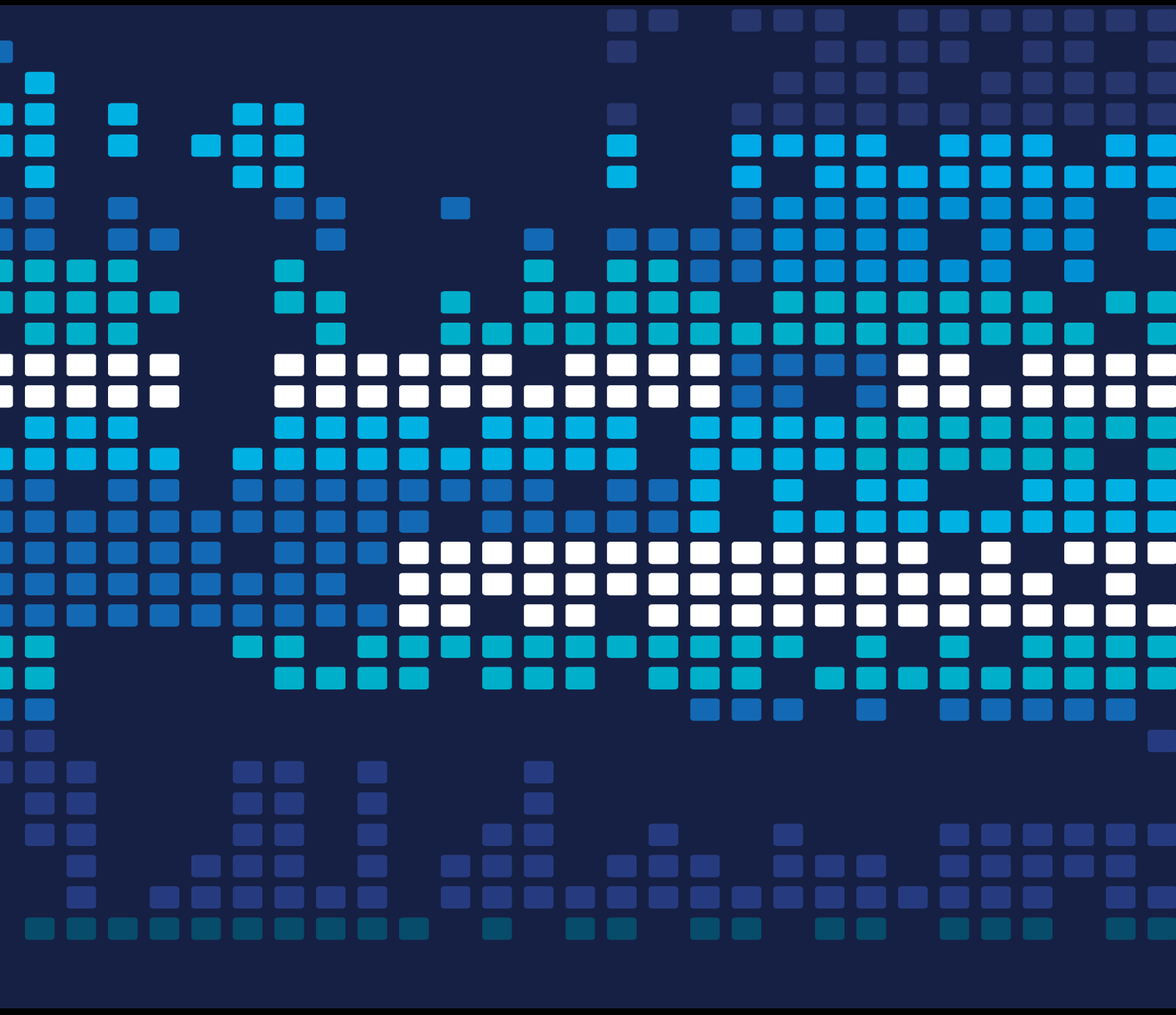
Scientific Programming

Big Data, Scientific Programming, and Industrial Internet of Things

Lead Guest Editor: Habib Ullah Khan
Guest Editors: Shah Nazir and Shaukat Ali



Big Data, Scientific Programming, and Industrial Internet of Things

Scientific Programming

Big Data, Scientific Programming, and Industrial Internet of Things

Lead Guest Editor: Habib Ullah Khan
Guest Editors: Shah Nazir and Shaukat Ali

Contents

Contents

Use of Big Data Tools and Industrial Internet of Things: An Overview
Yingzi Wang, Muhammad Nazir Jan, Sisi Chu, and Yue Zhu
Review Article (10 pages), Article ID 8810634, Volume 2020 (2020)

Evaluating the Role of Big Data in IIOT-Industrial Internet of Things for Executing Ranks Using the Analytic Network Process Approach
Xiaoqun Liao, Mohammad Faisal, Qing QingChang, and Amjad Ali
Research Article (7 pages), Article ID 8859454, Volume 2020 (2020)

Big Data, Scientific Programming, and Its Role in Internet of Industrial Things: A Decision Support System
Ju Li, Muhammad Nazir Jan, and Mohammad Faisal
Research Article (7 pages), Article ID 8850096, Volume 2020 (2020)

A New Scalable and Expandable Access Control Model for Distributed Database Systems in Data Security
Mehmet Guclu, Cigdem Bakir, and Veli Hakkoymaz
Research Article (10 pages), Article ID 8875069, Volume 2020 (2020)

Hindawi
Scientific Programming
Volume 2022, Article ID 8853448, 22 pages
https://doi.org/10.1155/2022/8853448

Research Article

Secure Smart Healthcare Monitoring in Industrial Internet of Things (IIoT) Ecosystem with Cosine Function Hybrid Chaotic Map Encryption

Jalaluddin Khan,[1,2] Ghufran Ahmad Khan,[3] Jian Ping Li,[2] Mohamed Fahad AlAjmi,[4] Amin Ul Haq,[2] Shakir Khan,[5] Naved Ahmad,[6] Shadma Parveen,[7] Mohammad Shahid,[8] Sultan Ahmad,[9] Mordecai Raji,[2] Bilal Ahamad,[10] Abdulrahman Abdullah Alghamdi,[10] and Amjad Ali[11]

[1]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh 522502, India*
[2]*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[3]*School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China*
[4]*Department of Pharmacognosy, College of Pharmacy, King Saud University, Riyadh 11451, Saudi Arabia*
[5]*College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh 11432, Saudi Arabia*
[6]*Head of Research Support Unit, AlMaarefa University, Riyadh 11597, Saudi Arabia*
[7]*School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, China*
[8]*Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan*
[9]*Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Alkharj 11942, Saudi Arabia*
[10]*College of Computing and Information Technology, Shaqra University, Shaqra 11961, Saudi Arabia*
[11]*Department of Computer Science and Software Technology, University of Swat, Saidu Sharif Swat 19200, Pakistan*

Correspondence should be addressed to Jalaluddin Khan; jalal4amu@yahoo.com, Jian Ping Li; jpli2222@uestc.edu.cn, and Shadma Parveen; yourshadma@yahoo.com

Received 10 August 2020; Accepted 28 January 2022; Published 29 March 2022

Academic Editor: Habib Ullah Khan

The technological progression is raised as a hybrid ecosystem with the industrial Internet of Things (IIoT). Among them, healthcare is also broadly unified with the Internet of Things to develop an industrial forthcoming system. Utilizing this type of system can be facilitating optimum patient monitoring, competent diagnosis, intensive care, and including the appropriate operation against the existing critical diseases. Due to enormous data theft or privacy leakage, security, and privacy towards patient-based informative data, the preservation of personal patients' informative data has now become a necessity in the digitized community. The produced article is underlined on handsomely monitoring, perceptively extracted keyframe, and further processed lightweight cosine functions using hybrid way chaotic map keyframe image encryption. Initially, a regular concept of extracted keyframe is deployed to salvage meaningful detected frames by transmitting an alert autonomously to the administration. Then, lightweight cosine function for encryption is employed. This encryption incorporates keyframe exceedingly secure and safe from the outside world or any adversary. Our proposed methodology validates effectiveness throughout the IIoT ecosystem. The produced outcome is highly acceptable and has minimum execution time, robustness, and reasonably adopted cost-effective, secure parameter than any other (keyframes) image encryption methods. Furthermore, this methodology has optimally reduced bandwidth, essential communicating price, transmission cost, storage, and immediately judicious analysis of each occurred activity from the outside world or any adversary to remain secure and confident about the real patient-based data in the smartly developed environment.

1. Introduction

Nowadays, the Internet of Things (IoT) has intensified its global omnipresence. The implementation of smart networks is supposed as an exposition of ubiquitous computing. The goal is to expand network edge enabling smart services with IoT system. These kinds of computing is the best way replacement of the upfront user intentness towards interconnected employed devices (sensor-enabled strategies). It deprived of human interaction and instigated industrial perspective such as industrial Internet of Things (IIoT) [1–8]. These applications are well-organized integral sensor-based IoT architecture which is behaving well-informed accosted systems, such as smart cities (SC), smart healthcare system (SHM), smart wireless-multimedia sensor network (SWMSN), smart homes (SH), remote monitoring on farm animals (RMFA), automobiles, drone monitoring system (DMS), smart industry surveillance system (SIS), agricultural crop monitoring system (ACMS), pain monitoring system (PMS), self-driving vehicles (SV), and smart transportation system (STS) [9–17]. These treasured research and innovation parturitions as a high processing skill set in terms of the intelligent operational IIoT ecology, which is well suited to meaningful communication through the environs.

WMSN (wireless-multimedia surveillance networks) are among the top innovative contributor. It is an essential parameter towards IIoT empowered ecology that operates visual sensor data uniformly. It is also measuring all the possible views and nonstop capturing visual images. These laboring methodologies are generating enormous multimedia visuals with too much redundancy into the system [18–22]. Due to availability of huge redundant visuals into the intelligent system, it is observed consensus of the researchers that developed a mechanism which can process meaningful as well as informative visuals from the surveillance networks. Add the best way recorded for forthcoming usages. It is tracing behavior or activities (simple interaction or abnormal), diagnosis liveliness of the doctors, operation theater activities, patients handling ventures of the staff as well as nurses, and observing hygienic facilities in the whole industrial healthcare setup. The approaching mentioned problems. The optimum emphasis is to capture every probable abnormal activity detection by accurate analysis with well-organized supervision and video abstraction. The foremost motive behind this, transferring all the visuals through the communication medium before processing is not the best way because it contains more bandwidth as well as energies. Additionally, it is very hard (difficult) along with time-consuming to recognize and ingeniously extracted action-oriented intelligence from a high volume of surveillance visuals [23, 24].

Therefore, it is essential to employ a piece of machinery or methodology that can accumulate each valued visual information individually incorporating high processing skills set and communication capabilities of the smart-enabled IIoT sensors. The quality of these methodologies is intelligently selecting accurate views from any different locations or multiple views and smoothly captured. The closely informative pursuits or core-specific data in real-time with the sensors, as a result, accurately transfer that visual data to the expert witness. Additionally, the key importance is to take action from the specialist by the investigation of original gigantic enlightening visuals. Then, a conventional methodology including efficient monitoring (surveillance) is shown in Figure 1. The enhanced conceptual framework is apprehending each perilous movement for accurate recognition of any possible happened actions in a quite shrewdly and reporting immediate to enforce action promptly and reduces any miss happening into the entire system. This approach also provides healthier achievement regarding disciplined resource utilization and robustness, which is the essential requirement when monitoring smart healthcare systems with the (WMSNs) wireless multimedia sensor networks by proper investigation to resolve any technological, nature-based, and human malfunctioning defects [25, 26].

Concerning the tendency of visual transmission at wireless environments as a WMSN from start to endpoint, vice versa, these communications are coming in a vulnerable way with the enormous security threads. Consequently, it is highly recommended that the visual data in terms of keyframe image can be securely transferred to a base station with enhanced and secure guided way without any modification from any untrusted parties. Additionally, we are guessing that the utilized dedicated communication of visual data supposed to problems due to jammed (congested) spectrum allocation methods into WMSN medium. Furthermore, the unfolded report is an emphasis on solving mentioned problems by incorporating an intelligent and power-efficient system. That can easily handle in better way by gathering informative data or relevant information in real-time and prompt action taken against the happened events. By employing this, our methodology can be proficient in reducing the cost of the transmission as well as consumption of congested bandwidths. In addition, adapted methodology has the primary intent to enrich the security prototype, in which system is fully taking care, protecting WMSNs, and improving utilization towards concern authority. Technically proposed system engendered encrypted keyframe images when transferring extracted keyframes to concern authority for accomplishing high rate of security during entire communication inside the smart healthcare IIoT enabled setup.

Core enrichment of the designed report is registered as follows:

(I) We proposed an effective architectural adjustment to monitor healthcare smartly with IIoT via utilizing cosine as a function targeted encryption based on chaotic hybrid map.

(II) The methodology emphasizes first to extract the most relevant or meaningful keyframes from the summarized video data into the keyframe extraction model.

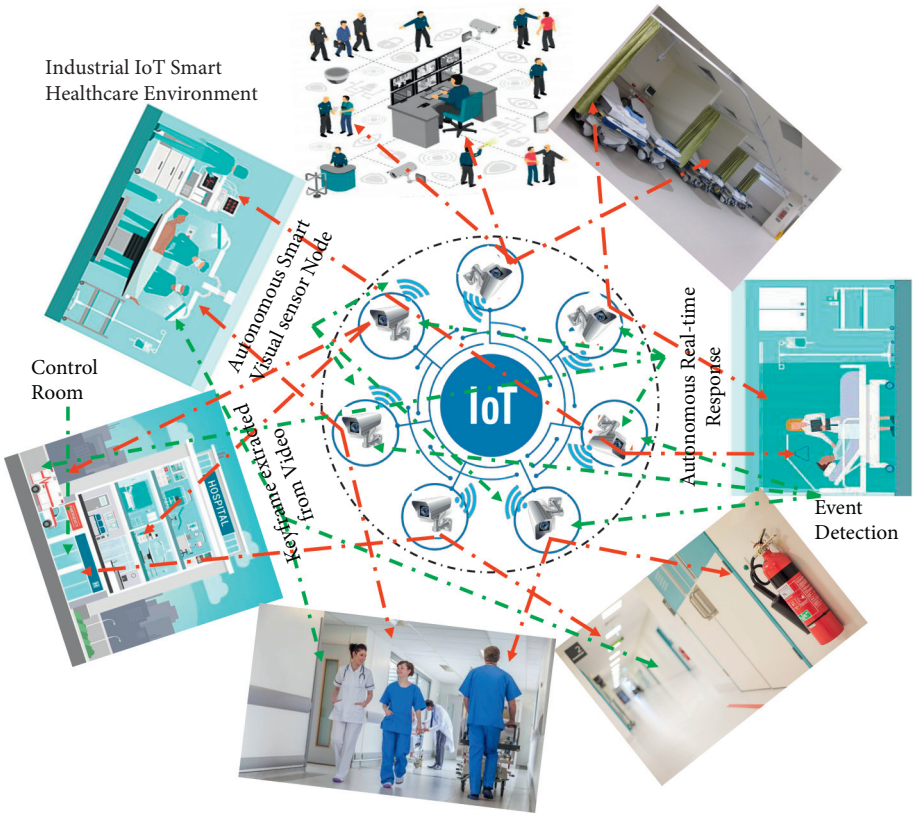


Figure 1: Secure monitoring within industrial IoT environment as a smart healthcare setup.

(III) After that, we incorporated a well-organized probabilistic as well as lightweight cosine function for encryption. This encryption approach is conquering strong security against any adversary.

(IV) The methodology is applied to hybrid technologies such as Python, TensorFlow, YOLOv3 for the extraction of keyframes, and cryptographic simulation done by MATLAB. The extracted keyframe evaluation and security analysis ensured that our methodology is proficient for reducing the cost of transmission as well as the efficiency of consumption bandwidth.

(V) The produced report endorses the commanding characteristics of the patient-based privacy in terms of an encrypted matrix to avoid any adversary outbreak.

(VI) The produced report also approved that the numerous rigorous security threads can withstand.

The rest of the article is summarized and pursues as follows: Section 2 has articulated preliminary work briefly about the smart healthcare system, monitoring and surveillance system, video summarization, and RGB image encryption. Section 3 has incorporated a novel planned monitoring (surveillance) tactic in which model extraction of keyframe from visual sensors is briefly discussed, and it describes implementation of lightweight cosine function hybrid chaotic map encryption methodology. Section 4 has investigated its experimental outcomes and relative discussion. Section 5 has briefly encountered numerous rigorous security examination with each possible security-based parameter. Section 6 is an efficiently concluding move toward entire work as well as a future direction.

2. Preliminary Work

Recent huge progressions in the territory of WMSNs are owned as a smart intelligence-based healthcare setup. It elevates finest incorporation at the hospital management with relating to patient privacy, security, and safety in the industrial perspective. Targeted mensuration about confidentiality, safety, and security, it is important to examine visual intelligence data and maternal encryption methodology. That can validate the genuineness of the system designed for complete patient protection or towards health safety. Although providing relief to human lives, the huge surveillance information poses the challenging task of spending time and energy while the footage is being picked up. Consequently, some methodology or approaches are required to facilitate as well as provide the most relevant data as a summaries video, including extraction of meaningful keyframe within that summaries video instead of viewing the whole surveillance visual data. The video summary received considerable study coverage over the past two decades. In a very concise way, it helps to abbreviate one long video or different images, such as video skimming as well as a static storyboard. Khan et al. reported on [27] an effective, co-evolutionary neural network-based summary procedure for

a resource-constrained system of the surveillance videos, in which the shot segmentation process is incorporated by utilizing deep features. That mechanism retains the interesting nature of the produced summary by utilizing image memorability as well as entropy properties. It also claimed that in each shot of the frame recorded highest score as a keyframe of the memorability as well as entropy. Hussain et al. [28] expressed his thought about MVS (multi-view video summarization), how it is challenging to accommodate gigantic volumes of data, light variation, redundancy, overlapping views as well as inter-view correlation as well. Their idea is to integrate MVS with a deep neural network as a two-tier method to adopt soft computing processes. Its first automated tier conducts segmentation of priority shots based on appearance and preserves them in a query table which is forwarded to all the cloud with further analysis. Its second tier captures deep properties from every frame of a series in the indexed table as well as transfers those to deep, long-term bidirectional memory to gain insightful odds and providing a summary.

Huang and Wang [29] explained the popularity of video summarization and its key point selection of the keyframes that can represent actual content as a video sequence. Their idea is to provide video summarization as well as motion of the video summarization. For achieving this, initially the capsule networks are accurately trained, such as extractor of spatiotemporal details, as well as focused on such spatiotemporal properties as an interframe movement curve is produced. A (TED) transition effect detection system is subsequently proposed to segment the video streams automatically into shots. At last, a self-attention framework is implemented for selecting keyframe sequences within videos. Therefore, selecting key static images as a summary of streaming content as well as measuring visual flows as a summary of video motion. Jei et al. [30] proposed an action-driven video synthesis paradigm focused on reinforcing learning. The framework is divided mainly into two sections: video cut through action parsing and video description focused on reinforcement learning. Within the first section, a chronological multi-instance learning framework is equipped using weekly interpreted data to crack the issue of the time-consuming maximum annotation as well as the uncertainty of the weak annotation. Throughout the second section, it built a deep recurrent video summarization model constructed on the neural network that selects the most recognizable frames compared with other behavior. In the meantime, the consistency of the mainframes extracted may be assessed by the correctness of the categorization. Yuan et al. [31] introduced a new (DSSE) deep side semantic embedding prototype to produce video summaries using the free side details. The DSSE establishes an implicit subspace through correlating the two unimodal autoencoders' hidden layers. Respectively, that embedded the side information and video frames. Furthermore, by dynamically reducing the loss of semantic meaning and the loss of the two unimodal autoencoders in function reconstruction, the comparable general facts between side information and video frames can be more thoroughly understood. Hence, its semantic significance can be evaluated quite efficiently. Ultimately, semantically relevant portions are picked from videos by eliminating its lengths to the side details and the latent subspace is constructed.

The cryptographic encryption of the digital images can be extensively recognized into two major groups. One of them is color image encryption, and another one is grayscale image encryption. The encryption methodology of a grayscale image is only based on one plane. But it can be prolonged by incorporating different color planes like red, green, and blue to adopt completely color image encryption procedures. Digital security is a vital problem limiting IIoT applications as well as cyber-physical systems' wider acceptance. In these regards, focusing various cryptographic concepts earlier, Wang et al. [32] expressed their methods to incorporate quantized logistic map-based stream encryption among the (WBAN) wireless body area networks. WBANs are committed to the transmission and processing of human-grown biomedical information. This encryption mechanism uses a chaotic method to execute the process for encryption. The assessment results demonstrate that the suggested encryption framework seems to have the advantages of success with high-security defense. Al-Khedhairi et al. [33] reported their hybrid cryptosystem for achieving three goals. One of which is to implement a new (2D) two-dimensional fractional-order map with very dynamic chaotic behavior as well as a unique true value of Lyapunov exponents across a wide variety of variables, particularly in comparison to other 2D maps. Secondly, for the first time, it is suggested a new robust, stable encryption system integrating the related chaotic pseudo-orbits of the suggested map with those of the benefits of elliptic curves through public-key cryptography while implemented to color images. At lastly, the hybrid structure is confident of verifying secure exchange of hidden keys as well as extremely obscure and concealing information communications. Zhang et al. [34] described an effective S-box, secure hash algorithm MD2, and fractional-order logistic map-based image encryption. Furthermore, the authors emphasize that the proposed algorithm is focused on a logistic fractional-order map which had benefits in their improved potential to withstand specific cryptanalyst attacks with reduced execution times. Huang et al. [35] articulated about the procedure for the color-based image encryption wherein permutation-diffusion happened concurrently. In addition, it is cooperating strangely among the color image matrixes. Accordingly, this approach is added security characteristics through high competence.

Almalkawi et al. [36] incorporated a lightweight image encryption method in which a hybrid kind of chaotic maps (logistic and henon) are used to encrypt the image matrix. His idea is to split digital images into blocks and effectively compress those blocks for reducing the size of the image matrix. After that then applied a logistic map for the creation of the key. The permutation and substitution operations are performed to attain shuffling and transportation of the image pixels. The henon chaotic map is implemented to adjust the pixel positions throughout the diffusion process to improve the necessary level of protection and to resist different security threats. Chai et al. [37] explained an RGB

image crypto-method founded on chaos as well as DNA oriented encryption. Firstly, his idea is to crumble color image as a red, green, blue channel. After that, intra-inter component permutation methods are reliant on a plain image which is functioned to scuffle them. Additionally, transformed and rejoined permuted classification in DNA matrix. Lastly, ornamental security employed a twice confusion process through scrambling images in receipt of encrypted images. They were creating pseudorandom chaotic orders, and the author accumulated a four-wing hyperchaotic process in the methodology. Wang and Li [38] proposed an innovative amalgamated chaotic color image encryption by using logistic and tent maps for preliminary key scanning. This method utilizes a chaotic hybrid map together via a structured logistic map and tent map. Afterward, it receives the necessary specification for Arnold mapping via a functionality transformation required to scramble that image matrix. The diffusion process utilized a chaotic neural network via Hopfield to produce a chaotic sequence of self-diffusion. At that time, the key is also created by a parameter reconstruction. Eventually, the mashed image is operated XOR with a key to acquire the penultimate encrypted cipher image. Hamza et al. [39] expressed an effective predictive cryptosystem to preserve the confidentiality of keyframes and the privacy of the patients. An innovative PRNG (pseudorandom number sequence) focused on chaos, concentrating on merging as well as cascading the positions of two of those same chaotic maps. By attaining safe and privacy-conserving communication, the medical professionals diagnostically enforced keyframes through limited bandwidth of energies and communication. Investigational testing from various viewpoints is done to ensure a high level of protection with more effective implementation relative to current methods. Hua et al. [40] proposed a CTBC (cosine-transform-based chaotic) system, in which two chaotic seed maps are used as a chaotic hybrid map to produce dynamic and a complex pattern of sequences. The encryption mechanism takes high-efficiency scrambling to isolate neighboring pixels and operates arbitrary pattern substitution to disperse a very small shift in the images towards all pixels of cryptographic matrixes. Additionally, the summative assessment shows that although the chaotic maps created by the CTBC demonstrate far more complex chaotic responses than the currently available. The experiment result shows how reliable the suggested image encryption framework is.

Hamza et al. [41] incorporated a safe system for video summarization of exterior patients under WCE protocol. In this method, keyframes are derived using only a lightweighted video description framework to make it fully WCE-fit. After that, a cryptosystem focused on Zaslavsky chaotic map. It is provided for the protection of derived keyframes. Observational findings confirmed the suggested cryptosystem's efficiency with respect to reliability and high-level protection relative to many other modern image cryptographic algorithms. During most of the distribution of essential keyframes to customized WCE healthcare providers and gastroenterologists, Hamza et al. [42] reported an effective cryptosystem for IoT-based monitoring systems. Firstly, a lightweight automated summary methodology is utilized. To retrieve the keyframes throughout the surveillance footage, based on a simple histogram-clustering method, then, it compressed the synthesized data size using a discrete cosine transform (DCT) methodology. Eventually, the implemented architecture executes effective image encryption with the introduction of discrete random fractional transformation (DFRT). Kaur et al. [43] incorporated pseudodominated genetic algorithm processing and specific chaotic image encryption techniques. It is introduced to change the 5D chaotic map hyperparameters. Firstly, the incoming image is broken down into thread bands that used a dual-tree dynamic wavelet transform (DTCWT) to perform the TFCM. Therefore, these thread bands are diffused, and the hidden key is used from just the engineered 5D-chaotic diagram. An inverse DTCWT is gradually enforced to get the ultimate encrypted image. Broumandnia [44] implemented 3D modular chaotic mapping that enhanced key size and acceleration for encryption of RGB images. To extend key space color picture encryption with modular arithmetic, he provided reasonable success with the study of the histograms as well as a correlation of the adjacency pixels. This method is used the confusion and diffusion function with the image encryption measures implemented in substitution and permutation.

Mondal et al. [45] introduced a highly efficient image encryption system for safe interaction and preservation of images. The framework is concentrated on a map of the chaotic skewed tent as well as cellular automatons (CA). The synthesis for both chaotic maps with CA respectively provides a model which a larger key size as well as quicker generator PRNS. The results of the experiment indicate good effects of encryption that can also withstand every kind of identified attack. The proposed concepts followed an additional parametric observation presented with the distinguished modification, which reflects a relatively faster encryption methodology at [46].

3. Proposed Work

The mounting standard of autonomous monitoring is highly progressed manufacturing-based visual sensor and technological progression of IIoT. It is bringing effective management tools that can make accurate as well as easy analysis of the digital world. It can also perform constant adoption of cumulative monitoring networks. This can be well-known as a smart setup of healthcare. These technological ecologies are empowered to be analyzed autonomously in an intelligent way in real-time of visual data (video source). VSN (visual sensor networks) have the capability to interact smartly and agreeing to conduct critical as well as very complex visual data processing in real-time through the processing capacities and increased storage. For multi-view observation videos recorded in the healthcare communities. Its computing competencies can be utilized to evaluate streaming video footage to categorize keyframes as well as afterward eliminate obsolete and meaningless visual data, thereby reducing the parameters for minimizing bandwidth. The advanced communication skills of sensor nodes can also be

utilized to cooperatively accomplish comprehensive scene assessment to produce real-time multi-view overviews of monitoring keyframes.

Highly intelligence-based sensors can be utilized to yield a mechanized response. Subsequently, anomalous events are spotted such as the need of the patients in emergencies, discomforting from high and low blood pressure among the patients, and feeling symptoms like heart attack among the cardio patients, unbearable pain among the patients, and required help by shouting, monitoring high-risk patients' activities like special care of cancerous patients treating in the special wards [47]. Besides, the core emphasis of our methodology, it is incorporating a very well-designed lightweight keyframe image encryption among the extracted keyframe. A well-organized, highly acceptable proposed smart setup healthcare IIoT is presented in Figure 1. Further coming sections are enforcing the best guidelines for achieving required results and its essential personifications.

3.1. Concept of Extracted Keyframes from Visual Information. In the concept of keyframe extraction, visual information is transversely produced by installing visual sensors and retrieved from VPH (visual-processing hub). This is retrieved in terms of video frames into the entire smart setup which consists a massive amount of visual data. These visual data transportations are unrealistic in terms of distance covering (more data traveling) in entire networks due to bandwidth and limitations of energy constraints in the wireless multimedia sensor networks. To handle these problems, many researchers have employed some technique to control and limits such a massive video information flow. Regarding this, researchers are adopted distinct compression schemes [48, 49] as well as video summary approaches to minimize a huge amount of visual contents at the visual-processing hub. So as to the appropriate frames, videos are traveled towards base stations [23]. Considering energies and bandwidth constraints, we employed an energy responsive keyframe extraction methodology to diminish data redundancy [50]. Our extraction methodology consists a lightweight keyframe extraction with an improved YOLOv3 algorithm [51]. In this methodology, our main emphasis is on extracting relevant keyframe from the summarized video, which is coming through installed wireless multimedia surveillance networks in the smart healthcare system. This algorithm adopted a backbone network which is called Darknet-53 [52].

In Figure 2, the up-sampling network, as well as detection layer, is recognized as a YOLO layer. YOLOv3 is completely dependent on Darknet-53 to extract keyframes from the visual input data. As an essential component, each network is using a residual. Each five layers of residual are selected differently in terms of depth and scale to perform residual among the produced output at diverse layer. All the convolutional in the number showed 53 with a residual block in terms of pair 3×3 and 1×1 layer. The spatial resolution is 1/32 smaller compare to visual input data at the final feature map size. YOLOv3 recognizes as a three-scale layer of YOLO which is basically answerable for the detection of the object at a different layer of scale. The First YOLO layer is based on grid resolution. It contains 1/32 of the visual input data and detecting big objects. The last layer has 1/8 resolution, and that can effectively detect a small extent of objects. Between the layers, one known as the up-sampling layer, and the other are several convolution layers that exist [53]. The resolution of this algorithm is initially set by default 416×416, and the algorithm is also favored height, as well as width, must be equal in size. But no essential to modify resolution before inputting images. We had designed appropriately in a manner to adjust the resolution of each image automatically when its obligatory.

The proposed extraction approach is administering a conceptual description at Figure 2. The whole setup includes a combination of four training modules. Firstly, the interaction of visual contents towards training module is defined. This module has been defined to train data with the help of the model module. Secondly, model module is responsible for modeling data towards received from first module. This module has capability to further model if there are some deficiencies from earlier process. Thirdly, modeled data are approaching to prediction module for accurate prediction. At the last foresight, prediction module is transferred data to detection module. Finally, the output comes as a human presence (patients and staff) via recorded video summary in terms of keyframe. A lightweight YOLOv3 algorithm [51] is quite efficient to detect visual data, either videos or images, in a real-time manner. For the proposed extraction of keyframe, we properly trained the model in the Darknet-53 platform [52]. Additionally, this model is also renewed into the TensorFlow atmosphere. Attaining high precision, which is much required for the experimental work, we trained this model with a vast collection of image datasets. One of them is a wider-face dataset [54]. Originally this model is initially set floating-point as a by default. Therefore, it is transformed in the fixed-point model through TensorFlow. By converting the model, this model behaves computationally competent approach. It is effortlessly incorporated into the smart healthcare system. This can also suit any such small devices such as visual sensors that consume processing, bandwidth constraints, and energy. We effectively tested the model of the extraction by utilizing Face Database (FDB) [55].

Our incorporated methods are proposed to increase gratitude precision at encouraging real-time exhibiting YOLOv3s bounding box, which is a more appropriate symbol of SSD (single-stage detection). The experimental results of the average accuracy are received 88–90% with 1–16 files per second (FPS). This model is operated at the environment of Intel (R) Core i5-6500 Microprocessor, CPU (Central Processing Unit) 3.20 GHz, RAM 8 GB, functioned Windows 10 Operating System. That is highly suitable concerning patient-based monitoring into smart setup healthcare IIoT-enabled architecture which is revealed extracted keyframe in Figure 2. This also indicates that the patient is correctly detected in terms of high precision as well as inside the bounding boxes. The process of extraction is incorporated among the patients for testing numerous clinical wards with the proper settings. After significantly produced keyframe from the model of extraction, the

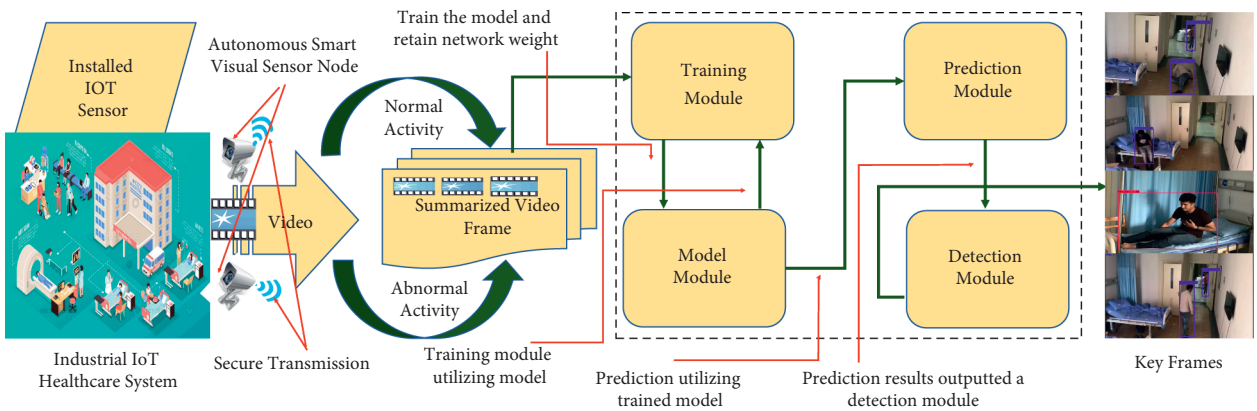


Figure 2: Concepts of extracted keyframes from an installed visual architecture.

keyframe is passed to lightweight cosine function through hybrid chaotic map encryption for the additional security process.

3.2. Lightweight Cosine Functions with Operation of Hybrid Chaotic Map Keyframe Encryptions. This subdivision acquaints with examines the planned cosine function of chaotic sequence (CCS) [40]. It is physiognomies along with effective encryption methodology into the smart healthcare IIoT ecosystem. It is caused by keyframes from the video streaming data by the visual sensor. Our approach is handling first to generate PRNG from CCS with the incorporation of a secret key into the keyframe images. Secondly, it performed confusion as well as diffusion operation into the keyframe image as an engagement of bitwise XOR and highly efficient scrambling algorithms [40] as shown in Figure 3. This methodology is achieved by fast RGB keyframe encryption, randomized complex sequences, and highly encrypted cipher keyframe data. This cipher matrix is impossible to recognize an adversary to detect actual keyframe images. This can also confine the approachability of the information required in a typical cryptosystem. These keyframes are emanating from the monitoring of video streaming, which is done in terms of RGB format with highly resolution visual sensors installed at smart setup of IIoT enabled healthcare.

3.2.1. Preliminaries about Cosine Function. The CCS has incipiently pleaded each existing drawback that frail in chaos as well as tendencies of the weakness in the chaotic maps. This can be explained better in terms of used mathematical notations as below:

$$X_{i+1} = Cos(\pi(Chaotic\ \text{Seed Map 1} + Chaotic\ \text{Seed Map 2} + \beta)), \tag{1}$$

where Chaotic Seed Map 1 and Chaotic Seed Map 2 are known as two different used chaotic maps. That are also well known as a combination of seed map with some control parameter and constant shifting β (set as $\beta = -0.50$).

Equation (1) enlightens that CCS integrates with two seed maps output with adding constant shifting operator to get cosine transformation in terms of output. Mixing with two seed maps are the outcomes of a complex and efficient sequence generation in the cosine function-based hybrid chaotic sequence (CCS). Using two seed maps can also provide flexibility in terms of making countless fresh chaotic maps employing different configurations of prevailing maps. There are so many chaotic seed maps available that can be a better alternative as a complex sequence provider with the consideration or nature of the requirement. Some of them are mentioned below throughout:

Logistic Map:

$$X_{i+1} = \{L(r, x_i) = 4rx_i(1 - x_i). \tag{2}$$

Arnold Map:

$$X_{i+1} = \{\Gamma\colon (x, y) \longrightarrow (2x + y, x + y)\text{mod}1. \tag{3}$$

Tent Map:

$$X_{i+1} = \mu(r, x_i) = \begin{cases} 2rx_i & \text{if } x_i < 0.5 \\ 2r(1 - x_i) & \text{if } x_i \geq 0.5, \end{cases} \tag{4}$$

Chirikov Taylor Map:

$$X_{i+1} = \begin{cases} P_{n+1} = P_n + K\,Sin\theta_n(\theta_n) \\ \theta_{n+1} = \theta_n + P_{n+1} \end{cases}, p_n, \theta_n \text{ mod}2\theta, \tag{5}$$

Sine Map:

$$X_{i+1} = \{S(r, x_i) = rSin(\pi x_i) \tag{6}$$

Zaslavskii Map:

$$X_{i+1} = \begin{cases} x_{n+1} = [x_n + v(1 + \mu y_n) + \varepsilon v\mu Cos(2\pi x_n)]\text{mod}1 \\ y_{n+1} = e^{-r}(y_n + \varepsilon Cos(2\pi x_n)), \mu = (1 - e^{-r}) \div r, \end{cases} \tag{7}$$

Henon Map:

$$X_{i+1} = \begin{cases} x_{n+1} = 1 - ax_n^2 + y_n \\ y_{n+1} = bx_n, \end{cases} \tag{8}$$

Arranging the above chaotic seed maps in terms of haybrid nature such as Equation 1. The incorporated cosine function-based hybrid chaotic sequences compile on the following:

Cosine-transform: Pair 1.

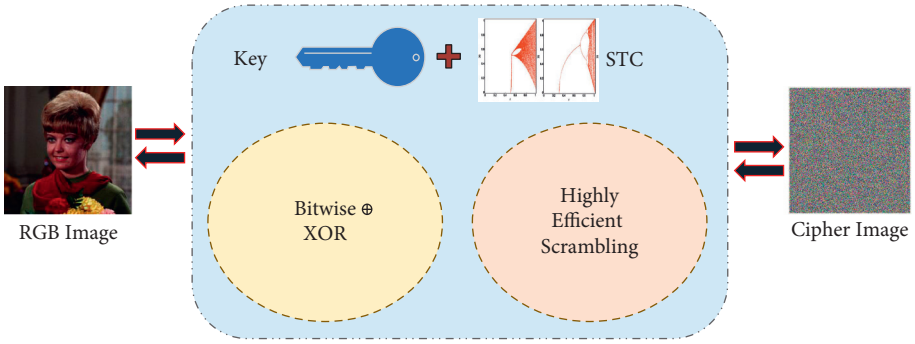


Figure 3: Encryption methods of the keyframe.

$$X_{i+1} = \{Cos(\pi(4rx_i(1 - x_i) + (1 - r)Sin(\pi x_i) - 0.5)). \tag{9}$$

Cosine-transform: Pair 2.

$$X_{i+1} = \begin{cases} Cos(\pi(rSin(\pi x_i) + 2(1 - r)x_i - 0.5)) & \text{for } x_i < 0.5 \\ Cos(\pi(rSin(\pi x_i) + 2(1 - r)(1 - x_i) - 0.5)) & \text{for } x_i < 0.5, \end{cases} \tag{10}$$

Cosine-transform: Pair 3.

$$X_{i+1} = \begin{cases} Cos(\pi(2rx_i + 4(1 - r)(1 - x_i) - 0.5)) & \text{for } x_i < 0.5 \\ Cos(\pi(2r(1 - x_i) + 4(1 - r)x_i(1 - x_i) - 0.5)) & \text{for } x_i < 0.5, \end{cases} \tag{11}$$

where $r \in [0, 1]$, α replaces r and b swaps $1 - r$. Equations (9)–(11) are the same as our standard notations, as mentioned in equation (1). The proposed method is used equation (10) STC maps (Sine Tent Cosine) as the actual combination of the CCS in the entire paper, which is the adaptation of Sine and Tent seed maps.

3.2.2. Structural Setup towards Producing Key. The reliable key controls the series of STC at primary states. Concerning the ideal state of key, the researchers emphasized on [56]. It is highly counseled to withstand diverse sorts of attacks whenever the key size of every chaos-based cryptographic protocol corresponds to the 2^{100} proportions. Concerning the effective length, which can provide freedom to maintain a higher security aspect, the key space of STC-IES used 256-bit long proportion that can be represented as 2^{256}. It is made of five key components such as $K = \{X_0, Y_0, U, g, d\}$, in which preliminary states are addressed with (X_0, Y_0), U is the best represented component as a constraint of disorder that befuddle the preliminary states, g is the coefficient of preliminary states, and d comprises four-parameter of disturbance coefficients represented as $\{d_1, d_2, d_3, d_4\}$. Individually of each X_0, Y_0, U, g, d_1, d_2, d_3, and d_4 are a set of 32-bit length proportion. The float variables are X_0, Y_0, U, within the range of [0, 1] and individually represented as the 32-bit stream as follows:

Float number $\sum_{i=1}^{32} Binary_i \times 2^{-i}$, where g, d_1, d_2, d_3, and d_4 are the coefficients of integer, which can be generated as follows: integer values $\sum_{i=1}^{32} Binary_i \times 2^{i-1}$.

These preliminary stages can be additionally calculated by the way of proclaimed setup of encryption operation form as follows:

$$\begin{cases} X_0^i = (X_0 \times g + U \times d_i)\text{mod}1, \\ Y_0^i = (r_0 \times g + U \times d_i)\text{mod}1. \end{cases} \tag{12}$$

The STC map is generated uniformly scattered chaotic sequences, which is shown in Figure 4 through Table 1 for the process of bitwise XOR and highly efficient scrambling at the preliminary stages (X_1, Y_1).

3.2.3. Steps of Lightweight STC-IES. The proposed STC-IES approach is the keyframe encryption as well as decryption algorithm suggested in Tables 1–3, respectively. The key approach for producing encryption is demarcated as follows:

(I) Each generated keyframe image is first permuted by sine tent cosine (STC) chaotic sequence through a randomly generated secret key to accomplishing a more complex chaotic sequence. Additionally, keyframe is reshaped through recently created complex nature chaotic sequence. After that, splitting each keyframe image as a matrix form of its color RGB channel such as red, green, and blue, an operation is performed as a bitwise XOR and highly efficient scrambling.

(II) Each matrix of color keyframe is diffused by bitwise XOR by combining highly efficient scrambling algorithms to generate the cipher keyframe images.

(III) The cipher keyframe is confused through the highly efficient scrambling procedures [40] to obtain the caused encrypted keyframe image.

The decryption process approach is fully followed through the inverse of highly efficient scrambling procedures and inverse bitwise XOR methods, that is, consecutively decrypt encrypted cipher image at the original keyframe images. Tables 1–3 are confirmed pseudocode. It is used for bitwise XOR, STC sequence generation with the addition of highly efficient scrambling algorithms, one-to-one, where S is the sine tent cosine (STC) chaotic sequence and the size of the keyframe image is $M \times N$. The produced secure key is used in entire cryptosystem and is formed with the random key generation procedures.

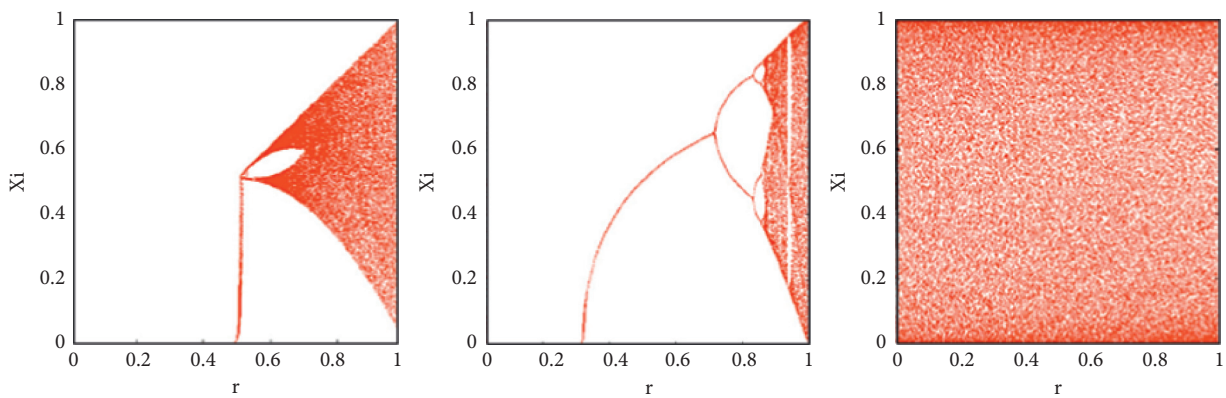


Figure 4: Bifurcation of tent map, sine map, and complex STC sequences [40].

Table 1: Algorithm 2, STC sequence generation.

Input: key and size of keyframe $(M \times N)$
Output: STC generation
(1) Reading secret key K (and allocate to x)
(2) Read r
(3) Resetting chaotic sequence
(4) Starting for loop from $m = 1:100$
(5) $X = Cos(\pi \times (r \times Sin \times (\pi \times x) + 2 \times (1 - \mathrm{r}) \times x - 0.5))$
(6) end of the for loop
(7) Starting for loop from $m = 1: M \times N$
(8) $X = Cos(\pi \times (r \times Sin \times (\pi \times x) + 2 \times (1 - \mathrm{r}) \times (1 - x) - 0.5))$
(9) $S\ (m) = X$
(10) end of the for loop
(11) STC generation

Table 2: Algorithm 1, encryption process through bitwise XOR.

Input: size of keyframe image (M×N) per initial state (X_0^i, Y_0^i)
Output: encrypted cipher keyframe C
(1) Read size (M×N) RGB keyframe image
(2) Resizing keyframe images
(3) Generation of random number for getting key
(4) Generating secret key by calling function
(5) Generating STC sequence by calling function
(6) Reshaping keyframe images with STC
(7) Splitting RGB keyframe image in each 3 channels IR, IG, and IB
(8) Bitwise XOR operation is performed in every channel
(9) Performing highly efficient scrambling algorithm obtained keyframe channel from step 9
(10) Merging every scrambled keyframe channel
(11) Encrypting cipher keyframe images C

Table 3: Algorithm 2, highly efficient scrambling.

Input: keyframe (IR, IG, IB) channels size $M \times N$, STC sequences, state
Output: cipher scrambled keyframe channels
(1) Obtain block size $H = \min$ {[M1/2], [N1/2]} of color channels separately (operated keyframe of bitwise XOR)
(2) Producing four STC sequences such as O, P, Q, and R
$O = \mathrm{STC}_{1:\ H^2}, P = \mathrm{STC}_{H2+1:\ 2H^2}, Q = \mathrm{STC}_{2H2+1:\ 3H^2}, R = \mathrm{STC}_{3H2+1:\ 4H^2}$
(3) Sorting each four STC sequences such as sort (O), sort (P), sort (Q), and sort (R) and obtaining their corresponding indexes
(4) Two matrixes S and T are initialized with the square of block size $H^2 \times H^2$
(5) Starting loop, for $j = 1$: H^2 (H^2 square of the block size within $M \times N$)
(6) Starting loop, for $i = 1$: H^2
(7) a = $((\mathrm{i} + \mathrm{IP}(\mathrm{j}) - 1) \bmod \mathrm{H}^2 + 1)$
(8) b = $((\mathrm{i} + \mathrm{IR}(\mathrm{j}) - 1) \bmod \mathrm{H}^2) + 1$, Si, $j = \mathrm{IOa}$, ti, $j = \mathrm{IQb}$

Table 3: Continued.

(9) end of the for loop
(10) end of the for loop
(11) If en = = state
(12) Starting loop, for $i = 1$: H^2
(13) Starting loop, for $j = 1$: H^2
(14) $m = \mathrm{Si}, j$; $n = \mathrm{Tm}, j$;
(15) $c_1 = ((\mathrm{m} - 1) \div \mathrm{H})$; $c_2 = ((n - 1) \div \mathrm{H})$;
(16) $\mathrm{b}_1 = (\mathrm{m} - 1) \bmod \mathrm{H}$; $\mathrm{b}_2 = ((n - 1) \bmod \mathrm{H}) + 1$;
(17) $x = c1 \times H + c2$ $y = b1 \times H + b2$
(18) Scramble channel (x, y) = operated keyframe of bitwise XOR channel (i, j)
(19) end of the for loop
(20) end of the for loop

4. Simulation Results and Discussion

This segment imposes as well as discourses the investigational consequences of the proposed STC-IES on MATLAB 2018a version software. It also analyzes security with the help of using test images from the eminent source USC-SIPI [57] repository database of digital images. An ideal and prominent encryption approach is always intelligent to encrypt dissimilar classes of images into highly secure cipher images. When an encrypted method is strongly enforced to reach an optimum security level in terms of cipher images, it is noticeable that any keyframe can only be retrieved by knowing the arcuate used top-secret key. Unless the known used secret key, not any information of the keyframe images can be extracted in smart healthcare IIoT-enabled system. Figure 5 is illustrated the encryption approach. It used USC-SIPI images A (as a test image). Image A is the pepper image, and its corresponding originated all the three-color frequency as mentioned at histogram R, G, and B, which is purely showing the maximum correlation of the pixels in each color channel. The encrypted cipher image of the pepper image B and its corresponding three-color channel histogram R, G, and B show uniform distribution in each color channel. Similarly, Image C is the keyframe image. It originated from all the color channels as shown in each color channel histogram R, G, and B, which originally showed maximum pixels of the correlation in individual color channels. The encrypted cipher image of the keyframe image is D. It is conforming that the entire three-color channel histogram R, G, and B are screening uniform distribution in the separate color channel.

Figure 5 is indicating that the proposed approach has a strongly encrypted methodology. Uniformly distributed pixels are also visible in the cipher image. It means this cipher image has adopted better encryption process; in this way, any intruder or attackers cannot effectively gain original keyframe information from cipher image. Therefore, STC-IES can easily convalesce every keyframe fully from consequently encrypted image data for the reason that all the handling process is completely reversible. Our methods have robust inefficiency, fast execution, and minimal computational overhead, employing bitwise XOR and highly efficient scrambling algorithms at a smart healthcare setup. Thus, this encryption methodology can be accomplished faster and efficiently in the IIoT ecosystem. To provide better security as well as privacy towards the patients in the respectively possible emergency requirement, Table 4 depicts the time complexity to encrypt keyframe, uses test images, and undergoes comprehensive comparison with the different types of available encryption algorithms.

The simulation and experimental cryptosystems are effectively conducted on the required platforms. For example, Intel (R) Core i5-6500 Microprocessor, CPU (Central Processing Unit) 3.20 GHz, RAM 8GB, operated Windows 10 Operating System. The demonstration of the fast keyframe encryption is exhibiting STC-IES from various image extent exemplified on the numerous results in terms of tabular form. It is always required to have a vintage high-quality of deciphered images because a noteworthy discrepancy of pixels into keyframe can disturb entire pixels into the encrypted images. Incorporation of the STC-IES can be a safeguard in terms of higher security aspects of the cipher images. A delicate modification of pixels into a ciphered encrypted image matrix can distress some pixels as a deciphered conclusion (results) into the decryption approach. In this crucial condition, if STC-IES encrypted data miss some pieces of information, the process of decryption can also easily recuperate tangible keyframe visual facts. Figure 6 demonstrates the quality-based approach after the decipherment process as soon as the STC-IES encrypted cipher images suffered from noises or various types of proportion data forfeiture. For example, in Figure 6(a), the decipherment methodology is fully recovered in pepper image and no data loss into encrypted cipher of pepper image are viewed. Even though the cipher images had also lost some data (info) or noise moreover, their deciphered findings incorporate most sensory information of the same individual pepper images, which can be presented in Figures 6(b) to 6(d). Accordingly, the employed STC-IES can fairly decrypt the superior quality in terms of encrypted cipher.

5. Security Analysis

With the intention of validating STC-IES preeminence, this unit scrutinizes efficient security in terms of resulting aspects with minimal encryption speed assessment, histogram analysis, information entropy analysis, differential attack analysis, correlation analysis, analysis of produced key, and comprehensive comparative analysis among surveillance systems. To supplementary exemplify the effectiveness of STC-IES, we equate our approach with the

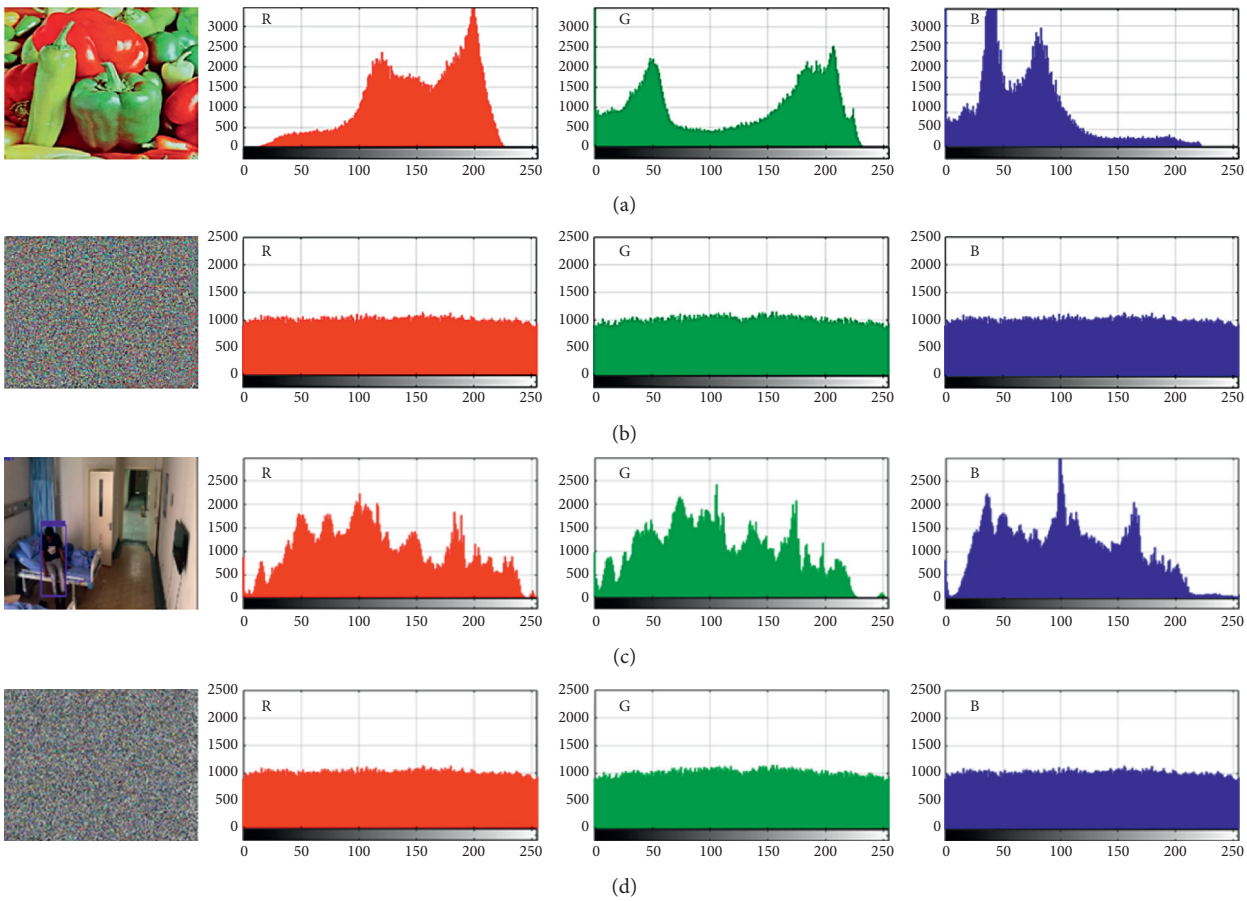


Figure 5: Fair encryption methodologies (STC-IES) operated over test and keyframe images.

Table 4: Encryption speed and its comparison with relative approaches.

Method	Encryption time (E.T) 256 × 256 keyframes (images)	Encryption time (E.T) 512 × 512 keyframes (images)	Encryption time (E.T) 1024 × 1024 keyframes (images)
Proposed	0.0670–0.08470	0.2205–0.2541	1.0079–1.1095
[20]	0.1616	0.6708	2.821
[37]	1.28	NA	NA
[40]	0.0949	0.4010	1.9857
[58]	0.080–0.082	0.327–0.333	NA
[59]	0.6212	NA	NA
[60]	0.3340	1.3357	5.3223
[61]	0.224	0.9731	3.8377
[62]	1.7874 (En & Dec)	NA	NA

other available advanced image encryptions. We cited directly, respectively, finding in terms of references reported at esteemed scientific journals by the author for the fair comparison at the area of highly noticeable image-based encryption algorithms.

5.1. Encryption Speed Assessment. This segment is familiarized with minimal computational-based overhead and speed assessment displayed in Table 4. The time computation segment of any encrypted system depends on the generation of chaotic sequences with the proper handling of permutation and diffusion methods in the operating algorithms, eEnforcing smart healthcare enabled IIoT system. We accentuated to achieve minimal computational overhead for constructing relatively better communication in speed with the adoption of the STC-IES mechanism. We observed that the generation of complex sequences is relatively quick in nature. The operation of bitwise XOR with highly efficient scrambling is operating in minimal computational time. The proposed approach is demonstrated as an average encryption time assessment for each set of various sizes of the keyframes. The production of numerical values of the encrypted keyframes is presented in Table 4 and compared with other past color image encryption methodology such as [20, 37, 40, 58–62]. Our proposed approach

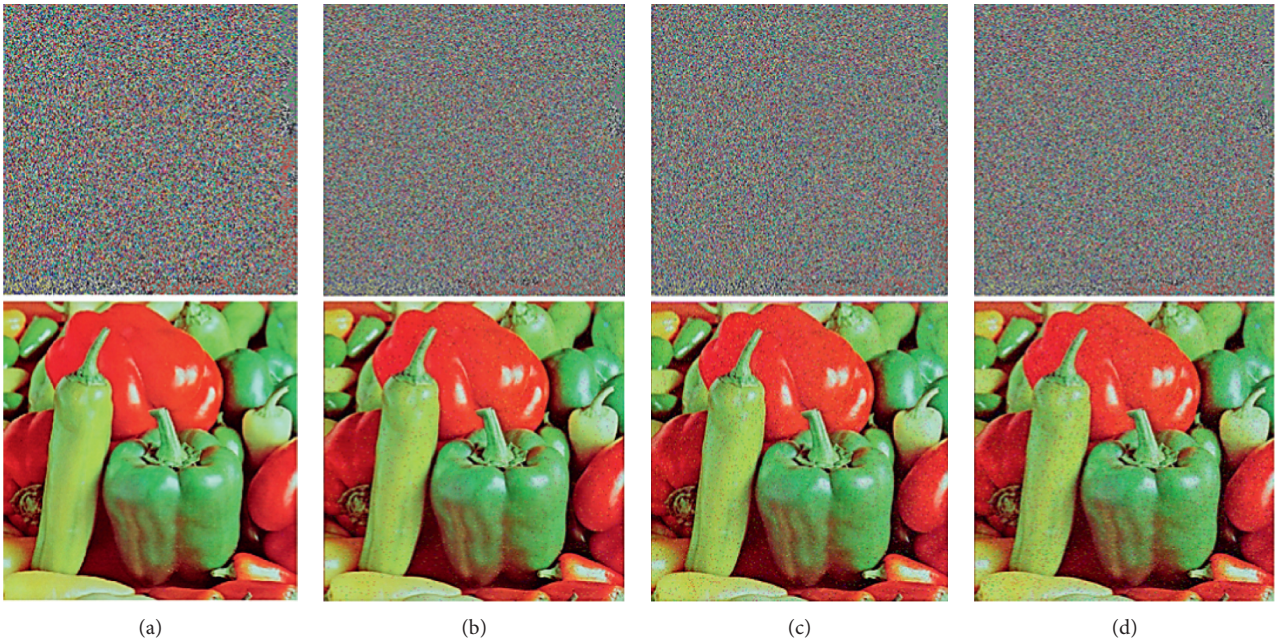


Figure 6: Quality-based analysis of decrypted image: (a) encrypted and simultaneously deciphered image; (b) noise-encrypted simultaneously deciphered image containing 1% salt and pepper; (c) noise-based encrypted simultaneously deciphered matrix containing 2% salt and pepper; (d) noise-based encrypted simultaneously deciphered matrix containing 3% salt and pepper.

is fast in running and has minimum computational effort, which enforces for real-time application setup like a smart healthcare IIoT-enabled architecture.

5.2. Histogram Analysis. Keyframe histogram is the best illustration of a typical graphical representation of the pixel rate distribution in the keyframe images. The keyframe/encrypted cipher keyframe image histogram contains the statistical data by which evaluation of the robustness recognizes the statistical data analysis. Really, histogram reports about the distribution of a keyframe's grey-level values, and relatively smooth distribution will expose the loopholes in the method that attackers can use to initiate a preferred-ciphertext attack using statistical analysis. However, for reliable cryptography, the dispersal of histograms must be uniform. Figure 5 demonstrates the histograms of the keyframe and its ciphered images, which can pictorially recognize that the arrangement amongst pixel correlation at ciphered-keyframe images has fairly unvarying color frequencies. Still, there are a few variations in the original keyframe image [44, 58, 63].

The histogram deviation is used to measure an encrypted image, and this also contributes to significant analysis. Suppose the quantity of variation in the encrypted cipher keyframe image is small, the greater its uniformity. Two encrypted keyframe images are produced using different surreptitious keys, even with similar image data. It shows better homogeneity, uniformity, of the encrypted keyframe images if the differences are similar enough. Figure 5 demonstrates color test image histograms, wherein it is clearly verified, specifically color frequency histogram representation and its encrypted image histogram accurately. We perceived that each color channel of the test images and keyframe images shows their natural behavior before encryption, and after encryption, each color channel behaves uniformly in nature. The uniform behavior of the histogram can be calculated numerically of a variance. The uniform behavior of a histogram is purely dependent on the variance. Lower variance is recognized as the higher nature of uniformity. The higher variance is lower uniformity. The variance of the histogram can be statistically obtained by Eq (13), in which grayscale numerical value is n, Y_i, and Y_j stands for the values of the pixels at the ith, j^{th} gray levels [58, 64]:

$$\text{Variance}(y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\left(y_i - y_j\right)^2}{2}. \tag{13}$$

Consequently, in such a scenario, if any attacker decrypts any piece of the image, assuredly, the attacker cannot find the entire original data because the proposed method has a powerful encryption methodology that keeps away any data leakage. In the same way, with the observation of the keyframe channels as well as encrypted keyframe channels remains identical keyframe, it is signifying no data loss during the communication into smart healthcare IIoT enabled system occurs. Therefore, the histogram analysis equally justifies that the proposed STC-IES methodology strongly avoided any statistical or numerical attacks. It also provided a statement for the integrity, including consistency during the transmission at smart setup healthcare system.

5.3. Information Entropy Analysis. The methodology of analyzing the amount of uncertainty and randomness behavior among the cipher images' correlated pixels is called information entropy. It can be well demarcated as

$$H(n) = -\sum_{i=1}^{255} P(n_i)\log_2 P(n_i), \tag{14}$$

where information entropy is $H(n)$ and $P(n_i)$ is the probability of the occurrences at each n_i. Generally, the possible information entropy's numerical value recognizes as 8 for the ideal scenario at each random keyframes or image. In an ideal, discrete image, probabilities of any containing n_i signs are identical. The probability among each n_i symbol is, therefore, 1/256:

$$H(n) = -\sum_{i=1}^{255} P(n_i)\log_2 P(n_i) = -\sum_{1}^{255} \frac{1}{256}\log_2\frac{1}{256} = -256 \times \frac{1}{256}\log_2\frac{1}{2^8} = -\log_2 2^{-8} = -(-8)\log_2 2 = 8. \tag{15}$$

After encryption, the encoded cipher image must continue to act as the best possible random image. This is validated that encrypted keyframe information entropy should be around numerical value 8, significantly proved from equation (14). The following is also demonstrated: the information entropy around 8 is highly random in nature and very low information leakage is present in our proposed smart healthcare IIoT-enabled system. Table 5 demonstrates information entropy of color channels with its extracted as well as encrypted keyframes (1–6). It is also relatively compared in Table 6 with earlier referenced encryption methodology such as [20, 50, 65–70]. Tables 5 and 6 exemplified by the projected approach have shown healthier results in terms of information entropy for each color channel, which almost have numerical value 8 as an encrypted keyframe image. This indicates that the proposed STC-IES methods deliver quite a suitable notch of security, privacy, and randomness behavior in contradiction to the information entropy attack. That is the furthermost requirement for the appropriate way of handling secure concern smart healthcare IIoT-enabled system.

5.4. Differential Attack Analysis. The differential attack is a well-known attack into keyframe image encryption. This attack is used as an active security attack. It focuses on associating a sturdy relationship among keyframes per their ensuing encrypted images via perceiving transformations at a keyframe. It can annoy encrypted data. If it preserves diffusion in an ideal way, the keyframe encryption methodology expresses high usefulness in the sidestepping differential attack. This diffusion process proves that the encrypted-cipher keyframe possibly will scatter a minor alteration throughout the keyframe image over the entire statistics or data. The proposed STC-IES mechanism is appropriately demonstrated as an innovative way diffusion process in Figure 5.

The differential attacks can be best examined among encryption methodologies to withstand attacks with the two well-known metrics such as NPCR (number of pixel change rate) and UACI (uniform average change intensity) [74]. NPCR calculates the change percentage amongst the pixel positions within an encrypted cipher keyframe by adjusting one pixel's worth at keyframe or original image. UACI evaluates the average strength difference between individual keyframes besides their respective encrypted-cipher images. Suppose two encrypted-cipher keyframe images denoted as C_1 and C_2 which have a slight one-bit discrepancy in their corresponding keyframe; at that moment, their NPCR and UACI can be governed from equations (16)–(18) as follows:

$$NPCR = \frac{\sum_{i,j} G(i,j)}{\mathrm{M} \times \mathrm{N}} \times 100\%, \tag{16}$$

$$G(i,j) = \begin{cases} 0 \; if \; C_1(i,j) = C_2(i,j) \\ 1 \; if \; C_1(i,j) \neq C_2(i,j) \end{cases}, \tag{17}$$

$$UACI = \frac{1}{\mathrm{M} \times \mathrm{N}} \left[\sum_{i,j} \frac{\left| C_1(i,j) - C_1(i,j) \right|}{255} \right] \times 100\%. \tag{18}$$

The magnitude of the keyframe matrix is denoted as $M \times N$. Equation (17) is calculated $G(i, j)$, which is pixel transformation between corresponding encrypted keyframe images. Table 7 pointedly illustrates the best possible NPCR and UACI produced values from keyframe in the proposed STC-IES algorithm, and reasonable comparison is shown in Table 8 through previously referenced encryption algorithms such as [20, 50, 71–73, 75]. The results indicate that the produced result of NPCR is very near to a hundred and produced the result of UACI approaching very near to one-third of the hundred. This is purely demonstrates encryption methodology of STC-IES. It incorporates effectively dissimilar randomize keyframe images as well as completely misses the effectiveness of differential attacks. In this manner, the planned cryptosystem ensured complete security and privacy from the attacker for not getting any information into smart healthcare IIoT-enabled system.

For example, investigation of ideal encryption and randomness nature of NPCR, UACI result into cryptosystem. Suppose the encrypted matrixes which size is M×N, C1, and C2 are encrypted cipher keyframes. An ideal encrypted cipher image is a certain image that cannot be discerned from a pseudorandom image. Simply cipher image is an arbitrary arena at the size of $M \times N$, where integer $i \in [1, M]$ and $j \in [1, N]$, arbitrary pixels price $C(i, j)$ identical nature and independently occurs isolated unvarying distribution taking place 0 to C's principal buttressed numeral E such as $\forall j \in [1, \mathrm{N}]$, $\forall i \in [1, \mathrm{M}]$, $\exists C(i,j) \sim i.i.d\, U(0, E)$ [74]. The

Table 5: Information entropy of keyframe and test images.

Methods	Images	Plain image			Encrypted image		
		R	G	B	R	G	B
Proposed	Pepper	7.3388	7.4963	7.0583	7.9962	7.9831	7.9981
	Lake	7.3124	7.6429	7.2136	7.9876	7.9878	7.9977
	P001 (keyframe 1)	7.8579	7.9484	7.8348	7.9971	7.9961	7.9892
	P002 (keyframe 2)	7.7876	7.7255	7.6506	7.9989	7.9975	7.9956
	P003 (keyframe 3)	7.7592	7.6845	7.6197	7.9978	7.9952	7.9959
	P004 (keyframe 4)	7.7925	7.7459	7.7153	7.9984	7.9898	7.9959
	P005 (keyframe 5)	7.7982	7.8063	7.7699	7.9988	7.9965	7.9959
	P006 (keyframe 6)	7.7754	7.7022	7.6331	7.9899	7.9897	7.9966

Table 6: Information entropy comparison through relative approaches.

Methods	Images	Plain image			Encrypted image		
		R	G	B	R	G	B
Proposed	P002 (keyframe 2)	7.7876	7.7255	7.6506	7.9989	7.9975	7.9956
[20]	Keyframe 6	6.4410	6.3789	6.4770	7.9978	7.9978	7.9979
[50]	Keyframe 4	7.0818	6.7460	7.1210	7.9969	7.9919	7.9954
[65]	House	7.4007	7.2312	7.4357	7.9985	7.9984	79985
[66]	Baboon	7.7326	7.7591	7.4557	7.9993	7.9993	7.9994
[67]	Girl	7.3490	7.1876	6.9857	7.9994	7.9995	7.9994
[68]	Image 2	4.7664	4.4860	5.0793	7.9021	7.9027	7.9023
[69]	G01	7.16399	7.16399	7.16399	7.99696	7.99696	7.99696
[70]	Pepper	NA	NA	NA	7.9993	7.9993	7.9991

Table 7: NPCR and UACI of the keyframe and test images.

Methods	Images	NPCR			UACI		
		R	G	Blue	R	G	B
Proposed	Pepper	99.5520	99.5735	99.6532	33.3396	33.3363	33.4394
	Lake	99.4546	99.5453	99.6868	33.32473	33.2669	33.3329
	P001 (keyframe 1)	99.6634	99.5383	99.5983	33.3782	33.3493	33.3483
	P002 (keyframe 2)	99.5989	99.6682	99.5959	33.3498	33.3467	33.3467
	P003 (keyframe 3)	99.5566	99.6536	99.5990	33.4086	33.3474	33.3364
	P004 (keyframe 4)	99.6343	99.5888	99.6699	33.3366	33.3495	33.3498
	P005 (keyframe 5)	99.6346	99.6456	99.6412	33.3332	33.3983	33.3351
	P006 (keyframe 6)	99.4996	99.5978	99.5892	33.3596	33.5477	33.3499

Table 8: NPCR and UACI comparison through relative approaches.

Methods	Images	NPCR			UACI		
		R	G	Blue	R	G	B
“	P003 (keyframe 3)	99.5566	99.6536	99.5990	33.4086	33.3474	33.3364
[20]	Keyframe 4	99.6009	99.5899	99.6311	33.4910	33.3394	33.4804
[50]	Keyframe 3	99.6136	99.6136	99.5960	33.4406	33.3564	33.2764
[66]	Baboon	0.9962	0.9962	0.9962	0.2988	0.2844	0.3104
[71]	Frame 5	99.6070	99.5808	99.6307	33.4251	33.4013	33.5713
[72]	House 2	0.999908	0.999908	0.999923	0.332855	0.332929	0.332986
[73]	Lena	99.6258	99.6183	99.6182	33.4635	33.4877	33.4749

Table 9: Randomness test of NPCR.

Algorithms	Image shape	NPCR	NPCR00.05	NPCR tests
STC-IES	512 × 512	99.6536	99.6034	Pass

Table 10: Randomness test of UACI.

Algorithms	Image shape	UACI	$UACI_{0.05}^{*-}$	$UACI_{0.05}^{*+}$	UACI tests
STC-IES	512 × 512	33.4086	33.3463	33.4568	Pass

hypothesis assessment of NPCR (C_1, C_2) by means of β-level consequence follows as [74] in

$$\begin{cases} H_0\text{: NPCR } (C_1, C_2) = \xi_{NPCR} \\ H_1\text{: NPCR } (C_1, C_2) < \xi_{NPCR}. \end{cases} \tag{19}$$

It is implicit at what time $NPCR(C_1, C_2) < \xi_{NPCR}$ [76], and H_0 hypothesis assessment is purely rejected. That is the vital worth designed for challenging NPCR. On the contrary, H_0 hypothesis assessment is acknowledged. The produced result of $NPCR_\beta^*$ is critically enlightened [74] in the following equation:

$$NPCR_\beta^* = \xi_{NPCR} - \frac{\eta_{NPCR}}{\Phi(\beta)} = \left(\mathrm{E} - \frac{\sqrt{\mathrm{E/MN}}}{\Phi(\beta)}\right)/(E+1), \tag{20}$$

where $\Phi(\beta) = 1/\sqrt{2\pi}\exp(-\beta^2/2)$ [77–79] is well known as CD (cumulative density) function of SD (standard normal distribution) which has range amongst zero towards one, expressed through N [0, 1], and E represented as a gray image that is incorporated numerical value 255 in the current study:

$$\begin{cases} H_0\text{: UACI } (C_1, C_2) = \xi_{UACI} \\ H_1\text{: UACI } (C_1, C_2) < \xi_{UACI}. \end{cases} \tag{21}$$

Once $\mathrm{UACI}(C_1, C_2) \notin (UACI_\beta^{*-}, UACI_\beta^{*+})$ [80], H_0 hypothesis assessment is virtuously rejected. That is the crucial worth designed for challenging NPCR. On the contrary, H_0 hypothesis assessment is legalized. The produced results of $UACI_\beta^{*+}$, $UACI_\beta^{*-}$ are critically explicated [74] in the following:

$$UACI_\beta^{*-} = \xi_{UACI} - \frac{\eta_{UACI}}{\Phi(\beta/2)}, \tag{22}$$

$$UACI_\beta^{*+} = \xi_{UACI} + \frac{\eta_{UACI}}{\Phi(\beta/2)}, \tag{23}$$

$$\xi_{UACI} = \frac{E+2}{3E+3}, \tag{24}$$

$$\eta_{UACI} = \sqrt{\frac{(E+2)\left(E^2+E+3\right)}{18(E+1)^2 MNE}}. \tag{25}$$

Tables 9 and 10, correspondingly, exhibit assessment results from level $\beta = 0.05$ of the NPCR and UACI.

5.5. Correlation Analysis. CC_{xy} correlation coefficient of the two neighboring pixels offered randomness information. That is a parameter for calculating a keyframe cipher's robustness and can only be quantified utilizing equations (26)–(32). CC_{xy} analyses the total amount of linear correlation between both the adjoining pixels in the keyframe images. For authentic images, each direction of the keyframe images (diagonal, horizontal, and vertical) amongst pixel and the corresponding pixel is highly correlated. The STC-IES method's primary function is likely to tinkle causal relationships between nearby pixels together at each direction, such as vertical (V), horizontal (H), and its diagonal (D). At the same time, keyframe data matrix achieved approximately zero correlation with the highest possible unpredictable nature and strangeness (randomness) [44, 81, 82]:

$$CC_{xy} = \frac{\mathrm{Covariance}(x, y)}{\sqrt{D(x)D(y)}}, \tag{26}$$

$$\mathrm{Covariance}(x, y) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - E(x)\right)\left(y_i - E(y)\right), \tag{27}$$

$$D(x) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - E(x)\right)^2, \tag{28}$$

$$D(y) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - E(y)\right)^2, \tag{29}$$

$$E(x) = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{31}$$

$$E(y) = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{32}$$

The scope for the coefficient's performance is between +1 (a positive one) and −1 (a negative one), minor the value of the coefficient for encrypted cipher keyframe image, higher the quality of the cipher to battle any the statistical attack. Scenario 1: If $CC_{xy} > 0$, it exposes a positive correlation into the encrypted matrix. Scenario 2: If $CC_{xy} < 0$, it exposes negative correlation into the encrypted cipher matrix. Scenario 3: If $CC_{xy} = 0$, it exposes no correlation into the encrypted keyframe cipher matrix,

Table 11: CCs of two bordering pixels, plain and ciphered images (including test images).

