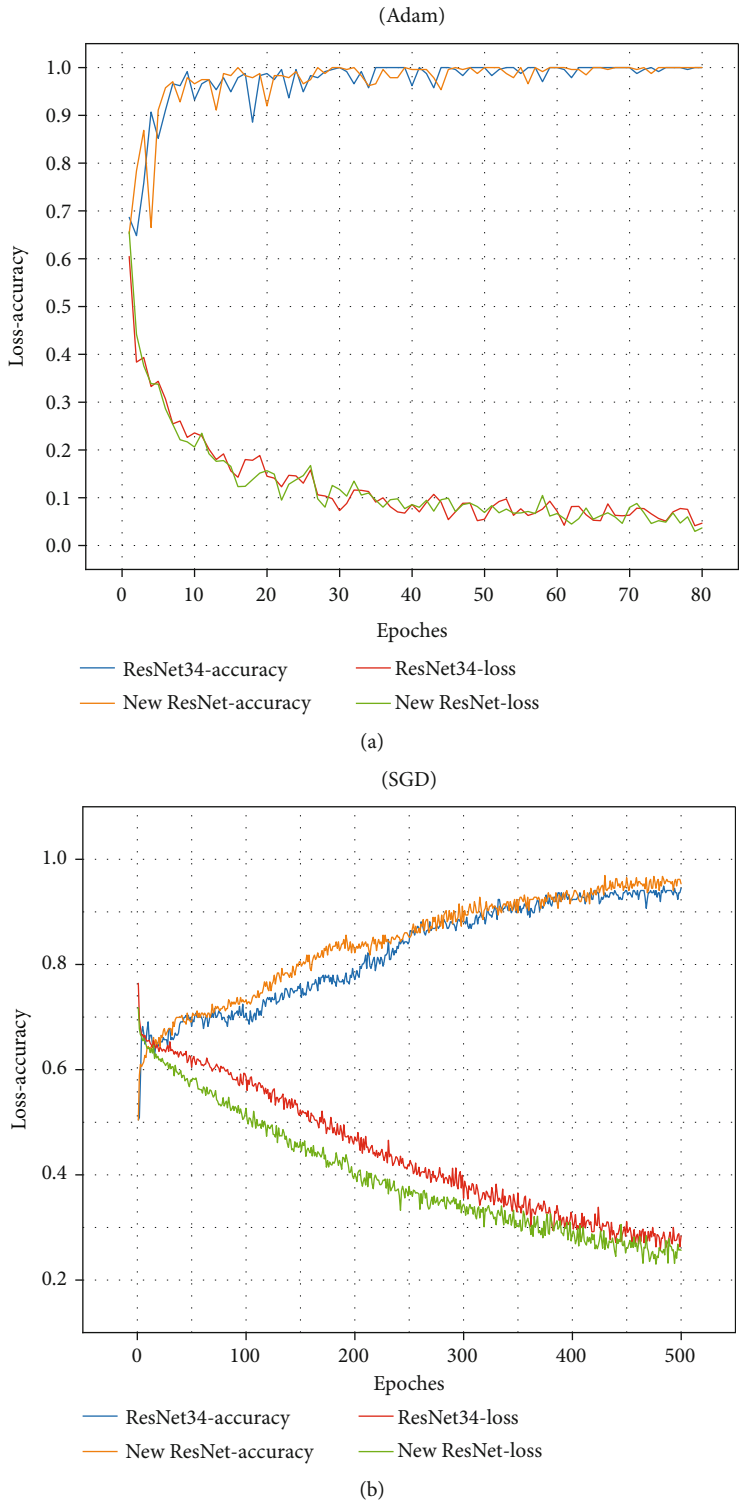



Figure 10: The loss and accuracy graphs of ResNet34 and improved ResNet34 (New ResNet): (a) network loss and accuracy curve diagram under the Adam algorithm; (b) graph of network loss and accuracy under the SGD algorithm.

has a greater improvement in the recognition rate compared to the ResNet34, and the recognition accuracy is increased by 1.84%. In general, the improved network is better in the accuracy of lining defect recognition. In addition to some improvements in recognition accuracy, the improved network has fewer model parameters and lower computational complexity and occupies less computer resources. The model parameters of the two models are shown in Table 4. Compared with the

Table 4: Parameters of different models.

| Network | Number of parameters |
|---|---|
| ResNet34 | 21,285,698 |
| Improved ResNet34 | 20,386,626 |

original network, the improved network reduces the parameter amount by 4.22%, which can reduce the training cost of the network to a certain extent.

## 5. Conclusions

In this paper, in order to automatically identify the defects in tunnel lining, a method based on deep learning is proposed. For this purpose, we established a defect image dataset and conducted theoretical research and experimental analysis on the existing VGG16 and ResNet34 network models. At the same time, this paper proposes an improved network based on the ResNet34 network, which further improves the accuracy of the recognition of lining defect images. The contribution of this article mainly includes the following aspects:

(1) This paper uses data enhancement and data preprocessing to process the lining defect image, effectively avoiding the network overfitting problem. The former is mainly used to increase the number of training samples, and the latter is used to adjust the data distribution in the network learning process. They greatly improve the learning effect of the network model

(2) This paper uses the deep learning method to identify the defects of the tunnel lining and verifies it by building an existing network model. The experimental results show the effectiveness and feasibility of the deep learning method in the identification of tunnel lining. It provides a new idea and new method for the detection of the interior of the tunnel lining

(3) Based on the existing network model, an improved network model is proposed. Compared with the original model, this model not only improves the recognition rate of lining defects but also reduces the number of model parameters

This article uses deep learning methods and improved network models to achieve very good results in the identification of tunnel lining defects. Since this research uses deep learning technology, the more data, the better the recognition effect. The more target types, the better the advantages of deep learning technology can be used. As there are relatively few data related to this research direction, collecting more data is one of the next steps in research work.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no competing interest.

## Acknowledgments

This study is supported by the Stable Supporting Fund of Acoustic Science and Technology Laboratory (JCKY2020604SSJS015).

## References

[1] M. Ukai, "Tunnel lining crack detection method by means of deep learning," *Quarterly Report of RTRI*, vol. 60, no. 1, pp. 33–39, 2019.

[2] Q. Ai, Y. Yuan, and X. Bi, "Acquiring sectional profile of metro tunnels using charge-coupled device cameras," *Structure and Infrastructure Engineering*, vol. 12, no. 9, pp. 1065–1075, 2016.

[3] J. Chen, M. Zhou, D. Zhang, H. Huang, and F. Zhang, "Quantification of water inflow in rock tunnel faces via convolutional neural network approach," *Automation in Construction*, vol. 123, p. 103526, 2021.

[4] Z. Gong and H. Zhang, "Research on GPR image recognition based on deep learning," *MATEC Web of Conferences*, vol. 309, p. 03027, 2020.

[5] Y. Liu, E. Yang, and S. Liu, "Detection of railway tunnel lining based on adaptive background learning," in *2020 15th IEEE International Conference on Signal Processing (ICSP)*, vol. 1, Beijing, China, 2020.

[6] Y. Dong, J. Wang, Z. Wang et al., "A deep-learning-based multiple defect detection method for tunnel lining damages," *IEEE Access*, vol. 7, pp. 182643–182657, 2019.

[7] L. Lv, J. Chen, Z. Zhang, B. Wang, and L. Zhang, "A numerical solution of a class of periodic coupled matrix equations," *Journal of the Franklin Institute*, vol. 358, no. 3, 2021.

[8] D. Han, J. Chen, L. Zhang, Y. Shen, Y. Gao, and X. Wang, "A deletable and modifiable blockchain scheme based on record verification trees and the multisignature mechanism," *CMES-Computer Modeling in Engineering & Sciences*, vol. 128, no. 1, pp. 223–245, 2021.

[9] L. Lv, C. Zheng, L. Zhang et al., "Contract and Lyapunov optimization based load scheduling and energy management for UAV charging stations," *IEEE Transactions on Green Communications and Networking*, vol. 5, 2021.

[10] L. Zhang, S. Tang, and L. Lv, "An finite iterative algorithm for sloving periodic Sylvester bimatrix equations," *Journal of the Franklin Institute*, vol. 357, no. 15, pp. 10757–10772, 2020.

[11] L. Xiong, D. Zhang, and Z. Yu, "Water leakage image recognition of shield tunnel via learning deep feature representation," *Journal of Visual Communication and Image Representation*, vol. 71, p. 102708, 2020.

[12] Q. Wei, F. Shao, and J. Liu, "Research summary of convolution neural network in image recognition," in *Proceedings of the International Conference on Data Processing and Applications*, New York, NY, USA, 2018.

[13] Q. Song, Y. Wu, X. Xin et al., "Real-time tunnel crack analysis system via deep learning," *Ieee Access*, vol. 7, pp. 64186–64197, 2019.

[14] G. Decor, M. D. Bah, P. Foucher, P. Charbonnier, and F. Heitz, "Defect detection in tunnel images using random forests and

deep learning," in *10th International Conference on Pattern Recognition Systems*, pp. 1–6, Tours, France, 2019.

[15] Y. Xue, X. Cai, M. Shadabfar, H. Shao, and S. Zhang, "Deep learning-based automatic recognition of water leakage area in shield tunnel lining," *Tunnelling and Underground Space Technology*, vol. 104, article 103524, 2020.

[16] X. Wang, H. Lu, X. Wei, G. Wei, S. S. Behbahani, and T. Iseley, "Application of artificial neural network in tunnel engineering: a systematic review," *IEEE Access*, vol. 8, pp. 119527–119543, 2020.

[17] J. Y. Chen, H. W. Huang, D. M. Zhang et al., "Deep learning based weak inter-layers segmentation and measurement of rock tunnel face," in *ISRM International Symposium-EUROCK 2020*, Trondheim, 2020.

[18] L. Zhang, Z. Huang, W. Liu, Z. Guo, and Z. Zhang, "Weather radar Echo prediction method based on convolution neural network and long short-term memory networks for sustainable e-agriculture," *Journal of Cleaner Production*, vol. 298, p. 126776, 2021.

[19] L. Zhang, C. Xu, Y. Gao, Y. Han, X. du, and Z. Tian, "Improved Dota2 lineup recommendation model based on a bidirectional LSTM," *Tsinghua Science & Technology*, vol. 25, no. 6, pp. 712–720, 2020.

[20] L. Zhang, Y. Huo, Q. Ge, Y. Ma, Q. Liu, and W. Ouyang, "A privacy protection scheme for IoT big data based on time and frequency limitation," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 5545648, 10 pages, 2021.

[21] X. Yan, G. Zhou, and X. Zhao, "Method for rapid detection and treatment of cracks in tunnel lining based on deep learning," in *Health Monitoring of Structural and Biological Systems XIV*, vol. 11381, California, United States, 2020.

[22] H. Ding, X. Jiang, K. Li, H. Guo, and W. Li, "Intelligent classification method for tunnel lining cracks based on PFC-BP neural network," *Mathematical Problems in Engineering*, vol. 2020, 12 pages, 2020.

[23] R. Kaluri, "A comparative study on image segmentation techniques," *International Journal of PT*, vol. 8, no. 2, pp. 12712–12717, 2016.

[24] E. Protopapadakis, A. Voulodimos, A. Doulamis, N. Doulamis, and T. Stathaki, "Automatic crack detection for tunnel inspection using deep learning and heuristic image post-processing," *Applied Intelligence*, vol. 49, no. 7, pp. 2793–2806, 2019.

[25] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi, and N. Ismail, "Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks," in *2020 6th International Conference on Wireless and Telematics (ICWT)*, Yogyakarta, Indonesia, 2020.

[26] A. Rehman, S. U. Rehman, M. Khan, M. Alazab, and T. Reddy, "CANintelliIDS: detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-based GRU," *IEEE Transactions on Network Science and Engineering*, vol. 8, pp. 1456–1466, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 4486494, 10 pages
https://doi.org/10.1155/2021/4486494

*Research Article*

# SW-LZMA: Parallel Implementation of LZMA Based on SW26010 Many-Core Processor

**Bingzheng Li, Jinchen Xu, and Zijing Liu**

*State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China*

Correspondence should be addressed to Bingzheng Li; francisleeha@163.com

Received 12 August 2021; Revised 2 September 2021; Accepted 7 September 2021; Published 18 September 2021

Academic Editor: Thippa Reddy G



With the development of high-performance computing and big data applications, the scale of data transmitted, stored, and processed by high-performance computing cluster systems is increasing explosively. Efficient compression of large-scale data and reducing the space required for data storage and transmission is one of the keys to improving the performance of high-performance computing cluster systems. In this paper, we present SW-LZMA, a parallel design and optimization of LZMA based on the Sunway 26010 heterogeneous many-core processor. Combined with the characteristics of SW26010 processors, we analyse the storage space requirements, memory access characteristics, and hotspot functions of the LZMA algorithm and implement the thread-level parallelism of the LZMA algorithm based on Athread interface. Furthermore, we make a fine-grained layout of LDM address space to achieve DMA double buffer cyclic sliding window algorithm, which optimizes the performance of SW-LZMA. The experimental results show that compared with the serial baseline implementation of LZMA, the parallel LZMA algorithm obtains a maximum speedup ratio of 4.1 times using the Silesia corpus benchmark, while on the large-scale data set, speedup is 5.3 times.

## 1. Introduction

With the improvement of high-performance computer performance, its scale is expanding. The large-scale high-performance computing cluster system must maintain long-term and stable uninterrupted operation. The amount of data transmitted, stored, and processed is increasing, and the amount of system log data is also increasing explosively. At present and in the future, it can be predicted that the scale of social computing data and scientific computing data will continue to grow with the improvement of informatization, which brings new challenges to big data processing. Effective compression is necessary to reduce the space required for data storage, make maximum use of the limited communication bandwidth, and make the high-performance computing cluster system give full play to its efficiency. With the increasing amount of data, blockchain applications need a lot of storage space. A fast big data compression algorithm can improve the efficiency of blockchain applications [1].

In the actual application of the Internet of Things, there are obvious shortcomings, such as the limited energy and bandwidth of the sensor nodes, which brings huge challenges to the network data transmission of the Internet of Things devices. The compression algorithm is currently an important technology to reduce the amount of transmitted data. It can appropriately remove the redundancy, reduce the data storage space of the IoT, and improve the speed and success rate of data transmission of the IoT. From the point of view of the server, the rapid development of information technology, especially IoT, has brought about the explosive growth in the amount of data on the server due to the demand for big data processing. This also requires efficient compression algorithms to reduce the amount of data storage and processing of algorithms such as distributed big data processing and machine learning [2, 3]

Lossless compression algorithms have a wide variety of open-source implementations. In the Sunway TaihuLight supercomputer, the existing data compression algorithms include zlib Deflate, XZ, and LZ4. None of the compression algorithms is optimized in parallel, and only a single processor core is used for compression and decompression, while the processing performance does not have much room for

improvement. In compression algorithms, there are many problems, such as the contradiction between compression rate and storage space and poor data locality. In order to achieve an effective performance improvement, deep algorithm reconstruction and optimization must be carried out for specific high-performance processors.

Many studies have used multicore processor architecture to parallelize compression algorithms. The parallelization of BWT (burrows Wheeler transform) compression algorithm appeared earlier. Pankratius et al. [4] first proposed a parallel implementation of BWT, which obtained a linear speedup ratio and was applied to Bzip software. Pigz is a parallel version of Gzip compression algorithm, which was proposed by Gristwood et al. [5] and has been widely used, but the compression rate of this parallel algorithm is low. Patel et al. [6] used GPU to parallelize the binary tree search process of the BWT lossless compression algorithm, and the acceleration effect was significant. Wu et al. [7] studied the compression algorithm based on CUDA (compute unified device architecture) and used the block parallel strategy to optimize the LZ77 compression algorithm on the GPU. Pankratius et al. [8] use MPI (message passing interface) programming to realize the distributed MPIBZIP compression algorithm, which is suitable for distributed memory computing. Wright [9] uses MPI and pthreads programming interfaces to implement the bzip2 parallel algorithm in the distributed memory structure and the shared memory structure, respectively. Although BWT-based compression algorithms are easy to parallelize, they are not as good as LZMA (Lempel Ziv-Markov chain algorithm) in terms of compression rate. In the process of multithreading parallelization of open-source compression software such as XZ and 7zip, only the character matching core function in the LZMA algorithm is parallelized. The acceleration effect is not ideal and is limited by the number of processors [10, 11] Leavline and Singh used FPGA to accelerate the LZMA algorithm [12, 13], which can obtain a higher speedup ratio, but the application cost is higher and does not have general applicability.

In the Sunway TaihuLight supercomputer system [14], the basic unit is a computing node composed of a SW26010 many-core processor, 32 GB of memory, and other control units. The processor architecture is shown in Figure 1. Four core groups (CGs) constitute a SW26010 processor, and there are 64 computing processing elements (CPEs) plus one management processing element (MPE), totally 260 computing units in SW26010. Among them, the CPE adopts a lightweight core design, and its instruction set function is very streamlined, does not support operations such as interrupts, and only runs in user mode. Each CPE contains 16 KB instruction L1 cache and 64 KB LDM (local directive memory, on-chip local data space) and supports 256-bit SIMD operations. The CPE can share memory with the MPE and use DMA (direct memory access) to exchange data between memory and LDM. In the CPE cluster, the CPEs in the same row or column can exchange data through register communication, the maximum amount of data transmitted each time is 256 bits, and the delay is low.

Figure 2 is a memory hierarchy diagram of CPE. The slave core can read data from memory in two ways: direct register access and register LDM access. Since there is no shared cache between CPEs and MPE, the delay of direct register access reaches nearly a hundred clock cycles. One of the ways to solve the problem is to copy data to LDM for memory access through DMA to improve memory access speed. This increases the difficulty of parallel program design and requires the programmers to set up DMA scheduling strategies reasonably, so as to achieve overlap of computing and communication as much as possible and to improve parallel efficiency. Data exchange between CPEs can be carried out by register communication. The parallel program on the SW26010 processor adopts the master-slave parallel programming model. The master thread runs on MPE, and the slave threads run on CPEs. The master thread mainly completes data input, memory copy, result output, and other operations, and the slave threads mainly perform computing tasks. According to the characteristics of the master-slave parallel programming model, Sunway TaihuLight supercomputer system provides the Athread accelerated thread library, which is divided into two parts: the MPE accelerated thread library and the CPE-accelerated thread library.

The main purpose of this paper is to design an LZMA parallel algorithm for Sunway TaihuLight supercomputer system and combine the characteristics of Sunway 26010 many-core processor to reconstruct and optimize the algorithm. We present SW-LZMA that can obtain a maximum speedup ratio of 4.1 times using the Silesia corpus benchmark while on the large-scale data set, speedup is 5.3 times.

## 2. Analysis of LZMA Algorithm Based on SW26010 Processor

In this section, we mainly analyse the characteristics of the LZMA algorithm that affect the performance of the algorithm such as space requirements, memory access methods, data locality, and hotspot functions. Combined with the analysis of the key technologies of SW26010 Processor, the algorithm can be reconstructed and optimized in a targeted manner.

*2.1. LZMA Workflow.* The LZMA compression algorithm was proposed by Pavlov in 1998 [15], and its core is based on the improvement of the LZ77 compression algorithm. LZMA uses a sliding window-based dynamic dictionary compression algorithm and interval coding algorithm, which has the advantages of high compression rate, small decompression space requirement, and fast speed. Figure 3 shows the LZMA workflow, including the sliding window algorithm based on LZ77 [16] and interval encoding [17, 18] (range encoding) two-stage compression.

The LZMA supports a dictionary space of 4 KB to hundreds of MBs, which increases the compression rate and also causes its search cache space to become very large. To reduce the time required to match the longest string and quickly search for matching characters, in the implementation of the LZMA algorithm, multiple possible longest matches are stored in the Hash table, and the data structure of the Hash linked list or binary search tree is used to search data. As

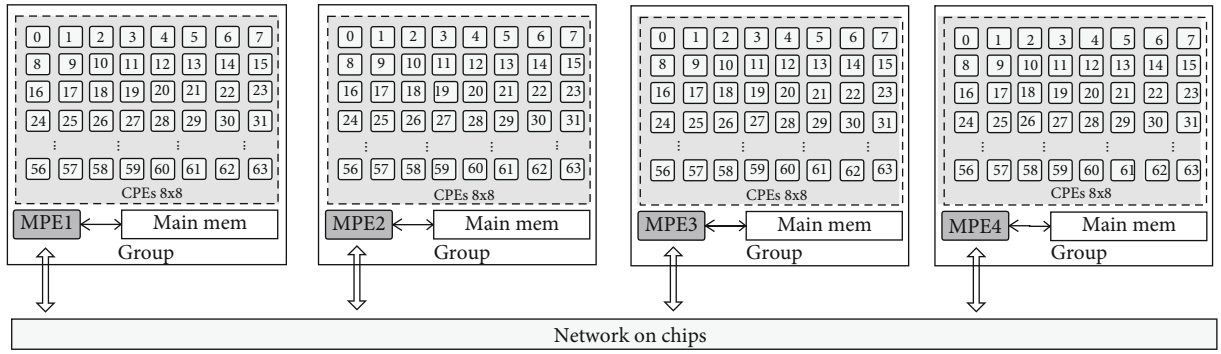



Figure 1: General architecture of SW26010 processor.

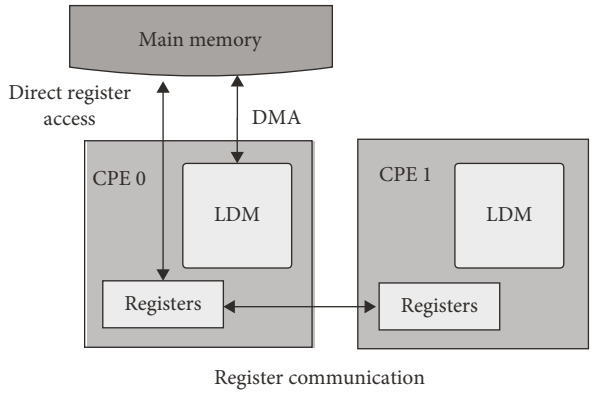



Figure 2: Computing processing element (CPE) memory hierarchy.

shown in Figure 4, in the Hash function, the hash value of the first two bytes of the search cache is used as the index of a hash array, and the hash array stores the starting position of the corresponding matching character group. The size of the hash array is a power of 2 that is half the size of the dictionary. The LZMA encoder sets up different levels of hash functions for 2, 3, and 4 adjacent bytes to achieve efficient positioning corresponding to different dictionary sizes.

*2.2. Memory Space Demand.* In the SW26010 processor, each CPE is equipped with a 64 KB LDM. In order to ensure that the CPR can obtain higher acceleration performance, it is necessary to copy the calculation data to the CPE's LDM space for memory access, which requires precise control of the use of the CPE's LDM variable memory space. Table 1 shows the usage of the local variables of the hotspot function of the LZMA algorithm, which mainly includes the local array size that takes up a large space, and the local scalar space takes up a small space and is negligible.

In the string-matching function based on the hash table, due to the large dictionary space, the hash table *hash_buf* reserves a larger hash space. This far exceeds the 64 KB LDM space of CPE and needs to be optimized to compress the use of local space. The range of the dictionary search can be reduced as much as possible within the allowable range of the compression loss, thereby reducing the size of the hash space of the hash table lookup function.

*2.3. Memory Access Characteristics.* In the LZMA algorithm, the data structure of the hash linked list is used to quickly find matching characters. Due to its relatively large search cache, its hash look-up table space has increased, with random access to memory in the range of 100 KBs to 10 MBs. At the same time, the LZ77 algorithm is based on sliding window streaming compression, because the uncoded data is continuously input, the coded data is discarded after reaching the upper limit of the search buffer space, and its data locality is poor.

Since it is impossible to prejudge the length and position of the repeated character string in the uncoded data, nor can it predict the distance of the matching character string, it is difficult to prefetch the data in the LZMA algorithm. During the compression process, the size of the dictionary gradually increases as the number of matching strings increases. Compressing the current data block depends on the dictionary obtained from the previous compression process. The LZMA algorithm has the characteristics of random access to memory and data dependence, which is a memory access-intensive algorithm. The key to its performance optimization is to combine the storage structure of the SW26010 processor to reconstruct the data structure and memory access of the algorithm to reduce memory access overhead and maximize the acceleration performance of CPEs.

*2.4. Hotspot Functions.* The main time-consuming functions of the LZMA compression algorithm are concentrated in the LZ77 string matching core function. The core function pattern matching process is shown in Algorithm 1. Among them, the time-consuming operations are mainly hash table lookup and character matching and hash table update.

In the hotspot function, the character matching process access to memory has a certain continuity, that is, starting from the current byte position, matching, and searching the same longest string in the cache. And each matched character is given by the input data. The position where the longest matching character may appear is stored in the hash table, and its look-up table access has certain randomness. In view of the characteristics of hotspot functions, there are mainly two optimization ideas. The first is to finely

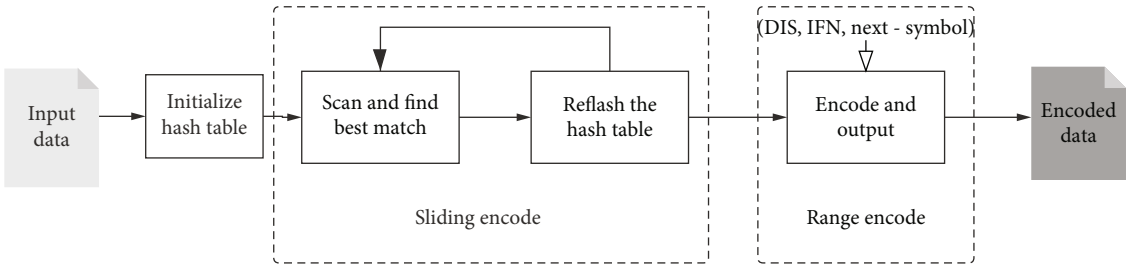



Figure 3: Flow diagrams of LZMA.

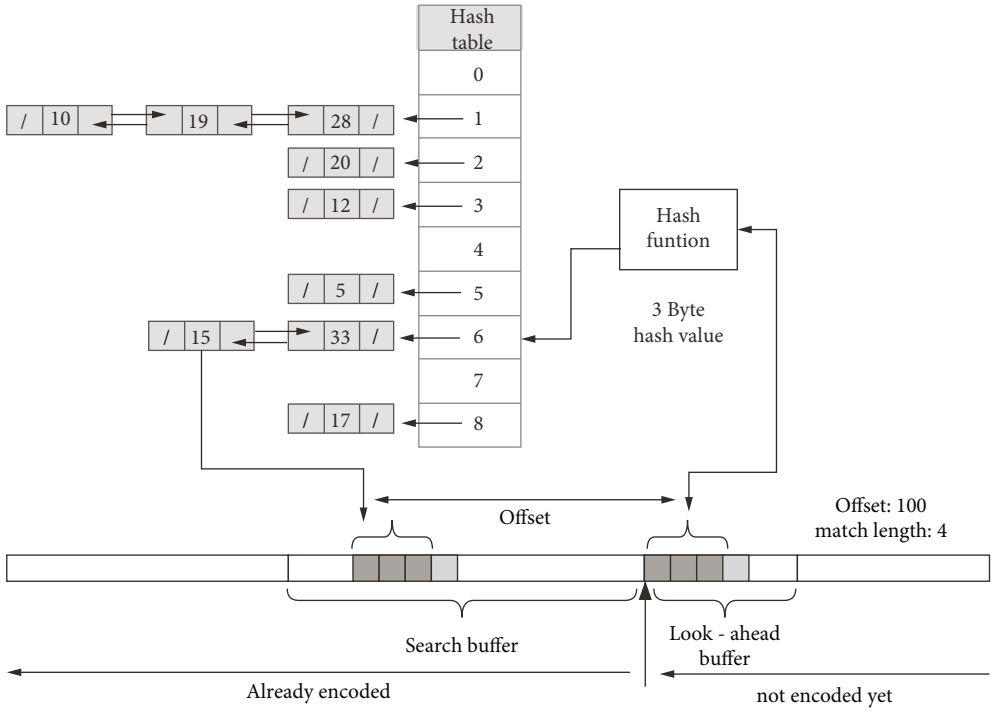



Figure 4: LZMA sliding window algorithm based on hash table.

Table 1: Local variable size of hotspot function.

| Variable | Types | Local memory size | Scope (hot spot function) |
|---|---|---|---|
| hash_buf | Int | 558 KB | Hc3Zip_MatchFinder_GetMatches<br>Hc3Zip_MatchFinder_Skip |
| data_stream | Char | 128 KB | GetOptimum |
| CRangeEnc_buf | Char | 48 KB | LitEnc_Encode<br>LenEnc_Encode |
| Match | Int | 24 KB | ReadMatchDistances<br>GetOptimum |
| isRepG0-G7 | Int | 4 KB∗8 | LzmaEnc_CodeOneBlock<br>GetOptimum |
| litProbs | Typedef struct | 16 KB | LitEnc_Matched_GetPrice<br>ReadMatchDistances<br>FillAlignPrices |
| g_FastPos | Int | 10 KB | FillDistancesPrices |

divide the space usage and focus the memory-intensive operations on LDM to the greatest extent; the second is to properly reconstruct the algorithm and replace the three-byte hash function with a two-byte hash function to further compress the use of LDM space. Both of those ways can reduce the memory access delay. At the same time, attention

**Require:**
*Hash_table*: Hash table for fast search entry
*cur_pos:* Pointer on first byte of the uncompressed data
**Process:**
1. Dictionary initialization
2. While(there are still having uncompressed data in *cur_pos*)
3. Calculate the *hash_value* of the first batch
4. If(the *hash_value* can be found in *hash_table*){
5. Update the value to the *hash_table*
6. Encode the maximum string as (*offset*, *len*, *cur*) from current position
7. }else{
8. Encode the value as (0, 0, *cur*) according to the current position
9. }
10. End while
**Output:** (*offset*, *len*, *cur*)

Algorithm 1: LZ77 compression algorithm based on sliding window.

should be paid to the LDM cache space size and the load balance of data transmission to achieve maximum hiding of computing communication.

## 3. Design and Implementation of SW-LZMA

*3.1. Parallel Design of SW-LZMA.* First, we designed the SW-LZMA multithreaded parallel algorithm on the SW26010 processor. The data to be compressed is evenly distributed to 64 CPEs cores. The CPE directly accesses the main memory to read the data to be compressed, and after adding header information to the compressed data, it directly outputs them to the main memory, and finally, the MPE writes the data blocks into the file in order. We adopted master-slave asynchronous parallelism and handed over the core computing tasks of the LZ77 compression algorithm and interval coding in the LZMA algorithm to the CPEs cluster. The MPE is only responsible for data partitioning and I/O operations. The steps of the thread parallel algorithm are as follows.

*Step 1.* Data segmentation. According to the number of CPEs, the data to be compressed are divided into several subblocks. We divide them according to the integer multiple of the memory page size. Since the amount of calculation in the compression algorithm is approximately proportional to the amount of input data, the parallel task load balance can be achieved only if the size of the divided data block is equal.

*Step 2.* Two-stage compression. Each data block is independently compressed by the CPE, including two-step compression. In the LZ77 algorithm, first, initialize the compression dictionary. As the sliding window advances, the data to be compressed continues to be input, and the dictionary size increases accordingly. Subsequently, the data structure compressed by the LZ77 algorithm is further compressed and output as the input data of interval coding.

*Step 3.* Data consolidation. After the CPEs complete the compression, the MPE is responsible for merging the compressed data. First, the MPE writes a 5 Byte header, and the content of which is compression parameter information such as dictionary size and maximum matching length. Then, each compressed block is output after adding 4 Byte header information *block_size* in the order of arrangement. The content of *block_size* is the size of the compressed data block.

*3.2. Implementation of DMA Double-Buffers.* Through the parallel method in Section 3.1, the core computing part of the serial LZMA compression algorithm can be transplanted to the CPEs cluster. However, when the CPE directly accesses the main memory, its memory access overhead will seriously reduce the performance of the parallel algorithm, and its acceleration effect is not enough to compensate for the performance loss caused by the memory access delay. In addition, in the serial version of the LZMA compression algorithm, the data to be compressed is stored in a dynamically allocated memory space, and the current compressed sliding window is determined by the address pointer. Due to the large scale of compressed data and the limited space of the CPE's LDM, even if it is divided into blocks according to thread tasks, its size is far greater than the 64 KB maximum capacity of LDM, and the data blocks to be compressed cannot be loaded all at once. Therefore, the algorithm needs to be reconstructed and optimized to compress the use of LDM space.

In order to improve the locality of data, we use the non-blocking DMA-based memory access double buffer technology based on the characteristics of the LZ77 sliding window algorithm and LDM space resources. As shown in Figure 5, the CPE does not directly access the main memory to read compressed data. Instead, the data in the current compression window and the data before and after it are transferred to the LDM buffer as a compression unit through the DMA method to achieve fast memory access. At the same time, the next compression unit has initiated DMA transfer to perform data prefetching. After the task of the current compression unit is completed, the compression calculation can be performed directly to achieve calculation and memory

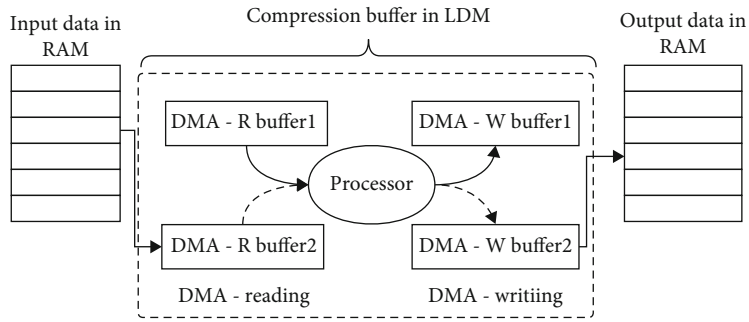



Figure 5: DMA double-buffers.

**Require:**
*Hash_table:* Hash Table for fast search entry
*cur_pos:* Pointer on first byte of the uncompressed data
*Src*: Pointer on input data (in main memory)
**Process:**

```
1.  DMA_get(buffer_base, src, buffer_size)
2.  While(src data is not empty){
3.      If(buffer block compression is finished){
4.          DMA_iget(buffer_base, src, buffer_size)
5.          DMA_iput(dest, compressed _data, compressed_size)
6.      }else{
7.          LZ77_Compress(hash_Table, cur_pos, buffer_base)
8.          Range_Encode(compressed _data)
9.      }
10.  }end while
```

**Output:**
*Dest: Compressed _data* output (in main memory)
*Compressed _data:* Output buffer (in LDM)

Algorithm 2: LZMA Athread on CPEs.

access overlap, which further reduces memory access overhead. At the same time, the output data is also buffered and copied to the memory through DMA. Algorithm 2 is an example of LZMA algorithm multithreaded parallel implementation using Athread interface.

*3.3. LDM Space Layout Optimization.* In the serial version of the LZMA algorithm, a pointer is used to directly point to the memory space of the data to be compressed, and a sliding window-based dictionary compression algorithm is implemented in the form of displacement. In the SW-LZMA algorithm, the compressed data needs to be copied to the LDM buffer area for memory access. In order to achieve DMA double buffering and make full use of the LDM space, we use manual methods for fine-grained management and allocation of the LDM address space and reconstruct the sliding window algorithm. We set up continuous double buffer space, and the pointer *buffer_base* points to the starting address of the address space, that is, the starting position of the first buffer. The pointer *buffer_middle* points to the middle position of the buffer space, that is, the starting position of the second buffer. The pointers *pos_start* and *pos_end* point to the start and end positions of the current sliding window, respectively.

At the beginning of the algorithm, as shown in Figure 6, the CPE initiates a blocking DMA request to read the data block to buffer 1, then calls the sliding window compression function, and initiates a nonblocking DMA request to read the next data block to buffer 2. When the sliding window pointer *pos_end* moves to the *buffer_middle* position, check that the nonblocking DMA request is completed, and then the compression can continue. Later, when the sliding window pointer *pos_start* moves to the *buffer_middle* position, a nonblocking DMA request is initiated to read the next data block to buffer 1. When the pointer *pos_start* and pointer *pos_end* move to the end position of the buffer, they move to the start position in a loop and continue to compress until all the data is compressed.

## 4. Evaluation

We mainly test and analyse the compression rate and compression time of the SW-LZMA algorithm. The benchmark performance is the compression ratio and compression time of the serial LZMA algorithm running on the main core. The timing method of the test is to use the Athread timing interface to count the number of CPU beats that the algorithm has run and calculate the operation time cost.

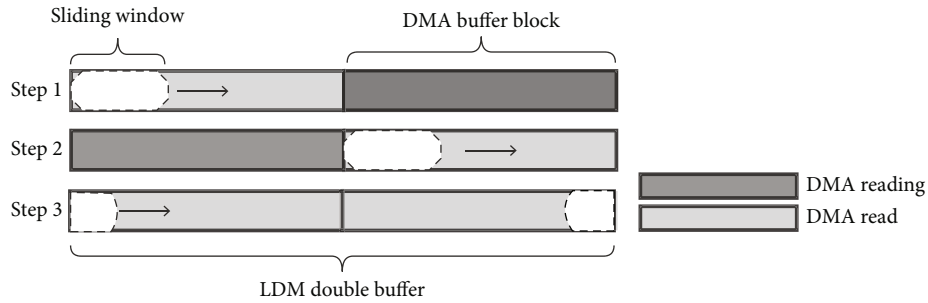



Figure 6: LDM address space partition of sliding window encoding based on DMA double buffer.

Table 2: Silesia corpus test contents.

| Filename | Description | Type | Raw size (byte) |
|---|---|---|---|
| dickens | Collected works of Charles Dickens | English text | 10192446 |
| mozilla | Tarred executables of Mozilla 1.0 (Tru64 UNIX edition) | Exe | 51220480 |
| mr | Medical magnetic resonance image | Picture | 9970564 |
| nci | Chemical database of structures | Database | 33553445 |
| ooffice | A dll from Open Office.org 1.01 | Exe | 6152192 |
| osdb | Sample database in MySQL format from open source database benchmark | Database | 10085684 |
| reymont | Text of the book by Władysław Reymont | Polish pdf | 6627202 |
| samba | Tarred source code of Samba 2-2.3 | Src | 21606400 |
| sao | The SAO star catalog | Bin data | 7251944 |
| webster | The 1913 Webster Unabridged Dictionary | HTML | 41458703 |
| xml | Collected XML files | HTML | 5345280 |
| x-ray | X-ray medical picture | Hospital image | 8474240 |

*4.1. Benchmark Corpus and Experiment Platform.* The Silesia compressed test corpus was proposed by Sebastian Deorowicz in 2003 [15], providing a file data set covering typical data types currently in use. The files' sizes are between 6 MB and 51 MB. The corpus is proposed to solve the problem of the lack of large files and single file types in the traditional Canterbury corpus. Table 2 shows the test example of the benchmark test set.

The experimental platform is the Sunway Taihulight supercomputing system, and its parameters are shown in Table 3 [18]. The compression algorithm benchmark test set used in the experiment is the Silesia corpus benchmark test set. At the same time, in order to test the compression performance of a large amount of data, we copied and packaged the Silesia corpus test set files to form GB-level data for compression testing.

*4.2. Performance Evaluation.* In order to test the acceleration effect of the SW-LZMA parallel algorithm on SW26010 processor, the serial version of the MPE LZMA algorithm was selected as a benchmark to compare the performance of different optimization schemes. The compression speed and compression rate of SW-LZMA in the Silesia corpus benchmark test are shown in Figure 7. Due to the large memory access bottleneck, the 64-thread parallel version that reads data directly from the main memory only obtains an average speedup of 2 times and even spends more time than the serial compression in some cases. In contrary, the optimized version

Table 3: Experiment environment.

| Item | Parameters |
|---|---|
| MPE | 1.45GHz, 32 KB L1 D-cache, 256 KB L2 cache |
| CPE | 1.45GHz, 64 KB LDM |
| CG | 1 MPE + 64 CPE |
| Single node | 1 CPU (4 CGs) +4∗8 GB DDR3 memory |

of communication overlaps using DMA double buffering obtained an average speedup ratio of 3.7 times and a maximum speedup ratio of 4.1 times, indicating that the parallel performance of using DMA double buffering has been greatly improved. In terms of compression ratio, the compression ratios of parallel and serial versions are basically the same.

Further analysis, we discussed the impact of the choice of single buffer size *buffer_size* on the number of message transfers and compression rate in the DMA double buffer design. As shown in Figure 8, when the *buffer_size* is less than 20 KB, due to the small amount of data copied by a single DMA, calculation and communication cannot be fully overlapped. At the same time, the number of DMA increases, the corresponding DMA overhead increases, and the compression speed decreases slightly. When the *buffer_size* is greater than 25 KB, the compression speed does not change much with the buffer size. Theoretically, the setting of the buffer should enable the DMA communication delay

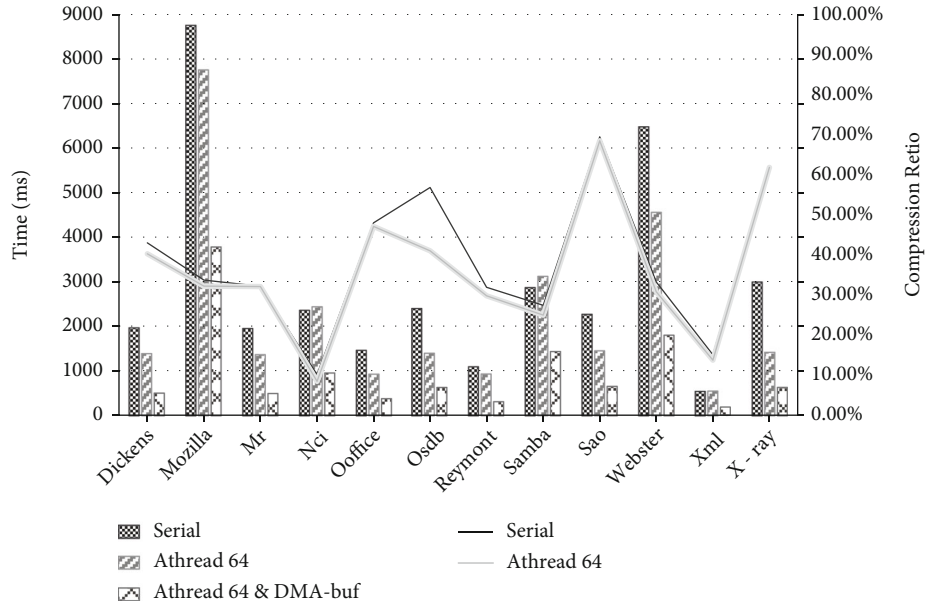



Figure 7: Test performance results of SW-LZMA.

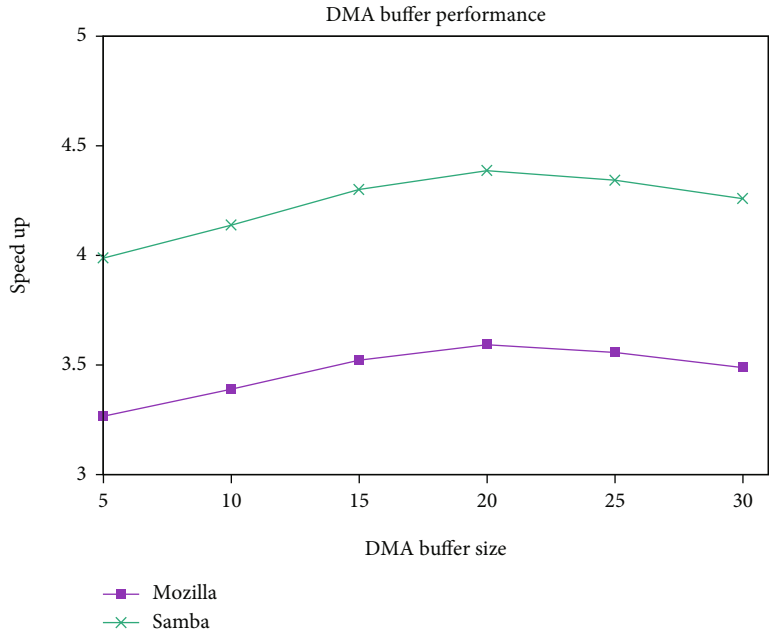



Figure 8: DMA buffer size influence on SW-LZMA performance.

and calculation to achieve load balance, but because the LDM space is limited and needs to be reserved for other local variables, the buffer cannot be expanded indefinitely.

In Section 2.2, we mainly discuss the memory space demand of the LZMA algorithm and try to satisfy it within the 64 KB LDM space of each CPE. We designed DMA double-buffers, and LDM address space partition of sliding window in Section 3 to make full use of LDM space. According to the experimental results, SW-LZMA parallel algorithm has reached the maximum utilization of the local memory space of CPEs and cannot be expanded to reach the maximum bandwidth utilization and frequency of the SW26010 processor mainly due to the LDM space limitation and memory access latency. Therefore, we take the buffer size with the best performance currently as the optimal parameter to maximize the overlap gain of computing communication optimization.

Most of the compressed test corpus data are small in scale, and no GB-level test cases are provided. We use Linux tar tool to package multiple copies of Silesia corpus to generate several large file test sets. We test the compression performance of the SW-LZMA parallel algorithm on big data based on the large file test sets. Table 4 shows the

Table 4: Large-scale data compression test results.

| File size (MB) | Serial | | Athread | | Speedup |
|---|---|---|---|---|---|
| | Time (s) | Compression ratio | Time (s) | Compression ratio | |
| 202.1387 | 34.01 | 32.96% | 15.88 | 30.24% | 2.142 |
| 404.2676 | 66.93 | 32.96% | 27.74 | 30.22% | 2.412 |
| 808.5254 | 268.55 | 32.96% | 51.48 | 30.21% | 5.216 |
| 1212.7832 | 409.78 | 33.23% | 77.11 | 30.21% | 5.314 |

Table 5: Performance test comparison between Intel x86 and Sunway 26010.

| Hardware architecture | Intel E5-1650 v2 3.5 GHz | Sunway 26010 | |
|---|---|---|---|
| Software environment | Linux 3.13, gcc 4.8.4 at -O2 level optimization, single thread | sw5CC -O2 level optimization, athread multithreads | |
| Compression corpus | Canterbury corpus | Silesia corpus | |
| Thread number | Single thread | Single thread | Athread |
| Compression ratio | 28.96% | 34.12% | 32.33% |
| Compression speed (MB/s) | 4.40 | 3.01 | 15.71 |
| Speedup | 1 | 0.68 | 3.57 |

performance comparison between the serial LZMA algorithm and the SW-LZMA parallel algorithm in large-scale data compression. It can be seen from the data table that when the data volume exceeds 500 MB, the parallel compression rate has a significant increase, with a maximum speedup of 5.3 times. The parallel compression algorithm has better adaptability in the compression of large-scale data.

*4.3. Related Work.* Alakuijala et al. used the Canterbury compression test corpus to perform performance testing and comparative analysis on compression algorithms such as Bzip2 and LZMA [19]. Their experiment platform is Intel E5-1650 v2 3.5 GHz processor. We compare it with the test results of the SW-LZMA parallel algorithm. Since the compressed data sets are different, we only compare the average compression ratio and average speedup ratio. The results are shown in Table 5. Because the CPU frequency gap is obvious, the performance of the LZMA serial algorithm on SW26010 is inferior to that on Intel CPU, and the parallel version of the LZMA has a more obvious acceleration effect than that of the Intel CPU, indicating that the SW-LZMA has better performance advantages.

## 5. Conclusions

The main work of this paper is to transplant the LZMA compression algorithm to the Sunway Taihulight supercomputer system and to reconstruct and optimize the parallel algorithm according to the characteristics of the Sunwei many-core processor. We use the Athread interface to parallelize the LZMA algorithm with multithreads and blocks and design a DMA-based double buffer mode to achieve overlap of computing communication. In further optimization, we perform fine-grained management and layout optimization on the LDM address space, set the buffer size reasonably, and obtain the best computing communication overlap effect. The test results show that in the Silesia Corpus benchmark test set, the SW-LZMA algorithm achieves a maximum speedup of 4.1 times. In the large file compression test, the SW-LZMA parallel algorithm achieved a maximum speedup of 5.1 times. Compared with mainstream CPU serial algorithms such as x86 CPU, the SW-LZMA algorithm has an obvious acceleration effect on SW26010 many-core processors, greatly reducing algorithm execution time, and has better performance. The SW-LZMA parallel algorithm not only can provide high-speed compression algorithms for applications in the field of high-performance computing but also is well known about its feasibility for more big data applications such as smart grid [20] and cloud computing [21].

In the future, there will be two research directions to further improve the performance of the LZMA algorithm: one is to upgrade the LZMA algorithm to further reduce the use of local space without affecting the compression rate; the other is to design more efficient parallel LZMA algorithm based on the new high-performance computing processors.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

## Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant no. 2018YF0804003).

## References


[1] N. Deepa, Q. V. Pham, D. C. Nguyen et al., "A survey on blockchain for big data: approaches, opportunities, and future directions," 2020, https://arxiv.org/abs/2009.00858.

[2] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[3] S. Rajadurai, M. Alazab, N. Kumar, and T. R. Gadekallu, "Latency evaluation of SDFGs on heterogeneous processors using timed automata," *IEEE Access*, vol. 8, pp. 140171–140180, 2020.

[4] V. Pankratius, A. Jannesari, and W. F. Tichy, "Parallelizing Bzip 2: a case study in multicore software engineering," *IEEE Software*, vol. 26, no. 6, pp. 70–77, 2009.

[5] T. Gristwood, P. C. Fineran, L. Everson, and G. P. Salmond, "PigZ, a TetR/AcrR family repressor, modulates secondary metabolism via the expression of a putative four-component resistance-nodulation-cell-division efflux pump, ZrpADBC, in Serratia sp. ATCC 39006," *Molecular Microbiology*, vol. 69, no. 2, pp. 418–435, 2008.

[6] R. A. Patel, Y. Zhang, J. Mak, A. Davidson, and J. D. Owens, "Parallel lossless data compression on the GPU," in *2012 Innovative Parallel Computing (InPar)*, San Jose, CA, USA, 2012.

[7] L. Wu, M. Storus, and D. Cross, *CUDA WUDA SHUDA: CUDA Compression Projects*, Stanford University, 2009.

[8] V. Pankratius, A. Jannesari, and W. F. Tichy, "Parallelizing Bzip2: a case study in multicore software engineering," *IEEE Software*, vol. 26, no. 6, pp. 70–77, 2009.

[9] C. Wright, "Hybrid programming fun: making bzip2 parallel with MPICH2 & Pthreads on the Cray XD1," in *Proceedings of the 48th Cray User Group meeting. CUG'06*, pp. 78–84, Lugano, Switzerland, 2006.

[10] S. D. Agostino, "Lempel–Ziv data compression on parallel and distributed systems," *Algorithms*, vol. 4, no. 3, pp. 183–199, 2011.

[11] X. Wang, L. Gan, J. Xu et al., "PLZMA: a parallel data compression method for cloud computing," in *Algorithms and architectures for parallel processing: 18th international conference, ICA3PP 2018*, pp. 504–518, Guangzhou, China, 2018.

[12] E. J. Leavline and D. Singh, "Hardware implementation of LZMA data compression algorithm," *International Journal of Applied Information Systems (IJAIS)*, vol. 5, no. 4, pp. 51–56, 2013.

[13] B. Li, L. Zhang, Z. Shang, and Q. Dong, "Implementation of LZMA compression algorithm on FPGA," *Electronics Letters*, vol. 50, no. 21, pp. 1522–1524, 2014.

[14] H. Fu, J. Liao, J. Yang et al., "The Sunway TaihuLight supercomputer: system and applications," *Science China Information Sciences*, vol. 59, no. 7, 2016.

[15] D. Salomon, *Data Compression: The Complete Reference*, Springer-Verlag New York, Inc., 2000.

[16] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 337–343, 1977.

[17] G. Martin, "Range encoding: an algorithm for removing redundancy from a digitised message," in *Proceedings of the Conference on Video and Data Recording*, pp. 24–27, Southampton, 1979.

[18] S. Deorowicz, *Universal Lossless Data Compression Algorithms*, Silesian University of Technology, 2003.

[19] J. Alakuijala, E. Kliuchnikov, Z. Szabadka, and L. Vandevenne, *Comparison of Brotli, Deflate, Zopfli, Lzma, Lzham and Bzip2 Compression Algorithms*, Google Inc, 2015.

[20] M. Alazab, S. Khan, S. S. R. Krishnan, Q. V. Pham, M. P. K. Reddy, and T. R. Gadekallu, "A multidirectional LSTM model for predicting the stability of a smart grid," *IEEE Access*, vol. 8, pp. 85454–85463, 2020.

[21] S. Senthilkumar, N. Kryvinska, S. Bhattacharya, and G. Reddy Bojja, *SCB-HC-ECC Based Privacy Safeguard Protocol For Secure Cloud Storage Of Smart Card Based Health Care System*, Frontiers in Public Health, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 2953827, 8 pages
https://doi.org/10.1155/2021/2953827

*Research Article*

# Optimization Method of Integrated Light-Screen Array with External Parameters Based on Genetic Algorithm

**Rui Chen, BoWen Ji, Ding Chen, and ChenXi Duan**

*School of Optoelectronic Engineering, Xi'an Technological University, Shaanxi, Xi'an 710032, China*

Correspondence should be addressed to Rui Chen; chenrui_xatu@163.com

Received 30 July 2021; Accepted 25 August 2021; Published 11 September 2021

Academic Editor: Thippa Reddy G



Due to the high sensitivity and fast response, the light-screen array measurement principle is suitable for the dynamic parameter measurement of small and fast targets including projectile. Since the spatial structures of the light-screen array determine the measurement accuracy, internal parameters such as the angles between the light-screens are usually calibrated and then directly used in the field. However, the effect of the measuring state is ignored in the test field. This paper takes the integrated light-screen array sky vertical target as the research object, and two rotation angles are introduced as external parameters to describe the deviation between the calibration state and measuring state of the target, so as to optimize the measurement model. Aiming at the problem that the external parameters cannot be measured directly, an external parameter inversion method of machine learning based on a genetic algorithm is designed under a complex engineering model. The deviation between the projectile hole and the light-screen array measurement coordinates is used to build an inversion database for the genetic algorithm during the machine learning process. The simulation and the live firing test show that the optimization method and parameter identification algorithm in this paper can optimize the measurement model and improve the measurement accuracy of the light-screen array principle directly and can also provide a reference for the optimization and parameter identification in other engineering problems.

## 1. Introduction

The flight parameters of some small moving targets, such as projectile, bullet, sport ball, and miniature unmanned aerial vehicle, attract much interest in many related fields. For example, it can help athletes improve their competitive performance, and it can also make the missions more smoothly completed for miniature unmanned aerial vehicles. Particularly, the measurement of projectile flight parameters is a highly concerning issue in the process of weapon development and production all over the world [1, 2].

The current measurement methods are mainly based on the acoustic principle [3], radar principle [4], CCD intersection principle [5], and light-screen array principle [6, 7]. Due to the fact that the acoustic principle cannot apply to subsonic projectiles, the radar principle is susceptible to electromagnetic interference, and the CCD intersection principle is expensive and difficult to operate, compared with other principles, the light-screen array principle has many advantages of high sensitivity, greater range of velocity measurement, fast response, and easy operation [8], and it can meet the measurement requirements of different characteristics of projectiles emitted by different types of barrel weapons, such as subsonic weapons, small minor-caliber weapons, and rapid-fire weapons, so it is necessary to improve its performance.

The measurement accuracy of the light-screen array depends on the precision of the description of the light-screen spatial structure [9, 10]. At present, the most effective method to obtain these structural parameters is calibration [11]. However, due to the influence factors such as environmental and human factors, the calibration state is difficult to be accurately reproduced in the test field [12], which leads to the structure deviation of the light-screen array; hence, the measurement model is inconsistent with the reality and the measurement accuracy is reduced. More importantly, the structure deviation is difficult to be directly measured due to the lack of reference, so they are ignored in the test field.

Therefore, research on model optimization and the method to identify the key parameters of the optimized model is necessary.

In recent years, big data has generated strong interest in various scientific and engineering domains over the last few years, which not only is being widely used in the field of healthcare, production, sales, IoT devices, Web, organizations, etc., but also plays an obvious role in the field of engineering, such as data analysis, calculation model parameter acquisition, and intelligent algorithm optimization [13]. And various algorithms have been applied to solve problems in these complex engineering fields, such as an advancing coupled multistable stochastic resonance method being realized, which considers the signal-to-noise ratio as the fitness function to optimize and determine the system parameters, so that the faults of motor bearings are detected in the field of engineering rotating machinery [14], using the cooperative coevolution framework simplifying the algorithm process by dividing the largescale and high-dimensional complex optimization problem into several low-dimensional optimization subproblems [15] and using an improved quantum-inspired cooperative coevolution algorithm based on combining the strategies of cooperative coevolution to optimize the airport gate allocation method and realized to effectively allocate airport gates to the flights [16].

In this paper, the parameters in the current measurement model are defined as internal parameters, and the external parameters from influence factors in the test field are introduced to improve the measurement model; furthermore, the inversion algorithm based on genetic algorithm is studied. The proposed algorithm establishes the objective function through the inversion idea; after the data dimension is reduced, the external parameters are acquired from the shooting database by machine learning based on genetic algorithm in reproduction. Comparing with the current light-screen parameter calibration method, the parameter acquisition does not depend on special measurement experiments, so it is more applicable. Finally, the method is verified by simulation and live firing tests, and the results of this paper can be directly used to optimize the measurement model and improve the measurement accuracy of photoelectric detection, and the algorithm can also provide a reference for the identification of model parameters in other engineering problems.

## 2. Measurement Principle and System Parameters

*2.1. Measurement Principle and Internal Parameters.* The basic measurement principle of the light-screen array is to form a specific shape (such as double N shape, double V shape, and double parallel shape) of the light-screen array in space by arranging the light-screens. After shooting, the time of the projectile arriving at each light-screen can be recorded; then, combined with the structure of the light-screen array, the flight parameters of velocity and coordinates can be calculated.

The method of light-screen generating equipment is basically the same. Take the typical sky screen vertical target as an example, it consists of a lens, a slit, a photodetector, etc., and forms the light-screens of a certain shape in the space above it. For example, the double N-shaped light-screen array is composed of two single N-shaped vertical targets, as shown in Figure 1.

In Figure 1, in order to constrain the light-screen structure, target I and target II are fixed on the pedestal after adjusting and aligning by the laser to form the integrated light-screen array. The trajectory direction is set as the $X$ -axis; then, the system of coordinate $OXYZ$ is established, and the projection view of the light-screen structure is shown in Figure 2.

The structure parameters of the light-screen array can be described by the vertical angles between light-screen planes $\alpha = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]$ and the horizontal angles between light-screen planes $\beta = [\beta_1, \beta_2]$. And $s_1$ is the distance between target I and target II. The current measurement model is fully described by these parameters above, which are collectively referred to as internal parameters in this paper.

The structure description of the double N-shaped light-screen array is shown in

$$A \cdot I = B, \tag{1}$$

where

$$A = \begin{bmatrix} \cos\alpha_1 & \sin\alpha_1 & 0 \\ \cos\beta_1 & 0 & \sin\beta_1 \\ \cos\alpha_2 & -\sin\alpha_2 & 0 \\ \cos\alpha_3 & \sin\alpha_3 & 0 \\ \cos\beta_2 & 0 & \sin\beta_2 \\ \cos\alpha_4 & -\sin\alpha_4 & 0 \end{bmatrix},$$

$$I = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{2}$$

$$B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ s_1 \cdot a_1 \\ s_1 \cdot a_2 \\ s_1 \cdot a_3 \end{bmatrix}.$$

The projectile flight velocity and the coordinates obtained can be expressed as follows in the measurement coordinate system $OXYZ$ by ignoring the gravity and air resistance:

$$X = M^{-1} \cdot N, \tag{3}$$

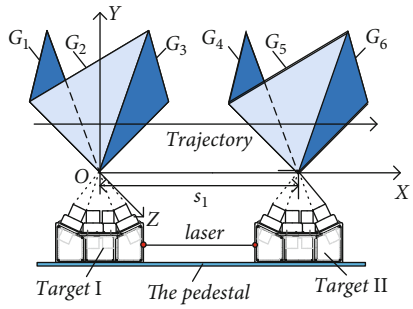



Figure 1: Measurement principle of double N shape light-screen array.

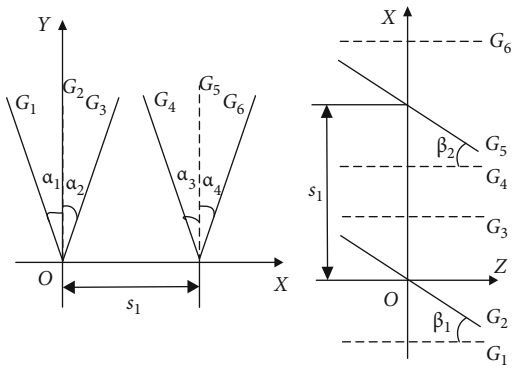



Figure 2: Projection view of light-screen and internal parameters.

where

$$M = \begin{bmatrix} \cos\alpha_1 & \sin\alpha_1 & 0 & 0 & 0 & 0 \\ \cos\beta_1 & 0 & \sin\beta_1 & \cos\beta_1 \cdot t_2 & 0 & \sin\beta_1 \cdot t_2 \\ \cos\alpha_2 & -\sin\alpha_2 & 0 & \cos\alpha_2 \cdot t_3 & -\sin\alpha_2 \cdot t_3 & 0 \\ \cos\alpha_3 & \sin\alpha_3 & 0 & \cos\alpha_3 \cdot t_4 & \sin\alpha_3 \cdot t_4 & 0 \\ \cos\beta_2 & 0 & \sin\beta_2 & \cos\beta_2 \cdot t_5 & 0 & \sin\beta_2 \cdot t_5 \\ \cos\alpha_4 & -\sin\alpha_4 & 0 & \cos\alpha_4 \cdot t_6 & -\sin\alpha_4 \cdot t_6 & 0 \end{bmatrix},$$

$$X = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ v_{x1} \\ v_{y1} \\ v_{z1} \end{bmatrix},$$

$$N = \begin{bmatrix} 0 \\ 0 \\ 0 \\ s_1 \cos\alpha_3 \\ s_1 \cos\beta_2 \\ s_1 \cos\alpha_4 \end{bmatrix}. \tag{4}$$

The internal parameters of light-screen angles $[\alpha_1, \alpha_2, \alpha_3, \alpha_4]$, $[\beta_1, \beta_2]$, and target distance $s_1$ are all constant values, which can be obtained by the calibration. Time series $t_i$ $(i = 1, 2, 3, \cdots, 6)$ indicates the time when a projectile arrives at each light-screen, respectively, which can be obtained by the chronograph.

*2.2. External Parameters and Model Optimization.* It is obvious that the structure of the light-screen array directly affects the measurement accuracy. However, it is difficult for the vertical target to reproduce its calibration state in the test field, and the influence is mainly manifested in Figure 3.

In Figure 3, the vertical target is not completely horizontal after placement in the test field, which causes the entire light-screen array to rotate angle $\tau$ along the $Z$-axis (as shown in Figure 3(a) and angle $\omega$ along the $X$-axis (as shown in Figure 3(b). These two angles are mutually independent, since they are orthogonally decomposed. The whole structure of the light-screen array is changed by these two angles which are called external parameters in this paper.

The measurement model is optimized by introducing external parameters into the measurement model shown in Formula (1), which is closer to the measuring state, as shown in

$$A \cdot I_\tau = B, \tag{5}$$

$$A \cdot I_\omega = B, \tag{6}$$

where

$$I_\tau = \begin{bmatrix} \cos\tau & \sin\tau & 0 \\ -\sin\tau & \cos\tau & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot I,$$

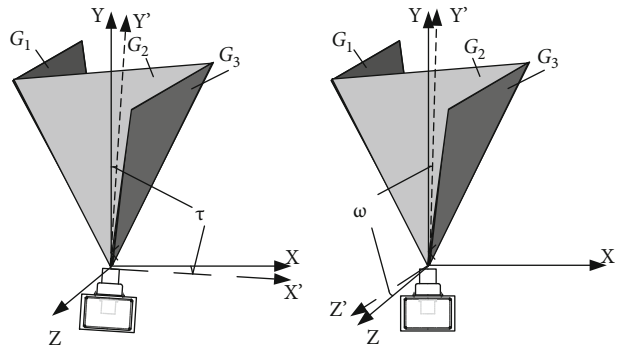
$$I_\omega = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\omega & \sin\omega \\ 0 & -\sin\omega & \cos\omega \end{bmatrix} \cdot I. \tag{7}$$

It is noteworthy that the external parameters are often ignored because they are difficult to directly measure in the test field. The external parameter inversion method of machine learning based on genetic algorithm is hereby studied.

## 3. Inversion Model and Simulation

Since there is no physical contact between the light-screen array and the projectile, a cardboard target is placed perpendicular to the trajectory direction after the vertical target; then, the coordinates of the projectile can be measured by the light-screen array and the cardboard target, simultaneously. The coordinates of the projectile hole on the cardboard target can be regarded as a quasitruth value, as shown in Figure 4.

The projectile coordinates $(z, y)$ are measured by the light-screen array and $(z*, y*)$ is the coordinates of projectile hole on the cardboard target plane. In Figure 4, $s_2$ is the distance between target II and the cardboard target, so that the coordinates can be inferred from the light-screen plane




(a) Rotate along the $Z$-axis (b) Rotate along the $X$-axis

Figure 3: The changes of light-screen array structure in measuring state.

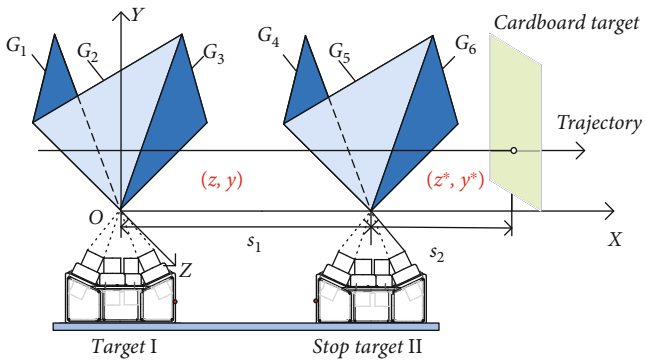



Figure 4: Structure parameter inversion model of light-screen array.

to the cardboard target plane via the flight parameters measured by the light-screen array. The simplified expression shown in Formula (8) causes the expression to be complicated to expand:

$$\begin{cases} z = f(\tau, \omega, s_{in}, t_n, s_2), \\ y = g(\tau, \omega, s_{in}, t_n, s_2). \end{cases} \tag{8}$$

In Formula (8), symbol $s_{\text{in}}$ represents the internal parameters, symbol $t_n$ $(n = 1, 2, 3, \cdots, 6)$ represents the time series of projectile arrival at each light-screen plane, and $\tau$ and $\omega$ are the external parameters to be inverted.

In the ideal case, the coordinates $(z, y)$ which are measured by the light-screen array should equal the projectile hole coordinates $(z*, y*)$ on the cardboard target plane. However, due to the influence of the external parameters and other factors, there is always some deviation between them, which represents the closeness between the measured coordinates of the light-screen array and the quasitruth value.

The optimization model contains many parameters and has a complex influence mechanism on the measurement results. Although some of the parameters are related to the results, they still can be regarded as constant values in the measurement process. These constant values have no effect on the coordinate deviation, so ignoring the influence of these constant values can reduce the dimension of the inversion parameters and reduce the algorithm burden [17].

Under the condition that other influencing factors remain unchanged, the larger the value of external parameters, the larger the deformation of the light-screen, and the larger the coordinate deviation of $(z, y)$ and $(z*, y*)$. The objective function is constructed by considering the deviation of horizontal and vertical coordinates comprehensively, which is shown as

$$\delta = \sum_{k=1}^{j} \frac{1}{j} \left( \left( f_k(\tau, \omega, s_{in}, t_n) - z_k^* \right)^2 + \left( g_k(\tau, \omega, s_{in}, t_n) - y_k^* \right)^2 \right)^{1/2}. \tag{9}$$

In order to reduce the influence of random error, the data obtained from multiple shots were used to establish the inversion database, which can be used to build a blockchain to provide services for obtaining model parameters in the algorithm below. Symbols $j$ is the total number of projectiles involved in inversion, and $k$ is the projectile serial number.

The parameters in the engineering model are decomposed according to the attributes, and the external parameters are taken as the parameters to be retrieved; then, the blockchain constructed according to the objective function serves as the inversion database in the genetic algorithm. Parameter values of the external parameter are inverted by machine learning based on a genetic algorithm.

Due to the random error being ignored and other conditions remaining unchanged, the data of blockchain is only affected by external parameters. The initial population with various characteristics is established within the range of external parameter values, and the population is reproduced by a genetic algorithm. In the breeding process, the objective function value can be calculated for each population value in the blockchain, and the population with a smaller objective function has a greater chance of survival. When the objective function is the smallest, the external parameter values are closest to their true values. The process of the inversion method of machine learning based on a genetic algorithm can be expressed in Figure 5.

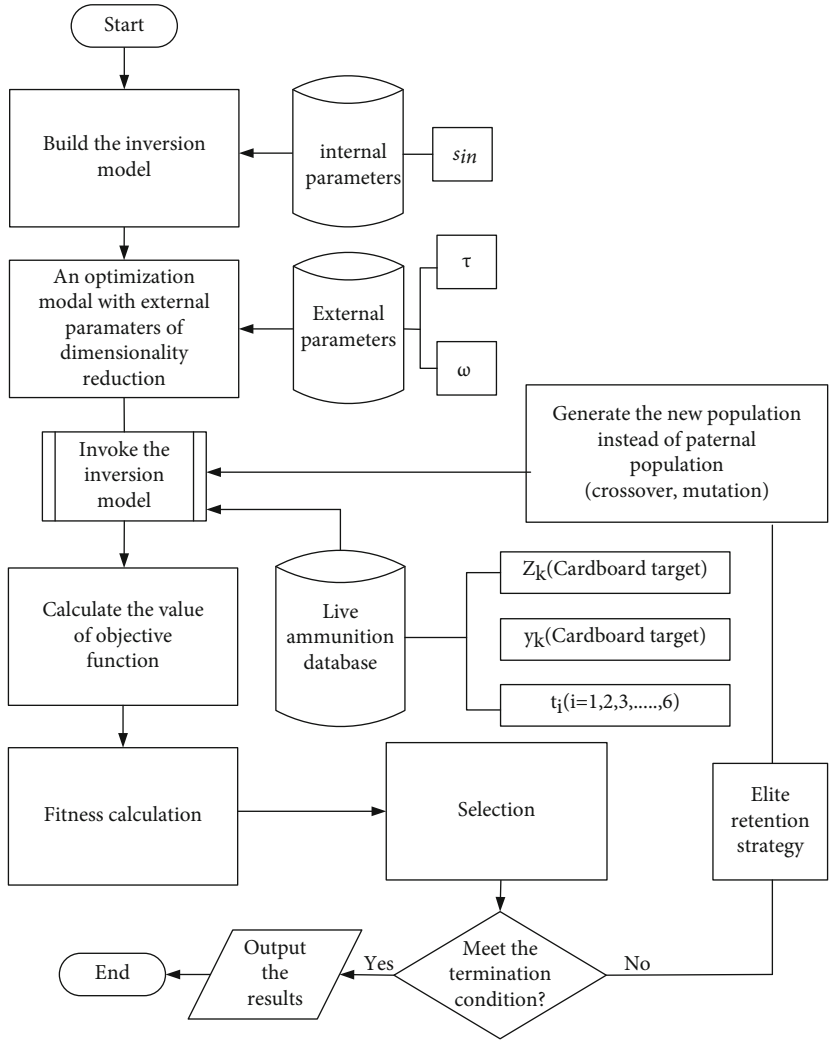



Figure 5: The diagram of machine learning based on genetic algorithm.

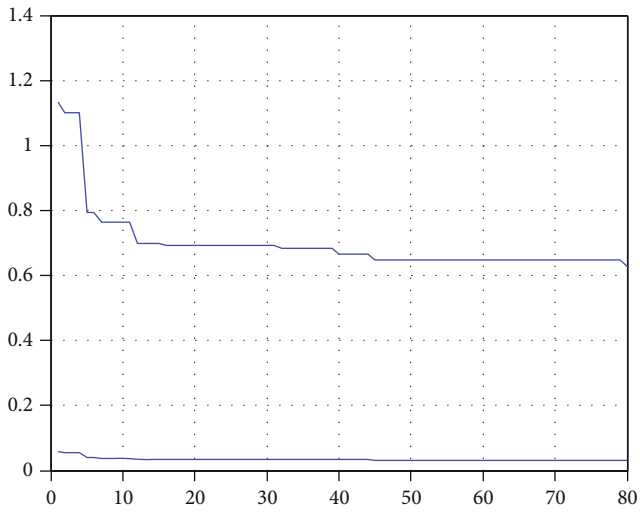



Figure 6: Simulation of genetic algorithm inversion method.

Table 1: The external parameters obtained by simulation.

| The value types | External parameters | |
|---|---|---|
| | $\tau$ (°) | $\omega$ (°) |
| Set values | 0.15 | 0.30 |
| Initial values | 0 | 0 |
| Inverse values | 0.18 | 0.26 |

Table 2: Experimental result of live firing.

| The value types | External parameters | |
|---|---|---|
| | $\tau$ (°) | $\omega$ (°) |
| Initial value | 0 | 0 |
| Inverse value | 0.05 | 0.09 |

The measurement model is optimized by introducing external parameters, and the model parameter dimension is reduced by inputting all the calibrated internal parameters and ignoring the constant parameters. Then, focus on the external parameters in which the initial value and value

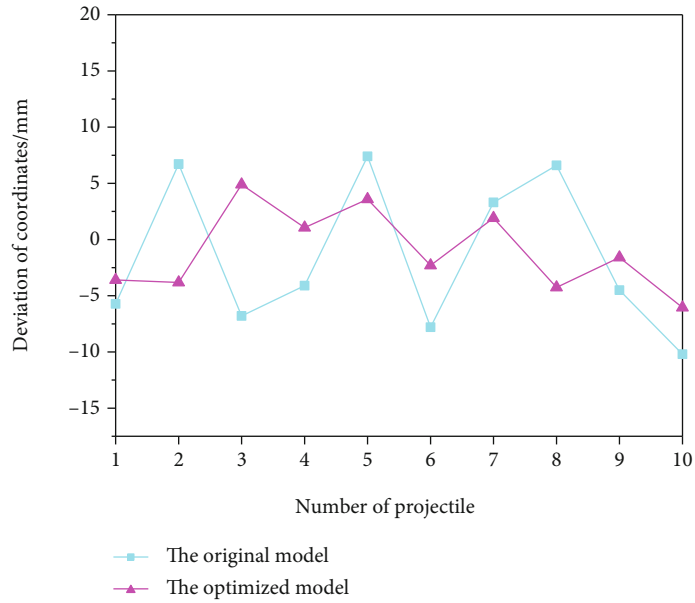



Figure 7: Distribution of coordinate deviation.

range are constraint in the test range. The constraint conditions are expressed as

$$\begin{cases} \tau = \tau_0 \pm \sigma_\tau, \\ \omega = \omega_0 \pm \sigma_\omega. \end{cases} \tag{10}$$

The shooting data is obtained by generating random numbers with the above range by simulation and using the objective function to build the blockchain. The fitness function is obtained on the basis of blockchain in the genetic algorithm section, and the elite strategy is introduced in the competition process to improve efficiency. The population gene with the small value of the objective function has a higher fitness and selected to be inherited to the next generation with a higher probability. When the objective function is less than a certain threshold and the generation number reaches the threshold value, the algorithm ends and the final inversion result is given.

The inversion model of the double N-shaped light-screen array was established in MATLAB, and the internal parameters are input according to the measured value of the field test; then, the value range of external parameters is limited to $\tau \in [-0.5^\circ, 0.5^\circ]$ and $\omega \in [-0.8^\circ, 0.8^\circ]$.

Take $j = 20$ times shot within the 500 mm × 500 mm rectangular target surface to establish the database to obtain the time series $t_i$ and target coordinates for simulation. Considering the diversity of the population and the amount of program calculation, the population number is set to 20. Take 2 significant digits to conduct binary coding according to

$$2^{m_j - 1} < (a - b) \times 10^n < 2^{m_j} - 1. \tag{11}$$

In Formula (11), $b$ and $a$ are the upper and lower bounds of the angle values, $n$ is the precision requirement, and $m_j$ is the length of the encoding string.

Start the MATLAB general toolbox and enter the genetic algorithm. The key parameters such as crossover probability are 85%, and mutation probability is 20% in the genetic algorithm, and set the reproduction generation GEN = 80 as a condition for the termination. The result is shown in Figure 6.

During the simulation process, the coordinate deviation of $(z, y)$ and $(z*, y*)$ is gradually reduced, and the efficiency of the inversion algorithm is obviously decreased with the increase of the reproduction algebra and gradually tends to be stable.

The simulation results are compared with the set values, and the results are shown in Table 1.

It can be seen from Table 1 that the external parameters after simulation are closer to the set values.

## 4. Verification Experiment

In order to verify the external parameter inversion method, the experiment is designed and carried out with live firing; then, these shooting data were used to modify the blockchain and participate in machine learning algorithms instead of simulation data. After the light-screen array vertical target is well arranged, the internal parameters are inputted and the measurement model is established; then, the external parameter measurement model is introduced for optimization, and the value range is estimated which is within the simulation value range.

After random shots on the 500 mm × 500 mm rectangular target surface, the database of the time series $t_i$ and target coordinates is obtained. The external parameters are learned by the machine learning, and the external parameters of the optimized model are identified. The experimental data is shown in Table 2.

After the optimization model is identified, 10 random shots are performed again. The coordinates of these projectiles are calculated separately under the original model and the optimized model; then, the calculated results are compared with the projectile hole coordinates on the cardboard target. The result is shown in Figure 7.

The experiment shows the deviation between cardboard target and the light-screen array measurement resulting under two different models which both fluctuate around zero. But the deviation of measurement resulting under an optimized model is smaller, which means that it is closer to the true value. The optimization model and the inversion method of obtaining external parameters are feasible, which makes the light-screen array have higher precision.

## 5. Conclusions

In order to effectively improve the measurement accuracy of the light-screen array principle, this paper proposes an optimized measurement model by introducing external parameters, whose spatial structure is more close to the actual measured state. Then, focusing on the problem of external parameter identification, the inversion method of machine learning based on genetic algorithm is further studied. The objective function is constructed by the deviation between the projectile hole and the light-screen array measurement coordinates, and the flight data of the projectile is collected to form the blockchain, which is used for the inversion database during the machine learning process. The dimension of the inverse model is reduced by inputting all the internal parameters and ignoring the constant parameters, so that the efficiency of the algorithm is improved. The simulation result shows that the inverted external parameter values are close to the given true values, and the feasibility of the proposed method is verified by the live firing test.

In the next step, air resistance, gravity, and other factors can be considered to modify the trajectory; then, the data interpretation basis was introduced (i.e., PaTa Criterion) to reject the data with gross error before writing into the database, in order to make the inversion model more accurate.

Although applying the optimized model and the algorithm to the separated light-screen array will increase the dimension of structural parameters in the inversion model, they are still theoretically applicable. The results of this paper can provide a reference for the optimization of the light-screen array measurement model and the improvement of the precision, and the correlation algorithm can also provide reference for the identification of model parameters in other engineering problems.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the Scientific Research Plan Projects of Shaanxi Education Department (grant number 20JK0692).

## References

[1] M. H. Ferdowsi and E. Sabzikar, "Optical target tracking by scheduled range measurements," *Optical Engineering*, vol. 54, no. 4, article 044101, 2015.

[2] N. A. Kazarinov, V. A. Bratov, N. F. Morozov et al., "Experimental and numerical analysis of PMMA impact fracture," *International Journal of Impact Engineering*, vol. 143, article 103597, 2020.

[3] A. Sedunov, A. Sutin, H. Salloum, and N. Sedunov, "Passive acoustic localization of small aircraft," *Journal of the Acoustical Society of America*, vol. 134, no. 5, p. 4076, 2013.

[4] Y. Li, Y. Wang, B. Liu, S. Zhang, L. Nie, and G. Bi, "A new motion parameter estimation and relocation scheme for airborne three-channel CSSAR-GMTI systems," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 4107–4120, 2019.

[5] W. H. Ma, T. Dong, H. Tian, and J. Ni, "Line-scan CCD camera calibration in 2D coordinate measurement," *Optik*, vol. 125, no. 17, pp. 4795–4798, 2014.

[6] H. S. Li and Z. Y. Lei, "Projectile two-dimensional coordinate measurement method based on optical fiber coding fire and its coordinate distribution probability," *Measurement Science Review*, vol. 13, no. 1, pp. 34–38, 2013.

[7] Z. C. Wu and X. L. Zhang, "On-sate calibration method of target distance of the sky screen target velocity measuring system," *Optik*, vol. 178, pp. 483–487, 2019.

[8] H. S. Li, Z. Lei, Z. Wang, and J. Gao, "Research object photoelectric characteristic and fire coordinate distributing probability in across screen system," *Nanoelectronics and Optoelectronics*, vol. 7, no. 2, pp. 199–203, 2012.

[9] J. P. Ni, *Technology and Application of Measurement of the Light Screen Array*, National Defense Industry Press, Peking, 2014.

[10] H. Tian, J. P. Ni, and M. X. Jiao, "Moment acquisition algorithm of a projectile passing through a trapezoidal screen," *Acta Photonic Sinica*, vol. 43, no. 12, article 1212001, 2014.

[11] R. Chen and J. P. Ni, "Optimization method of light-scree-array structure parameters of photoelectric target based on orthogonal test," *Acta Armamentarii*, vol. 38, no. 11, pp. 2234–2239, 2017.

[12] Z. C. Wu, J. P. Ni, X. L. Zhang, and Y. Wu, "Study on verification device of screen spatial location parameters of sky screen target," *Optik*, vol. 125, no. 14, pp. 3770–3773, 2014.

[13] N. Deepa, Q. V. Pham, D. C. Nguyen et al., "A survey on blockchain for big data: approaches, opportunities, and future directions," *ACM Computing Surveys*, vol. 1, no. 1, 2020.

[14] H. Cui, Y. Guan, H. Chen, and W. Deng, "A novel advancing signal processing method based on coupled multi-stable stochastic resonance for fault detection," *Applied Sciences*, vol. 11, no. 12, p. 5385, 2021.

[15] W. Deng, S. Shang, X. Cai et al., "Quantum differential evolution with cooperative coevolution framework and hybrid mutation strategy for large scale optimization," *Knowledge-Based Systems*, vol. 224, article 107080, 2021.

[16] X. Cai, H. Zhao, S. Shang et al., "An improved quantum-inspired cooperative co-evolution algorithm with muli- strategy and its application," *Expert Systems with Applications*, vol. 171, article 114629, 2021.

[17] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 7621692, 17 pages
https://doi.org/10.1155/2021/7621692

*Research Article*

# A Joint Optimization Model of $(s, S)$ Inventory and Supply Strategy Using an Improved PSO-Based Algorithm

**Huayang Deng, Quan Shi, and Yadong Wang**

*Army Engineering University, Shijiazhuang Campus, Shijiazhuang, Hebei 050003, China*

Correspondence should be addressed to Quan Shi; anshi2you@163.com

Received 9 July 2021; Revised 10 August 2021; Accepted 12 August 2021; Published 31 August 2021

Academic Editor: Thippa Reddy G



This paper mainly discussed the problem of a multiechelon and multiperiod joint policy of inventory and supply network. According to the random lead time and customers' inventory demand, the $(s, S)$ policy was improved. Based on the multiechelon supply network and the improved, the dynasty joint model was built. The supply scheme in every period with the objective of minimum total costs is obtained. Considering the complexity of the model, the improved particle swarm optimization algorithm combining the adaptive inertia weight and grading penalty function is adopted to calculate this model and optimize the spare part problems in various environments.

## 1. Introduction

As an important foundation of product maintenance, the research on spare parts in product maintenance is increasingly applied in industrial and military fields. Along with more and more research, it becomes more and more mature.

In the last practice, many researchers have studied many aspects in every field of spare parts. Sherbrooke in the establishment of the aviation spare for maximum availability considered the failure laws of a complex system and established a spare part demand model under complex factors [1]. Hu et al. analyzed the characteristics of various spare part inventory policies and distinguished the best use conditions for different inventory policies [2]. Ghobbar and Friend added dynamic coordination to the study of multistage spare part supply to improve the stability of the spare part supply process and improve system efficiency [3].

With the deepening of the research, the research of spare part work is not limited to one aspect and gradually begins to deepen the research on the whole process of spare part work. Among these, the researchers are more interested in the joint optimization of spare part inventory and supply process. In the study of joint optimization, it is resulting from the flexibility and random factors of maintenance mode, to consider multiperiod continuous process than to study spare part strategy for a certain period which is more valuable.

In the research of inventory policy, considering the characteristics of modern inventory, then the $(T, S)$ and $(s, S)$ policies are more in line with the actual inventory management experience [4]. According to the $(T, S)$ policy, the capacity will be replenished to $S$ in the same time interval of time $T$ [5]. As for the $(s, S)$ policy, when the capacity decreases less than $s$, the capacity is replenished whose capacity maximum is $S$. In contrast, the $(s, S)$ policy is more flexible and more complex in the research of inventory policy so that the joint inventory policy is less studied. This paper will make use of the characteristics and advantages of the $(s, S)$ policy in joint optimization to carry out joint optimization research [6].

In the project of joint optimization of inventory and supply, researchers focus on balancing transportation costs and breakdown losses caused by insufficient inventory based on satisfying demand and then determining the supply lead time, to achieve the purpose of maximizing benefits. In many previous works of literatures, fixed supply lead time is used to calculate, but in contrast to practical experience, this assumption is impractical, so much literature began to consider details of this aspect [7]. Reference [6] shows that the supply lead time set is an empirical formula, which is assumed to satisfy Poisson

distribution in the literature, but the lead time will change dynamically according to the influence of realistic objective, subjective, or random factors. Therefore, the lead time can still be further researched. The joint optimization in this paper can calculate the specific delivery time of each customer by optimizing the supply distribution process. The system optimizes the supply distribution process by adjusting the quantity of spare parts transporting between different nodes. So, the counts of the spare parts transporting between different nodes are regarded as the decision parameters and the optimization factors. The system can reduce the cost while ensuring the inventory consumption throughout the supply cycle as far as possible by adjusting these parameters.

The problem discussed in this paper is a multiechelon supply network optimization problem. This paper used an improved optimization algorithm to solve the problem. An adaptive particle swarm optimization (PSO) algorithm is proposed in this literature [8]. Based on the optimization framework of the traditional PSO algorithm, the improved algorithm can detect and respond to the changes in the optimization environment. Otherwise, we jointed the new parameters, such as the inertia weight and penalty function. The global exploration and local development ability of the algorithm is adjusted in time to improve the efficiency of the algorithm and adopt an adaptive neighborhood search policy when the environment changes.

The rest of this paper is arranged as follows: Section 2 outlines spare part supply, inventory policy, and joint optimization. Section 3 provides a multiperiod spare part supply optimization model based on product characteristics and applying the $(s, S)$ policy. Section 4 introduces the proposed improved PSO algorithm. Section 5 gave a numerical case to analyze the corresponding results. Section 6 combined numerical examples to analyze the sensitivity of the model and compare it with the traditional policy. Conclusions and future work were given in Section 7.

## 2. Literature Review

*2.1. Supply Network Optimization.* Many researchers study supply networks for a long time. The main research objectives are two aspects. The first one is the supply cycle. Most of research tend to study the supply of spare parts in a single echelon, while multistage supply should be studied in joint optimization. Sherbrooke [9] firstly builds the metric model by multiperiod resupply process. Vaughan studied the multiperiod process of supply to build the ordering policy of spare parts according to random failure possibility [10]. The research on the supply cycle is developed to multiple periods.

Another one is the supply structure. Cachon [11] built the two-echelon supply network. Kennedy et al. [12] research on multiechelon supply process. The research on the supply network structure began to develop from single-echelon to two-echelon and multiechelon.

*2.2. Inventory Policy.* The inventory strategy, aiming at the optimization goal, can be divided into two main aspects. On the one hand, the periodic inventory strategy, which is mainly represented by the $(T, S)$ and $(T, Q)$ inventory strategy, regards the inventory time as the optimization object. This kind of inventory strategy complements the inventory at the specified time node which is a difference by one period $T$ to replenish the inventory capacity to $S$ [13, 14]. On the other hand, the other inventory strategy is mainly represented by the $(s, S)$ and $(s, Q)$ inventory strategy. According to the $(s, S)$ and $(s, Q)$ inventory strategy, when the capacity level is equal to or less than $s$, the system will resupply the spare parts to increase inventory up to $S$ or resupply the stable quantity $Q$ of spare parts [6, 15, 16].

However, when analyzing the inventory strategy of the second kind, the inventory cannot be monitored in real-time under the actual situation, so the interval time of monitoring is still considered in the research process. In the process of optimization, more researchers' points focused on the study of time, so $(s, S)$ and $(s, Q)$ which stand for the second kind of inventory strategy are developed into $(s, S, t)$ and $(s, Q, t)$ inventory strategy [17, 18]. What is more, the maximum inventory in different nodes can be different.

*2.3. Joint Optimization.* For the research of joint strategy optimization, more researchers focus on the joint optimization of maintenance and inventory policy, and there are relatively few joint optimization studies on inventory and supply policy.

Federgruen and Zipkin [19] firstly began to consider the joint of inventory and supply. In the present, the research of inventory and supply joint model mostly starts from two aspects. On the one hand, Spanjers et al. [16] used two echelons in the structure of the joint model. Then, Aharon and Boaz studied establishing multiechelon and multiperiod joint models [17]. Furthermore, the joint model of decentralization is established by Aggarwal and Moinzadeh [18].

On the other hand, many researchers begin to study joint optimization through fixed-length check inventory strategy, such as $(T, S)$ and $(T, Q)$ [14, 20]. However, with the development of joint policy, the complexity of joint optimization is getting deeper and wider. This kind of fixed-length check strategy is not suitable for the developed joint policy. So, some researchers change to another kind of inventory strategy, such as $(s, S)$ and $(s, Q)$ [10, 21].

*2.4. Solution Algorithm.* There are many algorithms that emerged endlessly in the field of algorithm research. The new intelligent algorithm which combined the advantages of different algorithms has also been developed deeply [21–23].

Whether it is a traditional algorithm or a new intelligent algorithm, there are mainly two types, namely, heuristic and metaheuristic [24]. Among most supply models, particle swarm optimization (PSO) is mostly used. The PSO algorithm is a parallel algorithm, which makes use of the advantages of parallel computing of current processors efficiently and greatly improves the efficiency of optimization. Kennedy et al. [12] firstly adopted this algorithm. However, considering the diversity and complexity of the current supply model, the traditional PSO algorithm solves this kind of model problem for a long time and cannot obtain the result even. In order to improve this, many researchers begin to work. One hand is that Clerc and Kennedy [25] adopted the contraction factor

into the algorithm structure. Leong and Yen [26] adjust the inertia weight according to the particles' positions to speed up the convergence rate. On the other hand, Mezura-Montes and Coello Coello [27] improved the pental function in the algorithm to improve the optimization ability.

The improved algorithm in this paper combines the nonlinear dynamic inertia weight and the penalty factor of dynamic correction used to detect the global optimum. Adjusting the convergence weight of particles in global and local optimization adapts to the multiperiod iteration of the model. It is suitable for solving the multiperiod continuous optimization problem.

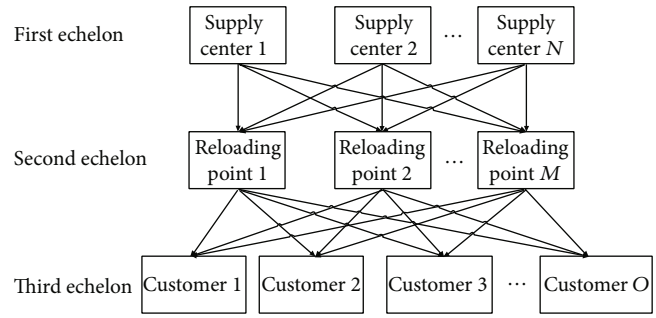



Figure 1: Three-echelon supply network.

## 3. Modelling

*3.1. Problem Description Assumptions.* The three-echelon spare part supply network consists of supply centers, distribution points, and customers, and the organizational structure is shown in Figure 1. Spare parts are sent from the first-echelon supply centers to the second-echelon reloading points and then from the reloading points to the third-echelon customers (the third-echelon customers include the used machines and the storage storehouses). Spare part transportation at all echelons does not affect the consumption of spare parts at the third echelon. The used machines at the third echelon are maintained by replacement, and the failure rate of spare parts is determined. The storage storehouses at the third echelon adopted the $(s, S)$ policy [28].

Because the demand for spare parts is multiperiod and the demand is intermittent, the purpose of the model is to minimize the total cost under the previous condition of a certain support rate. By adjusting the supply time of each period, the number of spare part supplies in each period is optimized and the optimal allocation scheme for each period is found.

The plan formulation process is as follows:

First of all, according to the life distribution of parts, the consumption of spare parts in time intervals can be calculated. Then, according to the requirements of the selected inventory policy, calculating the spare part demand is possible in the corresponding periods. Secondly, aiming at the minimum cost, the spare part supply network is constructed, supplemented by the corresponding constraints (node distance, capacity, transportation cost, transportation capacity, and others). According to the capacity level of each storage storehouse to supply at the third echelon, the conditions conclude the supply level of the system; above all, the optimal spare part supply policy will be obtained.

Because the demand for spare parts is multiperiod and the demand is intermittent, the purpose of the model is to minimize the total cost under the previous condition of a certain support rate. By adjusting the supply time of each period, the number of spare part supplies in each period is optimized and the optimal allocation scheme for each period is found.

(1) Assumptions and conditions

(a) The capacity of supply centers at the first echelon is unlimited

(b) The second-echelon reloading points only carry on the spare part transshipment and do not store up

(c) This joint model only considers one key kind of parts

(d) Customers at the third echelon have the same importance degree

(2) Notions

$i = 1, 2, 3 \cdots N$: index of supply centers at first echelon.

$j = 1, 2, 3 \cdots M$: index of reloading points at second echelon.

$k = 1, 2, 3 \cdots O$: index of customers at the third echelon.

$\tau = 1, 2, 3 \cdots \psi$: index of supply periods.

$n_k$: the quantity of customer $k$'s machines at the third echelon.

$t0^{\tau}$: arrival time of spare part resupply to a customer in period $\tau$.

$S_k$: the maximum inventory of customer $k$.

$s_k$: the inventory node of customer $k$.

$t^{\tau}$: lead time of resupply in period $\tau$.

$Tk^{\tau}$: the supply interval between period $\tau - 1$ and period $\tau$ for customer $k$ theoretically.

$Tk^{\tau'}$: the time from arriving at customer $k$'s storage storehouse in the previous period to it in period $\tau$ theoretically, as follows:

$$T^{k\tau\prime} = t^{k\tau-1} + T^{k\tau}. \tag{1}$$

$T^{\tau}$: the transporting time of system in period $\tau$.

$h(f)$: failure probability density function of the equipment resulting from the failure of the spare parts.

$F(h)$: failure cumulative distribution function of the equipment resulting from the failure of the spare parts.

$p_k$: supply support degree of the customer $k$ at third echelon.

$N_k{}^{Tk\tau}$: consumption of customer $k$'s spare parts in the supply interval of period $\tau$.

$N_k^{tk\tau}$: consumption of customer $k$'s spare parts in the lead time of period $\tau$.

$N^{k\tau}$: consumption of customer $k$'s spare parts in period $\tau$.

$C_{ij}^{p}$: unit transport cost from supply center $i$ to reloading point $j$.

$C_{jk}^{p}$: unit transport cost from reloading point $j$ and customer $k$.

$C_p$: total transport cost.

$C_k^{v}$: unit inventory cost of customer $k$ at third echelon.

$C_k^{o}$: ordering cost of customer $k$ at third echelon.

$C_\tau^{l}$: total delay loss by no spare part.

$C_k^{l}$: delay loss for customer $k$ at third echelon by no spare part.

$d_k^\tau$: the quantity of resupply spare parts to customer $k$ at third echelon in period $\tau$.

$T_{ij}^{p}$: the time from supply center $i$ at first echelon to reloading point $j$ at second echelon.

$T_{jk}^{p}$: the time from reloading point $j$ at second echelon to reloading point $k$ at second echelon.

$w_p$: weight of unit spare part.

$w_c$: limited weight of unit transport vehicle.

$b_{NM}^{c\tau}$: the count of transport vehicle from $N$ supply centers to $M$ reloading points in period $\tau$.

$b_{MO}^{c\tau}$: the count of transport vehicle from $M$ reloading points to $O$ customers in period $\tau$.

(3) Decision variable and notations

$X_{ij}^\tau$: the count of the spare parts transporting from supply center $i$ at first echelon to reloading point $j$ at second echelon in period $\tau$.

$X_{jk}^\tau$: the count of the spare parts transporting from reloading point $j$ at second echelon to customer $k$ at third echelon in period $\tau$.

*3.2. Calculation of Spare Part Requirements.* First of all, according to the principle of demand traction supply and the life distribution of spare parts, within the limited time range of each period, the replacement probability of spare parts (spare part consumption is $s$) is as follows:

$$P_x(s) = F^s(h) - F^{s+1}(h), \tag{2}$$

where $F^s(h)$ represents $F(h)$'s $s$-fold convolution [29].

The formula of supply support degree is as follows:

$$P = \sum_{s=0}^{N} P_x(s) = \sum_{s=0}^{N} \left[F^s(h) - F^{s+1}(h)\right]. \tag{3}$$

At the same time, the consumption formula of parts with different life distributions is given as follows:

(a) Assuming that the life of the part is exponentially distributed from the failure rate $\lambda$, the consumption is as follows:

$$\lambda h + Z_p\sqrt{\lambda h}, \tag{4}$$

where $Z_p$ is the quantile of the standard normal distribution.

(b) Assuming that the life of the component follows the normal distribution of mean $\mu$ and standard deviation $\sigma$, the consumption is as follows:

$$\frac{h}{\mu} + Z_p\sqrt{\frac{\sigma^2 h}{\mu^3}}. \tag{5}$$

(c) Assuming that the life of the component follows the Weibull distribution with shape parameter $\alpha$, scale parameter $\beta$, and position parameter $\gamma = 0$, the expectation of consumption is as follows:

$$\left[\frac{Z_p\Phi}{2} + \sqrt{\left(\frac{Z_p\Phi}{2}\right)^2 + \frac{h}{E}}\right]^2, \tag{6}$$

where the expectation $E$ and the variance $\Phi$ are as Equations (7) and (8) in the following:

$$E = \beta \cdot \Gamma\left(1 + \frac{1}{\beta}\right), \tag{7}$$

$$\Phi = \sqrt{\frac{\Gamma(1 + 2/\beta)}{\Gamma(1 + 1/\beta)^2 - 1}}. \tag{8}$$

Supply decision starts with spare part demand. Spare part resupply in each period $\tau$ shall meet all spare part consumption of the interval time from the former resupply arriving time $t_0^{\tau-1}$ in period $\tau - 1$ to the current resupply arriving time $t_0^\tau$ in period $\tau$. According to the definition of $T^{\tau'}$, it can be divided into the counts of consumption in supply interval $T^\tau$ and lead time $t^\tau$.