Methods	Images		Plain images			Encrypted images		
			H	*V*	*D*	*H*	*V*	*D*
Proposed	Pepper	R	0.9786	0.9820	0.9694	5.1e-04	1.2e-04	0.0017
		G	0.9775	0.9990	0.9650	0.0017	-6.7e0-4	-4.2e-04
		B	0.9760	0.980	0.9640	0.0031	1.6e-04	-3.5e-05
	Lake	R	0.9677	0.9663	0.9502	7.6e-04	-0.0016	0.0014
		G	0.9620	0.9640	0.9540	9.4e-04	-0.0011	-2.4e-04
		B	0.9670	0.9650	0.9520	0.0022	4.5e-04	0.0031
	P001 (keyframe 1)	R	0.9630	0.9760	0.9427	-0.0023	0.0017	5.5e-04
		G	0.9620	0.9740	0.9410	9.6e-04	0.0023	5.3e-04
		B	0.9605	0.9750	0.9430	2.2e-06	0.0022	-6.2e-04
	P002 (keyframe 2)	R	0.9578	0.9905	0.9511	2.3e-05	-4.3e-04	0.0032
		G	0.9560	0.9900	0.9505	0.0028	-9.1e-04	0.0033
		B	0.9570	0.9901	0.9500	5.2e-07	-3.3e-04	0.0011
	P003 (keyframe 3)	R	0.9554	0.9926	0.9498	6.16e-04	-0.0022	-3.1e-04
		G	0.9540	0.9910	0.9480	-5.3e-04	-0.0011	0.0012
		B	0.9545	0.9918	0.9490	0.0015	-0.0038	5.01e-04
	P004 (keyframe 4)	R	0.9578	0.9939	0.9525	3.12e-04	-0.0016	0.0011
		G	0.9500	0.9906	0.9510	0.0015	0.0024	-0.0022
		B	0.9565	0.9925	0.9520	-0.0013	0.0011	0.0023
	P005 (keyframe 5)	R	0.9494	0.9889	0.9386	3.23e-04	8.45e-05	0.0012
		G	0.9480	0.9875	0.9360	0.0016	0.0013	-0.0014
		B	0.9485	0.9870	0.9375	4.50e-05	0.0017	-0.0014

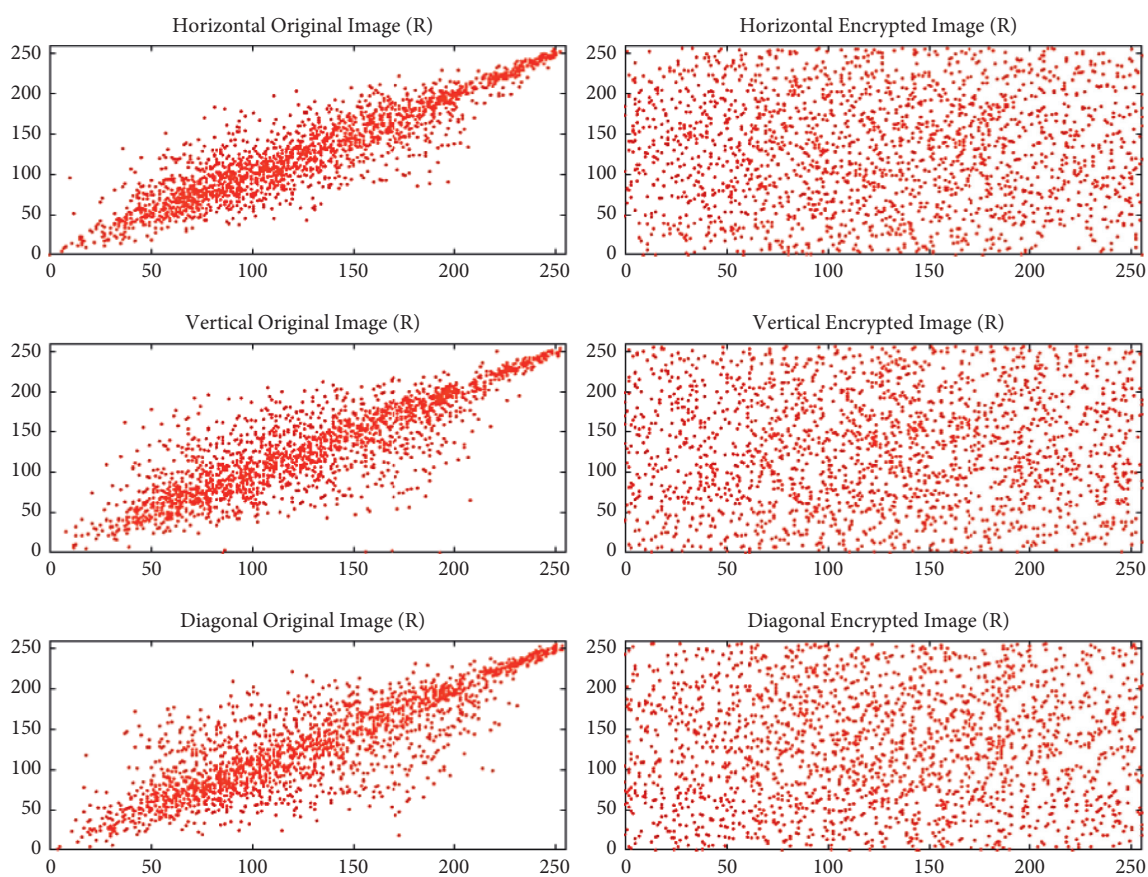


Figure 7: Distributed very close adjacent pixel (horizontal, vertical, and diagonal plane) at keyframe as well as encrypted keyframe (red channel of a baboon).

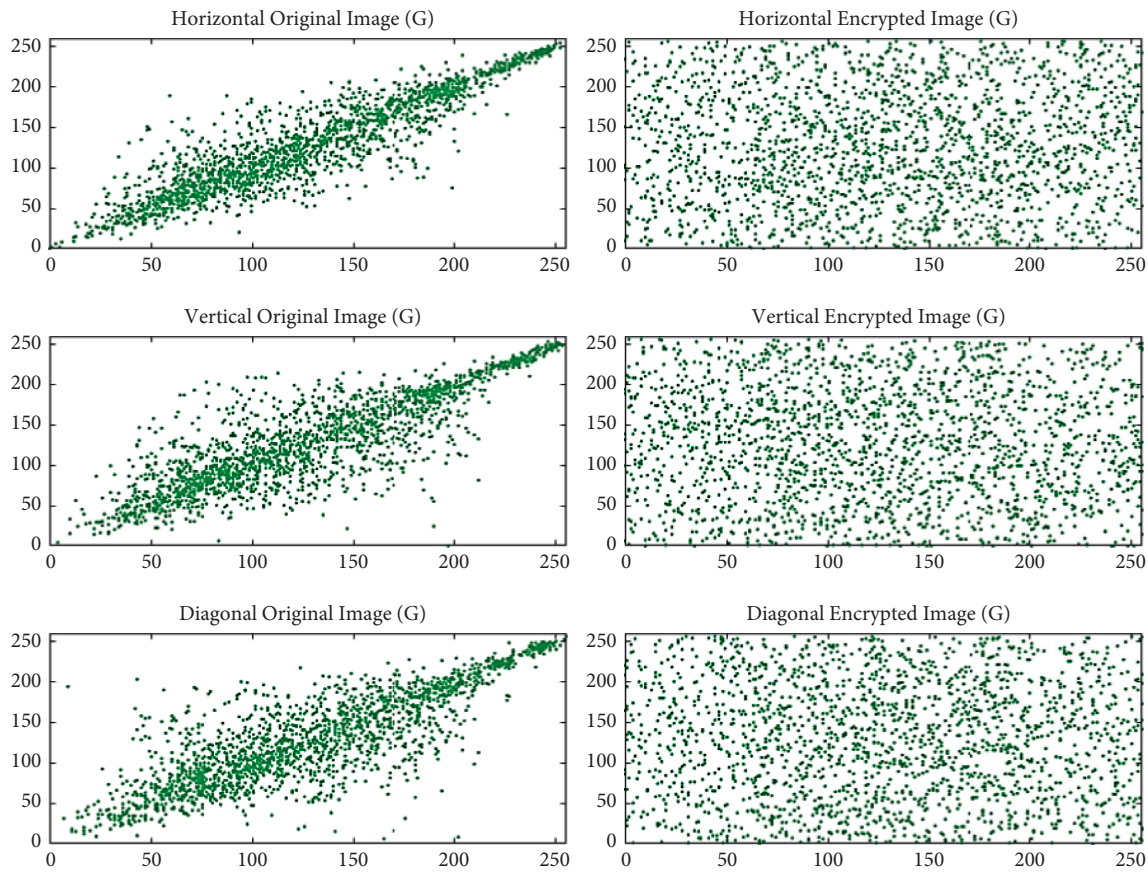


Figure 8: Distributed very close adjacent pixel (horizontal, vertical, and diagonal plane) at keyframe as well as encrypted keyframe (green channel of a baboon).

[44, 58, 83]. In this article, Table 11 demonstrates the random assortment of ten thousand combinations of two corresponding pixels at every three directions and correlation coefficient (CC_{xy}) values for test image pepper, lake, and extracted keyframe images with the size of (512×512). The discoveries of Table 11 and Figures 7–9 state significantly that the CC_{xy} of two adjoining pixels of extracted keyframe images in each direction, such as diagonal, vertical, and horizontal is just about to 1. However, the encrypted-cipher keyframe is almost zero. Produced result expressions of Table 11 including visual histogram representation at Figures 7–9, effectively validate the uppermost superiority of breaching the correlation link amongst the adjoining pixel at the test images including extracted keyframe images through planned STC-IES methods and fairly comparison on Table 12 through former referenced encryption approaches such as [39, 50, 66, 75, 84, 85, 87]. The results and comparison prove that the proposed STC-IES methodology is exceedingly unaffected to any sorts of statistical attacks into smart healthcare IIoT-enabled system.

5.6. Analysis of Produced Key. Configuration among the chaotic seed maps is susceptible towards preliminary surroundings. A cryptographic scheme is classified of superior quality or appropriate key size if it has sufficiently computational complexity with such a heightened sensitivity to modify the key. The proposed STC-IES has 2^{256} key size means 256-bit long proportion. That encounters the key performance requirements and is exceedingly operative in sidestepping dissimilar sort of security attacks [40, 56, 88]. Additionally, the proposed methodology, including key structure, behaves extremely sensitive to design architecture in five blocks. The STC-IES methodology articulated key size in Table 13 and equitably comparison with the past referenced encryption algorithms such as [37, 65, 71, 89–93]. It endorses that the proposed key size delivers a reasonably healthier variety of the key space to engender more complex chaotic actions. Consequently, STC-IES has the satisfactory key size to evade all the possibility of brute force attacks.

5.7. Comparative Analysis among the Monitoring and Surveillance System. This subdivision is comparatively based on the proposed methodology with the heretofore referenced surveillance as well as image encryption algorithms in Table 14. Table 14 expressively illustrates all the key characteristics in terms of security surveillance and encryption methodology that can prove secure and robustness constraint based on key analysis, encryption speeds, entropy analysis, correlation coefficients (CC_{xy}), NPCR and UACI produced outcomes. The outcomes of the proposed

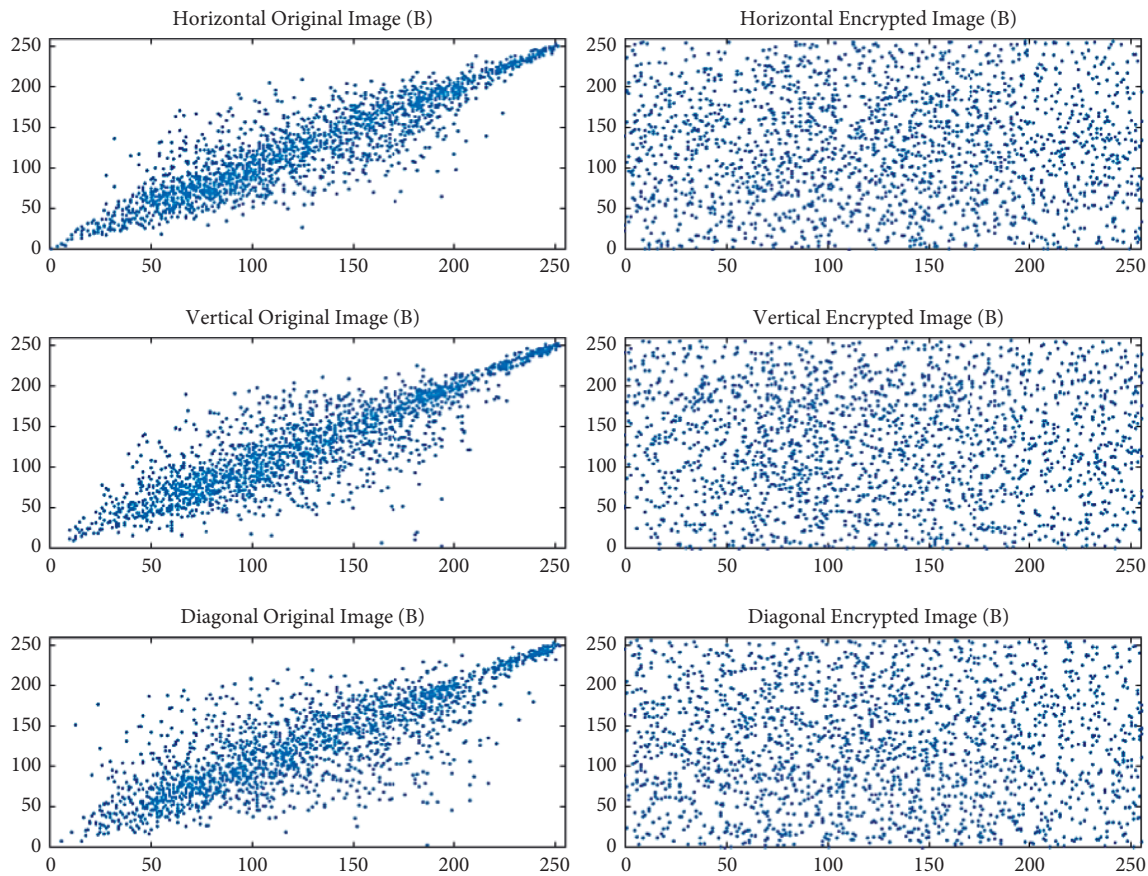


Figure 9: Distributed very close adjacent pixel (horizontal, vertical, and diagonal planes) at keyframe as well as encrypted keyframe (blue channel of a baboon).

Table 12: CCs of two bordering pixels and its comparison through relative approaches.

Methods	Images		Plain images			Encrypted images		
			H	*V*	*D*	*H*	*V*	*D*
Proposed	P002 (keyframe 2)	R	0.9578	0.9905	0.9511	2.3e-05	−4.3e-04	0.0032
		G	0.9560	0.9900	0.9505	0.0028	−9.1e-04	0.0033
		B	0.9570	0.9901	0.9500	5.2e-07	−3.3e-04	0.0011
[39]	RGB frame	R	0.99370	0.99010	0.98540	0.00120	−0.00270	0.00020
		G	0.99090	0.98340	0.98770	−0.00070	0.00210	−0.00100
		B	0.99310	0.98240	0.99010	0.00150	−0.00100	0.00070
[50]	Keyframe 5		0.9439	0.9880	0.9384	0.0051	−0.0033	0.0015
[84]	Jet plane		NA	NA	NA	0.0087	-0.0051	−0.0089
[66]	Baboon		0.9225	0.7860	0.8535	0.0012	0.0038	0.0011
[85]	Lena		0.967841	0.976007	0.954908	−0.001557	0.002225	0.000582
[86]	NA		0.9824	0.9632	0.9484	0.0016	0.0015	−0.0017
[73]	Airplane		0.9518	0.9751	0.9422	−0.0049	0.0019	0.0010
[75]	Lena		NA	NA	NA	0.0022	0.0022	−0.0019

methodology are fairly reasonable as well as nearly ideal standards. We compared proposed methods with past referenced schemes such as [20, 39, 41, 42, 50, 71]. We noticed that the referenced surveillance and encryption scheme is produced relative results to attain confidentiality of the security. We compared our results with numerous images employing diverse platforms and upbringing functionality, each restrained factor. The proposed methodology has comparatively better speed with minimal complexity in the execution, analogous better-quality entropy, lowermost correlation coefficient (CC_{xy}), satisfactory NPCR, and UACI statistics. Significantly improved results are resolutely signposted. That the proposed methodology has exceedingly adequate in the arena of smart healthcare IIoT enabled setup with monitoring (surveillance) as well as cryptographically secure ecosystem.

Table 13: Comparative key space analysis.

Algorithm	STC-IES	[37]	[65]	[71]	[89]	[90]	[91]	[92]	[93]
Key space	2^{256}	3.9402×10^{185}	2^{168}	2^{192}	2^{128}	2^{128}	10^{128}	2^{280}	2^{128}

Table 14: Fair comparative discussion amongst the monitoring and surveillance architecture.

Algorithms	Image size	Key length	Speed	Entropies	CC_{xy}	NPCR	UACI
STC-IES	512×512 [3]	2^{256}	0.2205–0.2541	7.9989	0.0011	99.6536	33.4086
[20]	512×512 [3]	10^{90}	0.6708	7.9998	0.0035	99.6125	33.4451
[39]	640×480 [3]	2^{372}	0.95/0.96	7.9994	0.0021	99.609	33.465
[41]	Keyframe 0065	2^{711}	2.58	7.9998	0.0019	99.609	33.450
[42]	1024×1024 [3]	2^{300}	NA	7.91	0.003	99.5826	33.4213
[50]	512×512 [3]	2^{256}	0.2811–0.3119	7.9991	0.0015	99.6212	33.4406
[71]	Keyframe 640×480 [3]	2^{192}	0.790	NA	0.0035	99.615	33.4658

6. Conclusion

The technological progression of the hybrid IoT ecosystem is referred to as the IIoT system. It is extensively organized with the industrial, medical healthcare system to deliver the finest services counting security and privacy about patients. The produced report formally focuses on security, privacy, and its challenges in smart setup healthcare IIoT architecture. We implemented secure monitoring (surveillance) with intelligently keyframe extraction and lightweight cosine function hybrid chaotic map encryption. It also assists, including security and privacy from the outside world or any adversary. At first, a well-disciplined model of keyframe extraction is employed with a lightweight YOLOv3 algorithm. This model is successfully operated and tested with a vast collection of image dataset Face Database to retrieve meaningful detected frames (normal/abnormal events) through the visual sensor. The average accuracy is received optimally around 90% with an acceptable frame per second rate. Second, a lightweight cosine function encryption is employed over approved extracted keyframe to remain exceptionally secure and safe from further attacks. The generated cosine function chaotic sequences are a non-linear transform to engender as well as exhibit expressively very complex chaos behavior. Our encryption methodology implies a better diffusion-confusion process of bitwise XOR operation and highly efficient scrambling algorithms, which are achieved to encrypt each color image channel separately and scatter neighboring pixels into different positions quickly. This proposed methodology ensured satisfactorily encrypted matrix as a cipher keyframe without identifying real keyframes into smart setup healthcare from adversarial threads.

Numerous security discussions and analysis results inveterate about the effectiveness of the proposed methodology. This has relatively higher security characteristics, and minimal computational processing agreed and contending with the earliest image encryption approaches. It also validates its accomplishment in the IIoT ecosystem with minimalizing bandwidth, storage space, communication cost, transmission expenses, and correspondingly waning time spending of specialists handling due to huge amount of monitoring data to take verdict over any such suspicious incidents detection or any suspicious action detection as an emergency need from the patients at smart setup healthcare architecture. The produced concept can also be utilized in many relatively analogous urgent responded real-time projects: traffic-control, fire-detection, crime-control, and smart transportation at smart cities.

For future concern, this approach can be supported to assimilate information from further systems, aimed at numerous applications and additional advanced security facet and privacy dealings in any exact areas within the healthcare sector. The innovative way can probably integrate dynamic key as an alternative applied process for additional ornamental security and privacy.

Data Availability

Previously reported are used to support this study and are available at [Wider-Face Data Set, Face Database (FDB), USC-SIPI Image Data Set Repository]. These prior studies (and datasets) are cited at relevant places within the text as references [51, 52, 54].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61370073), the National High Technology Research and Development Program of China (Grant No. 2007AA01Z423), and the project of Science and Technology Department of Sichuan Province. The authors are also acknowledging the generous support from the Research supporting Project number (RSP-2021/122), King Saud University, Riyadh, Saudi Arabia.

References

[1] R. A. Memon, J. P. Li, J. Ahmed, M. I. Nazeer, M. Ismail, and K. Ali, "Cloud-based vs. blockchain-based IoT: a comparative survey and way forward," *Frontiers of Information Technology & Electronic Engineering*, vol. 21, no. 4, pp. 563–586, 2020.

[2] X. Huang, "Intelligent remote monitoring and manufacturing system of production line based on industrial Internet of Things," *Computer Communications*, vol. 150, pp. 421–428, 2020.

[3] K. N. Qureshi, S. S. Rana, A. Ahmed, and G. Jeon, "A novel and secure attacks detection framework for smart cities industrial internet of things," *Sustainable Cities and Society*, vol. 61, p. 102343, 2020.

[4] B. Jiang, J. Li, G. Yue, and H. Song, "Differential privacy for industrial internet of things: opportunities, applications and challenges," *IEEE Internet of Things Journal*, vol. 8, p. 1, 2021.

[5] L. Nie, X. Wang, S. Wang et al., "Network traffic prediction in industrial internet of things backbone networks: a multi-task learning mechanism," *IEEE Transactions on Industrial Informatics*, vol. 17, pp. 7123–7132, 2021.

[6] M. Serror, S. Hack, M. Henze, M. Schuba, and K. Wehrle, "Challenges and opportunities in securing the industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, p. 1, 2020.

[7] T. Wang, H. Luo, W. Jia, A. Liu, and M. Xie, "MTES: an intelligent trust evaluation scheme in sensor-cloud-enabled industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2054–2062, 2020.

[8] S. Khan and M. Altayar, "Industrial Internet of things: investigation of the applications, issues, and challenges," *International Journal of Advances in Applied Sciences*, vol. 8, no. 1, pp. 104–113, 2021.

[9] R. A. Memon, J. P. Li, M. I. Nazeer, A. N. Khan, and J. Ahmed, "DualFog-IoT: additional fog layer for solving blockchain integration problem in internet of things," *IEEE Access*, vol. 7, pp. 169073–169093, 2019.

[10] J. Chi, Y. Li, J. Huang et al., "A secure and efficient data sharing scheme based on blockchain in industrial Internet of Things," *Journal of Network and Computer Applications*, vol. 167, p. 102710, 2020.

[11] J. Khan, J. P. Li, I. Ali et al., "An authentication technique based on oauth 2.0 protocol for internet of things (IoT) network," in *Proceedings of the 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 160–165, Chengdu, China, 14-16 Dec. 2018.

[12] J. Khan, H. Abbas, and J. Al-Muhtadi, "Survey on mobile user's data privacy threats and defense mechanisms," in *Procedia Computer Science*, 2015.

[13] S. Khan, "Modern Internet of Things as a challenge for higher education," *International Journal of Computer Science and Network Security*, vol. 18, no. 12, pp. 34–41, 2018.

[14] K.-K. R. Choo, S. Gritzalis, and J. H. Park, "Cryptographic solutions for industrial internet-of-things: research challenges and opportunities," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3567–3569, 2018.

[15] A. Alabdulatif, I. Khalil, X. Yi, and M. Guizani, "Secure edge of things for smart healthcare surveillance framework," *IEEE Access*, vol. 7, pp. 31010–31021, 2019.

[16] P. Sundaravadivel, K. Kesavan, L. Kesavan, S. P. Mohanty, and E. Kougianos, "Smart-log: a deep-learning based automated nutrition monitoring system in the IoT," *IEEE Transactions on Consumer Electronics*, vol. 64, no. 3, pp. 390–398, 2018.

[17] S. H. Alsamhi, O. Ma, M. S. Ansari, and F. A. Almalki, "Survey on collaborative smart drones and internet of things for improving smartness of smart cities," *IEEE Access*, vol. 7, pp. 128125–128152, 2019.

[18] P. Chanak and I. Banerjee, "Congestion free routing mechanism for IoT-enabled wireless sensor networks for smart healthcare applications," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 3, pp. 223–232, 2020.

[19] M. Shuai, L. Xiong, C. Wang, and N. Yu, "A secure authentication scheme with forward secrecy for industrial internet of things using Rabin cryptosystem," *Computer and Communications*, vol. 160, pp. 215–227, 2020.

[20] K. Muhammad, R. Hamza, J. Ahmad, J. Lloret, H. Wang, and S. W. Baik, "Secure surveillance framework for IoT systems using probabilistic image encryption," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3679–3689, 2018.

[21] M. Sajjad, I. Mehmood, and S. W. Baik, "Sparse representations-based super-resolution of key-frames extracted from frames-sequences generated by a visual sensor network," *Sensors*, vol. 14, no. 2, pp. 3652–3674, 2014.

[22] Z. Yang, B. Liang, and W. Ji, "An intelligent end-edge-cloud architecture for visual IoT assisted healthcare systems," IEEE Internet Things J., 2021.

[23] I. Mehmood, M. Sajjad, W. Ejaz, and S. W. Baik, "Saliency-directed prioritization of visual data in wireless surveillance networks," *Information Fusion*, vol. 24, pp. 16–30, 2015.

[24] S. Nazir, S. Khan, H. U. Khan et al., "A comprehensive analysis of healthcare big data management, analytics and scientific programming," *IEEE Access*, vol. 8, pp. 95714–95733, 2020.

[25] J. Ding, L. Jiang, and C. He, "User-centric energy-efficient resource management for time switching wireless powered communications," *IEEE Communications Letters*, vol. 22, no. 1, pp. 165–168, 2018.

[26] H. Yang, Y. Ye, X. Chu, and M. Dong, "Resource and power allocation in SWIPT-enabled device-to-device communications based on a nonlinear energy harvesting model," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 10813–10825, 2020.

[27] K. Muhammad, T. Hussain, and S. W. Baik, "Efficient CNN based summarization of surveillance videos for resource-constrained devices," *Pattern Recognition Letters*, vol. 130, pp. 370–375, 2020.

[28] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. De Albuquerque, "Cloud-assisted multiview video summarization using CNN and bidirectional LSTM," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 77–86, 2020.

[29] C. Huang and H. Wang, "A novel key-frames selection framework for comprehensive video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 577–589, 2020.

[30] J. Lei, Q. Luan, X. Song, X. Liu, D. Tao, and M. Song, "Action parsing driven video summarization based on reinforcement learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 2126–2137, 2018.

[31] Y. Yuan, T. Mei, P. Cui, and W. Zhu, "Video summarization by learning deep side semantic embedding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 226–237, 2019.

[32] J. Wang, K. Han, S. Fan et al., "A logistic mapping-based encryption scheme for Wireless Body Area Networks," *Future Generation Computer Systems*, vol. 110, pp. 57–67, 2020.

[33] A. Al-Khedhairi, A. Elsonbaty, A. A. Elsadany, and E. A. A. Hagras, "Hybrid cryptosystem based on pseudo chaos of novel fractional order map and elliptic curves," *IEEE Access*, vol. 8, pp. 57733–57748, 2020.

[34] Y.-Q. Zhang, J.-L. Hao, and X.-Y. Wang, "An efficient image encryption scheme based on S-boxes and fractional-order differential logistic map," *IEEE Access*, vol. 8, pp. 54175–54188, 2020.

[35] L. Huang, S. Cai, X. Xiong, and M. Xiao, "On symmetric color image encryption system with permutation-diffusion simultaneous operation," *Optics and Lasers in Engineering*, vol. 115, no. July 2018, pp. 7–20, 2019.

[36] I. T. Almalkawi, R. Halloush, A. Alsarhan, A. Al-Dubai, and J. N. Al-karaki, "A lightweight and efficient digital image encryption using hybrid chaotic systems for wireless network applications," *J. Inf. Secur. Appl.*vol. 49, p. 102384, 2019.

[37] X. Chai, X. Fu, Z. Gan, Y. Lu, and Y. Chen, "A color image cryptosystem based on dynamic DNA encryption and chaos," *Signal Processing*, vol. 155, pp. 44–62, 2019.

[38] X. Y. Wang and Z. M. Li, "A color image encryption algorithm based on Hopfield chaotic neural network," *Optics and Lasers in Engineering*, vol. 115, pp. 107–118, 2019.

[39] R. Hamza, Z. Yan, K. Muhammad, P. Bellavista, and F. Titouna, "A privacy-preserving cryptosystem for IoT E-healthcare," *Information Science*, vol. 527, pp. 493–510, 2019.

[40] Z. Hua, Y. Zhou, and H. Huang, "Cosine-transform-based chaotic system for image encryption," *Information Science*, vol. 480, pp. 403–419, 2019.

[41] R. Hamza, K. Muhammad, Z. Lv, and F. Titouna, "Secure video summarization framework for personalized wireless capsule endoscopy," *Pervasive and Mobile Computing*, vol. 41, pp. 436–450, 2017.

[42] R. Hamza, A. Hassan, T. Huang, L. Ke, and H. Yan, "An efficient cryptosystem for video surveillance in the internet of things environment," *Complexity*, vol. 2019, pp. 1–11, 2019.

[43] M. Kaur, D. Singh, K. Sun, and U. Rawat, "Color image encryption using non-dominated sorting genetic algorithm with local chaotic map," *Future Generation Computer Systems*, vol. 107, pp. 333–350, 2020.

[44] A. Broumandnia, "The 3D modular chaotic map to digital color image encryption," *Future Generation Computer Systems*, vol. 99, pp. 489–499, 2019.

[45] B. Mondal, S. Singh, and P. Kumar, "A secure image encryption scheme based on cellular automata and chaotic skew tent map," *J. Inf. Secur. Appl.*vol. 45, pp. 117–130, 2019.

[46] J. Khan, J. Li, A. Khan, G. A. Khan, and S. Ahmad, "Efficient secure surveillance on smart healthcare IoT system through cosine-transform encryption," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 1, pp. 1417–1442, 2021.

[47] A. U. Haq, J. P. Li, J. Khan et al., "Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data," *Sensors*, vol. 20, no. 9, p. 2649, 2020.

[48] L. Gong, K. Qiu, C. Deng, and N. Zhou, "An image compression and encryption algorithm based on chaotic system and compressive sensing," *Optics & Laser Technology*, vol. 115, pp. 257–267, 2019.

[49] J. Wu, B. Cheng, M. Wang, and J. Chen, "Energy-aware concurrent multipath transfer for real-time video streaming over heterogeneous wireless networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 2007–2023, 2018.

[50] J. Khan, J. P. Li, B. Ahamad et al., "SMSH: secure surveillance mechanism on smart healthcare IoT system with probabilistic image encryption," *IEEE Access*, vol. 8, pp. 15747–15767, 2020.

[51] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," *Computer Vision and Pattern Recognition*, 2018.

[52] J. Redmon, *Darknet Open Source Neural Network Framework*, [Online]. Available: https://github.com/pjreddie/darknet Accessed, 2016.

[53] J. Li, J. Gu, Z. Huang, and J. Wen, "Application research of improved YOLO V3 algorithm in PCB electronic component detection," *Applied Sciences*, vol. 9, no. 18, 2019.

[54] T. C. U. of H. K. Multimedia Laboratory, Department of Information Engineering, *WIDER FACE: A Face Detection Benchmark*, [Online]. Available: http://shuoyang1213.me/WIDERFACE/ Accessed.

[55] K. D. Mislav Grgic, *FACE RECOGNITION HOMEPAGE.*[Online]. Available: http://www.face-rec.org/databases/ Accessed.

[56] G. Alvarez and S. Li, "Some basic cryptographic requirements for chaos-based cryptosystems," *Int. J. Bifurc. Chaos*, vol. 16, no. 08, pp. 2129–2151, 2006.

[57] Digital Test Image, *Usc-sipi Image Database for Research in Image Processing, Image Analysis, and Machine Vision*, [Online]. Available: http://sipi.usc.edu/database/ Accessed, 2021.

[58] K. A. Kumar Patro and B. Acharya, "An efficient colour image encryption scheme based on 1-D chaotic maps," *J. Inf. Secur. Appl.*vol. 46, pp. 23–41, 2019.

[59] M. Alawida, A. Samsudin, J. Sen Teh, and R. S. Alkhawaldeh, "A new hybrid digital chaotic system with applications in image encryption," *Signal Processing*, vol. 160, pp. 45–58, 2019.

[60] P. Ping, F. Xu, Y. Mao, and Z. Wang, "Designing permutation–substitution image encryption networks with Henon map," *Neurocomputing*, vol. 283, pp. 53–63, 2018.

[61] A.-V. Diaconu, "Circular inter–intra pixels bit-level permutation and chaos-based image encryption," *Information Science*, vol. 355–356, pp. 314–327, 2016.

[62] D. Kumar, A. B. Joshi, and V. N. Mishra, "Optical and digital double color-image encryption algorithm using 3D chaotic map and 2D-multiple parameter fractional discrete cosine transform," *Results Opt*, vol. 1, p. 100031, 2020.

[63] M. A. Khan, J. Khan, A. A. Alghamdi, and S. M. A. B. Saidan, "Image encryption enabling chaotic ergodicity with logistic and sine map," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, pp. 437–442, 2021.

[64] X. Chai, Y. Chen, and L. Broyde, "A novel chaos-based image encryption algorithm using DNA sequence operations," *Optics and Lasers in Engineering*, vol. 88, pp. 197–213, 2017.

[65] X. Chai, J. Bi, Z. Gan, X. Liu, Y. Zhang, and Y. Chen, "Color image compression and encryption scheme based on compressive sensing and double random encryption strategy," *Signal Processing*, vol. 176, p. 107684, 2020.

[66] Z. Liu, C. Wu, J. Wang, and Y. Hu, "A color image encryption using dynamic DNA and 4-D memristive hyper-chaos," *IEEE Access*, vol. 7, pp. 78367–78378, 2019.

[67] X. Kang, X. Luo, X. Zhang, and J. Jiang, "Homogenized Chebyshev-Arnold map and its application to color image encryption," *IEEE Access*, vol. 7, pp. 114459–114471, 2019.

[68] D. Ravichandran, P. Praveenkumar, J. B. B. Rayappan, and R. Amirtharajan, "DNA chaos blend to secure medical privacy," *IEEE Transactions on NanoBioscience*, vol. 16, no. 8, pp. 850–858, 2017.

[69] A. A. Abd El-Latif, B. Abd-El-Atty, E. M. Abou-Nassar, and S. E. Venegas-Andraca, "Controlled alternate quantum walks based privacy preserving healthcare images in Internet of Things," *Optics & Laser Technology*, vol. 124, p. 105942, 2020.

[70] A. Shakiba, "A randomized CPA-secure asymmetric-key chaotic color image encryption scheme based on the Chebyshev mappings and one-time pad," *J. King Saud Univ. - Comput. Inf. Sci.*vol. 33, no. 5, pp. 562–571, 2021.

[71] R. Hamza, K. Muhammad, A. Kumar, and G. Ramirez-Gonzalez, "Hash based encryption for keyframes of diagnostic hysteroscopy," *IEEE Access*, vol. 6, pp. 60160–60170, 2018.

[72] X. Kang, A. Ming, and R. Tao, "Reality-preserving multiple parameter discrete fractional angular transform and its application to color image encryption," *IEEE Transactions on*

Circuits and Systems for Video Technology, vol. 29, no. 6, pp. 1595–1607, 2019.

[73] X. Ouyang, Y. Luo, J. Liu, L. Cao, and Y. Liu, "A color image encryption method based on memristive hyperchaotic system and DNA encryption," *International Journal of Modern Physics B*, vol. 34, no. 04, p. 2050014, 2020.

[74] Y. Wu, J. P. Noonan, and S. Agaian, "NPCR and UACI randomness tests for image encryption," *Cyberjournals.Com*, 2011.

[75] D. Herbadji, A. Belmeguenai, N. Derouiche, and H. Liu, "Colour image encryption scheme based on enhanced quadratic chaotic map," *IET Image Processing*, vol. 14, no. 1, pp. 40–52, 2020.

[76] A. El Halabi, A. Hachem, L. Al-Akhrass, H. Artail, and H. U. Khan, "Identifying the linkability between Web servers for enhanced Internet computing," in *Proceedings of the MELECON 2014 - 2014 17th IEEE Mediterranean Electrotechnical Conference*, pp. 1–5, Beirut, Lebanon, 13-16 April 2014.

[77] H. Khan and M. N. Faisal, "A Grey-based approach for ERP vendor selection in small and medium enterprises in Qatar," *International Journal of Business Information Systems*, vol. 19, no. 4, p. 465, 2015.

[78] O. A. Bankole, V. V. M. Lalitha, H. U. Khan, and A. Jinugu, "Information Technology in the maritime industry past, present and future: focus on LNG carriers," in *Proceedings of the 2017 IEEE 7th International Advance Computing Conference (IACC)*, pp. 759–763, Hyderabad, India, 5-7 Jan. 2017.

[79] H. U. Khan, V. V. M. Lalitha, and J. F. Omonaiye, "Employees' perception as internal customers about online services: a case study of banking sector in Nigeria," *International Journal of Business Innovation and Research*, vol. 13, no. 2, p. 181, 2017.

[80] V. F. Brock and H. U. Khan, "Are enterprises ready for big data analytics? A survey-based approach," *International Journal of Business Information Systems*, vol. 25, no. 2, p. 256, 2017.

[81] Y. Zhang, X. Wang, L. Liu, and J. Liu, "Fractional order spatiotemporal chaos with delay in spatial nonlinear coupling," *Int. J. Bifurc. Chaos*, vol. 28, no. 02, p. 1850020, 2018.

[82] A. U. Rehman, H. Wang, M. M. A. Shahid, S. Iqbal, Z. Abbas, and A. Firdous, "A selective cross-substitution technique for encrypting color images using chaos, DNA rules and SHA-512," *IEEE Access*, vol. 7, pp. 162786–162802, 2019.

[83] J. Zhang, D. Fang, and H. Ren, "Image encryption algorithm based on DNA encoding and chaotic maps," *Mathematical Problems in Engineering*, vol. 2014, pp. 1–10, 2014.

[84] L. Li, G. Wen, Z. Wang, and Y.-X. Yang, "Efficient and secure image communication system based on compressed sensing for IoT monitoring applications," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 82–95, Jan. 2020.

[85] A. Mansouri and X. Wang, "A novel one-dimensional chaotic map generator and its application in a new index representation-based image encryption scheme," *Information Science*, vol. 563, pp. 91–110, 2021.

[86] F. Yang, J. Mou, K. Sun, Y. Cao, and J. Jin, "Color image compression-encryption algorithm based on fractional-order memristor chaotic circuit," *IEEE Access*, vol. 7, pp. 58751–58763, 2019.

[87] L. Li, G. Wen, Z. Wang, and Y.-X. Yang, "Efficient and secure image communication system based on compressed sensing for IoT monitoring applications," *IEEE Transactions on Multimedia*, p. 1, 2019.

[88] P. Nigel, *Smart, "Algorithms, Key Size and Protocols Report (2018),"*, ECRYPT – CSA, [Online]. Available: http://www.ecrypt.eu.org/csa/documents/D5.4-FinalAlgKeySizeProt.pdf Accessed, 2018.

[89] X. Chai, Z. Gan, K. Yang, Y. Chen, and X. Liu, "An image encryption algorithm based on the memristive hyperchaotic system, cellular automata and DNA sequence operations," *Signal Processing: Image Communication*, vol. 52, pp. 6–19, 2017.

[90] A. Yaghouti Niyat, M. H. Moattar, and M. Niazi Torshiz, "Color image encryption based on hybrid hyper-chaotic system and cellular automata," *Optics and Lasers in Engineering*, vol. 90, pp. 225–237, 2017.

[91] J. Tamang, "Dynamical properties of ion-acoustic waves in space plasma and its application to image encryption," *IEEE Access*, vol. 9, pp. 18762–18782, 2021.

[92] X. Chai, X. Zhi, Z. Gan, Y. Zhang, Y. Chen, and J. Fu, "Combining improved genetic algorithm and matrix semi-tensor product (STP) in color image encryption," *Signal Processing*, vol. 183, p. 108041, 2021.

[93] H. Nematzadeh, R. Enayatifar, H. Motameni, F. G. Guimarães, and V. N. Coelho, "Medical image encryption using a hybrid model of modified genetic algorithm and coupled map lattices," *Optics and Lasers in Engineering*, vol. 110, pp. 24–32, 2018.

Hindawi
Scientific Programming
Volume 2021, Article ID 9948533, 13 pages
https://doi.org/10.1155/2021/9948533

Research Article

The Embedded IoT Time Series Database for Hybrid Solid-State Storage System

Tao Cai, Peiyao Liu, Dejiao Niu, Jiancong Shi, and Lei Li

School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China

Correspondence should be addressed to Tao Cai; caitao@ujs.edu.cn and Dejiao Niu; djniu@ujs.edu.cn

Received 25 March 2021; Revised 28 August 2021; Accepted 12 October 2021; Published 25 October 2021

Academic Editor: Shah Nazir

IoT time series data is an important form of big data. How to improve the efficiency of storage system is crucial for IoT time series database to store and manage massive IoT time series data from various IoT devices. Mixing NVM and SSD is an effective method to improve the I/O performance of storage systems. However, there are great differences between HDD and NVM or SSD. As a result, NVM and SSD cannot be directly used in the current time series database effectively. We design an IoT time series database with an embedded engine in storage device drivers for the hybrid solid-state storage system consisting of NVM and SSD. The I/O software stack of storing and managing IoT time series data can be shortened to improve the efficiency. Based upon the intrinsic characteristics of IoT time series data and different features of NVM and SSD, a redundancy elimination and compression fusion strategy, a hierarchical management strategy, and a heterogeneous time series data index are designed to improve the efficiency. Finally, a prototype of embedded IoT time series database named TS-NSM is implemented, and YCSB-TS is used to measure the IOPS. The results show that TS-NSM can improve the write IOPS up to 243.6 times and 174.3 times, respectively, compared with InfluxDB and OpenTSDB, and improve the read IOPS up to 10.1 times and 14.4 times, respectively.

1. Background

Time series data is the sequential data with time correlation, commonly generated from social networks, scientific experiments, Internet of Things (IoT), and log systems. This is an important form of big data. The continuous generation, storing, and processing are the main characteristics of time series data. Especially in IoT, a large number of IoT devices concurrently generate a large amount of time series data with a fixed time interval, which brings great challenges to computer system for storing and processing [1]. The storage wall problem is a serious problem for IoT with HDD due to low read and write speed. In recent years, flash-based SSD with large capacity has become more popular, which has the I/O performance advantage compared with HDD. At the same time, nonvolatile memory (NVM) [2], such as Phase Change Memory (PCM) [3], Shared Transistor Technology Random Access Memory (STT-RAM) [4], and the latest technology Intel 3D-Xpoint [5] and Intel Optane DC Persistent Memory [6], provides features such as byte-addressable, long life span, low dynamic energy consumption, and high I/O speed closing to Dynamic Random Access Memory (DRAM). Then, a hybrid solid-state storage system with SSD and NVM can offer a possible way to solve the storage wall of time series data for IoT. However, the SSD and NVM also bring huge challenges to the I/O stack of computer system. The researches have shown that 94% delay in NVM storage systems is caused by file system, general block layer, etc. [7]. Therefore, it is of crucial importance to design a native IoT time series data management engine for hybrid solid-state storage systems.

Current DBMS is devoted to ensure transaction in data processing and is difficult to meet the requirements of time series data storage and management due to the lack of optimization strategies for continuous, high-speed, and concurrent massive data management [8]. Particularly in IoT, there are a large number of time series data generated by many different IoT devices in the meantime, which further increases the difficulty to store and manage IoT time series data with current DBMS. Accordingly, the time series

database is developed based on the current DBMS, which accords with the characteristics of time series data and optimizes some strategies such as data compression and timeliness. Then, it can continuously and concurrently store and manage massive time series data. Regarding IoT time series database, some special features are as follows: (1) continuous high-speed writing. As a large number of IoT devices expose multiple time series data, the IoT time series database needs to continuously store data with high throughput; (2) fast concurrent reading: each IoT device is relatively independent, so the IoT time series database should have high concurrent query capability to meet the different characteristics of each IoT device; (3) data compression: although the amount of IoT time series data is very large, the value often remains unchanged or changes very little over a period of time by using IoT devices. Then, data compression is very useful and critical for the IoT time series database; (4) time range queries: the query within a certain period of time is an important task of the IoT system; (5) self-management over time: IoT time series data has distinct timeliness, and the data access frequency will decrease sharply over time. It is important for the automatically adjustment strategy to store and index. Nevertheless, the present time series database management system is generally implemented by adding a time series management mechanism on the basis of common DBMS, which further increases the complexity of the I/O software stack for IoT time series data storage and management and cannot take advantage of the high I/O speed of SSD and NVM. There are also some differences between SSD and NVM. In spite of lower read and write speed, SSD just uses the block interface, and NVM is byte-addressable. Meanwhile, the current time series databases are designed for block or byte access interface storage devices, but they do not have the optimization strategy for the hybrid solid-state storage system composed of SSD and NVM. NVM's high read and write speed advantage cannot be used to improve the throughput of the current time series database due to the complex I/O software stack by using NVM instead of storage devices directly [9]. In addition, the write times of SSD and NVM are limited. In particular, a single unit in TLC-based SSDs just can write several hundred times. Meanwhile, there are a large number of the same or similar values in IoT time series data, which is very easy and necessary to do some optimization to reduce write amplification. IoT time series data also has different characteristics compared with time series data in social networks. All IoT time series data should be stored, and there is no deletion or update. At the same time, the time interval of series data is relatively fixed, and the data format is regular, which brings some convenience for IoT time series data management. However, IoT is also more time-sensitive for time series data storage and management, which poses a high challenge to the storage system. Thus, an important and urgent issue is how to design a new IoT time series database for the solid-state storage system composed of SSD and NVM.

In this paper, a new IoT time series data management engine is designed and embedded into the device driver of the hybrid solid-state storage system constituted by SSD and NVM. The NVM is used as a fast storage device in our design. The summary of our contributions is as follows:

(1) The IoT time series data storage and management engine is embedded in the storage device driver, which shortens the I/O software stack of IoT time series data management based on the hybrid solid-state storage system and avoids frequent exchange of large amounts of data between the host and the storage system. It can better utilize the high-speed read and write capabilities of NVM and SSD to improve the efficiency of IoT time series data management.

(2) The redundancy elimination and compression fusion strategy is designed according to the characteristics of IoT time series data and the hybrid solid-state storage system, which can reduce the storage consumption and the wear of SSD and NVM.

(3) The hierarchical structure is made by the access and management characteristics of IoT time series data, which can combine the advantages of NVM and SSD to improve the economic and scalability of the storage system.

(4) The two-dimensional hybrid index is designed for NVM, which combines hash and ordered linked list to improve the I/O performance and efficiency of IoT time series database.

(5) Block note index is designed to improve the query efficiency of IoT time series data stored in SSD.

(6) The prototype of an embedded IoT time series database for hybrid solid-state storage systems named TS-NSM is implemented to be tested and compared with the current popular time series databases such as InfluxDB and OpenTSDB. The results showed that TS-NSM can improve the write IOPS up to 243.6 and 201.2 times, respectively, and improve read IOPS up to 10.1 and 14.1 times, respectively.

2. Related Work

NVM has high I/O speed and byte-addressable interface, which can be used to build the hybrid solid-state storage system with SSD. There are many researches of hybrid solid-state storage system and database optimization for NVM.

2.1. Hybrid Storage System Based on NVM. NVM can be used to construct a hybrid memory with DRAM, and it also has a hybrid storage with HDD or SSD. Peiquan Lin constructed a hybrid memory with PCM and DRAM, where DRAM was mainly used as the cache of PCM to improve the lifetime of PCM [10]. A double dynamic bucket linked list was used to manage the space of PCM, and the age-based lazy cache strategy was designed to manage the cache in DRAM. Hibachi proposed a collaborative hybrid cache strategy based on NVM and DRAM [11]. The read and write cache were separate, and different management mechanism was designed for them to improve the hit rate. Meanwhile, the page dynamic adjustment mechanism was to adjust the size

of clean and dirty cache for different workloads. Chen et al. designed an efficient KV storage engine named FlatStore for the hybrid memory constructed by NVM and DRAM [12]. The log was stored in NVM to ensure reliable of storage, and DRAM was used to store the index to improve the search efficiency. An OpLog for each core was proposed to cache the small KV pairs, and the pipeline horizontal batch processing mechanism was designed to improve throughput and reduce read latency. LosPem is a novel log-structure framework for persistent memory constituted by NVM and DRAM to address the performance challenge [13]. It deployed an efficient hash-index linked list to maintain the log contents and reduce the significant overhead of log content retrieval. In addition, LosPem improved the transaction throughput by decoupling a transaction into two asynchronous steps and creating a write buffer based on DRAM to cache frequent data writes. In terms of NVM hybrid storage. The NVMCFS constructed a hybrid storage system with NVM and SSD [14], which used the head-tail layout and space management based on two-layer radix tree to provide unified logic space between two type NVM devices and used complex file structures, dynamic file data distributed strategy, buffer for an individual file, and asymmetric call in strategy to speed up the access response and improve I/O performance. Strata are a file system for the hybrid NVM storage system [15], which utilized the byte addressable ability of NVM to merge logs and migrated them to the underlying SSD/HDD to minimize the write amplification. However, file data can only be allocated in NVM and be migrated from the NVM layer to the SSD/HDD layer. Ziggurat is a multitiered NVM-based file system that spans NVM and HDD [16]. It is based on NOVA. Ziggurat exploited the benefits of NVM through intelligent data placement during file write and data migration. Ziggurat included two placement predictors that analyze the file write sequences to predict whether the incoming writes were both large and stable and whether their update to the file is likely to be synchronous. Then, it steered the incoming writes to the most suitable tier based on the prediction and writes to synchronously updated files go to the NVM tier to minimize the synchronization overhead. Small random writes also go to the NVM tier to entirely avoid random writes to HDD. The remaining large sequential writes to asynchronously updated files go to HDD. vNVML addressed the problem of combining a smaller, faster byte-addressable NVM with a larger and slower storage device, such as SSD, to create the impression of a larger and faster byte-addressable NVM, which can be shared across multiple applications concurrently [17]. vNVML provided application transaction-like memory semantics that can ensure write ordering, durability, and persistency guarantees across system failures. vNVML exploited DRAM for read caching to improve performance and potentially to reduce the number of writes to NVM and extend the NVM lifetime, whereas these new technologies still have some problems when used in the current database systems [18]. For example, disk-oriented database systems (such as Oracle RDBMS, IBM DB2, and MySQL) use block devices for persistent storage, which need to maintain an in-memory cache for tuple blocks and try to maximize sequential reads and writes to HDD with the bad performance of random access. As for memory-oriented database systems (such as VoltDB and MemSQL), they contain certain components to overcome the volatility of DRAM, while, in a byte-addressable NVM system with fast random access, such components are unnecessary.

2.2. Database Management System Based on NVM. NVM can provide features such as persistent storage, byte-addressable, low dynamic energy consumption, and high I/O speeds close to DRAM, yet the traditional database management systems (DBMS) cannot make full use of this technology, because their internal architecture is designed for DRAM or HDD. If NVM is used directly to replace DRAM or HDD, many of the components in these database systems are unnecessary and will reduce the efficiency of data-intensive applications based on it. Therefore, researchers have reconsidered the data storage and management methods in both volatile memory and persistent storage in the current DBMS. For example, BzTree is a lock-free B-tree index structure for NVM [19], and it replaced the index structures such as the black-and-white tree and the jump table in memory databases. BzTree mainly optimized PMwCAS in memory by using the NVM to greatly reduce the complexity of the traditional single-word CAS-based lock-free index, to improve the indexing efficiency of the database. Arulraj implemented six engines based on different storage management architectures in the modular DBMS test platform [9]; they are in-place update (InP), copy-on-write update (CoW), and log structure update (Log), and their variants on NVM such as NVM-InP, NVM-CoW, and NVM-Log; these variants mainly remove or modify unnecessary protection mechanisms. Their test results indicated that the three variant engines have better performance than the basic structure, and NVM-InP engine achieved the best throughput with the least amount of wear on the NVM device. At the same time, they found that the NVM access latency has the most impact on the runtime performance of the engine, more than the workload skew or the number of modifications to the database in the workload. Wang et al. presented a passive group commit method for a distributed logging protocol extension of Shore-MT [20, 21]. Rather than issuing a barrier for every processor when committing, the DBMS is tracked when all records required to ensure the durability of a transaction are flashed to NVM and allocated log buffers from local memory, so as to avoid remote DRAM access. It makes transaction-level partitioning more advantageous. This work is based on the Shore-MT engine, which means that the DBMS records page-level undo logs before performing in-place updates and will result in high data duplication. Shimin et al. described the unique characteristics of PCM and analyzed their potential impact on database management system [22]. It particularly puts forward analytic metrics for PCM endurance, energy, and latency and illustrated that current approaches for common database index algorithms such as B+-trees and Hash Joins are suboptimal for PCM. And then it introduced a new B+-tree and hash join, which has reduced both execution time

on PCM and write time. wB+Tree focused on the performance overhead of NVM write and CPU cache refresh to design the write atomic B+ tree [23], which resulted in reducing the number data to store in NVM by using an indirection slot array to minimize the movement of index entries and adopted a redo-only logging algorithm to ensure consistency. HiKV is a KV storage engine based on NVM [24], which established and retained a hash index in NVM to maintain its inherent fast index search capabilities and built a B+ tree index in DRAM to support range queries, thereby avoiding long-term NVM write and guaranteeing the index consistency. RangeKV is a persistent KV storage engine based on RocksDB, which is built on a heterogeneous storage architecture [25]. It used RangeTab in NVM to manage L0 layer data and increased L0 capacity to reduce LSM tree level and compression time overhead. RangeKV prebuilt a hash index of RangeTab data to reduce NVM access time and used a double buffer structure to reduce LSM tree write amplification from compression. HiLSM is the KV storage engine of the hybrid storage system with NVM and SSD [26]. According to the features of hybrid storage mediums, HiLSM adopted the hybrid data structure consisting of the log-structured memory and the LSM-tree. It proposed a fine-grained data migration strategy, a multithreaded data migration strategy, and a data filter strategy to address the issues such as write stalls in write intensive scenario, the performance gap between NVM and SSD, and the LSM-tree's inherent issue of write amplification. However, the optimization of database management systems for NVM is only limited to relational databases and KV store. The emerging time series database is still designed for DRAM or HDD, so their internal architecture still cannot fully utilize the advantages of NVM.

2.3. Time Series Database Management System. In order to store and analyze massive time series data, researchers have developed the special time series management system (TSMS or TSDBMS) to overcome the limitations of using general DBMS to manage the time series data [8]. Log-Structured Merge Tree (LSM-Tree) is very popular in the current time series database management systems [27], such as KairosDB [28] and OpenTSDB [29]. Meanwhile, InfluxDB [30] storage engine TSM-Tree is also the optimization based on LSM-Tree. LSM-Tree is designed for block storage engine. Its advantage is the great write performance, while its disadvantages are the poor read performance and write amplification. It used an update log to convert random write operations into sequential write and takes full advantage of the sequential writing of HDD, but it must reorder the written data and compress it layer by layer to reduce the cost of reading. This operation caused more serious write amplification. There are many other improvements in time series database management system. Yagoubi et al. mentioned a distributed parallel indexing method for time series databases [31], which used the correlation between ordered dimensions and adjacent values to achieve high scalability when having high performance of similarity query processing. Gorilla is an in-memory distributed time series database [32]. It proposed a three-level in-memory data structure, which is a shared index in memory, which can achieve high throughput of searching. However, due to the volatility and high cost of memory, it must use HBase to regularly back up and map nodes to ensure its availability in the case of single node or regional failures, and it takes delta-of-delta and XOR-based compression mechanism to minimize memory space overhead. There are some time series database management systems based on HDD as well. To overcome the problems of slow read and write speeds of HDD, they have designed a series of optimization mechanisms. For instance, LittleTable is a distributed relational database optimized for time series data [1]. To improve write performance, LittleTable spends more than half of the time on seeking data in the tablets for flushing the data into the HDD at once time together. Therefore, it requires a large amount of buffer size. To enhance query efficiency, LittleTable used two-dimensional clustering of relational tables. One is to divide the rows by timestamp, so the latest time series data can be retrieved quickly. And the other is to divide the keys by hierarchical structure, so that each partition can be further sorted to realize a fast indexing of time series data and flexible data table model optimization. However, LittleTable's consistency and durability guarantees were weak, and there is no batch delete function, because its underlying layer is still a relational database storage engine. ModelarDB is a general-purpose modular distributed time series database management system [33], which stores time series data as a model; hence, all operations are optimized on the model's mode, and multiple models can dynamically adapt to the data set. ModelarDB relies on Spark and Cassandra to manage time series data for accomplishing high-speed writing capabilities, data compression, and scalable online aggregation query processing capabilities. Nevertheless, the model-based design also limits ModelarDB to only be used for fixed-frequency time series, and the compression of time series data is lossy compression. Meanwhile, ModelarDB's write performance dropped sharply when there were many preselected models, as it needed try all models for each segment of the time series. Peregreen is a distributed modular time series database management system deployed in a cloud environment [34]. It is designed for great-scale historical time series data in the cloud. Peregreen takes a dual storage method that divides the time series data into segments and blocks and then merges them into the three-tier index. It can achieve efficient query and calculation with small network overhead. Peregreen also requires a large amount of buffer size to provide great write performance, since it bundles the timestamp and value into a pair of compressions. Accordingly, when there is a mixed time series, its compression ability is poor, resulting in the fact that the performance of reading data will be sharply reduced when using a remote storage device.

In summary, although the current researches can improve the performance of the hybrid NVM storage system and can improve the efficiency of relation database management system, they are not available time series data management system based on NVM. Existing time

series data management systems are designed for memory or HDDs. They lack data read, write, and management strategies designed for NVM, and an optimization mechanism that integrates block and byte interface storage devices. Therefore, they cannot better utilize the advantages of the hybrid solid-state storage system. In addition, the existing time series data management systems generally are based on the traditional database management systems, which also limits the efficiency owing to lacking of the native storage engine with the characteristics of time series data.

3. Analysis of IoT Time Series Database

When the current time series database is used in the IoT system, the problems are as follows:

(1) Lack of native IoT time series data management engine. Most of the time series databases are developed from general databases, such as OpenTSDB and TimescaleDB, which are based on Hbase and PostgreSQL, respectively. An extra interface layer for reading and writing of time series data is achieved for the general database. It not only failed to improve the storage and management efficiency on the basis of the characteristics of IoT time series data, but also increased the complexity of the I/O software stack for IoT time series databases. Then, it is difficult to take advantage of the fast read and write speed of SSD and NVM, and IoT time series databases become the bottleneck of IoT time series data management efficiency.

(2) Lack of optimization strategy for solid-state storage systems. Compared with HDD, SSD and NVM have higher read and write speeds. Especially for NVM, its read and write speeds are close to DRAM and support byte-addressable, which brings many challenges to the current storage system, such as their cache mechanism and merging small data into blocks. SSD and NVM also have limited write lifetime, and there are some differences in access interface and I/O performance between SSD and NVM. The current time series databases lack the corresponding optimization strategy to improve the IoT time data management efficiency by the respective advantages of SSD and NVM.

(3) Serious write waste. Different from general data, the values of IoT time series data will be the same or close to an adjacent point during a period of time, so it causes great write waste by writing them all to the storage device. Although there are some data compression mechanisms in the current time series database such as InfluxDB, KairosDB, and OpenTSDB, they still lack optimization for IoT time series data due to the difference between IoT and social networks. This will not only affect the efficiency of IoT time series database, but also reduce the lifetime of SSD and NVM.

(4) Weak concurrency for read and write IoT time series data. There are a large number of IoT devices in the IoT system, and each IoT device will continuously write data to the time series database according to its frequency. The current time series databases usually mix multiple time series sequence data in one file for storage and management and cannot optimize the storage and management efficiency by the different characteristics of IoT devices, such as frequency and value range.

(5) Lack of optimization strategies for timeliness. Compared with time series data from other systems, IoT time series data is more time-sensitive, which will bring obvious changes of access frequency in the time range, while the current time series databases such as InfluxDB, RRDTool, and OpenTSDB still use static index mechanism and cannot optimize the store and index efficiency by the access frequency change.

4. Embedded IoT Time Series Database

4.1. Architecture. The IoT time series data is various and large, so a high I/O performance and large capacity of storage system are needed. However, the IoT time series data has timeliness, the write is more frequent than read, and the access frequency changes over time. Then, the static strategy is not adaptable for the IoT time series data. A new embedded IoT time series database structure for hybrid solid-state storage systems was designed as shown in Figure 1.

The new system call is designed to bypass the file system and shorten the I/O software stack for the embedded IoT time series database. The embedded database engine is used to manage the IoT time series data in the device driver; it consists of the redundancy elimination and compression fusion module, the hierarchical management module, and the heterogeneous time series data index module. The redundancy elimination and compression fusion module is used to reduce the size and amount of data that should be stored in the hybrid solid-state storage system. The hierarchical management module is used to distribute the IoT time series data between SSD and NVM. In our design, NVM is used as a fast storage device. The heterogeneous time series data index module is used to index the data stored in SSD and NVM with different ways on the basis of their characteristics.

4.2. Redundancy Elimination and Compression Fusion Strategy. The IoT time series data has some special features compared with time series data in social networks. As shown in Figure 2, each IoT time series data includes the generation time, the value of IoT time series data, and the IoT device identifier. Therefore, each IoT time series data requires such a large storage space. However, the time interval of time series data from the same IoT device is fixed, and the value of IoT time series data always has the regularity that its value range is small and does not change in a period of time. These characteristics can be used to reduce the storage space of IoT

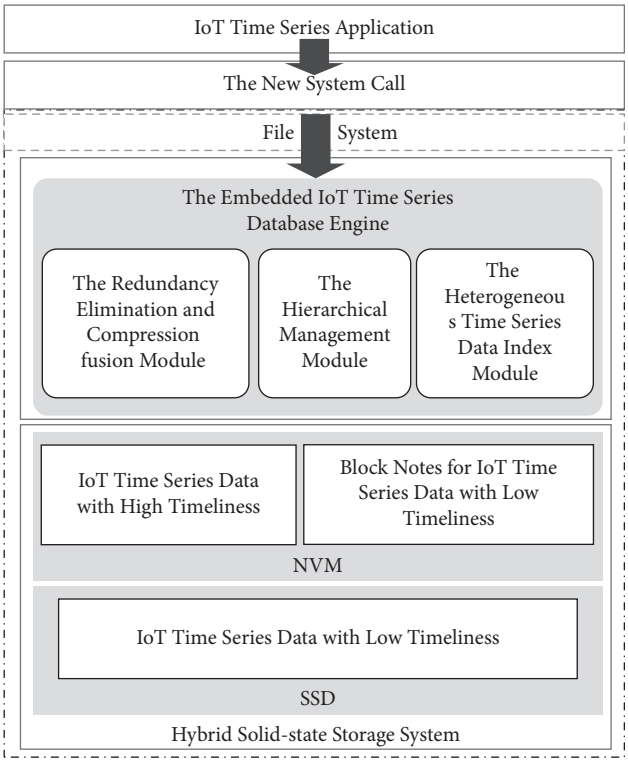


Figure 1: The structure of embedded IoT time series database for hybrid solid-state storage systems.

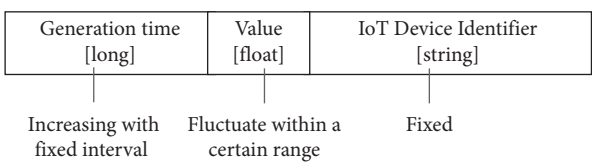


Figure 2: The original structure of an IoT time series data.

time series data. Then, a redundancy elimination and compression fusion strategy is designed.

Figure 3 shows the structure of the IoT time series data block. Each data block is 4 KB and consists of a block header area and a data area. The block header area includes start_time, start_value, p_num, last_point, and next_point, which, respectively, record the start time, the value of the first IoT time series data, repeat times of the first value, and pointers to the previous and next IoT time series data blocks. The data area includes 2032 short integers named D_0 to D_{2031}, which are used to store the compressed value of IoT time series data. In addition, the data in each IoT time series data block comes from the same IoT device.

On this basis, the redundancy elimination and compression rule is designed to reduce the size of the IoT time series data. Pseudocode 1 gives the flow of the rule.

When reading IoT time series data, the value can be restored according to formula (3) and formula (4) in 4.4.

By using the redundancy elimination and compression fusion strategy, most of the IoT time series data values represented by float numbers can be converted into a short integer for storage and thereby effectively will reduce the storage space consumption. Meanwhile, the first value of IoT time series data in each data block can be used to remove redundancy without complex calculation, which also can reduce the storage space consumption effectively. At the same time, the IoT time series data blocks still have good search performance, and the location of IoT time series data can be quickly obtained through simple calculations. In addition, it also can effectively improve the write performance of IoT time series data and reduce the consumption of storage device's lifetime.

4.3. Hierarchical Management Strategy. In IoT system, the access frequency of new IoT time series data is higher. The time interval and value range have differences for different IoT devices. On the basis of these characteristics, the hierarchical management strategy is designed.

Firstly, the storage space in NVM is divided into two parts, the IoT device area and data area. In the IoT device area, the structure shown in Figure 4 is used to store the features of each IoT device, which consists of IoT_device_ID, interval, value_precision, frequent_range, threshold, tags, block_address and start_time_SSD. IoT_device_ID is the IoT device identification. interval is the time interval of the IoT time series data. value_precision is the value precision of IoT time series data. frequent_range is the time domain of frequent query for each IoT device. *Threshold* is the migration threshold, which also means the maximum number of IoT time series blocks residing in NVM of each IoT device, calculated by formula (1). *Tags* are the address of IoT time series data's tags, which can be null, because some IoT time series data do not have tags. block_address is the address of the data area, the head IoT time series data block in NVM. start_time_SSD is the newest time of the IoT time series data stored in SSD corresponding to this IoT device.

Secondly, formula (1) is used to classify the IoT time series data as the high timeliness data category and low timeliness data category based on the access frequency for one IoT device. The high timeliness IoT time series data block is stored into NVM in the hybrid solid-state storage system, and the low timeliness IoT time series data blocks will be continuously migrated to the SSD from NVM.

$$\text{threshold} = \lceil \frac{\text{frequent\_range}}{\text{interval} * 2033} \rceil. \quad (1)$$

In formula (1), the migration threshold is the maximum number of high timeliness IoT time series data blocks. The earlier IoT time series data block will be migrated from NVM to SSD and converted to low timeliness data blocks.

Thirdly, when a high timeliness data block is converted into a low timeliness data block, the IoT_device_ID, start_time, and p_num stored in the IoT time series data block will be used to construct a key named Lblock_key. At the same time, the storage address of this IoT time series data block in the SSD is used as the value named Lblock_value. Then, as shown in Figure 5, the (Lblock_key, Lblock_value) KV pair is used as the block note of low timeliness IoT time series data block and stored in the NVM.

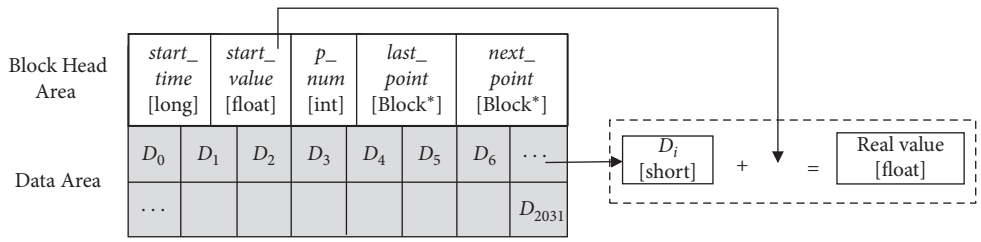


Figure 3: The structure of the IoT time series data block.

```
Void Block-compression (new_block, previous_block, current_time, current_value)
{    If (new_block is empty) //Initialized the IoT time series data block
{        The address of the previous IoT time series data block is filled by the value of next_point in the new_block.
   The current_time and current_value will be stored in the start_time and start_value in the new_block.
   block_start = 0; //Checking the top value of this IoT time series data block
   p_num = 0;}
else {
   if (new_block.start_value = = current_value) && (block_start = = 0) && (new_block is not full)
      p_num++; // The same value with the first IoT time series data is summarized and records in p_num.
   else
      if (new_block is not full)
         {The current_value and new_block.start_value will be amplified and the short integer delta of them will be calculated and
stored in D0 to D2031;
         block_start = 1; //Stopping to check the same value as the start_value. } }
```

Pseudocode 1: The redundancy elimination and compression rule.

IoT_device_ID [string]	*interval* [byte]	*value_precision* [byte]	*frequent_range* [int]	*threshold* [int]	*tags* [Block*]	*block_address* [Block*]	*Start_time_SSD* [time]

Figure 4: IoT device data structure diagram.

IoT_device_ID + start_time + p_num [string]	*Block_address_in_SSD* [unsigned int]

Figure 5: The structure of block note for low timeliness IoT time series data blocks.

On this basis, a migration algorithm for IoT time series data blocks is designed, which will migrate the older high timeliness IoT time series data blocks in the NVM into SSD to become the low timeliness IoT time series data blocks.

Figure 6 shows the diagram of IoT time series data distribution in the hybrid solid-state storage system. There are several independent storage areas for each IoT device in NVM to store the high timeliness IoT time series data blocks. The low timeliness IoT time series data blocks will be stored in SSD, but their block notes also are stored in NVM by several KV pairs.

By using the hierarchical management strategy, the IoT time series data can be distributed between SSD and NVM on the basis of their access frequency, which can be used to ensure the economics and efficiency of IoT time series database. At the same time, the time series data of different IoT devices are stored and managed independently, which can adapt to the characteristic difference of IoT devices and ensure the efficiency of the IoT time series data management. In addition, the block notes of IoT time series data block in SSD are stored in NVM, which can improve the search efficiency and reduce the write of SSD due to its block interface.

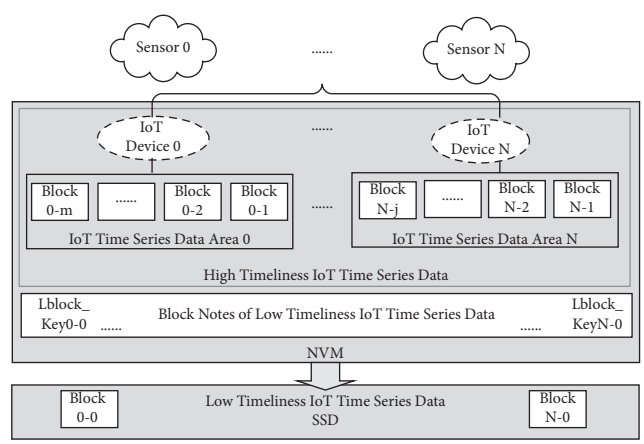


Figure 6: The diagram of IoT time series data distribution in a hybrid solid-state storage system.

4.4. Heterogeneous Time Series Data Index. In IoT systems, there are a large number of IoT devices, which bring a great challenge to search the IoT time series data in database. On the basis of 4.2 and 4.3, we design the heterogeneous time series data index mechanism, such that different methods are used to index the high and low timeliness IoT time series data. Its diagram is shown in Figure 7.

According to the distribution of IoT time series data in the hybrid solid-state storage system, the two-dimensional hybrid index is used to index the high timeliness IoT time series data on the NVM, and the block note index is used to the low timeliness IoT time series data on the SSD. At the same time, all indexes are stored in NVM to avoid the write amplification of SSD and use the I/O performance advantage of NVM, which can improve the efficiency of index for IoT time series database.

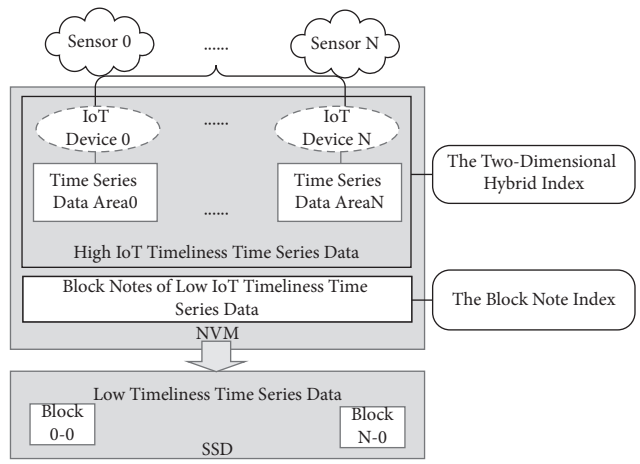


Figure 7: The diagram of heterogeneous time series data index.

4.4.1. Two-Dimensional Hybrid Index. Most of the searches in IoT systems target the high timeliness IoT time series data stored in NVM. The IoT time series database needs to identify the IoT time series data area at first and then search the corresponding IoT time series data block and the time series data. Meanwhile, most of the searches in IoT are the time range query.

A two-dimensional hybrid indexing is designed for the high timeliness IoT time series data. Firstly, the mid-square hash is used to identify IoT time series data area by the identifier of IoT device and get the address of the head IoT time series data block in this area. An inverted linked list is used to manage several IoT time series data blocks. Formula (2) is used to check whether the time series data block contains the time series data for search by start_time, start_value, and p_num stored in the head of each IoT time series data block.

$$\begin{cases} \text{search\_time\_start} \le \text{start\_time} + (p\text{\_num} + 2032) * \text{interval}, & (1), \\ \text{search\_time\_end} \ge \text{start\_time}, & (2). \end{cases} \tag{2}$$

In formula (2), search_time_start and start_time_end refer to the start and end time of the time range query. If formulas (1) and (2) both return true, this means that this data block is one of the target data blocks.

Then, formula (3) is used to calculate the position of the IoT time series data value in this block, and formula (4) is used to calculate the real value of the corresponding IoT time series data. If x is less than 0, the start_value is the value of query.

$$x = \frac{\text{search\_time} - \text{start\_time}}{\text{interval}} - p\text{\_unm} - 1, \tag{3}$$

$$\text{value} = \begin{cases} \text{start\_value} + D[x] * 10 - \text{value\_precision} & (x \ge 0), \\ \text{start\_value} & (x < 0). \end{cases} \tag{4}$$

In formula (3), search_time is the time point of IoT time series data query.

The two-dimensional hybrid index can be used to get the address of IoT time series data blocks efficiently. Compared with IoT time series data, the number of IoT devices is much smaller. The mid-square hash can be used to identify the IoT time series data area quickly, which can reduce the access times of NVM by using the powerful computing resource of CPU. The amount of high timeliness IoT time series data is also smaller compared to the total amount of time series data for each IoT device. Meanwhile, each time series data block can store more than 2033 values of IoT time series data, which also can reduce the time overhead of query. The access frequency of IoT time series data is decreased over time, and the interblock inverted linked list can reduce the time overhead of most query and avoid the frequent index change like B-tree.

4.4.2. Block Note Index. When the high timeliness IoT time series data block is converted to the low timeliness one, they will be migrated to the SSD. However, the low I/O

performance and block interface bring the challenge for query efficiency. Besides, the number of low timeliness IoT time series data blocks of one IoT device is much larger than that of high timeliness ones. However, the access frequency of the low timeliness IoT time series data is low as well, which provides some facilitates for the management of low timeliness IoT time series data blocks.

In 4.3, the block note of the low timeliness IoT time series data block is constructed and stored in NVM by KV pairs. As shown in Figure 8, the B-tree is used to index all block notes of low timeliness IoT time series data blocks in NVM. Then, all data for the query are stored in NVM. By the block note index, the start_time and p_num of the corresponding IoT time series data block can be got, and formula (2) can be used to check whether it contains the target time series data.

Therefore, the block node index can be used to reduce the access time of SSD to improve the efficiency of IoT time series database and improve the lifetime of SSD by using the NVM.

5. Prototype and Evaluation

Intel Optane DC Persistent Memory is a commercialized NVM storage device with the DIMM interface. PMEM [35] and NVMe [36] are open source device drivers for Intel Optane DC Persistent Memory and NVMe SSD. We modified the source code of PMEM and NVMe, and the redundancy elimination and compression fusion module, the hierarchical management module, and the heterogeneous time series data index module are added. Then, the IoT time series database engine can be embedded in PMEM and NVMe. Meanwhile, some new system calls are added in the Linux. Therefore, the prototype of the embedded IoT time series database for hybrid solid-state storage system is implemented, named TS-NSM. In TS-NSM, PEME is used as a fast raw storage device.

YCSB-TS is used as the test tool. It is a specialized performance testing tool for time series database. The Workloada and Workloadb in YCSB-TS are used as two workloads. There are a load stage and run stage in each workload. In the load stage, 1 million of time series data with one tag will be written into the database by the Workloada, and the same amount of time series data with three tags will be written into the database by the Workloadb. In the run stage, there are 1000 random queries in the Workloada. There are 1000 random range queries executed in the Workloadb, and there are 250 scans, 250 count, 250 sum, and 250 average operations for results.

The testing results will be compared with InfluxDB and OpenTSDB. InfluxDB is a time series database reimplemented from the bottom that does not rely on any other database. There is the batch mode to improve the write efficiency; we use InfluxDB-batch to delegate it. OpenTSDB is a distributed time series database based on HBase. During the test, HBase is run in all-in-one mode to avoid the network influence. The default configuration of InfluxDB and OpenTSDB is used.

The Intel Optane DC Persistent Memory is configured as App Direct mode and used as the fast block storage device for the prototypes of InfluxDB, influxdb-batch, and OpenTSDB. Meanwhile, the Ext4 is used as the file system for them. The hybrid solid-state storage system is used for TS-NSM constituent of Intel Optane DC Persistent Memory and NVMe SSD. Intel Optane DC Persistent Memory will retain 1000 of 4 KB IoT time series data blocks, and the rest of time series data will be stored in NVMe SSD.

One server is used for testing, which contains two 128 GB Intel Optane DC Persistent Memories and one NVMe SSD. Its configuration is shown in Table 1.

5.1. Write Performance. The load stage of Workloada and Workloadb in YCSB-TS is used to test the IOPS of write. The number of write threads is set to 1, 2, 4, 8, 16, and 32. The batch size of InfluxDB-batch is set to 10. The results are shown in Tables 2 and 3.

The results in Table 2 show that the written IOPS of TS-NSM is always the highest among all prototypes. Compared with InfluxDB, InfluxDB-batch, and OpenTSDB, its write throughput can increase by 86.5~243.6 times, 7.8~23.5 times, and 125.6~188.1 times, respectively. These results prove that TS-NSM can greatly improve the write IOPS of time series data. Although InfluxDB is a brand new time series database, its write throughput is still low. The batch mode can improve the write IOPS by about 10 times, but it is still lower than TS-NSM. OpenTSDB's write throughput is also lower compared with the all-in-one mode HBase. When there are two writing threads, OpenTSDB's write IOPS is close to that of the InfluxDB. The write IOPS of TS-NSM tends to increase firstly and then fall with more writing threads. When the number of writing threads is 8, the write IOPS of TS-NSM is the maximum. This is because the Intel Optane DC Persistent Memory still has performance limitations in multithreaded concurrent write. Meanwhile, the write IOPS of InfluxDB and InfluxDB-batch is continuously increased when increasing the number of write threads, but their write IOPS remains at a low level. OpenTSDB's write IOPS does not change significantly, despite the increase in the number of threads, and its highest and lowest throughput only differ by 54 ops. Meanwhile, OpenTSDB even has the lowest write IOPS after the write threads increased to 2. The reason is that OpenTSDB needs to store the data in Hbase, which seriously increases the complexity of the I/O software stack and extra time overhead. This also indicates that the I/O stack is a crucial factor affecting the write throughput of the time series database based on the solid state storage device. By contrast, TS-NSM's write IOPS increased by 29.3% after the number of writing threads increases from 1 to 8, and it can maintain high write IOPS after that. This is because TS-NSM embeds the time series database engine into the device drivers, which can shorten the I/O software stack and avoid unnecessary cache and replication.