According to Equation (2), the spare part consumption of customer $k$ in the supply interval $T_k^\tau$ of the third echelon is as follows:

$$N_k^{T^{k\tau}} = n_k \cdot \inf\left\{N_k^{T^{k\tau}} \mid p_k \geq \sum_{s=0}^{N_k^{T^{k\tau}}} \left[F^s\left(T^{k\tau'} - t^{\tau-1}\right) - F^{s+1}\left(T^{k\tau'} - t^{\tau-1}\right)\right]\right\}. \tag{9}$$

Spare part consumption during lead time $t^\tau$ at the third echelon is as follows:

$$N_k^{t^{k\tau}} = n_k \cdot \inf \left\{ N_k^{t^{k\tau}} \mid p_k \geq \sum_{s=0}^{N_k^{t^{k\tau}}} \left[ F^s(t^\tau) - F^{s+1}(t^\tau) \right] \right\}. \quad (10)$$

Therefore, according to Equations (9) and (10), the spare part consumption of customer $k$ in the period $\tau$ is as follows:

$$N_k^\tau = n_k \cdot \inf \left\{ N_k^\tau \mid p_k \geq \sum_{s=0}^{N_k^\tau} \left[ F^s\left(T^{k\tau} - t^{\tau-1} + t^\tau\right) - F^{s+1}\left(T^{k\tau} - t^{\tau-1} + t^\tau\right) \right] \right\}. \quad (11)$$

*3.3. The Joint Optimization Model Based on the $(s, S)$ and Supply Policy.* In the decisions of spare part inventory and supply, the model may involve the cost of transportation, inventory costs, ordering costs, downtime loss, and so on.

Due to equipment failure and the reduction of spare part inventory, it is necessary to resupply in advance to ensure the spare parts during maintenance. Otherwise, the huge cost of downtime brings a huge burden to customers.

The goal of the model is to reduce the equipment downtime loss and inventory cost as far as possible and to make the cost of the whole spare part inventory-supply process lowest under a certain equipment availability. In the discussion of this model, the functional model of spare parts is not discussed.

According to the characteristics of the $(s, S)$ policy, the cost of this model includes transportation cost, inventory cost, order cost, and downtime loss. The objective function is as follows:

$$\min C_\tau = C_\tau^p + \sum_k^O C_k^v \left( \sum_j^M X_{jk}^\tau \right) + \sum_k^O C_k^o + C_\tau^l. \quad (12)$$

Among them, the first item is the transportation cost, the second item is the inventory cost, the third item is the order cost, and the fourth item is the breakdown loss of machines.

According to the $(s, S)$ policy, the consumption of spare parts during the period $\tau$ should be equal to the difference between $S_k$ and $s_k$, as follows:

$$N_k^{T_{k\tau}} = S_k - s_k. \quad (13)$$

Combination with Equations (9) and (13) can be expressed as follows:

$$n_k \cdot \inf \left\{ N_k^\tau \mid p_k \geq \sum_{s=0}^{N_k^\tau} \left[ F^s\left(T^{k\tau} - t^{\tau-1} + t^\tau\right) - F^{s+1}\left(T^{k\tau} - t^{\tau-1} + t^\tau\right) \right] \right\} = S_k - s_k. \quad (14)$$

According to Equation (14), the system supply time of the third echelon in the period $\tau$ can be obtained, but in the actual supply, the model generally adopts the unified supply, and the supply time should be determined by the supply level of inventory capacity in the third echelon. The actual supply time should be the maximum theoretical time in the third echelon:

$$T^\tau = \max \left( T^{k\tau} \right). \quad (15)$$

Therefore, Equation (14) can be adjusted to

$$n_k \cdot \inf \left\{ N_k^\tau \mid p_k \geq \sum_{s=0}^{N_k^\tau} \left[ F^s\left(T^\tau - t^{\tau-1} + t^\tau\right) - F^{s+1}\left(T^\tau - t^{\tau-1} + t^\tau\right) \right] \right\} = S_k - s_k{}'. \quad (16)$$

From Equation (16), the resupply nodes in customers' inventories in fact is determined by which inventory reaching the supply point at the latest.

The resupply quantity of spare parts during the period $\tau$ should bring the inventory of each cunstomer back to the maximum. When there are spare parts in store, the quantity should be the difference between the maximum inventory and the remaining inventory. When the spare parts are used up, the quantity should be equal to the maximum. It should be expressed as follows:

$$d_k^\tau = \sum_j^M X_{jk}^\tau = \min\left(S_k, N_k^\tau\right). \quad (17)$$

Secondly, since the reloading points have no inventory capacity, the output of spare parts should be equal to the input of the reloading points, as follows:

$$\sum_i^N X_{ij}^\tau = \sum_k^O X_{jk}^\tau. \quad (18)$$

Furthermore, the quantity of transport is limited and is not higher than the maximum transfer capacity of each reloading point, as follows:

$$\sum_i^N X_{ij}^\tau \leq U_j. \quad (19)$$

At the same time, in the resupply process of spare parts, the part of downtime loss should be considered, as follows:

$$C_\tau^l = \sum_k^O C_k^l n_k \left( -\min\left( s_k - N_k^{t^{k\tau}}, 0 \right) \right). \quad (20)$$

When the remaining spare parts in inventory are sufficient to meet the spare part consumption in the lead time $t^\tau$, there is no downtime loss, as follows:

$$s_k - N_k^{t_{k\tau}} \geq 0. \quad (21)$$

When the remaining spare parts cannot do it, it is necessary to bear the downtime loss of machines that cannot replace spare parts, as in Equations (22) and (23) in the following:

$$s_k - N_k^{t_{k\tau}} < 0, \quad (22)$$

$$C_\tau^l = \sum_k^O C_k^l n_k \left(s_k - N_k^{t_{k\tau}}\right). \tag{23}$$

In the course of transport, taking into account the transport limit of vehicles, the total weight transported by a single means of transport shall not exceed the limit specified by the means of transport, as in Equations (24) and (25) in the following:

$$b_{NM}^{c\tau} = \sum_i^N \sum_j^M \left[X_{ij}^\tau \times \frac{w_c}{w_p}\right], \tag{24}$$

$$b_{MO}^{c\tau} = \sum_j^M \sum_k^O \left[X_{jk}^\tau \times \frac{w_c}{w_p}\right], \tag{25}$$

where the Gaussian function is used in the above formulas, which is rounding up. When the resupply weight exceeds the limit weight for a vehicle, an additional vehicle is needed.

The transport cost is related to the distance of transport carrying out transport, as follows:

$$C_\tau^p = b_{NM}^{c\tau} \times C_{ij}^p + b_{MO}^{c\tau} \times C_{jk}^p = \sum_i^N \sum_j^M \left(\left[X_{ij}^\tau \times \frac{w_c}{w_p}\right] \times C_{ij}^p\right) + \sum_j^M \sum_k^O \left(\left[X_{jk}^\tau \times \frac{w_c}{w_p}\right] \times C_{jk}^p\right).$$

In this model, it is one of the previous conditions that spare parts can start to be delivered to customers only after all spare parts arrive at the reloading points. For different customers, the delivery time between supply centers and reloading points is the same, but the delivery time between reloading points and customers is determined by the time of the spare parts arriving at the customers. Therefore, the customer's arriving time is different from each other, and the span of the transportation time is related to the quantity of spare parts transported.

In period $\tau$, the lead time $t^\tau$ for the entire system is as follows:

$$t^\tau = \max\left(X_{ij}^\tau T_{ij}^p\right) + \max\left(X_{jk}^\tau T_{jk}^p\right). \tag{27}$$

The lead time is determined by the amount of spare parts transported between nodes and transportation time. The system would not start the next resupply until the current supply ends. So, the lead time should be equal to the maximum resupply time between nodes.

At the same time, as shown in Equation (28) in the following:

$$t^0 = 0, \tag{28}$$

where the beginning of the system needs not a resupply.

There is a one more thing that the variable is the natural number and positive number, as follows:

$$X_{ij}^\tau, X_{jk}^\tau \in N^+. \tag{29}$$

Composed with the previous formulas, the multiperiod joint model of inventory and supply is as follows:

$$\begin{aligned}
&\min C_\tau = \sum_i^N \sum_j^M \left(\left[X_{ij}^\tau \times \frac{w_c}{w_p}\right] \times C_{ij}^p\right) + \sum_j^M \sum_k^O \left(\left[X_{jk}^\tau \times \frac{w_c}{w_p}\right] \times C_{jk}^p\right) + \sum_k^O C_k^r \left(\sum_j^M X_{jk}^\tau\right) + \sum_k^O C_k^o,\\
&\text{s.t.} X_{ij}^\tau, X_{jk}^\tau \in N^+,\\
&N_k^\tau = n_k \cdot \left\{N_k^\tau \mid p_k \geq \sum_{s=0}^{N_k^\tau} \left[F^s\left(T^{k\tau} - t^{\tau-1} + t^\tau\right) - F^{s+1}\left(T^{k\tau} - t^{\tau-1} + t^\tau\right)\right]\right\},\\
&d_k^\tau = \sum_j^M X_{jk}^\tau = \min\left(S_k, N_k^\tau\right),\\
&n_k \cdot \inf\left\{N_k^\tau \mid p_k \geq \sum_{s=0}^{N_k^\tau} \left[F^s\left(T^\tau - t^{\tau-1} + t^\tau\right) - F^{s+1}\left(T^\tau - t^{\tau-1} + t^\tau\right)\right]\right\} = S_k - s_k{}',\\
&\sum_i^N X_{ij}^\tau \leq U_j,\\
&\sum_i^N X_{ij}^\tau = \sum_k^O X_{jk}^\tau,\\
&t^\tau = \max\left(X_{ij}^\tau T_{ij}^p\right) + \max\left(X_{jk}^\tau T_{jk}^p\right),\\
&t^0 = 0,\\
&N_k^{t^\tau} = n_k \cdot \inf\left\{N_k^{t^\tau} \mid p_k \geq \sum_{s=0}^{N_k^{t_\tau}} \left[F^s(t^\tau) - F^{s+1}(t^\tau)\right]\right\},\\
&i = 1, 2, 3 \cdots N;\ j = 1, 2, 3 \cdots M;\ k = 1, 2, 3 \cdots O;\\
&\tau = 1, 2, 3 \cdots \psi,\\
&X_{ij}^\tau, X_{jk}^\tau \in N^+.
\end{aligned} \tag{30}$$

## 4. Proposed Algorithm

For calculating the proposed model, it is used the intelligent algorithm in this paper. Because of the learning ability of them, especially the PSO algorithm, it can make good use of the existing resources to search the optional decision variables fully, so the PSO algorithm is a good method to solve it. So, we decide to adopt the PSO algorithm to solve this problem.

On the other hand, there are many assumptions and conditions established in this article. Because of that, there are many model constraints, and the target environment structure is more complicated. As the number of data increases, the solution of the PSO algorithm is likely to enter the "local selection trap" and thus cannot obtain the global optimal solution. Therefore, in order to solve this problem, this paper used an improved PSO algorithm.

*4.1. Traditional PSO Algorithm.* Set the potential solution to the optimization problem as a group of particles in space. Suppose there are $N$ particles in a $D$-dimensional search space, and the vector of the $i^{\text{th}}$ particle in the $D$-dimensional space is expressed as Equation (31) [30].

$$X_i = (x_{i1}, x_{i2}, x_{i3} \cdots x_{iD}), \quad i = 1, 2, 3 \cdots N. \tag{31}$$

Each particles have an adaptive value (fitness value)

determined by its location and passing velocity. The "flying" velocity is as follows:

$$V_i = (v_{i1}, v_{i2}, v_{i3} \cdots v_{iD}), \quad i = 1, 2, 3 \cdots N. \tag{32}$$

Their location updates are as in Equations (33) and (34) in the following:

$$v_{id}(t+1) = w \cdot v_{id}(t) + c_1 \cdot r_1 \cdot [p_{id} - x_{id}(t)] + c_2 \cdot r_2 \cdot [p_{id} - x_{id}(t)], \tag{33}$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1). \tag{34}$$

Among them, $P^i_{\text{best}}$ is the $i^{\text{th}}$ particle that so far obtained the optimal position. The individual extreme point is recorded as follows:

$$P^i_{\text{best}} = (p_{i1}, p_{i2}, p_{i3} \cdots p_{iD}), \quad i = 1, 2, 3 \cdots N. \tag{35}$$

$G_{\text{best}}$ is the optimal position that the entire particle swarm has searched so far. The global extreme point is denoted as follows:

$$G_{\text{best}} = (g_{i1}, g_{i2}, g_{i3} \cdots g_{iD}), \quad i = 1, 2, 3 \cdots N. \tag{36}$$

For inertia weight $w$, it is used to represent the influence of the initial state on the particle motion. The acceleration degree $(c_1, c_2)$ and the acceleration weight coefficient $(r_1, r_2)$, which are composed of the last two terms of Equation (33), represent the influence of the particle's own historical experience data and the collective historical data on the particle motion. Find the position of the optimal fitness value particles, compare and adjust the local and global extremum points. From then on, repeat Equations (33) and (34), update the local and global extremum, and get the optimal extremum of the system [31].

However, the PSO algorithm also has its disadvantages. For an in-depth discussion of it, one of the biggest advantages lies in the application of the algorithm. The adjustment parameters are less, but they directly affect the performance and convergence of the algorithm [32]. For the weight parameter $w$, the traditional PSO algorithm can improve the global search ability and reduce the local search ability of the algorithm. Therefore, many researchers began to put forward many programmers that improved the weight parameter $w$, such as the adaptive weight method, the random weight method, and the linear recursive weight method.

*4.2. Adaptive Weight Method.* The adaptive weight method mainly has two kinds of optimization directions. Firstly, according to the early convergence degree of particle swarm, and the value of population fitness, the change of inertia weight $w$ is determined, and the population is divided into three subsets. It is known that the velocity of the algorithm's constringency depends on the dispersion degree of particles. Therefore, when the dispersion degree of particles is a little dispersed, the adaptive weight method reveals details that decreasing the value of $w$ does well in speeding up the velocity of algorithm's constringency. On the other hand, it can raise the value of $w$ to decrease the degree of falling into the "local optimal" trap either [32–34].

```
Begin
Input      acceleration degree c1, c2
           time of maximum iteration M
           search space dimension D
           number of individual groups N
           inertia weight w
Set        condition of stopping iteration
Initialize local extremum P_best^i and global extremum G_best
For        condition of stopping iteration
Do for     i ←1 to N
  Do for     j ←1 to D
    Do       initialize particle position x(i, j) randomly
             initialize particle position v(i, j) randomly
For i ←1 to N
Do for j ←1 to D
  Do       calculate fitness value of each particle
    If (fitness(x(i, j)) < fitness(P_best^i))
    Then [P_best^i ← x(i j)]
    If (fitness(P_best^i) < fitness(G_best))
    Then [G_best ← P_best^i]
    v_id ← w•v_id + c1•r1•(p_id - x_id) + c2•r2•(p_id - x_id)
    x_id ← x_id + v_id
End
```

Algorithm 1: Traditional PSO algorithm.

Secondly, there is also a method used in this paper to adjust the inertia weight $w$ for the global optimal distance from the current position, as shown in Figure 2. Because the PSO algorithm gets closer to the global optimal in theory through optimization, it decreases $w$ continuously with iterations. On the contrary, it will increase $w$ to strengthen the ability of global research [35].

According to the former discussion, the nonlinear dynamic inertia weight coefficient is determined by the current position. The formula is as follows:

$$w = \begin{cases} w_{\min} - \dfrac{(w_{\max} - w_{\max}) \times (\text{fitness} - \text{fitness}_{\min})}{\text{fitness}_{\text{avg}} - \text{fitness}_{\min}}, & \text{fitness} \le \text{fitness}_{\text{avg}}, \\ w_{\max}, & \text{fitness} > \text{fitness}_{\text{avg}}, \end{cases} \tag{37}$$

where fitness represents the current fitness value, $\text{fitness}_{\min}$ and $\text{fitness}_{\text{avg}}$ represent the minimum and average of the fitness values of all current particles. It can be seen from Equation (37) that when the adaptation value difference of each particle is getting larger, the inertia weight $w$ will be dereasing. When the adaptation value of each particle swarm is getting closer, the inertia weight $w$ will be increasing.

*4.3. Fitness Function.* In order to deepen the fluence of the constraints in the optimization, the penalty function is introduced into the adaptive function. As a common method to deal with constraints, the penalty function method can transform the constraint optimization problem into an unconstrained

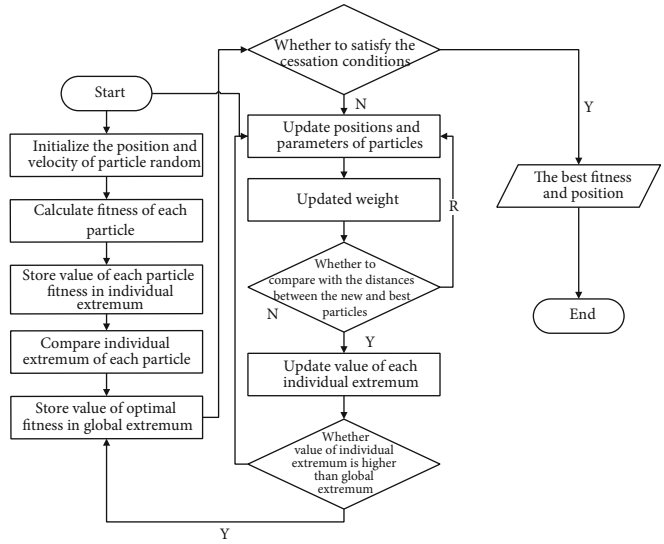



Figure 2: Adjusting the inertial parameter flow at the best global distance.

optimization problem. The punishment function method is mainly divided into an internal punishment method and external punishment method. Set the appropriate parameters according to the constraint conditions and establish the penalty function.

The adaptive function is composed of three parts, which conclude the set variables, the optimized objective function, and the variable constraint condition. The general formula is as follows:

$$L[x(i,j),k] = \text{fitness}[x(i,j)] + \sigma \cdot P[x(i,j)], \tag{38}$$

where $\text{fitness}[x(i,j)]$ is the objective function, $\sigma$ is the penalty factor, and $P[x(i,j)]$ is the limit function.

A limit function is a set of $n(n = 1, 2, 3 \cdots N)$ limiting conditions, as follows:

$$P[x(i,j)] = \sum_{n=1}^{N} G_n[x(i,j)]. \tag{39}$$

Among them, $G_n[x(i,j)]$ is the function of the corresponding constraint condition, which is affected by the constraint function, as follows:

$$P[x(i,j)] = \sum_{n=1}^{l} \max\{0, g_n[x(i,j)]\} + \sum_{n=1}^{m} \max\{0, |h_n[x(i,j)]| - \delta\}, \tag{40}$$

where $m$ is the quantity of constraints and $\delta$ is the tolerance value of the equality constraint. It is the unequal constraint, when the condition is as follows:

$$g_n[x(i,j)] \le 0. \tag{41}$$

It is the equivalent constraint, when the condition is as follows:

$$h_n[x(i,j)] = 0, (\delta \longrightarrow 0). \tag{42}$$

In most algorithms, a fixed penalty value method is generally used. With the iterations of the particle swarm algorithm, the parameters of the group will change. A dynamic correction method whose penalty value changes with the constraint value is shown in Reference [34]. The general formula of the penalty function is as follows:

$$L[x(i,j),k] = \text{fitness}[x(i,j)] + h(k) \cdot P[x(i,j)]. \tag{43}$$

Among them, the penalty factor $h(k)$ is obtained with the number of iteration $k$ increases, and the limit function $P[x(i,j)]$ is improved to satisfy the requirements of multistage iterative changes, as follows:

$$P[x(i,j)] = \sum_{n=1}^{m} \theta\{q_n[x(i,j)]\} \cdot q_n[x(i,j)]^{\gamma\{q_n[x(i,j)]\}}, \tag{44}$$

where $q_n[x(i,j)]$, $\theta\{q_n[x(i,j)]\}$, and $\gamma\{q_n[x(i,j)]\}$ are the corresponding series of violation constraint function, multiiteration distribution function, and penalty function.

The search space of the constrained optimization problem is composed of feasible points and infeasible points. The feasible points satisfy all the constraints, and the infeasible points violate at least one constraint. Penalty function technology solves the constrained optimization problem through penalty constraints. If the penalty value of penalty function is too high, the optimization algorithm is easy to converge to the local minimum solution. If the penalty value is too low, it is difficult to find a feasible optimization solution. The penalty function is divided into fixed penalty value and dynamic correction of penalty value. The penalty function depends on the constraint condition. The optimization result obtained by the dynamic modification of the penalty value with the change of the constraint value is better than that of the fixed penalty value.

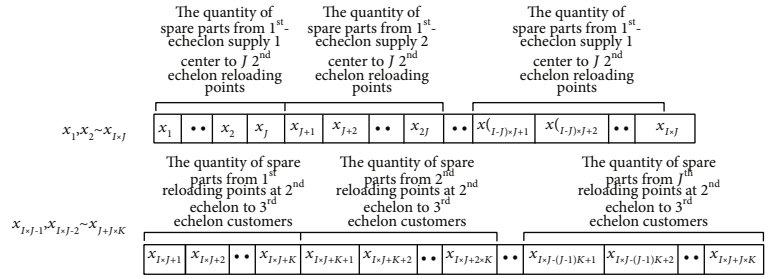



Figure 3: Schematic diagram of particle code.

Above the former discussion, the rules of corresponding parameter satisfied are as the following:

(1) When

$$q_n[x(i,j)] < 1, \gamma\{q_n[x(i,j)]\} = 1, \quad n = 1, 2, 3 \cdots m \tag{45}$$

(2) When

$$q_n[x(i,j)] < 1, \gamma\{q_n[x(i,j)]\} = 1, \quad n = 1, 2, 3 \cdots m \tag{46}$$

(3) When

$$q_n[x(i,j)] < 1, \gamma\{q_n[x(i,j)]\} = 1, \quad n = 1, 2, 3 \cdots m \tag{47}$$

(4) When

$$0.001 \le q_n[x(i,j)] < 0.1, \theta\{q_n[x(i,j)]\} = 20, \quad n = 1, 2, 3 \cdots m \tag{48}$$

(5) When

$$0.1 \le q_n[x(i,j)] < 1, \theta\{q_n[x(i,j)]\} = 100, \quad n = 1, 2, 3 \cdots m \tag{49}$$

(6) When

$$q_n[x(i,j)] \ge 1, \theta\{q_n[x(i,j)]\} = 300, \quad n = 1, 2, 3 \cdots m \tag{50}$$

It can be seen from the penalty function value corresponding to the function value listed above that for different limit function values, the penalty function is classified. The closer the function value is, the smaller the penalty amount. On the contrary, the farther the value is, the penalty the larger the value of the function, the stronger the convergence efficiency of the improved PSO algorithm mentioned in the article.

*4.4. Particle Coding.* In this article, an integer coding method is used to index each particle, and the variation of each dimension of each particle represents the number of spare parts transported between different nodes.

As shown in Figure 3, it represents the quantity of the spare parts transferring to every second-echelon reloading point in the first-echelon supply center *i*. Resulting from it, $x_1 \sim x_J$ represent the number of spare parts transferred from the first-echelon supply center 1 to the second-echelon reloading points, $x_{J+1} \sim x_{J\times 2}$ represent the number of spare parts transferred from the first-echelon supply center 2 to the second-echelon reloading points, and so on. While $x_{I\times J+1} \sim x_{I\times J+K}$ are the quantity of transporting from second-echelon reloading point 1 to the storage storehouses of third-echelon customers, $x_{I\times J+K+1} \sim x_{I\times J+K\times 2}$ are the quantity of transporting from second-echelon reloading point 2 to the storage storehouses of third-echelon customers, etc.

*4.5. Improved Algorithm Framework.* To sum up, this paper combined the inertial weight method and the multistage allocation penalty function method to adjust the global optimal distance from the current position and adopts the improved PSO algorithm. The following is the main process of the improved algorithm:

## 5. Case Analysis

*5.1. Case Description.* This paper will provide a case to show the actual optimization effect and test the correctness of the established model and algorithm. Data comes from a spare parts supplier in 2019.

At present, a kind of machine has widely used in a certain area, so that there is a demand that 6 customers need to be supplied spare parts to the three reloading points at first echelon by one supply center at first echelon. When the inventory level of spare parts' capacity reaches the set spare part supply level, the customers send out the required supply information. The first-echelon supply center supplies the same. In this case, all customers are factories assuming that six factories use the machine in the area. Considering the complexity of the distance and transportation

```
Begin
Input      acceleration degree c1, c2
           time of maximum iteration M
           search space dimension D
           number of individual groups N
           inertia weight w
Set        condition of stopping iteration
Initialize local extremum P_best^i and global extremum G_best
For        condition of stopping iteration
Do for     i ←1 to N
  Do for     j ←1 to D
    Do       initialize particle position x(i, j) randomly
             Initialize particle position v(i, j) randomly
For i ←1 to N
Do for j ←1 to D
  Do       calculate fitness value of each particle according to ((26)) and ((27))
       If (fitness(x(i, j)) < fitness(P_best^i))
       Then [P_best^i ← x(i j)]
       If (fitness(P_best^i) < fitness(G_best))
       Then [G_best ← P_best^i]
       Adjust inertial weight(w) according to ((24))
       v_id ← w•v_id + c1•r1•(p_id - x_id) + c2•r2•(p_id - x_id)
       x_id ← x_id + v_id
End
```

Algorithm 2: Improved PSO algorithm.

Table 1: Failure distribution of the discussed part.

| Name | Detail |
|---|---|
| Failure probability ($\lambda$) | $0.6 \times 10^{-4}$ time/hour |
| Failure distribution function | $F(t) = 1 - e^{-\lambda t}$ |
| Failure distribution convolutional function | $F^s(t) = \sum_n^{s-1} \frac{1}{n!}(\lambda t)^n e^{-\lambda t}$ |

conditions, the factories reserve spare parts for this kind of machine, but because of the particularity of the parts using this kind of spare parts, these kinds of parts can only be used in the factory for the maintenance and replacement of the machine. The lifetime distribution of parts is known, as shown in Table 1.

According to the practice, the production environment parameters, such as the number of machines put into production, the number of spare parts stored, the maximum inventory, and the supply support required by production are different. At the same time, due to the influence of regional, production products and policies, various factories pay the cost of ordering, inventory, or breakdown losses caused by spare parts differently. According to statistics, the relevant parameters and costs are shown in Table 2, among which the left five columns are production environment parameters, which are the number of machines put into production in each factory, the supply level, the maximum inventory, and the supply support. The three in the right are listed as economic costs, which are spare part storage cost, order cost, and delay cost.

Transportation costs are one of the most important aspects of the cost; as a result of transport conditions and policies, the cost of transporting spare parts from the first-echelon supply center to the second-echelon reloading points and the second echelon to the storage storehouse of third-echelon factories is different, as shown in Table 3. What is more, the cost of transporting spare parts from the second-echelon reloading point to the third-echelon factory is lower than the cost of transferring spare parts from the first-echelon supply center to the second-echelon reloading point due to the impact of transport conditions and policies. While the reloading points have their own limitations, the maximum transportation volume is set to 100, 100, and 120.

In spare part supply, another very important parameter is time. A large part of the time in the actual supply process is determined by the volume of spare part transportation. So, the transportation time of unit spare parts in the supply structure is provided in Table 4 in the parameter setting.

In this paper, we divided the whole supply process into five periods, ignoring the transmission time of spare part information. At the same time, the supply process adopted

Table 2: The correlative parameters of third-echelon factories.

| | $n_k$ | $s_k{}'$ | $S_k$ | $P_k$ | $Z_p$ | $C_k{}^v$ (unit: ¥) | $C_k{}^o$ (unit: ¥) | $C_k{}^l$ (unit: ¥) |
|---|---|---|---|---|---|---|---|---|
| Factory 1 | 10 | 20 | 60 | 0.90 | 1.28 | 200 | 2000 | 15000 |
| Factory 2 | 10 | 10 | 60 | 0.99 | 2.33 | 220 | 2000 | 25000 |
| Factory 3 | 8 | 30 | 70 | 0.95 | 1.65 | 200 | 2000 | 20000 |
| Factory 4 | 5 | 25 | 45 | 0.90 | 1.28 | 210 | 2000 | 15000 |
| Factory 5 | 10 | 20 | 50 | 0.80 | 0.84 | 250 | 2000 | 10000 |
| Factory 6 | 12 | 20 | 80 | 0.99 | 2.33 | 225 | 2000 | 10000 |

Table 3: The correlative parameters of third-echelon factories.

| | Supply center | Factory 1 | Factory 2 | Factory 3 | Factory 4 | Factory 5 | Factory 6 |
|---|---|---|---|---|---|---|---|
| Reloading point 1 | 5700 | 600 | 480 | 540 | 480 | 480 | 570 |
| Reloading point 2 | 6000 | 480 | 450 | 510 | 450 | 510 | 630 |
| Reloading point 3 | 4800 | 510 | 330 | 570 | 600 | 510 | 480 |

Table 4: Transportation time of spare parts (unit: km).

| | Supply center | Factory 1 | Factory 2 | Factory 3 | Factory 4 | Factory 5 | Factory 6 |
|---|---|---|---|---|---|---|---|
| Reloading point 1 | 13 | 5 | 2 | 2 | 4 | 3 | 2 |
| Reloading point 2 | 17 | 3 | 6 | 1 | 1 | 2 | 3 |
| Reloading point 3 | 16 | 1 | 1 | 3 | 2 | 2 | 4 |

road transportation which is widely used. The weight of spare parts is 0.5 tons, and the carrying weight of transporting vehicles is limited to 31 tons, without considering the high medium limit of transportation. The best supply scheme is obtained by calculating the above parameters.

Finally, set the PSO algorithm parameter: the maximum number of iterations $M = 300$, the number of individual groups $N = 200$, and the penalty factor is as follows:

$$h(k) = k\sqrt{k}, \tag{51}$$

learning factor $c_1 = 1.6962$, $c_2 = 1.8962$, and the range of inertia weight $[w_{\min}, w_{\max}] = [0.6, 0.8]$.

*5.2. Analysis Process.* For the joint model in this paper, Equation (18) shows that the transferring quantity of spare parts from first to second is the same as the quantity transferred from first to third echelon. In the calculation of this case, to decrease the decision parameters and programming space, the code schematic can be simplified to encode the second-level supply. The simplified code is shown in Figure 4.

The simplified code schematic transforms constraints into logical relationships between decision parameters to speed up the convergence rate. $X_1 \sim x_K$ represent the number of spare parts transported from the second-echelon reloading point 1 to the storage storehouses of third-echelon customers, $x_{K+1} \sim x_{2\times K}$ are the number of spare parts transported from the second-echelon reloading point 2 to the storage storehouses of third-echelon customers,

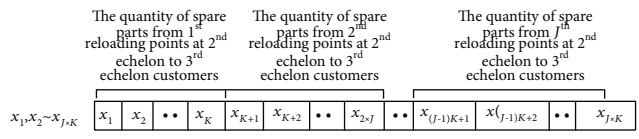



Figure 4: Simplified particle code schematic.

etc. What is more, the sum of each group is the number of spare parts transported to the second-echelon reloading points at the corresponding first-echelon supply center, such as the sum of spare parts which includes the number of $x_{K+1} \sim x_{2\times K}$ which is equal to the number of spare parts transported to the second-echelon reloading points at the first-echelon supply center 1.

*5.3. Result Analysis.* According to the improved PSO algorithm proposed in this paper, the corresponding calculation results are obtained in Table 5.

As shown in Figure 5, they are the change of the objective function, fitness function, and penalty function with the number of iterations.

In Figure 5, the longitudinal axis represents the function value, and the transverse axis represents the iteration period and the number of iterations, where Figure 5(b)) represents the change of the penalty function following the objective function and Figure 5(c)) represents the convergence of the objective function.

It is learning that the adaptive function and objective function tend to converge slowly with the increase of iteration

Table 5: Results of calculation by improved PSO.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Period 1 | 6 | 4 | 0 | 8 | 17 | 9 | 0 | 8 | 24 | 8 | 12 | 9 | 0 | 12 | 19 | 8 | 16 | 43 |
| Period 2 | 0 | 7 | 23 | 7 | 11 | 10 | 6 | 8 | 24 | 17 | 14 | 0 | 0 | 9 | 12 | 26 | 0 | 7 |
| Period 3 | 0 | 8 | 12 | 10 | 12 | 4 | 0 | 0 | 0 | 20 | 13 | 0 | 0 | 0 | 42 | 9 | 0 | 9 |
| Period 4 | 0 | 5 | 40 | 8 | 12 | 6 | 0 | 0 | 14 | 8 | 6 | 5 | 0 | 44 | 0 | 6 | 15 | 8 |
| Period 5 | 35 | 0 | 15 | 7 | 15 | 5 | 0 | 8 | 0 | 11 | 12 | 1 | 0 | 12 | 0 | 10 | 14 | 8 |

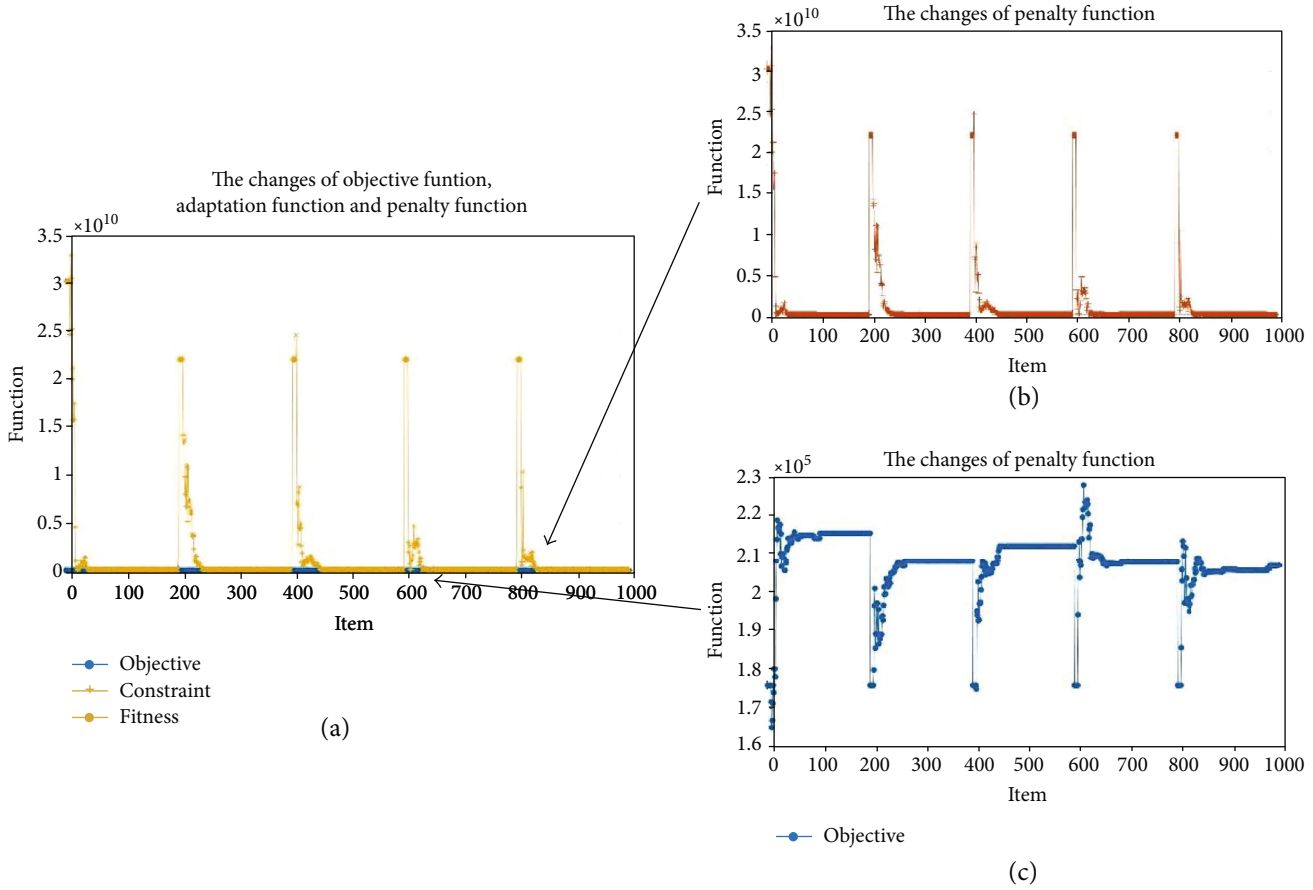



Figure 5: Simplified particle code schematic.

period and times from Figures 5(a) and 5(c)). They tend to be the same at the same time, which shows that the convergence of the results is better, and the convergence of the first stage is slow.

On the other hand, since the adaptive weights used in this paper are closely related to the value of the objective function, the curve of the penalty function in Figure 5(b)) is similar to the objective function. At the same time, the fitting value when all periods are finished tends to 0, indicating that the result is the minimized optimal solution.

The results of the decision variables are shown in Tables 5 and 6. As can be seen from Table 6, they are shown which are inventory node, the cost of supply at each period, the total cost of the system, and the quantity of spare part delays. Secondly, in Table 5, columns 1-6 are the volume of transport from the second-echelon reloading point 1 to the third-echelon factories, 7-12 are the volume of transport from the second-echelon reloading point 2 to the third-echelon factories, and 13-18 are the volume of transport from the second-echelon reloading point 3 to the third-echelon factories. According to Equation (18) and the former discussion, the volume of transportation from the first-echelon supply center to the second-echelon reloading points can be calculated.

As shown in Table 7, compared to the last three columns included in Table 6, the calculated volume of transport is from the first-echelon supply center to the second-echelon reloading points. Among them, column 19 is the sum of columns 1-6, that is, the transfer volume of the transfer of the

Table 6: Result of the best supply scheme.

| | Period 1 | Period 2 | Period 3 | Period 4 | Period 5 |
|---|---|---|---|---|---|
| Cost (unit: ¥) | 137967 | 139404 | 138765 | 138758 | 139907 |
| Total cost (unit: ¥) | | | 694801 | | |
| Consumption | [30,40,24,15,30,44] | [30,39,24,17,30,48] | [30,38,24,15,30,42] | [30,38,24,14,30,48] | [30,36,24,15,32,50] |
| Level of supply | | | 40,50,34,35,30,44 | | |
| Breakdown (unit: ¥) | [0,0,0,0,0,0] | [0,0,0,0,0,0] | [0,0,0,0,0,0] | [0,0,0,0,0,0] | [0,0,0,0,0,0] |

Table 7: Optimized and calculated optimal supply options.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Period 1 | 6 | 4 | 0 | 8 | 17 | 9 | 0 | 8 | 24 | 8 | 12 | 9 | 0 | 12 | 19 | 8 | 16 | 43 | 44 | 61 | 98 |
| Period 2 | 0 | 7 | 23 | 7 | 11 | 10 | 6 | 8 | 24 | 17 | 14 | 0 | 0 | 9 | 12 | 26 | 0 | 7 | 58 | 69 | 54 |
| Period 3 | 0 | 8 | 12 | 10 | 12 | 4 | 0 | 0 | 0 | 20 | 13 | 0 | 0 | 0 | 42 | 9 | 0 | 9 | 46 | 33 | 60 |
| Period 4 | 0 | 5 | 40 | 8 | 12 | 6 | 0 | 0 | 14 | 8 | 6 | 5 | 0 | 44 | 0 | 6 | 15 | 8 | 71 | 33 | 73 |
| Period 5 | 35 | 0 | 15 | 7 | 15 | 5 | 0 | 8 | 0 | 11 | 12 | 1 | 0 | 12 | 0 | 10 | 14 | 8 | 77 | 32 | 44 |

Table 8: Different optimization results of changing the failure degree.

| $\lambda(10^{-4})$ | Cost of each period (unit: yuan) | Consumption | Total cost (unit: yuan) | Breakdown loss (unit: yuan) |
|---|---|---|---|---|
| 0.6 | 137967 | 30,40,24,15,30,44 | 694801 | 0,0,0,0,0,0 |
| | 139404 | 30,39,24,17,30,48 | | 0,0,0,0,0,0 |
| | 138765 | 30,38,24,15,30,42 | | 0,0,0,0,0,0 |
| | 138758 | 30,38,24,14,30,48 | | 0,0,0,0,0,0 |
| | 139907 | 30,36,24,15,32,50 | | 0,0,0,0,0,0 |
| 0.8 | 144112 | 30,40,32,15,30,48 | 723862 | 0,0,0,0,0,0 |
| | 144509 | 30,38,32,15,30,45 | | 0,0,0,0,0,0 |
| | 145219 | 30,38,30,15,30,48 | | 0,0,0,0,0,0 |
| | 143910 | 30,40,32,18,25,47 | | 0,0,0,0,0,0 |
| | 146112 | 35,37,30,15,28,48 | | 0,0,0,0,0,0 |
| 1 | 186998 | 40,60,32,20,30,72 | 887221 | 0,0,0,0,0,0 |
| | 170912 | 40,50,32,20,30,60 | | 0,0,0,0,0,0 |
| | 170455 | 40,50,32,20,30,60 | | 0,0,0,0,0,0 |
| | 186751 | 40,60,32,20,30,72 | | 0,0,0,0,0,0 |
| | 172106 | 40,50,32,20,30,60 | | 0,0,0,0,0,0 |
| 1.2 | 199304 | 40,60,40,20,40,72 | 946741 | 0,0,0,0,0,0 |
| | 185191 | 40,60,32,20,30,72 | | 0,0,0,0,0,0 |
| | 188756 | 40,65,32,20,32,72 | | 0,0,0,0,0,0 |
| | 185761 | 40,60,32,25,30,58 | | 0,0,0,0,0,0 |
| | 187729 | 38,60,35,20,30,66 | | 0,0,0,0,0,0 |
| 1.4 | 3097202 | 40,70,48,20,40,84 | 3887792 | 0,0,0,0,0,480000 |
| | 199658 | 40,60,40,20,40,72 | | 0,0,0,0,0,0 |
| | 198548 | 40,65,40,20,35,72 | | 0,0,0,0,0,0 |
| | 193576 | 40,60,32,20,40,72 | | 0,0,0,0,0,0 |
| | 198807 | 40,60,38,20,40,76 | | 0,0,0,0,0,0 |

Table 9: Different optimization results of quantity of inventory to supply.

| No. | Quantity of inventory to supply | Cost of each period (unit: yuan) | Consumption | Total cost (unit: yuan) | Breakdown loss (unit: yuan) |
|---|---|---|---|---|---|
| 1 | 60 | 137967 | 30,40,24,15,30,44 | 694801 | 0,0,0,0,0,0 |
| | 80 | 139404 | 30,39,24,17,30,48 | | 0,0,0,0,0,0 |
| | 50 | 138765 | 30,38,24,15,30,42 | | 0,0,0,0,0,0 |
| | 45 | 138758 | 30,38,24,14,30,48 | | 0,0,0,0,0,0 |
| | 50 | 139907 | 30,36,24,15,32,50 | | 0,0,0,0,0,0 |
| | 80 | | | | |
| 2 | 30 | 170619 | 40,50,32,20,30,60 | 750160 | 0,0,0,0,0,0 |
| | 40 | 144536 | 30,40,32,15,30,48 | | 0,0,0,0,0,0 |
| | 26 | 144443 | 30,40,32,17,30,46 | | 0,0,0,0,0,0 |
| | 30 | 145630 | 30,35,32,20,30,50 | | 0,0,0,0,0,0 |
| | 30 | 144931 | 30,37,32,23,30,48 | | 0,0,0,0,0,0 |
| | 32 | | | | |
| 3 | 30 | 183044 | 40,50,40,20,40,60 | 889374 | 0,0,0,0,0,0 |
| | 40 | 178905 | 40,50,32,20,40,60 | | 0,0,0,0,0,0 |
| | 18 | 170547 | 40,50,32,20,30,60 | | 0,0,0,0,0,0 |
| | 30 | 177920 | 40,50,32,20,40,60 | | 0,0,0,0,0,0 |
| | 20 | 178959 | 40,55,32,28,30,72 | | 0,0,0,0,0,0 |
| | 32 | | | | |
| 4 | 20 | 209023 | 50,60,40,25,40,72 | 1007481 | 0,0,0,0,0,0 |
| | 30 | 199406 | 40,60,40,20,40,72 | | 0,0,0,0,0,0 |
| | 18 | 198954 | 40,65,40,35,40,72 | | 0,0,0,0,0,0 |
| | 25 | 199432 | 40,60,44,38,35,72 | | 0,0,0,0,0,0 |
| | 20 | 200665 | 40,60,40,20,40,64 | | 0,0,0,0,0,0 |
| | 20 | | | | |
| 5 | 20 | 215198 | 50,60,48,25,40,72 | 1054098 | 0,0,0,0,0,0 |
| | 30 | 210139 | 50,60,40,25,40,72 | | 0,0,0,0,0,0 |
| | 10 | 209519 | 50,65,40,32,40,72 | | 0,0,0,0,0,0 |
| | 25 | 209863 | 50,60,40,25,40,70 | | 0,0,0,0,0,0 |
| | 20 | 209379 | 50,55,40,30,40,72 | | 0,0,0,0,0,0 |
| | 20 | | | | |
| 6 | 20 | 3112742 | 50,70,48,25,50,84 | 15563254 | 0,0,0,0,0,480000 |
| | 20 | 3114210 | 50,70,48,25,60,78 | | 0,0,0,0,1000000,0 |
| | 10 | 3110595 | 50,70,60,25,50,80 | | 0,0,1600000,0,0,0 |
| | 25 | 3114297 | 50,70,50,25,50,82 | | 0,0,0,0,0,240000 |
| | 10 | 3111409 | 50,80,48,25,50,84 | | 0,2500000,0,0,0,480000 |
| | 8 | | | | |

first-echelon supply center to the second-echelon reloading point 1; column 20 is the sum of columns 7-12, that is, the transfer volume of the transfer of the first-echelon supply center to the second-echelon reloading point 2; column 21 is the sum of 13-18, that is, the transfer volume of the first-echelon supply center to the second-echelon reloading point 3.

## 6. Result Analysis

*6.1. Different Failure Degree ($\lambda$) Analysis.* As shown in Table 8, change the size and observe the different optimization results of the model and algorithm.

As can be seen from Table 8, with the increasing failure degree ($\lambda$), the consumption of the whole process is gradually increasing and the demand for spare parts is increasing. As a result, the preset inventory node is difficult to meet the consumption of spare parts during the process of supply. When $\lambda$ reached $1.4 \times 10^{-4}$, there is a delayed loss, and in the actual supply of spare parts, delay consumption should be eliminated or avoided as far as possible.

*6.2. Different Quantity of Inventory to Supply Analysis.* As can be seen from Table 9, when the inventory node is reduced, the consumption of spare parts is increasing continuously, the same as the failure degree, and the cost of each period is also increased. On the other hand, the reduction of the inventory node leads to insufficient advance supply time. Eventually, there is a breakdown loss when the inventory node drops to 20,20,10,25,10,8.

*6.3. Compared with Results of Traditional $(s, S)$ Policy.* In order to test the effect of the model and algorithm optimization, the model is compared with the traditional $(s, S)$ policy based on different inventory nodes.

Table 10: Comparison of No. 6 supply node with the traditional $(s, S)$ policy.

| | Cost of each period (unit: yuan) | Consumption | Total cost (unit: yuan) | Breakdown loss (unit: yuan) |
|---|---|---|---|---|
| Traditional model | 3078278 | 50,70,48,25,50,84 | 15395954 | 0,0,0,0,0,480000 |
| | 3078580 | 50,70,60,25,60,82 | | 0,0,160000,0,100000,240000 |
| | 3079982 | 55,70,52,30,50,82 | | 0,0,320000,0,0,240000 |
| | 3079691 | 50,76,48,40,53,80 | | 0,0,0,0,300000,0 |
| | 3079422 | 50,70,48,25,48,84 | | 0,0,0,0,0,480000 |
| Improved model | 199304 | 40,60,40,20,40,72 | 946741 | 0,0,0,0,0,0 |
| | 185191 | 40,60,32,20,30,72 | | 0,0,0,0,0,0 |
| | 188756 | 40,65,32,20,32,72 | | 0,0,0,0,0,0 |
| | 185761 | 40,60,32,25,30,58 | | 0,0,0,0,0,0 |
| | 187729 | 38,60,35,20,30,66 | | 0,0,0,0,0,0 |

Table 11: Comparison between the improved algorithm and the traditional algorithm.

| | | Cost of each period (unit: yuan) | Consumption | Total cost (unit: yuan) | Breakdown loss (unit: yuan) |
|---|---|---|---|---|---|
| Improved algorithm | | 137967 | 30,40,24,15,30,44 | 694801 | 0,0,0,0,0,0 |
| | | 139404 | 30,39,24,17,30,48 | | 0,0,0,0,0,0 |
| | | 138765 | 30,38,24,15,30,42 | | 0,0,0,0,0,0 |
| | | 138758 | 30,38,24,14,30,48 | | 0,0,0,0,0,0 |
| | | 139907 | 30,36,24,15,32,50 | | 0,0,0,0,0,0 |
| Traditional algorithm | 1 | 288765 | 70,40,24,15,30,35 | 982756 | 150000,0,0,0,0,0 |
| | | 279907 | 30,38,24,15,64,58 | | 0,0,0,0,140000,0 |
| | | 138569 | 30,37,24,15,30,42 | | 0,0,0,0,0,0 |
| | | 137659 | 30,38,25,14,50,48 | | 0,0,0,0,0,0 |
| | | 137856 | 30,36,24,15,32,40 | | 0,0,0,0,0,0 |
| | 2 | 186998 | 40,50,40,20,40,60 | 867222 | 0,0,0,0,0,0 |
| | | 160912 | 30,50,32,20,40,60 | | 0,0,0,0,0,0 |
| | | 160455 | 30,50,32,20,30,60 | | 0,0,0,0,0,0 |
| | | 186751 | 40,50,32,20,40,60 | | 0,0,0,0,0,0 |
| | | 172106 | 40,55,32,28,30,72 | | 0,0,0,0,0,0 |
| | 3 | 139967 | 30,37,26,17,30,46 | 696718 | 0,0,0,0,0,0 |
| | | 139204 | 30,36,24,15,30,50 | | 0,0,0,0,0,0 |
| | | 138765 | 30,38,24,15,30,42 | | 0,0,0,0,0,0 |
| | | 139258 | 30,38,24,15,31,48 | | 0,0,0,0,0,0 |
| | | 139524 | 32,39,24,17,30,48 | | 0,0,0,0,0,0 |

Table 10 is the model optimization comparison results based on the inventory node $s_k = [60,80,50,45,50,80]$. It can be seen from the table that the consumption of the adjustment model is lower so that the cost of the traditional model is lower than that of the improved model in this paper. However, from the comparison optimized results based on the inventory node $s_k = [20,20,10,25,10,8]$ in Table 9, the consumption of the adjustment model proposed in this paper is lower. At the same time, when the inventory node is in this state, the traditional model cannot meet the consumption in the supply period because the consumption of spare parts is not considered. This leads to delays and breakdown loss.

Although the cost of the traditional model is lower than that of this paper when setting lower supply nodes, considering the fault tolerance and stability of the whole model, the model proposed in this paper can meet the lower demand of inventory nodes while ensuring the stable supply of spare parts. It is in order to ensure that there are no delays and delay losses as much as possible and make it more stable and reliable.

*6.4. Comparison between Improved Algorithm and Traditional Algorithm.* In order to analyze the effect of the improved algorithm, the joint model is also calculated by the traditional PSO algorithm. The comparison result is shown in Table 11.

From Table 11, it is very obvious that the results of the first and second calculation through the traditional algorithm are in the position of local optimization. The system calculates the final result only once. Even the third result is

higher than the result of improved algorithm. However, with the effect of algorithm error, the final results of them can be seen as the same.

## 7. Conclusion

In this paper, the joint policy combines the inventory policy and spare part supply network. In the joint model, the multiperiod and multiechelon supply network is built, and the $(s, S)$ policy is improved by the random lead time and different customers' maximum inventory. Due to the nonlinear, nonmonotonic, and multiperiodic changes of the established model, an improved PSO algorithm is proposed. The algorithm used in this paper is optimized by adding adaptive inertia weight and penalty function to speed up the optimization efficiency and improve the convergence effect. A case is given. The optional supply scheme is obtained by the proposed algorithm. The sensitivity analysis is used to discuss the influence of important parameters on the model cost. The statistical characteristics of the model are summarized to provide a reference for the next intelligent decision. Except that, the comparison results concluding the traditional $(s, S)$ model and the traditional PSO algorithm are analyzed. We hope that the joint policy and used method can provide a reference for the spare part supply in industry and military.

In this paper, the parameters in the inventory policy are given as known, which can regard as optimized objects in the future research. In addition, it can continue to study the supply of multikind spare parts under different lifetime distribution and so on.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] C. C. Sherbrooke, "VARI-METRIC: improved approximations for multi-indenture, multi-echelon availability models," *Operations Research*, vol. 34, no. 2, pp. 311–319, 1986.

[2] Q. Hu, S. Chakhar, S. Siraj, and A. Labib, "Spare parts classification in industrial manufacturing using the dominance-based rough set approach," *European Journal of Operational Research.*, vol. 262, no. 3, pp. 1136–1163, 2017.

[3] A. A. Ghobbar and C. H. Friend, "Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model," *Computers & Operations Research.*, vol. 30, no. 14, pp. 2097–2114, 2003.

[4] R. Min, Q. Chen, and Z. Shen, *Spare Parts Supply Science*, National Defense Industry Press, 2013.

[5] R. P. Covert and G. C. Philip, "An EOQ model for items with Weibull distribution deterioration," *A I I E Transactions.*, vol. 5, no. 4, pp. 323–326, 1973.

[6] S. Bashyam and M. C. Fu, "Optimization of (s, S) inventory systems with random lead times and a service level constraint," *Management Science.*, vol. 44, no. 12-part-2, pp. S243–S256, 1998.

[7] S. Osaki, "An ordering policy with lead time," *International Journal of Systems Science.*, vol. 8, no. 10, pp. 1091–1095, 1977.

[8] M. Issa, A. E. Hassanien, D. Oliva, A. Helmi, I. Ziedan, and A. Alzohairy, "ASCA-PSO: adaptive sine cosine optimization algorithm integrated with particle swarm for pairwise local sequence alignment," *Expert Systems with Applications*, vol. 99, pp. 56–70, 2018.

[9] C. C. Sherbrooke, "Metric: a multi-echelon technique for recoverable item control," *Operations Research*, vol. 16, no. 1, pp. 122–141, 1968.

[10] T. S. Vaughan, "Failure replacement and preventive maintenance spare parts ordering policy," *European Journal of Operational Research.*, vol. 161, no. 1, pp. 183–190, 2005.

[11] G. P. Cachon, "Exact evaluation of batch-ordering inventory policies in two-echelon supply chains with periodic review," *Operations Research*, vol. 49, no. 1, pp. 79–98, 2001.

[12] W. J. Kennedy, J. Wayne Patterson, and L. D. Fredendall, "An overview of recent literature on spare parts inventories," *International Journal of Production Economics.*, vol. 76, no. 2, pp. 201–215, 2002.

[13] U. S. Rao, "Properties of the periodic review (R, T) inventory control policy for stationary, stochastic demand," *M&SOM.*, vol. 5, no. 1, pp. 37–53, 2003.

[14] Y. Wang and Q. Shi, "Improved dynamic PSO-based algorithm for critical spare parts supply optimization under (T, S) inventory policy," *IEEE Access.*, vol. 7, pp. 153694–153709, 2019.

[15] M. C. Reade, A. Delaney, M. J. Bailey et al., "Prospective meta-analysis using individual patient data in intensive care medicine," *Intensive Care Medicine*, vol. 36, no. 1, pp. 11–21, 2010.

[16] L. Spanjers, J. C. W. van Ommeren, and W. H. M. Zijm, "Closed loop two-echelon repairable item systems," *OR Spectrum*, vol. 27, no. 2-3, pp. 369–398, 2005.

[17] B.-T. Aharon, G. Boaz, and S. Shimrit, "Robust multi-echelon multi-period inventory control," *European Journal of Operational Research.*, vol. 199, no. 3, pp. 922–935, 2009.

[18] P. K. Aggarwal and K. Moinzadeh, "Order expedition in multi-echelon production/distribution systems," *IIE Transactions.*, vol. 26, no. 2, pp. 86–96, 1994.

[19] A. Federgruen and P. Zipkin, "A combined vehicle routing and inventory allocation problem," *Operations Research*, vol. 32, no. 5, pp. 1019–1037, 1984.

[20] J. Shu, C.-P. Teo, and Z.-J. M. Shen, "Stochastic transportation-inventory network design problem," *Operations Research*, vol. 53, no. 1, pp. 48–60, 2005.

[21] A. K. Saha, A. Paul, A. Azeem, and S. K. Paul, "Mitigating partial-disruption risk: a joint facility location and inventory model considering customers' preferences and the role of substitute products and backorder offers," *Computers & Operations Research.*, vol. 117, p. 104884, 2020.

[22] S. Ekinci, D. Izci, and B. Hekimoğlu, "Optimal FOPID speed control of DC motor via opposition-based hybrid Manta ray foraging optimization and simulated annealing algorithm," *Arabian Journal for Science and Engineering*, vol. 46, no. 2, pp. 1395–1409, 2021.

[23] J. F. Farfán and L. Cea, "Coupling artificial neural networks with the artificial bee colony algorithm for global calibration of hydrological models," *Neural Computing and Applications*, vol. 33, no. 14, pp. 8479–8494, 2021.

[24] A. S Sakthivel, A. D Mary, R. Vetrivel, and V. S. Kannan, "Optimal location of SVC for voltage stability enhancement under contingency condition through PSO algorithm," *International Journal of Computer Applications*, vol. 20, no. 1, pp. 30–36, 2011.

[25] M. Clerc and J. Kennedy, "The particle swarm - explosion, stability, and convergence in a multidimensional complex space," *IEEE Trans. Evol. Computat.*, vol. 6, no. 1, pp. 58–73, 2002.

[26] Wen-Fung Leong and G. G. Yen, "PSO-based multiobjective optimization with dynamic population size and adaptive local archives," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 5, pp. 1270–1293, 2008.

[27] E. Mezura-Montes and C. A. Coello Coello, "An improved diversity mechanism for solving constrained optimization problems using a multimembered evolution strategy," in *Genetic and Evolutionary Computation–GECCO 2004*, K. Deb, Ed., pp. 700–712, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2004.

[28] Q. Hu, J. E. Boylan, H. Chen, and A. Labib, "OR in spare parts management: a review," *European Journal of Operational Research.*, vol. 266, no. 2, pp. 395–414, 2018.

[29] M. Z. Ruan, Q. M. Li, Y. W. Peng, E. S. Ge, and A. L. Huang, "Model of spare part fill rate for systems of various structures and optimization method," *Systems Engineering and Electronics.*, vol. 33, pp. 1799–1803, 2011.

[30] A. Mahor and S. Rangnekar, "Short term generation scheduling of cascaded hydro electric system using novel self adaptive inertia weight PSO," *International Journal of Electrical Power & Energy Systems.*, vol. 34, no. 1, pp. 1–9, 2012.

[31] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95- International Conference on Neural Networks, IEEE*, pp. 1942–1948, Perth, WA, Australia, 1995.

[32] K. Deep, "Madhuri: application of globally adaptive inertia weight PSO to Lennard-Jones problem," *Proceedings of the International Conference on Soft Computing for Problem Solving (Soc ProS 2011) December 20-22, 2011*, K. Deep, A. Nagar, M. Pant, and J. C. Bansal, Eds., , pp. 31–38, Springer India, India, 2012.

[33] H. Shao and G. Zheng, "Boundedness and convergence of online gradient method with penalty and momentum," *Neurocomputing*, vol. 74, no. 5, pp. 765–770, 2011.

[34] P. Sincak, "Intelligent technologies-theory and applications: new trends in intelligent technologies," IOS Press, Ohmsha, Amsterdam; Washington, DC: Tokyo, 2002.

[35] A. C. Nearchou, "The effect of various operators on the genetic search for large scheduling problems," *International Journal of Production Economics.*, vol. 88, no. 2, pp. 191–203, 2004.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9956207, 11 pages
https://doi.org/10.1155/2021/9956207

*Research Article*

# TagNN: A Code Tag Generation Technology for Resource Retrieval from Open-Source Big Data

**Lingbin Zeng**,[1] **Xin Guo**,[1] **Cheng Yang**,[2] **Yao Lu**,[2] **and Xiao Li**[1]

[1]*Technical Service Center for Vocational Education, National University of Defense Technology, Changsha 410073, China*
[2]*College of Computer, National University of Defense Technology, Changsha 410073, China*

Correspondence should be addressed to Xiao Li; xiaoli@nudt.edu.cn

Received 4 June 2021; Accepted 2 August 2021; Published 26 August 2021

Academic Editor: Mamoun Alazab



With the vigorous development of open-source software, a huge number of open-source projects and open-source codes have been accumulated in open-source big data, which contains a wealth of code resources. However, effectively and efficiently retrieving the relevant code snippets in such a large amount of open-source big data is an extremely difficult problem. There are usually large gaps between the user's natural language description and the open-source code snippets. In this paper, we propose a novel code tag generation and code retrieval approach named TagNN, which combines software engineering empirical knowledge and a deep learning algorithm. The experimental results show that our method has good effects on code tag generation and code snippet retrieval.

## 1. Introduction

With the vigorous development of the open-source software, the resources contained in the open-source big data are becoming increasingly abundant [1–3], including not only open-source software artifacts but also software development behavior data and auxiliary documentation such as user manuals and technique reports. On the one hand, the rapid development of the open-source big data has provided software developers with a huge amount of software code snippets. On the other hand, the explosive growth of open-source resources has brought considerable challenges to the retrieval of open-source resources, especially code snippets. Part of the current research work has been aimed at the classification of open-source resources. For example, Stack Overflow (http://stackoverflow.com/) classifies query questions into different levels with different tags. These classification levels and tag settings are determined by specialized domain experts who spend considerable time and effort. Such a manual process is difficult to scale to the explosive development of the open-source big data [4]. Take the famous Linux kernel as an example. According to statistics (http://www.theregister.com/2020/01/06/linux_2020_kernel_systemd_code/), the Linux kernel has 27.8 million lines of code in the Git repository in 2020, increasing more than 1.7 million lines compared with 2019. Such a phenomenon is common in the open-source community. This brings great challenges to human annotation work.

Open-source code snippets are developing at such a rapid pace that traditional manual tagging is far from adequate. Thus, the automatic tagging and classification of open-source resources have begun to attract scholarly attention. Wang et al. [5] use the existing tag of an open-source community (e.g., Stack Overflow) to tag software using classification algorithms such as SVM. Zhou [6] puts forward a tool named *TagMulRec* which contained an efficient tag-based multiclassification algorithm that could deal with a mass of software. These tasks are aimed at constructing tags at the open-source software or project level.

However, with the development of the open-source big data, the reuse of open-source resources by software developers has dived into the level of specific source codes rather

than just using general mature open-source software products. Consequently, generating tags for natural language queries is an effective method to enable software developers to quickly retrieve high-quality open-source code snippets.

The traditional manual tagging model has difficulty in supporting the generation of code tags for massive natural language query statements. In recent years, the rapid development of deep learning methods has many mature applications, such as image recognition, speech recognition, and language translation. This also brings great opportunities for code tag generation [7, 8]. Unlike traditional machine learning models, training deep learning models require massive amounts of data. At this point, the open-source big data has massive resources for model training. The rapid development of deep learning in natural language processing has a certain enlightening effect on the tag generation of open-source resources.

Based on the above survey, we propose a technology to automatically generate code tags for natural language queries, in the hope of better guiding software developers to effectively retrieve code snippets, accelerating the application of open-source resources by software developers, and promoting the growth and expansion of open-source software. Specifically, we propose a novel code tag generation and code retrieval approach named *TagNN*. The model combines the term frequency-inverse document frequency (TF-IDF) algorithm and the recurrent neural network (RNN) framework. Our experiments found that simply generating a code sequence of a certain length through a short natural language description is not ideal. Therefore, we combine the deep learning method and the empirical knowledge of software engineering to generate the code tags for the corresponding natural language description, thereby facilitating the efficiency of code retrieval. The experimental results demonstrate that our method designed in this paper offers good performance results in improving the efficiency of code retrieval. The key contributions of this paper are as follows:

(i) A new code tagging framework that combines deep learning and software engineering empirical knowledge

(ii) A large-scale dataset of Java code tags that contains 717,980 pairs of text summaries and code snippets

(iii) A novel code tag mining framework that leads to a significant improvement of code retrieval compared to the state-of-the-art methods

The rest of this paper is organized as follows. Section 2 reviews previous works. Section 3 describes our methods. Section 4 shows the experimental design and results, and the last section concludes this paper.

## 2. Related Work

Many works have been proposed for code tagging and code retrieval. We will review these works in three categories including code retrieval, deep learning, and tag measurement.

*2.1. Code Retrieval.* In recent years, code retrieval has become a research topic of interest in software engineering. Researchers have proposed a variety of code retrieval methods. These studies cover multiple aspects, provide multiple forms of input, and recommend code resources at various levels. The following reviews typical research work about code retrieval.

INQRES [9] considers the relationship between each pair of words in the source code and interactively reconstructs the search query to optimize the query quality. Bajracharya [10] proposes a systematic model that takes a natural language query as input, finds the source code implementation of the corresponding function and the calling methods of existing code snippets, and uses the TF-IDF method and boosting technology to identify popular classes. XSnippet [11] takes a dedicated query statement as input. The advantage of XSnippet is that it divides the query into two steps to expand the scope of the query. Sourcerer [12] is a code retrieval tool based on Lucene that combines code attributes and code popularity as an indicator to evaluate the quality of recommended codes and then retrieves relevant code snippets.

The above works have their own characteristics in the study of code retrieval. They have contributed corresponding solutions to the code retrieval problem by analyzing the empirical knowledge of software engineering and the laws of natural language. However, retrieval patterns and code tag characteristics cannot be exhausted through manually constructed rules. The TagNN method proposed in this paper combines the characteristics of external rules with data-driven generalization ability through deep learning, which has certain innovations.

*2.2. Application of Deep Learning in Natural Language Processing.* The research area of deep learning in natural language processing has focused on sentence-level or document-level text representation and classification methods as follows.

UNIF [13] uses the attention mechanism to combine the embedding of each token in the code snippet and generates an embedding vector representation of the entire code fragment. Pennington et al. [14] propose using the global "word-word" co-occur matrix to obtain a word vector representation in the GloVe method. Hill et al. [15] propose learning distributed expressions corresponding to sentences from unlabeled data. Conneau [16] proposes using supervised learning to learn general sentence representations from natural language inference data. Tai et al. [17] improve the semantic representation model of long- and short-term memory networks with tree structures.

These works use deep learning methods to process and analyze natural language. However, few works applied deep learning methods to the code snippet data. Moreover, although the composition of the code is similar to natural language expression, there are still some differences. Different from other works, TagNN uses the empirical knowledge of software engineering to process the code snippets so that the deep learning method could effectively process the code snippets with good results.

*2.3. Tag Measurement.* The quality of tags is mainly measured from two aspects: similarity and generalization. The similarity measure indicates the distance between tags. The higher the value is, the closer the meaning or the stronger the association is. Wang et al. [18] integrated tag annotation through a labeling system that exists in the open-source community itself and then measured the textual similarity of open-source resources on this basis. Begelman et al. [19] measured the similarity of tags by the number of times the tags appear together.

The generalization degree of a tag represents the number of categories contained in the tag. The larger the value is, the higher the level in the tag hierarchy. Schmitz [20] adjusted the threshold to control the usage of special tags and increased the special vocabulary in the filter to improve the quality of the generated tags.

The above are all very classic works in this field which mainly judge the quality of the tag by measuring the attributes of the tag itself. In this paper, TagNN will judge whether tags are good or bad based on their effectiveness in assisting code retrieval that is different from the above work.

## 3. Overview of TagNN

In this section, we describe the framework of our approach. TagNN involves two main methods. One is a tag generation model based on deep learning methods, and the other uses TF-IDF to extract keywords from generated code snippets [21–24]. The deep learning model has a good effect on natural language processing, and TF-IDF has a good effect on extracting key information in the text. Thus, we combine the two types of methods for code tag generation. As shown in Figure 1, TagNN consists of five parts: data collection, training data processing, model training, model testing, and code retrieval.

Figure 1 shows the framework of our approach, which consists of five steps for code retrieval with code tags. This section describes the first three parts (i.e., data collection, training data processing, and model training), and Section 4 describes the last two parts (i.e., model testing and code retrieval).

We give a brief description of the first three parts. First, we obtain a large number of high-quality open-source projects from the open-source community, from which we extract code snippets and corresponding summary information. Then, we analyze the summary information according to the empirical knowledge of the natural language and process the code snippets according to the software engineering domain knowledge to obtain high-quality summary information and code snippets. Next, with the previously obtained dataset, we train the deep learning model. In addition, based on the trained model, we input the natural language description information to acquire the corresponding code tags. Finally, the generated code tags are passed to the corresponding code retrieval method to improve the efficiency of code retrieval.

*3.1. Data Collection.* We use the *Kraken* (https://forgeplus.trustie.net/projects/zenglingbin12/summer_all) [25] tool to acquire data, and the specific steps are shown in Figure 2.

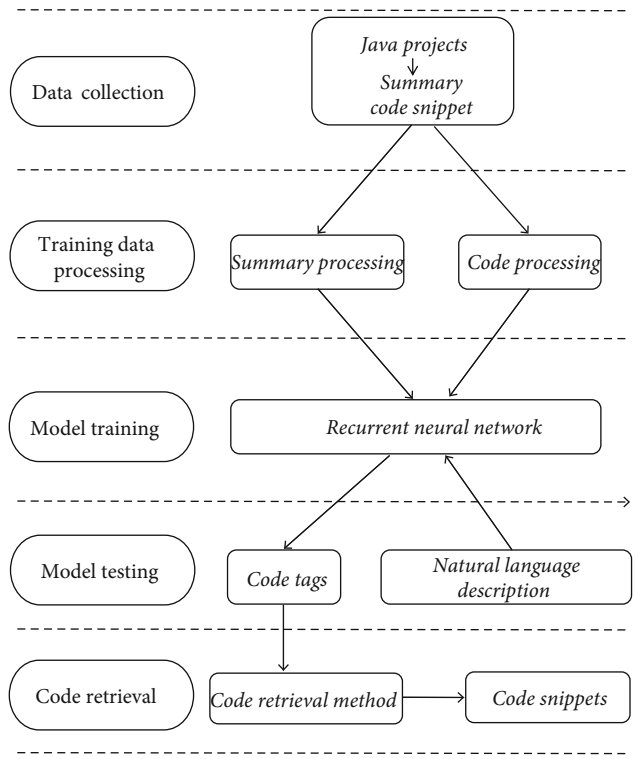



Figure 1: The framework of the TagNN.

*Step 1* (Project requirements). First, we need to clarify the type, quantity, and sorting criteria of the project to be obtained. However, measuring the quality of open-source projects is a complex system engineering problem. In this paper, we adopt the concept of crowd intelligence and use the star mechanism in GitHub to filter projects. Specifically, when users in the open-source community like the project, they can give the project a star. The more stars a project has, the more people acknowledge the project. Based on this concept, we collected the top-ranked projects in GitHub as the source of projects.

*Step 2* (Project lists). According to the requirements mentioned before, we use the API provided by GitHub to obtain the metadata of all projects through the *Kraken* tool and then analyze the data and sort out the project list.

*Step 3* (Cloning projects). According to the project lists, we use the protocol provided by the Git tool to clone and store the remote projects locally. Because the protocol provided by the Git itself is single-threaded, it is difficult to clone projects concurrently on a large scale. Thus, we design a multithreaded concurrent algorithm to clone projects.

*Step 4* (Extracting code files). Every project contains many different types of files, including documentation, technical manuals, and source codes. We need to filter out the source code files. In our work, we filter the corresponding code files by file name suffix matching. In addition, it should be noted that we believe that each project has only one main programming language. For projects composed of multiple programming

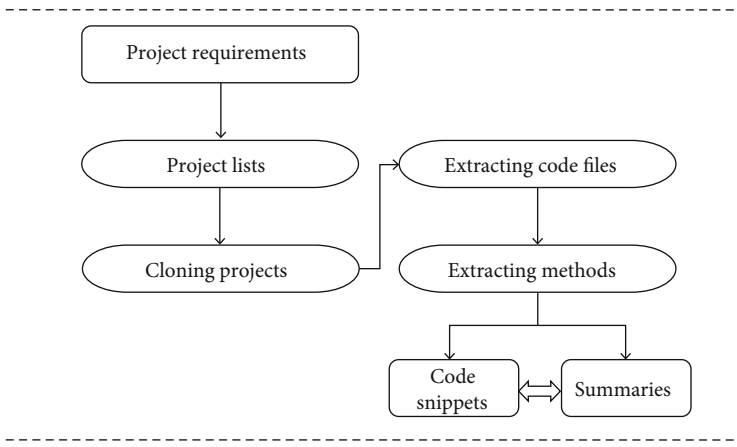



Figure 2: Data collection processing.

languages, we clarify the main programming language of the project through the information provided by GitHub.

*Step 5* (Extracting methods). For the obtained source code files, we use the characteristics of the programming language and the empirical knowledge of natural language to filter out method-level code snippets and corresponding description information in the code files.

*Step 6* (Paired dataset). The method-level code snippets and summaries are stored in a local database as a basic dataset for model training and testing.

### 3.2. Training Data Processing

*3.2.1. Summary Processing.* In the training data processing module, we process information of code snippets and code summaries separately. We believe that the quality of code snippets and code summaries extracted from high-quality projects in the GitHub community is generally high. However, the open-source community has a large number of contributors, and their mastery levels are uneven, which inevitably leads to uneven code summary quality. There could be invalid or nonfunctional descriptions. Based on this consideration, through the observation of the code summary information and the understanding of language grammar rules, we developed heuristic rules to filter code descriptions. Figure 3 shows our process and rules for filtering code summaries. We have a total of six steps for summary processing.

*Step 1* (Remove @ block information). For programmers who use the integrated development environment, as shown in Figure 4, if they create a comment after writing the code, the integrated development environment automatically adds some predefined information for the class-level and method-level code snippets. Take the well-known integrated development environment *Eclipse* as an example. It can automatically generate information such as "@author," "@data," and "@return." Although this information can help developers understand the code better, they are not functional descriptions. Therefore, in this step, we remove the code summaries that contain "@" block information.

*Step 2* (Remove other @ information). After removing the "@" block information in the first step, there is still some "@" information added by the software developer in the code summaries. As shown in Figure 5, "@link" indicates the class related to the object. In addition, there is still information similar to "@deprecated" and "@code." We use regular expressions to remove the code summary information that contains these "@" information.

*Step 3* (Remove web page information). Through the analysis of the code summary data, as shown in Figure 6, we found that there are many web page tag elements to better display the code summary information. However, these web page tags are noisy data for the code description data, which is not good for model training. We use regular expressions to remove these web page tag data.

*Step 4* (Remove punctuation information). The code summary is written by the software developer during the code development, which contains many punctuation marks (e.g., ",", ".", "?", "!", ":", and ";") and other symbols. To reduce the noisy influence of punctuation marks on the code summary information, as shown in Figure 7, we remove the punctuation marks and remove the summary information containing the question mark.

*Step 5* (Remove non-English vocabulary). For the method level, we only consider the functional summaries of the method. Therefore, as shown in Figure 8, we remove the summary information that contains non-English words. In this step, we use Python's *pyenchant* module to check each word described.

*Step 6* (Remove the description that is too short). In this step, we remove the ambiguous code description information.

As shown in Figure 9, this code summary information can only be understood by the code writer. Generally, a complete code description in English must have at least one verb

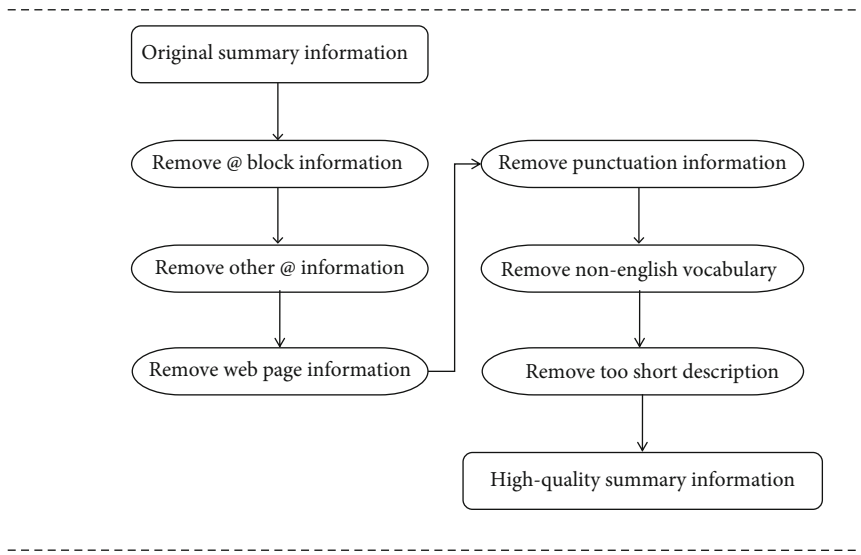



Figure 3: Summary information processing.

```
/**Write a localized message, using the default resource bundle.
   *@param key the key for the message to be localized
   *@throws IOException if there is a problem closing the underlying stream
 */
public void write I18N (String key) throws IOException {
  write (getString (i18n, key)) ;
}
```

Figure 4: Summaries containing "@" block information.

```
/**Returns the string for rendering the{@link IJavaElement#getElementName() element name} of
    * the given element.
   */
protected String getElementName (IJavaElement element) {
   return element.getElementName();
}
```

Figure 5: Summaries containing other "@" information.

```
/**Returns the coefficient of determination <em>R</em><sup>2</sup>.
  * @return the coefficient of determination <em>R</em><sup>2</sup>,
which is a real number between 0 and 1
 */
public double R2() {
   return R2 ;
}
```

Figure 6: Summaries containing web page information.

```
 /*??  '  ?  '   ?  '  /'   ?????* http://mavin-manzhan.oss.-cn-hangzhou
 aliyuncs.com/  ...
 */
 private static String getUrlFileName(String url) {
        String filename = null;
        String[] strings = url.split("/");
        ...
        return filename; }
```

Figure 7: Summaries containing punctuation information.

and one object. Therefore, we have deleted descriptions that are fewer than two words.

*3.2.2. Code Processing.* In this section, we deal with method-level code snippets to obtain code tags, as shown in Figure 10.

*(1) Code Segmentation.* As shown in Figure 11, a piece of the original code snippet is successfully segmented after six steps of processing.

*Step 1* (Remove parentheses). Parentheses in the code snippets have no actual special meaning. Thus, we replace them with blanks in the first step.

*Step 2* (Remove punctuation information). We delete the punctuation information of the code snippet as we did for the summary before.

*Step 3* (Remove underlining). When writing Java codes, some programmers are accustomed to using underscores in

```
/*writes data to a random filename
 (update_<per JVM random UUID>_<COUNTER>.tmp)
*/
private static DiskFileItem write ( String dir, byte[] data ) throws IOException, Exception
  {
      return makePayload(data.length + 1, dir, dir + "/whatever", data);
  }
```

Figure 8: Summaries containing non-English vocabulary.

```
// where
void findFiles (File dir, Set<File> files) {
    for (File f: dir.listFiles()) {
        if(f.isDirectory())
        findFiles(f, files);
     else
        files.add(f);
    }
}
```

Figure 9: Overly short description.

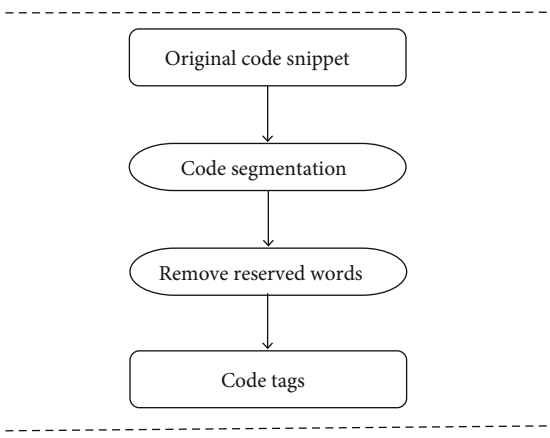



Figure 10: The key steps of code processing.

variable names to connect nouns. We delete the underscores and decompose the variable names.

*Step 4* (Dividing codes). The Java code writing rules follow the camel case naming rules, so we decompose the variable names based on the camel case rules and decompose each independent word from it.

*Step 5* (Remove Arabic numbers). The Arabic numbers themselves do not represent special meanings, so we remove the Arabic numbers that exist after the code participles.

*Step 6* (Lowercase vocabulary). After the final processing, we uniformly convert the vocabulary to lowercase and finally output high-quality code snippets.

*(2) Remove Reserved Words*. The keywords of the Java code exist in the system itself and have little meaning for the characterization of the function itself [26]. Therefore, when we obtain the code snippets processed in the first step, we process the code and delete the Java keywords.

*(3) Generate Tags*. After the first and second steps, the code snippets become code vocabulary sequences. We use the TF-IDF algorithm to select the most representative words for each code snippet in the entire training set. TF-IDF is a statistical method that evaluates the importance of a word to one of the documents in a document set. The importance of a word increases in proportion to the number of times it appears in the document, but at the same time, it decreases in inverse proportion to the frequency of its appearance in the corpus [27, 28]. Through the TF-IDF algorithm, we generate the top ten important words in each code snippet. The ten words are used as the tag of the code snippet to characterize it.

*3.3. Model Training.* After the data are processed in the second step, TagNN implements the construction of the model through a recurrent neural network (RNN) algorithm. We choose the classic Encoder-Decoder model that is often used in natural language processing. Next, we will give a general introduction to the selected model.

*3.3.1. Input.* We use the natural language description processed by heuristic rules as the input of the model.

*3.3.2. The Basic Theory of the Model.* The RNN is a classic neural network model. It is composed of an input layer, a hidden layer, and an output layer. For the convenience of description, we use $d$ to refer to the input layer, $t$ to refer to the output layer, and $h$ to refer to the hidden layer. The depth of the hidden layer can be set flexibly. The state of the hidden layer $h$ is changed by its previous state and the influence of the state $d$, and finally, $t$ is affected by the cumulative network weight propagation [29–31]. Besides, this paper uses the LSTM as an activation function [32], which has a good effect on natural language processing. Next, we will introduce the Encoder and Decoder in detail.

*(1) Encoder*. The Encoder is essentially an RNN. This paper takes the natural language description of the code snippet as the input sequence $D$ which inputs the words into the model one by one from the head position to the tail position, and the state of the corresponding hidden layer changes accordingly. After the $D$ sequence is input and processed by the hidden layer, the Encoder will output the intermediate state $m$, as shown in Figure 12 [30].

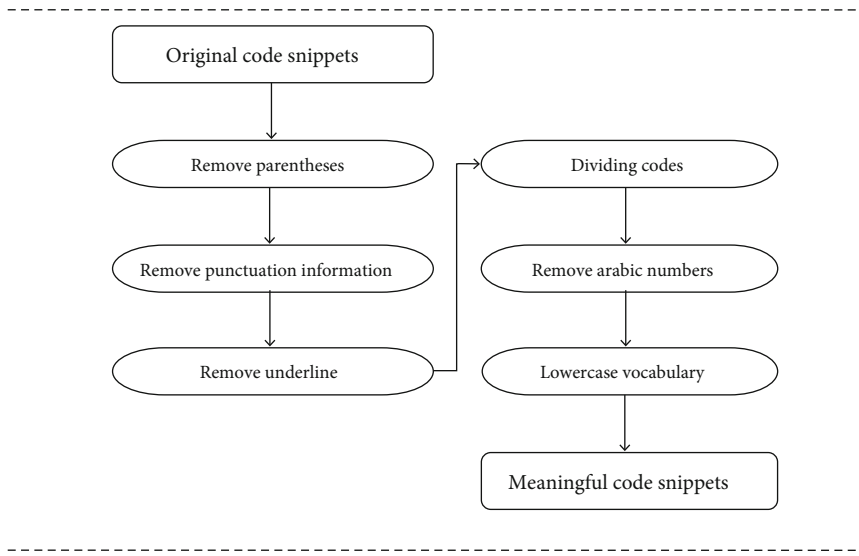



Figure 11: The key steps in code snippet processing.

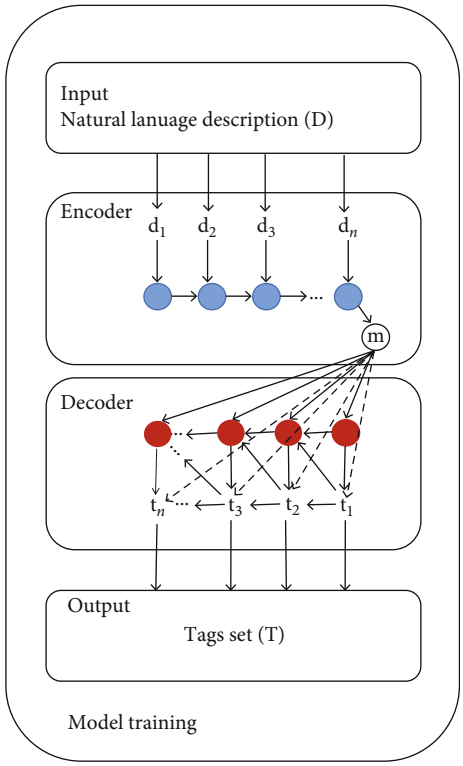



Figure 12: The architecture of model training.

*(2) Decoder*. The Decoder is also an RNN like the Encoder. When the Encoder outputs the state $m$ to the Decoder, the Decoder will output $t_i$ one by one, which is a tag that is used to measure the code snippets. Finally, the Decoder will output the sequence $T$ which is the set of tags.

*3.3.3. Output.* The output of the model is a set of tags that we want to measure the code snippets.

## 4. Experimental Design and Effect Verification

To demonstrate the validity of our TagNN model, we designed two related problems and conducted corresponding experiments. This section introduces experimental data and evaluates our experimental results.

### *4.1. Experimental Setup*

*4.1.1. Model Settings.* Table 1 shows the basic parameters of TagNN. We implement TagNN with the famous TensorFlow [33] framework. The model has six hidden layers, each of which has 128 neurons. The neuron type is LSTM, and the learning rate is set as 0.5.

*4.1.2. Data Settings.* We design an experiment based on GitHub's Java projects, from which we selected the top 5,000 Java projects based on the ranking of stars. After screening and distinguishing abstracts, we selected 717,980 summary-code pairs that met the conditions. As shown in Table 2, we use 80% of the data for the training set, 10% for the validation set, and 10% for the test set.

*4.2. Research Question.* To study the effect of TagNN on tag generation and whether the generated tags can help improve the search and retrieval of open-source codes through natural language, we propose the following two research questions:

*(1) Question 1*. What is the effect of generating code tags for natural language through TagNN?

*(2) Question 2*. Can the code tags generated by TagNN improve the accuracy of natural language retrieval codes?

For the first question, we analyzed the accuracy of the code tags generated by TagNN, and for the second question,

Table 1: Basic parameters of TagNN.

| Training tools | TensorFlow |
|---|---|
| Hidden layers | 6 |
| Number of neurons per layer | 128 |
| Neuron type | LSTM |
| Learning rate | 0.5 |

Table 2: Experimental dataset.

| The total amount of experimental data | 717,980 |
|---|---|
| Training set | 80% |
| Validation set | 10% |
| Test set | 10% |

Table 3: The effect of TagNN.

| The total amount of test data | 71,798 |
|---|---|
| Accuracy rate | 78.03% |
| Recall rate | 31.00% |

we used the traditional code search matching method plus the code tag data to see how it affects the search results.

*4.3. Metric Methods.* For problem one, the experiment uses the accuracy index. When the predicted tag generated by the model matches the true tag of the code snippet, the accuracy index is one; otherwise, the value is zero. In addition, we use the index of the recall rate, which is the probability that the true tags of the code snippets appear in the tags generated by the model.

*4.4. Results.* This section presents the final results of the above two experiments and targeted comparative analysis.

*4.4.1. The Effect of Generating Code Tags for Natural Language through TagNN.* We use the TagNN model trained by the RNN to read the 71,798 natural language descriptions in the test set and correspondingly generate 71,798 code tags. As shown in Table 3, the accuracy rate of tag generation is 78.03%, and the recall rate is 31.00%. The experimental results prove that our model has a high accuracy rate and a reasonable recall rate. To the best of our knowledge, this is the first time that code tags have been generated through natural language by deep learning methods to facilitate the retrieval of code snippets in natural language. The accuracy rate of 78.03% demonstrates that TagNN has a good effect in generating code tags for natural language.

*4.4.2. The Impact of Code Tags Generated by TagNN on Natural Language Retrieval Codes.* TagNN generates code tags by describing natural language to improve the efficiency of code retrieval. The role of TagNN tags is to allow existing code retrieval methods to achieve better results after using tags.

We selected the classic TF-IDF algorithm for code retrieval. Through the TF-IDF algorithm, we measure the similarity scores of a single natural language description and each piece of code and sort them with the similarity score from highest to lowest. We believe that if the target code snippet appears in the top ten code snippets, then the search task is successful. As shown in Table 4, before adding tags, the retrieval accuracy rate was 34.2%, while after adding tags, it was 40.03%, an increase of approximately 17.04%. This means that the code tags generated by TagNN are of great help in improving the accuracy of natural language retrieval codes.

Table 4: The boost effect of TagNN.

| Algorithm | Accuracy rate |
|---|---|
| Classic TF-IDF algorithm | 34.2% |
| TF-IDF algorithm with code tags | 40.03% |
| Boost effect | 17.04% |

Table 5: Response categories for relevance evaluation.

| Scale | Response category |
|---|---|
| 5 | Very relevant |
| 4 | Relevant |
| 3 | Neither relevant nor irrelevant |
| 2 | Irrelevant |
| 1 | Very irrelevant |

*4.5. User Study.* Through the above data analysis, we verified the accuracy of code tag generation and the effect of code tag-assisted code retrieval, which effectively proved the effectiveness of TagNN at the data level. To further measure the effectiveness of TagNN, we conducted a user study.

We invited 16 students with different backgrounds to evaluate the results. Among them, there are eight software development engineers, four master's students, and four doctoral students, all of whom have no less than five years of Java software development experience. They were asked to analyze the relevance of our tags and code snippets and the effectiveness of tags to help code retrieval.

*4.5.1. Relevance Evaluation.* We ask students to evaluate the relevance between the tags generated by TagNN and the corresponding code snippets. The evaluation uses the Likert-type method [34], which has a variety of different expression choices and can effectively measure users' agreement with the relevance of tags and code snippets. Users need to choose one of five candidate options to express their agreement with the degree of relevance. Table 5 indicates these options.

As shown in Figure 13, among the 10 samples, there were six samples with a median of 4, three samples with a median of 4.5, and one sample with a median of 3.5. The median average value is 4.1, which is more than 4. This shows that the tags generated by our TagNN method are highly correlated with the code snippets, reaching the relevant level.

*4.5.2. Usability Evaluation.* To determine whether the tags generated by TagNN are helpful for code retrieval, we organize users to evaluate the usability of tags for code retrieval. As before, we use the Likert-type method. As shown in

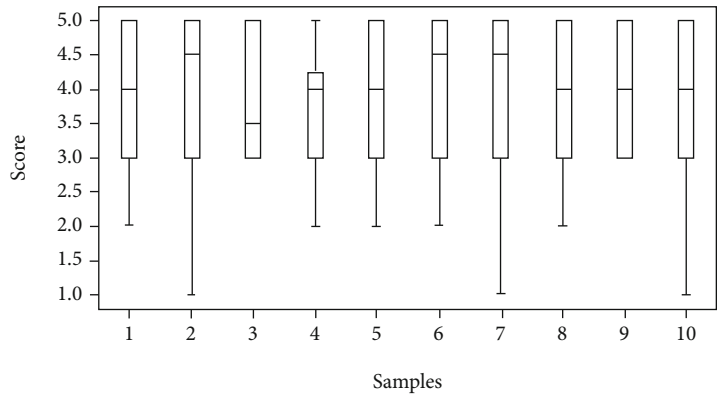



Figure 13: User evaluation on relevance.

Table 6: Response categories for usability evaluation.

| Scale | Response category |
|---|---|
| 5 | Very useful |
| 4 | Useful |
| 3 | Not sure |
| 2 | Useless |
| 1 | Very useless |

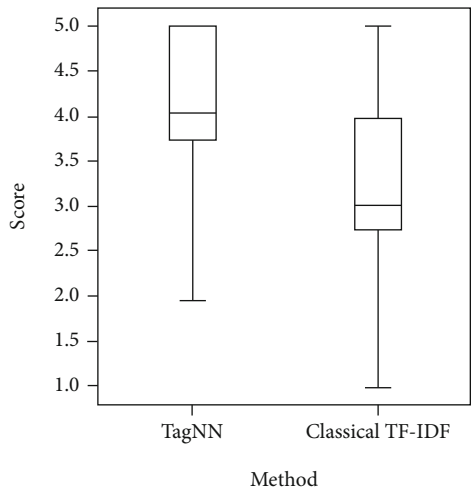



Figure 14: User evaluation on usability.

Table 6, users need to choose one of the five different options that best expresses their attitude.

As shown in Figure 14, the median value of the TagNN method is 4, and the average value is 4.06. The median value of the classic TF-IDF is 3, and the average value is 3.13. Experimental results show that from the views of users, the tags generated by TagNN have reached a useful level for code retrieval and are superior to the classic TF-IDF method.

*4.5.3. Discussion.* Through user evaluation, we found that the TagNN method demonstrates outstanding performance in relevance because the TagNN method combines the characteristics of deep learning methods with natural language and code language, which reflect the important characteristics of the code snippets.

As for the usability evaluation, TagNN has a better performance than the traditional TF-IDF. The reason is that the tags generated by the TF-IDF method are all derived from the code snippet itself, so there will be no vocabulary outside of the code snippet. However, TagNN uses deep learning methods and is trained based on a large amount of data to generate tags that may not be contained by the code snippet itself, which is more flexible and broad-sourced.

## 5. Conclusion and Future Work

In this paper, we aim to address the difficulty of retrieving massive codes in the open-source community to help developers quickly retrieve code resources, thereby speeding up the development efficiency of software developers and realizing the reuse and dissemination of high-quality code resources in the open-source community.

Based on the massive code snippets and natural language description information of the open-source community, we propose a novel code tag generation and code retrieval approach named TagNN, which combines software engineering empirical knowledge and a deep learning framework. Our method generates corresponding code tags for natural language descriptions through the RNN, thereby improving the retrieval effect. With large-scale experiments on the high-quality Java open-source project dataset collected from the GitHub community, we empirically evaluate the code tag generation effect of the model and the tags' role in improving the retrieval of code snippets.

There are still several shortcomings in our work. One is that when we selected the tags of code snippets in the training set, we simply applied the relatively rudimentary TF-IDF algorithm and did not choose a more effective weight measurement algorithm based on the actual situation of the code snippets. Second, we added the tag data directly to the natural language query to retrieve the code snippets without making more effective use of the tag data.

In future work, we will try to carefully analyze and observe the characteristics of the code snippets themselves and propose a more effective method of extracting code tags.

In addition, we will explore a more reasonable and effective application of tags to optimize the code retrieval effect.

## Data Availability

The data used to support the findings of this study are available from the corresponding author (email address: zenglingbin16@nudt.edu.cn) upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work is supported by the National Key Research and Development Program of China (2020AAA0108802), the National Natural Science Foundation of China (61702532), and the scientific research project of the National University of Defense Technology (No. ZK20-47).

## References

[1] A. Terminanto, A. N. Hidayanto, and F. B. Utomo, "Implementation open source system resource planning in sustainable supply chain management of small and medium enterprise," *International Journal of Supply Chain Management*, vol. 9, no. 3, pp. 472–495, 2020.

[2] N. Zöller, J. H. Morgan, and T. Schröder, "A topology of groups: what GitHub can tell us about online collaboration," *Technological Forecasting and Social Change*, vol. 161, article 120291, 2020.

[3] O. G. Glazunova, O. V. Parhomenko, V. I. Korolchuk, and T. V. Voloshyna, "The effectiveness of GitHub cloud services for implementing a programming training project: students' point of view," *Journal of Physics Conference Series*, vol. 1840, no. 1, article 012030, 2021.

[4] E. M. Kavuk and A. Tosun, "Predicting Stack Overflow question tags: a multi-class, multi-label classification," in *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, Seoul, Republic of Korea, 2020.

[5] T. Wang, H. Wang, G. Yin, C. Yang, X. Li, and P. Zou, "Hierarchical categorization of open source software by online profiles," *IEICE Transactions on Information and Systems*, vol. E97.D, no. 9, pp. 2386–2397, 2014.

[6] P. Zhou, "Scalable tag recommendation for software information sites," in *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, Klagenfurt, Austria, 2017.

[7] L. S. Ambati, "Factors influencing the adoption of artificial intelligence in organizations-from an employee's perspective," in *MWAIS 2020 Proceedings*, Des Moines, Iowa, 2020.

[8] S. Sophia and S. P. Rajamohana, "A survey on feature selection based spam review detection using deep learning techniques," *International Journal of Advanced Information and Communication Technology*, pp. 102–108, 2020.

[9] J. Lu, Y. Wei, X. Sun, B. Li, W. Wen, and C. Zhou, "Interactive query reformulation for source-code search with word relations," *IEEE Access*, vol. 6, 2018.

[10] S. K. Bajracharya, "Sourcerer: a search engine for open source code supporting structure-based search," Companion to the Acm Sigplan Symposium on Object-oriented Programming Systems ACM, 2006.

[11] N. Sahavechaphan and K. Claypool, "XSnippet: mining for sample code," in *Proceedings of the 21st annual ACM SIGPLAN conference on Object-oriented programming systems, languages, and applications*, Portland Oregon USA, 2006.

[12] S. Bajracharya, J. Ossher, and C. Lopes, "Sourcerer: an internet-scale software repository," in *2009 ICSE Workshop on Search-Driven Development-Users, Infrastructure, Tools and Evaluation*, Vancouver, BC, Canada, 2009.

[13] J. Cambronero, "When deep learning met code search," in *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Tallinn, Estonia, 2019.

[14] J. Pennington, R. Socher, and C. D. Manning, "GloVe: global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Doha, Qatar, 2014.

[15] F. Hill, K. Cho, and A. Korhonen, "Learning distributed representations of sentences from unlabelled data," 2016, https://arxiv.org/abs/1602.03483.

[16] A. Conneau, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen,Denmark, 2017.

[17] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *Computer Science*, vol. 5, no. 1, 2015.

[18] S. Wang, L. O. David, and L. Jiang, "Inferring semantically related software terms and their taxonomy by leveraging collaborative tagging," in *2012 28th IEEE International Conference on Software Maintenance (ICSM)*, Trento, Italy, 2012.

[19] G. Begelman, P. Keller, and F. Smadja, "Automated tag clustering: improving search and exploration in the tag space," in *collaborative web tagging workshop at WWW2006*, Edinburgh, Scotland, 2006.

[20] P. Schmitz, "Inducing ontology from Flickr tags," in *Collaborative Web Tagging Workshop at WWW2006*, vol. 50, Edinburgh, Scotland, 2006.