Table 3 shows the test results of write throughput with the Workloadb. Similar to Workloada, the write IOPS of TS-NSM is consistently higher than that of InfluxDB, InfluxDB-

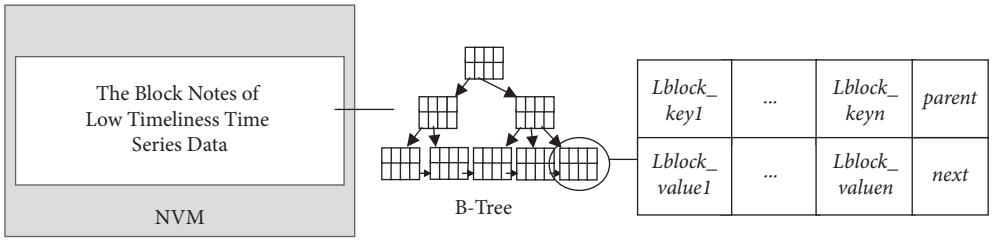


Figure 8: The block note index for low timeliness IoT time series data blocks.

Table 1: Test environment configuration.

Component	Configuration
CPU	Intel Xeon Platinum 8260 M 2.30 G
Memory	128 GB
NVDIMM	2 * 128 GB Intel Optane DC Persistent Memory
Disk	256 GB NVMe SSD
OS	CentOS 7.0 (Kernel Version 4.4.112)

Table 2: Write throughput on the Workloada.

The write throughput of workloada load stage (ops)						
TSDB	1thread	2thread	4thread	8thread	16thread	32thread
InfluxDB	459	981	1576	1442	1608	1652
InfluxDB-batch	4580	9248	15293	15192	15892	16112
OpenTSDB	887	894	867	840	853	870
TS-NSM	112283	130568	143730	158904	140697	151911

Table 3: Write throughput on the Workloadb.

The write throughput of Workloadb load stage (ops)						
TSDB	1thread	2thread	4thread	8thread	16thread	32thread
InfluxDB	440	665	1217	1322	1529	1601
InfluxDB-batch	4432	8145	13769	14897	15062	15867
OpenTSDB	803	841	821	820	809	816
TS-NSM	107415	118205	131873	143766	127855	139630

batch, and OpenTSDB. Compared with Workloada, there are more tags in each time series data in Workloadb, which brings more data written size. The write throughput of all prototypes is lower compared to that of Workloada. However, the difference of write IOPS between InfluxDB and InfluxDB-batch has increased; it increases from a maximum of 9.5 times under Workloada to 11.2 times under Workloadb, which indicates that the cache can reduce the write times and improve the efficiency surely when the amount of time series data is large. Compared with Workloada, the write IOPS of InfluxDB and InfluxDB-batch with Workloadb is reduced by 22.8% and 11.9%, while the write IOPS of TS-NSM is only reduced by 9.5%. This shows that TS-NSM has better adaptability, and the written IOPS is still high after increasing the amount of written data. However, the write IOPS of TS-NSM dropped even more significantly after the number of write threads increased more than 8 due to the bad adaptability of the Intel Optane DC Persistent Memory with concurrent write.

5.2. Query Performance. The run stages of Workloada and Workloadb in YCSB-TS are used to test the throughput of queries. The number of query threads is set to 1, 2, 4, 8, 16, and 32. The test results are shown in Figures 9 and 10.

Figure 9 shows the throughput of 1000 times random query with Workloada. It can be found that the random query throughput of TS-NSM is better than that of InfluxDB and OpenTSDB, and the query IOPS of TS-NSM improves 6.6~10.1 times and 2.2~14.4 times compared with that of InfluxDB and OpenSDB, respectively. The query IOPS of all prototypes shows a trend of first increasing and then decreasing with the increasing of query thread number. The differences are that InfluxDB's query IOPS reaches its peak when the number of query threads is 8, and it decreases by 4.5% and 9.0% when the number of query threads is 16 and 32. The query IOPS of TS-NSM and OpenTSDB reached its peak at 16 query threads, and the query IOPS of OpenTSDB decreased by 12.5% and that of TS-NSM only decreased by 4.2% with 32 query threads. Therefore, the TS-NSM has the

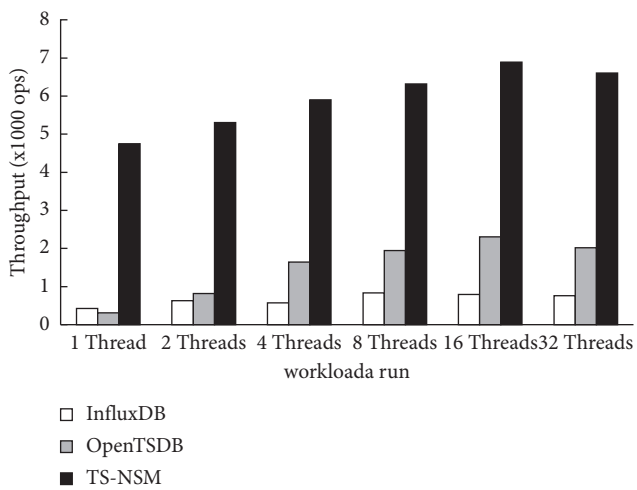


Figure 9: The query throughput with Workloada.

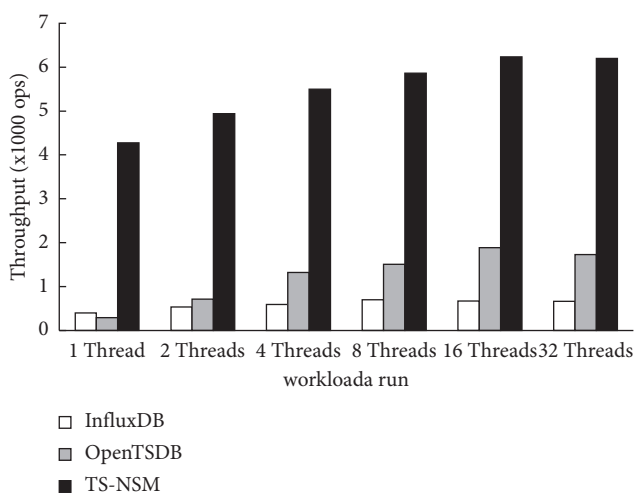


Figure 10: The query throughput of Workloadb.

more stable query throughput with several concurrent query threads, and it can better adapt to the large concurrent query for different IOT devices in the IoT system.

Time range query is very pervasive in the IoT time series database. Figure 10 shows the range query IOPS testing results by using Workloadb. Similar to Workloada, the range query IOPS of TS-NSM is much higher than that of InfluxDB and OpenTSDB, which is 7.5~9.8 times higher than that of InfluxDB and 2.4~13.8 times higher than that of OpenTSDB. This indicates that TS-NSM is also excellent in the range query throughput. Similar to the load stage, the query IOPS of using the Workloadb is generally lower than that of using Workloada. The range query IOPS of TS-NSM, InfluxDB, and OpenSDB also showed a trend of increasing first and then decreasing with the increase of query threads. The range query throughput of InfluxDB reaches a peak at 8 threads and then decreased by 4.3% and 5.8%. OpenTSDB and TS-NSM still reach a peak at 16 threads, and then OpenTSDB's IOPS is reduced by 8.2%, and TS-NSM's IOPS only decreases by 0.6%. Compared with Workloada, the

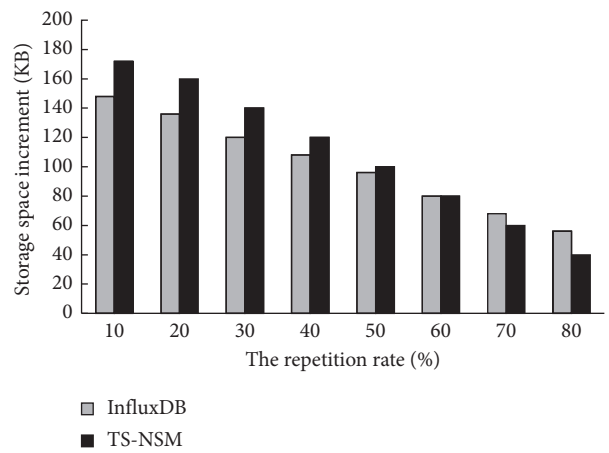


Figure 11: The storage space increment.

change rate of query IOPS is decreased with Workloadb, which indicates that TS-NSM has more advantages than InfluxDB and OpenSDB for range query. Meanwhile, the IOPS decrement of TS-NSM is also the lowest among the three prototypes, which further indicates that TS-NSM has better scalability of range query.

5.3. Compression Ratio. In general, the values of IoT time series data always do not change over a period of time. Therefore, a dataset containing 100, 000 of IoT time series data is built to test the compression ability of the prototype. The repetition rate of the dataset is from 10% to 80%. The storage space increments after inserting dataset are shown in Figure 11.

As shown in Figure 11, the compression ratio of TS-NSM is inferior to that of InfluxDB when the repetition rate of dataset is less than 50%, and the storage space increment of TS-NSM is much more than InfluxDB 4%–16%. This is because TS-NSM just uses the redundancy elimination and compression fusion strategy to achieve data compression and does not use common data compression algorithms such as XOR calculation difference of float point data and Snappy encoding of string used by InfluxDB. However, when the repetition rate of the dataset is more than 60%, the storage space increment of TS-NSM is always lower than that of InfluxDB. The space increment of TS-NSM is 29% less than that of the InfluxDB when the repetition rate of the dataset is 80%. Meanwhile, compared with InfluxDB, the compression of TS-NSM has less time overhead. These results indicate that TS-NSM can better adapt to the characteristics of IoT time series data to reduce the storage space overhead, because the repetition ratio is always high in IoT system, and there are a large number of IoT devices concurrently inserting time series data to the time series database.

6. Conclusion

Because IoT time series data is an important form of big data, storing and managing massive IoT time series data are a crucial task. NVM has the advantages of high I/O speed,

being nonvolatile, and being byte-addressable. And SSD has the advantages of high economy, large capacity, and higher I/O speed compared with HDD. Therefore, they can be mixed and constructed as a hybrid solid-state storage system to provide great support for the efficient storage and management of IoT time series data, while the current time series database lacks the corresponding optimization strategies. Based on the analysis of the characteristics of IoT time series data storage and management, an embedded IoT time series database for hybrid solid-state storage systems constructed by NVM and SSD is designed. Its structure and main algorithms are given. And the prototype named TS-NSM is implemented based on the device driver of NVM and SSD, which is verified by YCSB-TS, and the results show that the algorithms are effective.

Now, TS-NSM still lacks optimization mechanisms for multicore processors. In the future, we plan to make improvements in this regard to improve the storage and management efficiency of IoT time series database.

Data Availability

The data used to support the findings of this study are included within the article. For any further enquiries, the readers can contact the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded by the National Natural Science Foundation of China, grant no. 61806086, and the Project of National Key R&D Program of China, grant no. 2018YFB0804204.

References

[1] S. Rhea, E. Wang, E. Wong, N. Storer, and E. Atkins, "LittleTable: a time-series database and its uses," in *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 125–138, ACM, New York, May 2017.

[2] A. M. Caulfield, A. De, J. Coburn, T. I. Mollow, R. K. Gupta, and S. Swanson, "A high performance storage array architecture for next-generation, non-volatile memories," in *Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 385–395, IEEE, Washington DC, December 2010.

[3] J. Li and C. Lam, "Phase change memory," *Science China Information Sciences*, vol. 54, no. 5, pp. 1061–1072, 2011.

[4] K. Kuan and T. Adegbija, "Mirrorcache: an energy-efficient relaxed retention L1 STTRAM cache," in *Proceedings of the 2019 on Great Lakes Symposium on VLSI*, pp. 299–302, ACM, Tysons Corner VA USA, May 2019.

[5] U. D. Dadmal, R. S. Vinkare, P. G. Kaushik, and S. A. Mishra, "3D X point technology," *International Journal of Electronics, Communication and Soft Computing Science and Engineering*, pp. 13–17, 2017.

[6] Intel, "Intel optane dc persistent memory," 2019, https://www.intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory.html.

[7] S. Swanson and A. M. Caulfield, "Refactor, reduce, recycle: restructuring the I/O stack for the future of storage," *Computer*, vol. 46, no. 8, pp. 52–59, 2013.

[8] S. K. Jensen, T. B. Pedersen, and C. Thomsen, "Time series management systems: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 11, pp. 2581–2600, 2017.

[9] J. Debrabant, J. Arulraj, A. Pavlo, M. Stonebraker, S. Zdonik, and S. R. Dulloor, "A prolegomenon on OLTP database systems for non-volatile memory," in *Proceedings of the 40th International conference on Very Large Data Bases*, pp. 57–63, Hangzhou, China, 2014.

[10] P. Q. Jin, Z. L. Wu, X. L. Wang, X. Hao, and L. Yue, "A page-based storage framework for Phase change memory," in *Proceedings of the 2017 International Conference on Massive Storage Systems and Technology*, pp. 152–164, IEEE, Piscataway, NJ, USA, May 2017.

[11] Z. Q. Fan, F. G. Wu, D. Park, J. Diehl, D. Voigty, and D. Du, "Hibachi: a cooperative hybrid cache with NVRAM and DRAM for storage arrays," in *Proceedings of the 33rd International Conference on Mass Storage Systems and Technologies*, IEEE, Piscataway, NJ, USA, May 2017.

[12] Y. M. Chen, Y. Y. Lu, Y. Fan, Q. Wang, Y. Wang, and J. Shu, "An efficient log-structured key-value storage engine for persistent memory," in *Proceedings of the Architectural Support for Programming Languages and Operating Systems*, pp. 1077–1091, ACM, Lausanne, Switzerland, March 2020.

[13] S. Li and L. Huang, "LosPem," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 16, no. 3, pp. 1–17, 2020.

[14] T. Cai, D. J. Niu, Y. He, and Z. Yeqing, "NVMCFS: complex file system for hybrid NVM," in *Proceedings of the 2016 IEEE 22nd International Conference on Parallel and Distributed Systems*, pp. 577–584, IEEE, Wuhan, China, December 2016.

[15] Y. Kwon, H. Fingler, T. Hunt, S. Peter, E. Witchel, and T. Anderson, "Strata: a cross media file system," in *Proceedings of the 26th ACM Symposium on Operating Systems Principles*, pp. 460–477, ACM, Shanghai, China, October 2017.

[16] Z. Shengan, H. Morteza, and S. Steven, "Ziggurat: a tiered file system for non-volatile main memories and disks," in *Proceedings of the17th Conference of File and Storage Technologies*, pp. 207–219, USENIX Association, Berkeley, CA, USA, February 2019.

[17] C. C. Chou, J. Jung, A. L. N. Reddy, P. V. Gratz, and D. Voigt, "Virtualize and share non-volatile memories in user space," *CCF Transactions on High Performance Computing*, vol. 2, no. 1, pp. 16–35, 2020.

[18] J. Arulraj, A. Pavlo, and S. R. Dulloor, "Let's talk about storage & recovery methods for non-volatile memory database systems," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 707–722, ACM, Melbourne, Australia, May 2015.

[19] A. Joy, L. Justin, F. M. Umar, and P.-A. Larson, "BzTree: a high-performance latch-free range index for non-volatile memory," *Proceedings of the VLDB Endowment*, vol. 11, no. 5, pp. 553–565, 2018.

[20] T. Wang and R. Johnson, "Scalable logging through emerging non-volatile memory," *Proceedings of the VLDB Endowment*, vol. 7, no. 10, pp. 865–876, 2014.

[21] R. Johnson, I. Pandis, N. Hardavellas, A. Ailamaki, and B. Falsafi, "Shore-MT: a scalable storage manager for the multicore era," in *Proceedings of the 12th International Conference on Extending Database Technology*, pp. 24–35, ACM, Saint Petersburg, Russia, March 2009.

[22] C. Shimin, P. B. Gibbons, and S. Nath, "Rethinking database algorithms for Phase change memory," in *Proceedings of the 5th Biennial Conference on Innovative Data Systems Research*, pp. 21–31, CIDR, Asilomar, CA, USA, January 2011.
[23] S. Chen and Q. Jin, "Persistent B + -trees in non-volatile main memory," *Proceedings of the VLDB Endowment*, vol. 8, no. 7, pp. 786–797, 2015.
[24] F. Xia, D. Jiang, J. Xiong, and N. Sun, "HiKV: a hybrid index key-value store for DRAM-NVM memory systems," in *Proceedings of the 2017 USENIX Conference on USENIX Annual Technical Conference*, pp. 349–362, USENIX, Berkeley, CA, USA, July 2017.
[25] L. Zhan, K. Lu, Z. Cheng, and J. Wan, "RangeKV: an efficient key-value store based on hybrid DRAM-NVM-SSD storage structure," *IEEE Access*, vol. 8, no. 99, p. 1, 2020.
[26] W. Li, D. Jiang, J. Xiong, and Y. Bao, "HiLSM: an LSM-based key-value store for hybrid NVM-SSD storage systems," in *Proceedings of the 17th ACM International Conference on Computing Frontiers*, pp. 208–216, ACM, Catania, Italy, May, 2020.
[27] P. O'Neil, E. Cheng, D. Gawlick, and E. O'Neil, "The log-structured merge-tree (LSM-tree)," *Acta Informatica*, vol. 33, no. 4, pp. 351–385, 1996.
[28] D. B. Kairos, "Fast time series database on Cassandra," http://kairosdb.github.io/.
[29] "OpenTSDB scalable time series database (TSDB)," http://opentsdb.net/.
[30] "Influxdb.com: InfluxDB–Open Source Time Series, Metrics, and Analytics Database, ," , 2015.
[31] D. E. Yagoubi, R. Akbarinia, F. Masseglia, and T. Palpanas, "Massively distributed time series indexing and querying," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 1, pp. 108–120, 2020.
[32] T. Pelkonen, S. Franklin, P. Cavallaro, Q. Huang, J. Meza, and K. Veeraraghavan, "Gorilla: a fast, scalable, in-memory time series database," in *Proceedings of the VLDB Endowment*, pp. 1816–1827, Trondheim, Norway, September 2005.
[33] S. K. Jensen, T. B. Pedersen, and C. Thomsen, "ModelarDB," *Proceedings of the VLDB Endowment*, vol. 11, no. 11, pp. 1688–1701, 2018.
[34] V. Alexander, S. Alexey, Y. Semen et al., "Peregreen– modular database for efficient storage of historical time series in cloud environments," in *Proceedings of the 2020 USENIX Annual Technical Conference*, pp. 589–601, USENIX, Berkeley, CA, USA, July 2020.
[35] Peme: http://pmem.io/.
[36] NVMe: https://nvmexpress.org/.

Hindawi
Scientific Programming
Volume 2021, Article ID 5543714, 9 pages
https://doi.org/10.1155/2021/5543714

Research Article

Privacy Data Security Policy of Medical Cloud Platform Based on Lightweight Algorithm Model

JiMin Liu,[1] **HuiQi Zhao**,[1] **Chen Liu**,[2] **and QuanQiu Jia**[3]

[1]*Department of Intelligence Equipment, Shandong University of Science and Technology, Tai'an 271000, Shandong, China*
[2]*State Grid Nanjing Power Supply Company, Nanjing 210000, Jiangsu, China*
[3]*School of Taishan Technology, Shandong University of Science and Technology, Tai'an 271000, Shandong, China*

Correspondence should be addressed to HuiQi Zhao; zhqskd@163.com

Received 5 April 2021; Accepted 24 May 2021; Published 11 June 2021

Academic Editor: Shah Nazir

The deterioration of aging population has seriously hindered the development of society. Medical cloud platform has been widely used to alleviate the pressure of aging population on social economy. Most of them collect the user's sign information through the edge node and complete the disease prediction and diagnosis function combined with the cloud platform. However, the limited resources prevent the edge node from deploying the corresponding security policy after completing the data collection, storage, and calculation, which makes the edge data easy to be stolen. This paper proposes a security architecture of medical cloud platform based on lightweight algorithm model, which not only satisfies the needs of medical cloud platform to complete disease prediction and diagnosis accurately, but also creates a more secure edge node environment combined with other security strategies and hardware design. Finally, the prediction of cerebrovascular disease is used to verify the effectiveness of the proposed algorithm model.

1. Introduction

With the improvement of living standards, the aging population is becoming increasingly serious. It is estimated that the number of 55-year-old people in China will reach 300 million by 2025. The aging of population not only affects the health level of residents, but also greatly increases the medical expenses [1]. In view of the above problems, medical cloud platforms have sprung up to complete daily monitoring and disease prediction through the detection of users' data. Mora et al. constructed a health monitoring framework based on the Internet of Things and edge computing and applied it to prevent sudden death of athletes in the sports field [2]. Vilaplana et al. designed a control system for hypertensive patients, which uses hypertensive patients to send SMS to Cloud Computing Center for patient monitoring [3]. Shah Nazir et al. used seven popular machine learning algorithms to construct the heart disease prediction system [3]. But the application of edge computing brings more security problems. Hamid et al. proposed a Fog Computing Facility with Pairing-Based Cryptography, using edge computing tools to protect private data in cloud [4]. The new ice ++ framework proposed by Alberto et al. improves the security and availability of the whole medical environment by improving MCPs (medical cyber-physical system) [5]. Shaukat Ali et al. constructed a model-based security engineering for cyber-physical systems, $[n+1]$, but there are still many security problems in the edge layer [6]. The medical cloud platform uses edge nodes to collect information and due to the timeliness of medical cloud platform, more computing tasks are assigned to edge devices. However, the computing power of edge devices is limited, and there are not enough resources to configure security policies after the relevant computing tasks are allocated. Therefore, it is important to ensure the data privacy of users in the medical cloud platform without increasing the performance of edge devices.

This paper proposes a security framework for medical cloud platform, which deploys the improved lightweight computing model to edge devices. The edge node can deploy more security policies without increasing high-power devices. At the same time, data storage is deployed according to

the characteristics of the algorithm to ensure the security of user privacy data.

The paper objectives are as follows:

The medical cloud platform architecture is improved to ensure that the edge nodes have enough resources to deploy security policies while meeting the timeliness

The edge node structure based on smart phone is designed to ensure the security of the edge node

A reasonable computing framework is designed to meet the needs of medical cloud platforms and ensure that users' privacy data are stored in the cloud with higher security

A computing model based on lightweight framework is proposed to reduce the computing pressure of edge nodes

2. Proposed Platform Architecture

Different from the edge side, the cloud has enough computing power and storage capacity to meet the deployment and implementation of various security policies [7, 8], so most of the computing and data storage are arranged in the cloud. However, the medical cloud platform requires faster response and higher timeliness. Therefore, this paper proposes a security framework for high timeliness medical cloud platform, as shown in Figure 1.

A part of the computing is deployed to the edge node, and a feasible way is adopted to ensure the storage security of the edge node. The mature transmission protocol is used to upload the calculated basic attribute values and some physical signs data values to the cloud to ensure the data transmission security of edge nodes. A secure data storage and analysis system is established in the cloud to analyze the uploaded data and feed back the results, in order to further ensure the security of the framework, reduce the use of API as much as possible without affecting the accuracy of the algorithm, and avoid more security problems caused by complex API.

In addition, it has been a hot topic for medical data to be safely transmitted to the cloud through the edge layer after being collected by the Internet of Things devices [9, 10]. However, the resources occupied by the security policy are a burden for both the edge layer and the Internet of Things layer, and some carriers with relatively limited resources cannot run the security program perfectly. In the proposed framework, smart phones are used as the main edge nodes, which is not suitable. It not only has enough computing power and storage capacity, but also has the ability to collect biometrics. Combined with the algorithm model proposed for the privacy data problem, the data can be divided into basic information and sensitive information. The basic information is provided by the user when registering through the smartphone, mainly including the user's age, gender, date of birth, and basic physical conditions. The user's sign data are collected by the IoT layer sensor and uploaded to the cloud through the smart phone and stored in the cloud, or directly stored in the cloud by the hospital and other third parties. Cloud can deploy a variety of high-strength security policies relying on strong computing power and storage capacity, so the security of smart phone is crucial.

2.1. Sensors Represented by Smart Phones. IoT devices are used to collect the daily signs of user and convert them into digital signals, which have been widely used in the field of medicine and health [11–13]. With its low energy consumption and light use characteristics, IOT devices are used to collect the underlying information. The current high-sensitivity sensors, blood pressure meters, hand rings, and other instruments have high accuracy and low energy consumption and can complete the data collection work well. About IOT, this paper introduces the development of IOT devices. The safety of equipment has also achieved results [14, 15]. In the framework proposed in this paper, smart phones, as edge nodes, bear part of the computing power and storage capacity, and there are also a variety of security strategies. Ranadheer et al. deployed a new security service, EdgeSec, on the edge layer to improve the security of the entire Internet of Things system [16]. Xiao et al. analyzed several representative problems and security strategies of the edge layer Finally, the future research direction is proposed [17]. However, it is very difficult to deploy the security policy with the smart phone as the edge node without taking up too much resources so as not to affect the normal use of the smart phone. Therefore, the framework proposed in this paper adopts the mode of being equipped with a secure smart phone and adopts the existing open source projects OP-TEE (open trusted execution environment), and TrustZone hardware architecture protects user privacy data stored in edge nodes. The security smart phone is designed in Figure 2.

In the proposed framework, smart phones can also collect user data. Users provide basic information through registration and upload it to the cloud for storage and establish cloud PHR (personal health) and the sign information collected by the Internet of Things devices will be first transmitted to the smart phone. Prediction of chronic diseases does not require real-time uploading of data. The data collected by the Internet of Things will be stored in the security smart phone. In addition, the security smart phone will save the big data transmitted by the cloud. The analysis results are used to determine the disease. The edge node first determines and feeds back the user's data, which not only speeds up the efficiency from determination to feedback, but also avoids the peak of data transmission. The use of trust zone hardware architecture and open source OP-TEE can ensure the storage and computing security of edge nodes, which has been widely studied and applied [18, 19]. OP-TEE was first developed by Linaro company and released on GitHub in the form of high-quality open source code, which has realized the trusted kernel OPTEE_OS, supports trusted area allocation and multithreading, and supports a variety of platforms, such as ARM Juno Board and HiKey board. In addition, the OP-TEE project implements most of the API encapsulation provided by platform organizations around the world and has great advantages in platform portability.

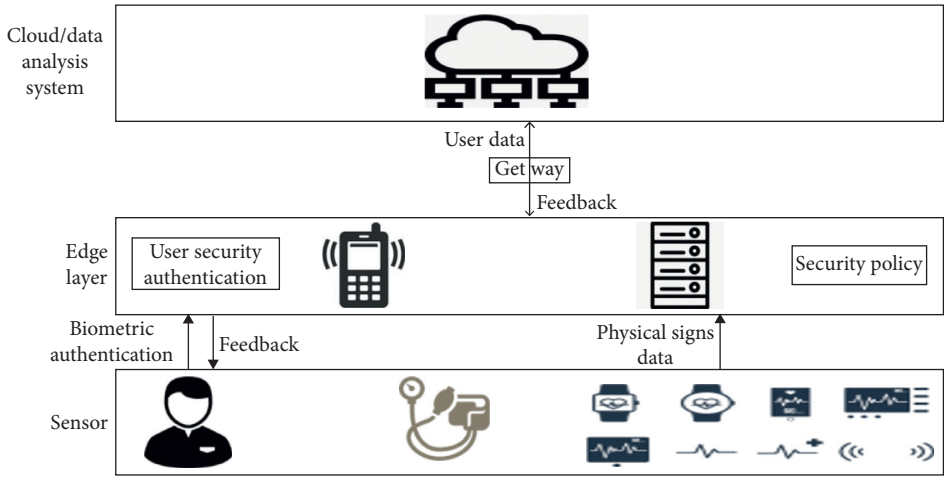


Figure 1: Architecture of security platform.

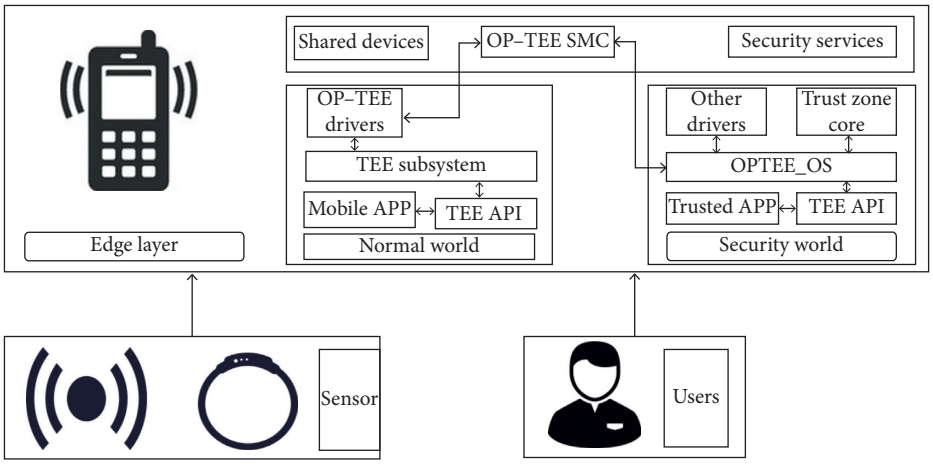


Figure 2: Architecture of security smartphone.

TrustZone technology was first introduced by ARM company, and now it has mature applications and unified standards. For example, Huawei mate 8 implements ARM TrustZone TZASC IP core in SOC, runs Huawei's operating system in the ordinary world, runs secure OS that is not open to the outside world in the secure world, or uses the antiloss function of Xiaomi mobile phone to realize arm in the bottom layer. The TZPC IP core of TrustZone technology can completely control the device and effectively prevent attackers from invading the device and stealing or modifying data. ARM TrustZone has a wide range of application scenarios with low cost, flexible and relatively simple programming environment, and can be designed according to the specific needs of users. ARM TrustZone and OP-TEE project can provide a high security environment for smart phones. The normal smart phone operating system can be stored in the ordinary world, and another operating system can be stored in the secure world. The security level of the secure world is higher, and the hardware resources accessible by the secure world are completely separated from the ordinary world. The two do not interfere with each other. EE communicates with general API and switches through SMC (secure monitor call) abnormal interrupt. In a secure world, it can protect the confidentiality, integrity, and access rights of resources and data. Users interact with the platform through normal world's mobile app, including information registration and receiving feedback. The user's sign data collected by sensors are also transmitted to the user's mobile phone through Bluetooth, ZigBee, and other ways and stored in the trust zone of secure world. In addition, the results generated by cloud big data analysis are also stored in the trust zone. In zone, the sign data accumulated with the user for a period of time are predicted in the trusted application of secure world and fed back to the user.

ARM TrustZone hardware architecture and open source OP-TEE make the smart phone as a reliable edge node with certain computing and storage capacity because of its flexible and highly independent characteristics. Huawei mobile phone based on HiSilicon chip can perfectly complete such tasks and provide a higher level of hardware security.

2.2. Edge Layer and Cloud Layer. As the product of cloud computing marginalization, edge layer lacks computing power and storage capacity compared with the cloud [20],

but the emergence of edge computing can not only reduce the load of the cloud, but also provide users with more rapid feedback [21, 22]. Edge nodes not only need to communicate with heterogeneous and resource limited Internet of Things devices in a short distance, but also upload data to the Internet as data collection in the middle layer of the cloud; message middleware or network virtualization technology is mostly used with the cloud server. Security is also very important. Although the storage and computing environment of edge node devices has been guaranteed by ARM TrustZone and OP-TEE, there is still a risk that the thief will attack or steal data by stealing the user's mobile phone or account. Therefore, a way to guarantee the user's credentials is needed, in which MFA (multifactor authentication) can provide effective authentication by adding at least one authentication factor in addition to password to the authentication process, which can not only ensure the security of user accounts, but also ensure that data can be safely and effectively transmitted from edge nodes to the cloud in the transmission protocol. There are many forms of MFA and applications [23–25]. The traditional identity authentication adopts the identification mechanism of "user account + static password", which is actually a single factor authentication. A static password has various disadvantages, which are easy to be guessed or illegally stolen by attackers. MFA can try and judge the user's identity and behavior in many aspects, among which biometric authentication technology uses the physiological information of human body. As a way of information authentication based on characteristics and behavior characteristics, each person's biometrics are different. Using biometric authentication technology can ensure the accuracy of verification results. With the support of today's intelligent devices, biometric collection and biometric-based authentication can be completed with high precision, which makes up for the lack of security authentication in various fields [26, 27]. In the proposed framework, intelligent devices are used as information collection devices, and sufficient resources are available to support the collection of biometrics. Fingerprint features are used to verify users, which has the characteristics of low cost and high security, as shown in Figure 3.

The user will first register through the app on the smart phone, and the app will be deployed to the normal world to interact with the user. The user will register by providing basic information, upload it to the cloud, and establish a PHR. The user's biological signs information, such as fingerprint information, will be stored secure in the world. When the user logs in again, the fingerprint information provided by the user will be compared. Only after the comparison is successful can the data access rights of the user account be obtained. In addition, after the user registration is successful, the secure smartphone will accept the model parameters sent from the cloud and save them in the secure world; when the user's data collection is completed, the initial disease judgment will be carried out at the edge node and timely feedback will be given to the user. When the user requests or needs to predict the disease, it will be temporarily stored in the secure smartphone. The user data in the world will be uploaded to the cloud. Otherwise, the cloud database will be uploaded and updated when the pressure of cloud resource occupation is low. At the same time, the model parameters will be updated after a period of accumulation. The data analysis system set up in the cloud will feed back the disease prediction results or healthy diet suggestions to users through the app.

For this reason, the model algorithm reduces the use of computing power of edge nodes, and the secure smartphone based on OP-TEE can ensure the storage security of edge nodes. The disease prediction model established in the cloud transmits the regularly updated model parameters to the smartphone, and the disease can be determined by using a small amount of storage and computing power of the smartphone.

In the process of designing the attributes required by the algorithm, firstly, users provide part of their basic information through registration and establish PHR in the cloud. At the same time, PHR receives the user's sign data and EMR (electronic medical) provided by hospitals and other third parties. The improved APC model is used to determine the disease through the cloud security environment, and the model parameters are stored in the secure world in the edge layer smartphone through the secure transmission protocol. When the user visits, the disease is determined in the edge layer first, and when it is necessary to predict the disease, the cloud is visited to predict and return the prediction results and exercise prescription.

3. Proposed Algorithm Model

In previous studies, it was found that, with the degradation of physical function, the elderly generally suffer from chronic diseases, and the physical signs data generally deviate from the normal value, which has caused great interference to the prediction and diagnosis of diseases. In addition, in the previous disease diagnosis and prediction algorithm design process, the security and functionality of the platform deployment architecture were rarely considered. To solve the above problems, a high-sensitivity disease prediction and diagnosis model for the elderly based on the security of edge computing is proposed.

3.1. Consideration of Algorithm Sensitivity. In order to solve the problem of the interference caused by the fact that the old people's body sign data are generally higher than the standard value to the judgment and prediction, this paper proposes a crowd queue algorithm based on the combination of basic information and sign data, which reflects the macro impact of social and historical development on the population through the basic information and uses the sign data to express the changes of personal physical function with the growth of time. It can ensure the sensitivity of the algorithm to the greatest extent.

Basic information is not only used to form cloud PHR, but also has many applications in disease judgment and prediction. APC (age period cohort) model studies the impact on outcomes from three dimensions of period, age, and cohort. Since frost proposed and applied to tuberculosis data research in 1939, it has been widely used. Pes et al. used

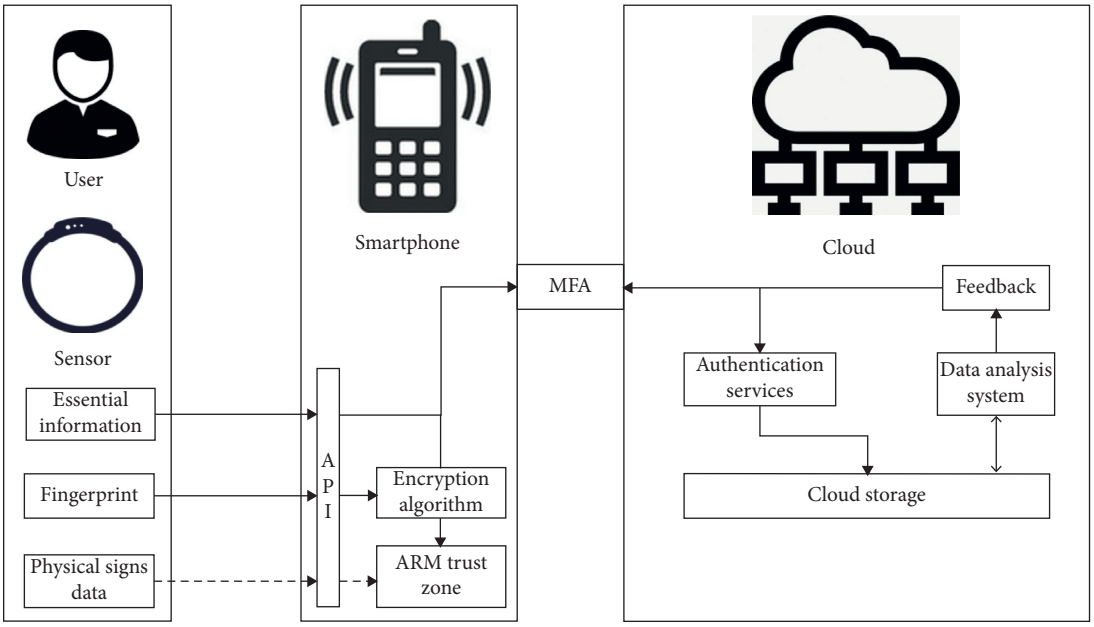


Figure 3: Edge layer architecture.

the APC model to analyze and study the time trend of cancer in the population after the socioeconomic transformation [28]. It has unique advantages in the prediction of chronic diseases and some infectious diseases.

In the previous study, we improved the APC model and studied the influence of smoking, drinking, and other habits on the outcome, which improved the sensitivity of APC model [29], but, at the same time, considering the irreplaceable value of sign data for disease prediction model, we also used basic data and sign data to ensure the sensitivity and accuracy of the algorithm to the greatest extent, which can be used in personal health. In the field of management, different diseases need different algorithm attribute selection, and attribute selection is often of a large number. The algorithm performance and complexity in massive high-dimensional datasets are the main basis for algorithm selection. The performance of clustering algorithm is excellent. But the application of pruning technology to reduce the number of dense unit candidate sets will lead to the loss of some datasets, so it is relatively easy to make mistakes in the field of disease prediction. Gayathri et al. combined PROCLUS algorithm with density-based algorithm to improve its performance in high-dimensional dataset clustering [30], but PROCLUS algorithm is easy to ignore small clustering, and it is easy to cause errors such as incomplete results in the field of disease prediction. Genetic clustering algorithm combines genetic algorithm with clustering algorithm and ensures accuracy through selection, exchange, and mutation operation, and the algorithm is simple, accurate, and effective, which is suitable for the research of cardiovascular and cerebrovascular diseases in the elderly.

Genetic algorithm was first proposed by Professor Holland of the University of Michigan in the 1960s and 1970s and was published in the first book on the basic theory and method of genetic algorithm in 1975. Because it only depends on the fitness function, it can demand the optimal solution, and it is simple and practical. It is suitable for parallel computing and has the characteristics of high efficiency and practicality. It is widely used in machine learning, neural network, and biological engineering. The genetic algorithm is simple, effective, and accurate, so it can be used to solve clustering problems. This paper uses genetic algorithm clustering. In previous studies, the traditional clustering algorithm is very sensitive to the selection of the preset clustering center and the input order of samples, and it is easy to fall into the problem of local optimization. Therefore, combining the local optimization of the traditional clustering algorithm with the global optimization of genetic algorithm, it has great value in the high-dimensional and low sensitive medical data of the elderly. With the advantages, genetic algorithm clustering has been widely used [31].

Then, the results of the improved APC model are quantized and input into the genetic algorithm clustering as an individual attribute. Because the disease prevalence is very different in gender, the analysis is carried out according to gender. The input attribute does not include gender. The input format is in Table 1.

In order to further improve the accuracy of clustering, clustering is carried out according to the age group, which can be divided into different groups according to age and gender. At the same time, the similar groups can also ensure the accuracy of the results. Because there are few data of people under 50 and over 75 years old in the dataset, which cannot meet the data requirements of cluster analysis. The data of 50- to 75-year-old people are processed for cluster analysis, and 50 groups of data are selected for each experiment, $T = 100$, $M = 100$, $P\_m = 0.01$, and $P\_C = 0.5$. Black represents patients with cardiovascular and cerebrovascular diseases, such as coronary heart disease and heart failure, and red represents normal people. The same attributes in different age groups and genders have different degrees of influence. Users are divided according to age

Table 1: Input attributes of genetic algorithm clustering.

Attribute	Age	BZ	APC	Pulse rate
Describe	50+	Difference from standard deviation	Quantification of APC model results	User pulse rate

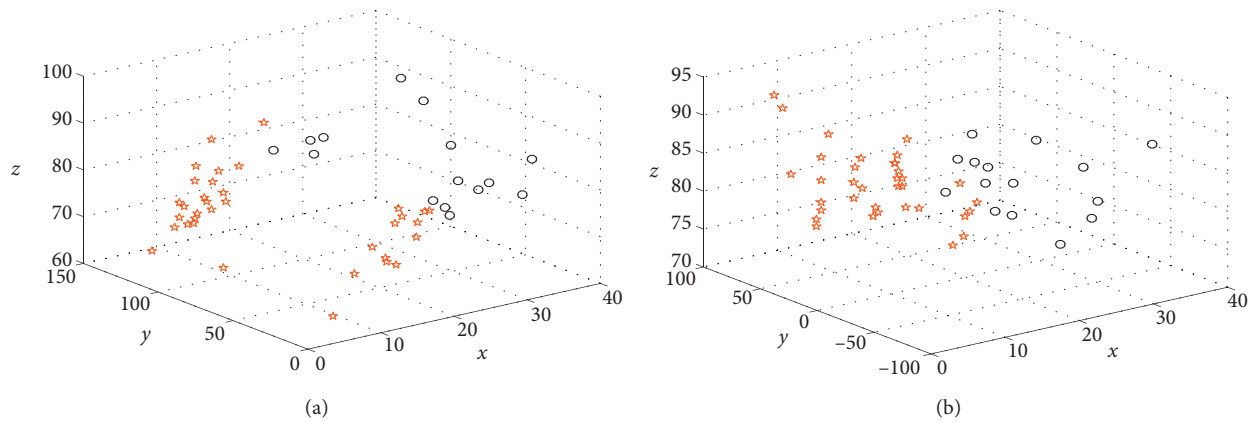


Figure 4: Cluster results of 50-year-old population ((a) male; (b) female).

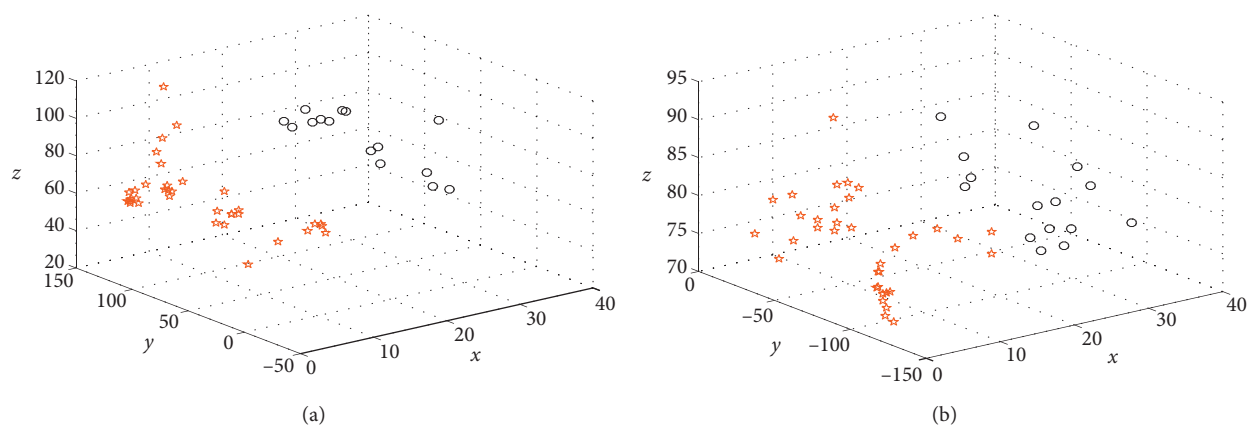


Figure 5: Cluster results of 55-year-old population ((a) male; (b) female).

groups to improve the accuracy and sensitivity of elderly data. The attributes include not only the influence of non-self-attributes on the body (APC model value), but also the main influencing factors of physical signs data (pulse rate and pressure difference). Genetic algorithm is based on the idea of population queue. The results of clustering model are as follows (Figures 4–9).

It can be seen from the analysis results in the figure that different age groups have different characteristics. With the growth of age, all attribute values have increased, which means that, with the growth of age, the immunity of the elderly gradually decreases, and the physical sign data are generally higher than the normal value, which is particularly prominent in patients with coronary heart disease and other diseases caused by hypertension. In addition, some women suffer from coronary heart disease. The results of cluster analysis are more discrete and lower than that of male patients, which indicates that female patients with hypertension are more likely to suffer from cardiovascular and cerebrovascular diseases caused by hypertension such as coronary heart disease and heart failure than male patients. According to the PHR format designed in this paper, the same user has the diagnosis results of different periods. Therefore, when the new user's data are input, the cluster analysis is first carried out according to the model, and then similar people are assigned according to the clustering results. The APC model value in the calculation attribute includes the macro impact of individual life and social changes on individuals, and the physical sign data value can reflect the individual. With the change of physical conditions, from two angles to find their belonging to the similar population, then, the probability of disease is expressed by the average probability of similar population.

Compared with the traditional genetic clustering algorithm, this model classifies users according to age and gender and uses the eigenvalues of the improved APC model as attributes to calculate. By classifying the population, the prediction accuracy of the model is improved, and the prediction of the onset period is carried out. Considering the age, gender, and other factors, the parameters of the

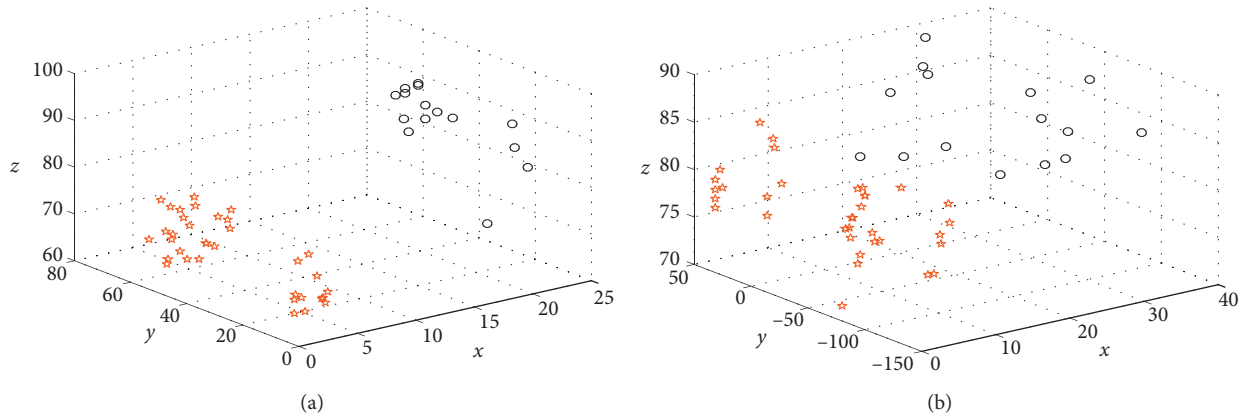


Figure 6: Cluster results of 60-year-old population ((a) male; (b) female).

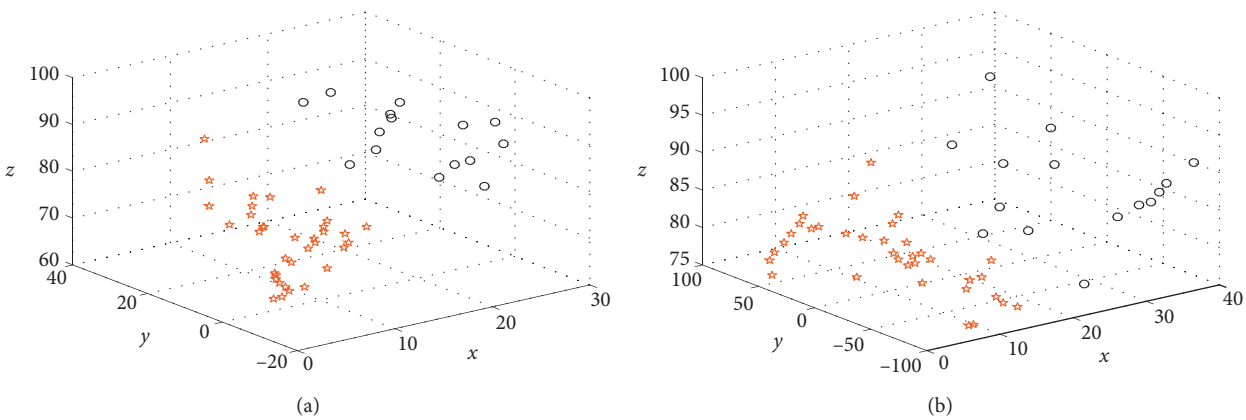


Figure 7: Cluster results of 65-year-old population ((a) male; (b) female).

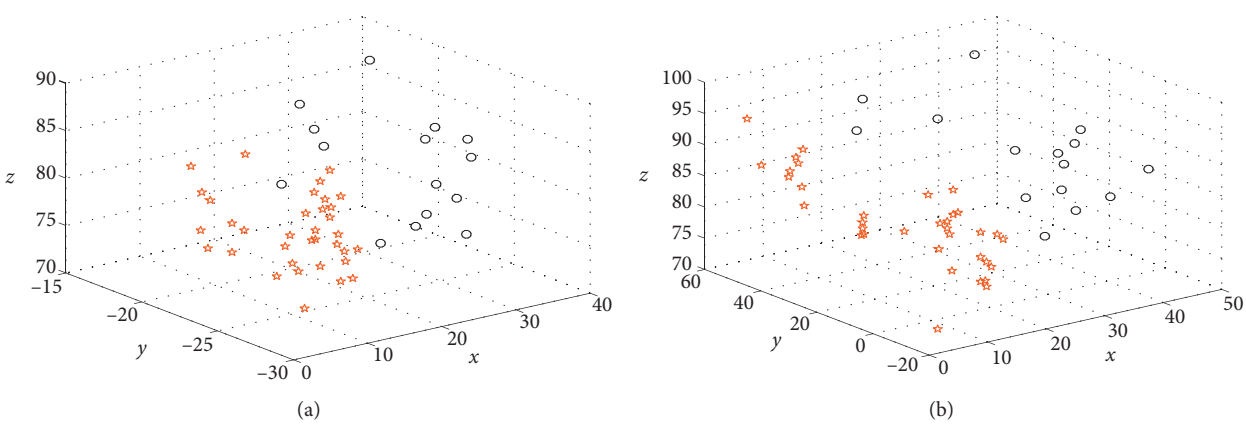


Figure 8: Cluster results of 70-year-old population ((a) male; (b) female).

improved APC model are taken as the macro parameters. The influence is added to the model analysis, and the following (Figure 10) is a comparison with the traditional genetic clustering analysis (traditional genetic clustering attribute selection: pressure difference, pulse rate, BMI).

It can be seen from Figure 10 that the model proposed in this paper has greater advantages than the AUC (area under curve) of the traditional model, and the model designed in this paper has higher accuracy, because it contains both the basic factors represented by the improved APC model and the physical factors after standardization, Although the traditional algorithm AUC also performs well, the model proposed in this paper will get higher TPR (true-positive comparison) under the same FPR (false-positive comparison) At the same time, the proposed model has higher TPR no matter how the initial point or the end point and the best

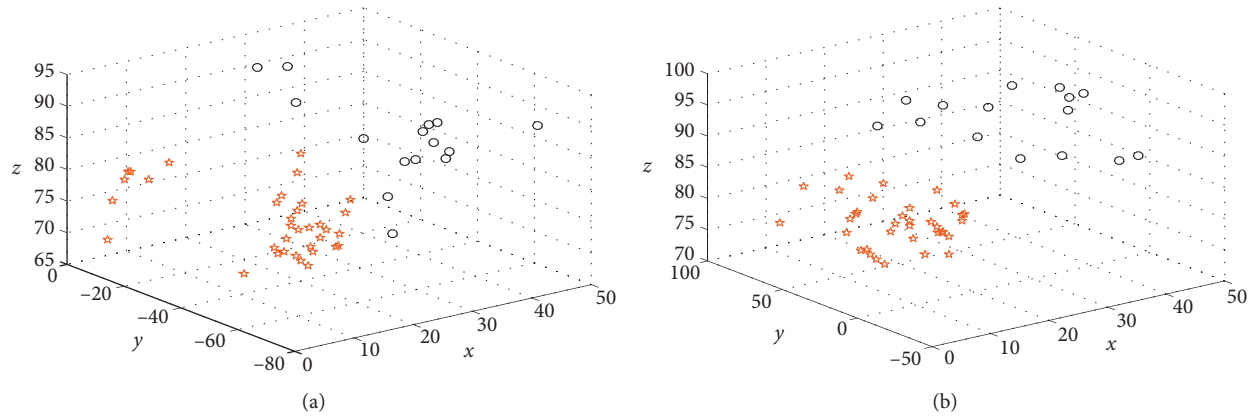


Figure 9: Cluster results of 75-year-old population ((a) male; (b) female).

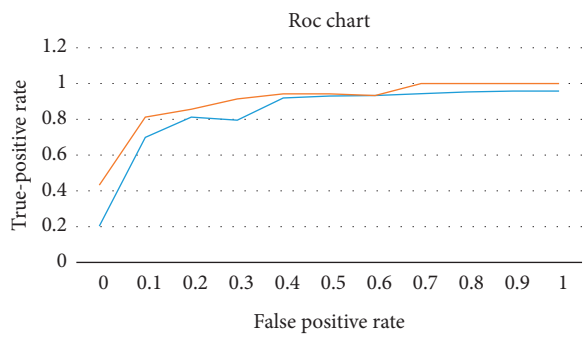


Figure 10: ROC curve comparison.

boundary point are, which indicates that the proposed model has higher adaptability and sensitivity than the traditional model in the field of cardiovascular and cerebrovascular disease prediction of the elderly.

The proposed algorithm model can ensure the accuracy of clustering to the greatest extent, so as to get the most similar people. Through the crowd queue analysis according to age, the most similar people of the same age can be obtained according to age. According to the information of similar people of the same age in PHR, the change trend of health status over time can be established and according to the changes of patients in similar people of the same age over time. The incidence rate of individual disease is the number of individuals. In order to facilitate calculation and user interaction, this article selects some similar populations and accurately sets the prevalence rate. According to the age of similar age groups, the prediction results of user onset period are expressed. The results are as follows: the probability of disease in the year is 23%, and the probability of illness is 47% in two to five years, in five to ten years. The probability of getting sick is 20%, and the probability of not getting sick within 10 years is 10%. In order to verify the accuracy of the results, this paper selects 100 test attributes to predict and analyze this method. The experimental results show that the accuracy rate reaches 91.34%, and in the case of wrong prediction results, the average deviation time is less than one year.

4. Conclusion

The framework proposed in this paper will store the calculation and storage of user sensitive information in the cloud and store the trained model parameters in the edge node, which ensures the storage of sensitive data and common data separately. The powerful computing power of the cloud is used to protect the privacy and data security of users. At the same time, the idea of using smart phones as edge nodes is proposed. ARM TrustZone and OP-TEE open source project are used to protect the security of the edge node. On the basis of not increasing the computing power of the edge node and ensuring the high timeliness of the platform, the security of the user's privacy data is guaranteed to the greatest extent.

Data Availability

The data of the paper are provided by the cooperative project for scientific research.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. Mora, D. Gil, R. M. Terol, J. Azorín, and J. Szymanski, "An IoT-based computational framework for healthcare monitoring in mobile environments," *Sensors*, vol. 17, no. 10, p. 2302, 2017.

[2] J. Vilaplana, F. Solsona, F. Abella et al., "H-PC: a cloud computing tool for supervising hypertensive patients," *The Journal of Supercomputing*, vol. 71, no. 2, pp. 591–612, 2014.

[3] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, Article ID 3860146, 21 pages, 2018.

[4] H. A. Al Hamid, S. M. M. Rahman, M. S. Hossain, A. Almogren, and A. Alamri, "A security model for preserving the privacy of medical big data in a healthcare cloud using a

Fog computing facility with pairing-based cryptography," *IEEE Access*, vol. 5, pp. 22313–22328, 2017.

[5] A. H. Celdrán, F. J. G. Clemente, J. Weimer et al., "ICE++: improving security, QoS, and high availability of medical cyber-physical systems through mobile edge computing," in *Proceedings of the 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pp. 1–8, IEEE, Ostrava, Czech Republic, September 2018.

[6] P. H. Nguyen, S. Ali, and T. Yue, "Model-based security engineering for cyber-physical systems: a systematic mapping study," *Information and Software Technology*, vol. 83, pp. 116–135, 2017.

[7] L. Fang, C. Yin, L. Zhou, Y. Li, C. Su, and J. Xia, "A physiological and behavioral feature authentication scheme for medical cloud based on fuzzy-rough core vector machine," *Information Sciences*, vol. 507, pp. 143–160, 2020.

[8] O. Kocabas and T. Soyata, "Towards privacy-preserving medical cloud computing using homomorphic encryption," *Virtual and Mobile Healthcare*, pp. 93–125, IGI Global, Hershey, USA, 2020.

[9] A. Celesti, M. Fazio, A. Galletta, L. Carnevale, J. Wan, and M. Villari, "An approach for the secure management of hybrid cloud-edge environments," *Future Generation Computer Systems*, vol. 90, pp. 1–19, 2019.

[10] M. Rehman, N. Javaid, M. Awais, M. Imran, and N. Nazeer, "Cloud based secure service providing for IoTs using blockchain," in *Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–7, IEEE, Waikoloa, HI, USA, December 2019.

[11] P. Verma and S. K. Sood, "Cloud-centric IoT based disease diagnosis healthcare framework," *Journal of Parallel and Distributed Computing*, vol. 116, pp. 27–38, 2018.

[12] P. P. Ray, "Understanding the role of internet of things towards smart e-healthcare services," *Biomedical Research*, vol. 28, no. 4, pp. 1604–1609, 2017.

[13] B. P. L. Lo, H. Ip, and G.-Z. Yang, "Transforming health care: body sensor networks, wearables, and the Internet of things," *IEEE Pulse*, vol. 7, no. 1, pp. 4–8, 2016.

[14] M. O'Neill, "Insecurity by design: today's IoT device security problem," *Engineering*, vol. 2, no. 1, pp. 48-49, 2016.

[15] J. Wurm, K. Hoang, O. Arias, A.-R. Sadeghi, and Y. Jin, "Security analysis on consumer and industrial IoT devices," in *Proceedings of the 2016 21st Asia and south pacific design automation conference (ASP-DAC)*, pp. 519–524, IEEE, Macao, China, January 2016.

[16] K. Sha, R. Errabelly, W. Wei, T. Andrew Yang, and Z. Wang, "Edgesec: design of an edge layer security service to enhance iot security," in *Proceedings of the 2017 IEEE 1st International Conference on Fog and Edge Computing (ICFEC)*, pp. 81–88, IEEE, Madrid, Spain, May 2017.

[17] Y. Xiao, Y. Jia, C. Liu, X. Cheng, J. Yu, and W. Lv, "Edge computing security: state of the art and challenges," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1608–1631, 2019.

[18] R. Liu and M. Srivastava, "PROTC: PROTeCting drone's peripherals through ARM trustzone," in *Proceedings of the 2017 3rd Workshop on Micro Aerial Vehicle Networks, Systems, and Applications*, pp. 1–6, Niagara Falls, NY, USA, June 2017.

[19] M. Gentilal, P. Martins, and L. Sousa, "TrustZone-backed bitcoin wallet," in *Proceedings of the 2017 Fourth Workshop on Cryptography and Security in Computing Systems*, pp. 25–28, Stockholm Sweden, January 2017.

[20] D. Dasgupta, A. Roy, and A. Nag, "Multi-factor authentication," *Infosys Science Foundation Series*, pp. 185–233, Springer, Cham, Switzerland, 2017.

[21] D. M. T. Ting, O. Hussain, and G. Laroche, "Systems and methods for multi-factor authentication," U.S. Patent 9,118,656, 2015.

[22] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[23] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[24] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: a survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–454, 2017.

[25] J. Armington and P. Ho, "Robust multi-factor authentication for secure application environments," U.S. Patent 10/086,123, 2003.

[26] F. Han, J. Hu, X. Yu, and Y. Wang, "Fingerprint images encryption via multi-scroll chaotic attractors," *Applied Mathematics and Computation*, vol. 185, no. 2, pp. 931–939, 2007.

[27] H. Chen and H. Chen, "A novel algorithm of fingerprint encryption using minutiae-based transformation," *Pattern Recognition Letters*, vol. 32, no. 2, pp. 305–309, 2011.

[28] G. M. Pes, F. Cocco, S. Bibbò, G. Marras, and M. P. Dore, "Cancer time trend in a population following a socio-economic transition: results of age-period-cohort analysis," *International Journal of Public Health*, vol. 62, no. 3, pp. 407–414, 2017.

[29] Z. Li, L. Wen, J. Liu et al., "Fog and cloud computing assisted IoT model based personal emergency monitoring and diseases prediction services," *Computing And Informatics*, vol. 39, 2020.

[30] S. Gayathri, M. Mary Metilda, and S. Sanjai Babu, "A shared nearest neighbour density based clustering approach on a proclus method to cluster high dimensional data," *Indian Journal of Science and Technology*, vol. 8, no. 22, pp. 1–6, 2015.

[31] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," *Pattern Recognition*, vol. 33, no. 9, pp. 1455–1465, 2000.

[32] Mof.gov.cn. (2019). Financial news. [online] Available at: http://www.mof.gov.cn/zhengwuxinxi/caizhengxinwen/.

Hindawi
Scientific Programming
Volume 2021, Article ID 9976209, 9 pages
https://doi.org/10.1155/2021/9976209

Research Article

Application of Data-Driven Iterative Learning Algorithm in Transmission Line Defect Detection

Yuquan Chen,[1] **Hongxing Wang**,[1] **Jie Shen**,[1] **Xingwei Zhang**,[1] **and Xiaowei Gao**[2]

[1]*Jiangsu Frontier Electric Technology, Nanjing 211102, China*
[2]*Beijing Imperial Image Intelligent Technology, Beijing 100085, China*

Correspondence should be addressed to Hongxing Wang; wanghxft@163.com

Received 15 March 2021; Accepted 3 May 2021; Published 13 May 2021

Academic Editor: Shah Nazir

Deep learning technology has received extensive consideration in recent years, and its application value in target detection is also increasing day by day. In order to accelerate the practical process of deep learning technology in electric transmission line defect detection, the current work used the improved Faster R-CNN algorithm to achieve data-driven iterative training and defect detection functions for typical transmission line defect targets. Based on Faster R-CNN, we proposed an improved network that combines deformable convolution and feature pyramid modules and combined it with a data-driven iterative learning algorithm; it achieves extremely automated and intelligent transmission line defect target detection, forming an intelligent closed-loop image processing. The experimental results show that the increase of the recognition of improved Faster R-CNN network combined with data-driven iterative learning algorithm for the pin defect target is 31.7% more than Faster R-CNN. In the future, the proposed method can quickly improve the accuracy of transmission line defect target detection in a small sample and save manpower. It also provides some theoretical guidance for the practical work of transmission line defect target detection.

1. Introduction

In recent years, deep learning methods produced an effective method for big data processing, and it has made a breakthrough in many different fields such as automatic speech recognition and target recognition [1–5]. At present, it has begun to promote the development of a new generation of artificial intelligence industry worldwide. In the notice of the three-year action plan, China proposed promoting the mutual promotion of the real economy and artificial intelligence technology [6, 7]. With the development of the intelligent industry, deep learning technology has begun to emerge in smart grid image recognition and defect detection applications [8–10].

It is widely known that complex tasks require high-intensity training to construct deep models, and deep learning techniques also require large-scale data for network training [11–13]. In transmission line defect target detection, it is difficult to construct training data due to the variable target shape. Using a small number of samples will lead to poor generalization of the training model and make it prone to overfitting. These problems bring some difficulties for transmission line defect detection. Combined with the above phenomenon and the important role of recognition algorithm in transmission line defect detection, this paper reviews the related work. For example, The improved SSD method is utilized to detect the transmission line foreign body and present an improved Faster R-CNN deep learning method to accomplish fault detection of the insulator [14, 15]. To solve the low detection accuracy of SSD for the small size object, we proposed an improved algorithm of SSD object detection based on the FP-SSD, which has greatly improved the precision [16]. Chang used SSD combined with binocular vision distance detection method to realize the detection of pantograph offset of transmission line [17]. Antwi-Beko uses Convolutional Neural Network (CNN) to detect and classify defective insulators in transmission line images, achieving high-precision defect

target recognition and location [18]. Wang Yixing trains the stackable automatic encoder (SAE) to initialize and train the deep learning neural network and observes the hidden features of defects from different dimensions, so as to make a preliminary judgment of defects [19]. The above series of works has greatly optimized the target recognition algorithm, but the data is rarely mentioned. The training algorithm designed in this paper provides different ideas to solve the abovementioned problems, mainly using professional knowledge to construct a small amount of labeled data to achieve data-driven iterative training. In order to learn the model incrementally and adapt to unlabeled data, we use the improved Faster RCNN to initialize and train the deep learning neural network and observe the hidden features of defects from different dimensions. In order to make preliminary judgments on the defects, a small amount of labeled data is used to initialize the training network in the early stage of training, and then we continue to perform sample mining in a large amount of unlabeled data in a human-machine collaborative manner [20–22]. Through this data-driven learning method, the recognition module can obtain the recognition model from the training module, and we use the acquired recognition model to label the inspection data, and then the training module uses the labeled raw data to update the model and iteratively improve the model precision. In addition, the deformable convolution and feature pyramid modules are added to the training algorithm to greatly improve the feature extraction capabilities.

2. Materials and Methods

2.1. Training Data. In general, a deep model with good generalization ability must depend on a certain amount of sample data. The richest possible training data including different situations and different forms can make the deep network get more accurate parameters in training. Considering the current lack of publicly available transmission line data sets, it is necessary to construct a professional inspection image insulator defect data set. In order to verify the robustness of the algorithm, a large number of transmission line raw data collected from a certain province and city using UAV inspection are collected. The image resolution is about $5000 * 3000$, covering four seasons. Voltage levels include 35 kV, 110 kV, 220 kV, 500 kV, and high voltage levels. Among them, pin-level defect detection is the most difficult type of defect detection in electrical defect detection, and it is a common electrical defect, so pin defect detection is selected for algorithm testing. To make the network easier to learn, the defects are divided into two categories, including lack of locking pin and rusted-on nut, and a detailed description of the pin-level defect is shown in Figure 1.

2.2. Improved Training Method. Faster R-CNN [23] is the basic two-stage target detection algorithm [24, 25], which has achieved good results in many target detection tasks. Fast R-CNN is the most basic network architecture to ensure the good scalability of the algorithm. Faster R-CNN introduces Region Proposal Net (RPN) on the basis of Fast R-CNN [26], to participate in the target recognition work, in order to improve the detection speed and realize the integration of the suggestion box generation model and the Fast R-CNN model by sharing the convolutional layer. Considering the complexity of transmission line defect samples, a deformable convolution module and a feature pyramid module are added to the Faster R-CNN algorithm to solve the problem of multiscale target detection to a great extent.

2.2.1. Deformable Convolutional Network. The key point to solve the high-precision recognition of transmission line defects in a complex background is how to adapt to the geometric changes of the proportion, posture, and viewpoint of the defective target in the image. We know that CNN's ability to model geometric transformations is limited, because the geometric structure in CNN is fixed and cannot be deformed during convolution operations. Based on this, a deformable convolution module is introduced to improve the transform modeling ability of CNN. The deformable convolutional network adds a 2-dimensional offset to the sampling point position of the receptive field in the standard convolution. It can freely match the receptive field with the target shape; that is, no matter how the transmission line defect target shape changes, the convolutional receptive field can always cover target range. These offsets are learned from the previous feature map through the additional convolutional layer. Each sampling point has learned to weight and redistribute the modified sampling area, which can achieve more accurate feature extraction, thereby effectively improving the training effect.

The general convolution operations can be divided into two major steps: (1) use a regular grid G on the input feature map for sampling; (2) perform weighting operations. G defines the size and expansion of the receptive field.

The operation of deformable convolution is different. In the regular grid R, it is expanded by adding an offset. For each position P_0 on the output feature map, the calculation equation is as follows:

$$y(p_0) = \sum_{p_n \in R} w(p_n) * f(p_0 + p_n + \Delta p_n), \tag{1}$$

$$G = \{(-1,-1), (-1,0), \ldots, (0,1), (1,1)\}. \tag{2}$$

Among them, f represents the input feature map, y represents the output feature map, and p_n is an enumeration of the positions listed in G. $p_n + \Delta p_n$ represents the offset of the sampling position relative to the center. Since Δp_n is generally not an integer, equation (1) can be realized by the bilinear interpolation method. The specific equation can be expressed as

$$f(p_0 + p_n + \Delta p_n) = \sum_q T(q,p) \cdot f(q), \tag{3}$$

where q represents any position on the input feature map f, and $T(p,q)$ is 0 in most positions of f.

(a) (b)

Figure 1: Pin-level defect. (a) Rusted-on nut. (b) Lack of locking pin.

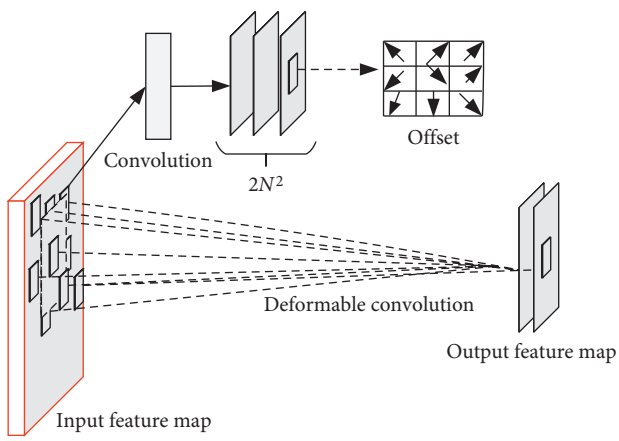


Figure 2: Deformable convolution operation.

In Figure 2 in the process of deformable convolution, if the convolution kernel size of the original input feature graph is $N \times N$, use the standard 2D convolution to calculate a new feature map consistent with the input feature map size, and its convolution kernel size is still $N \times N$. The number of channels is $2N^2$ and the $2N^2$-dimensional vector in the direction of the channel dimension represents N^2 number of 2D offsets, so each position on the new feature map represents the offset of the original convolution kernel on the input feature map. When using the deformable convolution operation, the offset of the corresponding position is superimposed with the sampling position of the original convolution kernel to obtain the offset sampling position. The process after feature sampling is consistent with the standard 2D convolution calculation, and the final result is output feature map.

2.2.2. Feature Pyramid Network. The feature pyramid network mainly aims at the recognition of multiscale targets in transmission line defects. By simply changing the network connection structure, the performance of small target recognition is greatly improved while maintaining the calculation scale of the original model. In traditional vision tasks, multiscale object detection is mainly realized by image pyramid method or hierarchical prediction method, but this method has higher requirements for hardware computing power and memory size, so it can only be used in limited fields. To solve the above problems, feature pyramid network uses the information of each layer in CNN network to generate the final expression feature combination. The feature pyramid network will process the feature output of each CNN layer in the model to generate features that reflect this dimensional information, and there is also an association relationship generated between the features after top-down processing; that is, the high-level features will affect the low-level feature expression. Finally, all the feature combinations are used as the input for the next task of target detection or category analysis. The basic structure of the feature pyramid network is shown in Figure 3. The network is directly modified on the original single

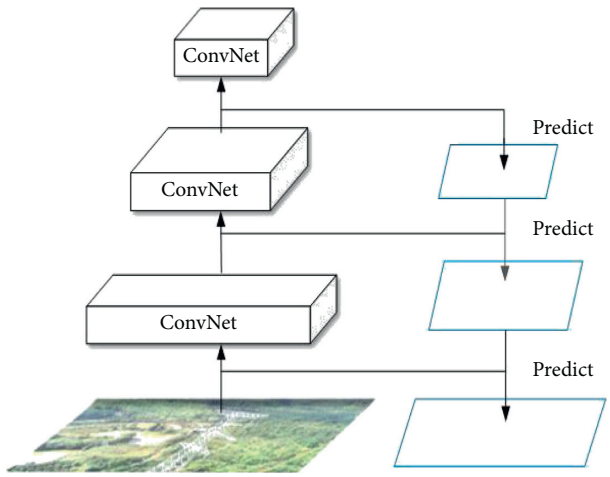


Figure 3: Feature pyramid network.

network, and the feature map is sampled from the bottom to the upper layer. The feature map of each resolution is added element by element with the feature map of twice its scale. Using this connection method, the feature maps predicted by each layer are fused with features of different semantic strength and resolution. Targets with different resolutions are detected by feature map with different resolution, so that each layer has moderately strong semantic and resolution features. Moreover, the feature pyramid network only adds additional cross-layer connections on the basis of the original structure. It is confirmed by practical applications that the use of the feature pyramid network hardly increases the amount of extra calculation and time.

3. Data-Driven Training System

In the process of continuous optimization of the model, inspired by the human learning model, a data-driven iterative learning method is constructed. Data-driven iterative learning method mainly solves two problems: one is the automatic construction of in transmission line defect sample library; the other is to solve the problem of overfitting of the training model under small samples.

The data-driven iterative learning method is based on the principle of deep convolutional neural network training. In the early stage, a small number of samples are used for model training. For a large number of unlabeled detection data, the detector with the highest accuracy model is used for detection, and the detection results are sorted by confidence. The threshold is set according to experience, and some of the detection results with higher confidence are extracted, and the active learning method is used for labeling, and the updated data is used to optimize the training model. In this way, by continuously rolling iterative training and detection processes, using self-supervision and active learning methods to cooperate with each other for sample mining, the two processes of training and detection complement each other and continuously improve the accuracy of the recognition model. Mining a large number of unlabeled samples makes the model not only improve the robustness of the classifier to noise samples or outliers, but also improve the accuracy of detection and finally realize the closed-loop structure of automatic labeling of unlabeled sample data and iterative update of model training, and the closed-loop structure diagram is shown in Figure 4.

How to complete the data connection between the training module and the detection module is a significant step in the application of the data-driven iterative learning algorithm in the transmission line inspection defect detection application. While each process is working on its part, the data dependency between them is resolved through asynchronous calls, so that the two modules can operate at the same time, and they can interact and penetrate each other. Considering the huge amount and variety of transmission line samples, the direct connection of data is mainly through database read and write operations, sample mining is performed in the recognition module part, and the training module reads data for model tuning.

3.1. Data Recognition Module. As the core of data mining, the recognition module provides necessary data support for the training module. The recognition module is based on the deep convolutional neural network model for defect identification, and the detection results are sorted according to the confidence level. If necessary, the human-computer interaction method is used to extract part of the recognition result data with higher confidence, so as to calibrate the type of defects that can be trained and recognized. The identified defect information is stored in the database, and an AI training library for inspection defect data is established. Through the cooperation of self-supervision and active learning methods, the massive uninterrupted image data is structured, labeled, and classified, and training data sources that can be easily found and used were established.

3.2. Data Training Module. With the application of defect recognition and the gradual accumulation of data, the training module can continuously perform feature learning

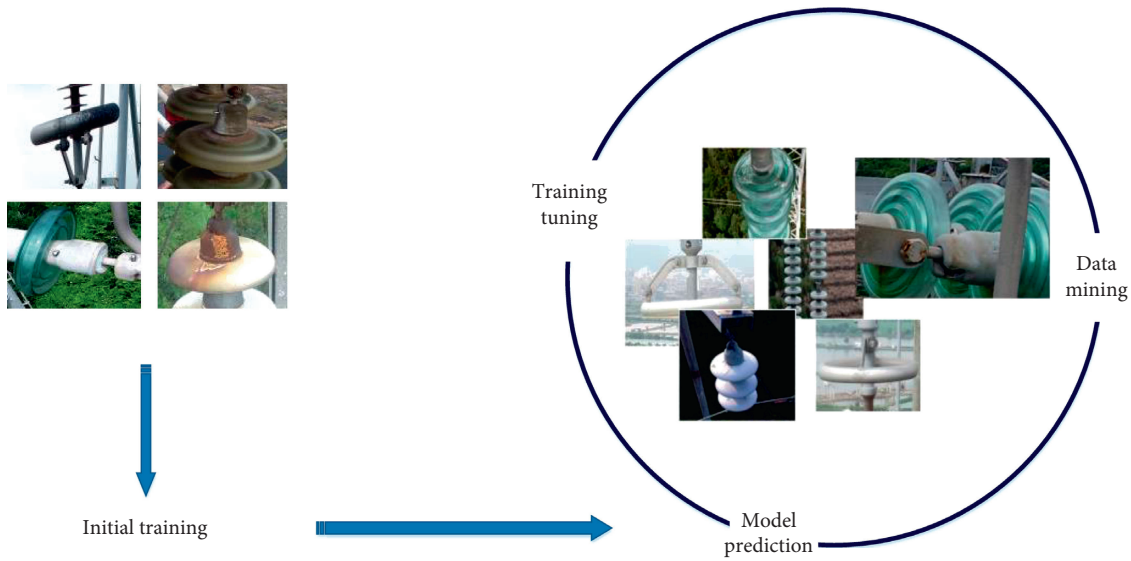


Figure 4: Data-driven iterative training closed-loop structure.

and model updates according to predetermined training strategies. It can not only quickly use artificial intelligence technology to improve the efficiency of defect recognition but also customize it according to its own specific needs. Only a small amount of data tags can complete the complex deep learning training task. Data-driven training is realized based on the principle of deep learning, and iterative inspection data and recognition models are continuously rolled over, so as to achieve a closed loop of image processing with high automation and intelligence, and to improve the accuracy of training image defect recognition.

The data-driven iterative learning algorithm ensures the continuous update of the recognition model through the data-driven training function, continuously optimizes the detection effect, and forms a closed-loop ecology from data to training. At the same time, the data-driven training can adaptively adjust training strategies and training modes according to training accuracy without excessive human intervention.

Figure 5 shows the overall process architecture of the data-driven iterative learning algorithm. The underlying environment of the algorithm is Mxnet [27]. The two parts of model recognition and model training are independent of each other. The data-driven iterative learning algorithm can be used to connect the two training and recognition modules in series to train and identify various types of defects independently. The data-driven method is used for sample mining and model iterative updating, and the model with higher accuracy can be selected for publishing.

4. Results and Discussion

Due to the high requirement of hardware configuration for deep learning training. The hardware environment of the algorithm in this experiment is Tesla V100-DGX WORK STATION, the running environment is Ubuntu16.04 operating system, and the computer processor is E5-2698 v4@2.20 GHz. The 101-layer ResNet [28] is selected as the backbone network, and different initial training parameters are selected in the experiment, such as solver type, initial learning rate, and learning step size. The network optimization function selected in this experiment is Stochastic Gradient Descent (SGD) [29], the learning step is set to 2, the image scaling scale is (1222, 800), the NMS threshold is 0.5, and the initial learning rate is 0.001.

In the initial stage, we made a small number of data samples. The Pascal VOC [30] data set format was directly imitated in the production stage, and we used Extensible Markup Language (XML) to record the location and type of defect target in detail. The experimental data in Table 1 show the initial stage of manually labeling the training set according to the defect characteristics, in which the number of lack of locking pin is 2682, and the number of rusted-on nut data is 1621, and in the test set, the number of lack of locking pin is 1899, and the number of rusts is 1412. Table 1 also shows the growth of data volume after two rounds of data-driven iterative training. Pascal VOC data set analysis uses Average-Precision (AP) as a comprehensive evaluation index and uses precision and recall to examine the model. The above indicators can be expressed as follows:

$$\text{precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}, \tag{4}$$

$$\text{recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}. \tag{5}$$

In equations (4) and (5), True Positives (TP) represent the number of defective targets that are correctly classified; FP represents the number of background interferences that are mistakenly regarded as defective targets; False Negatives (FN) represents that the defective targets are incorrectly classified as background quantity. AP, as the average accuracy, for the PR curve (Recall on the horizontal axis and Precision on the vertical axis), can be calculated as follows:

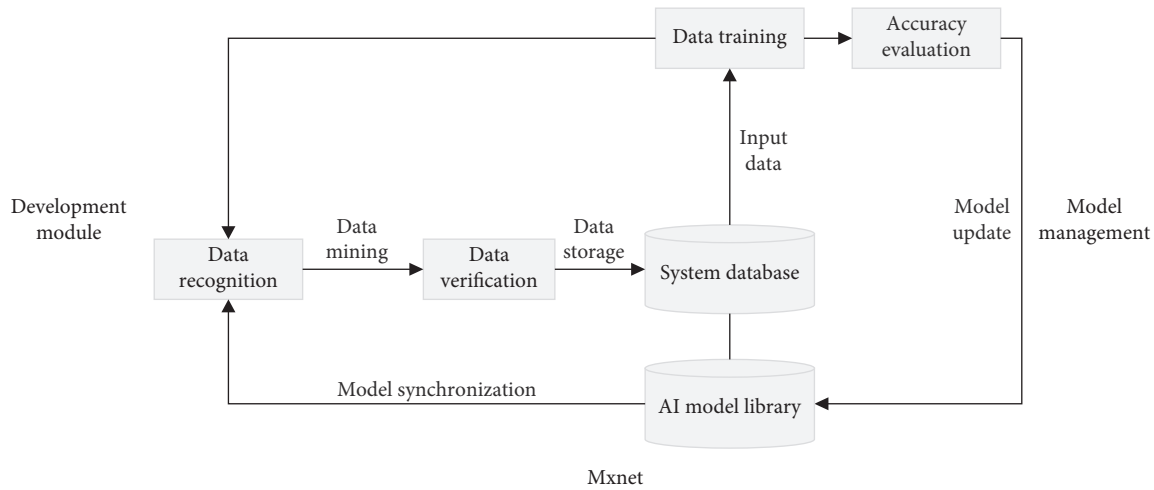


Figure 5: The overall architecture of data-driven iterative learning.

Table 1: Statistical table of data set quantity.

Datasets	Lack of locking pin (lsqxz)		Rusted-on nut (lmsx)	
	Number of pictures	Number of targets	Number of pictures	Number of targets
Initial training data	2682	3969	1621	2446
The first iterative learning + sample mining	19603	30841	7803	14514
The second iterative learning + sample mining	57272	73415	16238	24840
Test data	1899	2484	1412	1901

$$\mathrm{AP} = \int_0^1 p(r)\mathrm{d}r. \tag{6}$$

Among them, P and r are, respectively, the accuracy rate and the recall rate. In actual applications, the PR curve is not directly calculated, but the PR curve is smoothed; that is, for each point on the PR curve, the value of precision is the largest precision on the right side of the point.

In order to test the effectiveness of the platform algorithm, this paper designs a comparative experiment, which mainly includes two parts: one is the influence of the setting of the experience threshold of screening data in the process of data-driven training on the accuracy of the model. Based on the improved Faster RCNN algorithm, different experience threshold has been selected to test the accuracy of the model. The second is the analysis of the impact of improved algorithms on model accuracy. To compare the accuracy of training algorithms, first compare the accuracy of the Faster RCNN basic algorithm with the improved Faster RCNN algorithm while keeping the training data consistent, then verify the effectiveness of the data-driven iterative learning algorithm, and finally, test the influence of the number of training iterations on the results of the model.

The first part of the experimental results is shown in Figure 6. The experiment selects five values of 0.65, 0.7, 0.75, 0.8, and 0.85 as the threshold for screening data based on experience. The results show that when the threshold is set to 0.75, the training model obtained by data mining has the highest average precision. After analysis, when the threshold is set higher, some correct data will be screened out, and when the threshold is lower, it will affect the accuracy of the target. After all training methods tested, obtained when the threshold is about 0.75, the accuracy of the model can be maintained at a high level. Experiments show that the accuracy of the model can be increased to a certain extent by adjusting the appropriate threshold value. Table 2 shows the second part of the experiment content. Compared with Faster R-CNN method, the improved Faster R-CNN method performs more prominently in the detection of lack of locking pin and rusted-on nut defects. At the same time, it can be also seen in Figure 7 that the number of iterations in model training is also an extremely important parameter. As the number of iterations increases, the number of iterations required must increase accordingly to meet the needs of model learning. After the first iterative learning, the model accuracy reached the highest value of 0.704 at 2.06×10^4 iterations. After the second iterative learning, the model accuracy reached the highest value of 0.785 at 3.81×10^4 iterations, which is 25.6% higher than the model accuracy of the initial sample training. In the final experiment, the result shows that the improved Faster R-CNN method has improved the accuracy of the model to a certain extent. But after the second iterative learning, the accuracy is increased by 31.7% compared with Faster R-CNN, which proves that the proposed method is extremely effective.

Finally, the defect images of different scenes are selected to test the model. The detection results are shown in Figure 8. It can be seen that the final model can correctly identify most of the pin defects. In conclusion, the data-driven iterative learning algorithm proposed can effectively improve the accuracy of defect detection and can save manpower to a great extent, which is beneficial to users to independently train high-precision models.

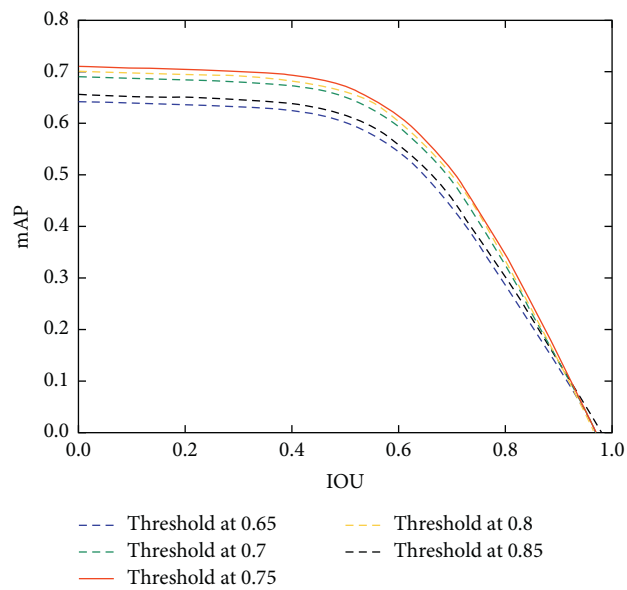


Figure 6: Model accuracy under different screening data thresholds.

Table 2: The precision of algorithm test.

Algorithm	Classes	Recall	Precision	AP	mAP	Training time (h)
Initial training data + Faster RCNN	lsqxz	0.751	0.684	0.65	0.596	6.11
	lmxs	0.643	0.685	0.541		
Initial training data + feature pyramid + deformable convolution + Faster RCNN	lsqxz	0.78	0.689	0.68	0.625	8.05
	lmxs	0.668	0.702	0.569		
The first iterative learning + sample mining + feature pyramid + deformable convolution + Faster RCNN	lsqxz	0.892	0.794	0.803	0.704	39.05
	lmxs	0.696	0.787	0.604		
The second iterative learning + sample mining + feature pyramid + deformable convolution + Faster RCNN	lsqxz	0.927	0.8	0.854	0.785	118.80
	lmxs	0.752	0.814	0.715		

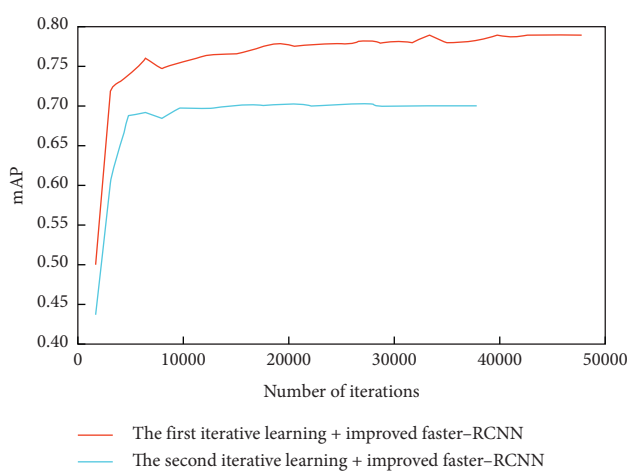


Figure 7: Model accuracy display under different iteration times.

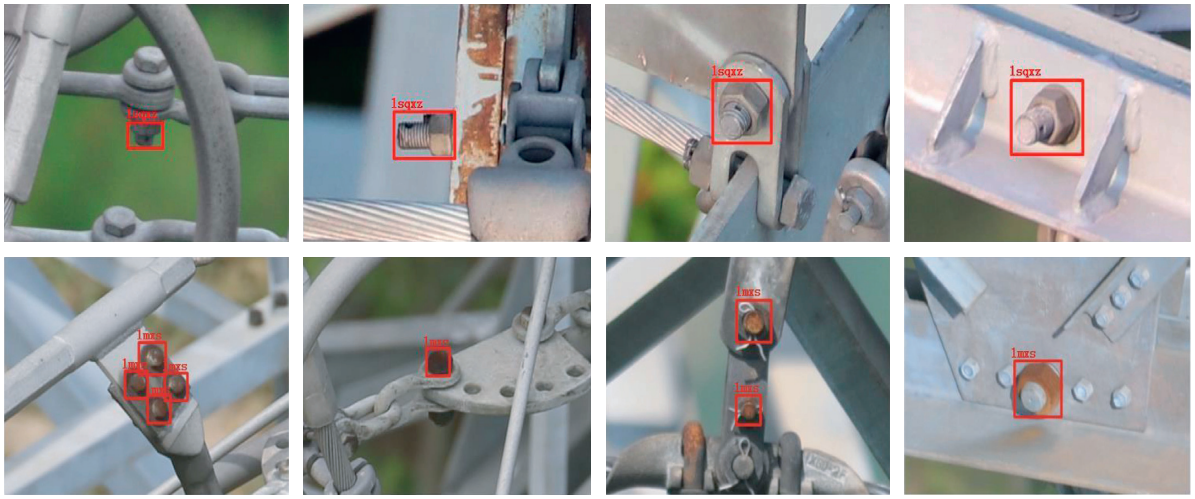


Figure 8: Pin defect identification effect.

5. Conclusions

The current work presents a data-driven iterative learning algorithm based on improved Faster R-CNN, which provides various applications such as accelerated training, model detection, and data mining for transmission line defect detection. The data-driven iterative learning algorithm proposes a data-driven training function for the application of transmission line defect detection. After comparison and verification, the proposed method can effectively improve the accuracy of inspection image defect detection. The main conclusions are as follows:

(1) Faster R-CNN with deformable convolution and feature pyramid modules is chosen as the optimal object detection architecture, so that the value of AP is increased to 0.625. The combined use of deformable convolution and feature pyramid modules makes the improved Faster R-CNN method more prominent in the detection of lack of locking pin and rusted-on nut defects.

(2) The advantages of data-driven iterative training and improved Faster R-CNN architecture are combined naturally. According to a large number of experimental comparisons, the result showed that the increase of the recognition of data-driven iterative training based on improved Faster R-CNN algorithms for the pin defect target is 31.7% more than Faster R-CNN. The proposed method can achieve a highly automated and intelligent image processing closed loop and improve the accuracy of more transmission line defect target inspections in the future.

(3) The proposed method provides a means of transmission line data mining. Due to the low cost of data collection and the convenient and efficient algorithm, it has strong practical value, which can meet the needs of users in the transmission line defect detection to a large extent.

Data Availability

This data comes from the original inspection data of the State Grid of China. Due to the limitations of the State Grid, it cannot be used as a public data set.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] L. Du, B. Liu, Y. Wang et al., "Target detection method based on convolutional neural network for SAR image," *Journal of Electronics Information Technology*, vol. 38, pp. 3018–3025, 2016.

[2] T. Zoughi, M. M. Homayounpour, and M. Deypir, "Adaptive windows multiple deep residual networks for speech recognition," *Expert Systems with Applications*, vol. 139, Article ID 112840, 2020.

[3] M. Tomaszewski, P. Michalski, and J. Osuchowski, "Evaluation of power insulator detection efficiency with the use of limited training dataset," *Applied Sciences*, vol. 10, no. 6, p. 2104, 2020.

[4] L. W. Sommer, T. Schuchert, and J. Beyerer, "Deep learning based multi-category object detection in aerial images," in *Proceedings of the SPIE Defense + Security*, International Society for Optics and Photonics, San Franciso, CA, USA, May 2017.

[5] S. Kamal, S. K. Mohammed, P. R. S. Pillai et al., "Deep learning architectures for underwater target recognition," in *Proceedings of the Ocean Electronics (SYMPOL)*, IEEE, Kochi, India, October 2013.

[6] G. Ling and L. Xiaohe, "Zhengzhou artificial intelligence industry research," *Innovation Science and Technology*, vol. 36, 2017.

[7] R. Yang, "The ministry of industry and information technology wants companies to unveil the list and take command of 17 key directions to tackle artificial intelligence," *Computer and Networks*, vol. 44, no. 22, 2018.

[8] W. Wang, Y. Wang, J. Han, and Y. Liu, "Recognition and drop-off detection of insula-tor based on aerial image," in

Proceedings of the 9th International Symposium on Computational Intelligence and Design (ISCID), vol. 1, pp. 162–167, Hangzhou, China, December 2016.

[9] Y. Zhai, D. Wang, M. Zhang, J. Wang, and F. Guo, "Fault detection of insulator based on saliency and adaptive morphology," *Multimedia Tools and Applications*, vol. 76, no. 9, pp. 12051–12064, 2017.

[10] U. Tudevdagva, B. Battseren, W. Hardt, and V. Ga-lina, "Image processing based insula-tor fault detection method," in *Proceedings of the XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, pp. 579–583, IEEE, Novosibirsk, Russia, October 2018.

[11] Y. Liu, B. Lubinski, Y. Shang et al., "Performance comparison of deep learning techniques for recognizing birds in aerial images," in *Proceedings of the 2018 IEEE Third International Conference on Data science in Cyberspace (DSC)*, IEEE Computer Society, Guangzhou, China, June 2018.

[12] V. Sharma and R. N. Mir, "A comprehensive and systematic look up into deep learning based object detection techniques: a review," *Computer Science Review*, vol. 38, 2020.

[13] A. Hazra, P. Choudhary, and M. S. Singh, "Recent advances in deep learning techniques and its applications: an overview," in *Proceedings of the ICBEST*, Raipur, India, September 2020.

[14] J. Yuan, B. Xue, W. Zhang, L. Xu, H. Sun, and J. Zhou, "RPN-FCN based rust detection on power equipment," *Procedia Computer Science*, vol. 147, pp. 349–353, 2019.

[15] X. Liu, H. Jiang, J. Chen, and J. Chen, "Insulator detection in aerial images based on faster regions with convolutional neural net-work," in *Proceedings of the IEEE 14th International Conference on Control and Automation (ICCA)*, pp. 1082–1086, Anchorage, AK, USA, June 2018.

[16] P. Qin, C. Li, J. Chen, and R. Chai, "Research on improved algorithm of object detection based on feature pyramid," *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 913–927, 2019.

[17] L. Chang, Z. Liu, and Y. Shen, "On-line detection of pantograph offset based on deep learning," in *Proceedings of the 2018 IEEE 3rd Optoelectronics Global Conference (OGC)*, pp. 159–164, IEEE, Shenzhen, China, September 2018.

[18] E. Antwi-Bekoe, Q. Zhan, X. Xie et al., "Insulator recognition and fault detection using deep learning approach," *Journal of Physics Conference Series*, vol. 1454, Article ID 012011, 2020.

[19] Y. Wang, M. Liu, and Z. Bao, "Deep learning neural network for power system fault diagnosis," in *Proceedings of the 2016, 35th Chinese Control Conference (CCC)*, pp. 6678–6683, IEEE, Chengdu, China, July 2016.

[20] T. O. Zander and C. Kothe, "Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general," *Journal of Neural Engineering*, vol. 8, no. 2, Article ID 025005, 2011.

[21] S. N. Young and J. M. Peschel, "Review of human–machine interfaces for small unmanned systems with robotic manipulators," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 99, pp. 1–13, 2020.

[22] J. Muesing, L. Burks, M. Iuzzolino et al., "Fully bayesian human-machine data fusion for robust dynamic target surveillance and characterization," in *Proceedings of the AIAA Scitech 2019 Forum*, San Diego, CA, USA, January 2019.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[24] L. Yan, M. Yamaguchi, N. Noro, Y. Takara, and F. Ando, "A novel two-stage deep learning-based small-object detection using hyperspectral images," *Optical Review*, vol. 26, no. 6, pp. 597–606, 2019.

[25] M. A. Al-Masni, W. R. Kim, Y. Kim et al., "Automated detection of cerebral microbleeds in MR images: a two-stage deep learning approach," *NeuroImage: Clinical*, vol. 28, 2020.

[26] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, Santiago, Chile, December 2015.

[27] T. Chen, M. Li, Y. Li et al., "Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems," 2015, https://arxiv.org/abs/1512.01274.

[28] K. He, X. Zhang, S. Ren et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Pittsburgh, PA, USA, June 2016.

[29] L. Bottou, "Stochastic gradient descent tricks," *Lecture Notes in Computer Science, Neural Networks: Tricks of the Trade*, Springer, Berlin, Germany, pp. 421–436, 2012.

[30] S. Vicente, J. Carreira, L. Agapito et al., "Reconstructing pascal VOC," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 41–48, Pittsburgh, PA, USA, June 2014.

Hindawi
Scientific Programming
Volume 2021, Article ID 5518309, 9 pages
https://doi.org/10.1155/2021/5518309

Research Article

Framework for Educational Domain-Based Multichatbot Communication System

Zojan Memon,[1] Hamideh Aghian,[2] Muhammad Shahzad Sarfraz,[3] Akhtar Hussain Jalbani,[4] Rozita Jamili Oskouei,[2] Khuda Bux Jalbani,[5] and Ghulam Hussain Jalbani[4]

[1]*Department of Information Technology, University of Sufism and Modern Sciences, Bhit Shah 70140, Pakistan*
[2]*Department of Computer Science and Information Technology, Mahdishahr Branch, Islamic Azad University, Mahdishahr, Iran*
[3]*Department of Computer Science, National University of Computer and Emerging Sciences, Chiniot-Faisalabad Campus, Chiniot, Islamabad, Pakistan*
[4]*Department of Information Technology, Quaid-E-Awam University of Engineering, Science and Technology, Nawabshah 67450, Pakistan*
[5]*Riphah Institute of Systems Engineering, Riphah International University, Islamabad 44000, Pakistan*

Correspondence should be addressed to Rozita Jamili Oskouei; rozita2020j@gmail.com

Received 22 January 2021; Revised 24 February 2021; Accepted 19 April 2021; Published 6 May 2021

Academic Editor: Shah Nazir

Education is an area where innovation moves slowly. In this study, we will propose a framework with a novel approach that will support the development of a multi-interactive chatbot's system for an educational area using AIML 2.0. The system will facilitate the students for their learning towards an outcome-based education domain. The proposed framework will be composed of a user module which consists of user and user interface, chat agents module which will respond to the user query, chatbot KB which will act as the brain for the chatbot system, and socket system for establishing the communication link. Finally, the proposed system will be evaluated using a confusion matrix. The multichatbot communication system will support text-based dialogues on a limited set of questions related to education. However, the system will be implemented in java. The outcomes of this research will be useful for the education sector where these intelligent systems will help the students in schools, universities, and other training scenarios.

1. Introduction

AI deals with the development of intelligent systems which can think and act as humans. There are various tasks that intelligent systems do, such as learning, planning, and speech recognition, while research related to AI is extremely specialized as well as technical.

Chatbot is one of the types of intelligent systems which communicate with the user like a human through voice commands or text or both. These chatbot systems are capable of generating an appropriate response or an appropriate action based on the user input. The quality of such kind of systems can be determined through the significance of their output which is being chosen by the chatbot in response. Basically, chatbots were intentionally designed to replicate the communication smartly with single or multiple users, while NLP and ML are the important aspects of a chatbot system [1].

Chatbots have changed the way we think and live because chatbots had the capability of being present and ready to provide help by performing tasks while conducting conversation anytime and anywhere [2]. These intelligent chatbot systems are being used in many areas such as banking, customer support, security, health, and education.

As the increasing utilization of technology has been changing the way students learn and understand the information as the chatbots have provided a personalized learning environment to the students [3]. In education, the chatbots can be used to teach the students by conducting a lecture in a series of messages in order to make it look like a

standardized chat conversation. These chatbots help to improve communication and productivity, minimize ambiguity, enhance the learning process, and assess the level of understanding of the student from interactions [4].

Various chatbot systems for education purposes exist but there are some situations where a single chatbot system is not sufficient to atomize various operations or handle various domain requests [5]. However, the use of AI technology to formulate most of the operations of an autonomous organization can be difficult for a single chatbot to handle. Therefore, there is a need of a platform which supports multichatbot communication to facilitate learner in a group learning environment where they can be able to learn through interactions between human and chatbot and between the chatbots, where not only the single chatbot agent but also multiple intelligent chat agents work together as a team to help leaners towards their educational problems. The purpose of developing such kind of systems is to reduce the agent's complexity for various domains and to observe the advantages of multiagents over a single chat agent in an interactive learning environment.

Therefore, in this research paper, we have proposed a framework which is an effective, simple, easy, and low-cost approach for designing text-based multi-interactive chatbots for education.

The proposed framework will be composed of major three modules: (1) user module, (2) agents module, and (3) module socket system. The user module will consist of the user interface through which the user will interact with the system, the agents module will consist of two chatbots that will respond to the user query, and the third module socket system will help the chat agents and users to communicate.

This paper is split up into six major parts: in Section 2, we will explore existing chatbots; Section 3 will demonstrate the framework along with its implementation, tools, and techniques that will be used to accomplish the research work; Section 4 will illustrate the results and discussion which will describe how do multichat agents simulate conversation in a quiz style on Outcome Based Education; Section 5 will explain the assessment of our multiagent chatbot communication framework which will demonstrate the accuracy of the system, and Section 6 will bring a conclusion.

2. Related Work

Various tutoring systems were built so as to provide an easy, quick, and all the time available learning environment to the learners.

Hilles and Naser [6] developed a tutoring system that teaches the fundamentals of the database known as MDB by using the ITSB authoring tool. The MDB teaches by conducting a series of lectures. To model the domain knowledge, the system had utilized constrained-based modeling and was capable of observing the actions of the students and learning capabilities.

Al-Hanjori et al. [7] designed and developed a computer-based coaching program that applies AI which helps students to learn computer networks. The system was developed using the ITSB tool. The system was able to provide educational content by presenting intelligently based on the information level, knowledge of the subject, preferred level of details, and assessment level of the student.

Jia [8] had developed a CSIEC English tutoring system which consists of a chatbot that chats with English learners anytime and anywhere. According to the user input, context, domain information, and common-sense information and with logic, the response was being generated by the bot. NLML forms were used to express the entire knowledge.

Danforth et al. [9] developed an intelligent medical tutoring system that simulates actual patients reliably; the system allowed the students to interact with patients in a native language to get an appropriate record of patient, symptoms, and so forth; and the system makes up diagnoses along with an appropriate treatment based upon the disease where these students can perform their professional practices in a simulated environment.

Holotescu [10] developed an educational MOOCBuddy chatbot which provides a personalized educational environment with MOOCS. The purpose of the chatbot was to assist the users by discovering news regarding MOOCs. The tutors were also able to incorporate MOOCs in their courses.

Dutta [11] proposed a web-based intelligent tutoring chatbot tool that was developed to assist high school students in order to learn their general knowledge subjects. Different chatbot platforms were evaluated: "api.ai, wit.ai, Luis.ai, and Pandorabots." On the basis of the evaluation results, Dialogflow.com (Api.ai) was selected to develop the chatbot. The intelligent chatbot was capable of engaging in small talks with the learners. Mahdi et al. [12] developed an information security tutoring system with the help of the "ITSB" tool. It helps the new learning students to become a good professionals in the area of security.

Hiremath et al. [13] developed a chatbot system for education which provides responses to user queries regarding education. To make the system scalable and highly interactive, user-friendly chatbots were using its local database as well as web database to provide responses. The techniques that were used by the chatbot system were ML, NLP, data processing algorithms, and pattern matching to increase the performance.

Kumar and Rose [14] proposed Basilica architecture for developing conversational agents that support collaborative learning where several learners interact with agents. The architecture carried out object-situated. The authors have created three explicit conversational specialists that were being created utilizing this design.

Graesser et al. [15] developed a conversational intelligent auto tutoring system that was holding diverse initiative conversational dialogue. The system helps college-level students to study regarding computer literacy. The system enhances the learning process by presenting complex contents to the learner which is being answered by students in English.

Holland et al. [16] developed J-LATTE, an intelligent constrained-based tutoring system that was teaching a subset of the java programming language. The system was composed of two modes, concept mode and code mode with concept mode the students designing the programs instead

of specifying contents of statements, whereas, with the coding mode, the students complete the code.

Shawar and Atwell [17] introduced machine learning methods for an Arabic chatbot, where the user input is being accepted in Arabic and responses are extracted from Quran, a java program that reads text from the corpus and converts it into A. I.M.L format used by the ALICE.

Doshi et al. [18] developed an intelligent Android chatbot application that has used program-O which is an AIML interpreter for the generation of the responses of users' input. The system works on text and voice mode and the response generation process is carried out in two phases, (1) preparation of pattern matching and (2) pattern matching behavior, and the chat system can answer the questions which are already trained in its dataset.

Alencar and Netto [19] developed TUCUMA, an intelligent virtual agent that carries out the tracking of students in the virtual environment. The system acts as distance learning and monitoring actions of the students and eliminates uncertainties through conversation while the system is composed of multiagents where these agents were responsible for producing gestures and monitoring of students' activities.

3. Proposed Framework

The proposed framework is composed of several components which will help to achieve the desired results as shown in Figure 1.

3.1. User Module. The user module consists of the user and user interface through which they will interact with the system.

3.2. Chat Agents Module. The agent module of the proposed framework is based on two major components: (1) agent manager and (2) chat agent A and chat agent B.

The agent manager is responsible for managing the overall conversational process between user and chat agents. When a user enters the query in its chat window and clicks the send button the agent manager is responsible to broadcast the user query to chat agent A and chat agent B, once the response for a given query is being generated by the chat agents and sent back to the agent manager which then makes it visible for the user.

3.3. Chat Agent A and Agent B. Chat agents A and B are the chatbots which act as tutors that will respond to the user query related to the trained educational domain or we can say that these agents produce an optimal response according to the user query from their KB. However, these agents are domain experts consisting of KB which acts as the brain of the chatbots.

3.4. Chatbots Knowledge Base AIML. Basically, AIML is an XML specification for programming chatbots like ALICE using program ab. AIML supports the development of high functioning chatbot system but it all depends upon how suspiciously we have mapped out the logic and conversational flow for our bot.

According to the Pandorabots report, over 300,000 chatbots have been developed using this platform. Pandorabots implement an AIML scripting language for the chatbots' development. However, the language powers the development of the most complex chatbots including multiaward winning, that is, Mitsuku chatbot [20]. AIML chatbots consist of AIML files.

AIML chatbots consist of the AIML files or we can say AIML KB is composed of AIML files which contain various AIML tags as mentioned in Figure 2. We can add the knowledge or train the agents by creating new files along with this AIML-based chatbot system extracting the responses from there.

The fundamental entity of knowledge inside AIML files is called categories where every category contains the user input which is a pattern and an output which is the response or template under which a given pattern is being matched and an optional context.

3.5. AIML Interpreter. Various AIML interpreters are there for different platforms, for example, Program O is an AIML interpreter for Android, Program ab is an AIML interpreter for Java, and so forth. AIML interpreters are capable of performing preprocessing functions in order to expand abbreviations, eliminate misspellings, and so forth. Before an AIML code is processed for the response generation, two important operations are being performed by the AIML interpreter program. These are (1) deperiodization and (2) normalization.

3.6. Graph Master. The Graph Master algorithm is being used by many AIML interpreters. Graph Master enables getting the most accurate or optimal time and provides control of stored memory.

In AIML, all categories are stored in a tree type, which is managed by an object called Graph Master as shown in Figure 3 while the Graph Master stores the AIML patterns along a path where each category is uniquely identified, which is from the root node "**r**" to a terminal node "**t**" where the AIML template is being stored.

3.7. AIML Pattern Matching. Basically, AIML employs the model of "pattern matching" or "pattern recognition"; AIML uses CBR case based reasoning where these cases are the categories the AIML finds the best match for a given pattern although the algorithm that the AIML implements for finding the best match for a given input is "k nearest neighbor classification" (KNN) as the model of learning within AIML is supervised learning [23].

Basically, in pattern recognition, the KNN algorithm calculates the distance between the test data and the given input where the input consists of k closest training

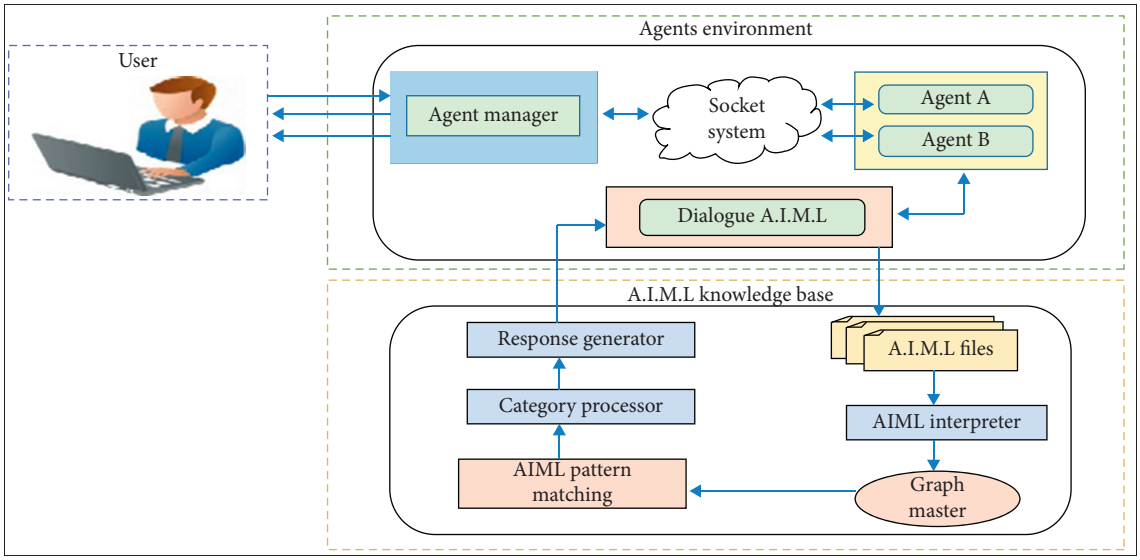


Figure 1: Multiagent chatbot communication framework.

<aiml>	Defines the beginning and end of an AIML document.
<category>	Defines the unit of knowledge in Alicebot's knowledge base.
<pattern>	Defines the pattern to match what a user may input lo an Alicebot.
<template>	Defines the response of an Alicebot to user's input.
<star>	Used to match wildcard * character (s) in the <pattern> Tag.
<srai>	Multipurpose tag used to call/match the other categories.
<random>	Used <random> to get random responses.
	Used to represent multiple responses.
<set>	Used to set value in an AIML variable.
<get>	Used to get the value stored in an AIML variable.
<that>	Used in AIML to respond based on the context.
<topic>	Used in AIML to store a context so that later conversation can be done based on that context.
<think>	Used in AIML to store a variable without notifying the user.
<condition>	Similar to switch statements in a programming language. It helps ALICE to respond to matching the input.

Figure 2: Basic AIML tags [21].

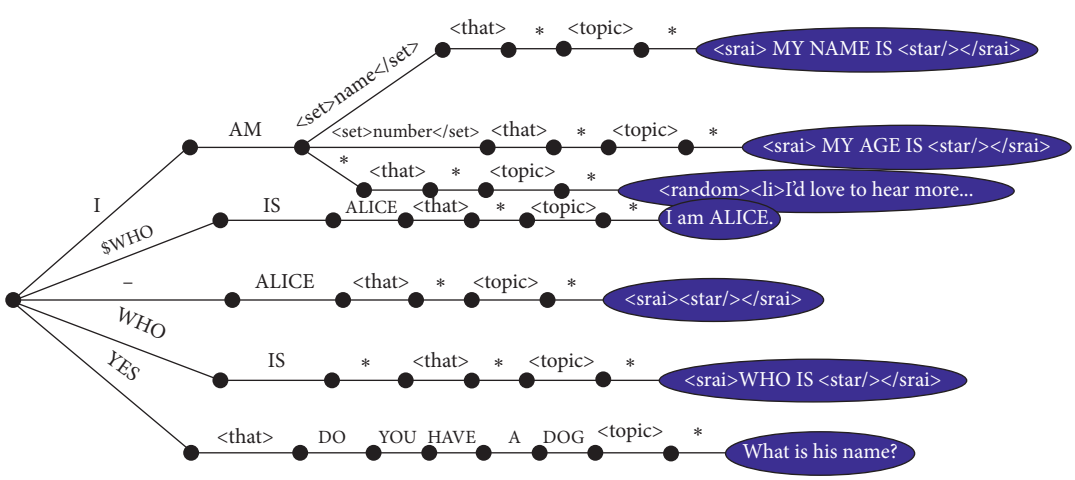


Figure 3: The working draft of an AIML Graph Master [22].

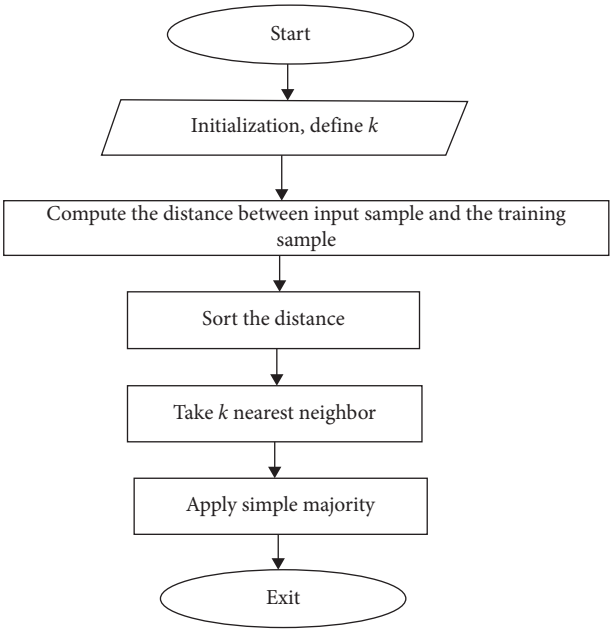


Figure 4: KNN classification algorithm [24].

examples and finds the closest match as shown in Figure 4.

3.8. Category Processor. Once the closest match is found for a given pattern, then the AIML category processor process categories within AIML files.

3.9. Response Generator. It reads the TEMPLATE and takes an appropriate action. The processing of dialog response which is being produced by a TEMPLATE diverges from being as simple as replicating the contents to the response [25].

3.10. Socket System. The client and server socket system establishes the Transmission Control Protocol (TCP) communication link among those Chat Agents in order to make them able to send and receive responses on a given port.

4. Results and Discussion

In this paper, we have presented an experimental study which analyzes how does multichat agents' communication will take place. Figure 5 shows the overall system flow chart; the system work flow chart is composed of various levels. The first step of the research flow chart is to collect the domain data in order to train our chatbot system once we collect the data for our desired domain; then, we prepare a possible list of questions with their corresponding responses.

After collecting domain information and preparation of questions and responses, the next level of the flow chart is to prepare an appropriate KB for our chatbot agents. Once the data training is completed, then, in the next level, we establish the communication link among those chat Agents in order to make them capable of sending and receiving information.

The next level of the flow chart is to generate the responses these chat agents use their KB and generate an appropriate response for a given input query. The last level of the flow chart is that we have performed testing on various data inputs so that to measure the accuracy and efficiency of our chat agents.

4.1. Data Description. The data for this study have been collected from various sources such as books, articles, websites, and OBE experts while the total training data sets are 500 categories as shown in Table 1.

The proposed system provides learners an interactive learning environment with multichatbot tutoring agents for solving their uncertainties.

The multichatbot system consists of the window with the text area reflecting the conversation and a textbox to introduce new requests so that the learners should be able to maintain the conversation with the proposed system to get better acurate results from this system.

Figure 6 shows the typical interface of a system where text-based conversation among the user, agent manager, and chat agent 1 and chat agent 2 and then accurate responses have been provided from chat agent 1 and chat agent 2 as shown in Figures 7 and 8.

5. System Evaluation

The multichatbot systems have been evaluated by performing testing on a range of inputs. Testing sample size, we have considered the "last 100 user messages to our chatbot system" using logs being marinated by the proposed system and the accuracy and other factors are being judged based on the responses generated by the chatbot system for the given user inputs. However, the system is being evaluated using the confusion matrix, which is a performance measurement tool for ML. Basically, it is a type of table which contains 4 diverse groups of predicted and actual values, and it is quite appropriate in this case, as the user's input will be classified and matched with the nearest matching category [26, 27].

In this scenario, the confusion matrix provides a way to reason about accuracy for our chatbot system. Let us classify the TP, TN, FP, and FN in terms of chatbot analogy. Correct mapping refers to the user's input been matched with the expected, or with proper category.

True positive which refers to a user input being matched to an expected category in the approved manner

True negative which refers to the user query not matched with an expected category as no such category is defined instead of generating nothing, it generates the default response

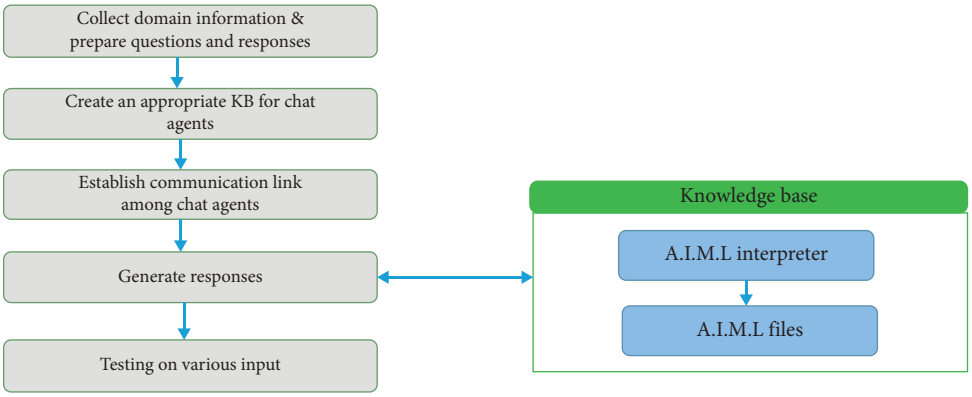


Figure 5: System development flow chart.

Table 1: Data description.

Total Categories	Total file	Total nodes	Total singletons	Total leaves	Total Branches
500	350 KB	2654	1954	453	2653

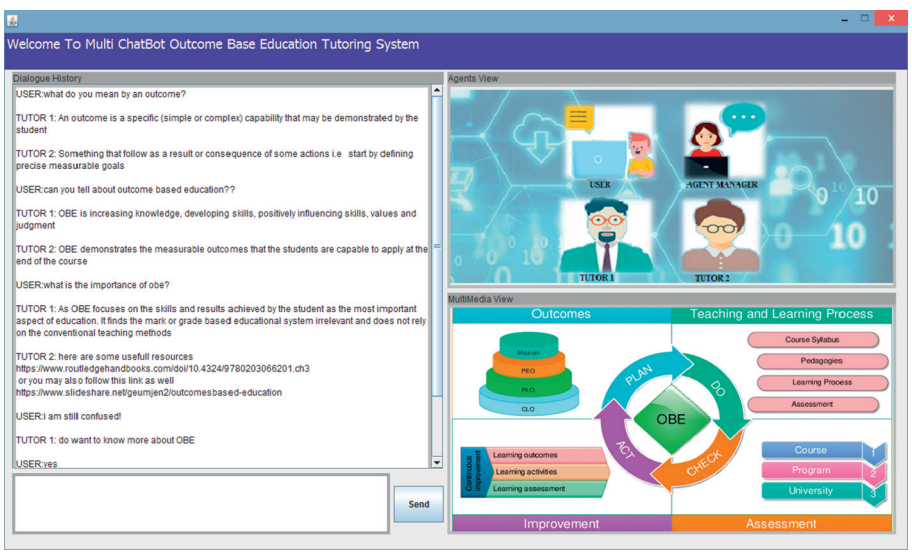


Figure 6: Conversational flow among chat agents and users.

False positive which refers to a user query matched either with another category or with the default category because we did not yet handle that phrase

False negative which refers to the user query not matched with the expected category as we have already defined a category to handle that phrase which means it does not work properly on that input

Figure 9 shows the total TP, TN, FP, and FN after the analyzing last 100 messages to our system.

Finally putting the given values for TP, TN, FP, and FN in a confusion matrix as shown in Figure 10, the results we got are as shown in Figure 11 while Figure 12 shows the overall accuracy of the responses generated by these chatbots.

5.1. System Security. The system utilizes the Java secure sockets layer (SSL) socket class in order to provide a secure communication link among the chat agents. However, SSL provides security using protocols such as the SSL or IETF Transport Layer Security (TLS) protocols.

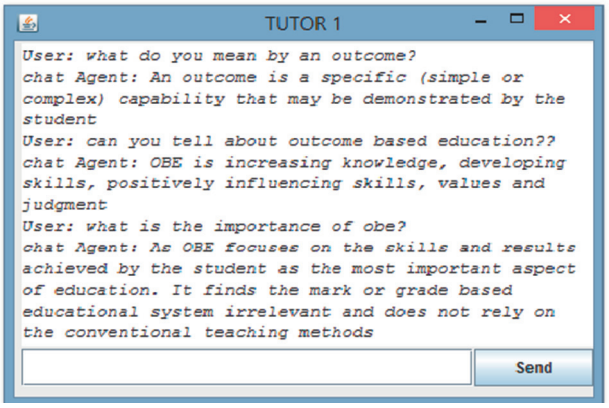


Figure 7: Response generation by Tutor 1.

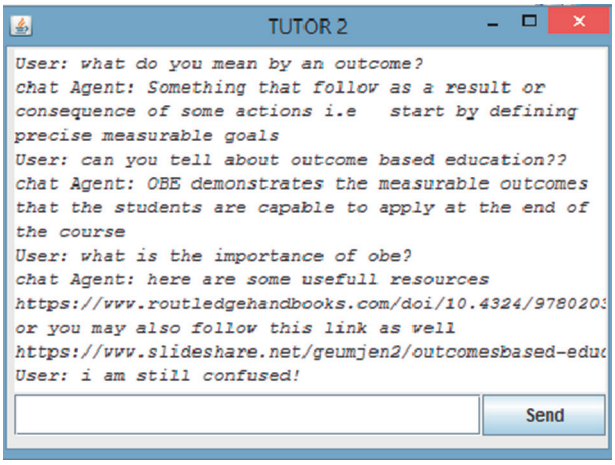


Figure 8: Response generation by Tutor 2.

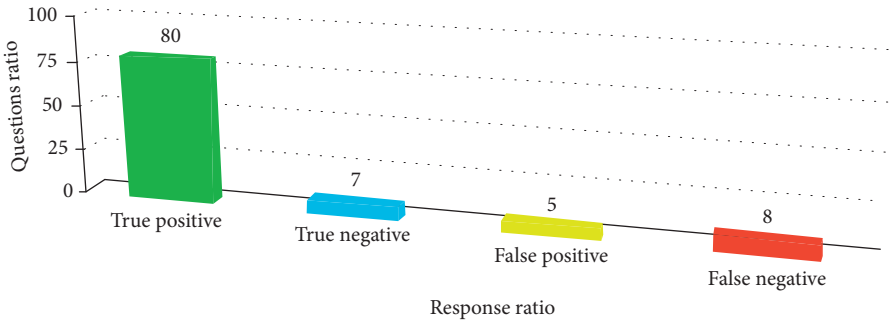


Figure 9: The total proportion of TP, TN, FP, and FN.

	True positive	True negative
Predicted positive	80	7
Predicted negative	5	8

Figure 10: The values which are inserted in the tool.

Measure	Value	Derivations
Sensitivity	0.9412	TPR = TP/(TP + FN)
Specificity	0.5333	SPC = TN/(FP + TN)
Precision	0.9195	PPV = TP /(TP + FP)
Negative predictive value	0.6154	NPV = TN/(TN + FN)
False positive rate	0.4667	FPR = FP/(FP + TN)
False discovery rate	0.0805	FDR = FP/(FP + TP)
False negative rate	0.0588	FNR = FN/(FN + TP)
Accuracy	0.8800	ACC = (TP + TN)/(P + N)
F1 score	0.9302	F1 = 2TP/(2TP + FP + FN)
Matthews correlation coefficient	0.5038	TP*TN – FP*FN/sqrt ((TP + FP)*(TP + FN)*(TN + FP)*(TN + FN)

Figure 11: The results generated by the confusion Matrix tool [28].

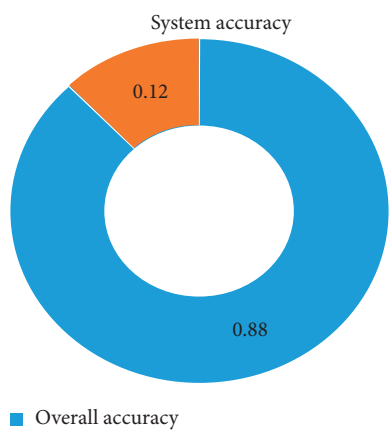


Figure 12: The overall accuracy of the system.

6. Conclusion

In this paper, we have proposed the framework with a novel approach that supports the development of a multi-interactive chatbot's system for an educational area using A.I.M.L 2.0. The system will facilitate the students for their learning towards an outcome-based education domain. The proposed system composed of user module consists of user and user interface, chat agents module which will respond to the user query, chatbot KB acting as the brain for the chatbot system, and socket system for establishing the communication link. The system has been evaluated by analyzing the last 100 user messages to our chatbot system. Using the confusion matrix, we found promising results with an overall accuracy of about 88'%. By and large, chatbots can be utilized to give essential lectures. The goal is that chatbots can fill in as virtual tutors and that in the process they adjust to the capabilities of the students. However, it has been implemented in java. The outcomes of this research will be useful for the education sector where these intelligent systems will help the students in schools, universities, and other training scenarios. The emergence of artificial intelligence applications, such as text-based virtual assistants (chatbots) especially in education, is a very new field. These types of systems can be useful in helping teachers and students to solve their educational problems and routine tasks efficiently. This article basically is the foundation for designing and developing the chatbots for educational systems and may provide justifiable results when applying in a particular situation in future implementations [29].

7. Future Work

In this study, these multichatbots systems can respond to text-based dialogues only but, in the future, these chat agents will interpret voice commands as well as images.

Furthermore, we can expand the scope of these chatbots for various domains as well. We can implement this framework in a mixed real enviroment.

Data Availability

The authors have used their own data. Data can be obtained through contacting jalbaniakhtar@gmail.com.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Schlesinger, K. P. O'Hara, and A. S. Taylor, "Let's talk about race: identity, chatbots, and AI," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, Montreal, QC, Canada, 2018 April.

[2] B. A. Shawar and E. Atwell, "Chatbots: are they really useful?" *LDV Forum*, vol. 22, no. 1, pp. 29–49, 2007.

[3] A. Kerlyl, P. Hall, and S. Bull, "Bringing chatbots into education: towards natural language negotiation of open learner models," in *Proceedings of International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pp. 179–192, Springer, London, UK, 2006 December.

[4] A. A. Georgescu, "Chatbots for education–trends, benefits and challenges," in *Proceedings of the Conference proceedings of eLearning and Software for Education (eLSE)*, vol. 2, no. 14, pp. 195–200, "Carol I" National Defence University Publishing House, Bucharest, Romania, 2018.

[5] Z. Peng and X. Ma, "A survey on construction and enhancement methods in service chatbots design," *CCF Transactions on Pervasive Computing and Interaction*, vol. 1, no. 3, pp. 204–223, 2019.

[6] M. M. Hilles and S. S. A. Naser, "Knowledge-based intelligent tutoring system for teaching mongo database," *European Academic Research*, vol. 4, no. 10, pp. 8783–8794, 2017.

[7] M. M. Al-Hanjori, M. Z. Shaath, and S. S. A. Naser, "Learning computer networks using intelligent tutoring system," *International Journal of Advanced Research and Development*, vol. 2, no. 1, 2017.

[8] J. Jia, "CSIEC: a computer assisted English learning chatbot based on textual knowledge and reasoning," *Knowledge-Based Systems*, vol. 22, no. 4, pp. 249–255, 2009.

[9] D. R. Danforth, M. Procter, R. Chen, M. Johnson, and R. Heller, "Development of virtual patient simulations for medical education," *Journal For Virtual Worlds Research*, vol. 2, no. 2, 2009.

[10] C. Holotescu, *MOOCBuddy: a Chatbot for Personalized Learning with MOOCs*, RoCHI, Pune, India, 2016.

[11] D. Dutta, *Developing an Intelligent Chat-Bot Tool to Assist High School Students for Learning General Knowledge Subjects*, Georgia Institute of Technology, Atlanta, Georgia, 2017.

[12] A. O. Mahdi, M. I. Alhabbash, and S. S. A. Naser, "An intelligent tutoring system for teaching advanced topics in information security," *Wo Rld Wide Journal of Multidisci Plinary Rese Arch and Development*, vol. 2, no. 12, pp. 1–9, 2016.

[13] G. Hiremath, A. Hajare, P. Bhosale, R. Nanaware, and K. S. Wagh, "Chatbot for education system," 2018.

[14] R. Kumar and C. P. Rose, "Architecture for building conversational agents that support collaborative learning," *IEEE Transactions on Learning Technologies*, vol. 4, no. 1, pp. 21–34, 2011.

[15] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan, and D. Harter, "Intelligent tutoring systems with conversational dialogue," *AI Magazine*, vol. 22, no. 4, p. 39, 2001.

[16] J. Holland, A. Mitrovic, and B. Martin, *J-LATTE: A Constraint-Based Tutor for Java*, University of Canterbury, Christchurch, New Zealand, 2009.

[17] A. Shawar and E. S. Atwell, "An Arabic chatbot giving answers from the Qur'an," in *Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles*, vol. 2, pp. 197–202, Fez, Morocco, April 2004.

[18] S. V. Doshi, S. B. Pawar, A. G. Shelar, and S. S. Kulkarni, "Artificial intelligence chatbot in android system using open source program-O," *Artificial Intelligence*, vol. 6, no. 4, 2017.

[19] M. Alencar and J. F. Netto, "TUtor collaborator using multi-agent system," in *International Conference on Collaboration Technologies*, pp. 153–159, Springer, Berlin, Heidelberg, 2014 September.

[20] A. Khanna, *Pandorabots Chatbot Hosting Platform*, SARANG Bot, Ratnagiri, Maharashtra, India, 2015.

[21] https://medium.com/@pemagrg/aiml-tutorial-a8802830f2bf%20 (Last Accessed on 20-10-2019).

[22] https://gist.github.com/onlurking/f6431e672cfa202c09a7c7cf92ac8a8b (Last Accessed on 21-4-2019).

[23] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient knn classification with different numbers of nearest neighbors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774–1785, 2017.

[24] https://www.slideshare.net/15koolneha/knearest-neighbor-classifier-74464857%20 (Last Accessed on 21-4-2019).

[25] R. Jindal, R. Kumar, R. Sahajpal, S. Sofat, and S. Singh, "Implementing a natural language conversational interface for Indian language computing," *IETE Technical Review*, vol. 21, no. 4, pp. 243–250, 2004.

[26] https://towardsdatascience.com/understanding-confusion-matrix%20a9ad42dcfd62?gi=6702c33e3ed6 (Last Accessed on 20-3-2019).

[27] https://miningbusinessdata.com/how-do-you-measure-your-dialogflow-bots%20accuracy (Last Accessed on 21-3-2019).

[28] http://onlineconfusionmatrix.com (Last Accessed on 22-3-2019).

[29] R. Chocarro, M. Cortiñas, and G. Marcos-Matás, "Teachers' attitudes towards chatbots in education: a technology acceptance model approach considering the effect of social language, bot proactiveness, and users' characteristics," *Educational Studies*, vol. 40, pp. 1–19, 2021.

Hindawi
Scientific Programming
Volume 2021, Article ID 6641832, 15 pages
https://doi.org/10.1155/2021/6641832

Review Article

Bidirectional Language Modeling: A Systematic Literature Review

Muhammad Shah Jahan,[1] **Habib Ullah Khan**,[2] **Shahzad Akbar**,[3] **Muhammad Umar Farooq**,[1] **Sarah Gul**,[4] **and Anam Amjad**[1]

[1]*Department of Computer Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Islamabad, 44000, Pakistan*
[2]*Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, Qatar*
[3]*Riphah College of Computing, Riphah International University, Faisalabad Campus, Faisalabad 3800, Pakistan*
[4]*Department of Biological Sciences, FBAS, International Islamic University, Islamabad, Pakistan*

Correspondence should be addressed to Shahzad Akbar; shahzadakbarbzu@gmail.com

Received 27 December 2020; Revised 21 April 2021; Accepted 26 April 2021; Published 3 May 2021

Academic Editor: Fabrizio Riguzzi

In transfer learning, two major activities, i.e., pretraining and fine-tuning, are carried out to perform downstream tasks. The advent of transformer architecture and bidirectional language models, e.g., bidirectional encoder representation from transformer (BERT), enables the functionality of transfer learning. Besides, BERT bridges the limitations of unidirectional language models by removing the dependency on the recurrent neural network (RNN). BERT also supports the attention mechanism to read input from any side and understand sentence context better. It is analyzed that the performance of downstream tasks in transfer learning depends upon the various factors such as dataset size, step size, and the number of selected parameters. In state-of-the-art, various research studies produced efficient results by contributing to the pretraining phase. However, a comprehensive investigation and analysis of these research studies is not available yet. Therefore, in this article, a systematic literature review (SLR) is presented investigating thirty-one (31) influential research studies published during 2018–2020. Following contributions are made in this paper: (1) thirty-one (31) models inspired by BERT are extracted. (2) Every model in this paper is compared with RoBERTa (replicated BERT model) having large dataset and batch size but with a small step size. It is concluded that seven (7) out of thirty-one (31) models in this SLR outperforms RoBERTa in which three were trained on a larger dataset while the other four models are trained on a smaller dataset. Besides, among these seven models, six models shared both feedforward network (FFN) and attention across the layers. Rest of the twenty-four (24) models are also studied in this SLR with different parameter settings. Furthermore, it has been concluded that a pretrained model with a large dataset, hidden layers, attention heads, and small step size with parameter sharing produces better results. This SLR will help researchers to pick a suitable model based on their requirements.

1. Introduction

Transfer learning encompasses the model training on large text corpus and utilization of obtained knowledge to downstream tasks [1]. Before the emergence of transformer architecture for transfer learning, unidirectional language models were used extensively but these models faced many limitations such as reliance on unidirectional recurrent neural network (RNN) architecture and limited context vector size. To overcome these gaps, bidirectional language models such that bidirectional encoder representation from transformer (BERT) is introduced to improve the performance of downstream tasks. Bidirectional language models can be applied in a wide variety of tasks such as natural language inference (NLI) [2, 3], paraphrasing at sentence-level [4], Question Answering (QA) systems, and entity recognition at token level [5]. In the beginning, pretraining of bidirectional language models was done via supervised learning [6] but human-labeled datasets are limited. To resolve this issue, the use of a large corpus-based unsupervised learning increased.

Language models are one of the most crucial components of natural language processing (NLP). A language model provides context to distinguish between words and phrases that sound alike in English such as "recognize speech" and "wreck a nice beach" but indeed very different. The language model is a probability distribution over sequences of words and used in information retrieval. There are many types of language models including n-gram, exponential neural network (ENN), and bidirectional. These language models are the backbone of Google Assistant, Amazon's Alexa, and Apple's Siri to analyze the data for the prediction of words. BERT is first deep bidirectional language models based on transformer architecture which means it reads the input from both sides left-to-right and right-to-left while existing models were unidirectional and just read the input from one side. BERT outperforms all existing models.

A large amount of data [7] such as text corpus, domain-specific data (e.g., PubMed, PubMed Central (PMC) [8]), and scientific dataset [9] is available for unsupervised learning. Also, different sentence tokens, e.g., span [10], semantic [11–13], lexical [14], and syntactic [15], are used to pretrain the models. In general, large pretraining objective, unlabelled datasets [16, 17], benchmarks [18, 19], and fine-tuning methods [20, 21] are beneficial in unsupervised learning. Pretrained models developed using unsupervised learning have produced state-of-the-art results due to better use of parallel computing. The resultant models are not only applicable to computer domains but also used in other specific domains, e.g., [22] business [23], medical [24, 25], and science [26]. The performance of downstream tasks directly depends on pretraining of the models which subsequently considers many significant factors such as dataset size, batch size, step size, sequence size, parameters, layers, hidden layers, attention heads, and cross-layer sharing for practical implications. These factors are used in different research studies to get better results of pretrained models, but there is no study available to the best of our knowledge which provides comprehensive review to these research studies. This paper attempts to find answers to the following five research questions (RQs) as follows:

RQ1: what are the significant model types and techniques used for sentence embedding learning?

RQ2: what is the effect of dataset size with different batch, step, and sequence size on the performance of the pretrained model?

RQ3: what is the effect of parameters with different input layers, hidden layers, and attention heads on the performance of the pretrained model in downstream tasks?

RQ4: what are the effective techniques for cross-layer parameter sharing in the pretraining of models?

RQ5: what are the leading datasets used in the pretraining of models?

To find answers to these research questions, we performed an exhaustive systematic literature review (SLR) of thirty-one (31) research papers as presented in Table 1. The contributions of this paper are as follows:

(i) Firstly, this research study discovers all bidirectional language models built upon transformer or Transformer-XL architecture during 2018–2020.

(ii) Secondly, all the important settings of the pretrained model such as size of the dataset, batch size, step size, sequence size, parameters, layers, hidden layers, attention heads, and cross-layer sharing are recognized in this paper.

(iii) Every model is compared with RoBERT that is a replicated BERT model with a large dataset and batch size but with a small step size. The analysis of existing models with RoBERTa is also carried out in this SLR.

Rest of the paper is organized as follows: in Section 2, the research methodology is developed which consists of selection and rejection criteria, search process, quality assessment criteria, data extraction, and synthesis. Section 3 and Section 4 present the results and answers of the five developed questions, respectively. Section 5 discusses the analysis of the selected research studies. Section 6 provides recommendations to the existing research studies. Lastly, Section 7 concludes the whole research study and provides future directions.

2. Research Methodology

This research study is performed based on the guidelines of the systematic literature review standard. Following features that distinguish the systematic literature review from conventional literature review are as follows [53]:

(i) To begin, review protocol is developed based on the research questions.

(ii) Selection and rejection criteria are developed to assess each primary study.

(iii) Search strategy is defined to provide the addition of the most relevant literature in the SLR. It is documented to ensure the completeness of research study.

(iv) Information from each research study is evaluated using quality assessment criteria.

(v) To perform quantitative meta-analysis, review protocol turns out to be the prerequisite.

(vi) This review protocol establishes the basis of SLR due to which it becomes possible to identify the research gaps from the selected area so that new research activities can be positioned.

Two sections (Background and Research Questions) are already provided in the introduction section. Therefore, we

Table 1: Overall hyperparameters.

Paper	Batch size	Max sequence	Learning rate	Step size	Parameters	Layers	Hidden	Attention head
[17]	2K	512	1e − 6	125K	360	24	1024	16
[10]	256	128	1e − 4	2.4M	340	24	1024	16
[11]	32	128	2e − 5	1M	340	24	1024	16
[14]	512	256	5e − 5	1M	114	6	768	12
[15]	400K	256	5e − 5	4K	114	24	1024	16
[27]	256	128	1e − 4	1M	110	12	768	12
[28]	2048	512	1e − 5	500K	340	24	1024	16
[29]	330	512	3e − 5	777K	340	24	1024	16
[20]	32	512	1e − 4	1M	330	24	1024	16
[30]	32	512	1e − 4	1M	340	24	1024	16
[31]	256	128	1	1M	14.5	4	312	12
[32]	256	128	1e − 4	1M	340	24	1024	16
[33]	4096	512	0.00176	125K	233	12	4096	128
[34]	1024	128	1.0e − 4	1M	3.9	48	2560	40
[35]	4096	512	0.00176	125K	233	12	4096	64
[36]	2048	128	0.01	2.1M	11	12	768	12
[37]	2K	512	10^{-3}	125K	356	24	1024	16
[38]	8K	512	1e − 6	500K	360	24	1024	16
[39]	32	128	2e − 5	1M	340	12	768	12
[40]	2048	512	2e − 4	1.75M	335	24	1024	16
[41]	1024	512	5e − 4	400k	33	12	768	12
[42]	32	256	2^{-5} to10^{-5}	1M	66	6	768	12
[43]	256	128	1e − 4	1M	108	12	768	12
[44]	128	128	1e − 4	1M	340	24	1024	16
[45]	256	128	1e − 4	1M	110	12	768	12
[46]	256	128	1e − 4	1M	340	24	1024	16
[47]	256	128	5^{-5} to 10^{-5}	1M	110	12	768	12
[48]	128	512	3e − 4	50K	9.5	24	1024	16
[49]	6	512	1.5e − 5	1M	340	24	1024	16
[50]	8000	512	1e − 6	500K	400	12	1024	12
[51]	5120	128	1.8e − 4	25K	340	24	1024	16
[52]	7680	128	6e − 4	0.5M	110	12	768	12

are omitting both sections and will describe the other four elements in subsequent sections.

2.1. Selection and Rejection Criteria. We defined logical rules for the selection and rejection of the research papers to achieve the objectives of SLR. These rules are as follows:

(i) The selected research studies must target the bidirectional language modeling and the BERT model.

(ii) Selected research studies for this SLR must be published between 2018 and 2020.

(iii) All the research studies selected in this SLR must be from one of these four scientific repositories, i.e., arXiv, Elsevier, ACM, IEEE, and two conferences including NIPS (neural information processing systems) and MLR (machine learning research).

(iv) Duplicate research studies are not selected. Similar content shared by more than one research study is discarded.

2.2. Search Process. Four scientific repositories and two conferences mentioned in Section 2.1 initiated the search process. Five defined keywords, i.e., (1) bidirectional language modeling, (2) pretrained language modeling, (3) biLM, (4) BERT, and (5) transformer, are used to search the research studies as shown in Table 2. We only use AND operator while searching because without AND operator, some keywords produced irrelevant searches. We also used some advanced options provided by databases to refine the search result. For example, while searching for research studies on Science Direct with the keyword "transformer," we receive a lot of results because "transformer" belongs to other domains as well. To generate relevant results, an advanced option is used for publication titles such as "Science of Computer Programming."

We used open coding like process which involves three phases (Phase 1, 2, and 3) and three authors (A1, A2, and A3):

(i) Phase 1: A3 selects all papers which are from mentioned databases.

(ii) Phase 2: A2 checks all papers selected in Phase 1 and checks either these papers are published in between 2018 and 2020.

(iii) Phase 3: A1 selects all the papers provided at the end of Phase 2 which targeted the bidirectional language modeling or share its properties. Phases 1 and 2 are

Table 2: Number of results using keywords.

Sr. no.	Keywords	Operator	Scientific repositories					
		IEEE	ACM	arXiv	Elsevier	NIPS	MLR	
1	Bidirectional language modeling	AND	5	1	24	6	0	1
2	Pretrained language modeling	AND	2	2	18	2	0	0
3	biLM	N/A	1	3	1	1	0	0
4	BERT	N/A	16	6	24	1	1	1
5	Transformer	N/A	1	4	4	0	0	0

straight forward, but in Phase 3, if any of other two authors A1 or A2 had any disagreement, then a voting procedure was followed and majority wins.

In Figure 1, overview of the search process is illustrated. By using keywords shown in Table 2, we received total 94,954 results:

(i) We have rejected 92,876 papers based on the title of the research studies.

(ii) Among 92876, we rejected another 1589 papers by applying selection and rejection criteria on abstract.

(iii) Another 364 research studies are discarded by performing the general study of papers.

(iv) Lastly, after detailed study of 127 research studies, 96 research studies are eliminated and only thirty-one (31) relevant research studies are selected to perform SLR.

2.3. Quality Assessment. For the reliable outcome of the SLR, a quality assessment checklist (QA 1 to QA 5) is developed. Every paper included in this study must satisfy the assessment criteria to ensure high-quality of the selected research studies by answering a few questions in Table 3.

All selected research studies target bidirectional language models or BERT by either using or improving these models. The selected repository-based distribution of research studies is shown in Figure 2. All of the papers are from internationally recognized scientific repositories such as arXiv, IEEE, ACM, Elsevier, NIPS, and MLR. We included arXiv with other databases because most of the work on the bidirectional language model is published in arXiv. Almost all top LMs are developed by big technology organizations, and their work is published in arXiv. It can be analyzed from Figure 2 that arXiv is the most cited database. Also, it is ensured that the selected research study must answer at least one question. We have developed this checklist presented in Table 3 to ensure high-quality findings of our research studies.

2.4. Data Extraction and Synthesis. A template for data extraction and data synthesis is developed in Table 4 to answer the research questions. Data extraction is used to extract the specific and most related data based on selection and rejection criteria (Section 2.1). For data extraction and synthesis, we have extracted the bibliography of the paper and then core findings of the paper such as methodology, pretraining, fine-tuning, and the results are synthesized. We have performed the data synthesis to answer our developed research questions for this SLR.

3. Results

After applying the review protocol (see Section 2), thirty-one (31) research studies published during 2018–2020 are selected to conduct this SLR. We compared all models with RoBERTa [17] which is the replication of BERT [27] with a large dataset, batch size, sequence size, parameter, layers, hidden layers, attention head but with small step size, no parameter sharing, and no sentence representation learning. The main advantage of comparison with RoBERTa is that it is a model built on BERT with slightly changed parameters and can generate fair comparison for all other models used in this research. In this section, the dataset with other parameters effecting the pretraining of models, results based on model structure, pretraining objectives, sharing parameters in pretraining, and model selection for testing are discussed in detail.

3.1. Model Type and Technique Used for Sentence Embedding Learning. We identified four (4) types of pretraining objectives for language representation and three (3) types of sentence representation learning presented in Table 5. Pretraining objectives are as follows: (1) autoencoding is a model type in which the model reconstructs the original data from corrupted inputs. (2) Autoregressive uses the probability distribution that remembers previous states while partially autoregressive uses only one previous state. (3) Autoencoding and autoregressive present objective in which corrupting and knowledge of previous values preserve. (4) Autoencoding and partially autoregressive present objective in which corrupting and partially knowledge of previous values preserve.

Three (3) sentence representation learning tasks are discussed in terms of pretraining objectives in Table 5: Next Sentence Prediction (NSP), Sentence Order Prediction (SOP), and None. (1) NSP is a binary classification that predicts whether two segments that appear consecutively are from the same document. (2) SOP focuses on intersentence coherence with positive examples the same as NSP but negative examples are different and achieved by swapping the documents. (3) If a model used None it means neither NSP nor SOP is used.

3.2. Pretraining Setup. We have divided the pretraining dataset into four categories presented in Table 6 with respect

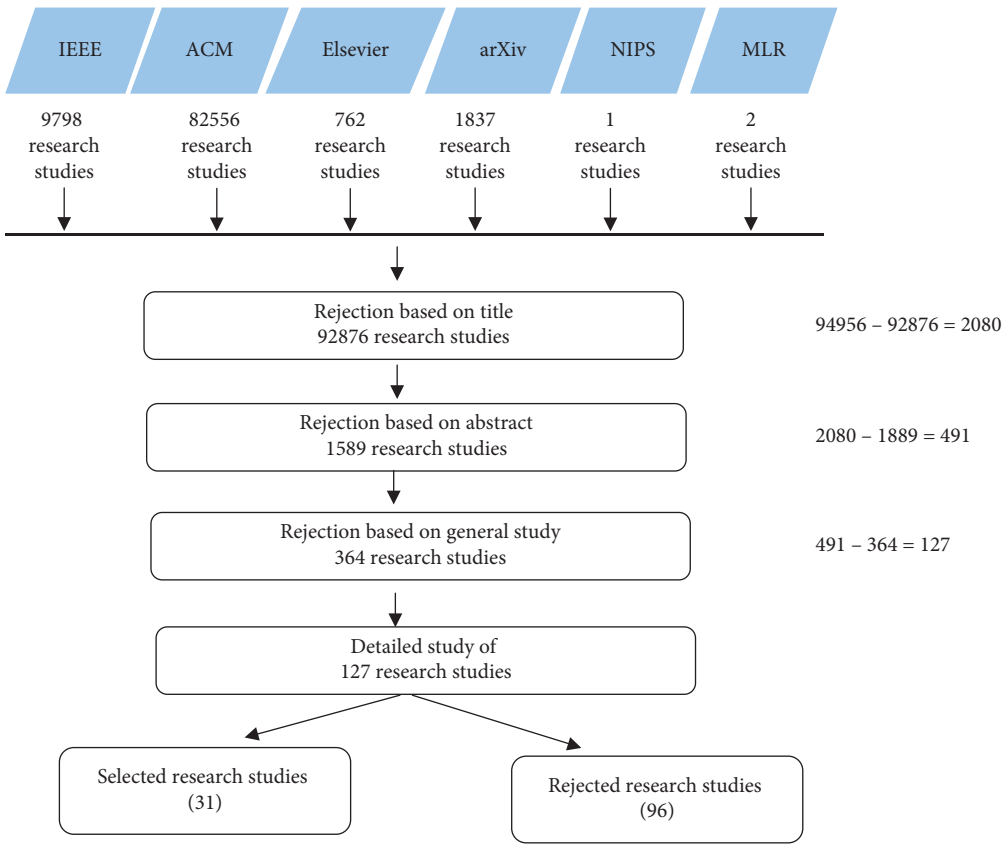


Figure 1: Overview of the search process.

Table 3: Quality assessment checklist.

Sr. no.	Quality assessment checklist
QA 1	Are models using bidirectional language modeling in selected research studies?
QA 2	Do all the papers use either BERT or improve it?
QA 3	Are selected research studies published from 2018 to 2020?
QA 4	Do the selected research papers are from the scientific repositories, NIPS and MLR?
QA 5	Do the selected research papers provide the required answers for developed research questions?

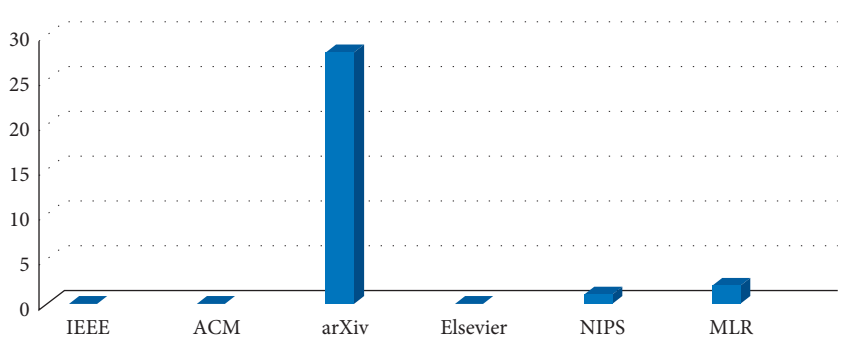


Figure 2: Summary of selected papers.

Table 4: Data extraction and synthesis.

Sr. no.	Description	Details
1	Bibliographic information	Authors, title, research type, publication year, etc.
2	Methodology	The main structure of our study is to extract the methodology of the paper.
3	Pretraining	Pretraining structure of each study is thoroughly analyzed.
4	Fine-tuning	Fine-tuning structure of each study is thoroughly analyzed.
5	Dataset	Datasets used in the selected research studies are identified.

Table 5: Model type and sentence representation learning.

Pretraining objectives	Sentence representation learning		
	NSP	SOP	None
Autoencoding (AE)	[14, 27, 29–32]	[33–35]	[10, 11, 15, 20, 36–49]
Autoregressive (AR)	—	—	[50]
Autoencoding and autoregressive	[51]	—	[28]
Autoencoding and partially autoregressive (PAR)	—	—	[52]

to size such as (1) 1 GB to 99 GB, (2) 100 GB to 149 GB, (3) 150 GB to 199GB, and (4) 200 GB to onwards. The main advantage of this categorization helps in visualization of dataset size used by different language models and the relation of dataset size with step size, batch size, and sequence size. Step size indicates how many steps a program will run and takes data points with respect to time. Batch size is the number of examples that can be utilized in one iteration. Sequence size defines the maximum size of an input. If one increases the sequence size, then it required a lot of computational power and resources to pretrain an LM. We divide the batch and step size into three categories (small, big, and same) while sequence size has two categories (small and same). We make two categories of sequence size because no model has a bigger sequence size than 512. Sequence size of 512 is used by RoBERTa [17] which has 160 GB of training dataset size with 2 K (small) of batch size and 125 K (medium) step size. The comparison of performance of different models is done using GLUE leaderboard. GLUE consists of ten (10) diversified tasks, but we use results of eight (8) of these tasks and leave WNLI and AX due to completely different behavior of these two tasks. Subsequently, SQuAD and then RACE are also used for comparison.

3.3. Effect of Parameters with Different Layers, Hidden Layers, and Attention Heads. As shown in Table 7, we have divided the parameter size of models into seven categories such as (1) 1 M to 99 M, (2) 100 M to 199 M, (3) 200 M to 199 M, (4) 301 M to 349 M, (5) 350 M to 399 M, (6) 400 M to 500 M, and (7) 501M to onwards. Every category contains the parameters used by models in pretraining. We have three categories for parameter size: (1) layers, (2) hidden layers, and (3) attention head, and every category has three subcategories: (1) less, (2) more, and (3) same. All results are compared against RoBERTa [17] which has 360 M parameters, twenty-four (24) layers, 1024 hidden layers, and sixteen (16) attention heads.

3.4. Cross-Layer Parameter Sharing. Cross-layer parameter sharing is a process in which models share some parameters during pretraining with purpose of gaining knowledge. We divide the cross-layer parameter sharing into four different categories as shown in Table 8. (1) In all-shared means both feedforward network (FFN) and attention are shared across layers, (2) in shared-attention, only attention is being shared, (3) in shared-FFN, only FFN is being shared across layers, and (4) not-shared shares nothing during pretraining. For the selected parameters, results are shown in Table 8.

3.5. Dataset Used. In this section, the datasets used are mentioned so that new models could be tested using these datasets. Four datasets are identified: (1) General Language Understanding Evaluation (GLUE): it consists of ten diversified tasks. Some tasks are single-sentence classification and some are sentence-pair classification. GLUE provides split training and testing data to test the performance of pretrained models. It allows us to submit our submissions on the GLUE leaderboard and compare our evaluation results on private held-out test data; (2) Bilingual Evaluation Understudy (BLEU) is used to evaluate the quality of machine translation text from one language to another; (3) the Stanford Question Answering Dataset (SQuAD) has two datasets with SQuAD v1.1 and SQuAD v2.0 with one has answerable questions and the other has unanswerable. SQuADv1.1 consists of 100 K questions while SQuADv2.0 consists of 150 K questions; (4) RACE is a comprehension dataset consists of 28 K passages and 100 K questions.