[21] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[22] N. Deepa, "A survey on blockchain for big data: approaches, opportunities, and future directions," 2020, https://arxiv.org/abs/2009.00858.

[23] M. Z. Asghar, F. Subhan, H. Ahmad et al., "Senti-eSystem: a sentiment-based eSystem-using hybridized fuzzy and deep neural network for measuring customer satisfaction," *Software: Practice and Experience*, vol. 51, no. 3, pp. 571–594, 2021.

[24] G. R. Bojja and J. Liu, "Impact of IT investment on hospital performance: a longitudinal data analysis," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, Hawaii, USA, 2020.

[25] L. B. Zeng, "Kraken: a continuous incremental data acquisition system for GitHub and Git repositorie," in *Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering*, Beijing, 2017.

[26] J. Gosling, B. Joy, G. Steele, and G. Bracha, *The Java Language Specification*, Addison-Wesley Longman Publishing Co. Inc., 1996.

[27] L. Havrlant and V. Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)," *International Journal of General Systems*, vol. 46, no. 1, pp. 27–36, 2017.

[28] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, Piscataway, NJ, 2003.

[29] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, 2014.

[30] K. Cho, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, https://arxiv.org/abs/1406.1078.

[31] J. C.-W. Lin, Y. Shao, Y. Djenouri, and U. Yun, "ASRNN: a recurrent neural network with an attention model for sequence labeling," *Knowledge-Based Systems*, vol. 212, article 106548, 2021.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] M. Abadi, "TensorFlow: a system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, Savannah, GA, 2016.

[34] S. Jamieson, "Likert scales: how to (ab)use them," *Medical Education*, vol. 38, no. 12, pp. 1217-1218, 2004.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5886119, 9 pages
https://doi.org/10.1155/2021/5886119

*Research Article*

# An Approach for a Next-Word Prediction for Ukrainian Language

**Khrystyna Shakhovska**,[1] **Iryna Dumyn**,[1] **Natalia Kryvinska**,[2] **and Mohan Krishna Kagita**[3]

[1]*Artificial Intelligence Department, Lviv Polytechnic National University, Lviv 79013, Ukraine*
[2]*Department of Information Systems, Comenius University in Bratislava, Bratislava 81499, Slovakia*
[3]*School of Computing and Mathematics, Charles Sturt University, Melbourne, Australia*

Correspondence should be addressed to Iryna Dumyn; irka.shvorob@gmail.com

Received 20 June 2021; Accepted 29 July 2021; Published 15 August 2021

Academic Editor: Rajesh Kaluri



Text generation, in particular, next-word prediction, is convenient for users because it helps to type without errors and faster. Therefore, a personalized text prediction system is a vital analysis topic for all languages, primarily for Ukrainian, because of limited support for the Ukrainian language tools. LSTM and Markov chains and their hybrid were chosen for next-word prediction. Their sequential nature (current output depends on previous) helps to successfully cope with the next-word prediction task. The Markov chains presented the fastest and adequate results. The hybrid model presents adequate results but it works slowly. Using the model, user can generate not only one word but also a few or a sentence or several sentences, unlike T9.

## 1. Introduction

A personal predictive text input system, also known as T9, has become an integral part of our text input lives. Therapists often recommend word prediction as a method of typing speed improvement for users with physical limitations. Even though papers claim that word suggestion affects writing skills, boosted speed could not be mentioned as its benefit if using a standard keyboard. Users should look away while typing to check predicted words; it could be a reason for word prediction failure because it could slow the user more than boosting given by word suggestion. For those cases when the typist must distract from the document, this effect is not related. The goal of research [1] was to conclude whether word prediction applications would improve typing speed when a user should look away from the document. The experiment shows that seven out of ten participants improved typing speed with word prediction. So the result indicates that word prediction could be helpful for users.

However, errors can often be observed when generating the next word. Therefore, this work is aimed at analysing the existing methods of the next-word prediction based on the entered text and testing them in Ukrainian language content. A possible obstacle will be the lack of Ukrainian language corpus and models supporting Ukrainian language (for example, GPT, BERT, and GPT2 for English). Also, a novelty in contrast to T9 will be the ability to predict not just one word but also a given number of words/sentences.

The work is aimed at developing a predictive system of Ukrainian language text input, which should improve the correctness of the next word. The research methods are machine learning algorithms: Markov chains and LSTM and their hybrid.

In the paper [2], the authors using a recurrent neural network trained language model. It was developed with the help of federated learning (FL). Federated learning is distributed, on-device learning framework currently called for next-word prediction. The significant advantage of method implementation is training on client devices because it helps to consolidate privacy. The federated averaging algorithm, which is used on client devices, shows better prediction recall than server-based training using stochastic gradient descent. One reason is training on a better quality dataset for this use case because a user customizes the dataset. Furthermore, one

more benefit is no need to export user data to servers. Long short-term memory (LSTM) with a Coupled Input and Forget Gate (CIFG) language model trained on the server and baseline n-gram model was compared to the federated learning model trained from scratch. It was shown that the created model outperforms the keyboard next-word prediction task. To add, this is one of the first federated language modelling applications in a commercial setting.

In the following paper [3], the authors had the same idea of using federated learning without direct access to the user data for commercial purposes. Apart from that, to improve search suggestion of virtual keyboard quality by global-scale setting training and evaluate results, the proposed method is a logistic regression model applied by federated learning. The first step is training the baseline model, which generates query candidates on the server side, the second step is triggering model trained by federated learning. The aim was to improve the query click through rate (CTR) by applying a baseline model for taking suggestions and the triggering model to remove low-quality recommendations. The comparison between wanted ΔCTR in training metrics and received ΔCTR in live experiments shows that real-life deployments significantly improve CTR. The result of this paper is also an early end-to-end example of FL in a commercial.

In paper [4], the authors research various sequences to sequence deep learning models in scripts for TV series generation: conversations and scenarios. LSTM, bidirectional recurrent neural network (RNN), and gated recurrent units (GRU) were analysed. Input sequence consists of $n$ characters. Therefore, targets consist of a similar amount of characters. One possible exception is that character could be shifted one to the right. Everything is stored in a pickle file what allows to train a model for script generation tasks with no need for human intervention. The LSTM shows the most reliable performance, and then, in the second place is GRU and last but not least bidirectional RNN. The loss analysis demonstrates that the least lost is in bidirectional RNN and then LSTM and the highest is in GRU. Execution time is the least in LSTM, GRU works a bit more, and bidirectional RNN has the highest execution time.

Sequence to sequence text generation is demonstrated in [5]. Bangla text generation is analysed using a deep learning approach. LSTM is used for analysing text sequences and next-word prediction. In this paper, Bangla Text was used. This model is in demand for this language because tools for Bangla are limited. No metrics are displayed in the article. Results are presented on two sentences example, which has a satisfactory accuracy rate.

Text mining and opinion mining are highly needed in practice. Sentiment analysis (opinion mining) is based on sentiment lexicon—a set of predefined keywords that correspond to a particular sentiment. This paper [6] proposed a new opinion mining method developed using hidden Markov models (TextHMMs) for text classification based on word sequence instead of a predefined sentiment lexicon. Text patterns are used for the representation of sentiment through ensemble TextHMMs. The semantic cluster represents hidden variables in TextHMMs based on the appearance of the words. Next, determine the sentiment score of sentences with the help of fitted TextHMMs. LSA (latent semantic analysis) is used in the method and provides boosting and clusters of words. The evaluation shows that the model outperforms some existing techniques and can classify implicit opinions. The method is applied to existing reviews from the online market.

In paper [7], the hidden Markov models were used for text prediction for the Polish language. Model is using an input text to learn the potential letters' sequence. For input, punctuation should be removed, and only spaces are considered as separate letters. From the cleaned text, transition matrix is created. The transition matrix is filled by occurrence. All sequences are sorted based on their frequency in the input text to increase the program's efficiency. The model evaluation showed that HMM could be successfully used for word generation but still is not enough for creating whole sentences. The displayed method can also be used in speech recognition.

Unsupervised word embeddings become highly popular, and it appears in many applications. It leads to the question of whether similar methods can improve semantic representations of word sequences. In paper [8] is presented efficient and, at the same time, a simple unsupervised objective for training distributed terms of sentences. A model called Sent2Vec can be interpreted as an extension from C-BOW to a bigger sentence context. A combination of a sentence with the help of an unsupervised objective function optimizes sentence words within a model. This method works better than all other unsupervised competitors except for SkipThought on most benchmark tasks. But SkipThought vectors demonstrate low performance on sentence similarity tasks while the proposed method is state of the art for these evaluations on average. It points out the robustness of the provided sentence embeddings for general purpose showing the relevance of well-grounded and straightforward representation models compared to the models using deep architectures.

Big data technics are used for natural text processing too. Particularly, linear discriminant analysis (LDA) is used in [9].

OpenAI pretrained a neural network GPT based on the Transformer architecture on a large amount of text. Based on this development, Google created its own neural network BERT. They significantly improved the result by making the neural network bidirectional, in contrast to GPT. OpenAI increased its GPT by a factor of 10 at once and trained it on an even larger volume of text—on 8 million Internet pages (a total of 40 GB of text). The resulting GPT-2 network is currently the largest neural network, with an unprecedented number of parameters of 1.5 billion (the BERT had 340 million in the largest model and the standard BERT 110 million). As a result, GPT-2 was able to generate entire pages of connected text. The main limitation of GPT-2 is English supporting [10, 11].

To summarize, federated learning for next-word prediction was used and showed fair results. RNN, in particular, LSTM, also demonstrated reliable results. The Markov model was used for sentiment analysis and gain satisfactory

accuracy rate. But it should be highlighted that the Markov chain has sequential nature to be used in next-word prediction.

This paper is aimed at comparing two sequential models: LSTM and Markov model in Ukrainian next-word prediction.

The contribution of the paper is given as follows:

(i) The dataset of Ukrainian poems with repeated patterns in texts is prepared. This allows to train models without Ukrainian language corpus

(ii) The structure of long short-term memory (LSTM) neural network with chosen hyperparameters is chosen

(iii) The Markov chain's parameters (frequency, probability, and transitions) are experimentally calculated

(iv) The hybrid model as combination of LSTM and Markov chain is developed. The prediction accuracy for all models is evaluated manually. The prediction accuracy of the proposed hybrid model is the best. However, time complexity is the highs

The paper is organized as follows. Materials and Methods describe the proposed approaches. Results and Discussion present the prepared dataset and experimental setup. Conclusions underline the main results and limitations of the work.

## 2. Materials and Methods

*2.1. Recurrent Neural Networks: LSTM.* Neural networks [12] are a collection of algorithms that is aimed at recognizing patterns. Their nature is very similar to the human brain. Machine perception, labelling, or clustering of source data helps to interpret sensory data. All real-world data (images, sound, text, or time series) must be converted into vectors for a neural network to recognize numerical patterns. Artificial neural networks (ANN) consist of many deeply interconnected processing elements (neurons) that collaborate to solve a problem.

ANN typically includes numerous processors that run parallel and are organized in tiers. The same as optic nerves get input in human visual processing first level gets the initial input. Each subsequent level accepts output from the level prefacing it, not from the raw input—similar to the neurons distant from the optic nerve that gets signals from those closest to it. The last level gives the system result.

Recurrent neural networks [13] are generalizations of a direct transmission neural network that has internal memory. The RNN is repetitive in nature because it performs the same function for each data entry, but at the same time, the current output depends on the previous calculation. After receiving the original data, it is copied and sent back to the periodic network. The decision is based on an analysis of the current input and the output from the previous input.

RNNs can use their internal state (memory) to process input sequences when direct communication neural networks cannot. The internal state helps them in duties such as speech recognition or unsegmented, connected handwriting recognition. Unlike RNN, other neural networks' inputs are independent of each other. All RNN inputs are interconnected.

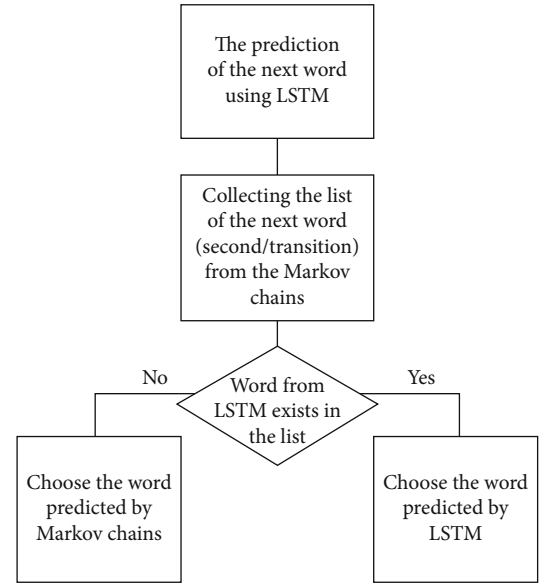



Figure 1: Hybrid model structure.

First take $X(0)$ from the input sequence and then output $h(0)$, which together with $X(1)$ is the next step input. Therefore, $h(0)$ and $X(1)$ are the input data for the next step. Thus, $h(1)$ is the input from $X(2)$ for the next step and so on. In this way, he continues to memorize the context while learning.

The current state formula is the following:

$$h_t = f(h_{t-1}, x_t). \tag{1}$$

Application of the activation function:

$$h_t = \tanh\left(W_{hh}h_{t-1} + W_{xh}x_t\right), \tag{2}$$

where $W$ is the weight, $h$ is the single latent vector, $W_{hh}$ is the weight in the previously latent state, $W_{xh}$ is the weight in the current input state, tanh is the activation function that implements nonlinearity that reduces activation to the range $[-1.1]$ .

Output:

$$y_t = W_{hy}h_t, \tag{3}$$

where $y_t$ is the output state and $W_{hy}$ is the output weight.

Advantages of the periodic neural network are as follows:

(i) In RNN's data sequence, each sample is supposed to depend on the previous ones

(ii) Extension of the effectiveness of neighbourhood of pixels is done by usage of convolutional layers

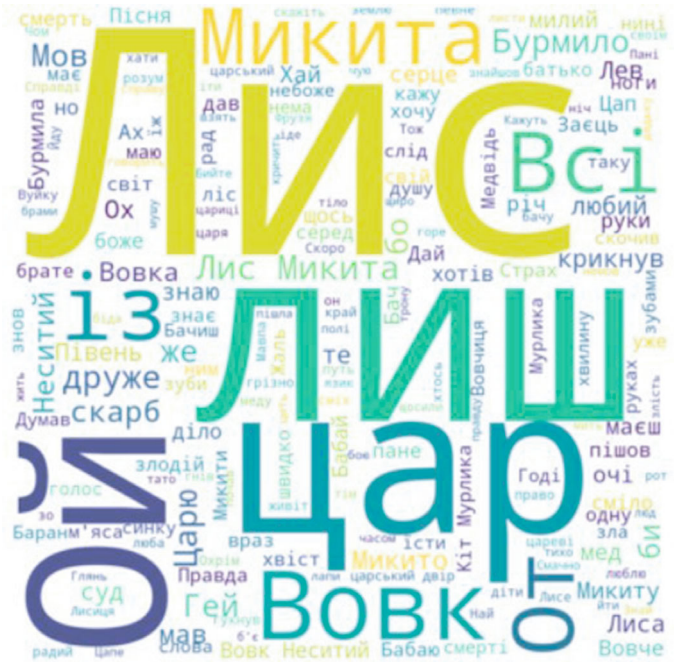



Figure 2: Wordcloud of original text.

Drawbacks of the periodic neural network are as follows:

(i) Gradient problems of disappearance and explosion

(ii) RNN training is a complicated task

(iii) In the case of tanh or relu activation function models cannot process very long sequences

Long-term short-term memory networks (LSTMs) [14] are a modified version of periodic neural networks that make it easier to remember past data in memory. Here, the problem of the vanishing RNN gradient is solved. LSTM [15] is well suited for classifying, processing, and forecasting time series based on time lags of unknown duration and trains the model with backpropagation. There are three gates in the LSTM network:

(1) Entrance gate: learn what value from the input should be used to modify the memory. The sigmoid function decides which values to pass through 0.1 and the tanh function provides a weighting to the values transmitted, determining the level of their importance in the range from -1 to 1:

$$\begin{aligned} i_t &= \sigma(W_i * [h_{t-1}, x_t] + b_i), \\ \tilde{C}_t &= \tanh\left(W_c * [h_{t-1}, x_t] + b_c\right). \end{aligned} \tag{4}$$

(2) Forget gates: learn what parts should be thrown out of the block. The sigmoid function determines this. Consider the previous state $(h_{t-1})$ and the input content $(x_t)$ and output a number from 0 (skip it) to 1 (save it) for each number in the state of cell $C_{t-1}$:

$$f_t = \sigma\left(W_f * [h_{t-1}, x_t] + b_f\right). \tag{5}$$

(3) Output gate: the input and memory of the unit are used to solve the output. The sigmoid function decides which values to pass through 0.1, and the tanh function provides a weighting to the transmitted values, determining their level of importance in the range from-1 to 1 and multiplying by the sigmoid output:

$$\begin{aligned} o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o, \\ h_t &= o_t * \tanh\left(C_t\right). \end{aligned} \tag{6}$$

*2.2. Markov Chains.* Markov chains [16] are one of the most valuable classes of stochastic processes:

(i) Supported by many sophisticated theoretical results but at the same time are flexible and simple

(ii) Helpful to have a clue about random dynamic models

(iii) Central to quantitative modelling by themselves

A stochastic matrix (or Markov matrix) [17] is an $n \times n$ square matrix $P$ such that each element $P$ is nonnegative and each row $P$ has the sum 1. Each row $P$ can be considered as a function of the probability mass for $n$ possible results. It

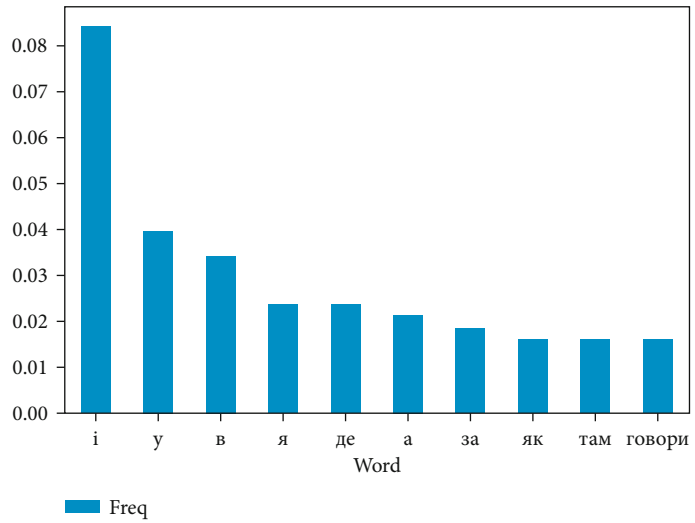



Figure 3: Frequency of words in original text.

Vocabulary size: 61
Total sequences: 110
Mode: "sequential_7"

| Layer (type) | Output shape | Param # |
|---|---|---|
| embedding_7 (Embedding) | (None, 1, 10) | 610 |
| lstm_7 (LSTM) | (None, 50) | 12200 |
| dense_7 (Dense) | (None, 61) | 3111 |

Total params: 15,921
Trainable params: 15,921
Non-trainable params: 0

Figure 4: Models' structure.

is easy to verify that if $P$ is a stochastic matrix, then the $k$th power $P_k$ is the same for all $k \in N$.

Markov chains are highly related to stochastic matrices. First, let $S$ be a finite set with $n$ elements $\{x_1, \cdots, x_n\}$. The set $S$ is called the state space, and $x_1, \cdots, x_n$ is the value of the state. A Markov chain $\{X_t\}$ on $S$ is a sequence of random variables on $S$ with a Markov property. So for any date $t$ and any state $y \in S$,

$$P\{X_t + 1 = y \mid X_t\} = P\{X_t + 1 = y \mid X_t, X_t - 1, \cdots\}. \quad (7)$$

In other words, knowledge of the current state is enough to know the probabilities for future states. In particular, the dynamics of the Markov chain are completely determined by a set of values:

$$P(x, y) \coloneqq P\{X_{t+1} = y \mid X_t = x\}(x, y \in S). \quad (8)$$

By construction,

(i) $P(x, y)$ is the probability of transition from $x$ to $y$ per unit time (one step)

Table 1: Model evaluation.

| Loss | Accuracy |
|---|---|
| 0.5750 | 0.7455 |

Table 2: The LSTM predicted results.

| Word | Prediction |
|---|---|
| Крила (1 word) | крила має |
| Крила (4 words) | крила має а крила має |
| Немає (4 words) | Немає то буде воля немає |

(ii) $P(x, \cdot)$ is the conditional distribution of $X_{t+1}$, given $X_t = x$

We can consider $P$ as a stochastic matrix, where

$$P_{ij} = P(x_i, x_j), \quad 1 \le i, j \le n. \quad (9)$$

Alternatively, if we take the stochastic matrix $P$, we can form a Markov chain $\{X_t\}$ as follows:

Table 3: The Markov chains' predicted results.

| Initial word | Second word | Transitions |
|---|---|---|
| {'а': 0.020997, 'землі': 0.0026247, 'немає': 0.0052493, 'в': 0.03412073, 'живе': 0.00262467, 'вони': 0.002624671, 'у': 0.039370, 'або': 0.00262467, 'людина': 0.002624671…} | {'а': {'й': 0.125, 'як': 0.125, 'крила': 0.25, 'з': 0.125, 'вона': 0.125, 'твоє': 0.125, 'на': 0.125}, 'землі': {'немає': 1.0}, 'немає': {'поля': 0.5, 'пари': 0.5}, 'в': {'цьому': 0.07692307692307693, 'сяйві': 0.07692307692307693, 'далеких': 0.076923, 'гріб': 0.076923, 'одному': 0.076923, 'степу': 0.076923, 'україні': 0.076923, 'семї': 0.076923, 'довгу': 0.076923, 'день': 0.3076923}, 'живе': {'на': 1.0}, 'вони': {'ті': 1.0}, …} | {('а', 'й'): {'правда': 1.0}, ('й', 'правда'): {'крилатим': 1.0}, ('правда', 'крилатим'): {'ґрунту': 1.0}, ('крилатим', 'ґрунту'): {'не': 1.0}, ('не', 'треба'): {'END': 1.0}, ('ґрунту', 'не'): {'треба': 1.0}, ('землі', 'немає'): {'то': 1.0}, ('немає', 'то'): {'буде': 1.0}, ('буде', 'небо'): {'END': 1.0}…} |

(i) Subtract $X_0$ from some specified distribution

(ii) For each $t = 0, 1, \cdots$, find $X_{t+1}$ with $P(X_t, \cdot)$

*2.3. Hybrid Model.* For achieving better results of the next-word prediction model, it was decided to develop a hybrid of LSTM and Markov chains.

The proposed algorithm for next-word prediction consists of the following (Figure 1).

## 3. Results

*3.1. Dataset Description.* As the work is aimed at developing personalized system in full cycle system, the model should be trained on user's input text. In such case, the base functionality of the system would be similar to T9 [18].

Each person has own linguistic style and uses the same repeated patterns in texts or messages.

Ukrainian poems contain a huge amount of different variations of phrase combinations. That is why such text corpus could be used as a training set for the developed prediction model. The specific group of Ukrainian poems [19] was used due to having no ability to gather the required type of dataset for training. Also, Ukrainian story was tested, but in story, there are not much repeated patterns which could help to train a model. Thus, the decision to use poems as a training set is substantiated by the fact that poems have more often used required repeated patterns for building prediction lines.

For the current investigation, a group of poems which consists of 1641 words was used. Based on this text corpus, it is possible to build up to 10 of phrase combinations [20–22].

*3.2. Data Preparation.* Although clearing the text affects the models' results, it is better to clean the text minimally because predicting the next word requires the original text.

The text before cleaning is demonstrated in Figures 2 and 3. The highest frequency has prepositions and connecting words. Regular cleaning process will remove them and predicted sentences could be incomplete.

So for data cleanup, it could be determined in the next steps:

(i) Lowering words

(ii) Punctuation removal

(iii) Tokenization (Python Natural Language Toolkit is used for this)

Stop-words removing, lemmatization, and stemming are not used in this phase, because the main idea is to save the structure of the sentence and to find consonant words.

*3.3. Model Testing*

*3.3.1. LSTM Model.* The figure below demonstrates the structure of the used LSTM (Figure 4).

The model consists of the embedding layer, LSTM, and neural layers. The embedding layer is highly important for NLP tasks as it helps to achieve better results focusing on keywords.

The training was based on Lina Kostenko's poem "Wings." Results of model evaluation on the validated dataset are displayed in Table 1.

Even though the text was small and the dimension of the dictionary was only 61 words, an accuracy of 75% was achieved, which is not bad for the start of the research. With increasing training text, greater accuracy will be achieved.

In the next step, the prediction on validated text is given.

Variants of one and several word prediction were tested (Table 2).

From the current demonstration, it is obvious that the prediction of the next word is correct.

*3.3.2. Markov Chain Model.* As the previous example showed, it is better to increase the sample; a collection of popular Ukrainian poems was used for training.

Below is a part of the trained model of Markov chains (Table 3).

The generated poem using Markov chain is given below (Table 4).

A graphic example of Markov chains on own text is shown in Figure 5.

*3.3.3. Hybrid Model.* The hybrid model was executed 100 times. The final model was generated Markov output only 44 times and combined output 56 times. The example of generated text is given in Table 5.

Table 4: The generated poem using Markov chain.

| Ukrainian poem | Translated poem |
|---|---|
| людина нібито не літає… | A man supposedly does not fly… |
| стане початком тоді мій кінець | Will be the beginning, then my end |
| і чом твій усміх – для мене весела весна | And why your smile is a happy spring for me |
| гей ви зорі ясні тихий місяцю мій | Hey you stars of light, my quiet moon |
| немає поля то буде небо | No field, it will be heaven |
| зпоза хмар | Outside the clouds |
| їй виспівує осанна | Hosanna sings to her |
| говори зі мною | Talk to me |
| серце мліло не хотіло | The heart did not want to beat |
| будуть приходити люди | People will come |

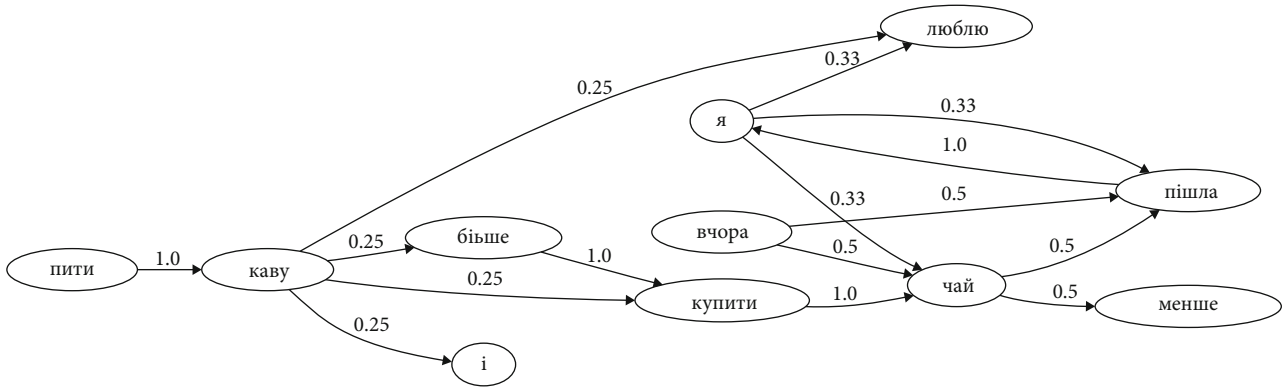



Figure 5: An example of the created model.

Table 5: Generated poem using the hybrid approach.

| Output generated by 1—Markov and 2—LSTM | Text phrase |
|---|---|
| 1 | чом ваші очі сяють тим чаром |
| 2 | і тремтить райдуги в крилі |
| 1 | в день вольній новій |
| 2 | радощі й тугу нестимуть мені |
| 1 | так ніхто не кохав через тисячі літ |
| 1 | завтра на цій землі |
| 2 | його вистачить всім – |
| 1 | я на лиш знаю і одне засвоїв |

*3.3.4. Model Comparison.* The content of generated texts was evaluated manually. Six students of Applied Linguistics Department at Lviv Polytechnic National University have evaluated the quality of generated text by three models. Unfortunately, GPT-2 adaptation to Ukrainian language is presented but it cannot be executed on a small dataset (https://kaif.revo.ua/). Each expert has analysed 10 texts generated by each model. The results of expert evaluation are given in Table 6.

Table 6: Generated poem using the hybrid approach.

| Expert no. | The best model | | |
|---|---|---|---|
| | LSTM | Markov chain | Hybrid |
| 1 | 2 | 4 | 4 |
| 2 | 1 | 4 | 5 |
| 3 | 2 | 4 | 4 |
| 4 | 1 | 4 | 5 |
| 5 | 3 | 4 | 3 |
| 6 | 1 | 4 | 5 |

As shown from Table 6, the accuracy of the hybrid model is adequate.

The three models' comparison by time is given in Table 7.

## 4. Discussion

The possible improvements of the proposed model include

(i) adding punctuation to the prediction

(ii) adding suggestion of word form

(iii) completing the system for a user with full functionality

Table 7: Comparison by time.

| Model | Training time (s) | Execution time (s) |
|---|---|---|
| LSTM | 70.53 | 0.38 |
| Markov chains | 0.01 | |
| Hybrid | | 0.31 |

The proposed algorithm is oriented on Ukrainian language. However, it can be retrained for other languages.

## 5. Conclusions

Word prediction is helpful for users because it can boost typing speed and help to omit errors. A personalized, predictive text input system is a relevant research topic for all languages, especially for Ukrainian, as currently products have limited tools which support the Ukrainian language.

Based on state-of-the-art research for analysis, three algorithms were chosen: LSTM, Markov chains, and hybrid. All algorithms are relevant for the next-word prediction task because both have a sequential nature (current output depends on previous). The Markov chains performed the task the fastest and qualitatively for development. Thus, it was chosen for the final outcome.

The novelty of the research is that, unlike T9, the developed system can generate not only the next word but also a few words, whole sentence, or several sentences.

Future improvement should be focused on punctuation and word form prepositions. Also, it will include a function for personalization of the system by developing a prediction model with users' linguistic style analysis.

Furthermore, provide a more accurate evaluation on marked data, currently working on their gathering.

## Data Availability

The row data used to support the findings of this study have been deposited in the repository (10.6084/m9.figshare.14844654.v1).

## Conflicts of Interest

The authors declare no conflict of interest.

## Acknowledgments

This research was funded by National Research Foundation of Ukraine and Comenius University, Slovakia.

## References

[1] D. Anson, P. Moist, M. Przywara, H. Wells, H. Saylor, and H. Maxime, "The effects of word completion and word prediction on typing rates using on-screen keyboards," *Assistive Technology*, vol. 18, no. 2, pp. 146–154, 2006.

[2] A. Hard et al., "Federated learning for mobile keyboard prediction," 2018, https://arxiv.org/abs/1811.03604.

[3] T. Yang, G. Andrew, H. Eichner et al., "Applied federated learning: improving Google keyboard query suggestions," 2018, https://arxiv.org/abs/1812.02903.

[4] S. Mangal, P. Joshi, and R. Modak, "Lstm vs. gru vs. bidirectional rnn for script generation," 2019, https://arxiv.org/abs/1908.04332.

[5] M. S. Islam, S. S. Sharmin Mousumi, S. Abujar, and S. A. Hossain, "Sequence-to-sequence Bangla sentence generation with LSTM recurrent neural networks," *Procedia Computer Science*, vol. 152, pp. 51–58, 2019.

[6] M. Kang, J. Ahn, and K. Lee, "Opinion mining using ensemble text hidden Markov models for text classification," *Expert Systems with Applications*, vol. 94, pp. 218–227, 2018.

[7] G. Szymanski and Z. Ciota, "Hidden Markov models suitable for text generation," in *WSEAS International Conference on Signal, Speech and Image Processing (WSEAS ICOSSIP 2002)*, pp. 3081–3084, Corfu, Greece, 2002.

[8] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," 2017, https://arxiv.org/abs/1703.02507.

[9] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[10] J. S. Lee and J. Hsiang, "Patent claim generation by fine-tuning OpenAI GPT-2," *World Patent Information*, vol. 62, article 101983, 2020.

[11] D. Ham, "End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 583–592, Seattle,WA, USA, 2020.

[12] M. Antony and P. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge university press, 2009.

[13] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," 2017, https://arxiv.org/abs/1702.01923.

[14] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *13th Annual Conference of the International Speech Communication Association*, Portland, OR, USA, 2012.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] P. Gagniuc, *A Markov Chains: From Theory to Implementation and Experimentation*, John Wiley & Sons, 2017.

[17] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer Science & Business Media, 2012.

[18] Y. V. Krak, A. V. Barmak, R. A. Bagriy, and I. O. Stelya, "Text entry system for alternative speech communications," *Journal of Automation and Information Sciences*, vol. 49, no. 1, pp. 65–75, 2017.

[19] "The most popular poems of Ukrainian poets known all over the world," https://maximum.fm/najpopulyarnishi-virshi-ukrayinskih-poetiv-yaki-znayut-u-vsomu-sviti_n169157.

[20] R. S. Bolia, W. T. Nelson, M. A. Ericson, and B. D. Simpson, "A speech corpus for multitalker communications research," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 1065-1066, 2000.

[21] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems*, vol. 117, pp. 47–58, 2021.

[22] P. Ashokkumar, G. Siva Shankar, G. Srivastava, P. K. R. Maddikunta, and T. R. Gadekallu, "A two-stage text feature selection algorithm for improving text classification," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 3, pp. 1–19, 2021.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3582053, 10 pages
https://doi.org/10.1155/2021/3582053

*Research Article*

# Do City Size and Population Density Influence Regional Innovation Output Evidence from China?

**Cai Shukai, Wang Haochen, and Zhou Xiaohong**

*Department of Economics and Management, Anhui Polytechnic University, Wuhu, Anhui 241000, China*

Correspondence should be addressed to Zhou Xiaohong; 017050@ahpu.edu.cn

Received 30 June 2021; Accepted 19 July 2021; Published 10 August 2021

Academic Editor: Thippa Reddy G



This paper proposed a substantial gap to a large-scale population density and city size on regional innovation output. To measure the impact of population density and city size on regional innovation output, this study employs the threshold effect model with panel data of 230 prefectures and cities from 2007 to 2016. Based on the econometric analysis, the results exhibit a positive and significant relationship between population density, city size, and innovation output. This correlation suggests that when one factor increases, the other increases in the parallel direction and vice versa. Moreover, when the city size expands the threshold value of 2.934 percent, the innovation promotes and increases the effects of urban-scale expansion. On the other hand, for medium- and low-density cities, the increase of urban population density has a significant and positive impact on urban innovation output. However, for high-density cities, the increase of population density has no significant impact on innovation output.

## 1. Introduction

With the advancement of the new urbanization strategy, the population mobility flows rapidly increase between urban and rural cities, while the population size of Chinese cities is constantly changing. In recent years, the population mobility flows from the first-tier cities such as Beijing, Guangdong, Shanghai, and Tianjin, which increase alongside with the second-tier cities, namely, Nanjing, Chengdu, Suzhou, Dalian, Tianjin, Zhongshan, and Dongguan. However, some uncompetitive urban population continued to flow out and even evolved into "shrinking cities." At the same time, the population density level of different cities also changed accordingly. Due to the unplanned allocation of infrastructure development in many Chinese cities, urbanization is expanding faster than the growing population in these cities, while the population density of Chinese cities shows a downward trend due to the lack of planned development. Nevertheless, the changes in urban population density in different scales are determined on basis of various administrative levels and region-to-region significant disparities [1]. In the urban population density in dominant administrative level, the study examines that the central and eastern regions are growing fast. The city is the core carrier of innovation, and scholars widely believe that innovation is mainly in the city. Moreover, larger cities are the backbone core centers of technological innovation; numerous research scholars extensively acknowledge that innovation is the first place phenomenon in a city. Pred (1996) found the per capita patents of 35 largest cities in the United States from 1860 to 1910 and examined that per capita of the four megacities was approximately four times larger than the national average [2]. According to Duncan et al. (1999), a study of 26 metropolises in the United States investigated that these metropolitan cities not only occupy a large number of territorial areas but also carry a large number of the urban population and generated 75% of the patents of the national metropolises [3]. Marella et al. (2017) and Wojcik-Popek (2019) explored the links between city size and patents. However, their findings cannot identify external effects [4, 5]. Oort and Bosma (2012) performed empirical analysis on the data of 14 countries and 110 regions in Europe; they examined the mechanism of the impact of urban scale on innovation, and their findings show that the impact of urban scale on innovation

is linked by human capital [6]. A study proposed by Carlino et al. (2007) with external factors remaining unchanged found that for every 2-fold increase in population density in metropolitan cities of the United States, the corresponding patent strength would increase by 20% [7]. A study performed by Yuan-quan and Jia-jia employed a cluster analysis to explore the impact of urbanization scale distribution on regional innovation efficiency; he found that there is an "inverted U" association between urbanization scale distribution and regional innovation efficiency. In addition, an optimal city size distribution can maximize the efficiency of regional innovation. Inefficient or excessive city size is not conducive to the improvement of regional innovation efficiency [8]. Gao (2015) performed empirical research of more than 287 prefecture-level cities; the larger the scale of the city, the stronger its innovation ability and performance [9]. Guo et al. (2015) described the city size in terms of population density, innovation output, and patent strength; they explored the relationship between population density and regional innovation output using spatial economic methods. The study found that when the population of the city is between 50 and 90 million, the innovation output level of the city is the highest, and there is a significant "inverted U" relationship between the city size and innovation [10]. In the two key factors such as agglomeration effect and crowding effect, many scholars explored that the relationship between population density and regional innovation output is not necessarily driving a linear relationship which depends on the comprehensive outcome of both agglomeration effect and crowding effect. Recent research work by Min and Changquan (2019) addresses that in certain regions when the population density increases excessively, it will lead to the phenomenon of excessive agglomeration and the generation of congestion effect in the region [11]. Additionally, Ye et al. (2016) propagate that there is a threshold effect between the population density and regional innovation efficiency by analyzing the relationship between the population density and regional innovation efficiency [12]. Some scholars discussed big data and smart cities such as analysis of dimensionality reduction techniques on big data and deep learning for future smart cities. Kumar et al. (2021) think that smart cities have become the mainstream of urbanization [13]. Ram et al. (2021) discussed security-by-design (SbD) concepts in the energy harvester technologies for sustainable and secured IoT with uninterrupted energy resource smart villages and smart cities [14]. Liu et al. (2017) discussed enormous data from IoT is stimulating our cities to become smarter than ever before [15].

From the above studies, a wide range of census by numerous scholars noticed that there is an extensive and complicated relationship existed between urban innovation output and urban scale, population density, and the nonlinear relationship between population density and regional innovation output. However, few studies have explored the threshold effect of urban size and population density on regional output. Therefore, based on the panel data of 225 cities at the prefecture level and above in 2007-2016, this study examined the impact of city size and population density on regional innovation output. In contrast, the existing literature on the innovation of this research lies in various ways. Firstly, this study considers an empirical analysis of the cross effect of innovation output on urban size and population density to examine whether there is an alternative or complementary relationship between urbanization size and population density. For this purpose, we add the cross terms of urbanization size and population density into the empirical analysis model to examine whether the effectiveness of one variable is conditional on another variable and if so whether there is a complementary or alternative relationship.

Secondly, this study considers an empirical analysis on whether there is a threshold effect on the regional innovation output of urban scale and population density. Considering that the improvement of urban scale and population density will promote the improvement of urban innovation capability, such improvement may depend on the development of urban scale and population density to a certain extent, which means that there may be a "threshold effect" in the improvement effects of urban scale and population density on innovation competency. Therefore, this study uses city size and population density as threshold variables and utilizes the bootstrap method to sample 500 times, to measure the threshold effect of regional innovation output of city size and population density.

## 2. The Internal Mechanism of the Impact of Urban Scale and Population Density on Innovation Output

Developed cities are the core hub of innovations such as human resources, scientific research, technological advancement, and information resources. The impact of urban-scale expansion on innovation output is primarily reflected in many aspects such as the expansion of urban scale helps to promote the agglomeration of regional innovation output, expertise, and boost innovation patterns. Many researchers believe that the expansion of city scale promotes regional innovation such as human resources, intellectual property protection, product professional test, and intermediate and technical information services. In addition, it has been widely acknowledged that the Chinese first-tier cities attracted millions of highly skilled workers that resultantly consolidate the city innovation output on a larger scale and also promote the technological innovation research to rapidly diminish the innovation cost and time. The latest study by Gerlach (2009) shows that when conducting innovation activities, most enterprises prefer to choose regions with higher concentration, because these enterprises have rich shared resources, compared with the regions with low concentration; these regions can enable enterprises to save innovation costs and reduce their risks [16].

Secondly, the expansion of the urban scale will help to enhance the matching degree and supply of the labor market. Consequently, the innovation subjects can match highly skilled talents and thus promote the development of regional innovation [17]. Another study by Parrotta et al. (2014) shows that the abundance of talent selection will significantly reduce the innovation robustness of enterprises [18]. The

refinement of the matching degree of the labor market will also help to promote the structural matching and accumulation velocity of knowledge creation and promote innovation output. Thirdly, the expansion of city scale means the expansion of market scale and the continuous expansion of the demand of innovation outcomes. All these factors encourage enterprises to increase the absorption capacity and integration of innovation resources that improve their innovation capabilities.

The impact of urban population density on innovation output is mainly through the following ways: (i) the spatial proximity effect brought by the increase of population density is conducive to innovation output. The spatial proximity effect can create highly skilled talents to better communicate and boost the frequency of communication that is conducive to the spillover and dissemination of knowledge, especially invisible knowledge and noncoding knowledge to promote the generation of innovation. (ii) The social network effect brought by the increase of population density is beneficial to innovation output with all kinds of innovative outcomes obtained in the city. Moreover, establishing a formal contact is becoming a perfect social network, in the long-term stable cooperative relationship. In addition, it helps to reduce the contract cost caused by the uncertainty and inconsistency of information, promote the sharing of information resources, and improve the regional innovation ability. A similar study by Glaeser (1999) shows that even with the development of modern technology, people can communicate in a more diversified and convenient way. However, in social networks, the impact of informal relations on knowledge and information dissemination will not be neglected, because it cannot replace face-to-face communication, which is inevitable in a formal contract [19].

## 3. Empirical Study on the Impact of Urban Scale and Population Density on Innovation Output

*3.1. Variable Selection and Data Analyzation.* In this paper, we add one explanatory variable that is innovation output (patent density) and describe it as the number of invention patents per 10000 people. Conventionally, the main indicators to measure the innovation output include the number of patent licenses, the number of patent applications, the transfer of patent use fees and royalties, and the technical market turnover. Moreover, it is difficult to measure accurately and effectively the regional innovation output, because the number of patent applications contains a large number of unlicensed patents. In addition, patents include invention licenses, utility model patents, and design patents. The amount of patent license can be considered to estimate the regional innovation output and information more effectively and efficiently. However, although the indicators such as royalty and technical market turnover with other data sources can more fully reflect the regional innovation output.

Explanatory variables include the size of a city that is represented by the resident population of the municipal district. Innovation is primarily a city phenomenon. The high concentration of various facilities in the city is more conducive to the development of innovation. The larger the city scale, the greater the agglomeration and attraction of innovative highly skilled talents. On the other hand, if the city scale is large, it is relatively easier to acquire relevant professional machinery and equipment. In addition to this, large-scale cities provide access to obtain relevant intellectual support and innovation-related services. At the same time, large city sizes are substantial to accommodate the division of labor and cooperation and various information exchanges related to innovation. Presumably, the city size helps to improve the process efficiency resulting from innovation output.

*3.1.1. Population Density-Regional Innovation Output Nexus.* This research uses the employment-population density of the municipal district to measure the employment-population of the municipal district, the built-up area of the municipal district. Because there are different categories and segments of the districts (mainly urban districts) and built-up areas in China, the population density is measured based on the urban district's population density and agglomeration due to the inclusion of a large number of rural areas. Substantially, most of the economic activities of a city are mainly concentrated in the urban start-up areas. Therefore, this paper uses the number of the urban population divided by the start-up area to better measure the regional population concentration [20, 21]. Knowledge exchange and accumulation are the basis of innovation. Moreover, the higher the population density, the more conducive the formation of a compact space distance and intensive social relationship network. Furthermore, it is convenient for people to exchange knowledge, disseminate implicit information and accumulate knowledge, and improve the frequency of contact between people. The higher frequency of contact increases the possibility of knowledge dissemination and innovation, which in turn promotes the enhancement of regional innovation efficiency and regional innovation output.

*3.1.2. Research and Development Investment.* In this section, we focus on the local financial science and technology expenditure of the municipal government to measure the government's financial support for technological innovation. On the one hand, the financial science and technology-related expenditure of the local government could directly provide financial support and related preferential policies for urban innovation activities, which increases the urban innovation input and thus improves the urban innovation output. On the other hand, the government's investment in innovation can serve as a "good gesture" and policy guidance. It is reflecting the local government's support capacity and image for innovation, which helps to strengthen the confidence of enterprises in investment and lead enterprises to increase more investment in research and development as well as create more innovation output. At the same time, the "leverage effect" brought by the government's innovation investment can leverage more private investment; thus, the urban innovation ability becomes stronger.

*3.1.3. The Proportion of the Third Industry.* In this section, we use the proportion of the third industry in GDP of municipal districts to measure the cluster development of the service industry that can provide targeted specialized services for innovation activities. It provides legal and advisory services in the process of the patent application that reduces the transaction costs of innovation activities. Furthermore, the development of financial services in the regional industry can effectively reduce the financial cost of enterprises, improve the financing efficiency, and provide more investment in innovation funds for enterprises. Therefore, the research by Yang and Bao (2019) also found that the specialization and diversification of producer services have a significant role in promoting urban innovation [22].

*3.1.4. Foreign Direct Investment.* In the amount of foreign direct investment in a municipal district compared with the domestic level, foreign direct investment can bring more advanced technology, highly skilled management, and more proficient innovation development. On the one hand, the proportion of high-quality foreign innovation capital has increased, which can directly improve the regional innovation efficiency with its efficient innovation ability. On the other hand, the foreign direct investment will bring technology spillover to the region and promote the innovation ability and innovation efficiency of the region through demonstration effect, competition effect, human capital flow effect, and correlation effect [23]. Foreign direct investment not only brings a large amount of capital investment but also brings advanced technology and highly skilled talent to boost domestic technological progress. It was in a relevant study where Miao (2009) used panel data to conduct spatial measurement research and found that FDI caused imitation effect and competition effect, which had a significant positive impact on innovation [24].

*3.2. Model Specification.* Based on the previous theoretical analysis, we constructed the following empirical model:

$$\text{Patent density}_{it} = \alpha_0 + \alpha_1 \text{size}_{it} + \alpha_2 \text{density}_{it} + \sum \alpha_k X_{kit} + \varepsilon_{it}, \tag{1}$$

where $i = 1, 2, \cdots, N$ represents the different cities, $t = 1, 2, \cdots, T$ represents the cities, $t$ represents the years, patent $\text{density}_{it}$ is the patent density, i.e., the number of invented patents authorized per 10000 people, and is the explained variable of the model, $\text{size}_{it}$ represents the city size, $\text{density}_{it}$ represents population density, and $X_k$ represents a series of control variables.

Likewise, we also investigate whether there is a substitution relationship between urban size and population density in the case of small city size. Nevertheless, the improvement of population density can make up for the low innovation efficiency brought by the small city size. In the case of insufficient population density, the expansion of the urban scale makes up for the problem of low innovation output caused by insufficient communication due to low population density. Therefore, we add the interactive terms of urban size and population density into Equation (1) to obtain the final expression of the theoretical model in this paper:

$$\begin{aligned}\text{Patent density}_{it} = {} & \alpha_0 + \alpha_1 \text{size}_{it} + \alpha_2 \text{density}_{it} + \alpha_3 \text{size}_{it} \\ & * \text{density}_{it} + \sum \alpha_k X_{kit} + \varepsilon_{it}.\end{aligned} \tag{2}$$

*3.3. Data Source and Descriptive Statistics.* In this paper, we used unbalanced data due to the lack of relevant observations for some prefecture-level cities. For the administrative region adjustment of relevant cities in the sample range of 2007-2016, it is necessary to adjust the relevant data to maintain the consistency and accuracy of the data. Therefore, this study conducts essential steps as follows. Firstly, the management area of some cities has changed significantly, which may be due to the adjustment of different zones. Therefore, we have eliminated the sample cities with land area change of more than 5% during the sample period. Secondly, for some indicators, data of individual cities are missing. To overcome this issue, this study uses the method of moving average to carry out the edge in processing by referring to the processing methods of relevant research. Thirdly, some sample cities with substantial data deficiency, such as Sansha City, Baying City, and Fangchenggang City, were excluded from the study. At the same time, taking into account the time of innovation (assuming 1 year), it generally takes 1 year to apply for and obtain patents. In this paper, the number of invention patents measured by regional innovation output since the regression equation is 2005-2014.

The data used in this paper are obtained from the *China Urban Statistical Yearbook*, *China Regional Economic Statistical Yearbook*, and *China Statistical Yearbook* from 2006-2016 issued by the National Bureau of Statistics as well as the relevant annual statistical yearbook of the provinces where the relevant cities are located. Due to the lack of some of the missing data supplemented by the statistical yearbook of the city where they are located, some of the data cannot be obtained and supplemented by the linear interpolation method. Taking into account the impact of inflation, the GDP deflator is added to take 2007 as the base period to conduct the deflator processing on R&D, FDI, and other data. The summary statistics are presented in Table 1.