4. Answers to Research Questions

RQ1: what are the significant model types and techniques used for sentence embedding learning?

Answer: as shown in Table 5, in thirty-one (31) identified models, only six models, ERNIE [14], BERT [27], UniLM [29], StructBERT [30], TinyBERT [31], and MT-DNN [32], use NSP as sentence learning technique while three models, ALBERT [33], Megatron-LM [34], and AlBERT (xxlarge-ensemble) [35], use SOP. It could be seen that twenty-one (21) out of thirty-one (31) research studies do not use any sentence embedding learning which means both NSP and SOP decrease the performance of these models.

Different models have different pretraining objectives such as BERT [50] is an autoregressive model without using NSP. Nezha [51] uses autoencoding and autoregressive with NSP while XLNet [28] uses autoencoding and autoregressive without NSP. Another model, UniLMv2 [52], is an autoencoding and partially autoregressive model. Rest of the models, i.e., twenty-seven (27) out of thirty-one (31) models, use autoencoding that is the most used pretraining objective with none sentence learning technique.

RQ2: what is the effect of dataset size with different batch, step, and sequence size on the performance of the pretrained model?

Answer: as shown in Table 6, seven (7) out of thirty-one (31) models, i.e., [15, 33–36, 38, 40], have outperformed RoBERTa. Among these seven models, three models [34, 36, 38] were trained on a larger dataset than RoBERTa while other three models [15, 33, 35] are trained on a smaller dataset whereas [40] trained on 126 GB dataset size which is

Table 6: Effect of training data size with different batch, sequence, and step size.

Training data size	Batch size			Sequence size		Step size			Performance	
	Small	Big	Same	Small	Same	Small	Big	Same	Slow	Better
200 GB-onwards	—	—	[36]	[36]	—	—	[36]	—	—	[36]
150–199 GB	[34]	[38, 50, 52]	[37]	[34, 52]	[37, 38, 50]	—	[34, 38, 50, 52]	[37]	[37, 50, 52]	[34, 38]
100–149 GB	—		[28, 40]	—	[28, 40]	—	[28, 40]	—	[28]	[40]
1–99 GB	[10, 11, 14, 20, 27, 29–32, 39, 41–49]	[15, 33, 35, 51]	—	[10, 11, 14, 15, 27, 32, 39, 42–47, 51]	[20, 29, 30, 33, 35, 41, 48, 49]	[15, 39, 48, 51]	[10, 11, 14, 20, 27, 29, 31, 32, 41–47, 49]	[30, 33, 35]	[10, 11, 14, 20, 27, 29–32, 39, 41–49, 51]	[15, 33, 35]

Table 7: Effect of parameters with different layers, hidden layers, and attention heads.

Parameters	Layers			Hidden			Attention head		
	Less	More	Same	Less	More	Same	Less	More	Same
1–99 M	[31, 41, 42]	—	[48]	[31, 41, 42]	—	[48]	[31, 41, 42]	—	[48]
100–199 M	[14, 27, 43, 45, 47, 52]	—	[15]	[14, 27, 43, 45, 47, 52]	—	[15]	[14, 27, 43, 45, 47, 52]	—	[15]
<200–300>M	[33, 35]	—	—	—	[33, 35]	—	—	[33, 35]	—
301–349M	[20, 39]	—	[10,28–30, 32, 40, 44, 46, 49, 51]	[39]	—	[10, 20, 28–30, 32, 40, 44, 46, 49, 51]	[39]	—	[10, 20, 28–30, 32, 40, 44, 46, 49, 51]
350–399	—	—	[11, 37, 38]	—	—	[11, 37, 38]	—	—	[11, 37, 38]
400–500	[50]	—	—	—	—	[50]	[50]	—	—
501-Ow	[36]	[34]	—	[36]	[34]		[36]	[34]	—

Table 8: Cross-layer parameter sharing.

Cross-layer parameter sharing	Paper	Performance	
		Decrease	Increase
All-shared	[15, 20, 29, 32–36, 40, 44, 46, 47, 51]	[20, 29, 32, 44, 46, 47, 51]	[15, 33–36, 40]
Shared-attention	[28]	[28]	—
Shared-FFN	[48]	[48]	—
Not-shared	[10, 11, 14, 27, 30, 31, 37–39, 41–43, 45, 49, 50, 52]	[10, 11, 14, 27, 30, 31, 37, 41–43, 45, 49, 50, 52]	[38]

close to RoBERTa of 160 GB. However, other three out of thirty-one (31) models [37, 50, 52] trained on the same size of the dataset but perform lesser than RoBERTa due to the large or same step size. Rest of the twenty-four (24) models trained on smaller datasets and among these 23 models, three models outperform the RoBERTa and all of these models use bigger batch size and same or small step size. A model trained on a larger dataset with larger batch size and smaller step size will generate better outcomes and saves pretraining time.

RQ3: what is the effect of parameters with different input layers, hidden layers, and attention heads on the performance of the pretrained model on downstream tasks?

Answer: performance of downstream tasks depends on the number of parameters used in training along with the layers, hidden layers, and attention heads. As shown in Table 7, seven models in [15, 33–36, 38, 40] outperform RoBERTa. Among these seven models, four models [15, 33, 35, 40] have less parameters than RoBERTa such as a model in [15] utilized a very low number of parameters, i.e., 114M parameter w.r.t 360M of RoBERTa. With less parameters, different combinations are also analyzed such as two research studies [33, 35] have less parameters but pretrained with deep hidden layers. On the other side, two out of seven models [34, 36] have very large parameters than RoBERTa. Lastly, one model [38] used the same number of parameters as RoBERTa. In the light of above parameters, it is analyzed that if a model has very large parameters or has large hidden layers and attention heads, it will produce better results.

RQ4: what are the effective techniques for cross-layer parameter sharing in pretraining of models?

Answer: cross-layer sharing helps the models to produce better results. As shown in Table 8, seven models [15, 33–36, 38, 40] produce a better output than RoBERTa and all of these models except for [38] shared both FFN and attention across layers during pretraining. On the other side, almost all the models except for [38] which are not shared during pretraining produce less efficient results. Among these seven models, different combinations with cross-layer parameters are used such as [33, 35] represents deeper model (i.e., including large hidden layers) while [34, 36, 38] are humongous models which means large input layer, hidden layer, and big attention head are involved in these models, whereas [40] is also close to RoBERTa in size, parameters, and other setting, but it shares parameters in pretraining and produces better results. It could be seen that all-shared parameters have a positive effect on the performance of models.

RQ5: what are the leading datasets used in pretraining of models during 2018–2020?

Answer: twenty-eight (28) out of thirty-one (31) models used GLUE for downstream tasks. GLUE consists of ten diversified tasks. These tasks could be seen on https://GLUEbenchmark.com/leadaerboard. As shown in Table 9, the higher number of models, i.e., seventeen (17), uses SQuAD for the downstream task. The SQuAD has two subtasks SQuAD v1.1 and SQuAD v2.0. Six models, [29, 36, 41, 50, 52, 54] research studies, used the BLEU dataset while [27, 28, 33, 34, 38, 43] research studies use the RACE dataset. Only [27] used all of these datasets while [28, 33] used GLUE, SQuAD, and RACE datasets altogether. GLUE and SQuAD are significant datasets for testing new models, and that is the reason that these datasets become the benchmark for future models.

5. Analysis

In this section, we discuss the analysis of the research studies based on the pretraining dataset and different settings such as data size, batch size, and step size.

5.1. Different Ways of Pretraining of BERT Model. Delvin et al. [27] introduced the first deeply bidirectionally pretrained model to train the unlabelled text. BERT also introduced the NSP which is used to predict the next sentence for the Question Answering system. Besides, BERT requires an additional layer to perform any type of downstream task. Wang et al. [30] incorporate the language structures (word and sentence) during pretraining which enable the model to reconstruct the right order of words and sentences. This model extends the NSP by predicting previous values. Joshi et al. [10] used Masked Language Modeling (MLM) at span level. It uses a novel span boundary objective which summarizes as required span as possible and uses a single contiguous segment instead of two segments in pretraining. Zhang et al. [11] introduced the model consisting of out-of-the-shelf labeler, a sentence encoder where semantic labels are mapped into embedding in parallel and semantic integration components to obtain joint representation for fine-tuning. Su et al. [39] presented squeeze and excitation to extract global information between layers and Gaussian blurring to capture the neighbor context in the downstream task. It also uses the Heuristic Analysis for NLI Systems (HANS) dataset which shows SesameBERT adopted shallow heuristic instead of a generalization.

Josefowicz et al. [55] fine-tuned on extreme multilabel text classification. This classification used semantic label

Table 9: Datasets used in selected research studies.

Dataset name	Research studies
GLUE	[10, 11, 14, 15, 20, 27, 29–33, 36–48, 50–52]
BLEU	[28, 29, 36, 41, 50, 52, 54]
SQuAD	[10, 11, 20, 27–31, 33–36, 40, 41, 49, 50, 52]
RACE	[27, 28, 33, 34, 38, 43]

clusters for better model dependencies and both label and input text to build label representation. It consists of semantic label indexing and ensemble ranking component. Jiao et al. [31] propose a transformer distillation method which transfers the linguistic knowledge from teacher BERT to TinyBERT. In distillation methods, we have two methods: first one consists of the big model called the teacher method, and other consists of the small model called the student model. When the teacher model is pretrained, it gains knowledge and transfers its knowledge to the student model. A two-stage learning framework uses transformer distillation on pre-training and fine-tuning and lets the TinyBERT capture general and specific knowledge of teacher BERT with 28% fewer parameters. Xu et al. [42] used progressive model replacing to compress the parameters; it first divides the BERT model and then builds their compact substitute. The probability of replacing was increased through training. Wang et al. [30] trained by the BERT model using stacking algorithm that observes the self-attention at different layers and positions for transferring the knowledge from shallow to deep model. It also finds local attention distribution and start-of sentence distribution. Goyal et al. [43] improved inference time with very little performance loss. PowerBERT removes the word-vector from the encoder pipeline which reduces the computation and directly improves inference time.

Furthermore, Chen et al. [48] task-oriented the compressed BERT model, called AdaBERT which uses differentiable neural architecture searches to automatically compress the BERT model into task-specific small models. AdaBERT incorporates task-oriented knowledge distillation (KD) loss for search hint and efficiency loss as search constraints. Beltagy et al. [26] built a new scientific vocabulary. The trained BERT model on large and in-domain-scientific data shows that the in-domain pretrained model performs better on downstream tasks due to in-domain vocabulary. Lee et al. [8] proposed a first domain-specific pretrained model that trained BERT on a medical dataset. The medical dataset includes PubMed abstracts and PMC full text instead of general datasets such as Wikipedia. Results show that the domain-specific pretrain model outperforms the BERT on Question Answering (QA) (12.24), Relation Extraction (RE) (2.8), and Named Entity Recognition (NER) (0.82). Chadha et al. [49] used set of modified transformer encoder units to add more focused query-to-context (Q2C) and context-to-query (C2Q) attention to BERT architecture. It also adds localized information to self-attention and skips connections in BERT. Kao et al. [35] boosted the BERT by duplicating some layers which makes BERT deeper without extratraining to increase the performance of the BERT model in downstream tasks.

5.2. Different Settings for Models. Liu et al. [17]pre-trained the BERT model on a 12 times bigger and diverse dataset with two changes in hyper-parameters during pre-training. First is a bigger batch size with small step size and second is dropping the NSP.. Lan et al. [33] reduced the size of the parameters of the model during training. Also, it uses two-parameter reduction techniques. The first one is factorized embedding parameterization to separate the size of hidden layers from the size of vocabulary embedding. The second one is cross-layer parameter sharing. It also replaces NSP with SOP. Yang et al. [28] used an autoregressive and autoencoding pretrained model that uses all possible permutations of factorization order. It uses relative positioning and segment recurrence mechanism borrowed from Transformer-XL (extension of transformer architecture with positional encoding). XLNet does not use MLM to remove pretraining and fine-tuning discrepancy and also leave the NSP which decreases the XLNet performance. Zhu et al. [38] presents the adversarial training algorithm which makes the transformer-based models better by adding adversarial perturbation to word embedding and minimize the maximum risk. Bao et al. [52] use Pseudo-Masked Language (PMLN) training procedure combining autoencoding (AE) and partially autoregressive (PAR).it follows BERT for encoding modeling. AE provides global PAR to learn interrelation between masked span. PMLN learns long-distance context better than the BERT.

Moreover, Lewis et al. [50] proposed denoising autoencoder to pretrain sequence-to-sequence models by corrupting the text with arbitrary noisy function and then reconstruct the original text. It proposes a novel in-filling scheme and is best to perform for generalization. It differs from BERT as additional cross-attention by decoder layers perform on the last hidden layer of the encoder. Chen et al. [56] proposed a unified framework that converts language problems into a text-to-text problem for training on a new dataset C4 with 11B parameters. Houlsby et al. [20] presented a novel adaptor tuning that uses only 3.6% of task-specific parameters of BERT instead of 100% use in fine-tuning. It provides a compact and extensible model adding only a small number of additional parameters per task because it remembers the previous values. Chang et al. [54] proposes a novel task conditional masked language to fine-tuned BERT on the text-generation dataset. It improves text generation by providing word probability distribution for every token in the sentence. Xu et al. [57] improved the BERT by using self-ensemble and self-distillation in fine-tuning without using external data. The self-ensemble model is an intermediate model at a different time which has average parameters of base-models. The distillation loss is used as regularization which improves the performance. Jiang et al. [37] overcome the limited downstream resources which make the model overfit, and it forgets the knowledge of the pretraining model. Smoothness-inducing regularization and Bregman proximal point optimization were applied on fine-tuning of models in which SMARTRoBERTa produces the SOTA results for many tasks. Zhang et al. [14] pretrained models on knowledge graphs and

large-scale textual. This model uses lexical, syntactic, and knowledge information with MLM and NSP loss.

Furthermore, Sun et al. [15] presented a pretraining framework that builds the task and then incrementally learn multitask learning. It extracts the lexical, syntactic, and semantic information as named entity, closeness relation from things corpus. Wei et al. [51] presented a new pretrained model trained on large Chinese corpus with functional relative positional encoding whole word masking strategy, LAMB optimizer mixed-precision training, and length of the training sequence. In Clark et al.'s study [40], the pretrained representation masked the input with plausible alternative sampled from a small genitor network. This model predicts whether each of the tokens in corrupted input was replaced by a generator sample or not. The computational speed of Electra was four times faster than RoBERTa and XLNet. Wang et al. [41] presented a compressed and small pretrained language model. This model contains two models: student and teacher. The student model is trained by deeply mimicking the self-attention model in the larger model (the teacher model). It performs distillation of the self-attention model from the last layer of the large model.

Shoeybi et al. [34] used billions of parameters by using efficient intralayer model parallelism attention in the placement of layer normalization in the BERT style model which increases the performance of model. Liu et al. [32] introduced a model that learned from multiple Natural Language Understanding (NLU) tasks. It uses cross-layer sharing and general representation which helps it to adapt to new tasks and domains. Clark et al. [44] introduced a model that uses knowledge distillation in which single-task trains the multitask. It proposes teacher annealing which takes the distillation to supervised learning which helps the multitask model to learn and surpass its teacher model. Dong et al. [29] trained the model on unidirectional, multidirectional, and sequence-to-sequence tasks. It fine-tuned for language understanding and generation tasks. It uses specific self-attention masks and a shared transformer network. Liu et al. [46] presented learning text representation across NLU tasks. An ensemble of teacher models is trained, and the student model is trained on the teacher model via learning distill knowledge.

Table 10 presents the overall data of all the models included in this study, training dataset, dataset size, tokens, model type, sentence learning, and cross-layer parameter sharing of every model. Table 1 shows the parameters and model setting, and Table 11 shows the result of every model. Every model in this study not just pretrained with different hyperparameters, different learning techniques, or sharing techniques. These models also pretrained differently, for example, some use MLM at the token level, some use at span level, some models use annealing, distillation method, and duplication of hidden layers, and others separate the hidden layers from model size. Effect of different ways of pretraining is minimum against the effect of parameters, for example, very few models perform better then RoBERTa. These better models solely depend on techniques which show the effect of hyperparameters, learning, and sharing on the performance of language models.

6. Discussion

In the above section, we have provided the answers to the question in Section 4. There are a few recommendations to existing/new models as follows:

(i) Small + FFN: small models (small input layers, hidden layers, and attention heads) with fewer parameters but with the sharing of both FFN and attention during pretraining improve the performance of the language model.

(ii) Deeper models: small models with very deep hidden layers and bigger attention heads using all-shared cross-layer sharing produce the best result among language models.

(iii) Bigger models: bigger models produce better results except when they are trained with fewer hidden layers and fewer attention heads. To increase hidden layers, we need to use fewer input layers to computationally compatible.

(iv) Dynamic masking: the use of dynamic masking allows changing the masking with every epoch to overcome the limitation of static masking. Static masking only masks the tokens with the same sequence affecting the performance of the model. If dynamic masking is used, then with every epoch, a new token will be masked.

(v) Larger batch: pretraining of the language model with larger batch size learns faster and improves results. It also saves us from large step size. If one increases the size of batch size, the step size decreases.

(vi) Sentence Order Prediction (SOP): use of SOP instead of NSP on other models such as XLNet and RoBERTa is beneficial. The reason is SOP can cover NSP tasks, but NSP cannot cover SOP task which means SOP has higher accuracy.

(vii) Domain-specific dataset: training on domain-specific datasets, such as medical and scientific, produces better results as the model will learn more about the specific domain better than the general domain.

(viii) Adversarial training: use of adversarial training on smaller models with cross-layer sharing is highly recommended. When it is applied on the fine-tuning step, it limits the maximum risks and could be applied to any model built upon the transformer architecture.

(ix) MLM: models can be pretrained with different MLM strategies on the span, lexical, syntactic, semantic, and knowledge information for pretraining the models.

(x) Distillation: the use of distillation methods having a student model and teacher model is highly recommended. The student models learn from the teacher model which saves it to pretrain on the

Table 10: Overall data.

Paper	Name	Training data size	Tokens	Training dataset name	Model type	Sentence learning	Cross-layer parameter sharing
[17]	RoBERTa (large)	160 GB	−2.2T	BooksCorpus + Wikipedia + CC-News + OpenWebText + Stories	Autoencoding	None	False
[10]	SpanBERT	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	False
[11]	SemBERT	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	False
[14]	ERNIE,	9 GB+	4.5B + 140M	English Wikipedia + Wikidata	Autoencoding	NSP	False
[15]	ERNIE2.0	13 GB+	8B	Encyclopedia + BooksCorpus + Dialog + Discourse Relation Data	Autoencoding	None	True
[27]	BERT (base)	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	NSP	False
[28]	XLNet	126 GB	32.89B	BooksCorpus + Wikipedia + Giga5+ ClueWeb 2012B + Common Crawl	Autoencoding + autoregressive	None	True
[29]	UniLM	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	NSP	True
[20]	—	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	True
[30]	StructBERT	16 GB	2.5B+	English Wikipedia + BooksCorpus	Autoencoding	NSP	False
[31]	TinyBERT	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	NSP	False
[32]	MT-DNN	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	NSP	True
[33]	AlBERT (xxlarge)	16 GB	—	BooksCorpus + Wikipedia	Autoencoding	Sop	True
[34]	Megatron-LM	174 GB	—	Wikipedia + CC-Stories + Real News + OpenWebText	Autoencoding	SOP	True
[35]	AlBERT (xxlarge-ensemble)	16 GB	—	BooksCorpus + Wikipedia	Autoencoding	None	True
[36]	T5	29 TB	—	Colossal Clean Crawled Corpus	Autoencoding	None	True
[37]	SMARTRoBERTa	160 GB	—	BooksCorpus + Wikipedia + CC-News + OpenWebText + Stories	Autoencoding	None	False
[38]	FreeLBRoBERTa	160 GB	−2.2T	BooksCorpus + Wikipedia + CC-News + OpenWebText + Stories	Autoencoding	None	False
[39]	SesameBERT	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	False
[40]	Electra1.75M	126 GB	33B	BooksCorpus + Wikipedia + Giga5+ ClueWeb 2012B + Common Crawl	Autoencoding	None	True
[41]	MiniLMa	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	False
[42]	SBERT-WK	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	False
[43]	PowerBERT	11 GB	3.4B	BooksCorpus + Wikipedia	Autoencoding	None	False
[44]	Bam	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	True
[45]	StackBERT	11 GB	3.4B	BooksCorpus + Wikipedia	Autoencoding	None	False
[46]	MT-DNNKD	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	True
[47]	HUBERT	16 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	True
[48]	AdaBERT	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	True
[49]	BERTQA	13 GB	3.8B	BooksCorpus + Wikipedia	Autoencoding	None	False
[50]	BART (large)	160 GB	2.2T	160 GB + Wikipedia	Autoregressive	None	False
[51]	Nezha	—	10.5B	Chinese Wikipedia + Baidu Baike + Chinese News	Autoencoding + autoregressive	NSP	True
[52]	UniLMv2	160 GB	—	BooksCorpus + Wikipedia + CC-News + OpenWebText + Stories	Autoencoding and partially autoregressive	None	False

Table 11: Results.

Paper	Batch size	Max sequence	Learning rate	Step size	Parameters (M)	Layers	Hidden	Attention head
[17]	2K	512	1e − 6	125K	360	24	1024	16
[10]	256	128	1e − 4	2.4M	340	24	1024	16
[11]	32	128	2e − 5	1M	340	24	1024	16
[14]	512	256	5e − 5	1M	114	6	768	12
[15]	400K	256	5e − 5	4K	114	24	1024	16
[27]	256	128	1e − 4	1M	110	12	768	12
[28]	2048	512	1e − 5	500K	340	24	1024	16
[29]	330	512	3e − 5	777K	340	24	1024	16
[20]	32	512	1e − 4	1M	330	24	1024	16
[30]	32	512	1e − 4	1M	340	24	1024	16
[31]	256	128	1	1M	14.5	4	312	12
[32]	256	128	1e − 4	1M	340	24	1024	16
[33]	4096	512	0.00176	125K	233	12	4096	128
[34]	1024	128	1.0e − 4	1M	3.9	48	2560	40
[35]	4096	512	0.00176	125K	233	12	4096	64
[36]	2048	128	0.01	2.1M	11	12	768	12
[37]	2K	512	10^{-3}	125K	356	24	1024	16
[38]	8K	512	1e − 6	500K	360	24	1024	16
[39]	32	128	2e− 5	1M	340	12	768	12
[40]	2048	512	2e − 4	1.75M	335	24	1024	16
[41]	1024	512	5e − 4	400k	33	12	768	12
[42]	32	256	2^{-5} to10^{-5}	1M	66	6	768	12
[43]	256	128	1e − 4	1M	108	12	768	12
[44]	128	128	1e − 4	1M	340	24	1024	16
[45]	256	128	1e − 4	1M	110	12	768	12
[46]	256	128	1e − 4	1M	340	24	1024	16
[47]	256	128	5^{-5} to 10^{-5}	1M	110	12	768	12
[48]	128	512	3e − 4	50K	9.5	24	1024	16
[49]	6	512	1.5e − 5	1M	340	24	1024	16
[50]	8000	512	1e − 6	500K	400	12	1024	12
[51]	5120	128	1.8e − 4	25K	340	24	1024	16
[52]	7680	128	6e − 4	0.5M	110	12	768	12

whole dataset. The student model just needs to learn from the teacher model.

(xi) Duplicating layers: A method of duplicating layers to make models deeper could save a lot of computational power. Duplicating layer is a method in which we keep smaller size of layers, and during the execution of pretraining, we duplicate the layers which directly make the model deep.

It is very hard to say which setting could be used to improve the performance of models due to trade-offs as larger models will use more resources while smaller models will cover fewer data. It is also recommended to use combinations such as bigger batch size with smaller step size, use of fewer input layers with largely hidden layers, and big attention heads. Subsequently, training of the hybrid combination on the domain-specific dataset and use of MLM on the span and lexical level is suggested. By doing so, performance of the bidirectional language models can be improved.

7. Conclusion and Future Work

This paper presents the SLR on a comprehensive study of thirty-one (31) pretrained language models to find the answers to five developed research questions. All models used in this paper are inspired by BERT and have a transformer or Transformer-XL architecture. The significant findings of this SLR are presented in Tables 1, 10, and 11. Table 10 presents the overall data used in these models, Table1shows the hyperparameter setting of these models, and Table 11 highlights the results of these models. These research papers show the effect of sentence embedding learning, size of the dataset, step, batch, parameters, layers, attention heads, and the effect of cross-layer sharing and also provide the most used benchmarks for future models. To conclude, whole focus of our study is about the pretraining of language models covering fine-tuning settings and the downstream task. The tables are created in two ways so that we could depict more accurate data by providing authentic information.

There are different ways to pretrain a model such that MLM on tokens or spans, etc. Besides, many models are pretrained on domain-specific datasets (e.g., business, medical, and physics) to improve the performance of models, but still the impact of these models is minimum when compared by dataset size, objectives, representation, sharing, and parameters. Therefore, it is important to consider these factors for language models before implementation because these factors can affect the performance

of any model. In this study, we only consider the models which were built on the transformer or Transformer-XL architecture and inspired by BERT, but in future, we intend to include models built on other architectures such as RNN.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] T. Mikolov, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 2, pp. 3111–3119, 2013.

[2] S. R. Bowman, "A large annotated corpus for learning natural language inference," 2015, https://arxiv.org/abs/1508.05326.

[3] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," 2017, https://arxiv.org/abs/1704.05426.

[4] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, Jeju Island, South Korea, October 2005.

[5] P. Rajpurkar, "Squad: 100,000+ questions for machine comprehension of text," 2016, https://arxiv.org/abs/1606.05250.

[6] D. Mahajan, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, September 2018.

[7] A. Radford, "Improving language understanding by generative pre-training," 2018.

[8] J. Lee, W Yoon, S Kim et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics (Oxford, England)*, vol. 36, no. 4, pp. 1234–1240, 2020.

[9] I. Beltagy, A. Cohan, and K. Lo, "Scibert: pretrained contextualized embeddings for scientific text," 2019, https://arxiv.org/abs/1903.10676.

[10] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[11] Z. Zhang, "Semantics-aware BERT for language understanding," 2019, https://arxiv.org/abs/1909.02209.

[12] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019.

[13] X. Yin, Y. Huang, B. Zhou, A. Li, L. Lan, and Y. Jia, "Deep entity linking via eliminating semantic ambiguity with BERT," *IEEE Access*, vol. 7, pp. 169434–169445, 2019.

[14] Z. Zhang, "ERNIE: enhanced language representation with informative entities," 2019, https://arxiv.org/abs/1905.07129.

[15] Y. Sun, "Ernie 2.0: a continual pre-training framework for language understanding," 2019, https://arxiv.org/abs/1907.12412.

[16] R. Zellers, "Defending against neural fake news," *Advances in Neural Information Processing Systems*, vol. 3, pp. 6000–6010, 2019.

[17] Y. Liu, "Roberta: a robustly optimized bert pretraining approach," 2019, https://arxiv.org/abs/1907.11692.

[18] A. Wang, "Superglue: a stickier benchmark for general-purpose language understanding systems," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1–9, 2019.

[19] A. Conneau and D. Kiela, "Senteval: an evaluation toolkit for universal sentence representations," 2018, https://arxiv.org/abs/1803.05449.

[20] N. Houlsby, "Parameter-efficient transfer learning for NLP," 2019, https://arxiv.org/abs/1902.00751.

[21] M. Peters, S. Ruder, and N. A. Smith, "To tune or not to tune? adapting pretrained representations to diverse tasks," 2019, https://arxiv.org/abs/1903.05987.

[22] Y. Iwasaki, "Japanese abstractive text summarization using BERT," in *Proceedings of the 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, IEEE, Kaohsiung, Taiwan, November 2019.

[23] M. G. Sousa, "BERT for stock market sentiment analysis," in *Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, Portland, OR, USA, November 2019.

[24] X. Yu, "BioBERT based named entity recognition in electronic medical record," in *Proceedings of the 2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, IEEE, Qingdao, China, August 2019.

[25] S. Jiang, "A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition," in *Proceedings of the 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, IEEE, Xiangtan, China, October 2019.

[26] I. Beltagy, K. Lo, and A. C.. SciBERT, "A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, November 2019.

[27] J. Devlin, "Bert: pre-training of deep bidirectional transformers for language understanding," 2018, https://arxiv.org/abs/1810.04805.

[28] Z. Yang, "Xlnet: generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 19, pp. 400–406, 2019.

[29] L. Dong, "Unified language model pre-training for natural language understanding and generation," 2019, https://arxiv.org/abs/1905.03197.

[30] W. Wang, "StructBERT: incorporating language structures into pre-training for deep language understanding," 2019, https://arxiv.org/abs/1908.04577.

[31] X. Jiao, "Tinybert: distilling bert for natural language understanding," 2019, https://arxiv.org/abs/1909.10351.

[32] X. Liu, "Multi-task deep neural networks for natural language understanding," 2019, https://arxiv.org/abs/1901.11504.

[33] Z. Lan, "Albert: a lite bert for self-supervised learning of language representations," 2019, https://arxiv.org/abs/1909.11942.

[34] M. Shoeybi, "Megatron-lm: training multi-billion parameter language models using gpu model parallelism," 2019, https://arxiv.org/abs/1909.08053.

[35] W.-T. Kao, "Further boosting BERT-based models by duplicating existing layers: some intriguing phenomena inside BERT," 2020, https://arxiv.org/pdf/2001.09309.

[36] C. Raffel, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019, https://arxiv.org/abs/1910.10683.
[37] H. Jiang, "SMART: robust and efficient fine-tuning for pre-trained Natural Language models through principled regularized optimization," 2019, https://arxiv.org/abs/1911.03437.
[38] C. Zhu, "Freelb: enhanced adversarial training for language understanding," 2019, https://arxiv.org/abs/1909.11764.
[39] T.-C. Su and H.-C. Cheng, "SesameBERT: attention for anywhere," 2019, https://arxiv.org/abs/1910.03176.
[40] K. Clark, "ELECTRA: pre-training text encoders as discriminators rather than generators," in *Proceedings of the International Conference on Learning Representations*, New Orleans, LA, USA, May 2019.
[41] W. Wang, "MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers," 2020, https://arxiv.org/abs/2002.10957.
[42] C. Xu, "BERT-of-Theseus: compressing BERT by progressive module replacing," 2020, https://arxiv.org/abs/2002.02925.
[43] S. Goyal and PoWER-BERT, "Accelerating BERT inference for classification tasks," 2020, https://arxiv.org/abs/2001.08950.
[44] K. Clark, "Bam! born-again multi-task networks for natural language understanding," 2019, https://arxiv.org/abs/1907.04829.
[45] L. Gong, "Efficient training of bert by progressively stacking," in *Proceedings of the International Conference on Machine Learning*, Long Beach, CA, USA, May 2019.
[46] X. Liu, "Improving multi-task deep neural networks via knowledge distillation for natural language understanding," 2019, https://arxiv.org/abs/1904.09482.
[47] M. Moradshahi, "HUBERT untangles BERT to improve transfer across NLP tasks," 2019, https://arxiv.org/abs/1910.12647.
[48] D. Chen, "AdaBERT: task-adaptive BERT compression with differentiable neural architecture search," 2020, https://arxiv.org/abs/2001.04246.
[49] A. Chadha and R. Sood, "BERTQA_Attention on steroids," 2019, https://arxiv.org/abs/1912.10435.
[50] M. Lewis, "Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, https://arxiv.org/abs/1910.13461.
[51] J. Wei, "NEZHA: neural contextualized representation for Chinese language understanding," 2019, https://arxiv.org/abs/1909.00204.
[52] H. Bao, "UniLMv2: Pseudo-masked language models for unified language model pre-training," 2020, https://arxiv.org/abs/2002.12804.
[53] B. Kitchenham, *Procedures for Performing Systematic Reviews*, Keele University, Keele, UK, 2004.
[54] Y.-C. Chen, "Distilling the knowledge of BERT for text generation," 2019, https://arxiv.org/abs/1911.03829.
[55] W.-C. Chang and X.- BERT, "eXtreme multi-label text classification using bidirectional encoder representations from transformers," 2019.
[56] R. Jozefowicz, "Exploring the limits of language modeling," 2016, https://arxiv.org/abs/1602.02410.
[57] Y. Xu, "Improving BERT fine-tuning via self-ensemble and self-distillation," 2020, https://arxiv.org/abs/2002.10345.

Hindawi
Scientific Programming
Volume 2021, Article ID 6641562, 11 pages
https://doi.org/10.1155/2021/6641562

Research Article

A Survey of Industrial Internet of Things Platforms for Establishing Centralized Data-Acquisition Middleware: Categorization, Experiment, and Challenges

Jin-Sung Ok,[1,2] **Soon-Do Kwon**,[2] **Cheol-Eun Heo**,[2] **and Young-Kyoon Suh**[1]

[1]*School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, Republic of Korea*
[2]*Smart Yard R&D Department, Daewoo Shipbuilding & Marine Engineering Co., Ltd. (DSME), Geoje 53302, Republic of Korea*

Correspondence should be addressed to Young-Kyoon Suh; yksuh@knu.ac.kr

Received 20 November 2020; Revised 1 February 2021; Accepted 15 April 2021; Published 28 April 2021

Academic Editor: Shah Nazir

The development of industrial Internet of Things (IIoT), big data, and artificial intelligence technologies is leading to a major change in the production system. The change is being propagated into the wave of transforming the existing system with a vertical structure into the corresponding horizontal platform or middleware. Accordingly, the way of acquiring IIoT data from an individual system is being altered to the way of being increasingly centralized through an integrated middleware of a scalable server or through a large platform. That said, middleware-based IIoT data acquisition must consider multiple factors, such as *infrastructure* (e.g., operation environment and network), *protocol heterogeneity*, *interoperability* (e.g., links with legacy systems), *real-time*, *and security*. This manuscript explains these five aspects in detail and provides a taxonomy of eighteen state-of-the-art IIoT data-acquisition middleware systems based on these aspects. To validate one of these aspects (network), we present our evaluation results at a real production site where IIoT data-acquisition loss rates are compared between wireless (long-term evolution) and wired networks. As a result, the wired communication can be more suitable for centralized IIoT data-acquisition middleware than wireless networks. Finally, we discuss several challenges in establishing the best IIoT data-acquisition middleware in a centralized way.

1. Introduction

Digital transformation, also known as *DT* or *DX*, is an important keyword for modern production systems. The utilization of technologies such as industrial Internet of Things (IIoT), big data, and artificial intelligence (AI) in existing systems enables digital transformation to immediately respond to customers' demands and build a production system that improves the current production efficiency [1, 2]. Thus, numerous research institutes and enterprises are conducting research on upgrading production systems that apply new technologies to the industrial environment.

Compared to building a new system from scratch, changing the existing system brings many considerations. One of the most time-consuming and costly processes is to acquire high-quality data. Most of the legacy IT and production systems, including Manufacturing Execution System (MES) and Supervisory Control and Data Acquisition (SCADA), have a vertical structure.

To flatten the vertical structure for better data acquisition, the new system must be able to *aggregate* each production data. To this end, numerous middleware platforms adopt a horizontal structure that integrates the data acquisition [3–12]. The proposed systems have been applied in actual industrial fields.

To establish centralized data-acquisition middleware, we must determine whether the above middleware platforms meet a set of major functionalities. This manuscript proposes the following functionalities: (i) wired and/or wireless network compatibility, (ii) support for a variety of compatible industrial protocols, (iii) automated real-time data collection, (iv) data integration and external transmission, and (v) security. As necessary functions and standards have

not been well standardized and established, the existing systems are based on their own criteria, which are non-consensual. Therefore, whether we are equipped to build high-quality IIoT data acquisition middleware is difficult to discern. Such ambiguous criteria may cause *duplicate development and increased development costs*.

To address this problem, we propose and describe a set of functionalities that must be addressed when developing centralized IIoT data acquisition middleware. We then review eighteen cutting-edge IIoT middleware systems and provide a taxonomy of these systems based on clearly motivational functionalities. One of these functionalities (communication type) was assessed in experiments at our real production site. The acquisition percentages of IIoT data under wired and wireless (long-term evolution, LTE) communications were 99.940% and 98.983%, respectively. From this result, we inferred that wired communication is more robust for centralized IIoT data acquisition than wireless communication. This empirical result sheds light on the potential validity of the proposed functionalities.

The main contributions of this manuscript are summarized as follows:

(i) We propose a number of considerations for building a centralized IIoT data-acquisition middleware

(ii) We elaborate on the distinctions between IoT and IIoT data-acquisition systems

(iii) We review a rich body of existing IIoT systems and qualitatively analyze them along with well-motivated criteria

(iv) We present our evaluation results obtained from a real industrial site with respect to IIoT data-acquisition loss between wireless and wired networks

(v) We draw several challenges for constructing IIoT data-acquisition middleware in a central server

The remainder of this manuscript is organized as follows. The following section proposes a set of considerations to establish the best IIoT data-acquisition middleware, classifies these considerations into five categories, and provides the key components of each consideration. The subsequent section reviews recent IIoT data-acquisition middleware systems. Thereafter, we present our experiment results showing different data-acquisition performances among IIoT devices (in this case, welding machines). Finally, we suggest the future research directions of our work.

2. Functionalities for Centralized IIoT Data-Acquisition Middleware

To build the Smart Factory or cyber-physical system (CPS) in a short time, the production data-acquisition system that serves as a backbone should be architected and well-designed. IIoT data-acquisition middleware enables fast and easy development of the applications. Most IoT systems develop applications for a new environment without integrating with existing systems. However, building IIoT systems often require upgrading existing production systems because IIoT data are not only obtained from existing sensors, gateways, and controllers but also fused with other application data. If the upgrade is necessary, modification of the existing system need to be minimized, and the core system of the current production system should remain unchanged. The reason is that upgrading the IIoT system incurs high investment cost.

To the best of our knowledge, data acquisition at industry sites has been little investigated. In this article, we fill this gap by exploring the various factors demanded of a solid and reliable middleware system for IIoT data-acquisition. A taxonomy of these factors is illustrated in Figure 1.

In the illustrated taxonomy, the first consideration is the *infrastructure*, including the operation and network environment. The infrastructure factor is divisible into two subfactors: operation environment and network. The first subfactor is further divided into on-premises, cloud, and hybrid environments. Most industrial sites have applied on-premises systems that satisfy the security and management requirements within the technical limitations. At present, numerous sites have adopted the cloud environment which allows users to gather and manage their IIoT data for further analysis and development [13]. Within the cloud environment, building systems can be quickly built and can be flexibly managed. However, the cloud incurs a security risk and requires additional hardware or programs for sending data to the cloud. For these reasons, most industry sites still prefer the on-premises environment. Other companies have built hybrid environments that combine the advantages of on-premises and cloud.

The second subfactor is *network*. The IIoT data-acquisition network environment is largely distinguished by wired and wireless networks. Wired communication is classified into analog signal, serial communication, and LAN communication. It has several advantages, such as cost-effectiveness, stability, and low maintenance. However, it can be disadvantageous when not installed in mobile environments. Recently, wireless communications have significantly expanded owing to technological advances and reduced system-development costs [14]. Wireless networks can utilize licensed frequency bands, such as 3 G, LTE, 5 G, and NB-IoT [15, 16], but licensed frequency standards and abilities vary among countries and local environments. If a network uses licensed frequency bands, it must use the demilitarized zone (DMZ) for safety purposes. Thus, numerous industrial sites have attempted to use unlicensed frequency bands in their local networks for IIoT data acquisition.

Short-distance local networks such as Wireless Fidelity (Wi-Fi), Bluetooth Low Energy (BLE), and ZigBee are also available. Recently, many industry sites have attempted to apply low-power wide-area networks (LPWAN), including Long Range (LoRa) and Sigfox, which are specialized for IoT and support small data transfer with low-power consumption [17–22]. In contrast to wired communication, wireless communication must guarantee stable data acquisition and control.

The second factor that must be considered is *heterogeneity* (in protocol). This factor can be divided into industrial protocol, communication protocol, and database driver. In

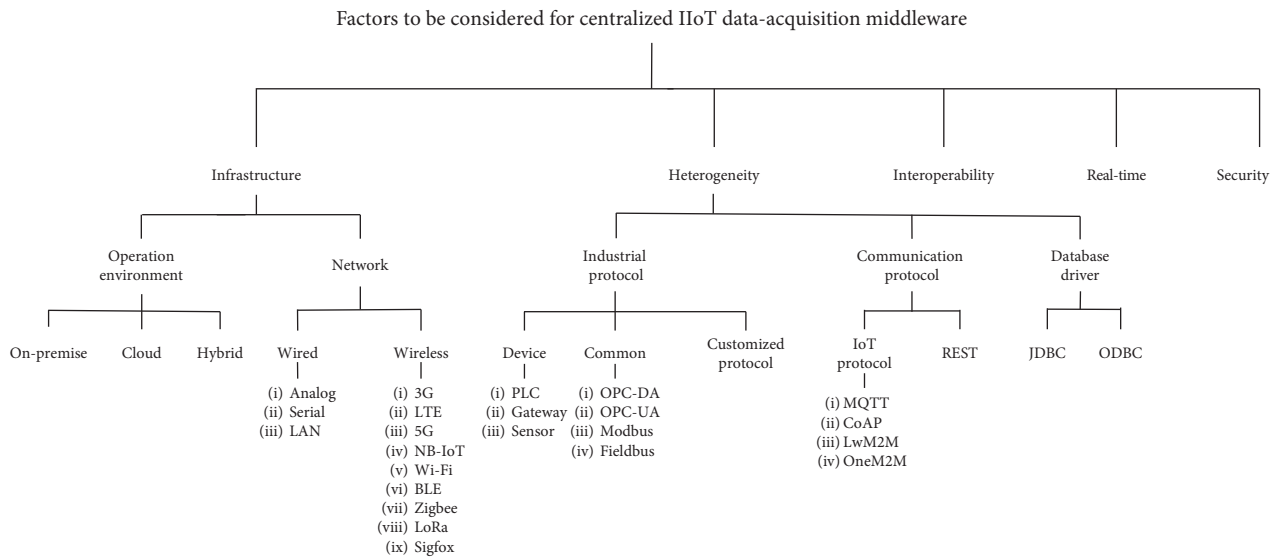


Figure 1: Proposed taxonomy of functionalities for centralized IIoT data-acquisition middleware.

general, most of the time and cost of an entire project is spent on setting and developing IIoT protocols and drivers. The first subfactor is industrial protocol. This can be further divided into device, common, and customized protocol levels.

At the device level, typically the gateway or controller uses programmable logic controllers (PLC). Some sensors and gateway have manufacturer-specific protocols. Therefore, a variety of PLC drivers, sensor protocols, and gateway protocols are required to obtain data from industrial equipment. Recently, the IIoT system is used as part of or in place of SCADA or MES (mentioned in the Introduction), so data-acquisition middleware with the device-level protocols is required.

At the common level, recently common protocols are adapted for many sites. Standard protocols are being introduced by several manufacturers and research institutes. The Open Platform Communications (OPC) Foundation developed two protocols—OPC-DA (Data Access) and OPC-UA (Unified Architecture)—for real-time monitoring and control systems. Again, it is very challenging to change the existing products and systems. Thus, protocols for existing equipment are necessitated. Moreover, because the existing applications including SCADA and MES use traditional industrial protocols such as Modbus and Fieldbus, the existing drivers must also be compatible.

At the customized protocol level, a specialized protocol for various purposes such as security and research needs to be developed.

The second subfactor is communication protocol. Two components associated with the communication protocol are IoT protocol and Representational State Transfer (REST).

The existing HTTP-based protocol is built for client-server architectures. Therefore, it may have limited ability to acquire real-time IIoT data. One such limitation is the request-response method, which cannot easily receive various IIoT data in real time. Moreover, a number of packets are needed to transmit and receive data. Thus, many institutes, companies, and researchers have developed their own IoT protocols.

In 2013, IBM developed Message Queuing Telemetry Transport (MQTT), which is a lightweight protocol using a publish/subscribe messaging model in a TCP/IP environment. MQTT provides a total of three quality of service (QoS) levels. In the adjustment of the QoS level, factors such as network quality and usage conditions should be considered. MQTT is increasingly used in embedded IIoT equipment, requiring light network environment.

Another protocol is Constrained Application Protocol (CoAP), a lightweight message-transfer protocol for use among devices on the same constrained network. OMA Lightweight M2M (LwM2M) is a device management protocol designed for sensor networks and machine-to-machine (M2M) environments. As an extensible resource and data model, LwM2M adopts an efficient secure data transfer standard called the CoAP.

The other is the oneM2M protocol, developed in July 2012 by eight organizations: (1) Association of Radio Industries and Businesses (ARIB), (2) the Alliance for Telecommunications Industry Solutions (ATIS), (3) China Communications Standards Association (CCSA), (4) European Telecommunication Standards Institute (ETSI), (5) Telecommunications Industry Association (TIA), (6) Telecommunication Technology Committee (TTC), (7) Telecommunications Technology Association (TTA), and (8) Telecommunications Standards Development Society, India (TSDSI).

The third subfactor is database driver. The database drivers, such as Java Database Connectivity (JDBC) and Open DataBase Connectivity (ODBC), for integrated system monitoring, are required to connect to the database.

The third component that must be considered is *interoperability*, which is the component for interchanging production data with legacy IT systems. An interface with

legacy IT applications is important. REST and MQTT protocols, which are widely used in IT systems, are needed as well.

Real-time is the fourth factor to be considered. This factor means the real-time equipment control and monitoring function. Equipment and a machine can be controlled manually and automatically. Remote control should be used in a wireless or wired network environment so that it can be controlled manually. The automatic and intelligent control should be able to perform real-time monitoring, analyze the current data set, and predict future situations for future systems, such as CPS.

The final factor is *security*. Security is divided into network, software, and hardware security. Network security aims to minimize the impact of unauthorized external disturbances by utilizing specific communication protocols [23–27]. Software security prevents other systems from accessing IIoT systems including sensors, gateways, and legacy systems. Software security assigns a security ID to each machine and sensor. Some recent security developments are based on blockchain technology [28, 29].

Nevertheless, there are many security challenges in the existing IIoT environment. For instance, most of the systems are trying to resolve the security hardware. To prevent physical access from the outside, the DMZ installations and local networks are utilized. Many companies have various policies on security. Depending on the environment of the production system, appropriate methods should be chosen to ensure security.

Note that to improve the production efficiency through AI and analysis using IIoT data, many industrial sites and research institutes have been actively conducting research on acquiring data quickly at a low cost.

The following section provides detailed descriptions of the discussed factors that need to be considered during data acquisition.

3. Key Components of IIoT Data-Acquisition Middleware

Recently, a production system is rapidly being changed to meet customers' demands. To make the system more flexible and intelligent, the system needs to collect and integrate information from a variety of IIoT devices. Figure 2 illustrates such a system centrally positioning IIoT data-acquisition middleware. The industrial data gathered through this centralized middleware can be used for data-driven decision making. Furthermore, other kinds of systems, such as intelligent and flexible systems as well as simulation systems, can utilize the collected data for further analyses and services.

To generate valuable information in an IIoT environment, real-time collection of consistent IIoT data is essential. Accordingly, middleware technology for robust data acquisition is solicited. Considering the fact that IIoT data obtained using such acquisition middleware usually come from many applications, building such middleware needs to consider the following key components: network bridge, licensed frequency band, LPWAN, industrial protocols, production IoT, and cloud.

3.1. Network Bridge and LPWAN. As mentioned earlier, networks can be classified into two broad categories: wired and wireless (See Figure 1). Many industrial sites adopt wired communication owing to its stability and speed. In a wired communication, data are often received from previously developed serial interfaces, such as RS232 and RS485. In this case, only a short-distance communication is possible. Thus, a network bridge is required to enable long-distance communication. For example, many production sites are heavily utilizing network bridges that can change serial communications to transmission control protocol/Internet protocol (TCP/IP).

Recently, with the increased use of IIoT systems, increasing data are received through wireless communication owing to the cost and deployment duration. In a wireless communication, BLE, ZigBee, Wi-Fi, etc., can be utilized for short-distance communication (See Figure 1). In this case, the data is sent to the central server by improving the distance using a dedicated network bridge. Furthermore, with the development of telecommunication infrastructures, both licensed frequency band (e.g., 3 G, LTE, NB-IoT, and 5 G) and unlicensed frequency band (e.g., LPWAN) have become widely used by many industrial sites. In the case of the licensed frequency band, certain fees are paid for use, as the frequency of the license plate is managed by a professional company or institution. Owing to its superior speed and capability to provide stable communication and large bandwidth, such a licensed frequency band is being used by many industries although it comes at high costs.

Conversely, regarding the wireless communication, batteries are considered as a critical factor, particularly in LPWAN enabling long-distance communication. When IIoT systems need the transfer of small data with low-power consumption, LPWAN has three types: Sigfox, LoRaWAN, and NB-IoT (see Figure 1). Its communication distance is in the range of 1–20 km. As described previously, the data-acquisition middleware requires a structure to make it possible to acquire data through both wired and wireless communications.

3.2. Industrial Protocols. Industrial devices are essential to achieve high reliability, durability, scalability, and ease of maintenance. PC-based controllers are used in complex operations. In fact, PLC—an industry-specific system that operates independently of the OS—is more widely used, thanks to its high compatibility with industrial protocols such as Fieldbus and Modbus (See Figure 1). Furthermore, PLC has the ability to easily acquire analog signals such as voltage or current and incurs lower cost compared to industrial PCs. Currently, the connection with IT systems has become a hot topic in PLC markets. Along with this wave, most PLCs provide common protocols to obtain and control variables over the TCP/IP environment.

However, it is costly to upgrade existing PLC programs for the purpose of sending data to other systems, in terms of expense and time. Therefore, it is of paramount importance to support various PLC protocols so that data can be

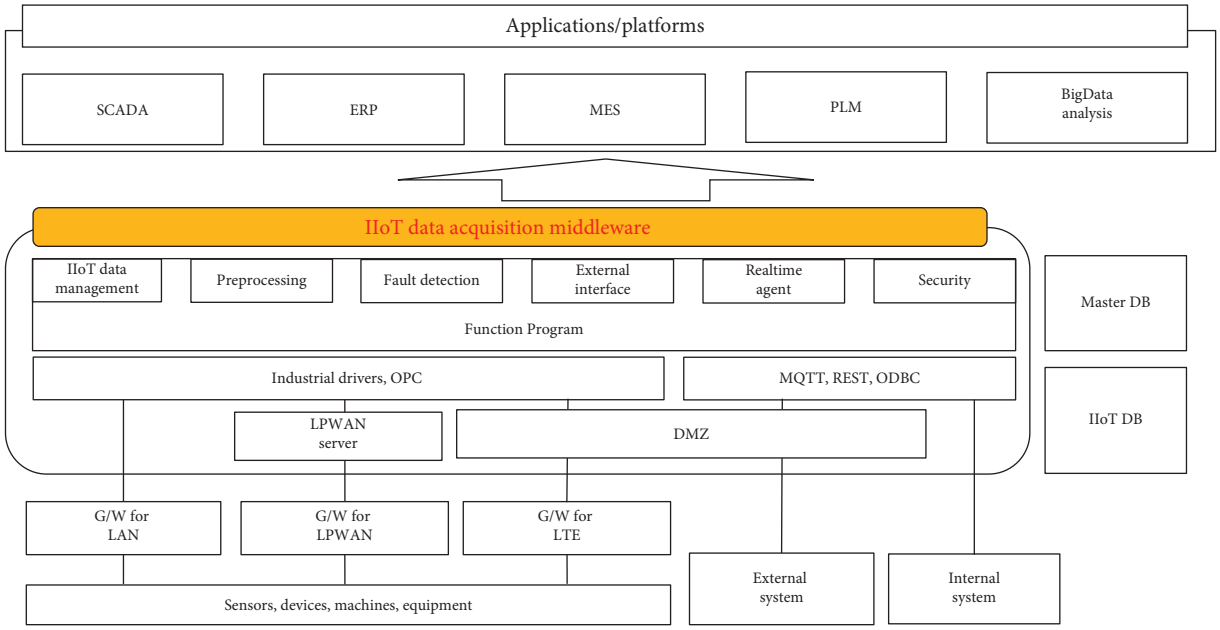


Figure 2: An overall architecture of a centralized IIoT server with data-acquisition middleware.

acquired without altering the existing PLC programs. Consequently, a number of commercial programs have been released for obtaining PLC data directly.

With the early development of PLC, a standard interface, OPC has been established. OPC enables real-time monitoring and links to automation systems, such as Human Machine Interface (HMI) and SCADA. OPC has improved the security and connection speed of the PLC protocol. In 2008, OPC-UA—vendor-dependent and highly secure protocol—was developed by the OPC Foundation [30]. It is used much in interworking with IT systems. The previous versions of OPC had a client-server architecture, which made it difficult to process multiple messages simultaneously. Conversely, OPC-UA provides publish/subscribe functions to enable 1 : N and N : N communications in real time. Moreover, it has a reliable version for cloud environments. In recent years, many researchers have been conducting research on Time-Sensitive Networking (TSN) linked with OPC to achieve 18 times faster real-time remote control and monitoring.

3.3. Production IT. Many companies have built ERP and MES for managing quality products. Before the emergence of Industry 4.0, the old systems used to operate in a vertical structure. In ISA-95 standard, the systems operate by sending and receiving data only at each of the front and rear levels. However, the development of IIoT has eliminated the boundaries of data. Data-acquisition middleware is needed to directly obtain data from MES, ERP, and the control process. The middleware requires protocols or drivers to obtain information from legacy IT systems. For instance, considering the fact that ODBC and JDBC are usually required to connect to the database, the middleware can support the drivers. In the case of a three-tier system, another interface such as REST can be used, particularly in the environment where there is little direct access to the data due to security reasons.

3.4. Cloud vs. On-Premises. Recently, the cloud has been widely adopted for its efficiency and cost-effectiveness. Many IT companies have customers who wish to use infrastructure and resources in the forms of SaaS, PaaS, and IaaS. For IIoT data acquisition, the cloud system acquires data in a different manner from an on-premises environment. The IIoT equipment, including sensors, actuators, and gateways, provide data while being located at the industrial site. Numerous existing equipment are mostly connected to networks such as LAN. This industrial equipment often has its own industrial protocols. In this case, the transfer of IIoT data to the cloud environment is required. Edge consists of a device or a program that converts the sensor's analog signal or serial signals to LAN communication. Edge also uses network bridges called protocol converters or gateways. The configuration of network bridges or the edge can be applied to industrial sites for industrial controllers, such as PLC, industrial computers, and dedicated converters, that change specific signals. Thus, an edge program that can connect to the cloud system needs to be installed on IIoT equipment. For example, offering an API is possible with standard protocols, such as MQTT and OPC-UA, or customized protocols of their own companies. Usually, due to installation of the edge program, OS-based products are needed. In this case, it is necessary to establish an environment where packages or APIs can be used, such as Linux OS or Windows OS.

Unlike the cloud systems using edge, the on-premises system makes it easy to obtain IIoT data in a centralized network management environment. Usually, the on-premises system uses network bridges for extended communication distance. The central server manages a variety of

information, including the IIoT device ID, protocols, acquisition rate, and resources. In the on-premises system, the network bridge has a wider range of configurable choices than that of the cloud system. Some cloud systems need to change equipment due to the requirement of some protocols such as MQTT, REST, and OPC-UA. However, the on-premises is more flexible than the cloud and can acquire data directly from IIoT equipment that are easy to use various gateways.

3.5. Qualitative Analysis of Existing IIoT Data-Acquisition Middleware Systems. In this section, we qualitatively analyze a variety of IIoT data acquisition middleware systems based on well-motivated criteria.

Table 1 lists the IIoT data-acquisition middleware systems developed by each vendor [31–48]. We comprehensively analyzed these systems in terms of the five major aspects in the second and third levels of Figure 1: (1) operating environment, (2) protocol, (3) driver, (4) real-time, and (5) security.

Edge software provided by IT vendors includes cloud-based middleware, such as Azure IoT Hub, AWS Industrial IoT, Oracle Internet of Things Cloud service, and Predix. These middleware systems provide software packages or APIs for the connection from IIoT equipment to their clouds using edge devices. The OPC-UA protocol is applied considering real-time control and monitoring, as well as the interface industrial system. In addition, due to the various conditions of industrial sites, IIoT data are acquired in cooperation with specialized partners in the field to suit the site situation. Kepware, PI Collect, AVEVA Edge, and MindSphere Connect that show strength in the current OT field can easily make connection of the current IIoT equipment to their systems. The companies are also increasing the ease of connectivity by providing various industrial protocols, such as the PLC interface, Modbus, and OPC-DA/UA. Moreover, some companies and research institutions use their own technology and thus create systems optimized for specialized environments [46–51]. In this case, although a middleware system does not have many functionalities, it offers great features that are specialized in the environment of operation.

Every middleware provides real-time "monitoring" functions, but some middleware services (such as Oracle Internet of Cloud Service, ThingPlug, and N-MAS) do not allow the control of IIoT devices in real time. Finally, all middleware systems well support security for communication from IIoT devices to their respective middleware.

4. Experiment: A Reliability Test of IIoT Data Acquisition at a Real Industrial Site

For convenient operation at industrial sites, IIoT data acquisition needs to be centralized. To minimize investment, we should have to determine the feasibility of acquiring the IIoT data from the legacy network infrastructure in a centralized way. To this end, we designed an IIoT data-acquisition experiment leveraging the wired and wireless networks used in office work. During this experiment, we measured data-acquisition rates for 24 h during weekdays and analyzed the network loads. Briefly, the results demonstrate that IIoT data can be "indeed" acquired by the networks used in general office work.

4.1. Environment Settings. By our intended design, we conducted two actual experiments in terms of central IIoT data acquisition in an on-premises environment via two methods, as shown in Figures 3 and 4. The difference between the two experiments was the communication environment through which the data was acquired.

For the wireless networks, we utilized LTE communication using a licensed band network (KT Corporation) in South Korea. In particular, we used a router with Private-LTE (P-LTE) for security purposes. The external LTE servers checked the router's IP and port number and switched to the designated IP and port number assigned by the customer. Subsequently, the data were sent to the internal DMZ server, which checked the IP and port number for security reasons. Finally, the data were safely sent to the internal server. The total processing time was one second.

The first method was to acquire IIoT data *via wired communication* (Figure 3(a)). The second method was to acquire data centrally *via LTE communication* (Figure 3(b)). The wired communication is the most adopted communication in the field, while LTE is now prevalent.

Configurations are exhibited in detail in Table 2. Test device is the welding machine used in a shipyard where ships and offshore plants are built. We used a total of 14 tags, including ID, voltage, current, temperature, and product information. The used protocol is a user-defined protocol. When data is requested, the welding machine sends the requested data (Figure 4). The requested data requires a total of 15 tags per a second. Because the data interface of the welding machine is RS232, the maximum transmission distance is 15 m. Thus, the machine requires a network bridge to transmit data at a long distance.

The data path of the welding machine is divided into two routes. In the first route, the IIoT data acquisition middleware requests tag data through the network bridge *via* TCP/IP communication, and the network bridge then sends tag data to the welding machine *via* RS232 interface. In the second route, the welding machine responds according to the command and then forwards all tag data back to the middleware.

The rate at which all data were acquired was once per second. Therefore, 86,400 s tag sets were acquired per a day. The experimental period was 10 days, excluding weekends, when the equipment was not in operation all day.

The applied network bridge model was NPORT-5610 in MOXA, which has eight ports that convert RS232 to TCP/IP communication. In the NPORT model, we used the TCP/IP server mode to communicate with the middleware.

The data acquisition middleware used PTC's KEP-ServerEX 6.4 with U-CON driver, which can handle the welding machine's customized protocol. The data acquisition middleware was linked to our IIoT platform to monitor, manage, and store the data being collected.

Table 1: Qualitative analysis of IIoT data acquisition middleware systems.

Middleware (company)	Operation environment	Industrial protocols	Communication protocol	Db driver	Real time	Security
Azure IoT Hub (Microsoft) [31]	Cloud	OPC-UA, modbus	MQTT, REST	O	O	O
AWS Industrial IoT (Amazon) [32]	Cloud	OPC-UA, modbus	MQTT, REST	O	O	O
IBM PSB (IBM) [33]	On-premises	OPC-UA	MQTT, REST	O	O	O
Oracle Internet of Things Cloud Service (Oracle) [34]	Cloud	X	REST	O	X	O
Predix edge (GE digital) [35]	Cloud	OPC-UA, modbus	MQTT, REST	O	O	O
Kepware (PTC) [36]	On-premises	PLCs, modbus, OPC-DA/UA	MQTT, REST	O	O	O
PI Collect (OSIsoft) [37]	On-premises	PLCs, modbus, OPC-DA/UA	MQTT, REST	O	O	O
AVEVA Edge (Aveva) [38]	On-premises	PLCs, modbus, OPC-DA/UA	MQTT, REST	X	O	O
Mind Sphere Connect (Siemens) [39]	Cloud	PLC (Siemens), modbus, OPC-UA	MQTT, REST	X	O	O
WISE-PaaS (Adventech) [40]	Cloud	OPC-UA	MQTT, REST	O	O	O
ThingPlug (SKT) [41]	Cloud	X	MQTT, REST, OneM2M	O	X	O
N-MAS (Ntels) [42]	Cloud, om-premises	X	MQTT, REST	O	X	O
ThingSPIN (Hancom MDS) [43]	On-premises	OPC-UA, modbus	REST	O	O	O
TeraONE (DataStreams) [44]	On-premises	OPC-UA	REST, OneM2M	O	O	O
MOBIUS (KETI) [45]	On-premises	X	REST, OneM2M	X	O	O
IoTEP [46]	On-premises	X	MQTT, REST, LwM2M	O	O	O
SEnviro Connect [47]	Cloud	X	MQTT, REST	O	O	O
SPLS [48]	Cloud	X	—	O	O	O

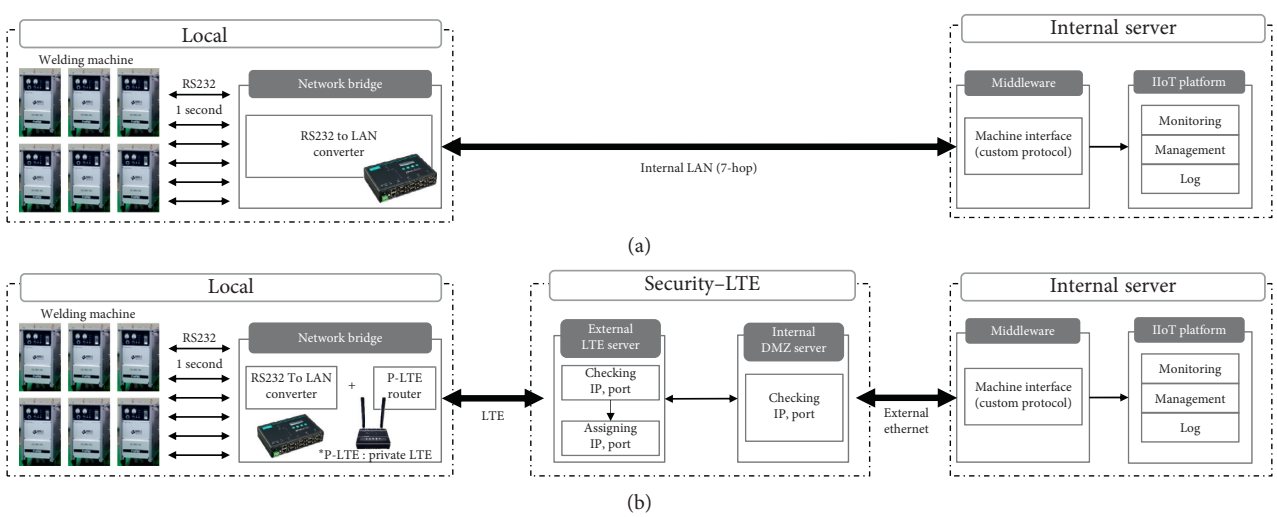


Figure 3: Experiment architecture of IIoT data-acquisition middleware using a welding machine. (a) On-premises environment using wired communication. (b) On-premises environment using LTE communication.

4.2. Result Analysis. In this section, we present and discuss our experiment results regarding the network sensitivity of IIoT data-acquisition middleware.

In our experiment the data-acquisition rates were calculated on a per-second basis. Thus, if the total count of data received reaches 864,000, its daily acquisition rate means 100% for 10 days.

As shown in Figure 5, we compare the data-acquisition ratios of wired and LTE communications for a total of 10 days. In the wired communication, the data-acquisition rate is from 99.984% to 98.537%. In LTE communication, on the contrary, the data-acquisition rate is 99.984% to 97.739%.

Figure 6 illustrates the per-hour data-acquisition rates for 24 h. Business hours are from 08 : 00 to 20 : 00 during the daytime and from 20 : 00 to 06 : 00 during the overnight. Most employees typically work during the daytime, so it is possible to confirm whether the network load was affected by the use of an internal network. According to our results, the

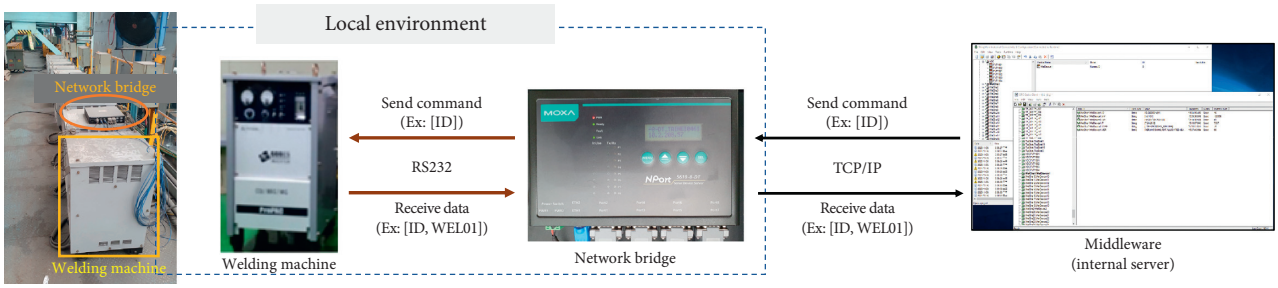


Figure 4: Environment settings and data descriptions for acquiring welding machine data at shipyard.

Table 2: Configuration settings in different communication methods.

	Wired communication	LTE communication
IIoT device	Welding machine	
Interface to IIoT device	RS232 (serial communication)	
Number of tags	15 EA	
Protocol	Customized protocol	
Network bridge	Used in RS232 to LAN converters	
LAN speed	100 Mbps	
Period of test	10 days	
Hop count (network distance)	7	8 or more
Latency (ping test)	Under 1 ms	Under 80 ms
Number of devices	10 EA	4 EA
Total number of dataset	9,929,890	3,276,842

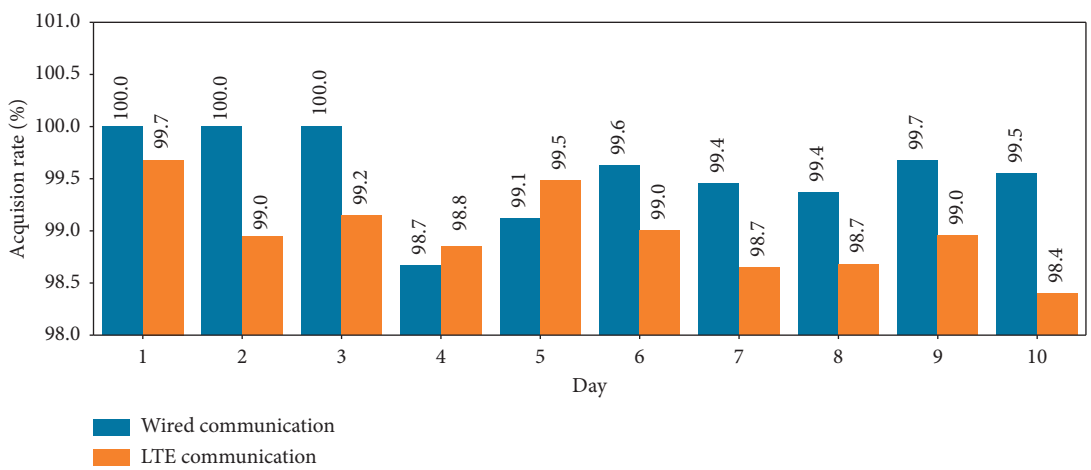


Figure 5: Average rate of IIoT data acquisition over 10 days.

network load turned out to be unaffected about acquiring IIoT data.

Table 3 exhibits our results about the interarrival time of data. A one-second interval (at the second row below the header of Table 3) is considered as *normal*, and *unstable*, otherwise. Because the ratio of the 1s interval differs by about 1% between wired and LTE communications (99.730% vs. 98.857%), we empirically confirmed that wired communication was not significant but more reliable than wireless communication in our IIoT environment (although this observation could be obvious).

Table 4 demonstrates the average data-acquisition rate of each of the welding machines used in our experiments. In the table, for all datasets, the averages were 99.940% for wired communication and 98.983% for LTE communication, respectively. In the case of wired communication, the hop count was seven, but the network in the case of LTE was more complex as it passed through eight hops or more through the external networks and the internal DMZ server. Thus, a lower data-acquisition rate was expected. Moreover, wired communication did not acquire 100% of the data due to communication errors in the device, middleware, and timer.

In this experiment, the overall average data-acquisition rate including wired and wireless communication was 99.701% despite centralized acquisition. We also confirmed

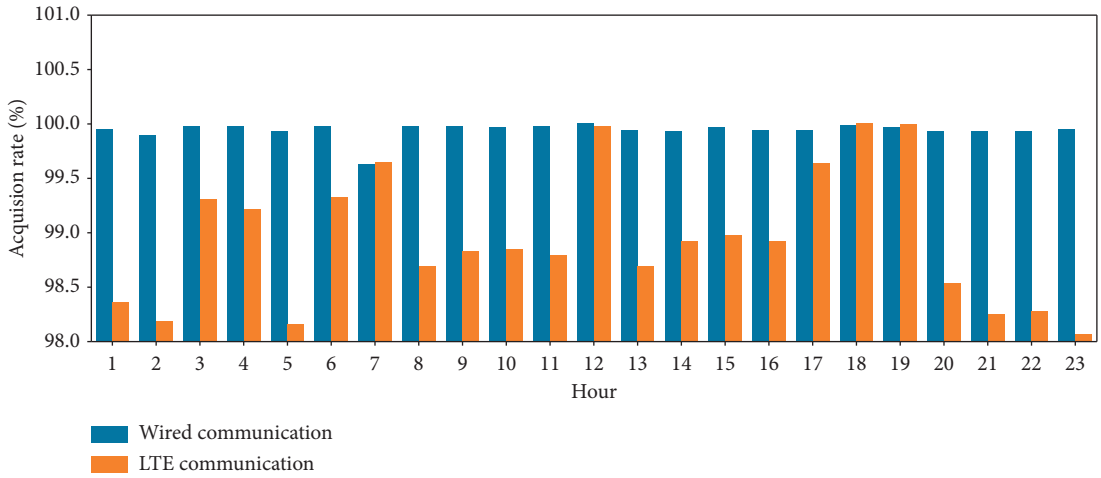


Figure 6: Average rate of IIoT data acquisition per hour for a day.

Table 3: Interarrival times between previous and current data packets.

Data Inter-arrival Time	Wired communication		LTE communication	
	Total count	Ratio (%)	Total count	Ratio (%)
0-1	20841	0.201	6858	0.200
1	10333828	99.730	3381751	98.857
2	4797	0.046	28386	0.830
3-5	1393	0.013	3541	0.104
5-10	518	0.005	183	0.005
>10	404	0.004	103	0.003

Table 4: Average data-acquisition rates of different welding machines.

	ID	Total count	Acquisition ratio (%)
Wired communication (average ratio: 99.940%)	Welder #1	863779	99.974
	Welder #2	863715	99.967
	Welder #3	863785	99.975
	Welder #4	863754	99.972
	Welder #5	863742	99.970
	Welder #6	860002	99.537
	Welder #7	863761	99.972
	Welder #8	863845	99.982
	Welder #9	863861	99.984
	Welder #10	863862	99.984
	Welder #11	863861	99.984
	Welder #12	863862	99.984
LTE communication (average ratio: 98.983%)	Welder #13	863862	99.984
	Welder #14	858271	99.337
	Welder #15	844462	97.739
	Welder #16	854244	98.871

that 98.983% of the data can be acquired although LTE communication was used.

To configure the same data acquisition middleware in the cloud, the use of an edge device is required to transfer data from the device to the cloud, taking into consideration security, as data is sent to external networks. In the cloud environment, the initial cost of infrastructure configuration is low. Thus, having a small number of IIoT equipment is advantageous. However, in the case of large-scale facilities, the operating costs increase with the increase in data transmission volume and data processing problems. Therefore, it seems that cost, maintenance, and security should be addressed well when an operation environment is selected. Currently, numerous hybrid systems combined with the on-premises and cloud are being used to do so.

5. Conclusion and Future Work

We conducted an in-depth survey of recent IIoT platforms with potentiality for horizontal data acquisition. We

reviewed various data-acquisition middleware products released by eighteen companies and research institutes. Through our investigation, we derived well-defined criteria by which the systems can be categorized. We also presented the major functionalities for building high-quality centralized IIoT data-acquisition middleware. To justify one of these criteria (network), we empirically evaluated the performance of centralized data acquisition via wired and LTE communications using an actual IIoT device (a welding machine). The overall average rate of 16 welding machines across the wired and wireless networks was 99.701%, validating the centralized IIoT data acquisition. Finally, we identified several challenges that must be resolved to construct the best data acquisition middleware in a centralized environment.

We expect that our work will help to clarify the criteria and the important considerations of high-quality IIoT data acquisition middleware systems. We plan to build our own data acquisition middleware that can fully meet the suggested functionalities. The middleware configuration and operation will be tested in a real production environment.

Data Availability

The experiment data are the property of Daewoo Shipbuilding & Marine Engineering Co., Ltd. (DSME). Therefore, the experiment data are proprietary to DSME.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported in part by DSME, Korea, under the DSME Industrial Application R&D Institute support research program and in part by Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (no. 2016-0-00145, Smart Summary Report Generation from Big Data Related to a Topic). The results of this experiment and research use DSME assets.

References

[1] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial internet of things: challenges, opportunities, and directions," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4724–4734, 2018.

[2] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital twin in industry: state-of-the-art," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2405–2415, 2018.

[3] M. Zdravković, M. Trajanović, J. Sarraipa et al., "Survey of internet-of-things platforms," in *Proceedings of the 6th International Conference on Information Society and Techology*, Kopaonik, Serbia, February 2016.

[4] J. Guth, U. Breitenbücher, M. Falkenthal et al., *A Detailed Analysis of IoT Platform Architectures: Concepts, Similarities, and Differences*, Internet of Everything, London, UK, 2017.

[5] J. Guth, U. Breitenbücher, M. Falkenthal, F. Leymann, and L. Reinfurt, *Comparison of IoT Platform Architectures: A Field Study Based on a Reference Architecture*, Cloudification of the Internet of Things, Stuttgart, Germany, 2016.

[6] A. A. Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: a survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[7] L. D. Xu, W. He, and S. Li, "Internet of things in industries: a survey," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233–2243, 2014.

[8] A. H. Ngu, M. Gutierrez, V. Metsis, S. Nepal, and Q. Z. Sheng, "IoT middleware: a survey on issues and enabling technologies," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 1–20, 2017.

[9] A. Atmani, I. Kandrouch, N. Hmina, and H. Chaoui, "Big data for Internet of Things: a survey on IoT frameworks and platforms," in *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019). AI2SD 2019. Lecture Notes in Networks and Systems*, M. Ezziyyani, Ed., vol. 92, Springer, Cham, Switzerland, 2020.

[10] H. Hejazi, H. Rajab, T. Cinkler, T. Cinkler, and L. Lengyel, "Survey of platforms for massive IoT," in *in Proceedings of the IEEE International Conference on Future IoT Technologies*, pp. 1–8, Eger, Hungary, January 2018.

[11] M. Aazam, S. Zeadally, and K. A. Harras, "Deploying fog computing in industrial internet of things and industry 4.0," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4674–4682, 2018.

[12] H. Cho and J. Jeong, "Implementation and performance analysis of power and cost-reduced OPC UA gateway for industrial IoT platforms," in *Proceedings of the 28th International Telecommunication Networks and Applications Conference*, pp. 1–3, Sydney, Australia, November 2018.

[13] H. Choi, J. Song, and K. Yi, "Brightics-IoT: Towards Effective Industrial IoT Platforms for Connected Smart Factories," in *Proceedings of the IEEE International Conference on Industrial Internet*, pp. 146–152, Seattle, WA, USA, October 2018.

[14] W. Wang, S. L. Capitaneanu, D. Marinca, and E.-S. Lohan, "Comparative analysis of channel models for industrial IoT wireless communication," *IEEE Access*, vol. 7, pp. 91627–91640, 2019.

[15] P. Duan, Y. Jia, L. Liang, J. Rodriguez, K. M. S. Huq, and G. Li, "Space-reserved cooperative caching in 5G heterogeneous networks for industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 6, pp. 2715–2724, June 2018.

[16] C. Hsu, Y. Hsu, and H. Wei, "Energy-efficient and reliable MEC offloading for heterogeneous industrial IoT networks," in *Proceedings of the 2019 European Conference on Networks and Communications (EuCNC)*, pp. 384–388, Valencia, Spain, June 2019.

[17] K. Mekki, E. Bajic, F. Chaxel, and F. Meyer, "A comparative study of LPWAN technologies for large-scale IoT deployment," *The Korean Institute of Communications and Information Sciences*, vol. 5, pp. 1–7, 2019.

[18] G. Premsankar, B. Ghaddar, M. Slabicki, and M. D. Francesco, "Optimal configuration of LoRa networks in Smart cities," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 12, 2020.

[19] A. Mahmood, E. Sisinni, L. Guntupalli, R. Rondón, S. A. Hassan, and M. Gidlund, "Scalability analysis of a LoRa network under imperfect orthogonality," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, 2019.

[20] E. Sisinni, P. Ferrari, D. Fernandes Carvalho et al., "LoRaWAN range extender for industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5607–5616, 2020.

[21] L. Leonardi, F. Battaglia, and L. Lo Bello, "RT-LoRa: a medium access strategy to support real-time flows over LoRa-based networks for industrial IoT applications," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10812–10823, 2019.
[22] M. Ballerini, T. Polonelli, D. Brunelli, M. Magno, and L. Benini, "NB-IoT versus LoRaWAN: an experimental evaluation for industrial applications," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 12, pp. 7802–7811, 2020.
[23] M. Zhu, L. Chang, N. Wang, and I. You, "A smart collaborative routing protocol for delay sensitive applications in industrial IoT," *IEEE Access*, vol. 8, pp. 20413–20427, 2020.
[24] R. Amin, S. Nazir, and I. García-Magariño, "A collocation method for numerical solution of nonlinear delay integro-differential equations for wireless sensor network and internet of things," *Sensors*, vol. 20, no. 7, p. 1962, 2020.
[25] L. Wang, Y. Ali, S. Nazir, and M. Niazi, "ISA evaluation framework for security of internet of health things system using AHP-TOPSIS methods," *IEEE Access*, vol. 8, pp. 152316–152332, 2020.
[26] X. Huang and S. Nazir, "Evaluating security of internet of medical things using the analytic network process method," *Security and Communication Networks*, vol. 2020, Article ID 8829595, 14 pages, 2020.
[27] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman, "A trustworthy privacy preserving framework for machine learning in industrial IoT systems," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 9, pp. 6092–6102, 2020.
[28] Editorial, "Blockchain in industrial IoT applications: security and privacy advances, challenges, and opportunities," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4119–4121, 2020.
[29] T. Kumar, E. Harjula, M. Ejaz et al., "Blockedge: blockchain-edge framework for industrial IoT networks," *IEEE Access*, vol. 8, pp. 154166–154185, 2020.
[30] OPC Foundation, "OPC unified architecture release 1.04," 2017, https://opcfoundation.org/, viewed.
[31] 2020 https://docs.microsoft.com/ko-kr/azure/architecture/guide/iiot-guidance/iiot-architecture.
[32] 2020 https://docs.aws.amazon.com/iot/latest/developerguide/what-is-aws-iot.html.
[33] S. Bonnaud, C. Didier, and A. Kohler, "Industry 4.0 and cognitive manufacturing," 2020, https://www.ibm.com/downloads/cas/m8j5ba6r.
[34] 2020, https://docs.oracle.com/en/cloud/paas/iot-cloud/.
[35] 2020, https://www.ge.com/digital/iiot-platform/predix-edge.
[36] 2020, https://www.kepware.com/en-us/products/kepserverex/.
[37] 2020, https://www.osisoft.com/pi-system/.
[38] 2020, https://www.aveva.com/en/products/edge/.
[39] 2020, https://siemens.mindsphere.io/en.
[40] 2020, https://wise-paas.advantech.com/ko-kr/marketplace.
[41] 2020, https://www.sktiot.com/iot/introduction/thingplug/thingplugMain.
[42] 2020, https://www.ntels.com.
[43] 2017, https://www.hancommds.com.
[44] 2020, http://www.datastreams.co.kr/.
[45] 2020, http://tech.iotocean.org/.
[46] F. Terroso-Saenz, A. González-Vidal, A. P. Ramallo-González, and A. F. Skarmeta, "An open IoT platform for the management and analysis of energy data," *Future Generation Computer Systems*, vol. 92, pp. 1066–1079, 2019.
[47] S. Trilles, A. González-Pérez, and J. Huerta, "An IoT platform based on microservices and serverless paradigms for smart farming purposes," *Sensors*, vol. 20, no. 20, p. 2418, 2020.
[48] Y. Zhang, Z. Guo, J. Lv, and Y. Liu, "A framework for smart production-logistics systems based on CPS and industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 9, pp. 4019–4032, 2018.
[49] S. Nazir, Y. Ali, N. Ullah, and I. García-Magariño, "Internet of things for healthcare using effects of mobile computing: a systematic literature review," *Internet of Things for Healthcare Using Wireless Communications or Mobile Computing*, vol. 2019, Article ID 5931315, 20 pages, 2019.
[50] R. S. Alonso, I. Sittón-Candanedo, Ó. García, J. Prieto, and S. Rodríguez-González, "An intelligent edge-IoT platform for monitoring livestock and crops in a dairy farming scenario," *Ad Hoc Networks*, vol. 98, Article ID 102047, 2020.
[51] A. R. Jadhav, S. Kiran, and R. Pachamuthu, "Development of a novel IoT-enabled power- monitoring architecture with real-time data visualization for use in domestic and industrial scenarios," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2021.

Hindawi
Scientific Programming
Volume 2021, Article ID 6645345, 10 pages
https://doi.org/10.1155/2021/6645345

Research Article

A New Big Data Feature Selection Approach for Text Classification

Houda Amazal and Mohamed Kissi

Computer Science Laboratory, Faculty of Sciences and Technologies, University Hassan II Casablanca, Mohammedia, Morocco

Correspondence should be addressed to Houda Amazal; houda.kamouss@gmail.com

Received 27 December 2020; Revised 16 March 2021; Accepted 4 April 2021; Published 19 April 2021

Academic Editor: Shaukat Ali

Feature selection (FS) is a fundamental task for text classification problems. Text feature selection aims to represent documents using the most relevant features. This process can reduce the size of datasets and improve the performance of the machine learning algorithms. Many researchers have focused on elaborating efficient FS techniques. However, most of the proposed approaches are evaluated for small datasets and validated using single machines. As textual data dimensionality becomes higher, traditional FS methods must be improved and parallelized to handle textual big data. This paper proposes a distributed approach for feature selection based on mutual information (MI) method, which is widely applied in pattern recognition and machine learning. A drawback of MI is that it ignores the frequency of the terms during the selection of features. The proposal introduces a distributed FS method, namely, Maximum Term Frequency-Mutual Information (MTF-MI), based on term frequency and mutual information techniques to improve the quality of the selected features. The proposed approach is implemented on Hadoop using the MapReduce programming model. The effectiveness of MTF-MI is demonstrated through several text classification experiments using the multinomial Naïve Bayes classifier on three datasets. Through a series of tests, the results reveal that the proposed MTF-MI method improves the classification results compared with four state-of-the-art methods in terms of macro-F1 and micro-F1 measures.

1. Introduction

Feature selection (FS) plays a key role in data mining [1], especially in text classification task that suffers from large dimensionality [2] in many application domains such as sentiment analysis [3], emotion identification [4, 5], and spam filtering [6]. Feature selection aims to select relevant and informative features (words) from large datasets [7]. Therefore, FS can reduce space dimensionality, decrease the running time in the classification process, and improve the efficiency of machine learning algorithms [8]. For this aim, FS is considered as a critical technique because it directly affects the accuracy of classification.

The FS methods can be divided into two major categories, namely, filter and wrapper methods [9]. Filter approach methods perform a statistical analysis of the feature space to select a distinguishing subset of features. Wrapper methods employ a search strategy to determine the goodness of a feature subset by providing it to the classifier and evaluating the performance. These two steps are repeated until reaching a suitable quality feature subset for a specific classifier. Wrapper methods primarily achieve better classification results than filter methods; however, they have a very high computational complexity [10] and are only efficient when the number of features is relatively small [11]. In contrast, the filter methods are efficient, scalable, and independent of any classifier interaction during the construction of the feature set. The need for classifier interaction may increase the execution time and make the FS method valuable only to a specific learning algorithm. Thus, filter methods are more suitable for large datasets [12].

Moreover, although most available FS methods for text classification are filter-based, these methods do not work when the datasets are large because they are based on the serial programming model. More precisely, classical FS algorithms need to read data into memory for analysis, but a limited memory cannot deal with the storage and processing of large datasets. Thus, FS methods are needed for

distributed environments, such as Hadoop, a powerful tool for distributed storage and distributed processing of large datasets [13]. Figure 1 presents a general overview of the distributed process of the filter FS approach for text classification.

Motivated by the above challenges, we introduce a parallel filter-based FS method for textual big data implemented on Hadoop. To this end, the proposed method focuses on the reduction of features using the term frequency (TF) [1] and mutual information (MI) techniques [14]. The MI technique is one of the most used filter FS techniques. However, the drawback of MI is that it chooses terms with high document frequency (DF) and low TF for features, which amplifies the importance of the low-frequency terms. Therefore, terms with low DF and high TF are not selected, which decreases the classifier performance because these terms are discriminative in classification.

(1) Documents are labeled and loaded into the Hadoop framework.

(2) An algorithm is introduced to calculate the TF values of features. Then, the average and maximum values of the TF for each feature are estimated based on the category under the Hadoop framework.

(3) An algorithm is proposed to calculate the MI value to evaluate the relationship between features and categories under the Hadoop framework.

In this paper, we present a hybrid distributed FS approach using the MapReduce paradigm to improve classification of textual big data. The proposal aims to select features with both high frequency and high feature-category dependency. Besides its independence from the classifiers, the proposed method is scalable and efficient for textual big data. The performance of the proposed method was compared with several state-of-the-art methods using three datasets, 20-Newsgroups, Reuters-21578, and WebKB, using multinomial NB as a classifier. According to the reported results, we can show that the proposal is outperforming standard methods.

The remainder of this paper is structured as follows. Section 2 introduces a brief literature review highlighting related work. In Section 3, the technical background in this work is discussed. The proposed method is explained in Section 4. Section 5 describes the experimental results, including the datasets, classifier, and performance measures used in the experiments. Finally, Section 6 presents the conclusion and future work.

2. Related Work

This work is focused on MI and parallel FS methods. Therefore, in the following context, we briefly present some related works on these two aspects.

Hadoop is the most used open-source MapReduce software to handle big data [15]. In [16], the authors presented a parallel FS method using MapReduce for text classification. Moreover, MI based on Renyi entropy was used to measure the correlation between features and classes.

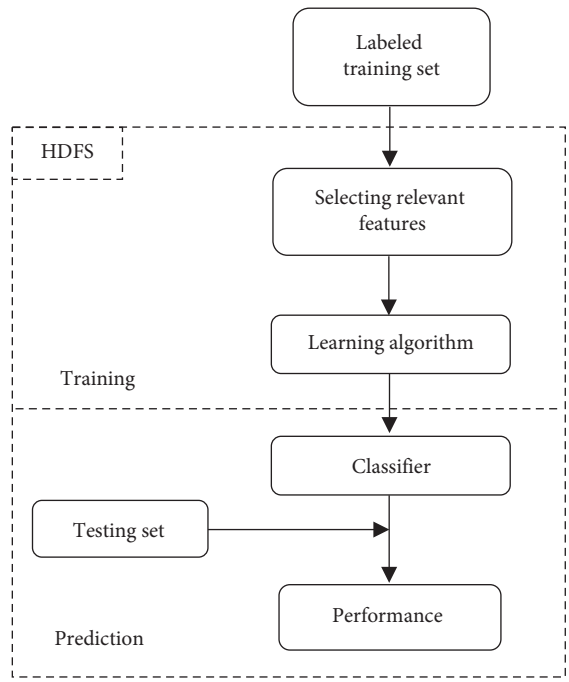


Figure 1: Feature selection process in HDFS.

Then, the maximum MI theory was used to generate the most distinguishing feature subset. In [17], the authors investigated the design and scalability of an MI-based algorithm, which is the minimum redundancy maximum relevance algorithm in MapReduce, and examined its performance in dense and sparse data.

In [18], the authors proposed a high-dimensional FS algorithm based on a variance study. The algorithm selects features by estimating their capacities to justify data variance. In [19], the authors explored a parallel FS method based on MI. However, the mentioned method is only applied to process discrete variables. In [20], the authors implemented a set of FS techniques based on a statistical test. All methods were parallelized using MapReduce on the Hadoop platform, and each feature was estimated independently. In [21], the authors introduced a MapReduce approach to derive a subset of features from large datasets. The proposed method was evaluated using classifiers, such as Support Vector Machine, Naïve Bayes, and Logistic Regression. The measurements revealed that the spark implemented framework was useful to perform evolutionary FS on massive datasets with improved classification precision and execution time.

In [22], the authors proposed a parallel FS algorithm, namely, the parallel forward-backward with pruning algorithm, for large datasets. The experimental study established increased scalability with running time. In [23], the authors proposed using MI to reduce dimensionality and improve accuracy for online streams. The proposed study focused on presenting a methodology to address the computational cost, the stability of the generated results, and the size of the final subset of selected features. In [24], the authors introduced a hybrid FS algorithm for a gene dataset by combining the MI maximization and adaptive genetic algorithm

(MIMAGA) to improve the competence of the MIMAGA algorithm. In [25], the authors proposed an evaluation of the MI-based FS methods.

In [26], the authors considered MI-based FS to increase the searching ability of the relevant subset of features. Based on MI, many studies have recently focused on maximizing the relevance of variables while minimizing variable redundancy to improve the quality of the selected features and reduce the space dimensionality [27–29].

In most of the works on FS, researchers have worked on binary classification rather than textual datasets. Selecting the most relevant features from a large volume of data has become the most significant challenge in many applications, especially in text classification [30]. As the amount of the data continues to grow, conventional algorithms cannot adapt in terms of memory requirements, execution time, and efficiency of the results. Thus, to address these large-dimensional problems, this work proposes selecting characteristics for text classification using the multicluster environment of Hadoop.

3. Technical Background

This section presents some basic concepts associated with the proposed FS approach, MTF-MI, and the parallelization technology used in our implementation (MapReduce).

3.1. Representation Phase. In this section, we denote $C = \{c_1, c_2, \ldots, c_k\}$ as the set of categories. Broadly, the documents from dataset are represented using word vectors. This representation is generated by the vector space model that uses the bag-of-words approach [31]. Thus, a text document of a category c_k is represented by a vector of features in this document. The jth document is denoted by vector $T_j = \{t_{1j}, t_{2j}, \ldots, t_{mj}\}$, where m is the number of terms in document d_j.

3.2. Mutual Information (MI). The MI is an essential concept in information theory. It is used to measure the degree of correlation between two random events [32]. In FS, MI is often used to represent the relationship between a feature and category. The MI between a feature t_i and a category c_k is defined as follows:

$$\mathrm{MI}(t_i, c_k) = \log \frac{p(t_i, c_k)}{p(t_i) \times p(c_k)}. \quad (1)$$

The approximate formula is the following:

$$\mathrm{MI}(t_i, c_k) = \log \frac{A \times N}{(A + C) \times (A + B)}, \quad (2)$$

where A is the number of documents in c_k containing t_i, B is the number of documents not in c_k containing t_i, C is the number of documents in c_k not containing t_i, and N is the total number of training documents.

Because MI does not consider the frequency of features in a text document, if two features appear in a document, their MI value is the same regardless of how often they occur. Thus, it is also necessary to consider the feature frequency in each document of the training dataset.

3.3. Hadoop Parallel Distributed Architecture. Faced with the continuous growth of data, traditional data analysis systems cannot store and process such a large volume of data. Thus, the best solution to manage the abundant data is to store it in the Hadoop distributed file system (HDFS). Due to its fault tolerance mechanism, the HDFS allows Hadoop to operate reliably and very efficiently. The HDFS can be viewed just like a regular file system; the only difference is that it handles larger datasets. This system splits data into 64 MB blocks by default, making it more efficient for large datasets. The data in HDFS are stored in two forms: the actual data and its metadata, such as file location and file size. Application data are stored in the data nodes of the HDFS, and the metadata are stored in the name node. The architecture of the parallel HDFS is illustrated in Figure 2.

The HDFS is the storage unit of Hadoop, and it follows the master-slave architecture. The master node includes three elements: the job tracker, name node, and secondary name node, whereas the slave node includes the task tracker and data node. The name node in the parallel HDFS architecture interacts with different data nodes residing in the slave nodes, whereas the job tracker in the master node organizes the task trackers on the slave nodes.

3.4. MapReduce. MapReduce is a programming model used in a distributed and parallel environment for processing large datasets [33]. The data processing in MapReduce is based on input data distribution; several tasks across many nodes execute the distributed data. A MapReduce program is divided into two main phases, map and reduce, and is executed in three steps: map, shuffle, and reduce. Figure 3 depicts the architecture of MapReduce. In the map step, input data are partitioned among nodes, and each partition of data is given as an input to a job that performs the map function. Each job processes the data and outputs key-value pairs. In the shuffle step, key-value pairs are grouped by key, and each group is sent to the reducer. The map and reduce functions are defined depending on the purpose of the application. The input and output of these functions are based on the key-value scheme. Thus, the MapReduce model allows the user to focus on the application without concern about issues, such as the program execution process on the distributed nodes, memory management, and fault tolerance. Apache Hadoop is a widely used open-source implementation of the MapReduce model.

4. Proposed Method

To deal with the problems described above, we introduced an improved MI FS approach called MTF-MI. This method introduces TF and term distribution to the classical MI method. The entire process of the proposed approach is described in Figure 4.

After the preprocessing step, including removing stop words, tokenization, and stemming, the documents are

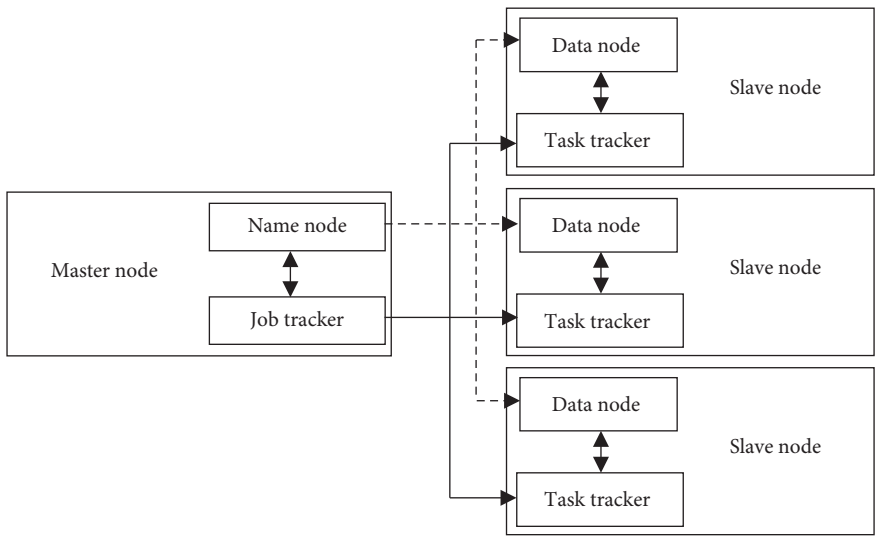


Figure 2: Hadoop distributed file system.

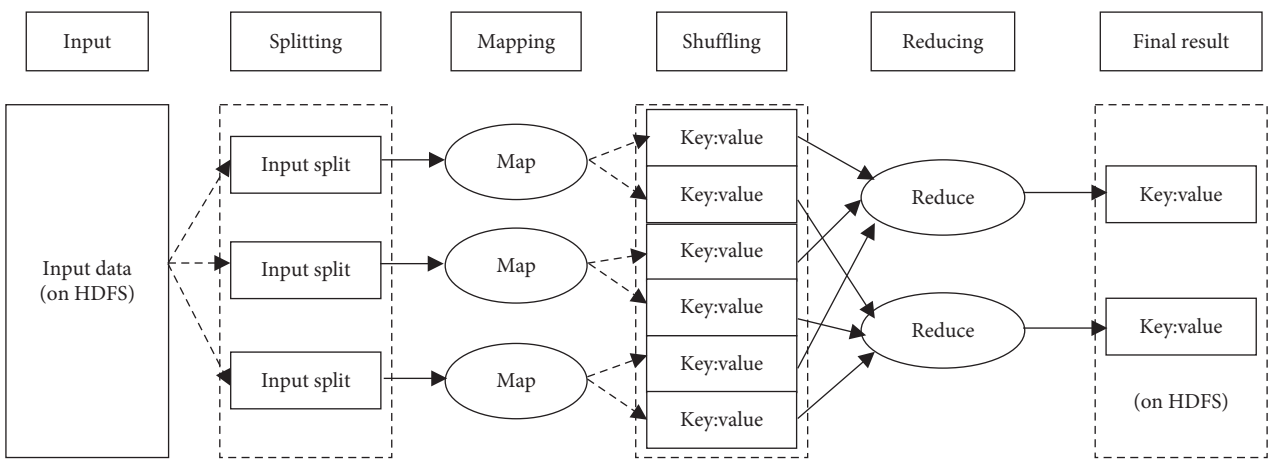


Figure 3: Phases of MapReduce.

loaded into the HDFS and are represented as described in Section 3. To redress the drawback of the traditional MI method, the TF is introduced first. tf_{ij} represents the TF of a term in a document d_j. Hence, the average term frequency $\overline{tf_i}$ and the maximum term frequency $tf_{i\max}$ for a specific category c_k can be calculated as follows:

$$\overline{tf_i} = \frac{1}{N_k}\sum_{j=1}^{N_k} tf_{ij},$$
$$tf_{i\max} = \max_{j=1}^{N_k}\{tf_{ij}\}, \tag{3}$$

where N_k is the number of documents belonging to category c_k. As the MI method is based on DF, according to its classical formula, if a term occurs many times in a document of a particular category when this type of document is rare, this term is not considered discriminative. Therefore, in this work, the TF is introduced in the MI formula. The term distribution is used to select more discriminative features. For a particular category, a feature has more discriminating power if it is regularly distributed. For this, the sample variance is used to evaluate the difference in term distributions. Sample variance is a commonly used statistics metric that measures the dispersion degree of a dataset. The sample variance is given as follows:

$$v(t_i, c_k) = \frac{1}{N_k - 1}\sum_{j=1}^{N_k}\left(tf_{ij} - \overline{tf_i}\right)^2 + \alpha. \tag{4}$$

The variable α denotes a very small real number. Finally, we introduce our approach based on the TF and term distribution to evaluate the feature t_i in category c_k as follows:

$$\mathrm{MTF-MI}(t_i, c_k) = \frac{tf_{i\max} \times \mathrm{MI}(t_i, c_k)}{v(t_i, c_k)}. \tag{5}$$

In the proposed method, to select terms with high discriminability power, as the TF is high and the DF is relatively low, we use the maximum TF $tf_{i\max}$, instead of the average. Based on the basic theory of MI, the greater the

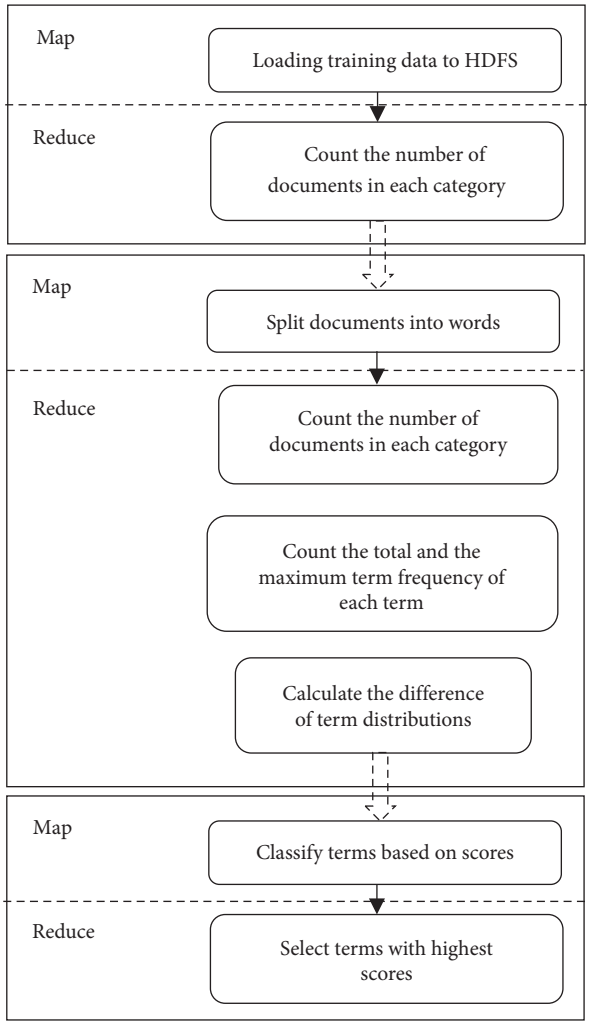


Figure 4: Proposed system.

value of MI is, the more category information the feature has. Hence, the formula is defined as follows:

$$\mathrm{MTF-MI}_{\max}(t_i) = \max_{k=1}^{M}\{\mathrm{MTF-MI}(t_i, c_k)\}, \quad (6)$$

where M is the total number of categories in the dataset.

Finally, the proposed method is implemented using MapReduce. The parallel implementation of the overall MTF-MI includes three process stages: job 1, job 2, and job 3.

Job 1 is achieved using Algorithm 1. Job 1 reads the incoming batch of training data and calculates the number of documents in each category. The results are used in job 2, which is achieved using Algorithm 2. For each term t_i belonging to category c_k, the total and maximum TF of t_i are calculated and stored in sum and $tf_{i\max}$, respectively. Then, using the value of sum, the average tf_i is calculated. Next, the difference in the term distributions is calculated for each term t_i in category c_k. Finally, the proposed approach calculates the value of term t_i in category c_k. Job 3 is achieved using Algorithm 3. Job 3 takes the values emitted by job 2 and assigns each term t_i to the category with the maximum score. Then, all features are sorted in descending order, and the x terms whose values are maximal are selected as the relevant features.

5. Experiments

The multinomial NB classifier [34] is used on three different datasets with different characteristics to validate the performance of the proposed MTF-MI. Two different commonly used measures, micro-F1 and macro-F1, were applied to observe the effectiveness of the MTF-MI method. The datasets and evaluation measures are briefly described in the following sections, and the experimental results are also presented.

5.1. Datasets. We achieved experiments with the Reuters-21578 [35], 20-Newsgroups [36], and WebKB [37] datasets. The Reuters-21578 dataset contains the news that appeared on the Reuters newswire in 1987 and belong to one out of eight possible categories. The 20-Newsgroups dataset contains around 20,000 documents that are taken from the Usenet newsgroup collection, and all documents were assigned uniformly to 20 different categories. The WebKB dataset is a subset of web documents, which contains 877 webpages from the computer science departments of four universities.

5.2. Naïve Bayes Classifier. The Naïve Bayes (NB) classifier is a simple probabilistic algorithm based on the Bayes theorem with a strong independence assumption [38]. The NB model is based on simplifying conditional independence assumption, which consists of, given a category, the words which are conditionally independent of each other. This assumption does not affect the accuracy of text classification and makes fast classification algorithms applicable to the problem. For this, NB is widely used in classification problems in real-world applications.

5.3. Performance Measures. In this study, two commonly used measures are employed, which are the macro-F1 and the micro-F1 [39]. In macro-F1, F-measure is calculated for each category within the dataset and then the average over all classes is obtained. Hence, the same weight is assigned to each category without regarding the class frequency. Macro-F1 can be formulated as follows:

$$\mathrm{Macro}-F1 = \frac{\sum_{k=1}^{c} F_k}{C}, \qquad F_k = \frac{2 \times p_k \times r_k}{p_k + r_k}, \quad (7)$$

where couple of (p_k, r_k) corresponds to precision and recall values of class k, respectively.

However, in micro-F1, the F-measure is computed globally without class discrimination. In this way, all classification decisions in the whole dataset are considered. If the classes in a dataset are biased, large classes dominate small

(i) Map
(ii) Input:
(iii) key: document name
(iv) value: document content
(v) Emit (ck, dj)
(vi) EndMap
(vii) Reduce
(viii) Input:
(ix) key: ck
(x) values: list $[dj]$
(xi) $N_k \longleftarrow 0$ //total number of documents in the category c_k
(xii) for each value in values do
(xiii) $\llcorner N_k++$
(xiv) Emit (c_k, N_k)
(xv) EndReduce.

Algorithm 1: Job 1.

(i) Map
(ii) Input:
(iii) key: Offset
(iv) value: line of document
(v) Emit $((t_i, c_k), dj)$
(vi) EndMap
(vii) Reduce
(viii) Input:
(ix) key: (t_i, c_k)
(x) values: list $[d_j]$
(xi) for each value in values do
(xii) $\llcorner \text{sum}(t_i) += tf_{ij}$
(xiii) $tf_{i\max} = \max\{tf_{i\max}, tf_{ij}\}$
(xiv) $\overline{tf_i} = \text{sum}(t_i)/N_k$ for each value in values do
(xv) $v(t_i, c_k) = (tf_{ij} - \overline{tf_i})^2/(N_k - 1) + \alpha$
(xvi) $\text{MTF-MI}(t_i, c_k) = tf_{i\max} \times \text{MI}(t_i, c_k)/v(t_i, c_k)$
(xvii) emit $((t_i, c_k), \text{MTF} - \text{MI}(t_i, c_k))$
(xviii) EndReduce.

Algorithm 2: Job 2.

(i) Map
(ii) Input:
(iii) key: (t_i, c_k)
(iv) value: MTF-MI (t_i, c_k)
(v) emit $(t_i, (c_k, \text{MTF} - \text{MI}(t_i, c_k)))$
(vi) EndMap
(vii) Reduce
(viii) Input:
(ix) key: t_i
(x) values: list $[(c_k, \text{MTF} - \text{MI}(t_i, c_k))]$
(xi) $\text{MTF} - \text{MI}(t_i) = \max \text{MTF} - \text{MI}(t_i, c_k)$
(xii) Emit $(t_i, \text{MTF} - \text{MI}(t_i))$
(xiii) EndReduce.

Algorithm 3: Job 3.

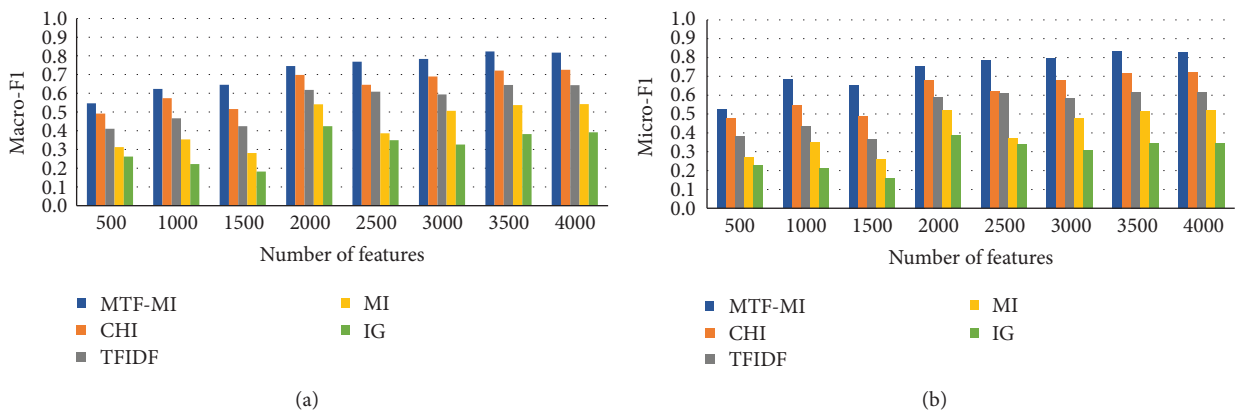


Figure 5: Macro-F1 and micro-F1 measures of multiclass text classification on the Reuters-21578 corpus with different numbers of features.

ones in microaveraging. Micro-F1 can be formulated as follows:

$$\text{Micro} - F1 = \frac{2 \times p \times r}{p + r}, \tag{8}$$

where pair (p, r) represents the precision and recall values, respectively, over all the classification decisions within the whole dataset, rather than the individual classes.

5.4. Results and Discussion. The macro-F1 and micro-F1 performances of MTF-MI are compared to four widely used feature selection techniques using Naïve Bayes classifier applied on three datasets (20 Newsgroups, Reuters-21578, and WebKB). The four feature selection techniques used for comparison are the classical MI, Chi-square (CHI), Term Frequency-Inverse Document Frequency (TF-IDF), and Information Gain (IG).

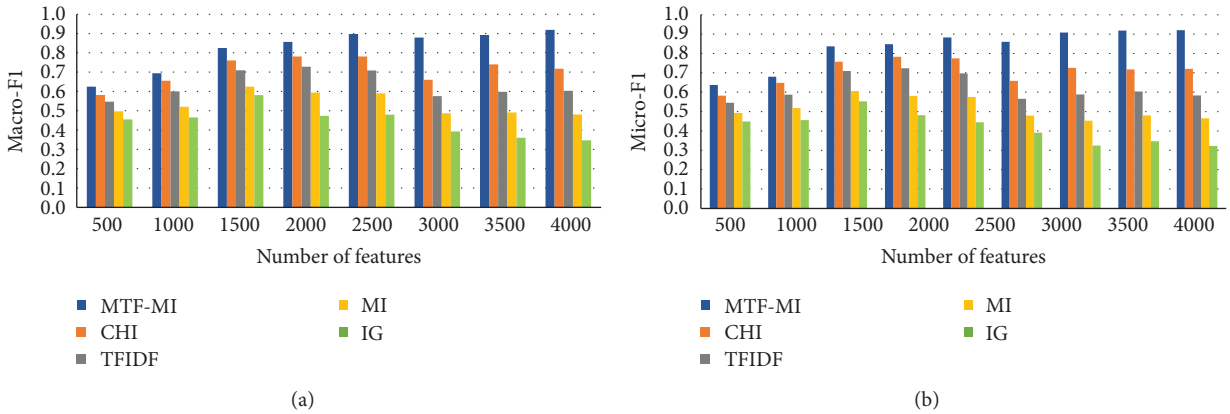
Figures 5–7 show the classification performance of the different feature selection methods for the three datasets. In all figures, the horizontal and vertical axes present the number of selected features and the corresponding classification performance, respectively.

Figure 5 presents the F1 classification performance based on 5 term weighting methods using NB classifier with different feature dimensionalities. It is noticeable that the proposed approach outperforms all other standard methods in terms of micro-F1 and macro-F1. Figure 5 shows that macro-F1 and micro-F1 of MTF-MI are close to those of CHI when 500 and 1000 features are selected. It is noticeable that the IG and MI techniques showed the lowest performance. The micro-F1 results are noticed to be high (more than 83%) using 3500 features, and the highest classification F1 value (82%) is achieved by the MTF-MI method. Moreover, it is noticeable that proposed method performs well for less than 1500 features as its F1 values range between 54% and 64%, while the performances of other methods were very weak on the same range of features. Although the categorical documents distribution in the Reuters-21578 dataset is highly skewed, the results show that NB classifier performs better on the representation of the proposed method MTF-MI. In the Reuters-21578 dataset, the boundaries between categories are apparent. Therefore, good classification performance can be achieved with a small number of features (3500). However, when the number of selected features increases, the classification performance decreases.

Figure 6 depicts the NB classification performance on the 20-Newsgroups dataset in terms of F1 measure, where it can be seen that the trend of the micro-F1 and macro-F1 performance is similar to that in Figure 5. Similar to the results of Reuters-21578 dataset, the proposed method outperforms other standard methods in micro-F1 and macro-F1. For instance, the best three micro-F1 and macro-F1 values (90%, 91%, and 92%) are reached by the MTF-MI method based on 4000 features. In contrast to the results in Figure 5, the performance of CHI method, as seen in Figure 6, is not competing with the performance of the proposed MTF-MI for features up to 3000. For example, the micro-F1 and macro-F1 values (66% and 65%, respectively) are reached by the CHI method on 3000 features, which are still less than the corresponding values of the proposed method (87% and 86%). Finally, the documents in 20-Newsgroups are almost uniformly distributed; therefore, the micro-F1 and macro-F1 performances of different schemes are noticed to be quite similar. In addition, the measure values increase as the feature number increases, which could be due to the similarity of some categories in the 20-Newsgroups dataset. Therefore, some terms are commonly present in more than one category, so when the number of selected features increases, it provides a better distinction between categories.

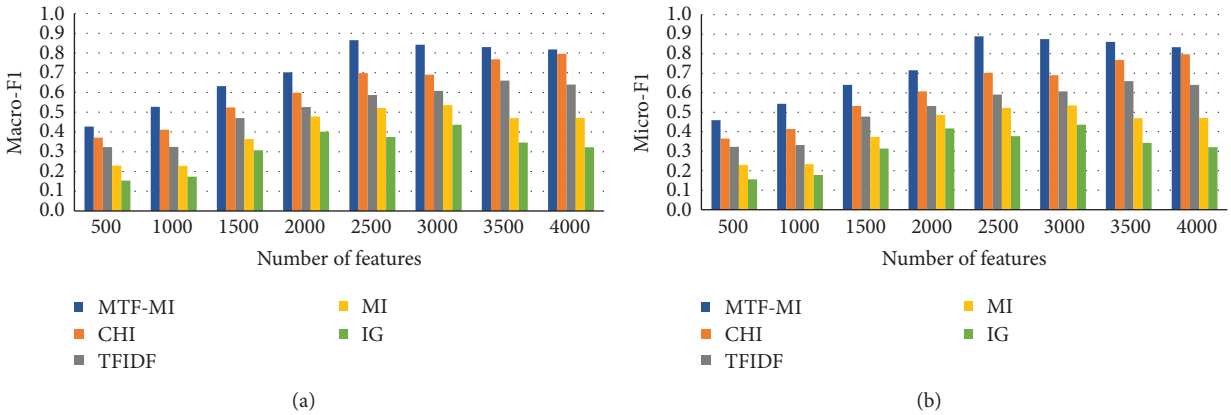
Figure 7 shows micro-F1 and macro-F1 classification performance on the WebKB dataset using NB classifier. Generally, the results in Figure 7 are similar to those in Figure 5 for standard weighting techniques, as the boundaries between categories are apparent. The proposed MTF-MI method outperforms other techniques in terms of micro-F1 and macro-F1, where the maximum micro-F1 value (86%) is achieved by MTF-MI on 3500 features. Moreover, similar to the results on Reuters-21578 and 20-Newsgroups datasets, the proposed MTF-MI has outperformed other methods with noticeable performance differences.

It can be concluded that the proposed method MTF-MI performs the highest on different corpora, which indicates



(a) (b)

Figure 6: Macro-F1 and micro-F1 measures of multiclass text classification on the 20-Newsgroups corpus with different numbers of features.



(a) (b)

Figure 7: Macro-F1 and micro-F1 measures of multiclass text classification on the WebKB corpus with different numbers of features.

that the proposed approach is effective in selecting the features and representing the data as well as their generality. Based on the experimental results on different datasets, the performance of the proposed method is more effective for the three datasets, which means that the maximum term frequency factor introduced to the classical MI plays a big role to reach high performance. Therefore, it can be concluded that the proposed MTF-MI method is more effective than the classical state-of-the-art method.

6. Conclusion

This paper introduces MTF-MI, a distributed feature selection approach designed upon the MapReduce programming model. The proposed approach, based on mutual information method, has been implemented using Apache Hadoop, and it has been applied over three different large datasets. The performance of resulting classification models generated by MTF-MI has been systematically evaluated using Naïve Bayes classifier, implemented in Hadoop framework, over a cluster of five computers. The experimental study has proved that MTF-MI efficiently improves the selection of the relevant features while discarding the selection of irrelevant ones. The proposed approach is the best in average of F-measure compared to four state-of-the-art methods, namely, CHI, TF-IDF, MI, and IG. However, this method becomes less performed for a given threshold of selected features. Although the results vary within the datasets, the general insights provided here help highlight the importance of the combination of the feature selection techniques with the distributed aspect that is added through Hadoop framework usage for the prediction tasks on large textual datasets.

As part of this work, we have also compared the proposed approach with a sequential version of MTF-MI implemented on a single machine using java. Our results showed that the sequential version is unable to handle large datasets due to memory requirements. Meanwhile, our version is fully scalable and yields better memory usage when dealing with very large datasets. Despite the multiple advantages of parallelism, it can be hazardous if not used appropriately. When large and complex datasets are used, overparallelism can cause the distribution to ignore certain meaningful relationships between features, which can negatively affect the accuracy of the results.

Data Availability

20-Newsgroups dataset is from https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups. Reuters-21578 dataset is from https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection. WebKB dataset is from http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/webkb-data.gtar.gz. 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 they have no conflicts of interest.

References

[1] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," *Concurrency and Computation: Practice and Experience*, vol. 10, Article ID e5909, 2020.

[2] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *Journal of Information Science*, vol. 43, no. 1, pp. 25–38, 2017.

[3] A. Onan, "Hybrid supervised clustering based ensemble scheme for text classification," *Kybernetes*, vol. 46, 2017.

[4] A. Onan, "Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering," *IEEE Access*, vol. 7, pp. 145614–145633, 2019.

[5] A. Onan and M. A. Tocoglu, "A term weighted neural language model and stacked bidirectional lstm based framework for sarcasm identification," *IEEE Access*, vol. 9, pp. 7701–7722, 2021.

[6] V. P. Deshpande, R. F. Erbacher, and C. Harris, "An evaluation of naïve bayesian anti-spam filtering techniques," in *Proceedings of the 2007 IEEE SMC Information Assurance and Security Workshop*, pp. 333–340, IEEE, New York, NY, USA, June 2007.

[7] L. M. Q. Abualigah, *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering*, Springer, Berlin, Germany, 2019.

[8] A. Onan, "Classifier and feature set ensembles for web page classification," *Journal of Information Science*, vol. 42, no. 2, pp. 150–165, 2016.

[9] J. Zhang, Y. Xiong, and S. Min, "A new hybrid filter/wrapper algorithm for feature selection in classification," *Analytica Chimica Acta*, vol. 1080, pp. 43–54, 2019.

[10] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature selection for text classification: a review," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3797–3816, 2019.

[11] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: a review," *Data Classification: Algorithms and Applications*, vol. 37, 2014.

[12] H. Amazal, M. Ramdani, and M. Kissi, "A parallel global tfidf feature selection using hadoop for big data text classification," in *Advances on Smart and Soft Computing*, pp. 107–117, Springer, Berlin, Germany, 2020.

[13] A. P. Rodrigues and N. N. Chiplunkar, "A new big data approach for topic classification and sentiment analysis of twitter data," *Evolutionary Intelligence*, pp. 1–11, Springer, Berlin, Germany, 2019.

[14] B. Venkatesh and J. Anuradha, "A hybrid feature selection approach for handling a high-dimensional data," in *Innovations in Computer Science and Engineering*, pp. 365–373, Springer, Berlin, Germany, 2019.

[15] D. Glushkova, P. Jovanovic, and A. Abelló, "Mapreduce performance model for Hadoop 2.x," *Information Systems*, vol. 79, pp. 32–43, 2019.

[16] Z. Li, W. Lu, Z. Sun, and W. Xing, "A parallel feature selection method study for text classification," *Neural Computing and Applications*, vol. 28, no. 1, pp. 513–524, 2017.

[17] C. Reggiani, Y. A. Le Borgne, and G. Bontempi, "Feature selection in high-dimensional dataset using mapreduce," in *Benelux Conference on Artificial Intelligence*, pp. 101–115, Springer, Berlin, Germany, 2017.

[18] Z. Zhao, R. Zhang, J. Cox, D. Duling, and W. Sarle, "Massively parallel feature selection: an approach based on variance preservation," *Machine Learning*, vol. 92, no. 1, pp. 195–220, 2013.

[19] Z. Sun and Z. Li, "Data intensive parallel feature selection method study," in *Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 2256–2262, IEEE, Beijing, China, July 2014.

[20] M. Kumar and S. Kumar Rath, "Classification of microarray using mapreduce based proximal support vector machine classifier," *Knowledge-Based Systems*, vol. 89, pp. 584–602, 2015.

[21] D. Peralta, S. Del Río, S. Ramírez-Gallego, I. Triguero, J. M. Benitez, and F. Herrera, "Evolutionary feature selection for big data classification: A mapreduce approach," *Mathematical Problems in Engineering*, vol. 2015, Article ID 246139, 11 pages, 2015.

[22] I. Tsamardinos, G. Borboudakis, P. Katsogridakis, P. Pratikakis, and V. Christophides, "A greedy feature selection algorithm for big data of high dimensionality," *Machine Learning*, vol. 108, no. 2, pp. 149–202, 2019.

[23] M. Rahmaninia and P. Moradi, "Osfsmi: online stream feature selection method based on mutual information," *Applied Soft Computing*, vol. 68, pp. 733–746, 2018.

[24] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification," *Neurocomputing*, vol. 256, pp. 56–62, 2017.

[25] C. Pascoal, M. R. Oliveira, A. Pacheco, and R. Valadas, "Theoretical evaluation of feature selection methods based on mutual information," *Neurocomputing*, vol. 226, pp. 168–181, 2017.

[26] M. Han and W. Ren, "Global mutual information-based feature selection approach using single-objective and multi-objective optimization," *Neurocomputing*, vol. 168, pp. 47–54, 2015.

[27] W. Gao, L. Hu, and P. Zhang, "Class-specific mutual information variation for feature selection," *Pattern Recognition*, vol. 79, pp. 328–339, 2018.

[28] W. Gao, L. Hu, P. Zhang, and J. He, "Feature selection considering the composition of feature relevancy," *Pattern Recognition Letters*, vol. 112, pp. 70–74, 2018.

[29] F. Macedo, M. Rosário Oliveira, A. Pacheco, and R. Valadas, "Theoretical foundations of forward feature selection methods based on mutual information," *Neurocomputing*, vol. 325, pp. 67–89, 2019.

[30] H. Amazal, M. Ramdani, and M. Kissi, "Towards a feature selection for multi-label text classification in big data," in *International Conference on Smart Applications and Data Analysis*, pp. 187–199, Springer, Berlin, Germany, 2020.

[31] B. Zhang, *Analysis and Research on Feature Selection Algorithm for Text Classification*, University of Science and Technology of China, Anhui, China, 2010.

[32] X. Tang, Y. Dai, and Y. Xiang, "Feature selection based on feature interactions with application to text categorization," *Expert Systems with Applications*, vol. 120, pp. 207–216, 2019.
[33] L. Abualigah, A. Diabat, S. Mirjalili, M. Abd Elaziz, and A. H. Gandomi, "The arithmetic optimization algorithm," *Computer Methods in Applied Mechanics and Engineering*, vol. 376, Article ID 113609, 2021.
[34] O. Aytuğ, "Sentiment analysis on twitter based on ensemble of psychological and linguistic feature sets," *Balkan Journal of Electrical and Computer Engineering*, vol. 6, no. 2, pp. 69–77, 2018.
[35] M. Jiang, Y. Liang, X. Feng et al., "Text classification based on deep belief network and softmax regression," *Neural Computing and Applications*, vol. 29, no. 1, pp. 61–70, 2018.
[36] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "Hybrid clustering analysis using improved krill herd algorithm," *Applied Intelligence*, vol. 48, no. 11, pp. 4047–4071, 2018.
[37] G. Beatty, E. Kochis, and M. Bloodgood, "The use of unlabeled data versus labeled data for stopping active learning for text classification," in *Proceedings of the 2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pp. 287–294, IEEE, Newport Beach, CA, USA, February 2019.
[38] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," *Journal of Information Science*, vol. 44, no. 1, pp. 28–47, 2018.
[39] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*, vol. 39, Cambridge University Press, Cambridge, UK, 2008.

Hindawi
Scientific Programming
Volume 2021, Article ID 5536884, 9 pages
https://doi.org/10.1155/2021/5536884

Research Article

EEWMP: An IoT-Based Energy-Efficient Water Management Platform for Smart Irrigation

Rafi Ullah,[1] Arbab Waseem Abbas,[1] Mohib Ullah,[1] Rafi Ullah Khan,[1] Irfan Ullah Khan,[2] Nida Aslam,[2] and Sumayh S. Aljameel[2]

[1]*Institute of Computer Science and Information Technology, The University of Agriculture, Peshawar, Pakistan*
[2]*Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia*

Correspondence should be addressed to Rafi Ullah Khan; rafiyz@gmail.com

Received 24 January 2021; Revised 17 March 2021; Accepted 26 March 2021; Published 8 April 2021

Academic Editor: Shah Nazir

Precision agriculture is now essential in today's world, especially for countries with limited water resources, fertile land, and enormous population. Smart irrigation systems can help countries efficiently utilize fresh water and use the excess water for barren lands. Smart water management platform (SWAMP) is an IoT-based smart irrigation project designed for efficient freshwater utilization in agriculture. The primary aim of SWAMP is to auto manage water reserves, distribution, and consumption of various levels, avoid over-irrigation and under-irrigation problems, and auto manage time to maximize production. This research proposed an energy-efficient water management platform (EEWMP), an improved version of SWAMP. EEWMP is an IoT-based smart irrigation system that uses field-deployed sensors, sinks, fusion centres, and open-source clouds. Both models' performance is evaluated in energy consumption, network stability period, packet sent to destination, and packet delivery ratio. The experimental results show that EEWMP consumes 30% less energy and increases network stability twice than SWAMP. EEWMP can be used in different irrigation models such as drip irrigation, sprinkler irrigation, surface irrigation, and lateral move irrigation with subtle alterations. Moreover, it can also be used in small farms of third-world countries with their existing communication infrastructures such as 2G or 3G.

1. Introduction

Water is an essential need for living things, and agriculture is the largest consumer of fresh water in the world with 70% consumption [1]. According to the researchers, the population growth has estimated to meet 10 billion in the twenty-first century. This rapid growth of population may create many demands and challenges for freshwater [2, 3]. Therefore, smart water management systems are essential to meet future demands of fresh water and food security. The irrigation system plays an essential role in crop yield as over-irrigation and under-irrigation may significantly affect productivity and result in power and water wastage [4]. Precision irrigation, on the other hand, is an intelligent method that can be used to avoid the wastage of power and water while increasing productivity.

Internet of Things (IoT) is a system of physical objects (appliances with software and sensors, data centres, and machines) whose purpose is to collect and exchange data with each other over the Internet [5]. IoT's primary purpose is to increase machine-to-machine communication and take optimal decisions according to the situations but with less human interaction [6]. The invention and involvement of IoT revolutionized different processes are involved in many domains such as home security, equipment manufacturing, health monitoring, automated transportation, and especially agriculture [7]. In modern agriculture, the IoT is efficiently utilized in many subdomains such as precision farming [8], smart crop monitoring [9], soil quality [10], smart irrigation systems [11, 12], and many others.

For precision agriculture, a smart irrigation system is an essential requisite. The smart irrigation decision support

system (SIDSS) is a novel technique that uses field deployed sensors to detect soil characteristics, weather and climate conditions, and crop conditions for irrigation [13]. In this regard, Kamienski et al. proposed an IoT-based smart water management platform (SWAMP) intelligent irrigation system [1, 14]. SWAMP is one of the best available IoT-based irrigation systems that use SPARQL-based [15] semantic reasoning features and open-source cloud-based IoT platform FIWARE [16, 17]. The SWAMP architecture comprises five different layers, including the device communication layer, acquisition security management layer, data management layer, water (irrigation) distribution layer, and water application services layer. In SWAMP, water management is further divided into three phases, i.e., reserve water, distribution, and consumption. Initially, SWAMP was implemented in Brazil and Europe at pilot project and produced promising results.

SWAMP is a big project and a combination of different technologies such as sensors, semantic computing, cloud services, communication protocols, drones, IoT, and many others [1, 14]. However, many inherited issues are not addressed, which may affect the SWAMP project's efficiency. One of the main issues in the SWAMP project is related to sensors' energy consumptions as they send continuous/redundant reports. An energy-aware, efficient sensor communication model will increase the sensors' lifetime and reduce the project's cost.

This research aims to improve the SWAMP system's performance by proposing energy-efficient utilization techniques for different sensors (soil moisture, temperature, and water level measuring sensors). Therefore, in this research, we proposed an energy-efficient water management platform (EEWMP) with reduced redundant data strategies. The experimental results show that EEWMP consumes 30% less energy and increases network stability twice than SWAMP. Similarly, due to increased network stability time, the destination's packets were 1.5 times more in EEWMP than SWAMP.

In Section 2, we introduce the SWAMP project briefly. In Section 3, we present a review of the latest literature on intelligent irrigation systems. In Sections 4 and 5, we discuss our proposed EEWMP model with its components and working. The methodology is discussed in Section 6. While in Sections 7 and 8, we discussed the results and conclusions with future work, respectively.

2. SWAMP Project

SWAMP is a collaborative project developed for intelligent irrigation and efficient freshwater utilization in agriculture [1, 14]. The primary aim of SWAMP is to auto manage water reserves, distribution, and consumption of various levels, avoid over-irrigation and under-irrigation problems, and auto manage time to maximize production. The proposed SWAMP architecture is divided into five layers where each layer is dedicated to a specific responsibility, and each layer communicates with other layers using RDF [18], NGSI, or NGSI-LD [19] protocols.

The first layer of SWAMP architecture is called the device and communication layer, where different types of sensors are deployed in the field to acquire various types of information such as moisture and temperature. The information is collected from sensors using drones. The second layer is called data acquisition, security, and management layer, responsible for data acquisition and management. The third layer is called a data management layer responsible for data storage, processing, and distribution. This layer also uses semantic computing engines to process the data for the next layer. The fourth layer is called the water irrigation and distribution model layer, in which different types of traditional agriculture irrigation models are used to estimate the water need. The last layer of the SWAMP model is called the water application services layer, where the water is irrigated according to the need based on the data collected previously.

3. Literature Review

Channe et al. proposed an IoT-based multidisciplinary model for precision agriculture [20]. They proposed various applications of their model such as online agriculture data analysis, agricultural cloud, agribusiness, soil and weather analysis and predictions, and mobile app for farmers, vendors, and government representatives. They aimed to improve the crop production process with updated information about fertilizer utilization, soil analysis, and future need predations. In 2017, FIGARO (Flexible and precIse irriGation plAtform to improve faRm scale water prOductivity) project was started. The FIGARO project is a decision support system proposed to manage freshwater irrigation and improve production. They used different sensors, software, and cloud and involved field experts for optimal decisions [21]. Popovic et al. presented a case study on IoT-based precision agriculture platforms that use different sensors, IoT protocols, and analytic tools for ecology monitoring and precision agriculture [22]. Similarly, Kamilaris et al. proposed a SWAMP [1, 14] like a theoretical framework for IoT-enabled smart farming that facilitates farmers by providing accurate information based on semantic reasoning and real-time stream processing for decision-making [23].

Jaiganesh et al. proposed an IoT-based elegant farming model that uses mobile devices, information processing systems, and cloud services. They also proposed an agriculture cloud (Agro Cloud) module to collect, process, and store the data [24]. Li et al. proposed an IoT-based greenhouse management system that uses various android applications, sensors, communication protocols, and different hardware. Their proposed system offers general control functions such as temperature, humidity, and light adjustment functions. Furthermore, the system also offers various monitoring functions and weather forecasting functions [25]. Kiani and Seyyedabbasi proposed a sensor and IoT-based small farm monitoring system that monitors the temperature, humidity, and soil moisture to efficiently schedule the irrigation, harvesting, and cultivation plan [26]. Nurellari and Srivastava implemented an energy-efficient agriculture field monitoring system using IoT-enabled

wireless sensors network. The system provides moisture, salinity, and soil temperature information to the farmers [27].

Karpagam et al. [28] proposed the IoT-enabled intelligent irrigation system for efficient water management and distribution. Their system monitors the water level in the field and supplies the water according to the need automatically with minimal human effort. Similarly, Gupta et al. [29] proposed an IoT-based intelligent irrigation system with a flood prevention system. They proposed a water level analysis system using well-maintained databases that measure the amount of rainfall and humidity level and predict future threats.

4. Energy-Efficient Water Management Platform (EEWMP)

The proposed energy-efficient water management platform (EEWMP) model is comprised of various sensors, computational devices, and services. The architecture of the proposed EEWMP is illustrated in Figure 1. The details of the proposed EEWMP system and architecture are discussed as follows.

4.1. Field Sensors. The sensor is a hardware device which is capable of collecting sensory information, processing it, and sending it to the base station using various communication technologies. In this research, we used three types of field sensors in the field, including soil moisture sensor, temperature sensor, and water level sensor, to collect the field data. This data of soil moisture, temperature, and the water level are then used in decision-making regarding irrigation. For experimentation, we used generic sensors available in Matlab.

4.2. In-Field Sink. The sink node's primary duty in wireless sensor network (WSN) is to collect the data from the sensors deployed in the vicinity using various strategies and conserve the nodes' energy by reducing communication traffic. In the EEEMP model, an in-field sink node is deployed to gather the whole field's data and send it to the outer sink. For experimentation, we used generic sensors available in Matlab.

4.3. Outer Sink and Fusion Centre. The outer sink node is used to collect the in-field sinks' data and then provide it to the embedded fusion centre. Fusion centre is a computational device precisely programmed for sharing information with the intellectual ability to remove redundant and empty data. The fusion centre's primary purpose is to reduce the communication traffic and conserve the energy consumed in various communication and computational tasks.

4.4. IoT Service Cloud. An IoT (Internet of Things) service cloud is an online service provided by different companies for different IoT-based services such as storage, processing, built-in and custom-built application to manage the data, and device management. Moreover, these clouds can also be accessible with desktops and handheld devices to get notifications and control devices. Several open-source and free IoT services are available such as FIWARE [30], Amazon, Microsoft, and Google's Cloud IoT [31]. In this research, the open-source cloud is used to reduce the cost. The provider cloud is used for registration/identification having a database used to save user data when the network is offline due to some reasons. The registration section is used to register the new client/end-user; identification is used to identify the end-user either the user has registered or not. All these devices of the provider cloud are connected to IoT connectivity.

4.5. Valves Controller. Valves controller is essential hardware used to control all the water values in the vicinity. Valves controller is IoT-enabled hardware connected with the IoT cloud to send the on/off signals to the small water channel valves according to the algorithm or user command.

4.6. Valve. Each farm is connected with the water canal through small water channels. Each small channel has an IoT-enabled valve used to control the water flow in the field when irrigation is needed. Each valve is connected with the valves controller.

4.7. User Connectivity. The user connectivity module is another important module of the whole proposed system. The user connectivity module is responsible for sending the notifications to the users regarding the farms' events and taking instructions from the users in response. The users can send and receive the information and commands using emails or cell phone applications.

5. Working of EEWMP

In the proposed energy-efficient water management platform (EEWMP) model, we introduced an in-field sink node whose duty is to collect the data from the sensors deployed in the field and thus save their energy. Moreover, the fusion centre can also reduce the communication traffic and can effectively conserve energy. In this section, we discuss the working of the EEWMP with the threshold value of different employed sensors.

Initially, all sensors are deployed in the 100-meter area with 10 meters distance and 10 to 15 cm depth with plant roots. One in-field sink node is deployed for each field to gather data from the deployed field sensors and send it to the outer sink. The outer sink node is embedded with fusion centre, responsible for filtering out the redundant and empty data and sending the clean data to the IoT service cloud. The IoT service cloud provides a device registration/identification process and application to manage and process the collected data. This application is also responsible for sending notifications to control the channel valves to auto start or stop the water if the irrigation is needed based on the collected data.

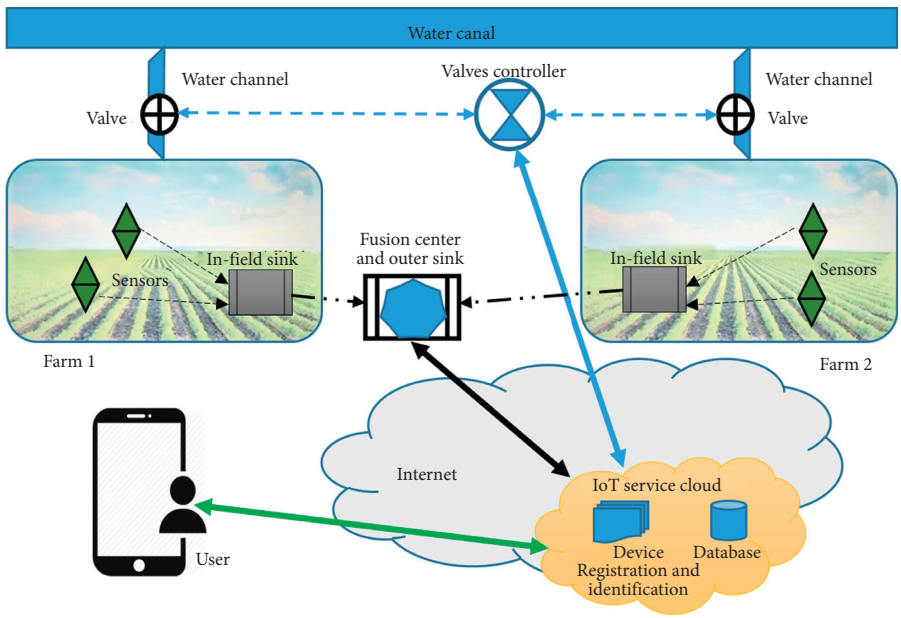


Figure 1: Architecture of proposed energy-efficient water management platform (EEWMP).

The threshold values of soil moisture for optimal field capacity are between 10% and 60%. When soil moisture is greater than 10%, the valves release water to the field until the moisture level reaches 60%. The sensors will keep on the soil moisture level to the IoT cloud for further processing (Algorithm 1).

In the case of temperature sensors, there are three threshold values for three different temperature ranges. If the temperature is less than 15°C, the water valve is programmed to release 3 mm to 4 mm water. If the temperature is between 15°C to 25°C, the water valve is programmed to release 5 mm to 6 mm water. Moreover, if the temperature is more than 25°C, the water valve is programmed to release 7 mm to 8 mm water (Algorithm 2).

The threshold of the water level sensor is based on two values: water level and tension test. Tension tests are used to check the soil moisture. If the water level is greater than or equal to 50% and the tension test value is greater than or equal to 20 cb (Centibar), the measure to expel additional water is taken. When the water level is less than or equal to 40% and the tension test value is less than or equal to 50 cb, the valves are instructed to release water in the field (Algorithm 3).

6. Methodology

In this research, the proposed EEWMP model's performance is evaluated and compared with the SWAMP model. Both the systems are implemented and simulated in the MATLAB 2019a tool. We considered 100 nodes scenario deployed in 400×400 meters field and sink nodes at 200×200 meters for experimentation. The details of the nodes energies for different tasks and points are given in Table 1. Both models' performance is evaluated using energy consumption, network stability period, packet sent to destination, and packet delivery ratio.

7. Results and Discussion

In this research, we proposed an IoT-based smart and energy-efficient irrigation system, EEWMP. The proposed EEWMP system's performance is compared with the SWAMP model using simulation conducted in a MATLAB environment. The models' performance is evaluated in energy consumption, network stability period, packet sent to destination, and packet delivery ratio. In this section, we discuss the results of the experiments.

With the introduction of sink nodes in the proposed EEWMP model, the network energy consumption is effectively reduced compared with the SWAMP model. Due to the efficient utilization of the nodes' energy, the proposed EEWMP model's network lifetime also increased compared with the SWAMP model. According to the results, after 250 seconds, both EEWMP and SWAMP networks utilized 8% energy. However, at 500 seconds, SWAMP utilized 35% energy whereas EEWMP utilized 30% energy, which is 5% less than SWAMP. Similarly, at 750 seconds, SWAMP utilized 69% energy, whereas EEWMP utilized 50% energy. At 1100 seconds, the SWAMP network is entirely exhausted, whereas EEWMP has more than 30% energy left at the same time. EEWMP network is completely exhausted at 2200 seconds, which is twice that of SWAMP. The energy consumption of the whole network in both models is illustrated in Table 2 and Figure 2.

Similarly, in network stability, the EEWMP model nodes are more stable than those in the SWAMP model. The results show that after 250 seconds, the number of alive nodes in the SWAMP model was 92, whereas the number of alive nodes in the EEWMP model was 93. Similarly, at 500 seconds, the number of alive nodes was 65 and 70 in SWAMP and EEWMP models. At 1000 seconds, the number of alive nodes in SWAMP was nine, but at the same time, the number of alive nodes in EEWMP was 36. Table 3 illustrates

```
Set time period P
Initialize:
  Monitor soil moisture SM (periodically P)
  If SM < 10%
    Open water valve for period P
    GOTO Initialize
  If SM > 60%
    Close water valve
    GOTO Initialize
```

Algorithm 1: Algorithm for soil moisture.

```
Set time period P
Initialize:
  Monitor temperature temp (periodically P)
  If Temp < 15°C
    Step 1: monitor water level WL ()
      If WL < 4 mm
        Open water valve for period P
        GOTO Step 1
      If WL > 4 mm
        Close water valve
        GOTO Initialize
  If Temp ≥ 15°C and Temp < 25°C
    Step 2: monitor water level WL ()
      If WL < 6 mm
        Open water valve for period P
        GOTO Step 2
      If WL > 6 mm
        Close water valve
        GOTO Initialize
  If Temp > 25°C
    Step 3: monitor water level WL ()
      If WL < 8 mm
        Open water valve for period P
        GOTO Step 3
      If WL > 8 mm
        Close water valve
        GOTO Initialize
```

Algorithm 2: Algorithm for soil temperature.

the number of alive nodes at different times in both models. The network stability period of both models is plotted in Figure 3.

In terms of the packet sent to the destination and packet delivery ratio, the EEWMP was more efficient and active than the SWAMP. According to the results (illustrated in Table 4), after 250 seconds, the total number of packets sent in the EEWMP scenario was 4900, whereas the total number of packets sent in the SWAMP scenario was 2500. The SWAMP model nodes were exhausted at 1100 seconds, and

```
Set time period P
Initialize:
  Monitor soil moisture SM (periodically P)
  Monitor tension test TT()
  If SM < 40% and TT > 50 cb (Centibar)
    Open water valve for period P
    GOTO Initialize
  If SM > 40% and TT > 20 cb
    Close water valve
    GOTO Initialize
```

Algorithm 3: Algorithm for water level and tension test.

Table 1: Simulations parameters.

Parameters	Values
Range of network	$400 \times 400\,m^2$
Location of sink	$200 \times 200\,m^2$
Number of nodes	100
Initial energy deployed to nodes	0.5 joules
Transmission energy per node	50 (nano jule) nJ/bit/m^2
Receiver energy per node	50 nJ/bit/m^2
Free space energy	10 pico-joule (pJ)
Amplification energy	0.0013 nJ/bit/m^2
Data aggregation energy	5 pJ/bit
Maximum no. of rounds	2500

Table 2: Energy consumption of the whole scenario in percentage.

Model	Time (sec)								
	250	500	750	1000	1250	1500	1750	2000	2250
EEWMP (energy consumption in %)	8%	30%	50%	64%	70%	79%	92%	97%	100%
SWAMP (energy consumption in %)	8%	35%	69%	91%	100%	100%	100%	100%	100%

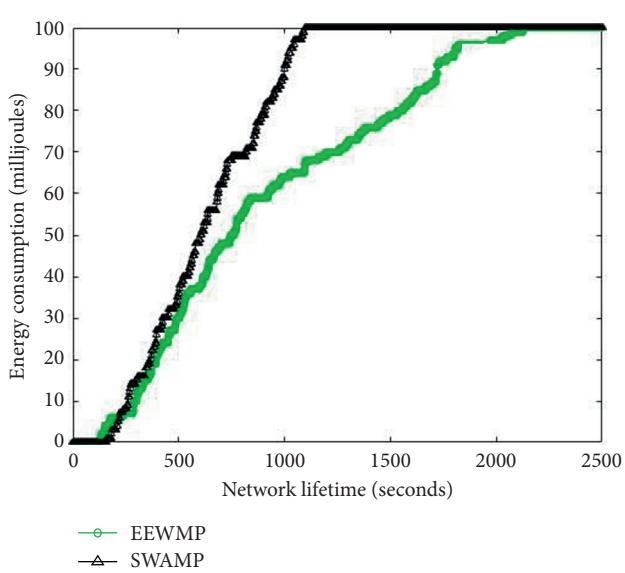


Figure 2: Energy consumption.

at the end of the simulation (2500 seconds), the EEWMP model nodes sent 11500 more packets than the SWAMP model. Figure 4 shows the packets sent to the destination in both models.

Similarly, the packet delivery ratio in the EEWMP model was found much better than the SWAMP model due to the in-field sinks and fusion centre. In the initial stages of the simulation, the packet delivery ratio of SWAMP was slightly better than EEWMP; however, after 750 seconds, the packet delivery ratio of the EEWMP became better. The packet delivery ratio of both the models is illustrated in Table 5 and plotted in Figure 5.

Overall, the performance of the EEWMP is found better compared with the network model of SWAMP in terms of energy consumption, network stability period, packet sent to destination, and packet delivery ratio. The EEWMP model nodes consume a small amount of energy due to sink nodes and fusion centre, helping them survive more. Similarly, the extended lifetime of the node also increases the packet creation and delivery ratio. Furthermore, the network's life

Table 3: Stability period.

Model	Time (sec)								
	250	500	750	1000	1250	1500	1750	2000	2250
EEWMP (alive nodes)	93	70	50	36	30	21	8	3	0
SWAMP (alive nodes)	92	65	31	9	0	0	0	0	0

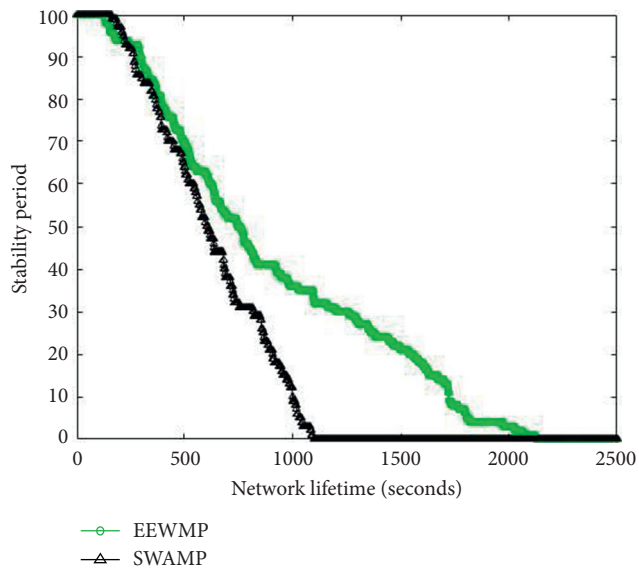


Figure 3: Stability period.

Table 4: Packets sent to destinations.

Model	Time (sec)								
	250	500	750	1000	1250	1500	1750	2000	2250
EEWMP	4900	9000	11900	14000	15600	17000	17900	18500	18500
SWAMP	2500	4400	5700	6700	7000	7000	7000	7000	7000

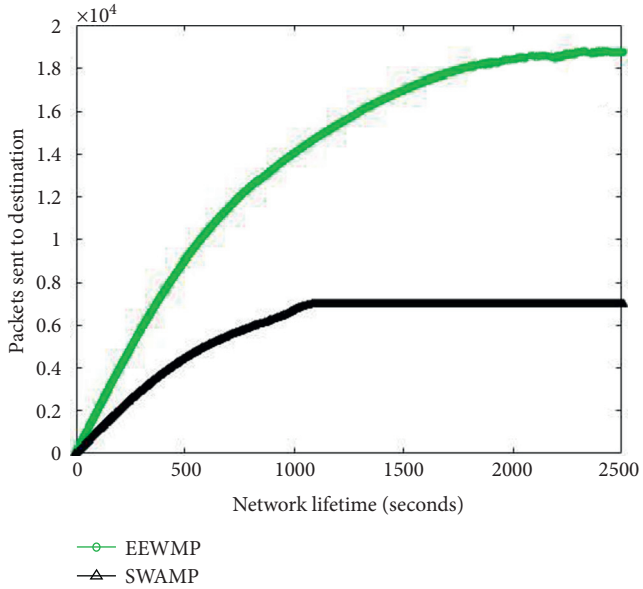


Figure 4: Packets sent to destinations.

Table 5: Packets delivery ratio.

Model	Time (sec)								
	250	500	750	1000	1250	1500	1750	2000	2250
EEWMP	1.95	3.59	4.77	5.60	6.26	6.76	7.06	7.12	7.12
SWAMP	2.22	3.97	5.03	5.51	5.53	5.53	5.53	5.53	5.53

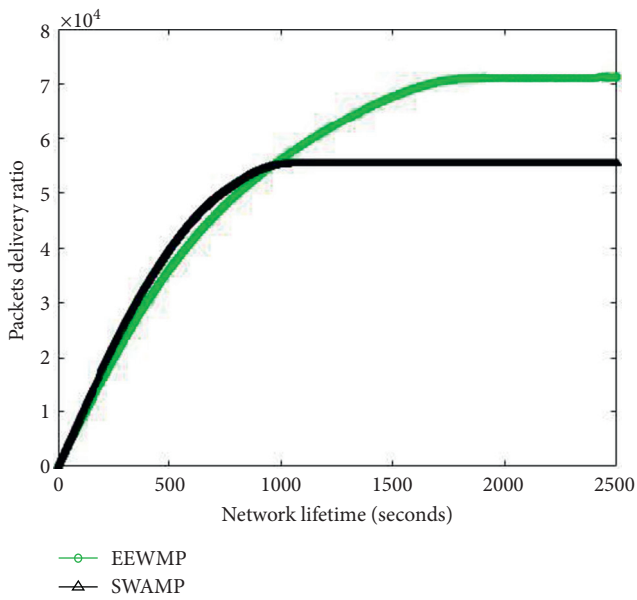


Figure 5: Packets delivery ratio.

can be further improved to find the optimized data sending frequency for nodes.

8. Conclusions and Future Work

The smart irrigation system is one of the essential needs for precision agriculture in the current time as we cannot afford fresh water wastage. IoT-based smart irrigation systems use field deployed sensors to detect soil characteristics, weather and climate conditions, and crop conditions for irrigation. SWAMP is a collaborative project developed for smart irrigation and efficient freshwater utilization in agriculture. The primary aim of SWAMP is to auto manage water reserves, distribution, and consumption of various levels, avoid over-irrigation and under-irrigation problems, and auto manage time to maximize production. This research improved the SWAMP network's performance by introducing an in-field sink and fusion centre use of open-source cloud to reduce costs. We called our improved model as energy-efficient water management platform (EEWMP). In the proposed EEWMP, the in-field sink node collects the data from the field's sensors and sends it to the fusion centre. The fusion centre aggregates the data and removes redundant information, thus reducing communication traffic and energy consumption. The results show that the performance of the EEWMP is found better compared with the network model of SWAMP in terms of energy consumption, network stability period, packet sent to destination, and packet delivery ratio. It was found that EEWMP consumes 30% less energy and increases network stability time twice as compared with SWAMP. Similarly, due to increased network stability time, the destination's packets were 1.5 times more in EEWMP than SWAMP.

Based on the simulation results, it is safe to conclude that the introduction of sink nodes and fusion centre enhances the network performance in terms of data generation and energy consumption. EEWMP can be used in different irrigation models such as drip irrigation, sprinkler irrigation, surface irrigation, and lateral move irrigation with subtle alterations. Moreover, it can also be used in small farms of third-world countries with their existing communication infrastructures such as 2G or 3G.

Precision agriculture is now essential in today's world, especially for countries with limited water resources, fertile land, and enormous population. Smart irrigation systems can help countries efficiently utilize fresh water and use the excess water for barren lands. In the future, we are interested in developing smart irrigation models for other irrigation systems, drip and sprinkler. We are also interested in utilizing other sensors to make smart irrigation models and algorithms for different soil types such as gravel, silt, loam, sand, and barren land.

Data Availability

It is a simulation-based research, and no data were used.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

References

[1] C. Kamienski, J.-P. Soininen, M. Taumberger et al., "Smart water management platform: iot-based precision irrigation for agriculture," *Sensors*, vol. 19, no. 2, p. 276, 2019.

[2] F. Aquastat, "Water uses," www. fao. org/nr/water/aquastat/water_use/index. stm, 2016.

[3] M. Rodell, J. S. Famiglietti, D. N. Wiese et al., "Emerging trends in global freshwater availability," *Nature*, vol. 557, no. 7707, pp. 651–659, 2018.

[4] R. G. Perea, A. Daccache, J. R. Díaz, E. C. Poyato, and J. W. Knox, "Modelling impacts of precision irrigation on crop yield and in-field water management," *Precision Agriculture*, vol. 19, pp. 497–512, 2018.

[5] P. P. Ray, "A survey on Internet of Things architectures," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 3, pp. 291–319, 2018.

[6] E. G. Popkova, E. N. Egorova, E. Popova, and U. A. Pozdnyakova, "The model of state management of economy on the basis of the internet of things," in *Ubiquitous Computing and the Internet of Things: Prerequisites for the*

Development of ICT, pp. 1137–1144, Springer, New York, NY, USA, 2019.

[7] P. Asghari, A. M. Rahmani, and H. H. S. Javadi, "Internet of Things applications: a systematic review," *Computer Networks*, vol. 148, pp. 241–261, 2019.

[8] R. Finger, S. M. Swinton, N. El Benni, and A. Walter, *Precision Farming at the Nexus of Agricultural Production and the Environment*, Annual Review of Resource Economics, vol. 11, no. 1, pp. 313–335, 2019.

[9] D. Popescu, L. Ichim, F. Stoican, and C. Dragana, "Hierarchical processing of signals for smart crop monitoring," in *Proceedings of the 2019 8th International Conference on Systems and Control (ICSC)*, pp. 265–270, Marrakech, Morocco, 2019.

[10] M. A. Ali, L. Dong, J. Dhau, A. Khosla, and A. Kaushik, "Perspective—electrochemical sensors for soil quality assessment," *Journal of The Electrochemical Society*, vol. 167, Article ID 037550, 2020.

[11] L. García, L. Parra, J. M. Jimenez, J. Lloret, and P. Lorenz, "IoT-based smart irrigation systems: an overview on the recent trends on sensors and IoT systems for irrigation in precision agriculture," *Sensors*, vol. 20, no. 4, p. 1042, 2020.

[12] S. Chaudhry and S. Garg, "Smart irrigation techniques for water resource management," in *Smart Farming Technologies for Sustainable Agricultural Development*, pp. 196–219, IGI Global, Philadelphia, PA, USA, 2019.