*3.4. Analysis of Empirical Results.* This study includes the panel data of 230 each prefecture-level city and over 10 years. We assume that $i > t$; the individuals of the samples in each period are the same, with balanced panel data. Since the number of individuals in this panel is much larger than the time dimension, therefore, we run a regression model without a unit root test. The basic model of panel data regression is as follows:

$$\text{Patent density}_{it} = \alpha_0 + \alpha_1 \text{size}_{it} + \alpha_2 \text{density}_{it} + \sum \alpha_k X_{kit} + \varepsilon_{it}. \tag{3}$$

In the empirical section, this paper employs a variety of panel data regression models, such as mixed-effects model, individual random-effects model, and fixed-effects model. The specific model shall be determined through relevant tests. Firstly, the appropriate estimation method is

Table 1: Summary of descriptive statistics.

| Variables | Observations | Mean | Std. | Max | Min |
|---|---|---|---|---|---|
| Patent density (piece/10000 persons) | 2300 | 17.08 | 26.74 | 290.34 | 0.01 |
| City size (10000 persons) | 2300 | 125.01 | 104.64 | 2440.82 | 26.3 |
| Population density (person/km$^2$) | 2300 | 13254.5 | 4402.50 | 28021.88 | 2688.73 |
| R&D investment (ten thousand yuan) | 2300 | 28242.89 | 66778.04 | 954447 | 44 |
| Third industry (%) | 2300 | 43.07 | 11.02 | 78.66 | 8.58 |
| FDI (ten thousand dollars) | 2300 | 78802.76 | 134140.10 | 983567 | 0.38 |

Table 2: Regression results of fixed-effects model, random-effects model, and mixed-effects model.

| Variable | Model 1 (fixed-effects model) | Model 2 (random-effects model) | Model 3 (mixed-effects model) | Model 4 (add cross product) |
|---|---|---|---|---|
| Size | 0.193*** (7.71) | 0.183*** (4.42) | 0.105** (3.65) | 0.164*** (4.42) |
| Density | 0.285*** (16.78) | 0.261 (0.90) | 0.424*** (8.95) | 0.296*** (7.10) |
| R&D | 0.235*** (14.46) | 0.250*** (12.20) | 0.271*** (9.70) | 0.246*** (6.15) |
| Third industry | 0.131*** (4.57) | 0.127*** (3.81) | 0.092 (2.81) | 0.091** (3.01) |
| FDI | 0.042 (1.31) | 0.045 (1.21) | 0.062 (1.16) | 0.116 (1.11) |
| Size ∗ density | | | | 0.121*** (8.8) |
| Constant | 0.035*** (2.80) | 0.2444 (6.20) | 0.011 (0.48) | 0.069** (3.52) |
| $R$-squared | 0.374 | 0.394 | 0.525 | 0.331 |

$t$ value in brackets. ∗ indicates significance under 10% significance level. ∗∗ indicates significance under 5% significance level, ∗∗∗ indicates significance under 1% significance level.

determined effectively: *B-P* test (chibar2 (01) = 2126.92, Prob > chibar2 = 0.0000) and LR likelihood ratio (LR test of sigma). The results of the ($u = 0$ : chibar2 (01) = 868.50, Prob ≥ chibar2 = 0.000) test indicated that it significantly rejected the original assumption at the level of 1%, indicating that individual random-effects model should be selected between mixed-effects model and individual random effect. Therefore, this paper employs the fixed-effects model with the best estimation effect as the main illustration object. As a comparison, the regression results of the random-effects model and the mixed OLS model are also presented in Table 2. Similarly, we add the cross terms of urban size and population density in model 4 to investigate whether there is a mutual or one-way relationship between urban size and population density.

The results indicate that urban scale has a significant and positive effect on innovation output at the level of 1%. It shows that the larger the city scale, the more innovative output factors such as people, enterprises, and industries accumulation. On the other hand, the urban scale provides a more suitable environment for innovation, which is conducive to the development of the latest science and advanced technology. At the same time, the larger the scale of the city, the better for people to learn knowledge, develop their skills, promote the formation of human capital, and ultimately promote the innovation ability of the city. The increase of the city scale is beneficial to the resembling and curtailing of the knowledge of different innovative groups, improving the efficiency and quality of knowledge cognition among different groups which helps in reducing the cost of science and technology. Likewise, innovation activities need to consolidate equipment and information to find people for division of work cooperation and information exchanges. The application and transaction of patents need the support of external service institutions. In large-scale cities with a large number of employees and a developed service industry, these demands are easier to meet and obtain high-quality services and support.

On the other hand, the results from the proposed model illustrate that population density has a significant and positive effect on regional innovation output at the level of 1%. The research and development (R&D) personnel in densely populated cities have relatively more person-in-person communication opportunities, which is beneficial to the development of innovative ideas. The higher the population density, the shorter the interaction distance and time, while more opportunities and lower costs for the communication between people in different industries and disciplines. In addition, it is easier for individuals with different knowledge to exchange ideas by searching for partners to improve the efficiency of knowledge exchange and promote the creation of knowledge. The increase in urban population density also contributes to the rapid expansion and exchange of innovative ideas and outputs. Geographical distance plays a distinct role in the process of knowledge exchange and innovation cooperation. The distance between people will affect the intensity of knowledge exchange, innovative cooperation behavior, and propagation of implicit knowledge. However, the theoretical part also supports that the spillover effect and proximity effect affect the probability and quality of innovation cooperation and information exchange. Moreover, the study finds that if the population density is higher,

Table 3: Results of the robustness test.

| Variable | Model 1 | Model 2 | | | Model 3 |
|---|---|---|---|---|---|
| | | Eastern | Middle part | Western | |
| Size | 0.121*** (3.21) | 0.281*** (2.29) | 0.213*** (2.87) | 0.121*** (2.67) | 0.298*** (3.12) |
| Density | 0.241*** (6.71) | 0.412*** (8.22) | 0.367*** (9.13) | 0.712*** (10.21) | 0.319*** (8.32) |
| R&D | 0.235*** (14.46) | 0.312*** (21.23) | 0.209 (1.54) | 0.281 (1.02) | 0.235*** (14.46) |
| Third industry | 0.131*** (4.57) | 0.131*** (4.57) | 0.131*** (4.57) | 0.131*** (4.57) | 0.131*** (4.57) |
| FDI | 0.043 (1.09) | 0.381 (0.32) | 0.124 (1.01) | 0.023** (1.96) | 0.121 (0.82) |
| Constant | 0.035*** (2.80) | 0.244*** (6.20) | 0.011 (0.48) | 0.069** (3.52) | 0.050*** (4.35) |
| $R$-squared | 0.374 | 0.394 | 0.525 | 0.331 | 0.479 |

$t$ value in brackets. * indicates significance under 10% significance level. ** indicates significance under 5% significance level. *** indicates significance under 1% significance level.

it will ultimately be favorable to promoting social proximity, enriching the knowledge exchange channels between innovation outputs, and improving the multidimensional exchange and interaction between innovation contents [25]. Likewise, the urban population density promotes the spillover of innovation knowledge among subjects for the purpose to improve the level of regional innovation. The higher the population density, the more easily the agglomeration economic externality will be obtained. The implicit knowledge related to innovation is often before some core employees, and the regions with high population density can effectively promote the flow and exchange of people and accelerate the spread of tacit knowledge.

Finally, the result of the cross product of urban size and population density is significant and positive, indicating that the impact of urban size and population density on innovation output is interdependent and contribution is interactive. There is a significant positive association between the two variables, one of which is essential to the other. It shows that the impact of urban population density cannot be ignored when considering the promotion of urban scale on innovation output. If the city size is too small and has a high population density, it is difficult to play a role in promoting the increase of innovation output. On the other hand, if the size of the city is large, but there is no certain population density support, it is difficult to form a sufficient agglomeration effect to promote the increase of regional innovation output.

*3.5. Robustness Test.* To further investigate the robustness of the results and ensure the rationality of the model accuracy and reliability, this study utilizes the following regression results to crosscheck the coefficients of all the selected variables. The robust estimates were performed for the following regression:

Firstly, we replace the main explanatory variable city size of this study. In the previous study, the resident population of municipal districts was used to measure the city size. Next, this study uses the total GDP of urban districts as the proxy variable to measure the city size and estimate the relevant measurement test results (model 1). Secondly, previous studies at the national level have shown that urban size and population density have a significant positive impact on technological innovation.

Thirdly, we exclude four municipalities from the model; the remaining 226 cities above prefecture level were brought into the panel data for estimation (model 3). From the estimation results, the robustness of the model is illustrated in Table 3. The results illustrate that the positive impact of urban size and population density on technological innovation still exists and is consistent with previous analysis. The results of most explanatory variables are robust and accurate in the fixed-effects panel model.

The main difference is that in model 2, the estimation results of subsamples in the eastern, central, and western regions are different to some extent. The impact of regional research and development (R&D) investment intensity on regional innovation efficiency is different between eastern, central, and western regions. From the sample regression results, the impact of regional R&D input intensity on regional innovation output is not significant in the subsample estimation results of the central region and western region, while the result of the eastern region is significantly positive. In this case, the input efficiency of innovation resources in the central and western regions is tropical except for the eastern region, while the intensity of R&D input has not been effectively transformed into actual innovation output. Since there are many historical debts of innovation investment in the central and western regions, hence, it is difficult to induce benefits from R&D investment. It is possible that the intensity of R&D investment has not yet reached a certain threshold and cannot play an effective role.

In this section of the paper, the result reveals that foreign direct investment (FDI) has a significant impact on the innovation output of cities in the western region at the level of 5%, but has no significant impact on cities in the eastern region and the central region, respectively. This may be the eastern and central regions are rich in urban innovation resources and have a relatively high level of innovation capacity and efficiency, resulting in an insignificant spillover effect of foreign direct investment on regional innovation output. Furthermore, for the cities in Western China, their innovation resources are relatively scarce, and their innovation ability and efficiency are relatively low. On the one hand, the inflow of FDI increases the regional innovation input; on the other hand, the technology spillover effect and competition effect enhance the regional innovation output.

Table 4: Threshold effect self-sampling inspection.

| Variable | | $F$ | $P$ | BS | 1% | 5% | 10% |
|---|---|---|---|---|---|---|---|
| Size | Single threshold test | $21.038^{***}$ | 0.033 | 300 | 20.479 | 18.500 | 18.292 |
| | Double threshold test | 7.023 | 0.197 | 300 | 11.276 | 10.574 | 8.842 |
| Density | Single threshold test | $34.622_{***}$ | 0.007 | 300 | 14.684 | 11.965 | 8.442 |
| | Double threshold test | $22.498^{**}$ | 0.023 | 300 | 15.269 | 8.201 | 6.066 |
| | Triple threshold test | 4.348 | 0.211 | 300 | 10.034 | 6.412 | 5.817 |

## 4. Research on the Threshold Effect of Urban Scale and Population Density on Innovation Output

Although the enhancement of urban scale and population density will promote the improvement of urban innovation ability, such improvement may depend on the development of urban scale and population density to a certain extent, i.e., the development effect of urban scale and population density on innovation ability may have a "threshold effect." Therefore, this paper takes city size and population density as threshold variables and uses the bootstrap method to sample 500 times to test the threshold effect of explanatory variables.

*4.1. Threshold Effect Test.* First, we need to determine whether the threshold effect exists, and select the appropriate threshold number as well as the threshold value. Based on $F$ statistics and 300 "self-sampling," the $P$ values reported in Table 4 indicate that when the city size is used as a threshold variable, the $F$ statistics value (21.038) of the single threshold test is greater than the 1% horizontal threshold (20.479). However, the $F$ statistics value of the double threshold test (7.023) is less than the 10% significance level threshold value (8.842). Therefore, it can be considered that there is only one threshold effect in size.

In addition, the study determines the estimate of the threshold and the 95% confidence interval. See Table 5 for further details. The single threshold estimate for city size is 293.430, and the 95% confidence interval is [39.000, 326.400]. The first threshold of population density is 17197.297, and the second threshold is 13136.986. The likelihood ratio function is illustrated in Figure 1 that shows the estimates and confidence intervals of the threshold.

*4.2. Threshold Model.* Equation (1) can only reflect the comprehensive static effect of urban scale on regional innovation output; the threshold panel model is further explored to analyze the dynamic effect of urban scale and population density variations on regional innovation output. According to the relevant literature, the two-stage least squares method proposed by Hansen (1999) is used to estimate the data model of the threshold panel. The setting equation of the single threshold regression model is as follows:

$$\text{Patent density}_{it} = \alpha_0 + \alpha_1 \text{size}_{it} + \alpha_2 \text{density}_{it} + \sum \alpha_k X_{kit} + \varepsilon_{it}, \quad (4)$$

Table 5: Threshold estimation results and confidence interval.

| Variable | The threshold value | Estimated value | 95% confidence interval |
|---|---|---|---|
| Size | Ito1 | 293.430 | [39.000, 326.400] |
| Density | Ito1 | 17197.297 | [13892.320, 18115.310] |
| | Ito2 | 13136.986 | [86432.016, 14341.831] |

where $1(0)$ is an indicator function and the value is 1 when the corresponding condition is raised; otherwise, it is 0.

In contrast, in the actual observation value with the threshold value, the sample observation value can be divided into two intervals: the coefficient of the corresponding range by looking at the marginal coefficients of different intervals. We examine the impact direction and extent of changes in urban size and population density on regional innovation output if there are multiple thresholds; we expand the above equation based on a single threshold.

*4.3. Panel Threshold Analysis.* Based on the threshold regression analysis, the urban innovation output is considered as the explanatory variable, while the urban scale and population density were added as the threshold variable. On the other hand, research and development (R&D) investment, the proportion of the tertiary industry, and international direct investment add as the core explanatory variables for threshold regression. The specific inspection results are illustrated in Tables 6 and 7.

Table 6 shows that there are significant innovation-promoting effects on urban-scale expansion. When the urban scale exceeds the threshold value of 293.430, the innovation promotion effect of urban-scale expansion increases significantly, and its effect coefficient increases from 0.128 to 0.233, indicating that although the expansion of urban scale is favorable to improving urban innovation output. However, for cities of different scales, when the city size is greater than the specific threshold (293.430 people/km$^2$), it will further release the innovation promotion effect brought by the expansion of the city scale. In theory, if the city size exceeds a certain threshold, the city will have complete information, communication, and other infrastructures. Moreover, the agglomeration effect of innovative resources, the accumulation effect of human capital, the formation effect of information exchange network, and the improvement effect of transaction efficiency can be effectively exerted. The larger the scale of the city, the easier it is to become the

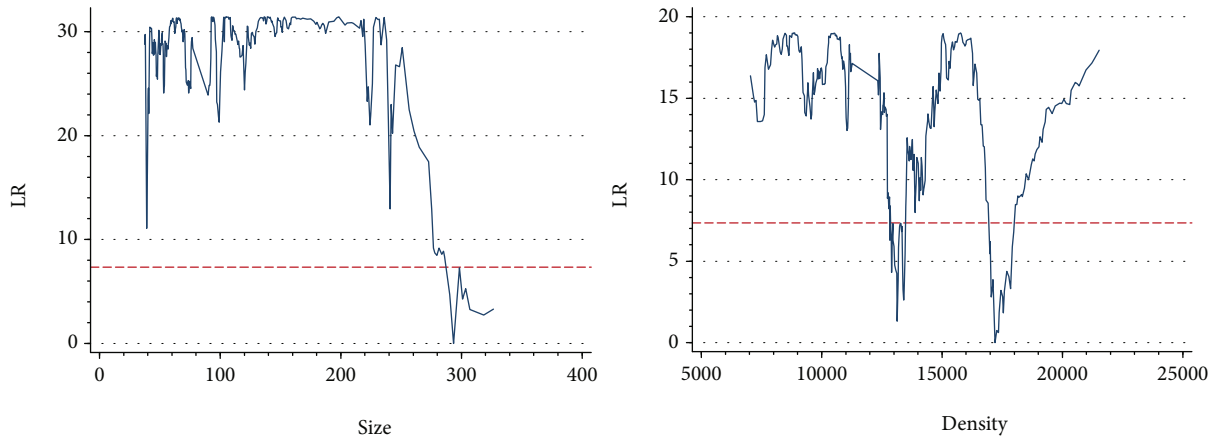



Figure 1: Threshold and confidence interval.

Table 6: Results of model estimation based on the urban size as a threshold.

| Variable | Coefficient | Standard error | $t$ |
|---|---|---|---|
| Size ≤ 293.430 | 0.128*** | 0.040 | 3.21 |
| Size > 293.430 | 0.233*** | 0.059 | 3.98 |
| Density | 0.347*** | 0.123 | 2.81 |
| R&D | 1.287*** | 0.247 | 5.21 |
| Third industry | 3.218** | 1.634 | 1.97 |
| FDI | 3.092 | 3.031 | 1.02 |
| Constant | 0.218*** | 0.068 | 3.21 |
| $R$-squared | | 0.372 | |

$t$ value in brackets. * indicates significance under 10% significance level. ** indicates significance under 5% significance level. *** indicates significance under 1% significance level.

Table 7: Model estimation results with population density as the threshold.

| Variable | Result | Standard error | $t$ |
|---|---|---|---|
| Density ≤ 13136.986 | 0.216*** | 0.049 | 4.37 |
| 13136.986 < density ≤ 17197.297 | 0.431*** | 0.110 | 3.92 |
| Density > 17197.297 | 0.033 | 0.046 | 0.72 |
| Size | 0.347*** | 0.025 | 13.64 |
| R&D | 0.821*** | 0.250 | 3.29 |
| Third industry | 1.921*** | 0.919 | 2.091 |
| FDI | 0.214 | 0.177 | 1.21 |
| Constant | 0.306*** | 0.062 | 4.91 |
| $R$-squared | | 0.461 | |

$t$ value in brackets. * indicates significance under 10% significance level. ** indicates significance under 5% significance level. *** indicates significance under 1% significance level.

concentration of innovation and invention as well as city size will affect the regional human capital accumulation [26]. On the other side, when the city scale is larger, then more various innovative resources will be gathered; also, more dynamic the innovation development as well as the innovation segments will boost likely to produce high-quality innovation results. The empirical study of Miao (2009) also shows that regional innovation activities have an externality [27]. The externality of innovation activities has a certain space scope that innovation activities have agglomeration. The expansion of urban scale is conducive to the agglomeration of innovation elements and the promotion of regional innovation output. The diversity brought by the expansion of urban scales, such as diversity of economic activities, diversity of innovation activities, and diversity of innovation groups, also helps to trigger innovation ideas and improve urban innovation output.

We refer to the population-density cities the first threshold (13136.986 people/km$^2$) as low-density cities, the population-density cities between the first threshold and the second threshold as medium-density cities (13136.986 people/km$^2$-17197.297 people/km$^2$), and the population-density cities above the second threshold (17197.297 people/km$^2$) as high-density cities. From the regression results in Table 7, we explore that for low-density cities and medium-density cities, the increase of urban population density has a significant positive impact on urban innovation ability. The increase in urban population density can significantly improve the technology of new knowledge in the society, further improve the level of specialization and division of labor [28], and is conducive to improving urban innovation.

However, in high-density cities, increasing population density does not have a significant impact on innovation output. This also indirectly proves that the density of employment agglomeration in some cities in China is too high, and there has been a significant agglomeration diseconomy [29]. This may be due to the lack of economic activities of congestion in high-density cities, such as the increase in commuting costs caused by the congestion effect, resulting in a large number of losses of effective labor commuting and damaging the city's innovation ability. From the perspective of coefficient, the innovation promotion effect of medium-sized cities is significantly greater than that of low-density cities with coefficients of 0.431 and 0.216, respectively. As

previously analyzed, the increase in urban population density can promote the sharing of public facilities, facilitate the matching between workers and positions, and promote the speed and scope of the dissemination and exchange of "hidden" knowledge related to innovation. Additionally, the urban city population facilitates the mutual exchange and imitation among organizations and talents. However, the larger population density will highlight crowding and other negative effects as well as make the marginal agglomeration effect caused by the increase of urban population density offset by the marginal crowding effect, which is not beneficial to the improvement of urban innovation output capacity. Therefore, for high-density cities, we should optimize the layout of industry and population within the city to develop multicenter cities. By decentralizing employment in peripheral subcenters, the scale of the urban population is increased, while the commuting distance is greatly reduced, the commuting cost is saved, and the negative impact of the increase in population density is reduced.

## 5. Conclusions and Policy Recommendations

The empirical research shows that the urban size and population density have a significant positive impact on regional innovation output. The impact of urban size and population density on innovation output shows a significant and positive relationship. The empirical results of the threshold panel model show that when the city size exceeds the threshold value of 293.430 million people, the regional innovation output development effect of urban-scale expansion increases significantly. There is a double threshold effect on population density that has a significant positive impact on urban innovation development with the increase of urban population density in low- and medium-sized cities. However, the result also reveals that for high-density cities, the increase in population density will not have a significant impact on innovation output.

Based on the empirical analysis findings, to further enhance the improvement of regional innovation level, this paper proposes the following countermeasures and suggestions: First, our findings suggest that local and central governments should continue to expand the scale of cities, especially small- and medium-sized cities. We follow the laws of urbanization development, urban-scale expansion, and industrial development that actively promote population urbanization. Specifically, starting from the solution of housing security, children's education, medical facilities, pension, and other issues, we should formulate relevant support and support policies to promote the "new citizens" to stay in the city. Particularly, in the context of rising housing prices, it is necessary to further improve the "government-led, market participation" housing system, increase the supply of affordable housing, and improve the quality of affordable housing. In addition, small- and medium-sized cities should cultivate pillar industries according to local conditions, gather population through industrial development, and improve the ability to create employment and acquire labor force. Secondly, while increasing the size of the city, the policymakers actively increase the population density of the city promote the expedition of population urbanization, consciously improve the population density of the city, avoid the large-scale inefficient outward spread of the city, improve the population and economic activities, and boost the utilization efficiency of land resources. Thirdly, smart growth theory and dense city concept should be merged into urban planning and development concept. To elude the crowding effect brought by megacities, the expansion of start-up area or restriction of population inflow currently adopted at the cost of damaging the regional innovation ability. It is necessary to integrate smart growth theory and dense city concept into urban planning and development concept into all aspects of urban development and management, to reduce the crowding effect while increasing urban population density.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no competing interest.

## Acknowledgments

This work was financially supported by the research projects of the social science and humanity on Young Fund of the Ministry of Education of China (18YJCZH163) and the Philosophy Planning Project of Anhui Province (AHSKZ2019D028).

## References

[1] Z. Rui, J. Limin, G. Xu, X. Zhibang, and D. Ting, "The relationship between urban space growth and population density," *Journal of Geography*, vol. 75, no. 4, pp. 695–707, 2020.

[2] A. Pred, "Interfusions: consumption, identity and the practices and power relations of everyday life," *Environment and Planning A: Economy and Space*, vol. 28, no. 1, pp. 11–24, 1996.

[3] B. W. Duncan, S. Boyle, D. R. Breininger, and P. A. Schmalzer, "Coupling past management practice and historic landscape change on John F. Kennedy Space Center, Florida," *Landscape Ecology*, vol. 14, no. 3, pp. 291–309, 1999.

[4] G. Marella, V. Antoniucci, and C. D'Alpaos, "How regulation affects innovation: the smart grid case at urban scale," in *22nd Annual European Real Estate Society Conference*, Istanbul, Turkey, 2015.

[5] A. Wojcik-Popek, "Green innovation in urban scale: activation of small cities through horticultural exhibitions in Berlin/-Brandenburg Metropolitan Region," *IOP Conference Series: Materials Science and Engineering*, vol. 471, p. 112100, 2019.

[6] F. G. van Oort and N. S. Bosma, "Agglomeration economies, inventors and entrepreneurs as engines of European regional productivity," *Annals of Regional Science*, vol. 51, no. 1, pp. 213–244, 2012.

[7] G. A. Carlino, S. Chatterjee, and R. M. Hunt, "Urban density and the rate of invention," *Journal of Urban Economics*, vol. 61, no. 3, p. 419, 2007.

[8] L. U. Yuan-quan and Q. Jia-jia, "Study on the influence of urban scale distribution of China on regional innovation efficiency," *Economic Survey*, vol. 35, no. 6, pp. 1–7, 2018.

[9] X. Gao, "Urban scale, human capital and urban innovation capacity in China," *Social Sciences*, vol. 3, pp. 49–58, 2015.

[10] J. Guo, H. Ning, and T. Shen, "Employment density and innovation-based on the spatial measurement of prefecture level cities in China," *Economic and Management Research*, vol. 36, no. 11, pp. 40–46, 2015.

[11] D. Min and L. Changquan, "Agglomeration effect, population mobility and urban growth," *Population and Economy*, vol. 21, no. 6, pp. 44–56, 2019.

[12] X. Ye, T. Changqi, and D. Hui, "Empirical study on the coupling of regional industrial innovation and industrial upgrading – taking the Pearl River Delta region as an example," *Scientific Research Management*, vol. 36, no. 4, pp. 111–119, 2016.

[13] P. Kumar, R. Kumar, G. Srivastava et al., "PPSF: a privacy-preserving and secure framework using blockchain-based machine-learning for IoT-driven smart cities," *IEEE Transactions on Network Science and Engineering*, p. 1, 2021.

[14] S. K. Ram, B. B. Das, K. Mahapatra, S. P. Mohanty, and U. Choppali, "Energy perspectives in IoT driven smart villages and smart cities," *IEEE Consumer Electronics Magazine*, vol. 10, no. 3, pp. 19–28, 2021.

[15] Y. Liu, X. Weng, J. Wan, X. Yue, H. Song, and A. V. Vasilakos, "Exploring data validity in transportation systems for smart cities," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 26–33, 2017.

[16] M. Gerlach, "Controlling in einem deutsch-chinesischen Joint Venture - Ein Erfahrungsbericht," *Controlling & Management*, vol. 53, no. 2, pp. 94–98, 2009.

[17] S. Yu, "Heterogeneous labor, matching effect and industrial agglomeration," *Economic and Management Review*, vol. 4, pp. 44–51, 2016.

[18] P. Parrotta, D. Pozzoli, and M. Pytlikova, "Labor diversity and firm productivity," *European Economic Review*, vol. 66, no. C, pp. 144–179, 2014.

[19] E. L. Glaeser, "Learning in cities," *Journal of Urban Economics*, vol. 46, no. 2, pp. 254–277, 1999.

[20] Q. Guo and H. Canfei, "Density, distance, segmentation and urban labor productivity: an empirical study based on China's 2004-2009 urban panel data," *China Soft Science*, no. 11, pp. 82–91, 2014.

[21] Z. Haoran and Y. Zhong, "Infrastructure, spatial spillover and regional TFP: an empirical study based on the Dupin model of 266 cities in China," *Economist*, vol. 2, no. 2, pp. 61–67, 2018.

[22] R. Yang and J. Bao, "Can agglomeration of productive services effectively promote urban innovation," *Discussion on Modern Economy*, no. 4, pp. 80–87, 2019.

[23] L. Zheng, S. Yang, and H. Bin, "Does FDI inhibit or improve the efficiency of regional innovation in China?– analysis based on the inter provincial spatial panel model," *Economic Management*, vol. 39, no. 4, pp. 8–21, 2017.

[24] T. Song, "Empirical Analysis on Regional Industry Brand Effect,Externality and Industrial Agglomeration—Based on Spatial Model of China's Provinces," *Economic Management Journal*, vol. 37, no. 8, pp. 35–44, 2015.

[25] L. Guoqing, Z. Gang, and G. Nana, "Dynamic evolution analysis of innovation network based on geographical proximity and social proximity – taking China's equipment manufacturing industry as an example," *China Soft Science*, vol. 5, pp. 97–106, 2014.

[26] C. Kaiming, Z. Yafei, and C. Long, "Effect decomposition and mechanism analysis of China's urbanization on energy consumption," *Geographic Science*, vol. 36, no. 11, pp. 1661–1669, 2016.

[27] F. Miao, *Spatial Measurement and Threshold Regression Analysis of Technology Spillover*, Huazhong University of Science and Technology, 2009.

[28] K. Shanzi and Y. Delong, "Causal relationship and determinants of industrial agglomeration and urban labor productivity: an analysis of the spatial econometric simultaneous equation of Chinese cities," *Research on Quantitative Economy, Technology and Economy*, vol. 25, no. 12, pp. 3–14, 2008.

[29] S. Hongjian and W. Houkai, "Density effect, optimal urban population density and intensive urbanization," *China Industrial Economy*, vol. 10, pp. 5–17, 2013.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5129396, 10 pages
https://doi.org/10.1155/2021/5129396

*Research Article*

# Source Routing for Distributed Big Data-Based Cognitive Internet of Things (CIoT)

**Seema Begum**,[1] **Yao Nianmin**,[2] **Syed Bilal Hussain Shah**,[2] **Asrin Abdollahi**,[3] **Inam Ullah Khan**,[4] **and Liqaa Nawaf**[5]

[1]*School of Computer Science and Technology, Dalian University of Technology, China*
[2]*School of Software, Dalian University of Technology, China*
[3]*Department of Electrical Engineering, University of Kurdistan, Sanandaj, Iran*
[4]*Department of Electronic Engineering, SEAS, Isra University, Islamabad, Pakistan*
[5]*Computer Science School of Technologies, Cardiff Metropolitan University, UK*

Correspondence should be addressed to Asrin Abdollahi; a.abdollahi@eng.uok.ac.ir

Received 18 June 2021; Accepted 15 July 2021; Published 1 August 2021

Academic Editor: Rajesh Kaluri



Dynamic opportunistic channel access with software-defined radio at a network layer in distributed cognitive IoT introduces a concurrent channel selection along with end-to-end route selection for application data transmission. State-of-the-art cognitive IoT big data-based routing protocols are not explored in terms of how the spectrum management is being coordinated with the network layer for concurrent channel route selection during end-to-end channel route discovery for data transmission of IoT and big data applications. In this paper, a reactive big data-based "cognitive dynamic source routing protocol" is proposed for cognitive-based IoT networks to concurrently select the channel route at the network layer from source to destination. Experimental results show that the proposed protocol cognitive DSR with concurrent channel route selection criteria is outperformed. This will happen when it is compared with the existing distributed cognitive DSR with independent channel route application data transmission.

## 1. Introduction

In order to utilize the natural available spectrum efficiently, the current static spectrum allocation needs to be switched to dynamic spectrum access [1–4]. To achieve that, the Federal Communications Commission (FCC) has proposed a novel way of accessing the static spread spectrum through "software-defined radio networks." With this, unused spectrum holes or TV white spaces (TVWS) in existing static spectrum can be opportunistically utilized by the secondary users through "cognitive radio networks." With the distributed cognitive radio, an unused primary spectrum band can be dynamically allocated to the secondary users for temporal basis [1, 5–8]. Hence, dynamic spectrum access through "cognitive radio networks" is a prominent solution to sustain for enhanced wireless technologies and increased number of radio users [3, 6, 9, 10]. In addition to this, next-generation Internet connectivity is extending to thing-to-thing connectivity through Internet of Things (IoT). With this, end-to-end application data will be transmitted from thing to thing without any human intervention. This brings new challenges in the end-to-end IoT network connectivity to transmit IoT data. Internet Engineering Task Force (IETF) proposed an open standard protocol stack for IoT with the introduction of different light-weight protocols to existing TCP/IP protocol stack.

IEEE 802.15.4 standard is used to provide the link connectivity among different IoT leaf nodes (sensor nodes). State-of-the-art IoT networks are interconnected with the nonconstrained heterogeneous networks through the wired backbone networks. For traditional wireless ad hoc networks,

interoperability with the wired backbone network to heterogeneous networks is mandatory to accommodate aggregated network traffic flows.

Devices linked to the Internet are growing gradually; the era of Internet of Things (IoT) and big data is coming. However, managing big data produced by the IoT networks will present substantial challenges for the conclusion makers. The IoT network is one of the big data sources in IoT. In such networks, a wide range of areas are monitored by thousands of smart sensors where assembled data are sent to the sink node. Unfortunately, IoT impose many challenges compared to other types of networks [11–13]. Data management is a mostly demanding task for IoT due to the huge amounts of data gathered in such networks.

Not all the attributes in the datasets produced are necessary for training the machine learning algorithms. Some attributes are perhaps not relevant, and some might not employ the outcome of the prediction. Removing or avoiding these irrelevant or less necessary attributes reduces the burden on machine learning algorithms [14]. In this article [42], we provide a comprehensive survey on blockchain for big data, focusing on up-to-date approaches, opportunities, and future directions.

However, for IoT application data, it is feasible to provide the backbone connectivity through the wireless broadband network to transmit the aggregate IoT application data to the nonconstrained networks. Since the existing static unlicensed wireless networks are deployed with multiple technologies, the available nonoverlapping channels are saturated and hard to accommodate with the new wireless technologies. Thus, it is effective to make use of dynamic spectrum to transmit the IoT application data from the IoT gateway to nonconstrained heterogeneous networks. To achieve that, this paper proposes a "reactive DSR source routing protocol for cognitive radio ad hoc networks" to transmit IoT data within the cognitive radio ad hoc networks from the IoT gateway to nonconstrained networks. Figure 1 explains the overview of the IoT gateway interconnected with the backbone distributed blockchain-based cognitive radio ad hoc networks to transmit IoT application data. [15, 16] explain how the blockchain technology gets integrated with cognitive IoT networks. Since cognitive ad hoc IoT network works on distributed coordination function, it is worthy to implement the blockchain module along with the distributed-coordination function to efficiently utilize the radio resources through reduced collisions and minimized channel saturations.

In this paper, our main contributions are as follows: distributed source routing protocol for big data-based [17] cognitive IoT is proposed to enhance the end-to-end throughput with minimized end-to-end delay. In addition, the gateway distributed CR routers will also act as IoT gateway nodes to aggregate and encapsulate the IoT data into the distributed cognitive radio ad hoc network.

The rest of the paper is organized as follows. Section 2 explains the pros and cons for the state-of-the-art cognitive routing protocols to transmit the aggregate IoT data. Section 3 briefly explains the proposed "reactive-based dynamic source routing protocol for cognitive-based IoT networks. Section 4 explains the experimental results whereas Section 5 ends with the conclusion and future work.

## 2. Related Work

End-to-end network performance of the routing protocol in cognitive radio-based IoT networks is mainly based on achievable network throughput, end-to-end packet delays, and node energy consumption. In order to achieve that, common control channel (CCC) [7, 18] plays a significant role to provide efficient end-to-end route discovery and route maintenance during the application packet transmission. It is noteworthy that channel route discovery is concurrently selected from the IoT gateway (source node of cognitive radio) to the destination whereas the default RPL route will be used as an end-to-end route from IoT leaf node to the IoT gateway.

Furthermore, a directional antenna is being proposed to provide increased number of simultaneous noninterfering transmissions within the cognitive radio network [19–21]. This will further enhance the achievable end-to-end throughput in multihop communication with efficient spatial reuse and reduced node power consumption. In other words, directional cognitive control and IoT application transmission will help to attain the increased end-to-end throughput by reducing the interference through directional antennas [21–25]. State-of-the-art routing protocols in "cognitive radio-based IoT networks" use omnidirectional transmission for application data (DATA/ACK) and cognitive control message exchange. But there will be great packet loss and frequent channel route failures due to cochannel interference with both primary users and secondary users [26, 27]. To overcome that, this paper designs a dynamic source routing protocol with directional antennas to transmit control and data transmission using directional antennas from the LBR gateway to the cognitive destination. In general, the IoT application data at the gateway may be destined to either IoT destination or destination of nonconstrained networks (cloud networks). The two different scenarios of IoT application data transmission from the IoT gateway to IoT destination and the IoT gateway to nonconstrained networks are briefly explained in Figures 2 and 3.

## 3. Proposed Work

State-of-the-art routing protocols in IoT networks are well explored in proactive-based routing protocols (RPL) to transmit the data from the leaf node (IoT end device) to the IoT boarder router. From LLN boarder router (IoT gateway router) to the non-IoT networks, a high-speed wired backbone network is being used to transmit the application data to the destination that are at nonconstrained networks. Using proactive-based source routing from the IoT gateway to the destination of the nonconstrained network is not a feasible solution in terms of control message exchange and the achievable application network throughput. In addition, it is beneficial to use wireless or opportunistic cognitive radio-based networks to retransmit the IoT application data from

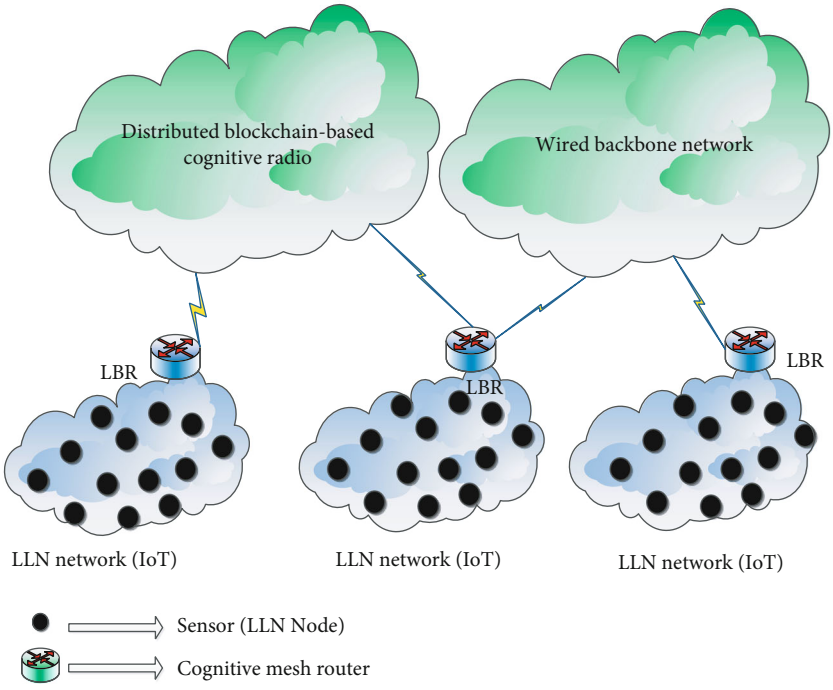



Figure 1: Overview of the cognitive IoT network.

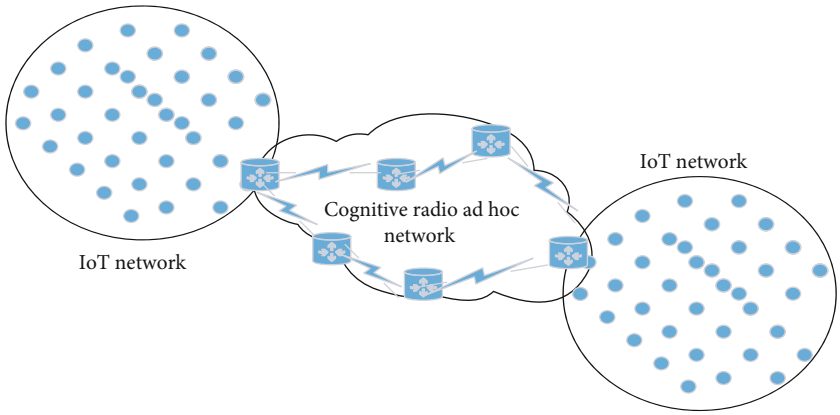



Figure 2: Communication between constrained IoT networks through CRAHNs.

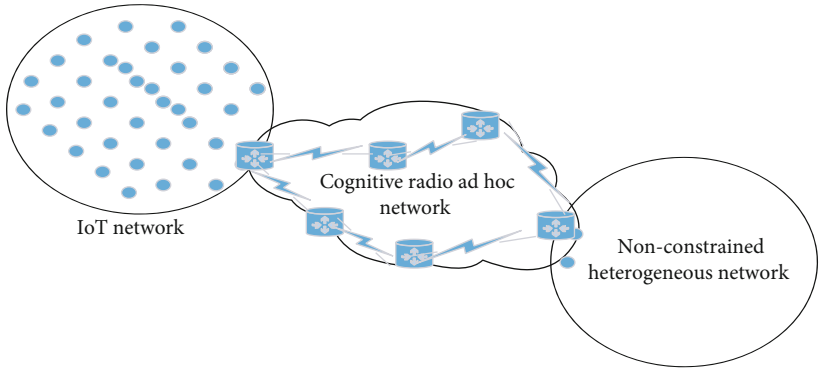



Figure 3: Communication between IoT and nonconstrained networks.

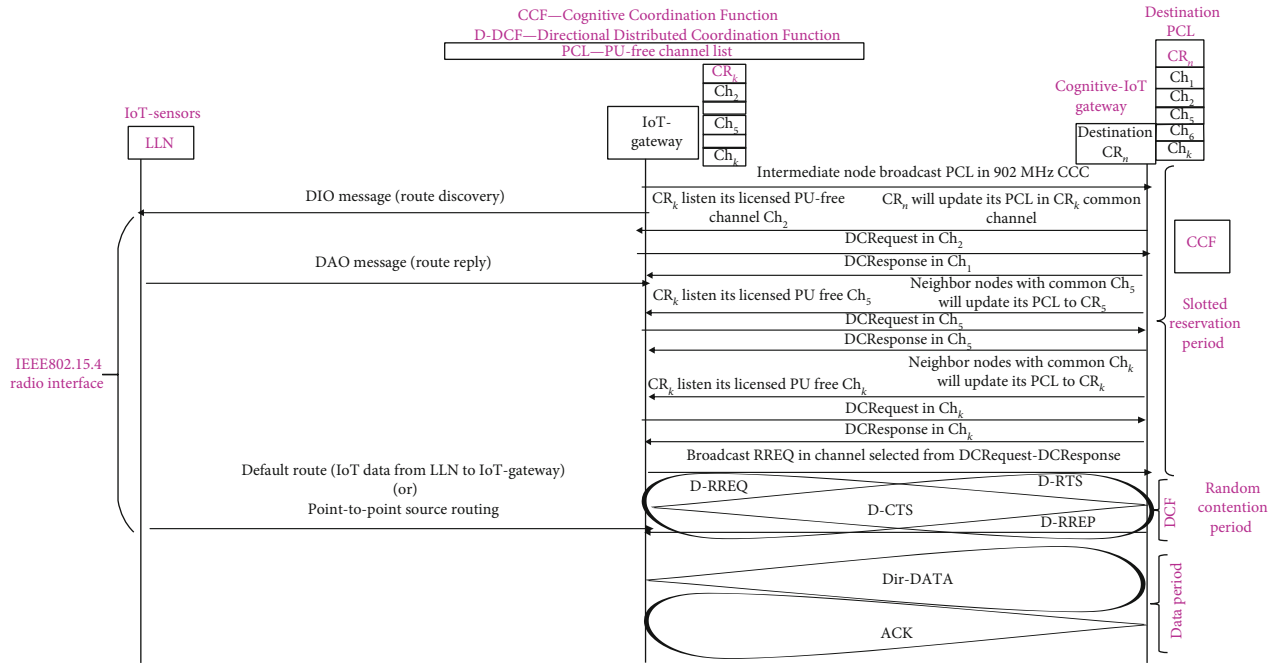



Figure 4: Overview of source routing in cognitive radio-based IoT networks.

the IoT gateway to the nonconstrained destination. To achieve that, this paper proposes a "reactive source routing protocol for IoT networks" to transmit the IoT application data from the IoT gateway to the destination of the nonconstrained network. IEEE 802.15.4 MAC protocol is being used to transmit the IoT application data from the IoT leaf node to the IoT gateway whereas the opportunistic TVWS (TV whitespace) is being used to transmit the IoT application from the IoT gateway to the destination. At the IoT gateway, the IP packet from the IoT network is being encapsulated with IP-in-IP encapsulation to provide the compatibility with the nonconstrained opportunistic-based cognitive radio ad hoc networks. 6lo and ROLL working groups in IETF worked on RFCs (Request For Comments) that enable the IP-based packet transmission from the IoT leaf node to the destination (within IoT network or outside of the IoT network). In general, with proactive-based routing, IoT leaf nodes will be periodically (trickle timer) sending the control messages to its one-hop neighbor nodes to maintain the route connectivity from the IoT leaf node to the IoT gateway. Whenever there is an application data at the IoT leaf node, then it will transmit to the IoT gateway through the proactive-based RPL routing protocol or any other proactive/reactive routing protocols. From the IoT gateway, the application data is being encapsulated and rerouted in dynamic spectrum access-based cognitive radio networks. Figure 2 depicts the channel route discovery overview within the cognitive radio ad hoc network for IoT application data transmission. Whenever there is an IoT application data at the CR source node (IoT gateway), then the CR node tries to find the shortest end-to-end channel route path towards the destination. The existing channel route path is being used if the intermediate node knows the channel path towards the destination CR node. In traditional reactive-based source routing protocols, nodes will record the IP address in its IP header while broadcasting the route discovery messages. Once the route control message reaches to the destination mobile node, it unicast the route reply message based on the IP address within the IP header. When it comes to cognitive radio ad hoc networks, the available PU-free channel needs to be concurrently selected along with the IP address for each and every intermediate CR node between cognitive source and destination. In this work, we considered that there can be a maximum of 256 PU channels that are available for opportunistic IoT application data transmission. In order to concurrently select the PU-free channel along with the channel route path, we have introduced channel ID information along with the 128-bit IP address in "source routing header." The step-by-step procedure to establish a channel route connectivity to transmit end-to-end IoT application data is explained in Figure 4. Firstly, DIO messages will be broadcasted within the LLN to provide the link and network connectivity with the LLN network. From the LLN gateway, the CR node will perform the IP-in-IP encapsulation and reinitiate the RREQ within the unlicensed opportunistic PU spectrum band. Once the RREQ message is reached to the destination CR node, then it will deencapsulate the encapsulated packet and send the LLN packet to the destination IoT node. Subsequently, destination nodes will uncast the RREP message back to the LLN gateway. From the LLN gateway (CR target node), the RREP packet will be encapsulated and sent back to the source CR node (source LLN gateway) through the cognitive radio network with opportunistic licensed spread spectrum.

Once the RREP message is reached back to the originating node, then it will start transmitting the IoT application data within the discovered channel route path. With this,

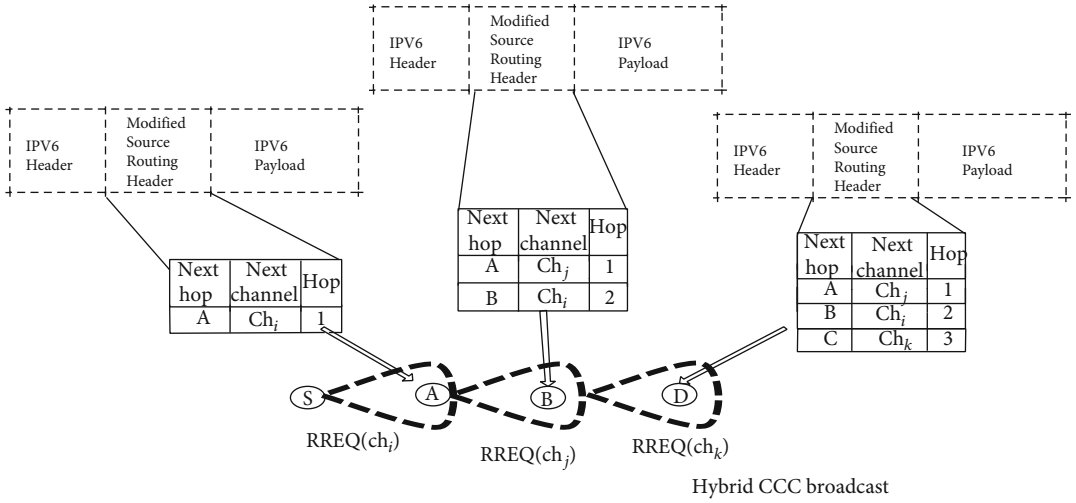



Figure 5: Channel RREQ from source CR node to destination-CR node through hybrid CCC-based source routing protocol.

the IoT application data will be transmitted using the cognitive radio network instead of the wired broadband network for end-to-end communication. Since cognitive radio makes use of the licensed PU-free spectrum band for secondary data transmission, it is crucial to select the reliable channel route during channel route discovery. As shown in Figure 4, there will a PU-free channel list (PCL) available at each and every CR node within the cognitive radio ad hoc network. Since CRAHN operation is based on distributed networking, each and every CR node should be capable of handling the resource allocation and network management. We assume that each and every CR node will have the PU channel list available through the spectrum sensing and spectrum management. In this paper, the authors assume that the PU-free channel is available to every next-hop neighbor CR nodes along with the per-hop link. This can be known through the distributed spectrum management or centralized spectrum management policies at the MAC layer of the cognitive radio ad hoc networks. The detailed operation of how exactly the end-to-end channel route discovery happens at the source (CR) node is explained in Figure 5. Once the channel route discovery is being initiated by the source CR node, then the channel RREQ will be broadcasted in the hybrid common control channel to the next-hop CR nodes. During the channel RREQ broadcast, the source CR node will initially update its 128-bit IP address along with the available PU-free channel (PCL) list to the next-hop neighbor nodes. Since the PCL list and 128-bit IP address occupy more space within the control packets, it is strongly recommended to compress the control information before broadcasting within the hybrid control channel. Once the channel RREQ is being broadcasted to the next-hop CR nodes, then next-hop CR nodes will check the PCL list of the source node with its PCL list. Whenever there is a common PU-free channel available between the source CR node and the next-hop neighbor CR nodes, then only the next-hop CR nodes will rebroadcast the received channel RREQ control messages. Since this paper proposes to work with the source routing-based cognitive radio networks to transmit IoT application data, every CR node will update its IP address within the source routing header of the channel RREQ message (see Figure 5).

Once the channel RREQ message is being reached to the destination CR node, then the destination CR node will initiate the unicast channel RREP message back to the source CR node. In general, when a channel RREQ message is being reached to the intermediate CR node and if the intermediate CR node is having the path to the destination CR node, then the intermediate CR node will send the channel RREP back to the source CR node. Subsequently, the intermediate CR node will initiate the gratuitous channel RREP message to the destination CR node so that the destination CR node will update its routing table with the IP address of the source CR node. Once the channel RREP message is transmitted back to the source CR node, then the source CR node will start transmitting the application data transmission. Figure 6 explains the application data transmission from the source CR node to the destination CR node through the source routing protocol. It is noteworthy that the source routing header will have the PU-free channel along with the 128-bit IP address of the next-hop CR node. In other words, the PU-free channel between every hop from destination to the source CR node will be updated at the time of transmitting the unicast channel RREP message. Later, this channel route from the source header of the encapsulated IPv6 packet will be used to forward the IoT application data from the source CR node to the destination CR node. In general, at the time of application packet transmission, there can be three types of packet failures, namely, spectrum handover packet failures, node mobility handover packet drops, and bandwidth degradation due to high network traffic flows. In this work, the performance of the CR network is being simulated with and without packet failure. In addition, packet drops at the edge of the PU receiver are being tested to check the performance degradation of the source CR routing protocol.

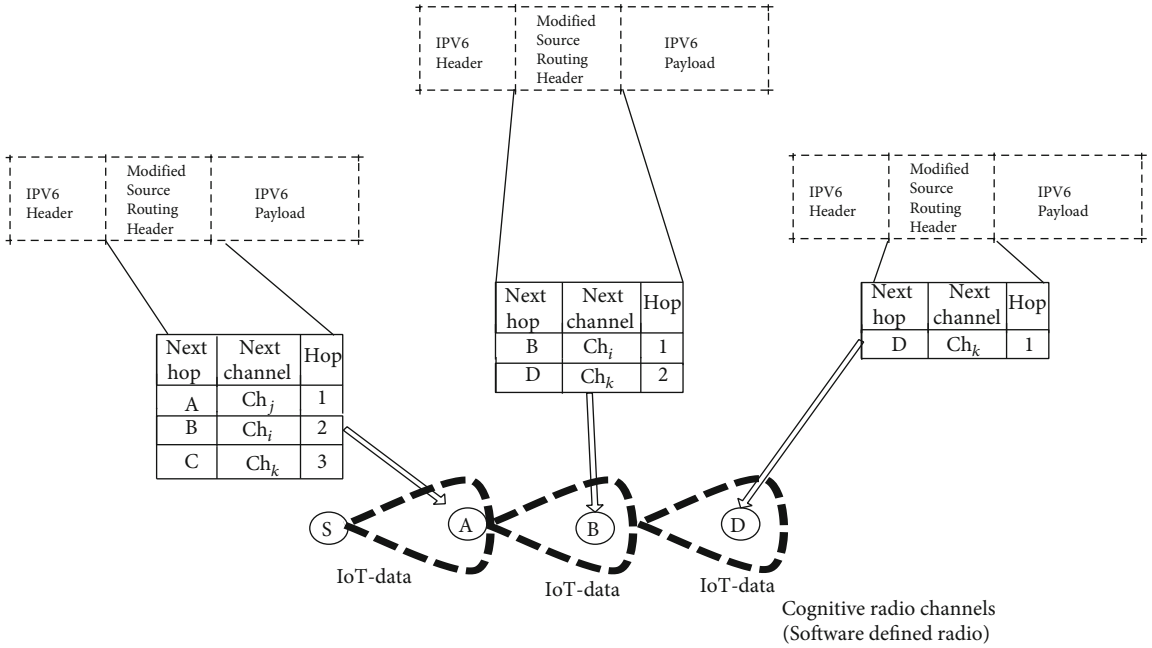



Figure 6: IoT application data transmission through source routing-based cognitive routing protocol.

*3.1. IP-in-IP Encapsulation at the Source CR Node.* Once the IoT application data is being transmitted from the LLN node to the LBR, then the constrained packet will be aggregated and encapsulated to transmit within the cognitive radio ad hoc networks. Once the message is reached to the destination CR node, then the received IP packet will be decapsulated and transmitted back to the constrained destination LLN node. In this paper, the authors assumed that the gateway routers support both the LLN and CR networks. In other words, the interfaces of both the CR network and LLN network will be deployed at the gateway nodes. Thus, the IEEE 802.15.4 standard protocol is being used within the LLN network whereas licensed PU-free spectrum bands are used at range of spectrum bands to transmit the data at the cognitive radio network.

*3.2. PU Receiver Protection at CRAHNS.* To implement dynamic and opportunistic spectrum access in software-defined cognitive radio, FCC introduces a fundamental requirement of "peaceful" coexistence between primary and secondary users. Hence, it is extremely important to protect the primary user communication during cognitive radio communication. In general, it is very hard to predict the PU receiver communication at the edge geographical location (see Figure 7). During cognitive radio communication, the highest spectrum utilization should be achieved by detecting all spectrum opportunities and accessing the spectrum so that collisions with the other secondary users get minimized. In addition, synchronization between the secondary transmitter and the primary receiver is also required to avoid the interference for both cognitive and primary networks. CR nodes that are close to the PU transmitter need to find out which PU spectrum bands are being used by the PU

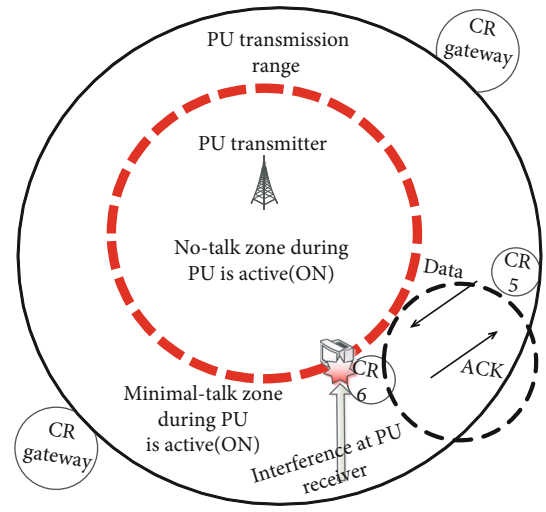



Figure 7: PU receiver protection for IoT data at CRAHNs.

transmitter. In addition, it is equally important to know whether there are continuous data transmissions occurring or noncontinuous data transmission is being done within the primary transmitter spectrum bands. It is extremely hard to concurrently transmit the primary and secondary data within the continuous transmission. But, in the case of non-continuous PU transmission, secondary nodes can make use of the PU spectrum band opportunistically without interrupting the primary user transmission. Let us consider that there are CR nodes that are located in 3 layers within the no-talk zone in Figure 7.

Level 1 area coverage is $\pi R1^2$.

Level 2 area coverage is $3\Pi r2^2$.

Level 3 area coverage is $5\Pi r3^2$.

Let us consider the area that is being covered by CR nodes within the minimal-talk zone:

Level 4 (minimal talk) area coverage is $7\Pi r4^2$.

In general, the area covered by the $N$ layers within the no-talk zone is as follows.

Level $N$ area coverage is $(2N-1)\Pi rN^2$.

The probability of the CR node that is being transmitted concurrently with the PU spectrum band in the no-talk zone is

$$\begin{aligned} P_T(r) &= \text{probability of CR transmission in level}_1 \\ &\quad + \text{probability of CR transmission in level}_2 + \cdots \\ &\quad + \text{probability of CR transmission in level}_n \\ &= \frac{(2r-1)\pi R_1^2}{n^2\pi D^2} + \frac{((2r+1)-1)\pi R_2^2}{n^2\pi D^2} + \cdots \frac{(2r-1)\pi R_n^2}{n^2\pi D^2}, \end{aligned}$$

$$P_T(r) = \frac{1}{n^2}\sum i = r^n\,(2i-1). \tag{1}$$

Equation (1) explains the general formula to calculate the level range with respect to the PU transmitter.

## 4. Experimental Results

The cognitive radio network simulator (NS-2.35) is used to simulate and check the performance of the proposed source routing-based cognitive source routing protocol for IoT application data [28–34]. In this work, we assume that the IoT network is being simulated in the Cooja simulator to reach the IoT data from the LLN node to the LNN gateway node (LBR) which also acts as a cognitive source node. Once the packet is reached to the CR source node, then the packet gets encapsulated with IP-in-IP encapsulation and is transferred through the cognitive radio network simulator. The simulation parameters are incorporated in Table 1. A practicable end-to-end cumulative network throughput within the CRAHNs is a subject matter to the channel route detection and restoration delays through local/global channel route recovery mechanisms. In addition, in the current CR communication channel, primary node transmitters are active randomly from 15th to 45th in available 8 MHz TV channels.

Due to this, the PU spectrum handoff probability is higher as compared to the nonexistence of active PU transmitters. Hence, the performance of source routing is being tested with different PU transmitter nodes to calculate the aggregate network throughput of IoT data within the CR ad hoc network. The performance of the cognitive source routing protocol is compared with the existing hybrid cognitive AODV routing protocols, licensed control channel-based AODV routing protocol, unlicensed AODV-based routing protocol, and traditional multichannel-based IEEE 802.11 DCF-based routing protocol. Figure 8(a) represents the comparison of delivery aggregate network throughput between existing protocols and the hybrid source routing protocol. It can be clearly seen that as the data rate increases, the network throughput for hybrid source routing performs better than hybrid cognitive AODV, licensed source routing, etc., since the routing table overhead is reduced by storing information in the network as well as packet.

Table 1: Simulation parameters.

| Parameters | Descriptions |
|---|---|
| Topology type | 1000 ∗ 1000 flat grid |
| Number of cognitive radio nodes | 10-100 nodes |
| Number of primary user channels | 8 MHz channels |
| Number of primary user transmitters (PUT) | 1-10 nodes |
| Unlicensed channels | ISM-902 MHz |
| Primary user active probability | 10, 15, and 20 msec |
| Type of mobility model | Random waypoint model |
| Input of CR transmit power | 10 mW |
| Receiver's threshold value | -95 dbm |
| Carrier sense (CS) threshold | -115 dbm |
| Cognitive radio transmitter (Tx) range | 200 m (licensed channel) |
| Primary user transmitter (Tx) range | 500 m (licensed channel) |
| Network data rate (DR) | 2 Mbps |
| Interface queue length | 50 |
| Simulation time (s) | 100 sec |

Also, in dynamic environment where PU behavior is unknown, the route recovery process is faster in the case of source routing. Similarly, Figure 8(b) compares the average throughput with data rate 1024 bytes/sec. Figures 8(c) and 8(d) demonstrate the average end-to-end Delay for the hybrid source routing (HSR) approach.

The simulation is done for 200 ms with data rate of 512 and 1024 bytes, respectively. The HSR approach discovers multiple routes to a given destination which takes time. However, in a cognitive radio channel, switching occurs frequently due to dynamic PU behavior which results in obsolete links. Thus, having an alternative route will help in node to resume its transmission with minimum switching delay. Also, route cache property helps in faster route discovery. Hence, the delay is comparatively less when compared with cognitive AODV which requires a large number of control packets which link failure occurs. Figures 8(e) and 8(f) demonstrate the performance of HSR comparing average throughput to no. of PU transmitters. With the increase in the number of transmitters, the channel occupancy increases. Hence, the coverage area of the channel is decreased resulting in the use of multiple channels to make end-to-end connectivity. The number of links increases the probability of link failure increases. Even with multiple backup links in the HSR approach, there is drop in the throughput. However, HSR performs better than cognitive AODV-based routing due to its feature of channel route caching, less channel route control overhead, and rapid discovery time.

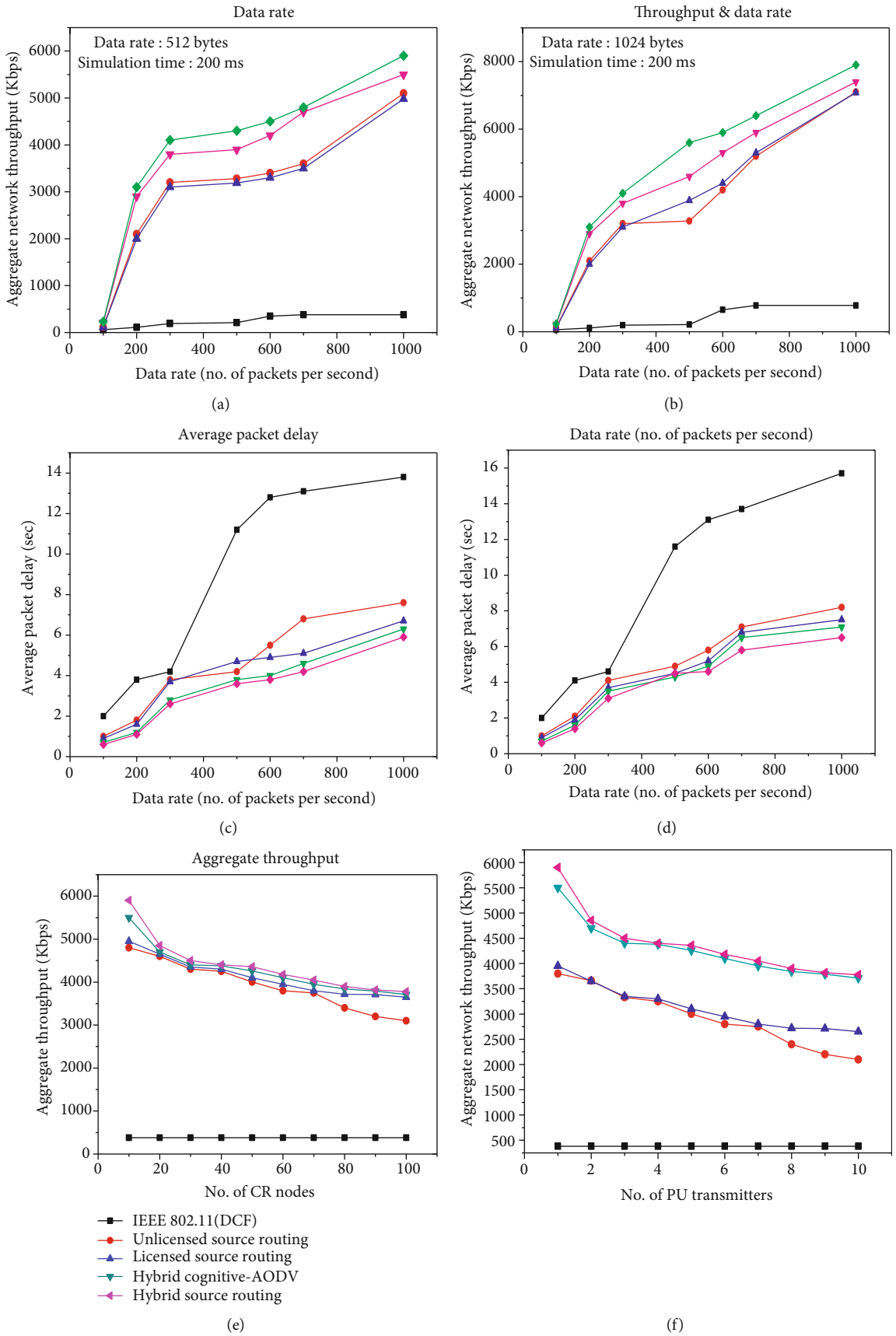



Figure 8: (a) Data rate, (b) throughput and data rate, (c) average packet delay, (d) average packet delay and data rate, (e) aggregate throughput, and (f) performance analysis of proposed source routing-based distributed cognitive IoT networks.

## 5. Conclusion

End-to-end channel route failure, spectrum mobility, and node mobility at the intermediate CR nodes during application data transmission play a significant role in the performance of distributed cognitive IoT networks [35, 36]. In this work, the distributed source routing protocol for big data-based [17] cognitive IoT is proposed to enhance the end-to-end throughput with minimized end-to-end delay. In addition, the gateway distributed CR routers will also act as IoT gateway nodes to aggregate and encapsulate the IoT data into the distributed cognitive radio ad hoc network [37]. The detailed simulation for each and every channel route failure with respect to different performance metrics will be analyzed as a future work. In the future, directional antenna-based source routing is planned to be implemented within the current cognitive source routing to efficiently reuse the unlicensed spectrum bands, minimize the interference, and enhance the achievable aggregate network throughput.

## Data Availability

All the data are available in the paper.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References


[1] S. H. Chun and R. J. La, "Secondary spectrum trading—auction-based framework for spectrum allocation and profit sharing," *IEEE/ACM Trans. Netw.*, vol. 21, no. 1, pp. 176–189, 2013.

[2] C. Lacatus, D. Akopian, Y. Prasad, and M. Shadaram, "Flexible spectrum and power allocation for OFDM unlicensed wireless systems," *IEEE Systems Journal*, vol. 3, no. 2, pp. 254–264, 2009.

[3] T. Baykas, M. Kasslin, M. Cummings et al., "Developing a standard for TV white space coexistence: technical challenges and solution approaches," *IEEE Wireless Communications*, vol. 19, no. 1, pp. 10–22, 2012.

[4] Y. S. Chen and J. S. Hong, "A relay-assisted protocol for spectrum mobility and handover in cognitive LTE networks," *IEEE Systems Journal*, vol. 7, no. 1, pp. 77–91, 2013.

[5] M. Faheem, S. B. H. Shah, R. A. Butt et al., "Smart grid communication and information technologies in the perspective of Industry 4.0: opportunities and challenges," *Computer Science Review*, vol. 30, pp. 1–30, 2018.

[6] S. Sengupta and K. P. Subbalakshmi, "Open research issues in multi-hop cognitive radio networks," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 168–176, 2013.

[7] R. Tandra, S. M. Mishra, and A. Sahai, "What is a spectrum hole and what does it take to recognize one?," *Proceedings of IEEE*, vol. 97, no. 5, pp. 824–848, 2009.

[8] B. F. Lo, "A survey of common control channel design in cognitive radio networks," *Physical Communication*, vol. 4, no. 1, pp. 26–39, 2011.

[9] H. Khalife, N. Malouch, and S. Fdida, "Multihop cognitive radio networks: to route or not to route," *Network, IEEE*, vol. 23, no. 4, pp. 20–25, 2009.

[10] N. Devroye, M. Vu, and V. Tarokh, "Cognitive radio networks," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 12–23, 2008.

[11] C. Iwendi, P. K. Reddy, T. R. Gadekallu, and M. J. Piran, "A metaheuristic optimization approach for energy efficiency in the IoT networks," *Software Practice and Experience*, 2020.

[12] M. Alazab, K. Lakshmanna, G. Thippa Reddy, Q.-V. Pham, and P. K. R. Maddikunta, "Multi-objective cluster head selection using fitness averaged rider optimization algorithm for IoT networks in smart cities," *Sustainable Energy Technologies and Assessments*, vol. 43, p. 100973, 2021.

[13] R. Kaluri, D. S. Rajput, Q. Xin, K. Lakshmanna, and S. Bhattacharya, "Roughsets-based approach for predicting battery life in IoT," *Intelligent Automation & Soft Computing*, vol. 27, no. 2, pp. 453–469, 2021.

[14] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[15] M. Bouaziz, A. Rachedi, and A. Belghith, "EC-MRPL: an energy-efficient and mobility support routing protocol for Internet of Mobile Things," in *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pp. 19–24, Las Vegas, NV, 2017.

[16] M. A. Ferrag, M. Derdour, M. Mukherjee, A. Derhab, L. Maglaras, and H. Janicke, "Blockchain technologies for the Internet of things: research issues and challenges," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2188–2204, 2019.

[17] M. A. Jan, J. Cai, X. C. Gao et al., "Security and blockchain convergence with Internet of multimedia things: current trends, research challenges and future directions," *Journal of Network and Computer Applications*, vol. 175, 2020.

[18] S. Anamalamudi and M. Jin, "Hybrid common control channel based MAC protocol for cognitive radio ad-hoc networks," *International Journal of Information and Electronics Engineering*, vol. 4, no. 3, pp. 216–224, 2014.

[19] E. F. Jesus, V. R. L. Chicarino, C. V. N. de Albuquerque, and A. de Rocha, "A survey of how to use blockchain to secure Internet of things and the stalker attack," *Security and Communication Networks*, vol. 2018, Article ID 9675050, 2018.

[20] Z. Yang, D. Niyato, P. Wang, and E. Hossain, "Auction-based resource allocation in cognitive radio systems," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 108–120, 2012.

[21] S. B. H. Shah, Z. Chen, F. Yin, and A. Ahmad, "Water rippling shaped clustering strategy for efficient performance of software define wireless sensor networks," *Peer-to-Peer Networking and Applications*, vol. 12, no. 2, 2017.

[22] G. A. Safdar and M. O'Neill, "Common control channel security framework for cognitive radio networks," in *IEEE conference on Vehicular Technology*, pp. 1–5, Barcelona, 2009.

[23] B. Hu and H. Gharavi, "Directional routing protocols for ad-hoc networks," *IET Communications*, vol. 2, no. 5, pp. 650–657, 2008.

[24] R. R. Choudhury and H. V. Nitin, "Performance of ad hoc routing using directional antennas," *International Journal of Ad Hoc Networks*, vol. 3, no. 2, pp. 157–173, 2005.

[25] D. Fei and W. Jie, "Efficient broadcasting in ad hoc wireless networks using directional antennas," *IEEE Transactions on*

*Parallel and Distributed Systems*, vol. 17, no. 4, pp. 335–347, 2006.

[26] S. B. Shah, C. Zhe, F. Yin et al., "3D weighted centroid algorithm & RSSI ranging model strategy for node localization in WSN based on smart devices," *Sustainable Cities and Society*, vol. 39, pp. 298–308, 2018.

[27] D. Ying and W. Jie, "Boundary helps: efficient routing protocol using directional antennas in cognitive radio networks," in *IEEE Conference on Mobile Ad-Hoc and Sensor Systems*, pp. 502–510, Hangzhou, China, 2013.

[28] S. B. Shah, Z. Chen, F. Yin, I. U. Khan, and N. Ahmad, "Energy and inter-operable aware routing for throughput optimization in clustered IoT-wireless sensor networks," *Future Generation Computer Systems*, vol. 81, pp. 372–381, 2017.

[29] State Radio Regulation of China, "Spectrum allocations overview in P.R. China," http://www.srrc.org.cn/NewsShow12074.aspx.

[30] Extension to nshttp://www.monarch.cs.rice.edu/.

[31] J. Zhong, *Development of NS-2 Based Cognitive Radio Cognitive Network Simulator*, MS Thesis, Michigan Technological University, 2009.

[32] C. Cordeiro, K. Challapali, D. Birru, and S. Shankar, "IEEE 802.22: the first worldwide wireless standard based on cognitive radios," in *First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005*, 2005.

[33] S. Mastorakis, T. Li, and L. Zhang, "DAPES: named data for off-the-grid file sharing with peer-to-peer interactions," in *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, 2020.

[34] X. Fu and Y. Yang, "Modeling and analysis of cascading node-link failures in multi-sink wireless sensor networks," *Reliability Engineering & System Safety*, vol. 197, p. 106815, 2020.

[35] S. Mastorakis, A. Mtibaa, J. Lee, and S. Misra, "ICedge: when edge computing meets information-centric networking," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4203–4217, 2020.

[36] R. Ullah, M. A. U. Rehman, M. A. Naeem, B. S. Kim, and S. Mastorakis, "ICN with edge for 5G: exploiting in-network caching in ICN-based edge computing for 5G networks," *Future Generation Computer Systems.*, vol. 111, pp. 159–174, 2020.

[37] M. A. U. Rehman, R. Ullah, B. S. Kim, B. Nour, and S. Mastorakis, "CCIC-WSN: an architecture for single channel cluster-based information-centric wireless sensor networks," *IEEE Internet of Things Journal*, vol. 8, no. 9, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9258411, 17 pages
https://doi.org/10.1155/2021/9258411

*Research Article*

# Active Fault-Tolerant/Active Passive Intrusion-Tolerant $H_\infty$ Cooperative Control of Discrete NCS under the Background of Big Data

**Wang Jun**[1] **and Meng Xiao-li**[1,2]

[1]*College of Electrical and Information Engineering, Lanzhou Univ. of Tech., Lanzhou 730050, China*
[2]*Key Lab of Advanced Control for Industrial Process in Gansu Province, Lanzhou 730050, China*

Correspondence should be addressed to Meng Xiao-li; mengxl0628@163.com

Received 20 April 2021; Accepted 16 June 2021; Published 19 July 2021

Academic Editor: Thippa Reddy G



For the linear discrete networked control system (NCS) which may suffer DoS attack on both sides of the controller, when the actuator has time-varying failure, the intelligent sensor unit uses wireless sensors to collect data. According to the large amount of data collected, the active fault tolerance/active passive capacity of linear discrete NCS under the discrete event-triggered communication mechanism (DETCS) is studied. The problem of cooperative controller design is discussed. Firstly, a linear discrete NCS model integrating DETCS, actuator fault, and network attack is established. Then, based on the idea of integral sliding mode control, an active fault-tolerant/attack active passive intrusion-tolerant cooperative controller is designed, and the actuator attack side network attack and sensor side network attack are extended to the state to obtain a new state vector. Then, an adaptive Kalman filter estimator (AKF) estimates the fault and attack information and then adjusts the initial fault-tolerant/intrusion-tolerant cooperative controller in real time according to the estimated information obtained by the adaptive Kalman filter estimator; finally, the MATLAB simulation example is used to verify the improvement of system performance by the designed control law and the saving of network resources by the introduction of DETCS.

## 1. Introduction

Compared with the traditional control system, NCS has the unique properties of remote operation, remote monitoring, security detection and resource sharing due to the existence of networks. However, due to the existence of networks, the structure of the control system is more complex, the data transmission will be affected to a certain extent, the system is more prone to component failure, and network transmission is more vulnerable to attack, so the physical fault tolerance is very important The research of intrusion control and network attack tolerance control has great practical significance and practical application value [1–3].

With the rapid development of various wireless sensor technologies, people can gradually perceive the real world through wireless sensors in every corner of the world, even in places where people rarely visit. These wireless sensors transmit the data to the computer on people's worktable through the network, which is convenient for scholars to carry out the corresponding research on all kinds of data. Generally speaking, fault-tolerant control (FTC) technology is divided into passive fault-tolerant control (PFTC) [4–6] and active fault-tolerant control (AFTC) [7–10]. PFTC uses a controller to make the closed-loop system insensitive to specific faults and maintain the stability and performance of the system. On the basis of acquiring fault information, AFTC designs a fault-tolerant controller to ensure the performance and stability of the system. For the actuator fault of the system, Zhang et al. proposed an adaptive observer [8, 9] by which the system state and fault can be obtained and the fault can be adjusted. Refer-

ence [10] studies the problem of adaptive fault-tolerant control for uncertain actuator fault compensation of linear time invariant NCS with unknown plant parameters and actuator fault parameters. Intrusion tolerance is developed on the basis of fault tolerance, but it is very different from fault tolerance. It is mainly reflected in that network attacks are malicious in motivation and difficult to judge in form. Denial of service (DoS) attack is a kind of attack means by weakening and preventing legitimate users from using legitimate network resources, which can affect the normal use of the network. Reference [11] studies the stability of control and measurement packets transmitted over a communication network under DoS attack. Reference [12] studies the event-triggered attack-tolerant tracking control problem of nonlinear NCS under sudden DoS attack. In Reference [13], the progress of security control and attack detection of industrial information physical fusion system is reviewed from the perspective of control theory. Although many achievements have been made in fault tolerance and attack tolerance of NCS, most of them are limited to independent design. For NCS with actuator failure and network attack coexisting, the research on comprehensive security control of fault tolerance and intrusion tolerance is rarely involved, and only literature [14] and literature [15] can be consulted at present. However, in actual NCS, it is inevitable that faults and attacks occur simultaneously.

In view of this, aiming at the actuator fault and DoS network attack on both sides of the controller, this paper establishes a closed-loop linear NCS fault/attack coexistence model by constructing the active fault-tolerant/active passive intrusion-tolerant system structure of NCS under DETCS with the help of the joint modeling method of network attack and controlled object. In addition, adaptive Kalman filter estimators are designed, respectively, for online monitoring of state, attack, and fault. The active fault-tolerant control is used to compensate the actuator network attack with the active intrusion-tolerant strategy, and the passive intrusion-tolerant mechanism is used to be robust to the sensor network attack, so as to realize the state feedback and active fault-tolerant and active passive intrusion-tolerant cooperative control.

## 2. Material and Method

*2.1. System Model.* Controlled object model:

$$\begin{cases} x(k+1) = Ax(k) + Bu(k) + Ef(k) + N_a a_a(k) + D_1 w(k), \\ y(k) = Cx(k) + N_s a_s(k) + D_2 v(k). \end{cases} \tag{1}$$

Among them, $x(k) \in R^n$, $y(k) \in R^m$, and $u(k) \in R^l$ are the state, output, and control input of the system, respectively; $f(k) \in R^p$ is the time-varying fault vector of the actuator; $a_a(k) \in R^q$ is the constant network attack vector suffered by the actuator side, referred to as the actuator attack vector; $a_s(k) \in R^q$ is the constant network attack vector suffered by the sensor side, referred to as the sensor attack vector; $A \in R^{n\times n}, B \in R^{n\times l}$, $C \in R^{m\times n}, E \in R^{n\times p}$, and $N_s \in R^{n\times q}$ are the coefficient matrices of proper dimension; and $w(k) \in R^p$ and $v(k) \in R^p$ are the perturbations of bounded energy. They are Gaussian white noise sequences with 0 mean value and independent of each other. Their covariance matrices are $Q \in R^{p\times p}$ and $R \in R^{p\times p}$; $D_1 \in R^{n\times p}$ and $D_2 \in R^{m\times p}$ are the perturbations of proper dimension.

Referring to the idea of state augmentation, the system state $x(k)$, actuator attack $a_a(k)$, and sensor attack $a_s(k)$ are augmented into a new state $\eta(k)$, which is called the augmented state. The augmented system model is

$$\begin{cases} \eta(k+1) = \tilde{A}\eta(k) + \tilde{B}u(k) + \tilde{E}f(k) + \tilde{D}_1\tilde{w}(k), \\ y(k) = \tilde{C}\eta(k) + D_2 v(k), \end{cases} \tag{2}$$

where

$$\eta(k) = \begin{bmatrix} x^T(k) \\ a_a^T(k) \\ a_s^T(k) \end{bmatrix}, \tilde{w}(k) = \begin{bmatrix} w^T(k) \\ b_a^T(k) \\ b_s^T(k) \end{bmatrix}, \tilde{A} = \begin{bmatrix} A & N_a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$
$$\tilde{B} = \begin{bmatrix} B \\ 0 \\ 0 \end{bmatrix}, \tilde{E} = \begin{bmatrix} E \\ 0 \\ 0 \end{bmatrix}, \tilde{D}_1 = \begin{bmatrix} D & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}, \tilde{C} = \begin{bmatrix} C \\ 0 \\ N_s \end{bmatrix}^T. \tag{3}$$

*2.2. DETCS Trigger Conditions.* This paper adopts the most representative discrete event trigger condition [16]:

$$[x\wedge(k) - x\wedge(t_k)]^T \mathbf{\Phi} [\hat{x}(k) - \hat{x}(t_k)] \le \sigma x\wedge^T(t_k) \mathbf{\Phi} \hat{x}(t_k). \tag{4}$$

Among them, $\hat{x}(k)$ is the estimated value of the state at the current sampling time, $\hat{x}(t_k)$ is the latest estimated value of the state at the network transmission time that satisfies equation (4) at the previous time, $\mathbf{\Phi}$ is the event trigger matrix with positive definite symmetry, and $\sigma \in [0, 1)$ is the event trigger parameter scalar, which can be set in advance and is related to the expected performance of the system. Therefore, only when the sampling data $\hat{x}(k)$ meets the trigger condition of equation (4) can it be transmitted from the sensor network to the controller. A lot of network resources is saved, so as to improve the utilization of the network. The system control structure is shown in Figure 1.

Define delay function $\tau(k) = k - t_k$, $k \in [t_k, t_{k+1})$, and $0 \le \tau(k) \le \tau_M$. Among them, $\tau_M \triangleq T_{k\max} + \tilde{\tau}$ are the maximum nonuniform transmission periods of data filtering, and $\tilde{\tau}$ is the upper bound of delay function $\tau(k)$ at $t_k + 1$.

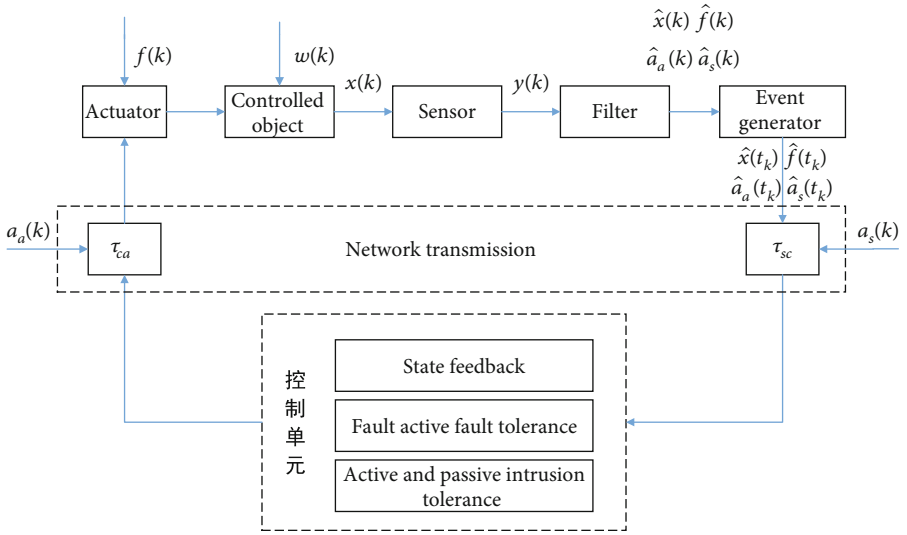



Figure 1: System control structure diagram.

*2.2.1. Related Lemmas*

**Lemma 1** (Schur complement lemma [17]). *For a given symmetric matrix*

$$Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{12}^T & Z_{22} \end{bmatrix}, \tag{5}$$

*the following three conditions are equivalent:*

$$\begin{aligned} &Z < 0, \\ &Z_{11} < 0, Z_{22} - Z_{12}^T Z_{11}^{-1} Z_{12} < 0, \\ &Z_{22} < 0, Z_{11} - Z_{12} Z_{22}^{-1} Z_{12}^T < 0. \end{aligned} \tag{6}$$

**Lemma 2** (Wirtinger inequality in discrete form [18, 19]). *For a given positive definite matrix $N \in R^{n\times n}$, scalar $0 \le \rho_1 \le \rho_2$ and vector function $\eta : [-\rho_2, \rho_1] \longrightarrow R^{n\times n}$ satisfy the following inequalities:*

$$-(\rho_2 - \rho_1) \sum_{s=k-\rho_2}^{k-\rho_1-1} \eta^T(s) N \eta(s) \le -\begin{bmatrix} \boldsymbol{\Omega}_1 \\ \boldsymbol{\Omega}_2 \end{bmatrix}^T \begin{bmatrix} N & 0 \\ 0 & 3N \end{bmatrix} \begin{bmatrix} \boldsymbol{\Omega}_1 \\ \boldsymbol{\Omega}_2 \end{bmatrix}. \tag{7}$$

*Among them,*

$$\begin{aligned} &\eta(s) = x(s+1) - x(s), \\ &\boldsymbol{\Omega}_1 = x(k-\rho_1) - x(k-\rho_2), \\ &\boldsymbol{\Omega}_2 = x(k-\rho_1) + x(k-\rho_2) - \frac{2}{\rho_2 - \rho_1 + 1} \sum_{s=k-\rho_2}^{k-\rho_1} x(s). \end{aligned} \tag{8}$$

*2.3. Preliminary Design of Active Fault-Tolerant/Active Passive Intrusion-Tolerant Cooperative Control Law.* For system (1), the control law can be designed as follows:

$$u(k) = u_n(k) + u_{fa}(k). \tag{9}$$

Among them, $u_n(k)$ is the nominal control law when there is no fault and no attack, and $u_{fa}(k)$ is the compensation control law when fault and attack occur simultaneously.

Considering the time delay problem, the sliding mode function can be designed as follows:

$$S(k) = Gx(k - \tau(k)) + \sum_{i=k-\tau(k)}^{k-1} GBu(i) + \boldsymbol{\Gamma}(k). \tag{10}$$

Among them, $G$ is a constant matrix with proper dimension, and $GB$ is a nonsingular matrix; $\boldsymbol{\Gamma}(k)$ is a sliding mode discrete compensator, which satisfies the following requirements: $\boldsymbol{\Gamma}(k+1) = \boldsymbol{\Gamma}(k) - G(A - BK)x(k - \tau(k)) + Gx(k - \tau(k))$, where $K$ is the gain matrix of the controller; the sliding mode of the system can be obtained by introducing $\sum_{i=k-\tau(k)}^{k-1} GBu(i)$.

Omitting $w(k)$ and $v(k)$, combining formula (1) and formula (10), we can get

$$\begin{aligned} \Delta S(k) &= S(k+1) - S(k) Gx(k+1-\tau(k+1)) + \sum_{i=k+1-\tau(k)}^{k} GBu(i) \\ &\quad + \Gamma(k+1) - Gx(k - \tau(k)) + \sum_{i=k-\tau(k)}^{k-1} GBu(i) + \Gamma(k) \\ &= GAx(k-\tau(k)) + GBu(k-\tau(k)) + GEf(k-\tau(k)) \\ &\quad + GN_a a_a(k-\tau(k)) - Gx(k-\tau(k)) + GBu(k) \\ &\quad - GBu(k-\tau(k)) + Gx(k-\tau(k)) - GAx(k-\tau(k)) \\ &\quad + GBKx(k-\tau(k)) = GBKx(k-\tau(k)) + GBu(k) \\ &\quad + GEf(k-\tau(k)) + GN_a a_a(k-\tau(k)). \end{aligned} \tag{11}$$

If the order is $\Delta S(k) = 0$, there will be

$$u(k) = -(GB)^{-1}[GBKx(k-\tau(k)) + GEf(k-\tau(k)) + GN_a a_a(k-\tau(k))]. \quad (12)$$

Furthermore, the sliding mode control law $u(k)$ based on the reaching law method is

$$u(k) = -Kx(k-\tau(k)) - (GB)^{-1}[qS(k) + \varepsilon \operatorname{sgn} S(k)] - (GB)^{-1}[GEf(k-\tau(k)) + GN_a a_a(k-\tau(k))] + N_s a_s(k-\tau(k)), \quad (13)$$

where $q, \varepsilon$ is a constant scalar and $q > 0, \varepsilon > 0$.

When there is no fault and no attack, the nominal control law $u_n(k)$ is

$$u_n(k) = -Kx(k-\tau(k)) - (GB)^{-1}(qS(k) + \varepsilon \operatorname{sgn} S(k)). \quad (14)$$

When faults and attacks occur simultaneously, the compensation control law $u_{fa}(k)$ is

$$u_{fa} = -(GB)^{-1}[GEf(k-\tau(k)) + GN_a a_a(k-\tau(k))]. \quad (15)$$

**Theorem 3.** *When constant $q, \varepsilon$ satisfies $q > 0, \varepsilon > 0$, respectively, the system (1) satisfies the existence and reachability conditions of the sliding mode.*

*Proof.* Select Lyapunov function: $V(k) = S^T(k)S(k)$.

Combining formulas (1) and (10), we can get

$$\begin{aligned}\Delta V(k) &= S^T(k)\Delta S(k) = S^T(k)[S(k+1) - S(k)] \\ &= S^T(k)[GBKx(k-\tau(k)) + GBu(k) \\ &\quad + GEf(k-\tau(k)) + GN_a a_a(k-\tau(k))].\end{aligned} \quad (16)$$

According to the control law (13),

$$\Delta V(k) = S^T(k)[-qS(k) - \varepsilon \operatorname{sgn} S(k)]. \quad (17)$$

When the constants $q, \varepsilon$ satisfy $q > 0, \varepsilon > 0$, there are

$$\Delta V(k) = -S^2(k) - \varepsilon S(k) \le 0. \quad (18)$$

Therefore, the reachability condition of the sliding surface is satisfied. So, the system state can reach the sliding surface in finite time.

When the system is on the sliding mode surface, i.e., $S(k) = 0$, the equivalent control term of the sliding mode control law can be obtained by substituting it into equation (13):

$$u_{eq}(k) = -Kx(k-\tau(k)) - (GB)^{-1}[GEf(k-\tau(k)) + GN_a a_a(k-\tau(k))] + N_s a_s(k-\tau(k)). \quad (19)$$

By substituting equation (19) into equation (1), the closed-loop model of NCS on a sliding surface can be obtained:

$$\begin{cases} x(k+1) = Ax(k) - Kx(k-\tau(k)) - B(GB)^{-1}[GEf(k-\tau(k)) + GN_a a_a(k-\tau(k))] + BN_s a_s(k-\tau(k)) + Ef(k) + N_a a_a(k) + D_1 w(k), \\ y(k) = Cx(k) + N_s a_s(k) + D_2 v(k), \\ x(k) = \phi(k), -\tau_M \le k \le 0, \end{cases} \quad (20)$$

where $\phi(k)$ is a real valued initial function on $[-\tau_M, 0]$. □

*2.4. Design of Adaptive Kalman Filter Fault/Attack Estimator.* For augmented system (2), the following adaptive Kalman filter fault/attack estimator is designed.

The prediction part of the adaptive Kalman filter fault/attack estimator is as follows:

$$\begin{cases} \hat{\eta}(k+1 \mid k) = \tilde{A}\hat{\eta}(k) + \tilde{B}u(k) + \tilde{E}\hat{f}(k), \\ P(k+1 \mid k) = \tilde{A}P(k-1 \mid k-1)\tilde{A}^T + \tilde{D}_1 Q \tilde{D}_1^T, \\ \tilde{y}(k+1) = y(k+1) - \tilde{C}\hat{\eta}(k+1 \mid k). \end{cases} \quad (21)$$

Among them, $P \in R^{n \times n}$ is the state error covariance matrix, and $\tilde{y}(k)$ is the error between the system output and the estimated output.

The gain matrix of adaptive Kalman filter fault/attack estimator is as follows:

$$\begin{cases} \Sigma(k) = \tilde{C}P(k \mid k-1)\tilde{C}^T + D_2 R D_2^T, \\ K_1(k) = P(k \mid k-1)\tilde{C}^T \Sigma^{-1}(k). \end{cases} \quad (22)$$

Among them, $K_1(k)$ is the gain matrix of Kalman filter.

The error gain matrix of adaptive Kalman filter fault/attack estimator is as follows:

$$\begin{cases} \mathbf{\Omega}(k) = \tilde{C}\gamma(k) + \tilde{C}\tilde{E}, \\ \gamma(k+1) = \left(\tilde{A} - K_1(k)\tilde{C}\right)\gamma(k) + \tilde{E}, \\ F(k) = J^T(k-1)\mathbf{\Omega}^T(k)\mathbf{\Lambda}^T(k), \\ J(k) = \dfrac{1}{\lambda}\left[I - J(k-1)\mathbf{\Omega}^T(k)\mathbf{\Lambda}(k)\mathbf{\Omega}(k)\right]J(k-1), \\ \mathbf{\Lambda}(k) = \left[\lambda\Sigma(k) + \mathbf{\Omega}(k)J(k-1)\mathbf{\Omega}^T(k)\right]^{-1}. \end{cases} \tag{23}$$

Among them, $\gamma(k)$, $F(k)$ is the error gain matrix, and $\lambda \in [0, 1)$ is the forgetting factor.

The update part of adaptive Kalman filter fault/attack estimator is as follows:

$$\begin{cases} \hat{\eta}(k+1) = \hat{\eta}(k+1 \mid k) + K_1(k)\tilde{y}(k) + \gamma(k+1)\left(\hat{f}(k+1) - \hat{f}(k)\right), \\ \hat{f}(k+1) = \hat{f}(k) + F(k)\tilde{y}(k), \\ P(k+1 \mid k+1) = \left[I - K_1(k)\tilde{C}\right]P(k+1 \mid k). \end{cases} \tag{24}$$

Suppose 4. Matrix $[\tilde{A}, \tilde{C}]$ is completely observable, $[\tilde{A}, Q^{1/2}]$ is completely controllable, if matrix $\xi(k)$ satisfies

$$\xi(k+1) = \left(\tilde{A} - K_1(k)\tilde{C}\right)\xi(k). \tag{25}$$

Then, for Kalman gain matrix $K_1(k)$, $\xi(k)$ are exponentially convergent.

Suppose 5. Under initial condition $J(0) = \omega I(\omega > 0)$, matrix $J(k)$ is strictly positive definite.

System error definition:

$$\begin{cases} \tilde{\eta}(k) = \eta(k) - \hat{\eta}(k), \\ \tilde{f}(k) = f(k) - \hat{f}(k), \\ z(k) = \tilde{\eta}(k) - \gamma(k)\tilde{f}(k). \end{cases} \tag{26}$$

From equations (2), (21), and (26),

$$z(k+1) = \left(\tilde{A} - K_1(k)\tilde{C}\right)z(k) - K_1(k)D_2v(k) + \tilde{D}_1\tilde{w}(k). \tag{27}$$

Further, we can get

$$E[z(k+1)] = \left(\tilde{A} - K_1(k)\tilde{C}\right)E[z(k)], \tag{28}$$

where $E[z(k)]$ is the mean of $z(k)$.

From hypothesis 1 and equation (28), we can see that 14 is exponentially convergent.

From equations (21) and (24),

$$\tilde{f}(k+1) = (I - F(k)\mathbf{\Omega}(k))\tilde{f}(k) - F(k)\left(\tilde{C}z(k) + D_2v(k)\right). \tag{29}$$

Because of $E[z(k)] = 0$, $E[v(k)] = 0$,

$$E\left[\tilde{f}(k+1)\right] = [I - F(k)\mathbf{\Omega}(k)]E\left[\tilde{f}(k)\right]. \tag{30}$$

From hypothesis 2, we can see that $J(k)$ is strictly positive definite.

**Theorem 6.** *When matrix $J(k)$ is strictly positive definite and $E[\tilde{\eta}(k)]$ exponentially approaches 0, the adaptive Kalman filter is convergent. Note $M(k) = J^{-1}(k)$.*

*Proof.* Define Lyapunov function:

$$V(k+1) = \left(E\left[\tilde{f}(k+1)\right]\right)^T M(k)E\left[\tilde{f}(k+1)\right]. \tag{31}$$

From equation (23), we can get

$$J(k) = \frac{1}{\lambda}[I - F(k)\mathbf{\Omega}(k)]J(k-1). \tag{32}$$

Then,

$$M(k) = \lambda M(k-1)[I - F(k)\mathbf{\Omega}(k)]^{-1}. \tag{33}$$

From equations (32) and (33),

$$\begin{aligned} V(k+1) &= \left(E\left[\tilde{f}(k)\right]\right)^T [I - F(k)\mathbf{\Omega}(k)]^T\lambda M(k-1)E\left[\tilde{f}(k)\right] \\ &= \lambda\left(E\left[\tilde{f}(k)\right]\right)^T M(k-1)E\left[\tilde{f}(k)\right] \\ &\quad - \lambda\left(E\left[\tilde{f}(k)\right]\right)^T \Xi E\left[\tilde{f}(k)\right]. \end{aligned} \tag{34}$$

In equation (34),

$$\begin{aligned} \Xi(k) &= \mathbf{\Omega}^T(k)\mathbf{\Gamma}^T(k)M(k-1) \\ &= \mathbf{\Omega}^T(k)\mathbf{\Lambda}(k)\mathbf{\Omega}(k)J(k-1)J^{-1}(k-1) \\ &= \mathbf{\Omega}^T(k)\mathbf{\Lambda}(k)\mathbf{\Omega}(k). \end{aligned} \tag{35}$$

Since $\mathbf{\Lambda}(k)$ is a positive definite matrix, then $\Xi(k)$ is also a positive definite matrix. So,

$$V(k+1) \le \lambda\left(E\left[\tilde{f}(k)\right]\right)^T M(k) E\left[\tilde{f}(k)\right] \le \lambda V(k). \quad (36)$$

It can be seen from equation (36) that $V(k)$ is exponentially close to 0. Since matrix $M(k)$ is strictly positive definite, $E\tilde{f}(k)$ is exponentially close to 0.

From equation (26),

$$E[\tilde{\eta}(k)] = E[z(k)] + \gamma(k) E\left[\tilde{f}(k)\right]. \quad (37)$$

□

Then, $E[\tilde{\eta}(k)]$ is also an index close to 0, so the adaptive Kalman filter is convergent.

## 3. Analysis

The online fault-tolerant/intrusion-tolerant control for system (1) can be described as follows: using the estimated value of system state $\hat{x}(k)$, actuator fault $\hat{f}(k)$, and network attack $\hat{a}_a(k)$ on the actuator side, considering the event triggering conditions, substituting the estimated value satisfying the conditions into control law $u(k)$ (13), the adjusted fault-tolerant/intrusion-tolerant control law $u(k)$ is

$$\begin{aligned} u(k) = &-K\hat{x}(t_k) - (GB)^{-1}[qS(k) + \varepsilon \operatorname{sgn} S(k)] \\ &- (GB)^{-1}\left[GE\hat{f}(t_k) + GN_a\hat{a}_a(t_k)\right] + N_s a_s(t_k). \end{aligned} \quad (38)$$

Suppose 7. Actuator fault $f(k)$ is bounded, i.e., $\|f(k)\| \le \rho_1$; $a_a(k)$ is bounded actuator attack, i.e., $\|a_a(k)\| \le \rho_2$, where $\rho_1, \rho_2$ are known functions greater than 0.