[13] A. Srilakshmi, J. Rakkini, K. Sekar, and R. Manikandan, "A comparative study on Internet of Things (IoT) and its applications in smart agriculture," *Pharmacognosy Journal*, vol. 10, 2018.

[14] C. Kamienski, J.-P. Soininen, and M. Taumberger, "Swamp: an IoT-based smart water management platform for precision irrigation in agriculture," in *Global Internet of Things Summit*, IEEE, New York, NY, USA, 2018.

[15] A. Fiorentino, J. Zangari, and M. Manna, "DaRLing: a Datalog rewriter for OWL 2 RL ontological reasoning under SPARQL queries," *Theory and Practice of Logic Programming*, vol. 20, no. 6, pp. 958–973, 2020.

[16] V. Araujo, K. Mitra, S. Saguna, and C. Åhlund, "Performance evaluation of FIWARE: a cloud-based IoT platform for smart cities," *Journal of Parallel and Distributed Computing*, vol. 132, pp. 250–261, 2019.

[17] M. A. Rodriguez, L. Cuenca, and A. Ortiz, "FIWARE open source standard platform in smart farming-a review," in *Proceedings of the Working Conference on Virtual Enterprises*, pp. 581–589, Cardiff, UK, 2018.

[18] M. Wylot, M. Hauswirth, P. Cudré-Mauroux, and S. Sakr, "RDF data storage and query processing schemes," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–36, 2018.

[19] F. Viola, F. Antoniazzi, C. Aguzzi, C. Kamienski, and L. Roffia, "Mapping the NGSI-LD context model on top of a SPARQL event processing architecture: implementation guidelines," in *Proceedings of the 2019 24th Conference of Open Innovations Association (FRUCT)*, pp. 493–501, Moscow, Russia, 2019.

[20] H. Channe, S. Kothari, and D. Kadam, "Multidisciplinary model for smart agriculture using internet-of-things (IoT), sensors, cloud-computing, mobile-computing & big-data analysis," *International Journal Computer Technology & Applications*, vol. 6, pp. 374–382, 2015.

[21] L. Doron, "Flexible and precise irrigation platform to improve farm scale water productivity," *Impact*, vol. 2017, no. 1, pp. 77–79, 2017.

[22] T. Popović, N. Latinović, A. Pešić, Ž. Zečević, B. Krstajić, and S. Djukanović, "Architecting an IoT-enabled platform for precision agriculture and ecological monitoring: a case study," *Computers and Electronics in Agriculture*, vol. 140, pp. 255–265, 2017.

[23] A. Kamilaris, F. Gao, F. X. Prenafeta-Boldu, and M. I. Ali, "Agri-IoT: a semantic framework for Internet of Things-enabled smart farming applications," in *Proceedings of the 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, pp. 442–447, Reston, VA, USA, 2016.

[24] S. Jaiganesh, K. Gunaseelan, and V. Ellappan, "IOT agriculture to improve food and farming technology," in *Proceedings of the 2017 Conference on Emerging Devices and Smart Systems (ICEDSS)*, pp. 260–266, Piscataway, NJ, USA, 2017.

[25] Z. Li, J. Wang, R. Higgs, L. Zhou, and W. Yuan, "Design of an intelligent management system for agricultural greenhouses based on the internet of things," in *Proceedings of the 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, pp. 154–160, Piscataway, NJ, USA, 2017.

[26] F. Kiani and A. Seyyedabbasi, "Wireless sensor network and internet of things in precision agriculture," *International Journal of Advanced Computer Science and Applications (Ijacsa)*, vol. 9, 2018.

[27] E. Nurellari and S. Srivastava, "A practical implementation of an agriculture field monitoring using wireless sensor networks and IoT enabled," in *Proceedings of the 2018 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS)*, pp. 134–139, Hyderabad, India, 2018.

[28] J. Karpagam, I. I. Merlin, P. Bavithra, and J. Kousalya, "Smart irrigation system using IoT," in *Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 1292–1295, Coimbatore, India, 2020.

[29] S. Gupta, V. Malhotra, and V. Vashisht, "Water irrigation and flood prevention using IOT," in *Proceedings of the 2020 10th International Conference on Cloud Computing*, pp. 260–265, Data Science & Engineering, Coimbatore, India, 2020.

[30] M. A. Rodriguez, L. Cuenca, and A. Ortiz, "FIWARE open source standard platform in smart farming-a review," in *Working Conference on Virtual Enterprises*, pp. 581–589, Springer, Cham, Switzerland, 2018.

[31] P. Pierleoni, R. Concetti, A. Belli, and L. Palma, "Amazon, google and microsoft solutions for iot: architectures and a performance comparison," *IEEE Access*, vol. 8, pp. 5455–5470, 2019.

Hindawi
Scientific Programming
Volume 2021, Article ID 5544071, 9 pages
https://doi.org/10.1155/2021/5544071

Research Article

Technical and Tactical Command Decision Algorithm of Football Matches Based on Big Data and Neural Network

Lei Fang,[1] Qiang Wei,[2] and Cheng Jian Xu [3]

[1]*Northwest University for Nationalities, Lanzhou 730124, Gansu, China*
[2]*Department of Physical Education, Tangshan Normal University, Tangshan, China*
[3]*Department of Competitive Sports, Guangdong Sports Vocational and Technical College, Guangzhou 510663, Guangdong, China*

Correspondence should be addressed to Cheng Jian Xu; haohaoxuexi86668@163.com

Received 21 January 2021; Revised 2 March 2021; Accepted 20 March 2021; Published 8 April 2021

Academic Editor: Shah Nazir

A successful football team not only consists of more than a dozen people on the field but also includes a complete training, analysis, coaching team behind it, and the same basic education and youth training system. With the development of scientific concepts and the advancement of computer technology, more people have begun to study the use of modern technology to replace part of the traditional human work with low creativity and the use of more convenient quantitative analysis, prediction and other technologies to assist football professionals' decision-making. Based on big data and neural network technology, this paper has designed a novel football technical and tactical command decision algorithm. First, the use of big data technology for analyzing the characteristics of the historical big data of football competitions provides valuable data for the work of this article. Secondly, to formulate scientific and reasonable football technical and tactical command, it requires learning effective offensive or defensive strategies from the big data of football competitions. This article uses deep neural networks to learn massive amounts of football competition data, which can effectively predict the offensive and defensive tactics of each position of the team to a certain extent. In addition, in order to better learn the timing video data of football matches, this paper also has proposed to use long- and short-term memory networks to improve the algorithm of this paper. The proposed method has achieved good results in football technical and tactical command and decision-making and also provides some new ideas for the subject of football combined with computer technology.

1. Introduction

Football match is a popular sport activity around the world. The prosperous development of sports business and the application of Internet technology allow fans to easily obtain high-quality videos of almost all professional football matches. Data providers can easily collect and archive these video data as well as game and player statistics. However, the resulting big data has brought both challenges and opportunities for sports professionals and related companies. In order to make the most of these data, football scouts and coaches need advanced analytical tools to evaluate, select, and train players. Researchers have been developing tools for scientific analysis to help evaluate the performance of players and teams [1].

Although the relevant research is quite popular and has made certain achievements, its scientificity and effectiveness still need to be improved [2–6]. Existing research is usually difficult for professionals and amateurs to examine. It is not easy to conduct professional and reliable quantitative analysis of competition data. The football game is a complex whole, affected by many factors. On the other hand, statistics are sparse and lack relevance. It is not wise to use only shooting or passing statistics to describe a game. Most existing systems only use statistics instead of considering the whole, which may lead to a misunderstanding of the game. This is why many coaches still believe in their eyes rather than statistics. Football professionals and computer and data researchers often lack effective communication, which makes the development of related tools inseparable from the

long-term close cooperation between football professionals and computer and data researchers. Finally, big data and neural network [7–12] technologies still need indepth research in the football field. As a football game with many players on the field, rich tactics, and a large range of activities, its flexibility and complexity make many existing methods unable to directly hit the core of football. Researchers need to develop relevant neural network algorithms specifically for football sports, taking full account of the particularity of football sports and filling the gaps in the industry.

Therefore, this paper designs a novel football technical and tactical command decision algorithm based on big data and neural network technology [13–17]. First of all, the use of big data technology to analyze the characteristics of the historical big data of football competitions provides a lot of valuable data for the work of this article. Secondly, to formulate scientific and reasonable football technical and tactical command, it requires learning effective offensive or defensive strategies from the big data of football competitions. This article uses deep neural networks to learn massive amounts of football competition data, which can effectively predict the offensive and defensive tactics of each position of the team to a certain extent. In addition, in order to better learn the timing video data of football matches, this paper also proposes to use long- and short-term memory networks to improve the algorithm of this paper. Following are the main innovative points of this paper:

(1) This paper uses big data technology to analyze the characteristics of massive historical data of football matches, extracts a lot of valuable data, and designs some new interactive methods, quantitative methods, and evaluation methods for football matches

(2) We have innovatively introduced a long- and short-term memory network (LSTM) to learn the time series data of football matches and built a deep neural network model to assist football technical and tactical command decisions

(3) We have conducted sufficient comparative experiments and ablation studies to prove the effectiveness of the football technique and tactics command decision algorithm based on big data and neural network proposed in this paper, which can provide a scientific basis for sports coaches to formulate reasonable football techniques and tactics

The organization of the paper is as follows: Section 2 shows the related work to the proposed study. Section 3 represents the methodology section of the paper. Section 4 depicts the experiment and results of the paper. The paper is concluded in Section 5.

2. Related Work

Various approaches and techniques are existing in literature associated with diverse decisions of football matches. Krzysztof and Pawet [18] conducted a multidimensional study on the characteristics of goals in the game. The author believes that insufficient sample size will make the research results inaccurate. In the 08/09 season, the Premier League's close-range shots accounted for more than 50%, 53% of the goals were below the goalkeeper's stock, and the second half of the goal accounted for 57%. The author believes that the inattention of players due to fatigue is a goal loss. Key factor: the goal rate after dribbling was only 12.8%. The goalkeeper failed to save a low shot (53% of goals). In general, the author believes that the lower area of the upper goal is the most effective area for scoring. Michalls et al. [19] through the way of creating goals in the Premier League, La Liga, Bundesliga, and Serie A, the purpose of the study is to compare the creation of goals in the four major European football leagues. The sample includes 80 random matches in the 2017-2018 season (20 La Liga; 20 Premier League; 20 Bundesliga; 20 Italian Serie A). 914 teams created scoring opportunities in the game. The study uses a multidimensional observation method to evaluate multiple tactical dimensions and the three stages of the team's beginning, development, and end. Kruskal–Wallis analysis showed that there were significant tactical differences in these 4 games. La Liga shows a greater proportion of long-distance offenses and teamwork. The Premier League appears to be quick and direct in attacking methods. Compared with cooperating offense and quick offense, Bundesliga has the most counterattacks, Serie A has the shortest offensive distance, and the proportion of counterattacks and direct attacks is also larger. Kubayi [20] studied the goal pattern of the 2018 World Cup. All goals in the game are analyzed with the InStat video analysis system. The results showed that a total of 169 goals were scored in this game (sports battle: 60.9%; setup: 39.1%). Eighty-five goals (82.5%) came from the team's offense, and 18 goals (17.5%) came from the counterattack. Chi-square test shows that there is a significant difference in the type of possession (card 2 (1, $n = 103) = 43.58$, $p \leq 0.001$). Compared with the first third (33%) and the middle third (32%), the last third (35%) has the most goals. The results also showed that most goals came from short passes (69.9%), 13.6% came from long passes, and 16.5% came from mixed passes (chi 2(2, $n = 103) = 62.12$, $p \leq 0.001$).

Perl et al. [21] believes that "different technologies have different data collection methods. The technical collection methods can be divided into two types of technologies: video capture and wearable transmission." Video capture uses panoramic photography and game video recording. The latter places chips in clothes and sneakers to analyze sports performance. Reilly and Williams [22] found that the Cologne University of Sport used a computer system to process video matches and found that, before 80% of a goal, the number of passes is no more than 3 times. Fast attack and teamwork have always been the direction of modern football development. Pollard and Reep [23] found that the analysis of football performance is divided into two types: predata collection and postanalysis. Compared with physical load, technical and tactical analysis can determine the outcome of a match.

3. Methodology

Football technique is the general term for reasonable actions and movements used in football matches. According to the analysis characteristics of large-scale competitions in the physical education field and the realistic possibilities of video statistics, this research divides football skills into two categories: offensive technology and defensive technology. Offensive skills include passing, ball possession, dribbling breakthrough, and shooting. Defensive skills include steals, clearances, and tackles. Among these concepts, the concepts of passing, possession, and shooting are very clear. Figures 1(a)–1(d) show an example of dribble breakthrough attack.

The algorithm proposed in this paper is shown in Figure 2. We define the other concepts in detail as follows:

(1) Dribble breakthrough: the player uses dribbling or passing to pass one or more opponents' pressing and blocking forward

(2) Steal: when the attacking player holds the ball or the attacking player passes the ball to his partner, the defensive player uses defensive skills to get the ball before the opposing player

(3) Rescue: in order to prevent the attacking side from attacking, the defensive player directly destroys the ball out of the defensive area without passing the ball.

3.1. Stadium Modeling. The purpose of stadium modeling is to convert the position of the player from the video perspective to the position of the top view (aerial view). One of the purposes of this is that redundant video information is often not needed in football visualization analysis, but top view and representative players are used. The dynamic icon is replaced. The second reason is that, in order to obtain information about the physical fitness of the player in the future, the player's position information needs to be obtained first. Due to the reason that the distance is small, the position information obtained by simply tracking the original video cannot be used for calculation. Therefore, perspective transformation is required for football field modeling. In this study, both the video provided by the Shanghai Institute of Physical Education and the video of the National Games adopted the FIFA standard show that the vertical length of the field is 105 meters, the horizontal width is 68 meters; the goal: the length is 7.32 meters, the height is 2.44 meters; the big penalty area (penalty area): horizontal 40.32 meters, vertical 16.5 meters, on the bottom line 5 meters away from the goal post; small restricted area (goal area): 18.32 meters long, 5.5 meters wide, 5.5 meters from the goal post on the bottom line; center circle area: a radius of 9.15 meters; 4 corner kick area: radius of 1 meter, 13.84 meters from the penalty area; penalty kick arc: a semicircle with a radius of 9.15 meters centered on the penalty kick point; penalty kick point: 11 meters from the goal line. For the convenience of debugging in the development process and the applicability of the subsequent software release, our video size and court plane model diagram are both $1920 * 1080$ pixels in size.

After the target detection and tracking part of the video, we successfully obtained the pixel position (x, y) of each player. Before we can use this location, there is one more step that needs further research. When the human eye sees things nearby, they appear larger than those far away. This is often referred to as a perspective phenomenon. The perspective transformation refers to the transfer of an object from one state to another. Popular application examples such as in the picture, there is a trapezoidal test paper viewed from an oblique direction. Its actual shape is undoubtedly a rectangle. Through perspective transformation, we select several points on the test paper and then select the corresponding one on the plan view. To demonstrate, the rectangular test paper in the illustration can be turned into a rectangle when viewed from the front. At this time, if you want to recognize the words on the test paper, the effect will be better. At the same time, the points on the plane where the test paper is not are used because of the number. The correct perspective transformation produces distortion. In general, the perspective conversion process converts a 3D world to a 2D image, but we need to determine a plane that is the only concern. Points outside the plane will be distorted. In this article, we need to use such a transformation matrix to adjust the pixel position as follows. Through perspective transformation, we obtain a top view. In the top view, any two points with the same distance in the real world have the same Euclidean distance between their pixels. We can superimpose the distance between two points between adjacent frames to calculate the running distance of the player. The essence of perspective transformation is to change the projection of the image to a new perspective plane. The general conversion formula can be written as, where (u, v) are the pixel coordinates of the original image, $(x = (x'/w'), y = (y'/w'))$. It is the pixel coordinate transformation after the image. The perspective transformation matrix is shown as follows:

$$[x', y', w'] = [u, v, w]\begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}. \quad (1)$$

We choose four pixels in the original image, which are usually easily distinguishable signs such as the corner, intersection, or inflection point of the court. Then, we find their corresponding points in the model image. We use these two sets of points to calculate the transformation matrix. After the transformation matrix is obtained, we select the position of the player's foot that is on the same plane as the court for transformation to obtain the position of the top view.

In the game, football can easily be in the air. If we use the same method as calculating the player's position, the height of the football will cause a big error: the position obtained by the perspective transformation is the position of the football projected on the ground plane in the video. How do we find the exact position of football? First of all, we need to calibrate the camera parameters of the camera used to obtain the correspondence between camera coordinates and world coordinates. After having at least two different views and calibrating the camera, we can calculate the 3D position of

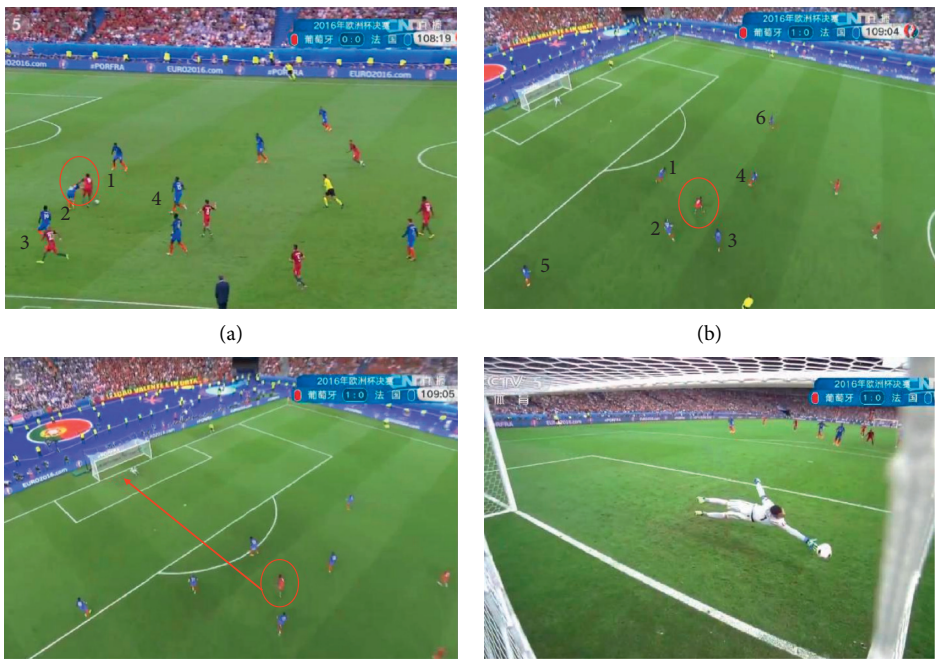


Figure 1: Example of dribble breakthrough attack. (a) Striker takes the ball. (b) Striker gets rid of. (c) Forward shot. (d) Goal.

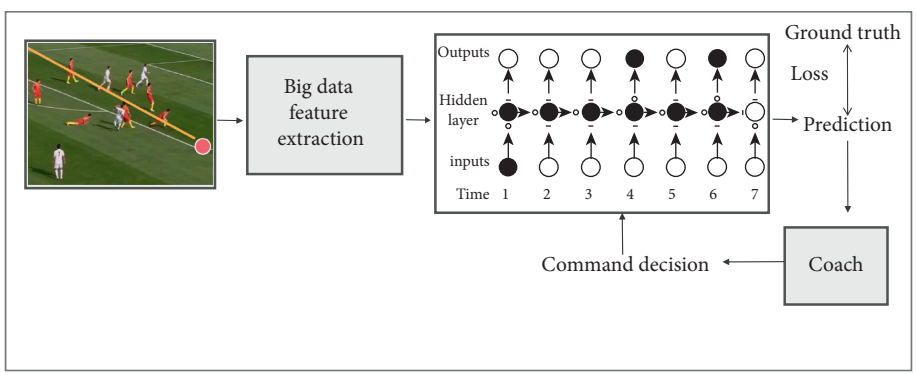


Figure 2: The flowchart of the overall architecture of our algorithm.

the ball. This is a simple application of 3D reconstruction. Suppose the straight line L can be expressed by the following formula:

$$L = t + w_3{}^* d_3, \tag{2}$$

where t is the position of the camera, w_3 is a constant, and direction 3 represents the direction of the line. We have two cameras, which form two nonparallel lines in the coordinate system:

$$\begin{cases} L_1 = t_1 + w_1{}^* d_1, \\ L_2 = t_2 + w_2{}^* d_2. \end{cases} \tag{3}$$

Calculate the intersection point:

$$\begin{cases} d^* d_3 = 0, \\ d_2{}^* d_3 = 0. \end{cases} \tag{4}$$

After the calibration operation is completed, with the help of perspective transformation, we can have two lines, which can be understood as a light taken by the football by the camera. Due to the influence of the height of the football mentioned above, the actual position of the football is on this line. At some point, ideally, the 3D position (x, y, z) of a football can be calculated from the intersection of two lines. In practical applications, it is impossible because the two lines intersect. We use the midpoint of the shortest distance line of the two straight lines to simulate the intersection of the two straight lines to obtain the football position.

3.2. Recurrent Neural Network. Deep neural networks [24] refer to neural networks with deep network structure, which have made outstanding achievements in match prediction [25–29] and computer vision [30–34]. Deep neural networks have more neural network layers, so they can extract more

features from the input and have a stronger ability to portray reality. According to the core network used, deep neural networks can be divided into deep convolutional neural networks (deep CNN) and deep recurrent neural networks (deep RNN). In the field of computer vision and image and video processing, convolutional neural networks have greater advantages than other neural networks. In the field of sports competition data analysis, because the input data is dependent and sequential mode, there is no correlation between the front and back inputs of CNN, and all outputs are independent of each other, so the performance of CNN is not good. For the task of automatic generation of Chinese couplets studied in this article, all outputs are related to the previous output, and some biases based on the previous output information are required. Therefore, the recurrent neural network RNN is more suitable.

3.2.1. Standard Recurrent Neural Network. The basic neural network includes a three-layer structure of input layer, hidden layer, and output layer. It only establishes connections between layers, while the standard recurrent neural network RNN (recurrent neural network) is based on this, in the same layer. Connections are also established between neurons. The neural network structure of the RNN is shown in Figure 3. The right side of the equal sign is the time-expanded diagram of the neural network, and the left side of the equal sign is its simplified diagram. Assuming that x_t is the input of time step t in the sequence and h_t is the hidden state vector of that time step. According to the neural network structure in Figure 3, the current h_t is

$$h_t = F\left(x_t W_{xh} + h_{t-1} W_{hh} + b_h\right), \tag{5}$$

where h_{t-1} is the hidden state vector of the previous time step. In simple terms, the hidden state vector of time step t is determined by the input x_t of the current time step and the hidden state vector h_{t-1} of the previous time step. F is the activation function of the neural network, and w_{xh} and w_{hh} are the weight matrices of the neural network.

3.2.2. Long- and Short-Term Memory Network. The long- and short-term memory network (As shown in Figure 4) is more suitable for solving the time series problem with long-term dependence, which is to solve the shortcomings of the general RNN network. As shown in Figure 5, compared with ordinary RNN, LSTM adds a memory unit and three controllers: input control, forgetting control, and output control. The function of the memory unit is to store the network state; the input controller decides how much input is kept at the current moment; the forgetting controller decides the degree of retention of the neural network state at the current moment; the output controller determines the output information according to the neural network state at the current moment.

The core of LSTM is the memory cell state, which is composed of a sigmoid network and a multiplier. Using this structure, the information in the memory cell can be added or deleted. The output range of the sigmoid network is [0, 1], which indicates how much information can pass through the memory unit, 0 means none can pass, and 1 means all can pass.

Forgetting gate: use a sigmoid layer to receive the output of the previous time node t-1 and the input of the current time node t, merge it into a tensor, and then apply a linear transformation afterwards. After the sigmoid activation function, the output of the forgetting gate is a value between 0 and 1. This value will be multiplied by the internal state, which is why it is called the forget gate. If $f_t = 0$, the previous internal state is completely forgotten, and if $f_t = 1$, it will pass without any change. The calculation equation of the forget gate is as follows:

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right). \tag{6}$$

Input gate: accept previous output and new input and pass it to another sigmoid layer. The input gate returns a value between 0 and 1. Then, the value returned by the input gate is multiplied with the output of the candidate layer.

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right). \tag{7}$$

This layer mixes the input with the output of the previous layer, then applies hyperbolic tangent activation, and returns a candidate vector to add to the internal state. The internal status update rules are as follows:

$$\begin{aligned} C_t &= \tanh\left(W_C\left[h_{t-1}, x_t\right] + b_C\right), \\ i_t &= f_t{}^* C_{t-1} + i_t{}^* C_t. \end{aligned} \tag{8}$$

The output gate controls how many internal states are passed to the output, and it works like other gate structures. The output gate calculation equation is as follows:

$$\begin{aligned} O_t &= \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right), \\ h_t &= O_t{}^* \tanh C_t. \end{aligned} \tag{9}$$

The above is the structure principle of LSTM.

4. Experiments and Results

The following sections show the experiments and results of the paper.

4.1. Experimental Environment. Since the experiment in this article needs to train a deep neural network, the scale is large, the structure is more complex, and the calculation scale is large. The neural network training process needs to use GPU to accelerate the calculation. The experimental environment configuration is shown in Table 1.

The experiments in this article are all done on this machine, where the neural network training is accelerated by GPU, the programming language used is Python, the version is 3.6, the deep learning framework used is Keras2.1.5, and the IDE for program deployment is PyCharm.

4.2. Hyperparameter Settings. Use LSTM as the neural network structure of the encoder and decoder and share parameters between the SoftMax layer and the word vector

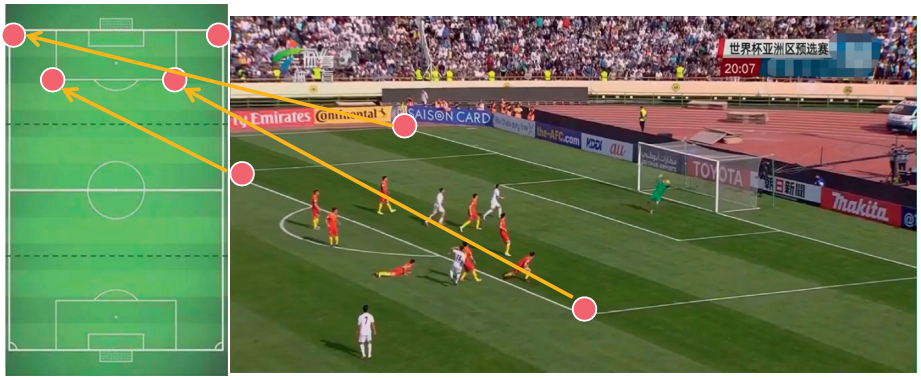


Figure 3: Goal retrieval.

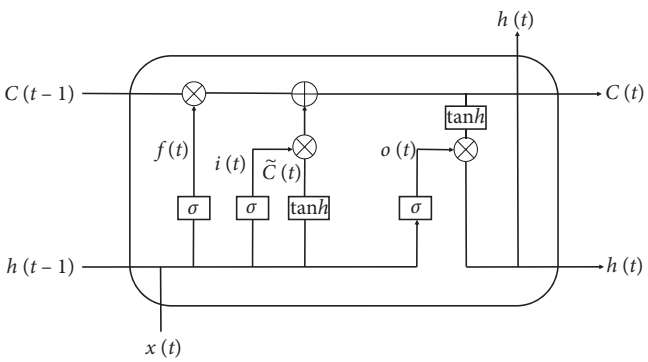


Figure 4: Basic structure of long- and short-term memory network (LSTM).

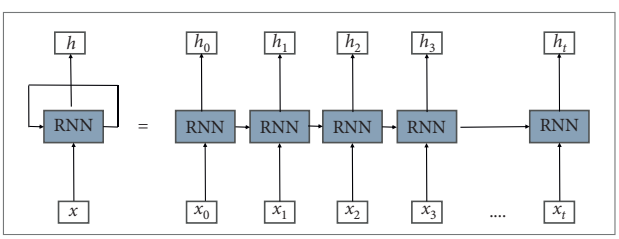


Figure 5: Basic structure of standard recurrent neural network.

layer to reduce the number of parameters [[45][46]}. The size of the LSTM hidden layer is 1024, the encoding and decoding neural network, the number of layers in the LSTM structure is 6 layers, the batch size is 128, the probability of a node being dropped out is set to 0.1, and the model is saved every 2000 steps in training.

4.3. Experimental Results of Different Methods. In this section, experimental results are presented to verify the proposed system. This paper presents experimental results to verify the proposed system. In order to ensure the effectiveness of the method, the detection experiment was applied to the football video data of the 13th China National Games. In order to prove the effectiveness of the width learning system, we compare our method with existing mainstream methods, including SVM, SRC, LRC, LCCR, and RDBLS.

Table 1: Experimental environment configuration.

Operating system	Windows
CPU	Intel (R) Core (TM)i7-8700 CPU @ 3.20 GHz
GPU	NVIDIA GeForce GTX 1080
RAM	16 GB
Deep learning library	Keras2.1.5

Table 2 gives a quantitative comparison of different methods. In Table 2, we find that our method can achieve less training time, while the accuracy remains at a level similar to other methods. The results of the number of different mapped features are given in Table 3. In Table 2, we tested three features: low-frequency Fourier transform feature (FFT), Gabor amplitude, and local binary pattern (LBP) instead of using the original pattern unaligned image. These methods are all suitable for identification with the information of the previous sequence. We tested the success rate of the tracker when it lost the target. In short, the width learning method showed acceptable accuracy and excellent running time.

Passing is one of the most basic techniques in football. Players need to pass the ball to connect in series. Passing is also the basis of collective cooperation in football. To complete tactical coordination and even create shooting opportunities, a team needs to pass the ball. Since the European Cup in 2008, the World Cup in 2010, and the European Cup in 2012, after the Spanish team won successively, a passing style has been set off in the world football. The world's strong teams have paid more and more attention to the passing ability of the players. Teams generally have a pass success rate of more than 90%.

The statistics of the number of passes, the number of long passes and the success rate of passes of the Portuguese team in each game of this European Cup are made. The specific results are shown in Table 2. The overall pass success rate of the Portuguese team is not high, with an average success rate of only 80.86%. In the last game of the group match, the Portuguese team had a traditional success rate of only 64% against the weak Hungary team. The passing success rate is not high but the game can be won, which has become a common phenomenon in recent world high-level football

Table 2: Statistics of passing characteristics.

Competitor	Total number of passes	Long passes	Pass success rate (%)
Iceland	421	62	88.8
Austria	403	53	83.9
Hungary	437	67	64.0
Croatia	489	100	80.9
Poland	561	83	82.3
Wales	587	58	82.7
France	506	98	83.4

Table 3: Predict the success rate of retrieval.

Methods	FFT		Gabor		LBP		Mean
	Probability (%)	Running time	Probability (%)	Running time	Probability (%)	Running time	
SVM	15.8	22.6	49.8	16.4	27.7	20.36	31.1
SRC	37.4	3682.9	65.7	3587.2	58.6	3608.2	53.9
LRC	20.3	31.72	26.4	30.02	28.4	33.79	25.0
LCCR	27.8	20.09	58.6	21.65	64.2	21.23	50.2
RDBLS	45.7	0.58	70.2	0.73	69.3	0.67	61.7
Ours	**46.9**	**0.55**	**72.2**	**0.68**	**70.1**	**0.66**	**62.9**

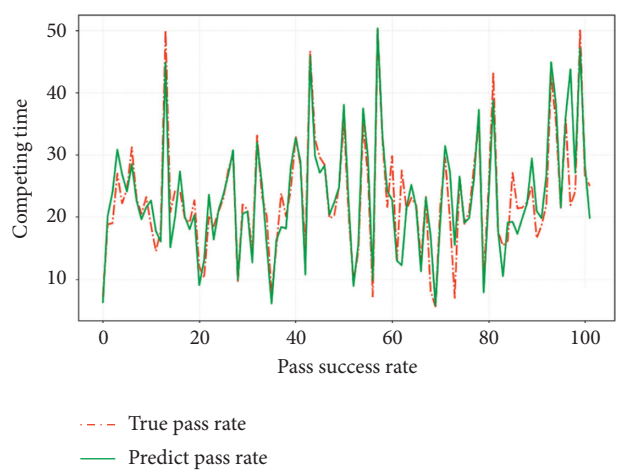


Figure 6: The comparison result of the Portuguese team's passing rate prediction using our algorithm.

matches. The reason is that the teams are not blindly pursuing a safer backcourt in the offense. It is daring to pass the threatening ball in the offense. In this case, although the turnover rate is high, the threat is not reduced at all.

4.4. Visualization of Results. Judging from the research results in Figure 6, the Portuguese team used more dribble breakthrough techniques in the game, and the success rate of dribble breakthroughs was also relatively high.

It is the excellent offensive players such as Ronaldo and Nani who can often create opportunities through sharp personal dribble breakthroughs. In fact, modern high-level football games require more comprehensive and three-dimensional player skills. A team needs to have players with strong passing and control skills, as well as players with personal dribbling and breakthrough capabilities. This type of staffing will enrich the team's tactical play, making it more difficult for opponents to pin them.

4.5. Technical and Tactical Decision Based on Neural Network. Section 4.4 proved our algorithm to accurately predict the pass success rate of a single team. For opponents, the prediction of the time and fulcrum of the tackle (Figure 7 shows an example of tackle) will become extremely important.

This article uses neural network to make model prediction. Figure 8 below shows the predicted position and actual position. From Figure 8, we can intuitively see that the football technical and tactical command decision algorithm based on big data and neural network in this article has achieved effective results. It can provide a certain reference

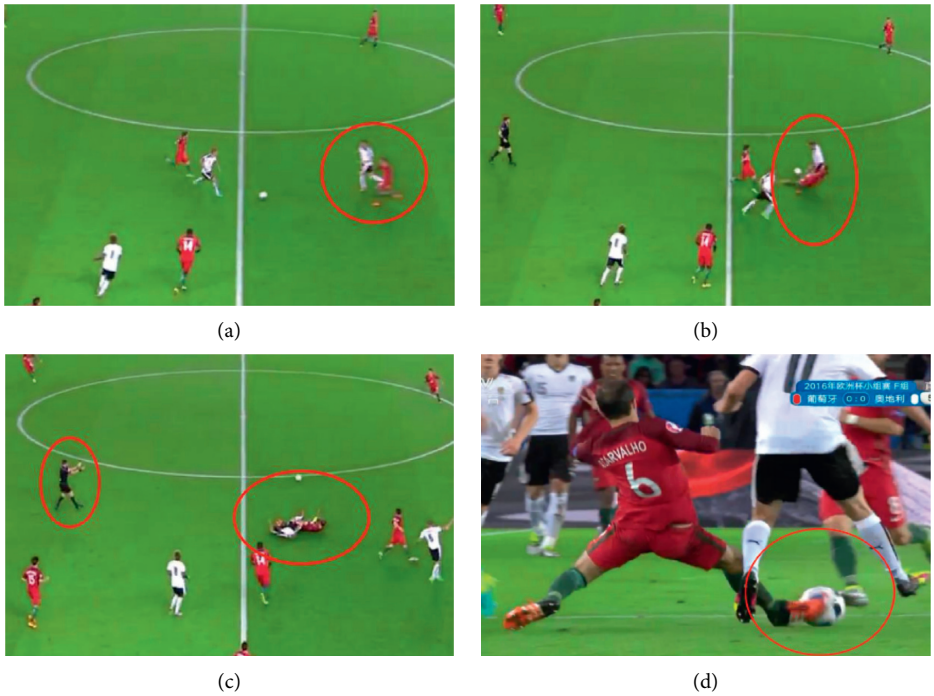


Figure 7: The neural network learns the time and location of the tackle. (a) Defensive position. (b) Tackle. (c) Referee's gesture. (d) Tackle slow motion replay.

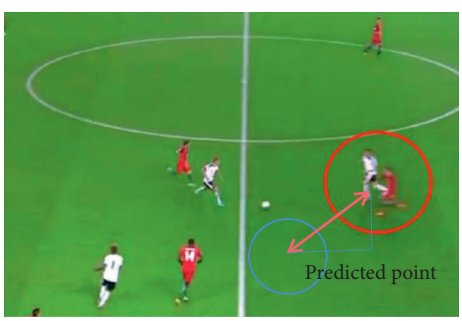


Figure 8: Predicted points and actual points.

value for the coaches and the training team's command and decision-making.

5. Conclusion

Based on big data and neural network technology, this paper designs a novel football technical and tactical command decision algorithm. Through, the experimental research proved that big data analysis can extract effective feature information. Secondly, we established a neural network model based on long- and short-term memory and proved through experiments that the algorithm can predict the pass success rate and defensive tackle position. It can formulate reasonable training plans for practitioners in the sports industry, especially coaches, and provide scientific command and decision-making advice on football standing skills. The method proposed in this paper has achieved good results in football technical and tactical command and decision-making and also provided some new ideas for the subject of football combined with computer technology.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts regarding the publication of this paper.

Acknowledgments

This work was supported by the Social Science Development Research Project of Hebei Province under Grant 20200502100.

References

[1] D. Memmert, "Data analytics in football: positional data collection, modeling, and analysis," *Journal of Sport Management*, vol. 33, p. 574, 2019.

[2] H. Sarmento, A. Figueiredo, C. Lago-Peñas et al., "Influence of tactical and situational variables on offensive sequences during elite football matches," *Journal of Strength and Conditioning Research*, vol. 32, no. 8, pp. 2331–2339, 2018.

[3] F. Sors, M. Grassi, T. Agostini, and M. Murgia, "The sound of silence in association football: home advantage and referee bias decrease in matches played without spectators," *European Journal of Sport Science*, pp. 1–9, 2020.

[4] E. Elyakim, E. Morgulev, R. Lidor, Y. Meckel, M. Arnon, and D. Ben-Sira, "Comparative analysis of game parameters

between Italian league and Israeli league football matches," *International Journal of Performance Analysis in Sport*, vol. 20, no. 2, pp. 165–179, 2020.

[5] R. Izzo, T. D'isanto, G. Raiola, A. Cejudo, N. Ponsano, and C. H. Varde'i, "The role of fatigue in football matches, performance model analysis and evaluation during quarters using live global positioning system technology at 50 Hz," *Sport Science*, vol. 13, no. 1, pp. 30–35, 2020.

[6] M. J. Zammit, "Predictive analysis of football matches using in-play data," Master's thesis, University of Malta, Msida, Malta, 2018.

[7] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, and S. Tian, "Feature refinement and filter network for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[8] W. Cai and Z. Wei, "PiiGAN: generative adversarial networks for pluralistic image inpainting," *IEEE Access*, vol. 8, pp. 48451–48463, 2020.

[9] X. Ning, P. Duan, W. Li, and S. Zhang, "Real-time 3D face alignment using an encoder-decoder network with an efficient deconvolution layer," *IEEE Signal Processing Letters*, vol. 27, pp. 1944–1948, 2020.

[10] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[11] X. Ning, K. Gong, W. Li, and L. Zhang, "JWSAA: joint weak saliency and attention aware for person re-identification," *Neurocomputing*, 2020.

[12] W. Cai, B. Liu, Z. Wei, M. Li, and J. Kan, "TARDB-Net: triple-attention guided residual dense and BiLSTM networks for hyperspectral image classification," *Multimedia Tools and Applications*, pp. 1–22, 2021.

[13] X. Ning, W. Li, B. Tang, and H. He, "BULDP: biomimetic uncorrelated locality discriminant projection for feature extraction in face recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2575–2586, 2018.

[14] Z. Wang, C. Zou, and W. Cai, "Small sample classification of hyperspectral remote sensing images based on sequential joint deeping learning model," *IEEE Access*, vol. 8, pp. 71353–71363, 2020.

[15] W. Cai and Z. Wei, "Diversity-generated image inpainting with style extraction," 2019, http://arxiv.org/abs/1912.01834.

[16] Z. L. Yang, S. Y. Zhang, Y. T. Hu, Z. W. Hu, and Y. F. Huang, "VAE-Stega: linguistic steganography based on variational auto-encoder," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 880–895, 2020.

[17] X. Ning, X. Wang, S. Xu et al., "A review of research on co-training," *Concurrency and Computation: Practice and Experience*, 2021.

[18] D. Krzysztof and B. Pawet, "Analysis of goals and assists diversity in English premier league," *Journal of Health Sciences*, vol. 4, no. 5, pp. 47–56, 2014.

[19] M. Michalls, R. Joaquín González, A. Vasilis, and A. Rafael, "The creation of goal scoring opportunities in professional soccer," *International Journal of Performance Analysis in Sport*, vol. 19, no. 3, pp. 452–465, 2019.

[20] A Kubayi, "Analysis of goal scoring patterns in the 2018 FIFA world Cup," *Journal of Human Kinetics*, vol. 71, no. 1, pp. 201–210, 2019.

[21] J. Perl, A. Grunz, and D. Memmert, "Tactics analysis in soccer–an advanced approach," *Dshs*, vol. 12, 2013.

[22] T. Reilly and A. M. Williams, "International research in sports and exercise science including physiology, psychology, sports medicine and biomechanics, coaching and talent identification," *Journal of Sports Sciences*, vol. 23, no. 6, 2013.

[23] R. Pollard and C. Reep, "Measuring the effectiveness of playing strategies at soccer," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 46, no. 4, pp. 541–550, 1997.

[24] L. R. Medsker and L. C. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, 2001.

[25] S. M. Arabzad, M. E. Tayebi Araghi, S. Sadi-Nezhad, and N. Ghofrani, "Football match results prediction using artificial neural networks; the case of Iran Pro League," *Journal of Applied Research on Industrial Engineering*, vol. 1, no. 3, pp. 159–179, 2014.

[26] F. Ameri, S. Moradian, T. M. Amani, and K. Faez, "The use of fundamental color stimulus to improve the performance of artificial neural network color match prediction systems," *Jounal of Chemistry and Chemical Engineering*, vol. 24, no. 36, 2005.

[27] C. D. M. Bezerra and C. J. Hawkyard, "Computer match prediction for fluorescent dyes by neural networks," *Coloration Technology*, vol. 116, no. 5-6, pp. 163–169, 2000.

[28] H. Li, "Analysis on the construction of sports match prediction model using neural network," *Soft Computing*, vol. 24, pp. 1–11, 2020.

[29] S. Booth, A. Shah, Y. Zhou, and J. Shah, "Sampling prediction-matching examples in neural networks: a probabilistic programming approach," 2020, http://arxiv.org/abs/2001.03076.

[30] Z. Wang, C. Long, G. Cong, and C. Ju, "Effective and efficient sports play retrieval with deep representation learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 499–509, Anchorage, AK, USA, July 2019.

[31] M. A. Russo, L. Kurnianggoro, and K. H. Jo, "Classification of sports videos with combination of deep learning models and transfer learning," in *Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–5, IEEE, Cox's Bazar, Bangladesh, February 2019.

[32] D. Tang, "Hybridized hierarchical deep convolutional neural network for sports rehabilitation exercises," *IEEE Access*, vol. 8, pp. 118969–118977, 2020.

[33] Y. C. Huang, I. N. Liao, C. H. Chen, T. U. İk, and W. C. Peng, "TrackNet: a deep learning network for tracking high-speed and tiny objects in sports applications," in *Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–8, IEEE, Taipei, Taiwan, September 2019.

[34] J. Lee, S. Moon, D. W. Nam, J. Lee, A. R. Oh, and W. Yoo, "A study on sports player tracking based on video using deep learning," in *Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1161–1163, IEEE, Jeju Island, South Korea, October 2020.

Hindawi
Scientific Programming
Volume 2021, Article ID 5526971, 10 pages
https://doi.org/10.1155/2021/5526971

Research Article

Injury Risk Prediction of Aerobics Athletes Based on Big Data and Computer Vision

Dongdong Zhu,[1] Honglei Zhang,[2] Yulong Sun,[3] and Haijie Qi[4]

[1]*Institute of Physical Education, Dezhou University, Dezhou 253023, Shandong, China*
[2]*Hengshui University, Hengshui 053000, China*
[3]*Hebei North University, Zhangjiakou 075000, Hebei, China*
[4]*Hebei Vocational College of Politics and Law, Shijiazhuang, 050061, China*

Correspondence should be addressed to Haijie Qi; qihaijie@cumt.edu.cn

Received 21 January 2021; Revised 23 February 2021; Accepted 4 March 2021; Published 2 April 2021

Academic Editor: Shah Nazir

In recent years, competitive aerobics has been rapidly popularized and developed, and the level of sports skills has also been greatly improved. The performance of some events has gradually approached and reached the advanced level. Therefore, it is vital to invest in the quantitative analysis and cross-disciplinary comprehensive research of aerobics performance and related factors. This paper adopts big data analysis technology and computer vision technology based on convolutional neural network, according to the related theories of sports biomechanics and computer image recognition, to establish a loss risk prediction model for aerobics athletes. The approach firstly has used technology of big data analysis for analyzing the characteristics of competitive aerobics sports data. Secondly, the approach combines the convolutional neural network to visually recognize the aerobics sports images and establish a two-branch prediction model. Finally, the output can be fused to accurately diagnose and evaluate the level of physical fitness development of aerobics athletes, the focus and goal of training content are clarified, and the scientific degree of aerobics training is improved. The study can help injury risk prediction of aerobic athletes based on applications of big data and computer vision.

1. Introduction

In recent years, competitive aerobics has developed rapidly in our country, and the corresponding sports injury risks have gradually increased. A number of studies have shown that due to the characteristics of aerobics itself, such as strict time requirements, more difficult movement requirements, fast-paced music accompaniment, and coherent coordinated movements, athletes will suffer sports injuries [1–3] if they are not paying attention. The shoulders, elbows, wrists, waist, thighs, knees, calves, and ankles are the parts that are more prone to injury during aerobics training. Among them, the most prone to injury is the ankle joint [4, 5]. In addition, the types of injury most likely to occur for competitive aerobics athletes is closed injury, most of which are joint strain, sprain and muscle strain, and chronic injuries [6–9], which are the main ones. However, the current scholars' research on aerobics injuries usually uses questionnaire surveys or expert interviews to determine the injured parts and possible mechanisms of aerobics exercises, and there is a lack of objective empirical research.

In addition, teenagers are in the golden stage of physical development, and various physical qualities will be significantly improved during this period. However, in interviews with young aerobics athletes and coaches, it is found that young athletes have more injuries than adult athletes. . This is due to the weakness of the muscles and joints of young athletes, which limits the ability to develop skills and long-term training of irregular technical movements and body postures. In the adolescent period of aerobics athletes, scientific and reasonable training can not only promote the physical development of adolescents but also improve their athletic ability more effectively [10]. Therefore, timely discovery of the causes of injury to young athletes and timely

prevention are critical to improving the skill level of young athletes and extending their sports life.

Due to the characteristics of competitive aerobics, athletes are required to complete a series of high-intensity movements in a short time, which requires a higher level of physical fitness and physical flexibility. Studies have confirmed that long-term high-intensity repetitive exercise training and asymmetric sports skills and postures will increase the risk of athletes' injuries. At the same time, the adolescent stage is a special period of physical development and a period of high incidence of sports injuries [11]. Therefore, this research is based on the development characteristics of adolescents' physical ability and the special characteristics of aerobics. It takes college students in a certain city as the research object, and the tests are conducted in groups of different genders and aerobics sports grades, and uses the receiver operating characteristic (ROC) curve to formulate evaluation standards, analyze the physical weakness of young aerobics athletes, and evaluate the risk of noncontact injury to provide a certain theoretical reference for subsequent aerobics training for young people. This paper presents a method of motion recognition for aerobics athletes based on machine vision, which recognizes joint strain, sprain, and muscle strain caused by their movements, and uses big data to train the vision-based deep learning [12–16] algorithm of this article. Following are the main innovative points of this paper:

(i) This article proposes a method of motion recognition for aerobics athletes based on machine vision, which recognizes joint strain, sprain, and muscle strain caused by their movements.

(ii) This paper constructs a dual-branch injury risk prediction model for aerobics athletes. One branch uses big data to analyze the characteristics of aerobics athletes' movement injuries, and the other branch builds a deep convolutional neural network model to identify joint strain, sprain, and muscle strain and perform counting and prediction.

(iii) We have conducted sufficient comparative experiments and ablation studies to prove the effectiveness of the algorithm based on big data and computer vision proposed in this paper. It can be used to discover the causes of injury to young athletes in time and prevent them in time, which is useful for improving the skills of young athletes. and prolonging sports life.

The organization of the paper is as follows: Section 2 shows the related research to the proposed area of research. Section 3 represents the methodology section of the proposed study with details of the given approach. Section 4 shows the experiments and results of the current study. The paper is concluded in Section 5.

2. Related Research

Various approaches and techniques have been devised in the literature for injury risk prediction of aerobics athletes. Fanian et al. [17] studied 6755 cases of sports injuries and investigated and found that the sport with the highest incidence is football; the common injury sites are knee joints (15%), calves (11.9%), and wrist joints (11%). The most common types of injuries are fractures (12.8%), cartilage injuries (6.87%), and contusions (2.5%); 12% of the patients underwent surgery; the average hospital stay was 3-4 days. Michael [18] surveyed athletes in various sports such as gymnastics, basketball, football, and running. After summarizing and analyzing the types and characteristics of athletes' sports injuries in different sports, they proposed the diagnosis and treatment of various types of sports injuries. Method aims at the occurrence of injury and points out two major steps to reduce the occurrence of sports injuries: one is to wear protective gear in vulnerable parts to reduce the probability of injury and the degree of injury; the other is to improve the athletes as much as possible. Physical fitness, including strength, agility, and flexibility, reduces the chance of athletes' injuries. Wiese-Bjomstal et al. [19] pointed out that sports injury is a relatively unacceptable thing for athletes. After a sports injury occurs, athletes will have complex psychology during training and rehabilitation, which is mainly reflected in cognition, emotion, and behavior. On the one hand, when athletes experience the process of recovering from injury and returning to training, they often have certain cognitive and emotional reactions, which are mainly affected by the individual and the environment. In addition, the time of injury will also affect the psychology of injured athletes to varying degrees. When sports injuries occur shortly before major competitions, the athletes' sense of disappointment and despair will be greatly enhanced [20].

Malliou et al. [21] used questionnaire surveys and conducted statistics and analysis on the injuries of athletes engaged in aerobics training and found that, in the injured population, lower limb injuries accounted for 97.3% and ankle and knee injuries were the most common in aerobics training. At the same time, it is pointed out that training time, years, and training level will have an impact on injury. Bintoudi et al. [22] investigated two aerobics pedal athletes and found that they had knee joint pain and fat pad edema. The article discussed the possible pathogenic factors and mechanisms of aerobics. Kiesel et al. [23] found in a study of professional American rugby players during nonseason that the scores of the FMS test are directly proportional to the athlete's sports injury risk. Through the calculation of the receiver operating curve, a benchmark score of 14 points is delineated for rugby players. Athletes with a season average FMS score of less than 14 points have a much higher risk of injury than those with more than 14 points. In his subsequent similar research, he found that after targeted intervention training for athletes, the number of FMS scores greater than 14 increased significantly, and athletes with movement asymmetry problems could also be improved. Dennis Rex [24] conducted FMS tests on 67 college football players and combined their lower limb explosive power and season injury data. They believed that the FMS test can be used as an assessment tool to predict serious sports injuries and reminded that the FMS score was less than 11 points.

The sports injury risk of athletes is 9 times that of other athletes, and timely corrective training is required. Dorrel et al. [25] followed up on 257 college athletes' injuries after conducting FMS tests and found that those athletes who scored less than 15 on the FMS test were more likely to get injured.

3. Methodology

This section mainly introduces the big data platform technology and related theories related to the research of this article. Figure 1 shows the examples of aerobics sports injuries.

The proposed approach firstly introduces the core architecture, working principle, and related ecosystem of the Hadoop platform and then introduces the basic ideas of the data mining theory of big data analysis and the common algorithms of data mining are outlined. Finally, the principle of convolutional neural network [26–31] and the two-branch algorithm proposed in this paper are explained. Figure 2 shows the overall framework of the proposed approach.

3.1. Data Mining and Analysis. The following subsections show the data mining and analysis section of the paper.

3.1.1. Big Data Platform. Hadoop is currently one of the big data platforms widely used by relevant research institutions in the industry. It is an open-source top-level project maintained by the Apache Foundation. Hadoop inherits the idea of distribution and fully applies it to data storage and processing technology. The Hadoop ecosystem is also composed of many open source software. The open source of the entire ecosystem allows the software in it to be supervised and maintained by everyone, and the stability and security of the system will have a relatively high guarantee. Moreover, the open source of the entire system greatly facilitates the use of users. Big data computing is no longer the unique ability of some professional organizations and has gradually become a field that every developer can set foot in. Figure 3 is its ecosystem architecture.

HDFS is the distributed file system of Hadoop. As the data foundation of the entire ecosystem, HDFS is located at the bottom of the ecosystem. It has very low requirements for hardware resources, can run on computers with low configuration, and can ensure the safety and stability of data in the event of hardware failure through redundant storage. The file system divides each file into multiple parts, each part is called a data block, and each data block is copied into three copies and stored in different locations, which not only provides high throughput of data access but it also has extremely high fault tolerance. Figure 4 shows the HDFS architecture diagram. As the role of the manager in HDFS, NameNode undertakes the core tasks. It is responsible for managing the mapping information of data blocks. As the hot backup storage of the NameNode, the significance of the existence of the secondary and the NameNode is to cope with the single point of failure that may occur in the NameNode. HDFSClient is a client that accesses HDFS, and all client requests will first interact with the NameNode. Each DataNode maintains interaction with the NameNode and serves as a block where the slave node stores data.

MapReduce is the computing engine of Hadoop. JobTracker is the core of MapReduce, responsible for all task allocations and job scheduling. JobTracker decomposes tasks and distributes them to each TaskTracker node for execution. TaskTracker runs maps and reduce tasks and reports the status of tasks to JobTracker regularly. There are many important components in the Hadoop ecosystem. HBase is a high-performance, scalable column storage database for structured data. Hive originated from Facebook, which is a Hadoop-based data warehouse, mainly used to solve the problem of massive structured data statistics. The biggest feature of Hive is that it can convert SQL into MapReduce jobs so that it can be executed on Hadoop. Zookeeper is used to solve the problem of application coordination in a distributed environment. Sqoop is responsible for data transmission tasks between traditional databases and Hadoop. Pig can convert scripting languages into MapReduce jobs and execute them on Hadoop. Mahout is a library of machine learning algorithms for Hadoop. Flume is an open source log collection system.

3.1.2. Data Mining. As a data processing technology, data mining aims to discover the laws and knowledge behind the phenomena hidden in reality through a series of operations and calculations on data. It is also one of the current research hotspots in the scientific field. Data mining technology draws on the advantages of a variety of related technologies, which can extract hidden valuable information from real data and provide references for actual activities. From the methodological point of view, data mining can be divided into two categories: description and prediction. The similarity between the two is that the law is calculated through the existing data. The difference is that the purpose of the description is to provide interpretation support for the law of the data, and the prediction is to provide forecasts for actual activities.

The functions of data mining are divided into the following categories:

(1) Concept or class description:

This kind of data mining mainly describes the class or concept of data through the method of data differentiation and characterization. Data classification classifies the data to be mined by constructing a comparative dataset. Data characterization is by first querying existing related datasets and then summarizing their characteristics.

(2) Predictive modeling:

Data mining prediction methods are divided into two categories, among which classification and regression methods are applied to discrete and continuous variables. Predictive modeling derives a model through training so that the error between the predicted value and the actual value of the specified variable reaches the global minimum. The

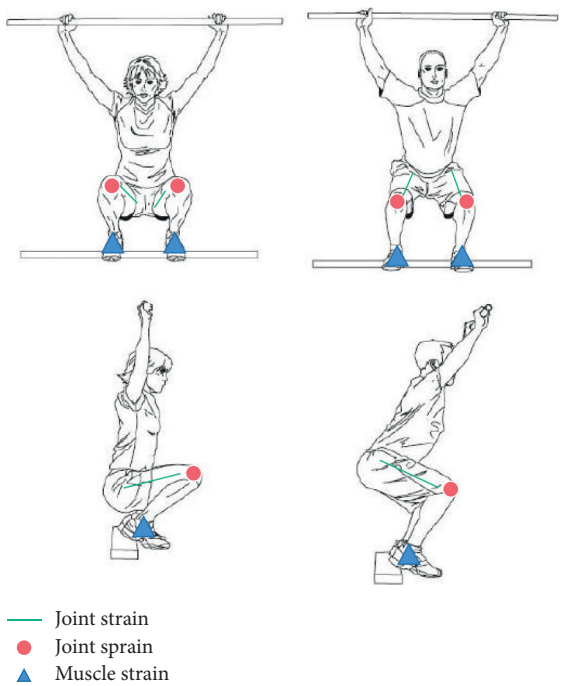


Figure 1: Examples of aerobics sports injuries.

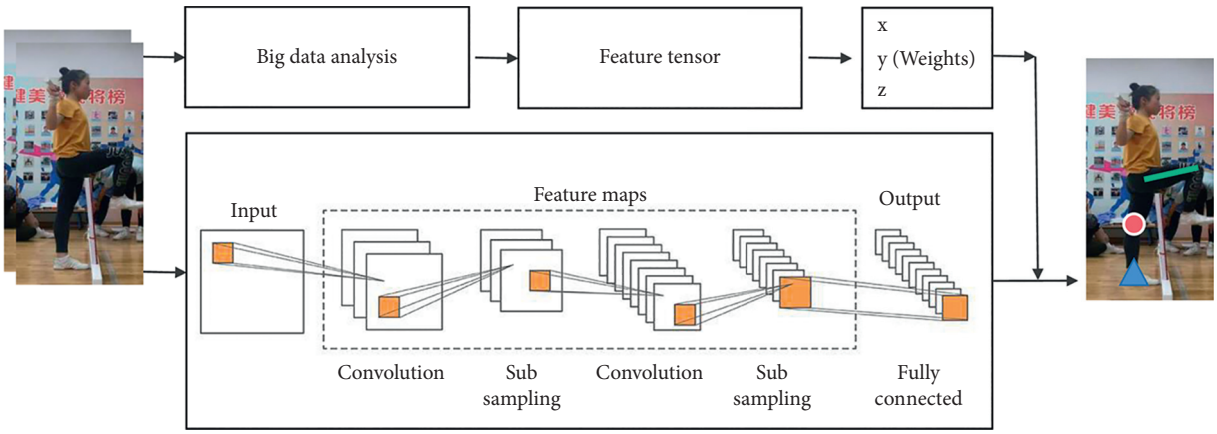


Figure 2: Overall framework of our proposed algorithm.

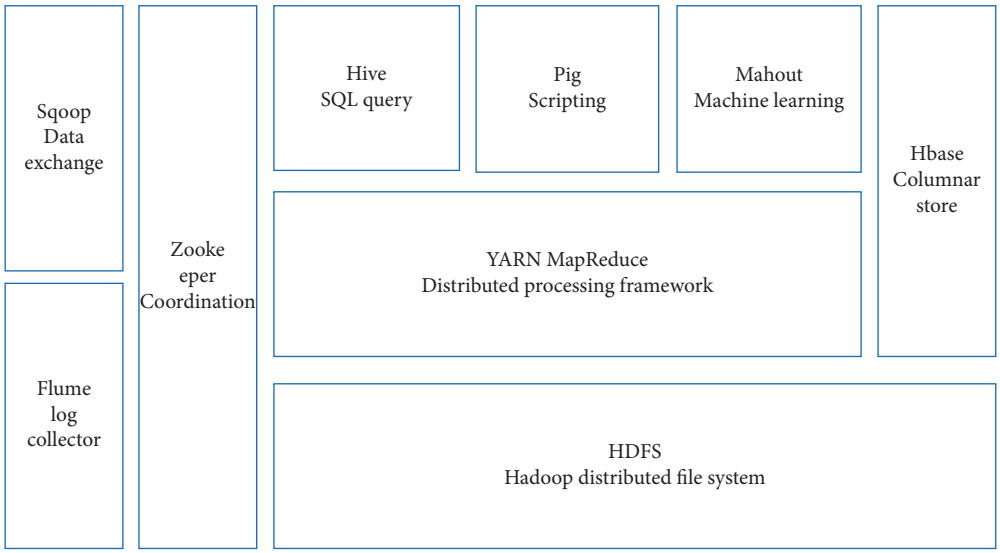


Figure 3: Hadoop ecosystem architecture.

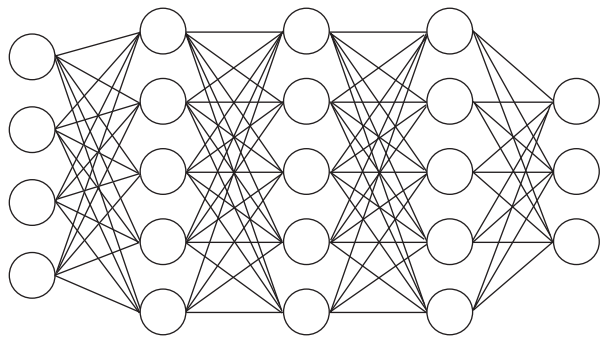


Figure 4: Schematic diagram of fully connected neural network.

application of this type of algorithm is commonly used in disease risk prediction.

(3) Association analysis:

Used to describe the associated features in the data, the discovered patterns are usually expressed in the form of implicit rules or feature subsets. The association rules derived from the association analysis can reveal the dependence of each element in the dataset and indicate the conditions under which the attributes appear in the dataset with a specific frequency. For example, the correlation analysis of multiple symptoms of patients can find the correlation rules that occur between the symptoms.

(4) Cluster analysis:

The main function of this type of data mining algorithm is to divide the dataset into valuable or meaningful groups. Cluster analysis and classification analysis are both similar and different. The difference is that cluster analysis belongs to unsupervised classification. The class label needs to be obtained from the data, and the number of classes is not known before clustering.

(5) Abnormal detection:

Based on the analysis of the dataset, abnormal characteristics or abnormal data are obtained. It is often used in the diagnosis of abnormal diseases and the detection of abnormal network traffic.

3.1.3. Data Mining Algorithm. As data mining continues to incorporate new domain knowledge, data mining algorithms are constantly evolving and mining methods are becoming more abundant. People can choose specific mining algorithms based on mining needs. The following are several commonly used algorithms for data mining:

(1) Neural network:

Neural network is an intelligent algorithm that simulates human brain nerve transmission. It is generally composed of three parts: input, output, and implicit. It is divided into three models. They are feedforward network model, feedback network model, and self-organization network model. The neural network has strong ability to process nonlinear data, good fault tolerance performance, and high classification accuracy, but low performance in the interpretation of the results. The most commonly used in the medical field is a multilayer feedforward neural network, the BP neural network. It is worth noting that the neural network used in this article is a convolutional neural network.

(2) Decision tree:

Decision tree is a tree structure model with classification rules, and its logical branch relationship is top-down. The decision tree selects the root node according to the variable attributes according to parameters such as information gain and Gini coefficient and then divides down according to the variable attributes of the root node to form branches; then, each branch node retests the variable attributes and continues to branch down and so on continue until the node's category is homogenized or reaches the set threshold. The algorithm can be converted into classification rules to classify diseases according to their symptoms, thereby predicting diseases. Common algorithms include ID3 algorithm, C4.5 algorithm, and CART algorithm.

(3) Clustering algorithm:

The purpose of clustering algorithm is to obtain several classes through computational analysis, which is an unsupervised learning method. Among them, the data between different classes are usually uncorrelated or there are certain differences, and the data between the same classes have a certain correlation or similarity. The k-means algorithm is the most common clustering algorithm.

(4) Association rules:

Association rule mining is the process of finding strong association rules through the specified minimum support and minimum confidence. Usually, it consists of two parts, one is to find all frequent item sets, and the other is to find the association rules in frequent item sets. Common algorithms include Apriori algorithm and FP-growth algorithm.

(5) Association classification:

Association classification algorithm is one of the important classification methods. The characteristic of this classification method is to first extract the association classification rules and then build a model to predict unknown instances and combine association rule mining and classification. Association classification algorithms usually consist of three parts: rule generation, rule sorting and pruning, and prediction of new instances. Common algorithms include CBA algorithm, CMAR algorithm, and ACSER algorithm.

3.2. Convolutional Neural Networks. Figure 4 is a structural diagram of a fully connected neural network, and Figure 5 is a structural diagram of a convolutional neural network.

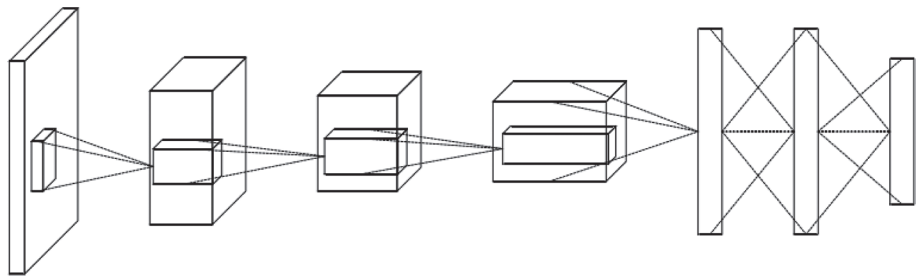


Figure 5: Schematic diagram of convolutional neural network.

Although the two are quite different on the surface, they are actually very similar in structure. Convolutional neural networks are also connected through layers of nodes, and each node also represents a neuron. The difference is that a fully connected neural network usually has a connection to every neuron in every two adjacent layers, while only some nodes in the adjacent layers of a convolutional neural network are connected, which can effectively alleviate the problem of excessive parameters of the fully connected neural network. Too many parameters of the neural network will cause the calculation speed to be slow, the calculation time is too long, and at the same time, it is more prone to overfitting. Convolutional neural networks can effectively reduce the number of parameters and speed up model training.

In the first few layers of the convolutional neural network, the data nodes are transformed into a three-dimensional matrix, and only some nodes are connected to adjacent layers. A convolutional neural network usually consists of the following structure.

3.2.1. Input Layer. The input layer is the input of the neural network. In the image-oriented convolutional neural network, it represents the pixel matrix converted after a picture is read by a computer. The pixel matrix is a three-dimensional matrix. The length and width represent the size of the image, and the depth represents the number of color channels of the image. When the input picture is a black and white picture, the depth is 1; when the input picture is a color picture, the depth is 3. Starting from the input layer, the three-dimensional matrix is transformed into another three-dimensional matrix through a different network structure until the final fully connected layer.

3.2.2. Convolutional Layer. The convolutional layer is the most important part of the convolutional neural network and the key to extracting features.

Figure 6 is a schematic diagram of the convolutional layer transformation. The small matrix of the input layer and the convolution kernel are convolved to obtain the small matrix of the output layer. The convolution kernel is also a three-dimensional matrix. Its length and width are manually set. The convolution kernel is also a three-dimensional matrix, its length and width are manually set, and the size is $3 * 3$ or $5 * 5$. Since the depth of the two matrices of the convolution operation must be the same, the depth of the filter cannot be changed and must be the same as the depth of the input layer matrix. The convolution kernel also needs to manually set the number, and its number determines the depth of the output layer matrix. The process from the input layer to the output layer is called forward propagation. Assuming that $a_{x,y,z}$ is used to represent the value of a certain point in the input matrix, $w_{x,y,z}$ represents the value of the position of the convolution kernel, and b represents the node's corresponding bias term parameter, then

$$g = f\left(\sum_{x=1}^{m}\sum_{y=1}^{n}\sum_{z=1}^{q} a_{x,y,z} \times w_{x,y,z} + b\right), \tag{1}$$

where g is the output of the corresponding point, (m, n, q) are the length, width, and number of the convolution kernel, respectively, and f is the activation function. The activation function acts as a nonlinear transformation of the output. The commonly used activation function consists of a linear rectification unit (ReLu) and a hyperbolic function (sigmoid).

3.2.3. Pooling Layer. It can be seen from Figure 7 that a pooling layer is usually added after the convolutional layer. The pooling layer can reduce the size of the input data, reduce the number of parameters, and speed up the network calculation at the same time. Preventing the occurrence of overfitting problems is similar to the forward propagation process of the convolutional layer. The pooling layer is also completed by sliding a filter structure on the matrix. The difference is that the pooling layer uses a simple maximum value or the average method.

Figure 7 shows the calculation process of the pooling layer using maximum pooling. Here, the size of the filter is 2×2, that is, the target area of a pooling operation is 2×2 nodes, and the step size is 2, that is, the calculation is done in every 2 steps.

3.2.4. Fully Connected Layer. The fully connected layer of CNN, like the fully connected neural network, maps the features learned by the previous convolutional layer and the pooling layer to the sample label space, which acts like a classifier. The fully connected layer is not necessary; it can be replaced by a convolutional layer using a convolution kernel of size 1×1.

3.2.5. Softmax Layer. In the use of multiclassification, the softmax layer is often used as the last layer of the deep neural network, which makes the output of the network straightforward to the probabilities of various classifications. The three aerobics sports injury categories in this article are multiclassification tasks.where n is the number of training samples, y is the label value of the training data, L is the number of network layers, and $a^L(x)$ is the output of the network.

Introduce the mean square error function as the cost function in back propagation,

$$C = \frac{1}{2n}\sum_x \left\| y - a^L(x) \right\|^2, \tag{2}$$

Calculate the error of each unit of the output layer, the definition of the error of the jth unit of the L-th layer:

$$\delta_j^L = \frac{\partial C}{\partial z_j^L}. \tag{3}$$

According to the chain rule, we have

$$\delta_j^L = \frac{\partial C}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L}\frac{\partial a_j^L}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L} f'\left(z_j^L\right). \tag{4}$$

Its matrix form is

$$\partial^l = \left(\left(w^{l+1}\right)^T \delta^{l+1}\right) \otimes f'\left(z^l\right). \tag{5}$$

After calculating the error of the output layer, the error of the previous layer should be calculated. According to the chain rule, we can get

$$\begin{aligned} \partial_j^l &= \frac{\partial C}{\partial z_j^l} \\ &= \sum_k \frac{\partial C}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial z_j^l} \\ &= \sum_k \frac{\partial z_k^{l+1}}{\partial z_j^l} \delta_k^{l+1}. \end{aligned} \tag{6}$$

Therefore,

$$\delta_j^l = \sum_k w_{kj}^{l+1} \delta_j^{l+1} f'\left(z_j^l\right). \tag{7}$$

In this way, errors can run through the entire network by back propagation. Let joint strain, sprain, and muscle strain be x, y, and z. The output equation through the softmax function is

$$\text{Soft max}\,(\text{input\_image}) = \max[x, y, z]. \tag{8}$$

4. Experiments and Results

The following subsections show the experiments and results of the paper.

4.1. Experimental Environment and Hyperparameter Settings. This experiment uses the PyCharm compiler and the TensorFlow deep learning framework in the Windows environment and uses the small batch learning method, the Adam optimizer, the learning rate is 0.0001, 100 images are selected for training at a time, and a total of 1000 iterations are performed. The experimental model parameters are shown in Table 1.

4.2. Subjects of the Experiment. Taking the freshman and sophomore aerobics students in a certain city as the research object, there are a total of 60 girls. At the beginning of the experiment, there were 2 people with acute sports injuries, and they were finally confirmed as 58 people. There was no significant difference in the height, weight, age, and training years of the 58 aerobics professional girls.

4.3. Experimental Methods. The visual recognition experiment based on convolutional neural network in this paper is divided into training phase and testing phase. The training phase has the following contents.

4.3.1. Aerobics Squat (Deep Squat). Stand upright with your feet shoulder-width apart or slightly wider than shoulder-width apart, with your toes pointing straight ahead. Place the test rod on the top of the head and adjust your hands to make the elbow joint 90°; extend both arms at the same time to lift the test rod to the head. Then, slowly squat until your thighs are parallel to the ground, keep your heels from leaving the ground, raise your head and chest, keep your back straight, do not arch your waist, and hold the test rod as far above your head as possible. During the squat process, the knee joints on both sides are in the same plane as the feet. Do not buckle your knees inward, and always keep your toes pointing forward. This squat test was repeated three times. This experiment takes photos of the entire training process for the neural network to learn the characteristics of joint strain, sprain, and muscle strain.

4.3.2. Active Straight Leg Raise. The subject lies on his back, with his arms on his side, palms facing up, and lying flat; the test board is placed under the knees with the toes pointing directly up.

It is placed in the middle of the line between the anterior superior iliac spine and the midpoint of the patella, perpendicular to the ground. The subject lifted one leg to the maximum extent and kept the knee joint straight during the lifting process. The knee joint on the other side should be kept in contact with the test board as much as possible, with the toes pointing upward, and the opposite side was also the same. This experiment takes photos of the entire training process for the neural network to learn the characteristics of joint strain, sprain, and muscle strain.

The training process is 1 week, and the test actions are 50 sets. The experimental results are shown in Table 2.

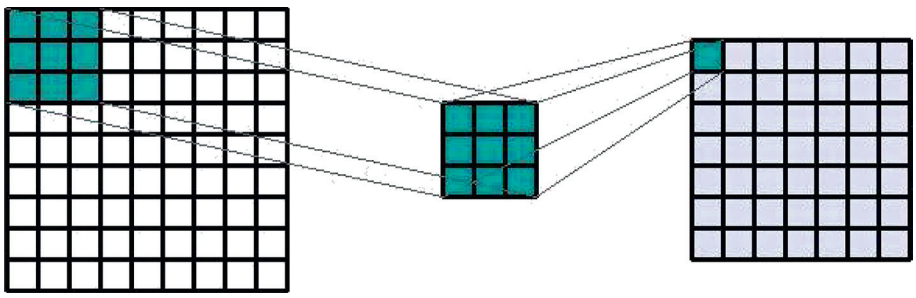


Figure 6: Convolutional layer.

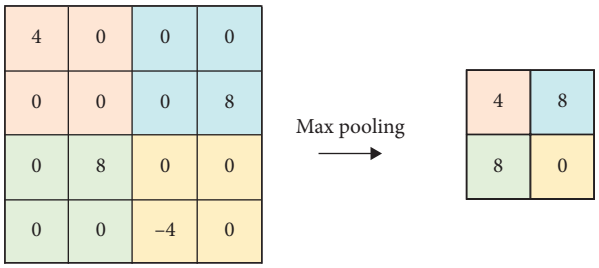


Figure 7: Max pooling layer.

Table 1: Experimental environment configuration.

	Filter size	Number of filters	Sliding step	Zero padding
Conv layer	5 ∗ 5	32	2/1	2
Pooling layer	2 ∗ 2		2	0
Conv layer	3 ∗ 3	64	1	1
Pooling layer	2 ∗ 2		2	0
Conv layer	3 ∗ 3	128	1	1
Pooling layer	2 ∗ 2		2	1
FC layer		512		
FC layer		256		
FC layer		128		
FC layer		100		
Softmax layer		3		

Table 2: Aerobics sports injury prediction results.

	Joint strain (%)	Joint sprain (%)	Muscle strain (%)
Group 1	74	68	71
Group 2	69	64	77
Group 3	78	64	69
Group 4	79	79	71
Group 5	80	61	69
Group 6	74	71	79

The CNN training results are shown in Figure 8. When the number of training steps is greater than 900, the loss function value drops below 0.1.

4.4. Visualization of Results. It can be seen from Figure 9 that the acute sports injury rate of aerobics in the low-risk group is lower than that of the high-risk group. Therefore, regardless of whether there is a corrective training intervention, the injury rate of the two low-risk groups will not be very high. The members of the high-risk group have a higher risk of injury during the training process. Therefore, the injuries of the two high-risk groups can better verify the effectiveness of the corrective training. This experimental

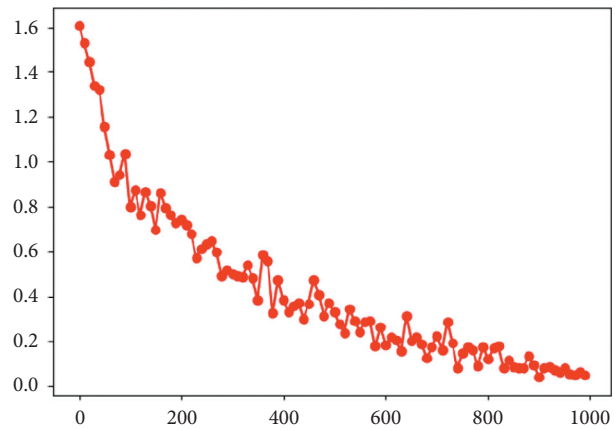


Figure 8: Declining curve of loss function.

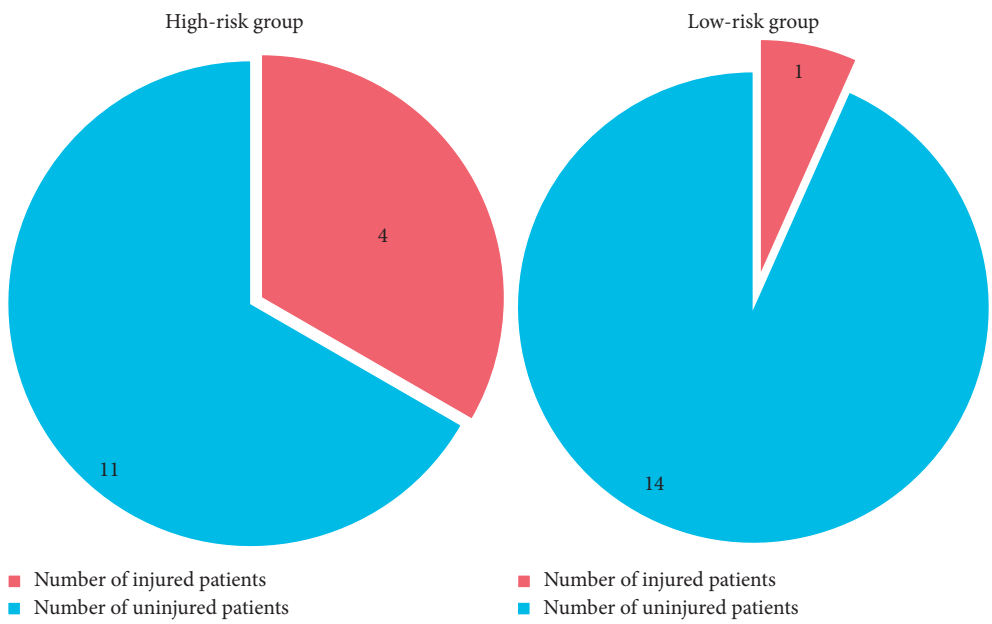


Figure 9: Comparison results of injury rate prediction between high-risk group and low-risk group.

test proves the effectiveness of the algorithm in this paper and can improve the scientific degree of aerobics training.