The existence condition and global reachability condition of sliding mode are

$$S^T(k)\Delta S(k) < 0. \quad (39)$$

By substituting equation (38) into equation (39), we find that

$$\begin{aligned} S^T(k)\Delta S(k) &= S^T(k)[GBKx(t_k) + GBu(k) + GEf(t_k) + GN_a a_a(t_k)] \\ &= S^T(k)\Big[GBKx(t_k) + GEf(t_k) + GN_a a_a(t_k) \\ &\quad + GB\Big(-K\hat{x}(t_k) - (GB)^{-1}\Big(qS(k) + \varepsilon \operatorname{sgn} S(k) \\ &\quad + GE\hat{f}(t_k) + GN_a\hat{a}_a(t_k)\Big)\Big)\Big] \\ &= S^T(k)\Big[GBK(x(t_k) - \hat{x}(t_k)) + GE\Big(f(t_k) - \hat{f}(t_k)\Big) \\ &\quad + GN_a(a_a(t_k) - \hat{a}_a(t_k)) - qS(k) - \varepsilon \operatorname{sgn} S(k)\Big]. \end{aligned} \quad (40)$$

From hypothesis 3,

$$\left\|f(t_k) - \hat{f}(t_k)\right\| \le \rho_1, \|a_a(t_k) - \hat{a}_a(t_k)\| \le \rho_2 \quad (41)$$

can be obtained.

Then,

$$\begin{aligned} S^T(k)\Delta S(k) \le &-qS^2(k) - \varepsilon\left|S^T(k)\right| + \left|S^T(k)\right|(GBKe \\ &+ GE\rho_1 + GN_a\rho_2). \end{aligned} \quad (42)$$

When the constant $q, \varepsilon$ satisfies the following inequality:

$$\varepsilon \ge (GBKe + GE\rho_1 + GN_a\rho_2);\ q > 0. \quad (43)$$

Then, $S^T(k)\Delta S(k) \le -qS^2(k) \le 0$.

Therefore, the existence and reachability conditions of the sliding mode of the system (20) are satisfied, which means that the starting point from any initial state other than the sliding mode surface $S(k) = 0$ can return to the sliding mode surface in a finite time.

When the system is on the sliding surface, the equivalent control term of the controller can be obtained by combining equation (54) with equation $S(k) = 0$:

$$u_{eq}(k) = -K\hat{x}(t_k) + N_s a_s(t_k) - (GB)^{-1}\left[GE\hat{f}(t_k) + GN_a\hat{a}_a(t_k)\right]. \quad (44)$$

By substituting equation (44) into equation (1), the state space model of the closed-loop system on the sliding surface can be obtained as follows:

$$\begin{cases} x(k+1) = Ax(k) + Ef(k) + N_a a_a(k) + D_1 w(k) + B, \\ \left[-K\hat{x}(t_k) - (GB)^{-1}\left(GE\hat{f}(t_k) + GN_a\hat{a}_a(t_k)\right)\right] + N_s a_s(t_k), \\ y(k) = Cx(k) + N_s a_s(k) + D_2 v(k), \\ x(k) = \phi(k), k \in [-\tau_M, 0], \end{cases} \quad (45)$$

where $\phi(k)$ is a real valued initial function on $[-\tau_M, 0]$.

Since the system has satisfied the reachability condition of the sliding mode surface, in order to ensure the stability, the sliding mode described by system (45) can be stabilized by designing an appropriate gain matrix $K$:

$$\begin{bmatrix} X_1 & * & * & * & * & * & * & * & * & * & * & * & * \\ Y_1 & -Q_1 & * & * & * & * & * & * & * & * & * & * & * \\ 2S_1 & 0 & -4S_1 & * & * & * & * & * & * & * & * & * & * \\ 6S_1 & 0 & 6S_1 & -12S_1 & * & * & * & * & * & * & * & * & * \\ \tau_M^2(BK)^T & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * & * & * & * \\ \tau_M^2(BK)^T & 0 & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * & * & * \\ \tau_M^2(BK)^T & 0 & 0 & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * & * \\ \tau_M^2(BK)^T & 0 & 0 & 0 & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma\mathbf{\Phi} & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\mathbf{\Phi} & * & * & * \\ \tau_M AS_1 & -\tau_M BK_1 & 0 & 0 & -\tau_M BK_1 & -\tau_M E & -\tau_M N_a & \tau_M D_1 & 0 & 0 & -S_1 & * & * \\ AS_1 & -BK_1 & 0 & 0 & -BK_1 & -E & -N_a & D_1 & 0 & 0 & 0 & -2S_1 + P_1 & * \\ S_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -I \end{bmatrix} < 0. \tag{46}$$

Among them,

$$\begin{cases} Y_1 = \tau_M^2 K_1^T B^T - 2S_1, \\ X_1 = -P_1 + Q_1 + R_1 + \left(\tau_M^2 - 4\right)S_1 - \tau_M^2 S_1 A^T - \tau_M^2 A S_1, \end{cases} \tag{47}$$

and $K_1 = K\tilde{S}$.

**Theorem 8.** *Given scalar $\sigma > 0, \tau_M > 0, \gamma > 0$, if there are matrix K, symmetric positive definite matrix $P_1, Q_1, R_1, S_1$, and event triggering matrix $\mathbf{\Phi}$ satisfying the following matrix inequalities (46), then the sliding mode (45) of the closed-loop system is asymptotically stable at $H_\infty$ disturbance rejection level $\gamma$.*

It is proved that the stability of mode (45) of the closed-loop system is investigated in the presence of external disturbance $w(k)$, actuator failure $f(k)$ and network attack $a_a(k)$. The asymptotic stability condition and robust performance index of the closed-loop system are as follows

$$\sum_{k=0}^{\infty} x^T(k)x(k) < \gamma^2 \sum_{k=0}^{\infty} \left(w^T(k)w(k) + (t_{k+1} - t_k)\left(e_x^T(k)e_x(k) + e_f^T(k)e_f(k) + e_a^T(k)e_a(k)\right)\right). \tag{48}$$

The Lyapunov-Krasovskii function is constructed as

$$V(k) = \sum_{i=1}^{4} V_i(k). \tag{49}$$

Among them,

$$\begin{cases} V_1(k) = x^T(k)\tilde{P}x(k), \\ V_2(k) = \sum\limits_{i=k-\tau(k)}^{k-1} x^T(i)\tilde{Q}x(i), \\ V_3(k) = \sum\limits_{i=k-\tau_M}^{k-1} x^T(i)\tilde{R}x(i), \\ V_4(k) = \tau_M \sum\limits_{i=-\tau_M}^{-1} \sum\limits_{j=k+i}^{k-1} \xi^T(j)\tilde{S}\xi(j), \\ \xi(k) \triangleq x(k+1) - x(k). \end{cases} \tag{50}$$

For $k \in [t_k, t_{k+1})$, define $\Delta V(k) = \sum_{i=1}^{4} \Delta V_i(k)$, where

$$\Delta V_1(k) = x^T(k+1)\tilde{P}x(k+1) - x^T(k)\tilde{P}x(k), \tag{51}$$

$$\begin{aligned} \Delta V_2(k) &= \sum_{i=k+1-\tau(k+1)}^{k} x^T(i)\tilde{Q}x(i) - \sum_{i=k-\tau(k)}^{k-1} x^T(i)\tilde{Q}x(i) \\ &= x^T(k)\tilde{Q}x(k) - x^T(k-\tau(k))\tilde{Q}x(k-\tau(k)), \end{aligned} \tag{52}$$

$$\begin{aligned}\Delta V_3(k) &= \sum_{i=k+1-\tau_M}^{k} x^T(i)\tilde{R}x(i) - \sum_{i=k-\tau_M}^{k-1} x^T(i)\tilde{R}x(i) \\ &= x^T(k)\tilde{R}x(k) - x^T(k-\tau_M)\tilde{R}x(k-\tau_M),\end{aligned} \tag{53}$$

$$\begin{aligned}\Delta V_4(k) &= \tau_M \sum_{i=-\tau_M}^{-1}\sum_{j=k+1+i}^{k} \xi^T(j)\tilde{S}\xi(j) - \tau_M \sum_{i=-\tau_M}^{-1}\sum_{j=k+i}^{k-1} \xi^T(j)\tilde{S}\xi(j) \\ &= \tau_M \sum_{i=-\tau_M}^{-1}\left(\sum_{j=k+1+i}^{k} \xi^T(j)\tilde{S}\xi(j) - \sum_{j=k+i}^{k-1} \xi^T(j)\tilde{S}\xi(j)\right) \\ &= \tau_M \sum_{i=-\tau_M}^{-1}\left(\xi^T(k)\tilde{S}\xi(k) - \xi^T(k+i)\tilde{S}\xi(k+i)\right) \\ &= \tau_M^2 \xi^T(k)\tilde{S}\xi(k) - \tau_M \sum_{j=k-\tau_M}^{k-1} \xi^T(j)\tilde{S}\xi(j).\end{aligned} \tag{54}$$

According to Wirtinger's inequality, it can be concluded that

$$\begin{aligned}&-\sum_{j=k-\tau_M}^{k-1} \tau_M[x(j+1)-x(j)]^T\tilde{S}[x(j+1)-x(j)] \\ &\le [x(k)-x(k-\tau_M)]^T\tilde{S}[x(k)-x(k-\tau_M)] \\ &\quad -3\left[x(k)+x(k-\tau_M)-\frac{2}{\tau_M+1}\sum_{j=k-\tau_M}^{k} x(j)\right]^T\tilde{S}\left[x(k)\right. \\ &\quad \left.+x(k-\tau_M)-\frac{2}{\tau_M+1}\sum_{j=k-\tau_M}^{k} x(j)\right].\end{aligned} \tag{55}$$

The event triggers the condition, when $k \in [t_k + \tau_{t_k}, t_{k+1} + \tau_{t_{k+1}})$, there is

$$e_l^T(k)\Phi e_l(k) \le \sigma x\wedge^T(t_k)\Phi\hat{x}(t_k). \tag{56}$$

Among them, $e_l(k) = \hat{x}(k) - \hat{x}(t_k)$ is the state estimation error.

Note $\varphi^T(k) = [x^T(k)\, x^T(t_k)\, x^T(k-\tau_M)\, \psi^T(k)\, e_x^T(t_k)\, e_f^T(t_k)\, e_a^T(t_k)\, w(k)\, x\wedge^T(t_k)\, e_l^T(k)]$.

Among them, $\psi(k) = 1/\tau_M + 1\sum_{j=k-\tau_M}^{k} x(j)$ can be obtained by combining formulas (51)–(55) and (45):

$$\begin{aligned}&\Delta V(k) + x^T(k)x(k) - \gamma^2\left[w^T(k)w(k) + \sum_{k=0}^{\infty}(t_{k+1}-t_k)\right. \\ &\left.\cdot\left(e_x^T(t_k)e_x(t_k) + e_f^T(t_k)e_f(t_k) + e_a^T(t_k)e_a(t_k)\right)\right] \\ &+ \sigma x\wedge^T(t_k)\Phi\hat{x}(t_k) - e_l^T(k)\Phi e_l(k) \le \varphi^T(k)\Theta\varphi(k).\end{aligned} \tag{57}$$

Among them,

$$\Theta = \begin{bmatrix} \Theta_{11} & * & * & * & * & * & * & * & * & * \\ \Theta_{21} & \Theta_{22} & * & * & * & * & * & * & * & * \\ 2\tilde{S} & 0 & -4\tilde{S} & * & * & * & * & * & * & * \\ 6\tilde{S} & 0 & 6\tilde{S} & -12\tilde{S} & * & * & * & * & * & * \\ \Theta_{51} & \Theta_{52} & \Theta_{53} & \Theta_{54} & \Theta_{55} & * & * & * & * & * \\ \Theta_{61} & \Theta_{62} & \Theta_{63} & \Theta_{64} & \Theta_{65} & \Theta_{66} & * & * & * & * \\ \Theta_{71} & \Theta_{72} & \Theta_{73} & \Theta_{74} & \Theta_{75} & \Theta_{76} & \Theta_{77} & * & * & * \\ \Theta_{81} & \Theta_{82} & \Theta_{83} & \Theta_{84} & \Theta_{85} & \Theta_{86} & \Theta_{87} & \Theta_{88} & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma\Phi & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\Phi \end{bmatrix},$$

$$\begin{aligned}
&\Theta_{11} = A^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)A - \tilde{P} + \tilde{Q} + \tilde{R} + \tau_M^2\tilde{S} - \tau_M^2A^T\tilde{S} - \tau_M^2\tilde{S}A - 4\tilde{S} + I, \\
&\Theta_{21} = -(BK)^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)A + \tau_M^2(BK)^T\tilde{S} - 2\tilde{S}, \\
&\Theta_{51} = -(BK)^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)A + \tau_M^2(BK)^T\tilde{S}, \\
&\Theta_{61} = -E^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)A + \tau_M^2E^T\tilde{S}, \\
&\Theta_{71} = -N_a^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)A + \tau_M^2N_a^T\tilde{S}, \\
&\Theta_{71} = D_1^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)A - \tau_M^2D_1^T\tilde{S}, \\
&\Theta_{22} = (BK)^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)BK - \tilde{Q}, \\
&\Theta_{52} = (BK)^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)BK, \Theta_{62} = E^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)BK, \\
&\Theta_{72} = N_a^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)BK, \Theta_{82} = D_1^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)BK, \\
&\Theta_{53} = \Theta_{63} = \Theta_{73} = \Theta_{83} = 0, \Theta_{54} = \Theta_{64} = \Theta_{74} = \Theta_{84} = 0, \\
&\Theta_{55} = (BK)^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)BK - \gamma^2I, \\
&\Theta_{65} = E^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)BK, \Theta_{75} = N_a^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)BK, \\
&\Theta_{85} = -D_1^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)BK, \Theta_{66} = E^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)E - \gamma^2I, \\
&\Theta_{76} = N_a^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)E, \Theta_{86} = -D_1^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)E, \\
&\Theta_{77} = N_a^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)N_a - \gamma^2I, \Theta_{87} = -D_1^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)N_a, \\
&\Theta_{88} = -D_1^T\left(\tilde{P}+\tau_M^2\tilde{S}\right)D_1 - \gamma^2I,
\end{aligned} \tag{58}$$

supposing that $\Delta V(k) + x^T(k)x(k) - \gamma^2[w^T(k)w(k) + \sum_{k=0}^{\infty}(t_{k+1}-t_k)(e_x^T(t_k)e_x(t_k) + e_f^T(t_k)e_f(t_k) + e_a^T(t_k)e_a(t_k))] < 0, \Theta$

$< 0$. By using Lemma 1 twice, we can get the following results:

$$\begin{bmatrix} X & * & * & * & * & * & * & * & * & * & * & * \\ Y & -\tilde{Q} & * & * & * & * & * & * & * & * & * & * \\ 2\tilde{S} & 0 & -4\tilde{S} & * & * & * & * & * & * & * & * & * \\ 6\tilde{S} & 0 & 6\tilde{S} & -12\tilde{S} & * & * & * & * & * & * & * & * \\ \tau_M^2(BK)^T\tilde{S} & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * & * & * \\ \tau_M^2 E^T\tilde{S} & 0 & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * & * \\ \tau_M^2 N_a^T\tilde{S} & 0 & 0 & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * \\ \tau_M^2 D_1^T\tilde{S} & 0 & 0 & 0 & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma\boldsymbol{\Phi} & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\boldsymbol{\Phi} & * & * \\ \tau_M\tilde{S}A & -\tau_M\tilde{S}BK & 0 & 0 & -\tau_M\tilde{S}BK & -\tau_M\tilde{S}E & -\tau_M\tilde{S}N_a & \tau_M\tilde{S}D_1 & 0 & 0 & -\tilde{S} & * \\ \tilde{P}A & -\tilde{P}BK & 0 & 0 & -\tilde{P}BK & -\tilde{P}E & -\tilde{P}N_a & \tilde{P}D_1 & 0 & 0 & 0 & -\tilde{P} \end{bmatrix} < 0. \quad (59)$$

Among them,

$$\begin{cases} Y = \tau_M^2(BK)^T\tilde{S} - 2\tilde{S}, \\ X = -\tilde{P} + \tilde{Q} + \tilde{R} + \left(\tau_M^2 - 4\right)\tilde{S} - \tau_M^2 A^T\tilde{S} - \tau_M^2\tilde{S}A + I. \end{cases} \quad (60)$$

If (59) is multiplied by diag $\{\tilde{S}^{-1}, \tilde{S}^{-1}, \tilde{S}^{-1}, \tilde{S}^{-1}, \tilde{S}^{-1}, \tilde{S}^{-1}, \tilde{S}^{-1}, I, I, I, \tilde{P}^{-1}, \tilde{P}^{-1}\}$ on the left and the right, then Lemma 1 is applied again to the inequality obtained. If $K_1 = K\tilde{S}^{-1}$, $S_1 = \tilde{S}^{-1}$, $P_1 = \tilde{S}^{-1}\tilde{P}\tilde{S}^{-1}$, $Q_1 = \tilde{S}^{-1}\tilde{Q}\tilde{S}^{-1}$, $R_1 = \tilde{S}^{-1}\tilde{R}\tilde{S}^{-1}$ is used at the same time, then

$$\begin{bmatrix} X_1 & * & * & * & * & * & * & * & * & * & * & * & * \\ Y_1 & -Q_1 & * & * & * & * & * & * & * & * & * & * & * \\ 2S_1 & 0 & -4S_1 & * & * & * & * & * & * & * & * & * & * \\ 6S_1 & 0 & 6S_1 & -12S_1 & * & * & * & * & * & * & * & * & * \\ \tau_M^2(BK)^T & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * & * & * & * \\ \tau_M^2 E^T & 0 & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * & * & * \\ \tau_M^2 N_a^T & 0 & 0 & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * & * \\ \tau_M^2 D_1^T & 0 & 0 & 0 & 0 & 0 & 0 & -\gamma^2 I & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma\boldsymbol{\Phi} & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\boldsymbol{\Phi} & * & * & * \\ \tau_M AS_1 & -\tau_M BK_1 & 0 & 0 & -\tau_M BK_1 & -\tau_M E & -\tau_M N_a & \tau_M D_1 & 0 & 0 & -S_1 & * & * \\ AS_1 & -BK_1 & 0 & 0 & -BK_1 & -E & -N_a & D_1 & 0 & 0 & 0 & -\tilde{P}^{-1} & * \\ S_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -I \end{bmatrix} < 0. \quad (61)$$

Among them,

$$\begin{cases} Y_1 = \tau_M^2 K_1^T B^T - 2S_1, \\ X_1 = -P_1 + Q_1 + R_1 + \left(\tau_M^2 - 4\right)S_1 - \tau_M^2 S_1 A^T - \tau_M^2 A S_1. \end{cases} \tag{62}$$

Similarly, for symmetric positive definite matrix $\tilde{P}, \tilde{S}$, we have

$$\tilde{S}\tilde{P}^{-1}\tilde{S} - 2\tilde{S} + \tilde{P} = \left(\tilde{S} - \tilde{P}\right)\tilde{P}^{-1}\left(\tilde{S} - \tilde{P}\right) \geq 0. \tag{63}$$

Further, we can get

$$-\tilde{P}^{-1} \leq -2S_1 + P_1. \tag{64}$$

Combining formulas (61) and (64), formula (46) holds. If (46) is true, then

$$\Delta V(k) + x^T(k)x(k) - \gamma^2 \left[ w^T(k)w(k) + \sum_{k=0}^{\infty}(t_{k+1} - t_k) \cdot \left(e_x^T(t_k)e_x(t_k) + e_f^T(t_k)e_f(t_k) + e_a^T(t_k)e_a(t_k)\right)\right] < 0. \tag{65}$$

We can go further:

$$\begin{aligned} &\sum_{k=0}^{\infty}\left[x^T(k)x(k) - \gamma^2\left(w^T(k)w(k) + (t_{k+1} - t_k)\left(e_x^T(t_k)e_x(t_k) + e_f^T(t_k)e_f(t_k) + e_a^T(t_k)e_a(t_k)\right)\right)\right] \\ &< -\sum_{k=0}^{\infty}\Delta V(k) = -V(k) \leq 0. \end{aligned} \tag{66}$$

Obviously,

$$\sum_{k=0}^{\infty} x^T(k)x(k) < \gamma^2 \sum_{k=0}^{\infty}\left(w^T(k)w(k) + (t_{k+1} - t_k)\left(e_x^T(t_k)e_x(t_k) + e_f^T(t_k)e_f(t_k) + e_a^T(t_k)e_a(t_k)\right)\right). \tag{67}$$

Namely,

$$\|x(t)\|_2 < \gamma_{\min}\left(\|w(t)\|_2 + (t_{k+1}\text{-}t_k)\left(\|e_x(t_k)\|_2 + \left\|e_f(t_k)\right\|_2 + \|e_a(t_k)\|_2\right)\right). \tag{68}$$

In other words, when equation (46) holds, sliding mode (45) of the closed-loop system is asymptotically stable and has $H_\infty$ disturbance rejection performance.

Theorem 8 is proved.

## 4. Result and Discussion

In this paper, the control system model given in Reference [20] is taken as the research object:

$$\begin{aligned} A &= \begin{bmatrix} 0.9879 & 0.0098 \\ -0.0837 & 0.9908 \end{bmatrix}, B = \begin{bmatrix} -0.0029 & -0.0005 \\ -0.1919 & -0.0378 \end{bmatrix}, \\ E &= \begin{bmatrix} -0.0029 \\ -0.1919 \end{bmatrix}, N_a = \begin{bmatrix} -0.1341 \\ -0.1243 \end{bmatrix}, N_s = \begin{bmatrix} -0.0051 \\ -0.0078 \end{bmatrix}, \\ D_w &= \begin{bmatrix} 0.1 \\ 1 \end{bmatrix}, D_v = \begin{bmatrix} 0.1 \\ 1 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned} \tag{69}$$

In this simulation, we make the following assumptions about actuator failure and network attack.

Actuator failure:

$$f(k) = 2\sin 0.01\pi k + 2, \quad 230 \leq k \leq 500. \tag{70}$$

Network attack:

$$\begin{cases} a_a(k) = 2, & 100 \leq k \leq 500, \\ a_s(k) = -2, & 100 \leq k \leq 500. \end{cases} \tag{71}$$

The expanded coefficient matrices are as follows:

$$A_{11} = \begin{bmatrix} 0.9879 & 0.0098 & -0.0029 & 1 \\ -0.0837 & 0.9908 & -0.1919 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$B_{11} = \begin{bmatrix} -0.0029 & -0.0005 \\ -0.1919 & -0.0378 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, E_{11} = \begin{bmatrix} -0.0029 \\ -0.1919 \\ 0 \\ 0 \end{bmatrix},$$

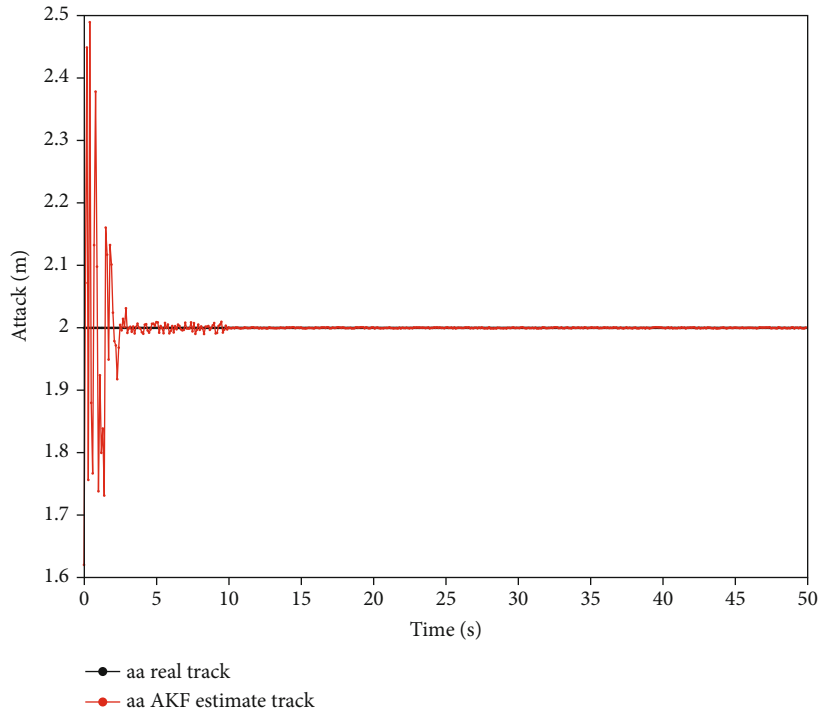



Figure 2: Actuator attack and its estimation for active fault-tolerant/active passive intrusion-tolerant systems.

$$D_{w11} = \text{diag}\{D_w, I, I\} = \begin{bmatrix} 0.1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$C_{11} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, D_{v11} = D_v = \begin{bmatrix} 0.1 \\ 1 \end{bmatrix}. \quad (72)$$

Let the initial condition of the system be $P(0 \mid 0) = I_4$, $\lambda = 0.9$, $x(0) = \hat{x}(0) = [-30 \quad 20 \quad 2 \quad -2]^T$, $R = 0.0025$, $r(k) = 0_{2\times 1}$, $Q = 0.0001$.

From equation (26),

$$K_1 = \begin{bmatrix} 1.1555 & -0.1168 \\ 1.6243 & -0.1598 \\ 0.4433 & -0.0447 \\ -0.0156 & -0.0018 \end{bmatrix}. \quad (73)$$

In Theorem 6, let the maximum delay upper bound $\tau_M = 1.2$, trigger parameter $\sigma = 0.85$, take $q = 5, \varepsilon = 2, T = 0.1$. Based on the sliding mode trigger matrix, we can get the event gain matrix $G$ and $\boldsymbol{\Phi}$, respectively,

$$G = \begin{bmatrix} -5.6995 & 1.0126 \\ -0.656 & 0.8053 \end{bmatrix}, \boldsymbol{\Phi} = \begin{bmatrix} 1.5812 & 0.2811 \\ -0.2657 & 2.6766 \end{bmatrix}. \quad (74)$$

The state feedback gain matrices are

$$K_{c1} = \begin{bmatrix} 63.72 & -48 \\ 23.28 & 23 \end{bmatrix}, K_{c2} = \begin{bmatrix} -273.6 & -998.25 \\ 832.23 & 389.58 \end{bmatrix}. \quad (75)$$

The simulation results are shown in Figures 2–11. Figures 2 and 3 show the attack and its estimation curve and attack estimation error diagram, respectively. Figures 4 and 5 show the actuator fault and its estimation curve and error diagram, respectively. Figures 6–9 show the state of the system and its estimation curve and state estimation curve, respectively. Figure 10 shows the output response curve of the system when actuator failure and network attack occur simultaneously. Figure 11 shows the data transmission time and the transmission interval of the system nonuniform transmission NCS.

It can be seen from Figures 2–9 that the adaptive Kalman filter fault/attack estimator can well estimate the state, fault, and attack of the system. It can be seen from Figure 10 that when the system encounters actuator

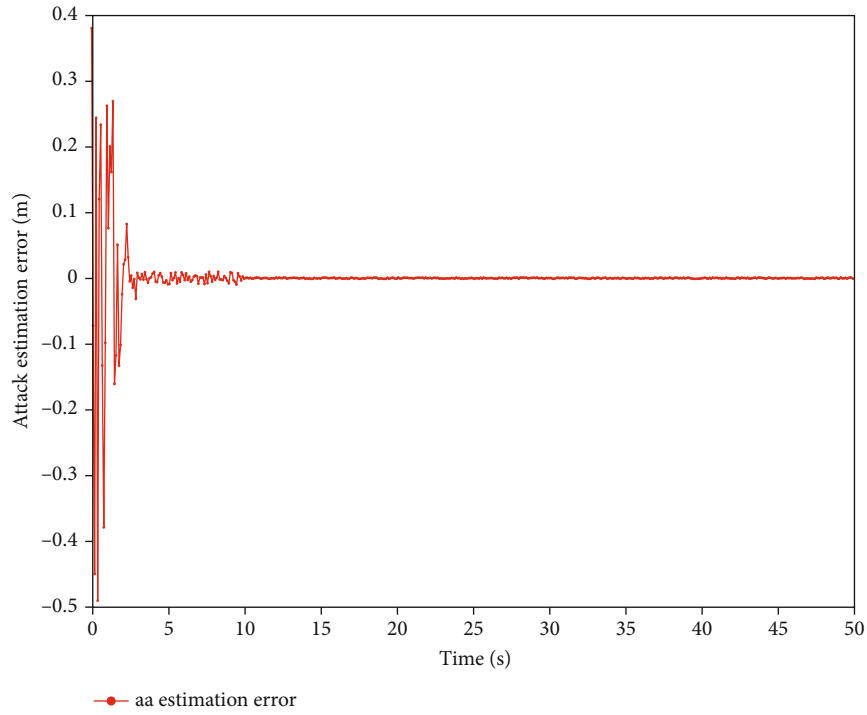



Figure 3: Actuator attack estimation error of active fault-tolerant/active passive intrusion-tolerant system.

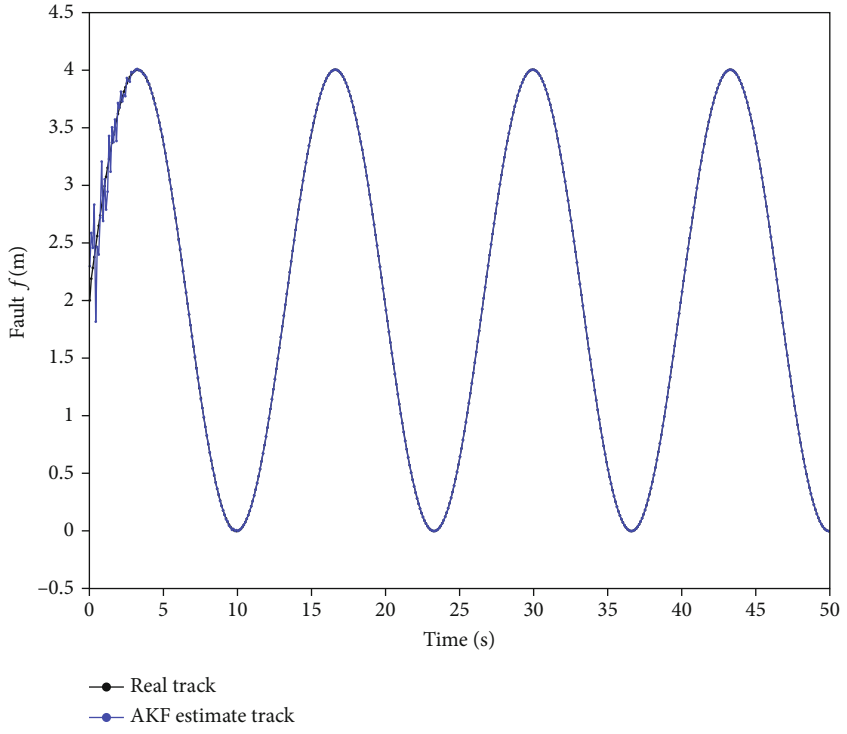



Figure 4: Fault estimation of active fault-tolerant/active passive intrusion-tolerant system.

failure and network attack, the vibration is gradually attenuated and tends to balance by using the cooperative controller designed in this paper and the compensation strategy of failure and attack. Simulation results show that the active fault-tolerant/active passive intrusion-tolerant cooperative controller designed in this paper is

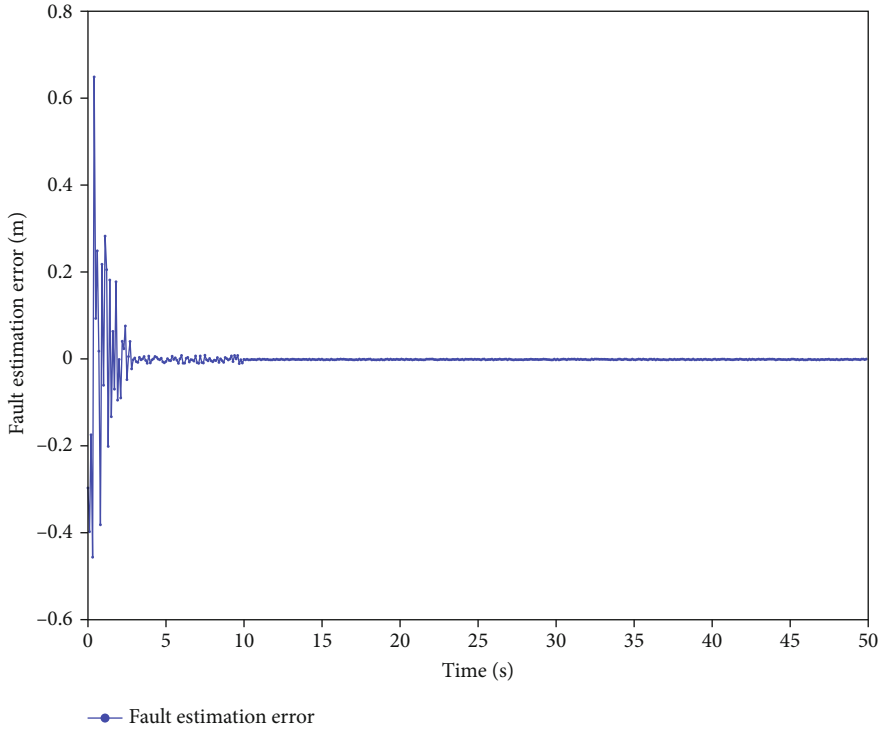



Figure 5: Fault estimation error of active fault-tolerant/active passive intrusion-tolerant system.

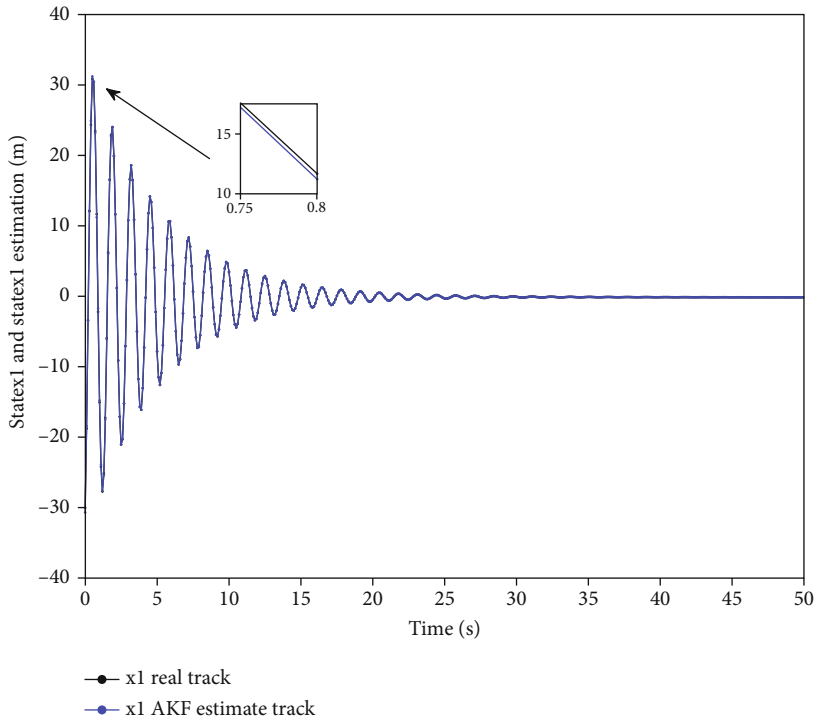



Figure 6: State 1 and estimation of active fault-tolerant/active passive intrusion-tolerant system.

effectively fault-tolerant and active and passive intrusion-tolerant for network attacks and also suppresses the influence of disturbance and noise.

According to the analysis in Figure 11, when the event trigger parameter is $\sigma = 0.85$, compared with the traditional PPTCS which needs to transmit 500 data in the

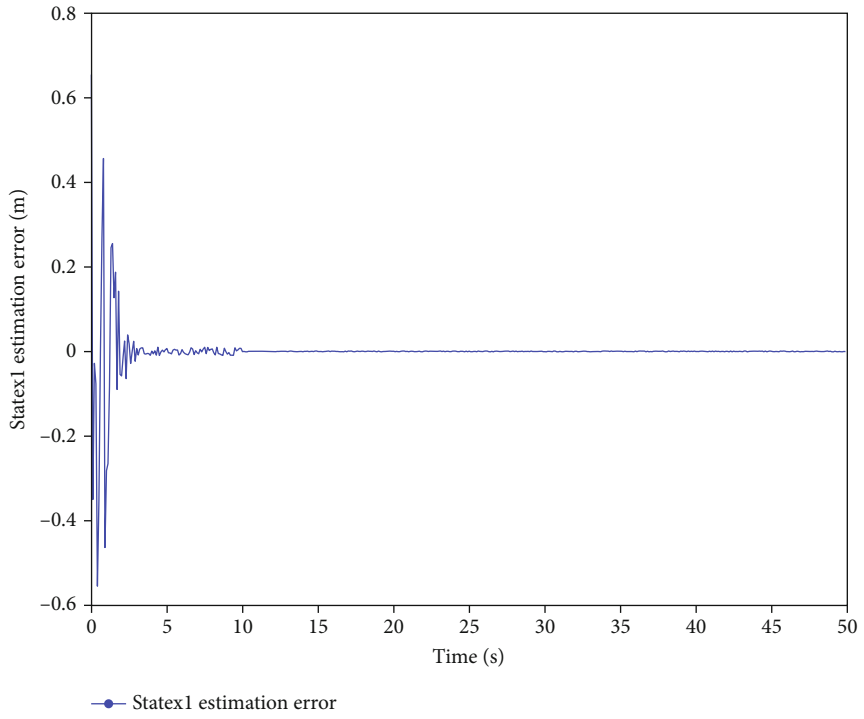



Figure 7: State 1 estimation error of active fault-tolerant/active passive intrusion-tolerant system.

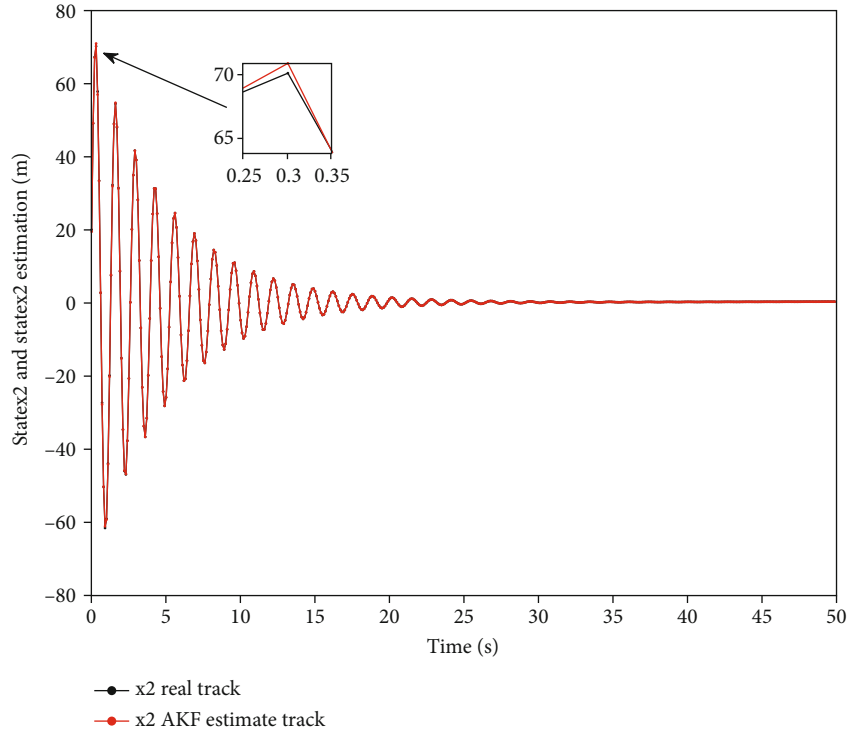



Figure 8: State 2 and estimation of active fault-tolerant/active passive intrusion-tolerant system.

simulation time of 50 s, 249 data are transmitted in the DETCS, the data transmission rate is 49.8%, the average transmission cycle is $\bar{T} = 0.2006$ s, and the maximum transmission cycle is $T_{\max} = 0.6$ s. This shows that the control method proposed in this paper not only ensures the excellent performance of the system but also effectively saves the network communication resources and improves the efficiency of network utilization.

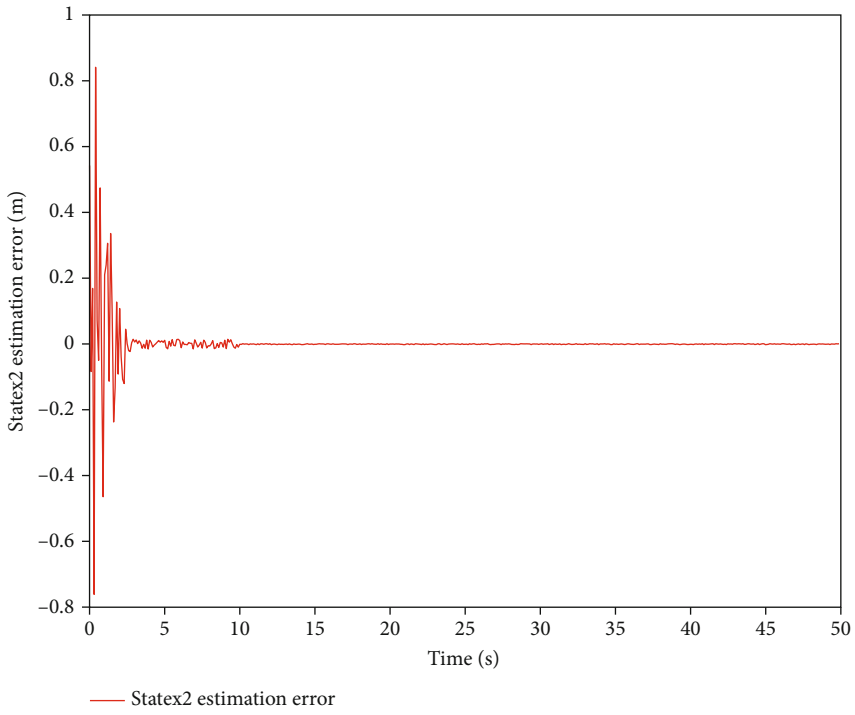



Figure 9: State 2 estimation error of active fault-tolerant/active passive intrusion-tolerant system.

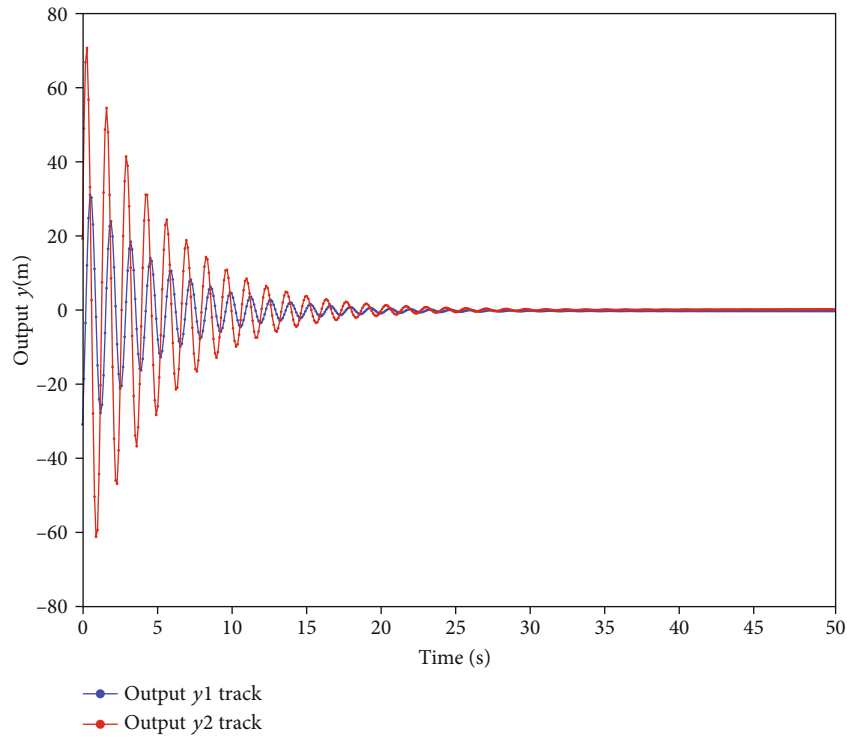



Figure 10: Output response of active fault-tolerant/active passive intrusion-tolerant system.

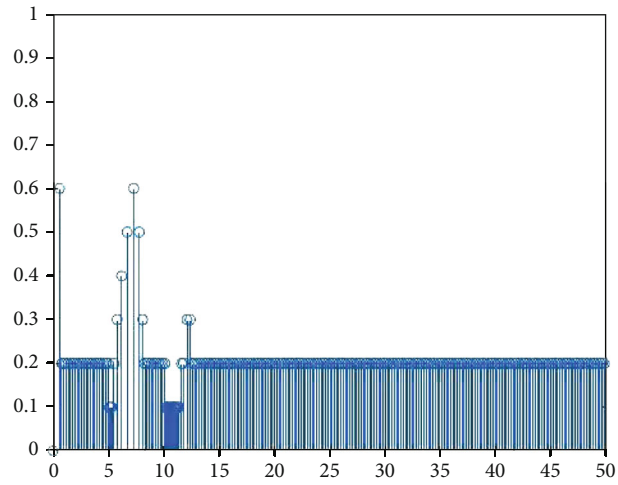



Figure 11: Transmission time and interval of nonuniform transmission.

## 5. Conclusion

In this paper, an active fault-tolerant/active passive intrusion-tolerant cooperative controller design method is proposed for linear discrete NCS with time-varying delay based on discrete event triggering mechanism when actuator failure and network attack occur. In this method, the attack is extended to the state, and the adaptive Kalman filter is used to estimate the state, fault, and attack. Then, the integral sliding mode control method is used to design the active fault-tolerant/active passive intrusion-tolerant cooperative controller, so that the system can keep normal operation and have $H_\infty$ disturbance rejection performance in the case of actuator failure and network attack. Finally, a simulation example is given to illustrate the effectiveness and applicability of the proposed method.

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

In this paper, the research was sponsored by the National Natural Science Foundation of China (Grant Nos. 61463030 and 61563031).

## References

[1] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—part I: fault diagnosis with model-based and signal-based approaches," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 6, pp. 3757–3767, 2015.

[2] D. N. Jiang, L. I. Wei, and J. Wang, "Robust-guaranteed fault-tolerant control of nonlinear networked control system based on T-S cloud model," *Journal of Lanzhou University of Technology*, 2012.

[3] M. Zhong, T. Xue, and S. X. Ding, "A survey on model-based fault diagnosis for linear discrete time-varying systems," *Neurocomputing*, vol. 306, pp. 51–60, 2018.

[4] A. R. Merheb, H. Noura, and F. Bateman, "Design of passive fault-tolerant controllers of a quadrotor based on sliding mode theory," *International Journal of Applied Mathematics and Computer Science*, vol. 25, no. 3, pp. 561–576, 2015.

[5] G.-H. Yang, J. L. Wang, and Y. C. Soh, "Reliable LQG control with sensor failures," *IEE Proceedings - Control Theory and Applications*, vol. 147, no. 4, pp. 433–439, 2000.

[6] X.-J. Li and G.-H. Yang, "Robust adaptive fault-tolerant control for uncertain linear systems with actuator failures," *IET Control Theory & Applications*, vol. 6, no. 10, pp. 1544–1551, 2012.

[7] X. Y. Cao, C. H. Hu, and Q. L. Ma, "Research on active fault-tolerant control for sensor failures of missile attitude control systems," *Control and Decision Making*, vol. 27, no. 3, pp. 379–382, 2012.

[8] B. Jiang, M. Staroswiecki, and V. Cocquempot, "Fault accommodation for nonlinear dynamic systems," *IEEE Transactions on Automatic Control*, vol. 51, no. 9, pp. 1578–1583, 2006.

[9] K. Zhang, B. Jiang, and P. Shi, "Observer-based integrated robust fault estimation and accommodation design for discrete-time systems," *International Journal of Control*, vol. 83, no. 6, pp. 1167–1181, 2010.

[10] G. Tao and S. M. Joshi, *Adaptive Control of Systems with Actuator Failures*, World Congress on Intelligent Control & Automation, 2008.

[11] R. Cao, J. Wu, C. Long, and S. Li, "Stability analysis for networked control systems under denial-of-service attacks," in *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 7476–7481, Osaka, Japan, 2015.

[12] X. Chen and Y. Wang, "Event-triggered attack-tolerant tracking control design for networked nonlinear control systems under DoS jamming attacks," *Science China Information Sciences*, vol. 63, no. 5, 2020.

[13] D. Ding, Q. L. Han, Y. Xiang, X. Ge, and X. M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[14] W. Li, Y. Shi, and Y. Li, "Research on secure control and communication for cyber-physical systems under cyber-attacks," *Transactions of the Institute of Measurement and Control*, vol. 41, no. 12, pp. 3421–3437, 2019.

[15] A. A. Yaseen and M. Bayart, "Cyber-attack detection in the networked control system with faulty plant," in *2017 25th Mediterranean Conference on Control and Automation (MED)*, pp. 980–985, Valletta, Malta, 2017.

[16] C. Peng and T. C. Yang, "Event-triggered communication and Hco control co-design for networked control systems," *Automatica*, vol. 49, no. 5, pp. 1326–1332, 2013.

[17] T. Q. Fang and Y. Wei, "Research on robust H8 fault-tolerant control for uncertain suspension system with time delay," in *2012 International Conference on Industrial Control and Electronics Engineering*, pp. 1994–1997, Xi'an, China, 2012.

[18] A. Seuret, F. Gouaisbaut, and E. Fridman, "Stability of discrete-time systems with time-varying delays via a novel summation inequality," *IEEE Transactions on Automatic Control*, vol. 60, no. 10, pp. 2740–2745, 2015.

[19] P. T. Nam, P. N. Pathirana, and H. Trinh, "Discrete Wirtinger-based inequality and its application," *Journal of the Franklin Institute*, vol. 352, no. 5, pp. 1893–1905, 2015.

[20] H. Dai, X. Luo, C. Bo, and S. Dai, "Integrated fault estimation and fault tolerant control for discrete systems based on linear matrix inequality," *Chinese Journal of Inertial Technology*, vol. 28, no. 1, p. 134, 2020.