5. Conclusion

Competitive aerobics has been rapidly promoted and established, and the level of sports expertise has also been significantly enhanced. The performance of some events has progressively approached and reached the advanced level. Therefore, it is vital to invest in the quantitative analysis and cross-disciplinary wide-ranging research of aerobics performance and associated factors. This paper constructs a novel dual-branch aerobics athlete injury risk prediction algorithm based on big data and computer vision technology, and through experimental research, it has been proved that big data analysis can extract effective features from competitive aerobics data. Secondly, combined with a convolutional neural network to visually recognize aerobics images, it can accurately diagnose and evaluate the physical fitness development level of aerobics athletes, clarify the focus and objectives of the training content, and improve the scientific degree of aerobics training.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] C. Bolling, W. Van Mechelen, H. R. Pasman, and E. Verhagen, "Context matters: revisiting the first step of the 'sequence of

prevention' of sports injuries," *Sports Medicine*, vol. 48, no. 10, pp. 2227–2234, 2018.

[2] A. Carbone and S. Rodeo, "Review of current understanding of post-traumatic osteoarthritis resulting from sports injuries," *Journal of Orthopaedic Research*, vol. 35, no. 3, pp. 397–405, 2017.

[3] A. Olmedilla-Zafra, V. J. Rubio, E. Ortega, and A. García-Mas, "Effectiveness of a stress management pilot program aimed at reducing the incidence of sports injuries in young football (soccer) players," *Physical Therapy in Sport*, vol. 24, pp. 53–59, 2017.

[4] X. Xiao, W. Xiao, X. Li, B. Wan, and G. Shan, "The influence of landing mat composition on ankle injury risk during a gymnastic landing: a biomechanical quantification," *Acta of Bioengineering and Biomechanics*, vol. 19, no. 1, 2017.

[5] J. R. Miller, K. W. Dunn, L. J. Ciliberti Jr., S. W. Eldridge, and L. D. Reed, "Diagnostic value of early magnetic resonance imaging after acute lateral ankle injury," *The Journal of Foot and Ankle Surgery*, vol. 56, no. 6, pp. 1143–1146, 2017.

[6] S. Malone, M. Roe, D. A. Doran, T. J. Gabbett, and K. Collins, "High chronic training loads and exposure to bouts of maximal velocity running reduce injury risk in elite Gaelic football," *Journal of Science and Medicine in Sport*, vol. 20, no. 3, pp. 250–254, 2017.

[7] X. Deng, X. Zhang, W. Li et al., "Chronic liver injury induces conversion of biliary epithelial cells into hepatocytes," *Cell Stem Cell*, vol. 23, no. 1, pp. 114–122, 2018.

[8] Y. A. Tuakli-Wosornu, E. Mashkovskiy, T. Ottesen, M. Gentry, D. Jensen, and N. Webborn, "Acute and chronic musculoskeletal injury in para sport," *Physical Medicine and Rehabilitation Clinics of North America*, vol. 29, no. 2, pp. 205–243, 2018.

[9] L. Bowen, A. S. Gross, M. Gimpel, and F.-X. Li, "Accumulated workloads and the acute:chronic workload ratio relate to injury risk in elite youth football players," *British Journal of Sports Medicine*, vol. 51, no. 5, pp. 452–459, 2017.

[10] S. J. Biddle, C. J. Wang, N. L. Chatzisarantis, and C. M. Spray, "Motivation for physical activity in young people: entity and incremental beliefs about athletic ability," *Journal of Sports Sciences*, vol. 21, no. 12, pp. 973–989, 2003.

[11] W. Van Mechelen, H. Hlobil, and H. C. G. Kemper, "Incidence, severity, aetiology and prevention of sports injuries," *Sports Medicine*, vol. 14, no. 2, pp. 82–99, 1992.

[12] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, and S. Tian, "Feature refinement and filter network for person Re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, In press.

[13] W. Cai and Z. Wei, "PiiGAN: generative adversarial networks for pluralistic image inpainting," *IEEE Access*, vol. 8, pp. 48451–48463, 2020.

[14] X. Ning, P. Duan, W. Li, and S. Zhang, "Real-time 3D face alignment using an encoder-decoder network with an efficient deconvolution layer," *IEEE Signal Processing Letters*, vol. 27, pp. 1944–1948, 2020.

[15] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[16] X. Ning, K. Gong, W. Li, and L. Zhang, "JWSAA: joint weak saliency and attention aware for person re-identification," *Neurocomputing*, 2020, in Press.

[17] H. Fanian, M. Khalilian, A. Zamanietal et al., *XXIX F1MS World Congreves of Sports Medicine*, pp. 14–16, Beiting, Sanya, China, 2006.

[18] L. Michael, *Tuggy & Cora Collette Brenner. Taylor Musculoskeletal Problems and Injuries a Handbook*, pp. 205–231, Springer, New York, NY, USA, 2006.

[19] D. M. Wiese-Bjomstal, A. M. Smith, S. M. Shaffer et al., "An intergraded model of response to sport injury: psychological and societal dynamics," *Journal of Applied Sport Psychology*, vol. 10, no. 10, pp. 46–69, 1993.

[20] T. Bianco, S. Malo, and T. Orlick, "Sport injury and illness: elite skiers describe their experiences," *Research Quarterly for Exercise and Sport*, vol. 70, no. 2, pp. 157–169, 1999.

[21] P. Malliou, A. Gioftsidou, G. Pafis, A. Beneka, and G. Godolias, "Proprioceptive training (balance exercises) reduces lower extremity injuries in young soccer players," *Journal of Back and Musculoskeletal Rehabilitation*, vol. 17, no. 3-4, pp. 101–104, 2004.

[22] A. Bintoudi, M. Goumenakis, and A. Karantanas, "Suprapatellar fat pad inflammation in step aerobics athletes: MR imaging evaluation of two cases," *Open Medicine*, vol. 7, no. 6, pp. 813–816, 2012.

[23] K. Kiesel, P. J. Plisky, and M. L. Voight, "Can serious injury in professional football be predicted by a preseason functional movement screen," *North American Journal of Sports Physical Therapy*, vol. 2, no. 3, pp. 147–158, 2007.

[24] B. Dennis Rex, *Strength Flexibility Functional Movement and Injury in Collegiate Men Football Players*, Georgia College and State University, Milledgeville, Georgia, 2010.

[25] B. Dorrel, T. Long, S. Shaffer, and G. D. Myer, "The functional movement screen as a predictor of injury in national collegiate athletic association division II athletes," *Journal of Athletic Training*, vol. 53, no. 1, pp. 29–34, 2018.

[26] W. Cai, B. Liu, Z. Wei, M. Li, and J. Kan, "TARDB-Net: triple-attention guided residual dense and BiLSTM networks for hyperspectral image classification," *Multimedia Tools and Applications*, pp. 1–22, 2021, In press.

[27] X. Ning, W. Li, B. Tang, and H. He, "BULDP: biomimetic uncorrelated locality discriminant projection for feature extraction in face recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2575–2586, 2018.

[28] Z. Wang, C. Zou, and W. Cai, "Small sample classification of hyperspectral remote sensing images based on sequential joint deeping learning model," *IEEE Access*, vol. 8, pp. 71353–71363, 2020.

[29] W. Cai and Z. Wei: (2019). Diversity-Generated Image Inpainting with Style Extraction. https://arxiv.org/pdf/1912.01834.pdf.

[30] Z. L. Yang, S. Y. Zhang, Y. T. Hu, Z. W. Hu, and Y. F. Huang, "VAE-Stega: linguistic steganography based on variational auto-encoder," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 880–895, 2020.

[31] X. Ning, X. Wang, S Xu et al., *A Review of Research on Co-Training. Concurrency and Computation: Practice and Experience*, Wiley, Hoboken, NJ, USA, 2021, In press.

Hindawi
Scientific Programming
Volume 2021, Article ID 5547237, 8 pages
https://doi.org/10.1155/2021/5547237

Research Article

An Empirical Investigation of the Challenges of Cloud-Based ERP Adoption in Pakistani SMEs

Mujtaba Awan,[1] **Niamat Ullah**,[2] **Sikandar Ali**,[3,4] **Irshad Ahmed Abbasi**,[5] **Muhammad Shabbir Hassan**,[1] **Hizbullah Khattak**,[6] **and Jiwei Huang**[3,4]

[1]*Department of Software Engineering, Riphah International University Islamabad, Rawalpindi, Pakistan*
[2]*Department of Computer Science, University of Buner, Buner 17290, Pakistan*
[3]*Department of Computer Science and Technology, China University of Petroleum-Beijing, Beijing 102249, China*
[4]*Beijing Key Lab of Petroleum Data Mining, China University of Petroleum-Beijing, Beijing 102249, China*
[5]*Department of Computer Science, Faculty of Science and Arts, Belgarn, P.O. Box 60, Sabt Al-Alaya 61985, University of Bisha, Saudi Arabia*
[6]*Department of Information Technology, Hazara University Mansehra, Khyber Pakhtunkhwa, Pakistan*

Correspondence should be addressed to Sikandar Ali; sikandar@cup.edu.cn and Jiwei Huang; huangjw@cup.edu.cn

Received 14 February 2021; Revised 21 March 2021; Accepted 23 March 2021; Published 1 April 2021

Academic Editor: Habib Ullah Khan

Cloud-based ERP solutions offer many benefits to small and medium enterprises (SMEs) and help them to integrate their activities, such as improve communications and reduce operational and maintenance costs. Primarily, it was only adopted by large organizations, but now SMEs are also keen on adoption. However, the motivation regarding the adoption of these systems in SMEs is relatively low in developing countries. This fact urges us to investigate the challenges faced by Pakistani SMEs. A qualitative research approach along with unstructured interviews was conducted by means of face to face. Interview methods are used to extract understanding, opinions, and challenges faced by SMEs on their way to adopt the cloud-based ERP system. The data were collected from eight well-reputed organizations, directly involved in the adoption. The study found ten (10) themes that are reluctant to adopt cloud ERP among Pakistani SMEs. The main benefit of these themes is to provide results that can be easily accessible to enterprises who want to adopt a cloud-based ERP. This can also contribute to the lack of the literature of cloud ERP and delivers insight for future study by practitioners and researchers.

1. Introduction

The most significant and impacting counterpart of the world economy is the business industry, majority of which consists of small and medium enterprises (SMEs). These enterprises are the key enabler for socio-economic development, job creation, poverty reduction, entrepreneurship, and rural development. SME's business environment has changed significantly in recent years for customer satisfaction and market flexibility. Various measures are taken by organizations to meet these challenges such as business model innovation, improved customer service, job automation as well as the development of the IT system such as enterprise resource planning (ERP) [1, 2]. ERP provides a safe strategy for SMEs in terms of organization operation processes and data management [3]. It is an integrated system known for improving processes and product quality, reducing production cycle time, and enhancing the decision-making process [4]. Despite the recognized significant benefits, traditional ERP is mostly a costly resource [5]. Cloud technology made ERP routed on a cloud platform with lower cost, scalability, and resource sharing [6]. Cloud ERP has grown exponentially in the last few decades due to the new delivery model for ERP, providing various advantages to organizations [7].

In contrast to traditional ERP, cloud ERP offers reduced capital expenditure, rapid execution, increased platforms, and improved storage and data processing. Cloud ERP has

become an interesting research area among researchers and experts due to its potential advantages to both SMEs and large enterprises [8–10].

2. Literature Review

Previous studies have addressed several perspectives while adopting cloud ERP software by SMEs. These include the economical perspective, technological perspective, and the people perspective, which are, respectively, dealt with financial issues, software evaluation, and its effect on people within the organization.

Studies conducted earlier exclusively revealed benefits in implementing cloud ERP in SMEs, such as efficient business processes, real-time access, visibility, accuracy of the information, and effective information management [11, 12].

Seethamraju examined the potential determinants using a cross-sectional field study. These include the low total cost of ownership, low investment cost, vendor participation in value cocreation, and product improvement [13].

Salim applied a content analysis method through which they identified 27 transition factors that contribute to the adoption of cloud-based ERP. The study aims to explore which transition factors are important to the different stages of the adoption for the organization to progress to the next level, and these transition factors are defined as "necessary" or "sufficient," where "necessary" transition factors need to occur, while "sufficient" means aiding in the movement [14].

The study reported by Peng et al. explored potential benefits and barriers associated with the adoption of cloud-based ERPs. A set of in-depth interviews are conducted with 16 enterprises and consultants. The results, derived from thematic analysis, showed that while the economic and technological benefits offered by cloud vendors are appealing, the success of cloud-based ERP adoption can be affected by critical challenges related to diverse organizational factors as well as with current legal and technical complexity in the cloud environment [15].

Albar and Hoque conducted ERP investigations in Saudi Arabian context utilizing the TOE framework and DOI theory. Factors identified include the competitive environment, complexity, observability, relative advantage, ICT infrastructure, regulatory environment, ICT skill, and top management support [16].

The practicality of cloud ERP systems for SMEs was further investigated by Zadeh et al. conducting a case study of the USA food industry. Identified benefits include increased performance, agility, flexibility, and cost savings. They also identified that vendor lock-in, security, and compliance issues are the major risk in this practice over the long run [17].

Usman et al. developed the TOE framework and DOI theory to explore determinant factors for cloud ERP adoption in the manufacturing sector of Nigeria. The most influential factors affecting Nigerian SMEs are compatibility, cost savings, lack of data security, competitor's pressure, regulatory support, and cloud ERP knowledge [18].

Ali et al. conducted a study which guided software development organization for cloud-based testing adoption. The study explores determinants and predictors of cloud adoption for software testing, and a list of predictors (main criteria) and factors (subcriteria) are identified using SLR approach. Seventy subcriteria are identified in this study and also known as influential factors [19].

Moh'd Anwer investigated the main logistical factors that are having impacts on the cloud ERP adoption among SMEs in developing economies of various countries. Overall, 14 factors such as relative advantage, security concerns, compatibility, complexity, value creation, technology readiness, technical barriers, enterprise readiness (ER), enterprise size (ES), enterprise status, top management support (TMS), competitive advantage, government support, and infrastructure/telecommunication were identified using logistic regression analysis [20].

Tongsuksai et al. published an SLR report from 81 articles on cloud ERP implementation. They investigated 32 critical success factors as well as an integrative model based on the organizational, environmental, technological, and individual characteristics. The identified CSFs and factors offer more clarity to (IT) practitioners and help organizations to achieve successful implementation of cloud ERP systems [21].

Ahn and Ahn conducted a comprehensive analysis using DOI theory and TOE frameworks, identifying important influencing factors like trial ability, vendor lock-in, regulatory environment, and relative advantage. This study's findings provide meaningful guidance for cloud-based ERP adoption by companies, agencies supporting enterprise digitalization, and cloud ERP vendors [22].

Ali and Li proposed a study that aims at developing a technology acceptance model. The model provides decision support to software development organizations in various predictors and determinants that will guide organization towards cloud adoption for software testing [23].

The implementation of cloud ERP in business enterprises is a relatively new phenomenon in terms of its influencing factors, inhibitors, and organizational determinants that affect its adoption and management. Understanding the determinants that encounter in the cloud ERP adoption in SMEs helps these organizations to accomplish better results upon their IT investments. Citing the previous literature, it becomes obvious that the main focus of the research in this area is the determinations of challenges and benefits of cloud ERP and factors affecting its adoption in small and medium enterprises. This results in exploring the fact that SMEs are not much prompt to adopt a cloud ERP system. Major investigation upon the low rate of cloud ERP adoption in this side of the business industry has been conducted in the context of developed countries [14].

Literature survey revealed that minimal to no research has been done to inspect the factors influencing the adoption of cloud ERP in developing countries like Pakistan. The present investigations aimed to address the major constraints and challenges faced by Pakistani SMEs while adopting cloud ERP systems. The acceptance and interest rate in the cloud ERP among SMEs are relatively slow and discouraging because of the lack of success stories. Hence, there is a need to identify the factors and determinants as well as other critical influencing factors [24–26].

The business practices embedded in the ERP system most likely reflected the US and European organization and their national culture. Upon implementation of such systems in Asian countries, problems may occur due to a mismatch between the cultural assumption and practices embedded in the software. Therefore, the investigations in this area of research will identify the factors that affect the adoption of cloud ERP systems in Pakistani SMEs, which may help the business industry to overcome the major challenges faced in implementing cloud technologies [27].

3. Methodology

This section describes the descriptions of the research methodology of the study that includes research questions, research method, data collection techniques, validation, and the detailed processes of the research.

3.1. Research Process. The first level of the research study was the literature review, conducted to understand the available information and data presented by previous researchers. Being an innovative research area, the cloud ERP was previously studied as a cloud and ERP under the titles separately. The second level of the present qualitative study was to conduct interviews from various organizations in Pakistan working on the said concept and was to evaluate different problems faced while adopting a cloud ERP system. The third level of the research was to present the data finding and data analysis [28].

3.2. Data Processing. The major objective of the current study was to investigate the constraints that impact the small and medium enterprises in adopting cloud ERP. The qualitative research methodology was utilized to accomplish the targeted objective as the literature review revealed that cloud ERP is still an innovative and unexplored area of the research as well as very limited empirical studies have been done. The selected methodology also helps to facilitate the researcher towards people understandings regarding the beliefs, behaviors, experience, social, and cultural experiences within the community [29]. Additionally, the proposed research idea aims to explore the field deeply, practicing in the small and medium organizations of Pakistan.

3.3. Data Collection. Data collection is one of the central parts of every research methodology, and hence it is necessary for the researchers to evaluate the previous information carefully and to adopt an appropriate strategy for data collection regarding the available facts. Cloud ERP being an unexplored area technology, the nature of the proposed study was quite complex, requiring a deep understanding of the problem area. For this reason, we decide to collect our empirical data utilizing the concept of the qualitative interview. This method of data collection is one of the professional conversations with precise structure and purpose in which two parties discuss a theme of mutual interest. It is through the interaction between the researcher and the respondent that knowledge is produced [30].

The unstructured interview type was selected to gather the necessary information about the organizations to collect the empirical data. This allowed asking a set of predefined open-ended questions regarding each theme to which the participants could provide clear answers. This approach is quite flexible that did not require any interview guide to follow. It also allows to change the order of the questions to critically follow-up on respondent's answers as well as the further probe and inquire through additional questions. We could therefore keep the interviews open without losing control or direction [31].

3.4. Data Analysis. The process of data analysis starts with transcribing the recorded interviews and translating them into phrases or text to analyze them using the content analysis technique [32]. Qualitative content analysis generally uses individual themes that might be conveyed in a single word, a phrase, a sentence, a paragraph, or an entire document. This technique helps to develop an understanding of collecting data and delivers the ability to test theoretical issues. In the following section, an analysis of the collected data from the Pakistani SMEs on cloud ERP system is presented and discussed. This analytical study emerged ten (10) most important themes which are discussed in the next section in detail.

4. Results and Discussion

This section presents the empirical data collected through interviews with various participants to identify the challenges, faced by an organization adopting the cloud ERP solution. The pie chart in Figure 1 shows the overall correlation between consumer concern and identified challenges and the frequency of each challenge/barrier faced by participants within their enterprises. The final discussion of the collected data indicated the suitability and aptness of cloud ERP in Pakistani small and medium organizations as well as future perspectives of the current research study.

4.1. Enterprises. Participants from 8 different organizations who contributed to this research are presented in Table 1.

4.2. Challenges Faced by SMEs. Challenges were identified through the analysis of recorded data. We listen to each interview very carefully again and again and note down important answers and claims by all participants; later on, all these points and claims are translated into common text/labels and stored in the word document. There were 10 challenges extracted from all participants during the interview session, the percentages and frequencies of the extracted challenges were established and are shown in Figure 1, and the challenges along with their frequencies are shown graphically.

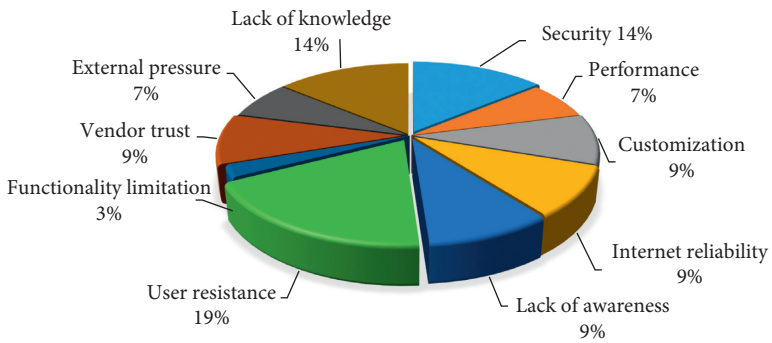


Figure 1: Pie chart showing a correlation between consumer concern and identified challenges.

Table 1: Overview of conducting interviews.

S.#	Name	Size	Position	Business location	Duration (min)
1	Participant 1	200	Manger products	National	60
2	Participant 2	20	Director, IT	National	55
3	Participant 3	20	Business analyst	National	45
4	Participant 4	100	Manager	International	50
5	Participant 5	30	Lead	National	40
6	Participant 6	250	IT manager	National	60
7	Participant 7	200	Manager	National	35
8	Participant 8	50	Manager	National	30

4.2.1. Security Risk. Security risk was found to be the most common challenge to the adoption of cloud-based ERP around 14% of all the interviewees. Security is one of the core concerns, mentioned by most of our participants during the interview session. Perceived security is revealed mostly as a psychological issue aided by a lack of knowledge and trust. The participants mentioned that their organization and many other organizations expressed their major concerns about their data which is stored off-premises. This practice gives them a big sense of insecurity as well as a lack of control over their company data. Most organizations believe that their data can only be safe in their place and therefore are much reluctant to move toward cloud technology. They worried that their data may be stolen or leaked out during this implementation. Participant P2 of the interview session stated that "...Organization feel insecure about an information disclosure which company keeps hidden from the competitors in the market. Financial information is most of the stored data of any organization and is considered to be the heart of the company. Therefore, it is difficult for any organization to trust the 3rd party." Participants P3 and P6 mentioned that "... organizations are worried whether the cloud service provider will properly run the services or not." Furthermore, according to participant P5, "...security threat is not real for SMEs but just psychological issue. A large organization may have a concern about the real security threat while moving their data to the cloud." Participant P8 mentioned the cost problem, saying that "...ensuring the security is a responsibility of cloud service provider they charged for the service."

According to various participants, security concerns have more to do with human behavior and psychology rather than a real organizational issue. Participant P1 revealed that the "...security concerns that make organization reluctant about data transfer to the cloud, are mentality and psychological issues, it is not because of there is some kind of security problems but actually, the perception is negative, and they feel fear of threat." Participant P6 stated that the security of the cloud ERP system varies from vendor to vendor. Not all vendor provides sufficient security measures, not all vendors are ISO certified and not even made enough level of investment on the infrastructure. Furthermore, Participant P8 mentioned that "... the whole ERP security part deal with the service legal agreement (SLA). Many providers don't offer consistent and reliable services on security." He further explored the main reason for the data security issue is the lack of knowledge in this field and also not sufficient SLA, which makes cloud ERP fail.

From all these arguments and claims, it is concluded that the organizations are not much confident to adopt cloud ERP to secure their data. Some organizations, although adopted this technology but have concerns about data security as well as lack of control over data, which is still questioning by the organizations.

4.2.2. Performance. Performance was discussed in 7% of the selected interviews. Cloud ERP is one of the most online technology, requiring high-speed Internet sources. Participants P1, P3, and P6 indicated the performance of this system is the major challenge faced by cloud ERP implementation, accessed via the Internet. According to participant P1 "...delay in response time come because cloud ERP system run with a lot of processes as well as a huge amount of data to process online." This fact makes the cloud ERP much dependent on Internet availability and speed, suffering its

performance. Further, he questioned that "... Is SQL server separate from application server? These resources directly impact application performance." You need to work with your IT team and cloud provider to ensure the servers running your applications have the proper resources for the number of users using that software. Participant P3 stated that ... "it is important to ensure that the activities and processes involved in cloud ERP would be able to work online efficiently. This fact can only be worked out if the cloud ERP system is made on modern architectures and the framework." He further pointed out that it is not necessary that all vendors made the same investment in the software to work over the Internet efficiently. The online efficiency of a system can only be increased if the system is based on modern technologies. All these problems make any organization reluctant to adopt ERP solutions. Participant P6 also added to this issue mentioning that "... the performance of ERP based system may depend on various factors, but the most important factor is the Internet connectivity because poor database performance slows down the whole system."

4.2.3. Customization. Customization is also a barrier towards the adoption of cloud-based ERP systems among the participants and reported 9%. Customization being an issue for the cloud ERP system has been concisely mentioned by participants P1, P5, P6, and P7. According to Participant P1 and P6 "... although cloud-based ERPs offer many technological advantages, their customization is limited. In this case, if the customization is done, it charges the organization with a huge cost, affiliated with that specific change." Some vendors also allow their users to make changes and customize the system. In the case of small and medium enterprises, this practice is not much applicable as they do not have many resources to customize the system on their own because of a lack of skill and knowledge. As ask a question on the customization, according to participant P5, "...customization is important concerns which can be a barrier, organization need, and requirements change time by time you need a system that can easily be customized and not take much money on it. I suggest SME who can bear the cost at one time and keep relax will go for a hybrid model." The hybrid model allows organizations to protect their most valuable data on their terms; they are the ones who decide where important data are stored, and how that data are protected from external and internal threats. It puts control back in the best hands. In contrast to the participant (P7), "... When we said vendor to customize the solution then first they said that the system which we build is a generic system in which we apply best business practices according to the business needs and requirements, initially, we said ok when the system was deployed and accessed the needs and requirements are not too much similar as we doing before, because of not properly work on the needs and requirements of the organization into the meeting taken into the account with the vendor again this is happened because of lacking in knowledge." The customization of cloud ERP is limited but some of the vendors allow their clients if they want to customize the system, and they have the freedom to do but lack of skilled peoples brings the barriers to avail of that opportunity.

4.2.4. Internet Reliability. Internet reliability has been also reported by 9% of most of the interviewed participants. According to P1, P3, and P8, the cloud-based ERP mostly depends on the Internet in developing countries the reliability of Internet connectivity is a question. As asked from participants P3 and P5, the infrastructure of the network is not well facilitate as compare to developed countries, then people blame the system itself as slow whether then the problem lies with the Internet connectivity. In contrast to participants P1 and P8, "...traditional ERP does not rely on the Internet reliability but when we moved to cloud platform there will be a huge dependency over the Internet connectivity, so for this concern, we installed multiple Internet connections to cater." From all the above arguments, we believe that in Pakistan, there is a reliable solution to the Internet reliability organization which does not only rely on a single Internet service provider but also cloud ERP vendors can come with a different and better solution.

4.2.5. Lack of Awareness. Around 9% of the participants mentioned lack of awareness as a significant barrier faced by an organization while adopting the cloud-based ERP. Participant P2 mentioned that "... How will we adapt if we do not know anything about cloud ERP so awareness is a challenge in Pakistani industry due to the lacking knowledge how to adopt system" and participant P4 declared that "...In Pakistan and our organization, we do not have much quality in education and have skilled people and knowledge about cloud ERP only a person in our organization who have little knowledge about this technology, Before moving to cloud ERP system one should be studied about the cloud ERP system, basic cloud knowledge is essential for the smoother adoption for the organization like SMEs, because they don't have much margin to make this adoption fail, this takes SMEs back." According to participant P5, "... Lack of awareness is an issue in Pakistan not everyone known about cloud ERP, IT training will be given to the head of the departments which will remove the hesitation to interact with the system" in addition to participant P3, "... We are not aware of cloud ERP our competitor pushes us to be aware of cloud ERP after that we are familiar with this new technology." From all concerns mentioned by participants how SMEs force to move towards enterprise solutions if they have no enough knowledge, enterprises are not well-informed benefits of cloud ERP, and they afraid to adopt due to negative perceptions. Therefore, a lack of awareness and knowledge left SMEs towards adoption.

4.2.6. In-House Resistance. User resistance is a challenge to the adoption of a cloud-based ERP system by organizations, approximately 19% of the selected participants. When an organization moves towards a cloud solution, in-house resistance to change from internal staff act like a suspending

factor. Resistance to change is a natural thing that comes from internal staff because of not taking into consideration the selection process of the cloud ERP. Participant P1, P6, and P8 expressed the same "... When we shifted to a cloud ERP solution, there will be resistance from staff about their job security"; this is just because much IT-related work will be shifted to cloud ERP provider and employee feel the treat of their job. Participants P4, P3, and P5 said that "...When a new system was adopted in the organization resistance come from the internal staff because they are lacking in knowledge and skills. And I want to add one thing very important don't keep an employee in the dark, communication with the employee in the workplace is always essential." When an organization is willing to adopt cloud-based ERP system, resistance occurred from the staff members, but this can be resolved by providing training to the employee and should be taken into consideration while adopting new technologies. In contrast to participants P2, P6, and P7 "... Resistance will come from staff end, they oppose to adopt, athletic training to the employee will remove uncertainty and after that, they are more comfortable to use such system more efficiently." Therefore, it appears that in-house resistance will come up from all our interviewee comments and claims resistance to change is a natural thing but can be resolved by taking employees into the selection process of ERP and provide some training to the employee by the organization.

4.2.7. Functionality Limitation. Functionality limitation was the least discussed factor and reported only 3% of the participants. Cloud ERP is a newbie in the world of technology, and enterprises still arise questions on its stability, reliability, and functionality. According to Participant P2 "... Cloud ERP are ready-made solutions, sometimes enterprise business process not mapped so the organization revised their business processes accordingly to that software" and added that further "... The ability to perform tasks and functionality still fear for many enterprises. Several enterprises are not very well aware and due to the lack of knowledge, they had a problem adopt smoothly. Cloud ERP functionality they think traditional ERP systems are old and mature, so time is taken to mature cloud ERP systems. However, some enterprises adopt cloud ERP and agreed their enterprises quickly shifted to the next level of advancement."

4.2.8. Vendor Trust. Vendor trust was discussed in 9% of the participant's end. Participants P2 and P3 mentioned their concern about the ability and capability of local vendors. Vendors must provide support to their clients, but not enough support is provided thus some organizations preferred to build their solution. As stated by participant P6, the relationship of trust not only based upon vendor reputation but also on the agreement between them. In contrast to participant P7, he believes that "... Vendor trust is an issue because local vendor in here is not much mature, so NDA will sign with a vendor to our data should not be compromised and this is the responsibility of vendor." From all the above comments and discussion, vendor trust is a challenge when an organization adopts a cloud ERP system due to immature local vendors in the market.

4.2.9. External Pressure. External pressure is also a challenge towards the adoption of cloud-based ERP among organizations and reported 7%. Previous studies on the adoption of technology define that external pressure is an indicator of adoption. In our study, when participants asked about the suggestion on the adoption of cloud ERP solution, P2, P3, and P6 believed that "... We cannot go to adopt new technology because of so many risk factors involved in it, when our competitors using the solution, so we are generally forced by different pressure to adopt new technologies" in addition to participant P6 "... In our surroundings technology is evolving very quickly and no other option to not adopt new technology, and especially when competitors use that technology, so it's become necessary to adopt due to competitive pressure." All participants during the interview agreed that they are forced to adopt new technologies by their competitor.

4.2.10. Lack of Knowledge. Lack of knowledge is a barrier towards the smooth adoption of cloud ERP solutions and reported 14% from the participants. As stated by P1 and P6 "...Enterprise who decide to shift towards cloud may hinder while adoption because of less experience and knowledge." As declared by P3 and P4 that "...we will be in trouble because we don't have enough knowledge about the software." Interviewee P5 "...Told us initially we are not aware of the cloud technology that is why we faced some challenges like security, customization, resistance, etc., lacking in knowledge is a big challenge and due to this we faced several challenges in adopting phase."

5. Conclusion and Future Work

SMEs are showing great interest in cloud ERP technology and are considered the best alternative rather than the traditional system; thus, it is essential to understand the challenges that can negatively impact the adoption in SMEs. In this regard, we identify ten (10) possible challenges of cloud ERP in the Pakistani context which influence the adoption decision. We used qualitative methodology along with unstructured interviews for extracting the challenges; and data collected from participants were successfully analyzed to ascertain the specific challenges. The study found challenges such as data security, customization limitation, external pressure, awareness, resistance to change, vendor competence, and lack of knowledge are the main problems to be mentioned when making a decision either to adopt or reject the cloud ERP systems. The reported work in this paper contributes to the area of information system adoption and especially in the domain of cloud ERP by identifying the challenges which impact small and medium enterprises. The research on cloud ERP systems required more exploration when it comes to the implications of cloud ERP, and much existing work focused on SMEs and did not

interrogate whether the different size of enterprises relates to these implications. In this, we feel that we have at least begun to bridge the gap. Future research addresses this problem by comparing small- and large-sized enterprises and finds the difference in how they perceive cloud-based ERP.

Data Availability

The data collected during the data collection phase are available from the corresponding authors upon request.

Conflicts of Interest

The authors declare that they have no potential conflicts of interest.

Acknowledgments

This study was supported by the National Key Research and Development Plan (no. 2016YFC0303700), National Natural Science Foundation of China (no. 61972414), Beijing Natural Science Foundation (no. 4202066), Beijing Nova Program (no. Z201100006820082), and Fundamental Research Funds for Central Universities (nos. 2462020YJRC001 and 2462018YJRC040).

References

[1] K. H. Salum and M. Rozan, "Barriers and drivers in cloud ERP adoption among SMEs," *Journal of Information Systems Research and Innovation*, vol. 9, no. 1, pp. 9–20, 2015.

[2] T. N. Mahara, "Indian SMEs perspective for election of ERP in cloud," *Journal of International Technology and Information Management*, vol. 22, no. 1, p. 5, 2013.

[3] S. S. Shahawai and R. Idrus, "Malaysian SMEs perspective on factors affecting ERP system adoption," in *Proceedings of the 2011 Fifth Asia Modelling Symposium*, pp. 109–113, IEEE, Kuala Lumpur, Malaysia, May 2011.

[4] H. Klaus, M. Rosemann, and G. G. Gable, "What is ERP?" *Information Systems Frontiers*, vol. 2, no. 2, pp. 141–162, 2000.

[5] S. Fuller and T. S. McLaren, "Analyzing enterprise systems delivery modes for small and medium enterprises," in *Proceedings of the 16th Americas Conference on Information Systems, AMCIS 2010*, p. 380, Lima, Peru, August 2010.

[6] M. Armbrust, A. Fox, R. Griffith et al., "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[7] J. Duan, P. Faker, A. Fesak, and T. Stuart, "Benefits and drawbacks of cloud-based versus traditional ERP systems," *Proceedings of the 2012-13 course on Advanced Resource Planning*, 2013.

[8] M. S. Amalnick, A. Ansarinejad, S.-M. Nargesi, and S. Taheri, "New perspective to ERP critical success factors priorities and causal relations under fuzzy environment," *Journal of Mathematics and Computer Science*, vol. 2, no. 1, pp. 160–170, 2011.

[9] H. U. Khan and H. S. I. A. Samad, "Enterprise strategic shift of technology: cloud-based systems verses traditional distributed system," *International Journal of Enterprise Network Management*, vol. 11, no. 4, pp. 320–346, 2020.

[10] A. Tahir, F. Chen, H. U. Khan et al., "A systematic review on cloud storage mechanisms concerning e-healthcare systems," *Sensors*, vol. 20, no. 18, p. 5392, 2020.

[11] R. Seethamraju and J. Seethamraju, "Adoption of ERPs in a medium-sized enterprise-a case study," *ACIS 2008 Proceedings*, vol. 104, Portland, OR, USA, May 2008.

[12] H. Samad and H. Khan, "Adoption of cloud in enterprises environment," in *Proceedings of the 2017 DSI Annual Meeting*, pp. 18–20, Washington, DC, USA, November 2017.

[13] R. Seethamraju, "Adoption of software as a service (SaaS) enterprise resource planning (ERP) systems in small and medium sized enterprises (SMEs)," *Information Systems Frontiers*, vol. 17, no. 3, pp. 475–492, 2015.

[14] S. A. Salim, "Cloud ERP adoption-a process view approach," in *Proceedings of the PACIS*, p. 281, Jeju Island, Korea, June 2013.

[15] G. C. A. Peng and C. Gala, "Cloud ERP: a new dilemma to modern organisations?" *Journal of Computer Information Systems*, vol. 54, no. 4, pp. 22–30, 2014.

[16] A. M. Albar and M. R. Hoque, "Factors affecting cloud ERP adoption in Saudi Arabia: an empirical study," *Information Development*, vol. 35, no. 1, pp. 150–164, 2019.

[17] A. H. Zadeh, B. A. Akinyemi, A. Jeyaraj, and H. M. Zolbanin, "Cloud ERP systems for small-and-medium enterprises," *Journal of Cases on Information Technology*, vol. 20, no. 4, pp. 53–70, 2018.

[18] U. M. Z. Usman, M. N. Ahmad, and N. H. Zakaria, "The determinants of adoption of cloud-based ERP of Nigerian's SMES manufacturing sector using toe framework and doi theory," *International Journal of Enterprise Information Systems*, vol. 15, no. 3, pp. 27–43, 2019.

[19] S. Ali, N. Ullah, M. F. Abrar, Z. Yang, and J. Huang, "Fuzzy multicriteria decision-making approach for measuring the possibility of cloud adoption for software testing," *Scientific Programming*, vol. 2020, Article ID 6597316, 24 pages, 2020.

[20] A.-S. Moh'd Anwer, "Towards better understanding of determinants logistical factors in SMEs for cloud ERP adoption in developing economies," *Business Process Management Journal*, vol. 25, no. 5, pp. 887–907, 2019.

[21] S. Tongsuksai, S. Mathrani, and N. Taskin, "Cloud enterprise resource planning implementation: a systematic literature review of critical success factors," in *Proceedings of the 2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pp. 1–8, IEEE, December 2019, Melbourne, Australia.

[22] B. Ahn and H. Ahn, "Factors affecting intention to adopt cloud-based ERP from a comprehensive approach," *Sustainability*, vol. 12, no. 16, p. 6426, 2020.

[23] S. Ali and H. Li, "Moving software testing to the cloud: an adoption assessment model based on fuzzy multi-attribute decision making algorithm," in *Proceedings of the 2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA)*, pp. 382–386, IEEE, Tokyo, Japan, April 2019.

[24] R. Small, "Factors affecting the adoption of enterprise resource planning (ERP) on cloud among small and medium enterprises (SMES) in Penang, Malaysia," *Journal of Theoretical and Applied Information Technology*, vol. 88, no. 3, 2016.

[25] E. O. Yeboah-Boateng and K. A. Essandoh, "Factors influencing the adoption of cloud computing by small and medium enterprises in developing economies," *International Journal of Emerging Science and Engineering*, vol. 2, no. 4, pp. 13–20, 2014.

[26] K. H. Salum and M. Rozan, "Exploring the challenge impacted SMEs to adopt cloud ERP," *Indian Journal of Science and Technology*, vol. 9, no. 45, pp. 1–8, 2016.

[27] J. Rajapakse and P. Seddon, "ERP adoption in developing countries in Asia: a cultural misfit," in *Proceedings of the 28th Information Systems Seminar in Scandinavia*, pp. 6–9, Kirstiansand, Norway, 2005.
[28] M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*, Pearson Education Limited, Harlow, UK, 2009.
[29] A. Alajbegovic, V. Alexopoulos, and A. Desalermos, Factors influencing cloud ERP adoption: a comparison between SMES and large companies, 2013.
[30] S. Kvale and S. Brinkmann, *Interviews: Learning the Craft of Qualitative Research Interviewing*, Sage, Thousand Oaks, CA, USA, 2009.
[31] M. Bengtsson, "How to plan and perform a qualitative study using content analysis," *NursingPlus Open*, vol. 2, pp. 8–14, 2016.
[32] V. Minichiello, R. Aroni, and V. Minichiello, *In-Depth Interviewing: Researching People*, Longman Cheshire, Harlow, UK, 1990.

Hindawi
Scientific Programming
Volume 2021, Article ID 6632450, 10 pages
https://doi.org/10.1155/2021/6632450

Research Article

A Spatiotemporal Change Detection Analysis of Coastline Data in Qingdao, East China

Muhammad Yasir,[1] Sheng Hui,[2] Zheng Hongxia,[2] Md Sakaouth Hossain,[3] Hong Fan,[4] Li Zhang,[5] and Zhao Jixiang[2]

[1]*School of Geosciences, China University of Petroleum Qingdao, Qingdao 266580, China*
[2]*College of Oceanography and Space Informatics, China University of Petroleum Qingdao, Qingdao 266580, China*
[3]*Department of Geological Sciences, Jahangirnagar University, Dhaka 1342, Bangladesh*
[4]*State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China*
[5]*Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China*

Correspondence should be addressed to Zheng Hongxia; zhenghongxia@upc.edu.cn

Received 22 October 2020; Revised 24 November 2020; Accepted 8 March 2021; Published 29 March 2021

Academic Editor: Habib Ullah Khan

This study focuses on the coastal features, environments, and dynamics to accurately describe and regularly monitor the Qingdao shoreline in eastern China. It collects categorical ETM+ and OLI data from 2000, 2010, and 2019 on the mainland coastline and explores the characteristics and spatiotemporal differences across the past 19 years by using remote sensing and geographic information system (GIS) technologies. The results show that the length of the Qingdao coastline has increased continuously over the last two decades, for a total increase of 18.14 km. There are different natural and artificial coastlines that have undergone major changes. The human-induced deterioration of coastlines has gradually and substantially risen from 53.63% in 2000 to 68.40% in 2019, while the length of the natural coastlines has decreased dramatically. Jiaozhou Bay focuses on areas with significantly changing coastlines, and major changes have occurred in the west and east of the Qingdao coast. The coastline has largely expanded seaward because of the comprehensive impact of natural and anthropogenic factors. The leading factor in coastal evolution is coastal engineering constructions. In addition, the top three other construction activities are the restoration of the aquaculture pond, salt field, and harbor edifices. The driving force that triggered the shift in the coastline reveals significant temporal heterogeneity.

1. Introduction

The coastline makes up the border between sea and land contact [1], a significant human survival and growth base and carrier, and a special natural resource [2]. Instability [3–7], functional diversity [8], geographic disparity [9, 10], and other significant features are included in these characteristics. Coastal countries have converted their economic growth centers into coastal areas since the twentieth century, and almost 50% of the world's population has settled in areas within 100 km of the coastline [11]. More emphasis is focused on coastal changes as they are major environmental factors that directly affect coastal economic growth and land management [12, 13]. The relocation of economic centers, however, has triggered rapid changes in coastal resources, which have had an immense effect on the economic, social, environmental, and ecological aspects of coastal areas [14]. For instance, the proportion of natural coastlines has been significantly reduced by large-scale reclamation projects [15]. The biological ecosystem of coastal areas has been degraded by the intensive use of coastlines, artificial correction of naturally curved coastlines, and disorderly aquaculture. The coastal zone has become a crucial region for coastal economic growth with the continual development

of marine resources. An efficient way to research the environmental and ecological changes of the coastal zone is to track coastline changes. Nayak [16] suggested that, for understanding different coastal processes, detailed demarcation and monitoring of shorelines (long-term, short-term, and seasonal) are important. Coastal monitoring has therefore become an important activity for sustainable development and the conservation of the environment [17, 18].

Because of their high coverage and low cost, remote sensing (RS) image-based methods have now become popular for monitoring coastline changes [19–25]. Many scientists have documented changes to the coastline, and the causes of these changes have been investigated and recognized [26]. For instance, through the use of open-access multitemporal satellite imagery, Mishra et al. [27] evaluated the long-term to short-term dynamics of the coastal position of the Uri district in India over the past 25 years (1990–2015), deriving the explanation behind the changes such as human buildings and coastline erosion. Kale et al. [28] researched the erosion rates and coastal variations on the Yesilirmak River on the northern coast of Turkey before 2017 to clearly demonstrate the effect of dam buildings on the data. Thoai et al. [29] studied the changes in the coastline of the Ca Mau Cape in Vietnam over the past 20 years and concluded that forest area loss, river dredging, and aquaculture and infrastructure growth were the most significant factors influencing the changes in the coastline in the region. Wu et al. [30] took China's Shenzhen Special Economic Zone as an example to study coastline changes in the region from 1988 to 2015 and found that coastline stability characteristics were entirely different between the eastern and western coastlines of Shenzhen, with regional differences primarily reflected in morphological changes and changing coastline laws.

Researchers around the world performed their studies mainly from two viewpoints on coastline changes: (i) measuring rates of coastline shift and shifting areas of land or sea to describe coastline spatiotemporal variability [31–36]; (ii) based on a study of spatiotemporal trends of coastal changes, the exploration of the impact of geology, geomorphology, climate, and economy on coastline changes [37–40]. Many research studies on improvements to coastlines have been performed. However, the coastline derived from satellite images by several processing methods [24, 41–43] is the land-sea boundary that existed at the particular time of the acquisition of the image, namely, the instantaneous coastline [44, 45], rather than the actual geographical description of the coastline [46–48]. It is important to transform this instantaneous coastline into a tide-coordinate coastline [44]. Unless adequate ground control and photogrammetric coverage are available, this approach has restricted use for historical coastline determination [49].

In different countries, the use of remote sensing and geographic information system technologies to track changes in the coastline has become the focus of much scholarly study. The integration of the two technologies is seldom explored in many studies. The challenge of collecting data over a long-term period is one explanation for this. Other explanations are that several extraction rules have been used to extract the coastline for remote sensing details, and the extraction process is complicated. Most researchers study the coastline, or coastal region, and its driving powers. Therefore, based on five remote sensing images of Qingdao from 2000 to 2019, this study uses a visual interpretation method to extract remote sensing information of the coastline. Quantitative analysis of the Qingdao coastal long time series of different spatiotemporal evolution characteristics and response relationship is conducted. The research findings can provide basic data for the production and use of coastal protection and coastal zone and are of great importance for solving the coastal zone health ecosystem problems.

1.1. Study Area. Qingdao is in the southeast of the Shandong Peninsula, at 119°30′–121°00′ east longitude, 35°35′–37°09′ north latitude, bordering the Yellow Sea in the east and south, next to Yantai city in the northeast, bordering Weifang city in the west, and bordering Rizhao city in the southwest (Figure 1), a total area of 11,282 km^2 [50].

2. Materials and Methods

2.1. Database. Landsat multitemporal satellite data (Operational Land Imager-OLI and Enhanced Thematic Mapper Plus-ETM+ sensors) with a spatial resolution of 30 m are used in the research work to cover the area of the study in 2000, 2010, and 2019. The images were collected from the EarthExplorer website (http://earthexplorer.usgs.gov) of the US Geological Survey (USGS), all of which were rectified and projected with a geographical error of 0.5 pixels using the Universal Transverse Mercator method in the World Reference System (WGS84) results. Table 1 depicts the details regarding the data.

2.2. Extraction and Classification of Qingdao Coastline. First, in this research paper, the preprocessing steps, including atmospheric correction and radiometric correction stages, have been applied to satellite images using ENVI 5.3 software. ArcGIS 10.5 was used for the classification of coastlines. The method of coastline interpretation includes visual interpretation and automated remote sensing image interpretation [51]. Methods of automatic analysis, such as the edge detection algorithm, the mathematical morphology algorithm, the region-growing algorithm, the data mining method, and the neural network method proposed in recent years, are widely used around the world, but conventional methods, such as the visual interpretation of human-computer interaction, are still widely used by experts to extract coastal information [52]. The method of human-computer integrated visual perception was therefore adopted in this article. ENVI 5.3 [53] software was used to preprocess the remote sensing images, including hypercube fast line-of-sight atmospheric analysis (FLAASH), radiometric correction, image registration, and different scene image mosaics. Most of the images were corrected systematically and geometrically. After preprocessing, the remote sensing images were separated by the normalized difference water index (NDWI) [54], the boundary line was enhanced, and the exact location of the instantaneous water edge was

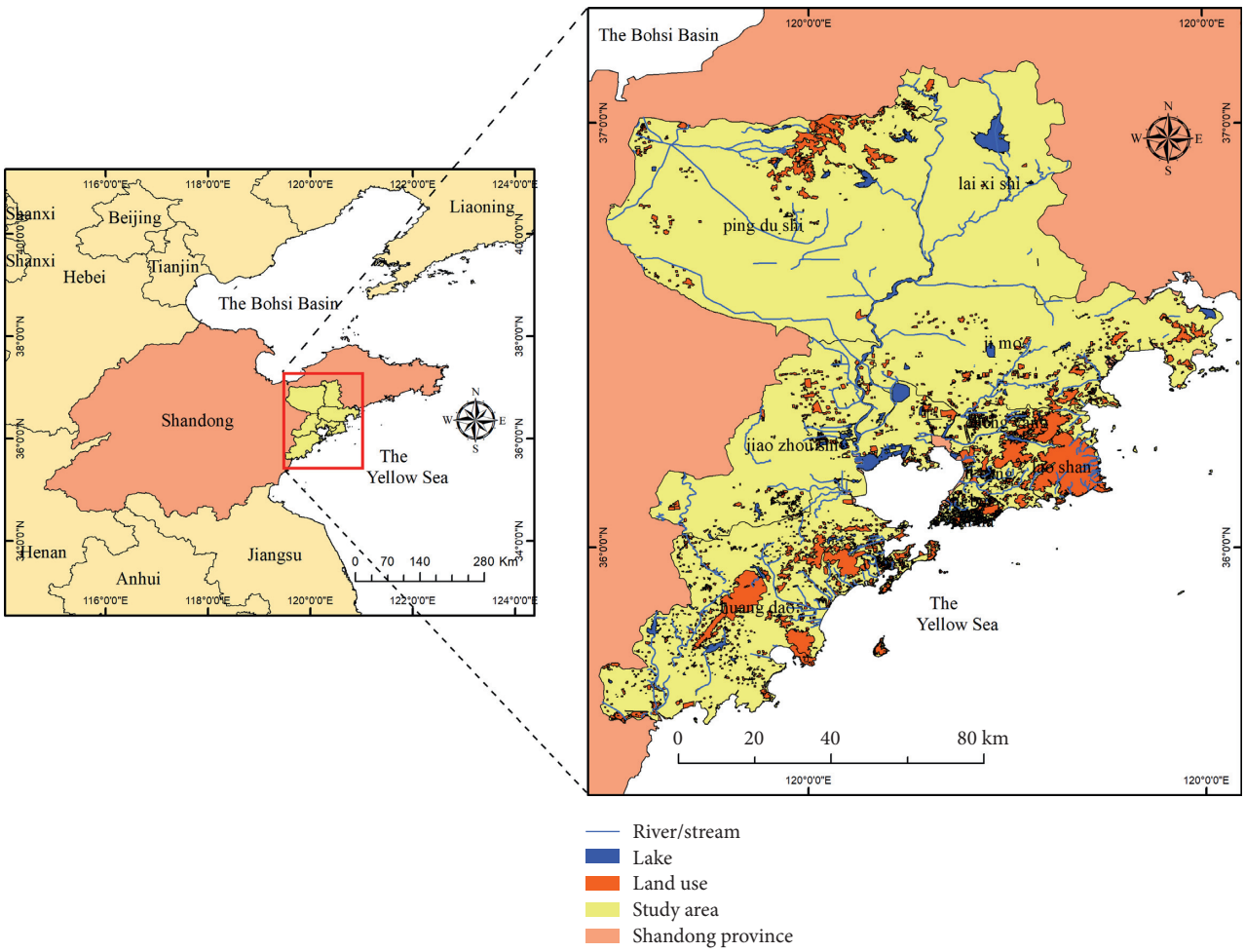


Figure 1: Geographical location map of Qingdao and shoreline, East China.

Table 1: Data source and its specifications.

Year	Satellite	Sensor	Path/row	Date	Resolution (m)	Cloud	No. of bands	Format
2000	Landsat_7	ETM+	120/35	08/9/2000	30	0	7	Geo TIF
2010	Landsat_7	ETM+	120/35	29/10/2010	30	0	7	Geo TIF
2019	Landsat_8	OLI	120/35	29/9/2019	30	0	11	Geo TIF

obtained by Otsu's threshold segmentation method based on the clarification of the characteristics of the differential reflection spectrum near the coastline [55]. The classification of the coastline acquired was mainly based on information gained from Google Earth images and field surveys; subsequently, topographical map digitization and visual analysis of remote sensing images were categorized. The coastlines were divided into coastlines on the island and coastlines on the mainland, and we only examined the mainland coastlines in this article.

2.3. Accuracy Validation of Extracted Coastlines. To test the accuracy of the extracted coastline, the statistical method of manually choosing random points was used [56]. First, on the initial image along the edge of the coastline, we manually picked 900 points. For each data cycle, we selected 900 points, for 2700 points for accuracy evaluation. Then, from each random point to the extracted coastline, we determined the shortest distance. The distance value was positive if the random point was within the land; otherwise, it was negative. The image resolution of TM/ETM+ and OLI was 30 m (Table 1). Therefore, according to the statistical results of the histogram (Figure 2), the proportions of random points chosen in 2000 within a one-pixel distance were 87.55% in the three cycles of 2000, 2010, and 2019, 90.78% in 2010 within a one-pixel distance, and 94.45% in 2019 within a one-pixel distance, and the extraction results were satisfactory.

3. Results

3.1. Classification of Coastlines. The distribution of coastal types for the years 2000, 2010, and 2019 was obtained through the acquisition and subsequent analysis of images of

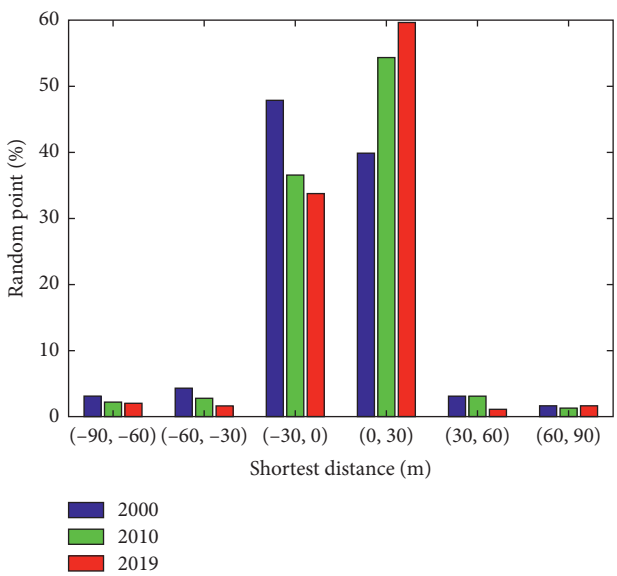


Figure 2: Histogram of randomly selected reference points (%) and the shortest distances (m) to the extracted coastlines.

the coastal areas of Qingdao covered by remote sensing satellites over the past 19 years. The Qingdao coastline was split into natural coastlines and artificial coastlines based on detailed research findings [3, 57–60]. The natural coastlines were divided into bedrock coastline, sandy coastline, muddy coastline, and estuaries; furthermore, the artificial coastlines were divided into harbor and wharf, revetment and seawall, aquaculture, and salt field. On the Landsat images, the corresponding coastline was shown in 543, 432, and 321 bands combined with false color displays (Figure 3), and finally, the coastline classification was achieved through interactive visual representation between humans and computers. Detailed classification information is shown in Table 2.

According to the statistics of the coastline classification results shown in Table 3 and Figure 4 the total length of Qingdao's coastline in 2000 was 507 km, of which the total length of the natural coastline was 235.22 km, and the total length of the artificial coastline was 272.34 km. As shown in Table 3, from 2000 to 2019, because of the comprehensive effects of natural and anthropogenic factors, the total length of coastlines in Qingdao increased by 4.35% in which the rates of change in 2000–2010 and 2010–2019 were 2.01 and 2.29, respectively, with a trend of accelerating growth. Anthropogenic variables influenced the coastline of Qingdao from 2000 to 2019. The total length of artificial coastlines increased from 272.34 km in 2000, accounting for 53.63% of the total length of the coastline, to 359.81 km in 2019, accounting for 68.40% of the coastline (Figure 4(a)).

The length of the natural boundary lines quickly decreased from 235.52 km to 180.15 km and then decreased steadily to 166.19 km. On the opposite, the length of the artificial boundary line grew quickly at first and then increased steadily. In 2019, its size grew to 526 km from 507.86 km in 2000. Throughout the long coastal region of Qingdao, bedrock has continuously deteriorated along the natural coastline (Figure 5). From 2000 to 2019, the bedrock coastline was substantially reduced from 29.04% to 13.65% by proportion. The sandy and estuarine coastline types have probably shown a growing pattern from 2000 to 2019 over the entire period. The muddy coastline was the only form of coastline where no change was observed overall. The largest altered portion among the artificial coastlines was aquaculture, followed by revetment and seawall (Figure 5). The area of the aquaculture coastline grew steadily over the entire period from 2000 to 2010, but its length decreased between 2010 and 2019. In the entire time from 2000 to 2019, harbor and wharf have progressed along the entire coastline. The salt field was the only coastline that proportionally showed a declining trend in its duration from 2000 to 2019, and it was 7.43% to 4.64%. Construction showed a decreasing trend from 2000 to 2010, but it increased rapidly after 2010. Despite the fact that the seawall and revetment only cover a small portion of the total coastline, their growth rate accelerated after 2000, and they were particularly visible from 2010 to 2019. In Figure 5, the spatial distributions are shown.

4. Discussion

A complicated interaction of different human-induced and natural coastal processes triggers the shoreline changes. The shorelines are changed by natural processes because of geology and geomorphology, the cumulative motion of tides and winds, sea level fluctuations, tectonics, and storms. Manipulation of hydrological cycles by dam construction, building on beaches, coastal structures such as harbors, beach-protecting structures, jetties, mining of beach sand, and degradation of coastal vegetation are the human activities that could exacerbate coastal changes. Most of the anthropogenic growth was the key reason for the coastline change in Qingdao over 19 years, while the natural factors were the secondary reasons [50]. The overall changes of shorelines along the coast of mainland China increased over the ~1990–2019 period primarily for land reclamation and quay construction [61].

Climatic conditions and natural processes are important in determining shoreline changes. These processes are influenced not only by changes in tides, waves, etc., on an hourly or daily basis but also over longer timescales, such as sea level and climate change. Relative sea level rise, which may occur because of a rise in the water volume of the seas or the subsidence or emergence of land by natural processes, is a long-term geological phenomenon of paramount significance to the shoreline. In the Qingdao coastal zone, the relative change in the sea level might be involved in modifying the shorelines. Over the past decades, the length of natural shorelines of each subtype has decreased, while artificial construction and quay shorelines of the artificial group have increased dramatically [61].

The shoreline changes have included different coastal landform features such as headland and bays, mudflats, estuaries, beaches, and bedrocks in the study area. The present study shows that the coastal areas of the bedrocks are under significant threat of decline. More changes have occurred in the narrow continental shelf and exposed rocky

Figure 3: Rocky coastline (a, b), sandy coastline (c, d), estuarine coastline (e, f), and artificial coastline (g, h) superimposed on Landsat (2019) and Google Earth images (2019).

coast regions, resulting in the accumulation of sandy and muddy sediments along beaches and estuaries. More sediments are produced, and cliffs are created by the high wave energy acting on the soft part of the bedrocks, which eventually changes the shoreline along the entire study area.

In the present results, rocky shorelines were lost, which evinces how anthropogenic changes that are often extreme to local ecosystems will destroy natural environments and prevent the accessibility of future generations from experiencing these lively, living landscapes. Sand deposited along the beaches is for the influence of materials transported by continental rivers and waves [62]. Coastal shifts have also been witnessed by low-lying sandy beaches and dunes, resulting in an overall seaward rise along the shoreline. The

Table 2: The coastline classification system of Qingdao, East China.

Primary categories	Secondary categories	Classification definition	Location name
Natural	Bedrock boundary line	It is the demarcation line on the bedrock coast between the land and sea	Liuqinghe
	Estuary boundary line	The border between the ocean and the estuary and between the ocean and the river	Mengjiatan
	Sandy boundary line	Composed of loose, very soft, and fine materials such as sand, silt, and sludge; comprising relatively straight coastline and wide, long beaches; sand in the bay at high tide	Jinshatan
	Muddy boundary line	Coastline on the muddy or silt coast, which is usually at the vegetation line where salt-tolerant plant distribution patterns change	Wangtai
Artificial	Salt field boundary line	The saline-alkali drying sea-land dividing line	Taitou
	Construction boundary line	A land-to-sea demarcation line for the growth of the maritime industry and for other constructions	Suliu
	Harbor and wharf boundary line	A demarcation line for the practical usage of port terminals between the land and sea	Xiangyang
	Breeding boundary line	An artificially built land-sea border for aquaculture	Yinghai
	Road boundary line	The land-sea dividing line of the artificially built highway	Xiaozhuang
	Other artificial boundary lines	Lines of artificial boundaries, which do not belong to the classification above	

Modified from [3, 57–60].

Table 3: Qingdao coastline statistics for 2000, 2010, and 2019.

Coastline types	2000		2010		2019		Net change area (%)	
	Area (km)	Area (%)	Area (km)	Area (%)	Area (km)	Area (%)	2000–2010	2010–2019
Natural coastline								
Bed rock	147.49	29.04	86.55	16.70	72.35	13.65	12.34	3.05
Sandy	80.64	15.87	88.15	17.01	88.74	16.74	1.14	0.27
Muddy	3.81	0.75	—	—	—	—	—	—
Estuary	3.58	0.70	5.45	1.05	5.10	0.96	0.35	0.09
Subtotal	235.52	46.37	180.15	34.77	166.19	31.35	11.6	3.42
Artificial coastline								
Aquaculture	117.19	23.07	167.01	32.23	137.29	26.08	9.16	6.15
Salt field	37.76	7.43	31.87	6.15	24.51	4.62	1.28	1.53
Construction	21.82	4.29	16.87	3.25	35.71	6.92	1.04	3.67
Harbor and wharf	31.34	6.17	55.72	10.75	66.18	12.67	4.58	1.92
Revetment and seawall	64.25	12.65	66.46	12.82	96.12	18.32	0.17	5.5
Subtotal	272.34	53.63	337.93	65.23	359.81	68.65	11.6	3.42
Total	507.86	100	518.08	100	526.00	100	—	—

Qingdao coastal zone has headlands and bays, which result in changes in the shoreline along the coast. There are few areas with curved sandy beaches and multiple estuaries that easily trap the sediments and contribute to the shoreline changes. Coastal plains are often formed depending on the rivers and their catchment area [63]. Development and its associated sedimentary processes have been and are some principal contributors to coastal geomorphology in much of the coastal region. The deposition of sediment shapes coastal features such as beaches, mudflats, estuaries, mangroves, and sand dunes. During 2000–2019, the Qingdao coastal region has experienced more accretion for the excess amount of sediment discharged through the rivers [47]. Sediment discharges from rivers are considerably reduced due to the building of dams, infrastructures, development activities, and encroachments (Figure 3). This makes the sandy beaches along the shoreline more vulnerable to destruction.

Ocean waves are one of the prime movers for the littoral processes which bring changes along the shoreline. For example, the greatest factor affecting the evolution of the shoreline shift is the long-shore sediment transport rate [64]. In the formation of such coastal features, such as bars and spits, coastal transport plays a major role and causes substantial shoreline changes [65]. The coastal zones of Qingdao have experienced changes because of the littoral drift of sediments transported through the Yang, Feng, Long Quan, and Wu Long Rivers. Most of the rivers have failed to develop true deltaic characteristics because of the strong offshore current in the study area.

Human interventions and anthropogenic activities have a major influence on shifts in shorelines. Consistent with previous researchers [3, 66], the key driving force behind the shoreline changes was human activities, including land reclamation [67–69], aquaculture development [60], and

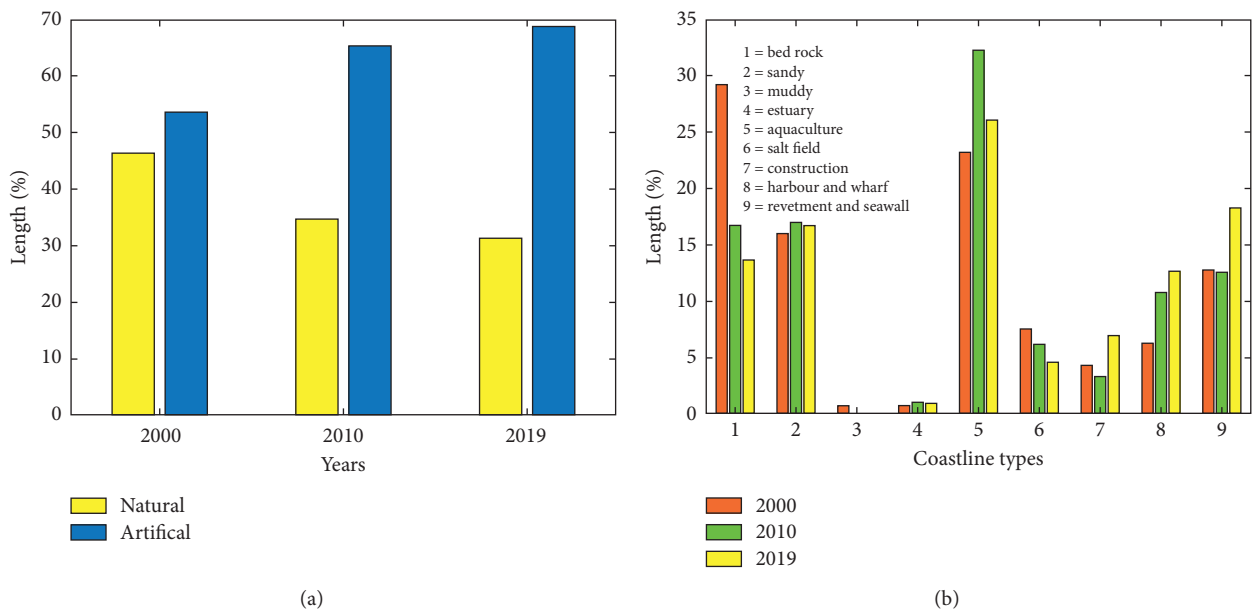


Figure 4: (a) Artificial and natural coastline length (%) from 2000 to 2019 and (b) percentage of different coastline lengths from 2000 to 2019 of Qingdao coast.

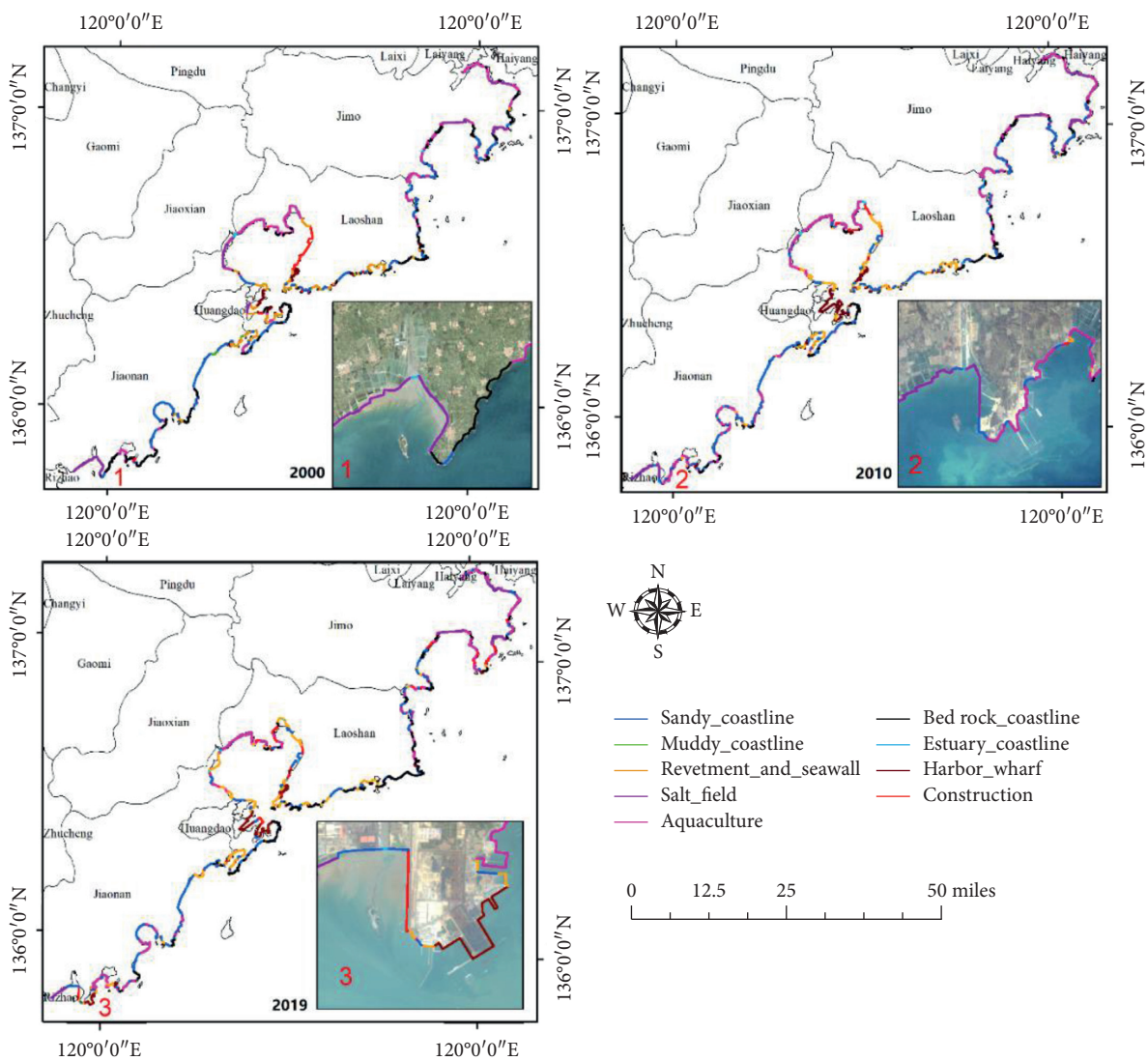


Figure 5: The spatial distribution on the Qingdao coast of the secondary coastlines.

quay construction [70]. The shoreline morphology also results in coastal urbanization, extraction, shifting of coastal sands, and dredging. The reduced supply of sediment caused by offshore sand mining, shoreline disturbances, and damming of sediment-rich rivers has led to the further loss and deterioration of coastal ecosystems, including beaches, estuaries, and mudflats.

The construction of coastal structures such as seawall, revetment, harbors, wharf, and aquaculture in Qingdao coastal regions may interfere with the process of long-shore drift, changing the sediment budget and exacerbating shoreline changes of the adjacent bedrocks, estuaries, mudflats, beaches, or beach head in a down-drift direction. The current study shows that, for anthropogenic processes along the coast, the Qingdao coast undergoes greater changes. Shoreline changes are exacerbated by unsustainable coastal construction programs. Different construction works are carried out along the shore that can cause major changes in the surrounding shoreline coastal area. The present studied coast of Qingdao is heavily experienced by constructions, revetment, seawall, and aquacultural activities (Figures 3(g) and 3(h)). Because of anthropogenic activities, this coast has high change rates, and it discloses the fact that the existence of river input, salt fields, and wave patterns determines the rate of shoreline changes. The disruption of river input is apparent here, which results in the beach's direct risk of survival.

There are several factors such as sand dunes and vegetation which protect shorelines from changes and help to preserve the coastal environment. The human-induced activities that cause shoreline changes have significantly affected these natural protections. Tian et al. [61] mentioned in 2020 that low-economic value shorelines, such as rocky, sandy, muddy, and estuarine shorelines, were transformed into high-economic value shorelines, such as aquaculture, which were transformed into shorelines with even more economic values, such as shorelines of construction and quay. Growing populations and development works along the coasts put the coastal ecosystems under unprecedented pressure. Meanwhile, ongoing manifestations of climate change, such as increasing sea levels and the severity of storm surges, increase the threats to the coastal population's habitats, natural resources, infrastructure, and well-being.

5. Conclusion

In this study, Landsat ETM+ images for 2000 and 2010 and Landsat OLI images for 2019 are used to delineate and classify the coastline along the East China coast of Qingdao. Because of both natural and human-induced activities, the Qingdao coastal zone has undergone significant coastline modifications. The spatial and temporal variations of the coastline in the study area have been extensive in the past 19 years: the length of the coastline has increased substantially from 2000 to 2019. The increased types of coastline are predominantly artificial frontier lines. The current study shows that the total length of artificial coastlines grew from 272.34 km in 2000, representing 53.63% of the total length of the coastline, to 359.79 km in 2019, representing 68.40% of the coastline. The rocky coast of the natural coastline decreased dramatically from 29.04% to 13.65% by proportion, whereas the largest portion covered by aquaculture in the artificial coastline changed from 23.07% in 2000 to 26.08% in 2019.

The growth of the coastline has been significantly influenced by human activities. The length of the natural coastal boundary line has deteriorated, converting a substantial number of natural coastlines into artificial coastlines. The geomorphological and geological effects, sediment supply, and relative increase in sea level promote shoreline development. The analysis of the shoreline change rate often shows both short- and long-term changes along the studied area's different coastal zones. Eventually, the environmental and resource shifts of the studied Qingdao coastal zone influence the coastal landforms. The types of coastal landforms change with the changes in the artificial and natural features along the coastline, particularly the infrastructure, aquaculture, harbor structures, bedrocks, beaches, estuaries, and mudflats. Human activities influence not only the shifts in the coastline but also the coastal zone's types and areas of surface coverage. A realistic management method is provided by the study of shoreline change and modes based on geomorphic concepts and thus understanding the reasons for shoreline changes. The present research thus clearly focuses on the impacts of the coastal processes produced in the study area, both natural and anthropogenic.

Data Availability

The authors used the USGS and field data along Qingdao coastline.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key Research and Development Program (Grant no. 2017YFC145600) and the National Natural Science Foundation of China (Grant no. 41776182).

References

[1] P. S. Mujabar and N. Chandrasekar, "Shoreline change analysis along the coast between Kanyakumari and Tuticorin of India using remote sensing and GIS," *Arabian Journal of Geosciences*, vol. 6, no. 3, pp. 647–664, 2011.

[2] X. Huang, *Resource Economics*, Vol. 1–3, Nanjing University Press, Nanjing, China, 2010.

[3] X. Hou, J. Liu, Y. Song, and X. Li, "Environmental-ecological effect of development and utilization of China's coastline and policy recommendations," *Bulletin China Academy Science*, vol. 31, pp. 1143–1150, 2016.

[4] C. Salmon, V. K. E. Duvat, and V. Laurent, "Human- and climate-driven shoreline changes on a remote mountainous tropical Pacific Island: tubuai, French Polynesia," *Anthropocene*, vol. 25, Article ID 100191, 2019.

[5] J. H. List, A. S. Farris, and C. Sullivan, "Reversing storm hotspots on sandy beaches: spatial and temporal characteristics," *Marine Geology*, vol. 226, no. 3-4, pp. 261–279, 2006.

[6] Z. Umar, W. A. A. W. M. Akib, and A. Ahmad, "Monitoring shoreline change using Remote sensing and GIS: a case study of padang coastal area, Indonesia," in *Proceedings of the IEEE 9th International Colloquium on Signal Processing and its Applications*, pp. 280–284, Kuala Lumpur, Malaysia, March 2013.

[7] R. Ranasinghe, T. M. Duong, S. Uhlenbrook, D. Roelvink, and M. Stive, "Climate-change impact assessment for inlet-interrupted coastlines," *Nature Climate Change*, vol. 3, no. 1, pp. 83–87, 2012.

[8] T. Liao, T. Cai, Y. Liu, and X. Xia, "Continental shoreline change in Zhejiang during the last one hundred years," *Journal of Marine Science*, vol. 34, pp. 25–33, 2016.

[9] B. Wang, L. Liang, F. Hui et al., "Spatiotemporal changes of the Chinese coastlines: landsat imagery from 1975 to 2015," *Journal of Beijing Normal University (Natural Science)*, vol. 55, pp. 83–100, 2019.

[10] X. Sun, M. Wu, J. Tian et al., "Driving forces and spatio-temporal variation of Weihai coastline in recent 30 years," *Journal of Oceanography*, vol. 38, pp. 206–213, 2019.

[11] J. Li, M. Ye, R. Pu et al., "Spatiotemporal change patterns of coastlines in zhejiang province, China, over the last twenty-five years," *Sustainability*, vol. 10, no. 2, p. 477, 2018.

[12] R. Welch, M. Remillard, and J. Alberts, "Integration of GPS, remote sensing, and GIS techniques for coastal resource management," *Photogrammetric Engineering and Remote Sensing*, vol. 58, pp. 1571–1578, 1992.

[13] H. Stokkom, G. Stokman, and J. Hovenier, "Quantitative use of passive optical remote sensing over coastal and inland water bodies," *International Journal of Remote Sensing*, vol. 14, pp. 541–563, 1993.

[14] A. D. Switzer, C. R. Sloss, B. P. Horton, and Y. Zong, "Preparing for coastal change," *Quaternary Science Reviews*, vol. 54, pp. 1–3, 2012.

[15] Y. Yan, Z. Zhang, C. Wang, L. Zhang, Y. Huang, and J. Zhang, "Analysis of recent coastline evolution due to marine reclamation projects in the Qinzhou Bay," *Polish Maritime Research*, vol. 24, no. s2, pp. 188–194, 2017.

[16] S. R. Nayak, "Use of satellite data in coastal mapping," *Indian Cartographer*, vol. 23, pp. 147–157, 2002.

[17] E. Sener, A. Davraz, and S. Sener, "Investigation of aksehir and eber lakes (SW Turkey) coastline change with multitemporal satellite images," *Water Resources Management*, vol. 24, no. 4, pp. 727–745, 2009.

[18] G. Zhu and X. Xu, "Annual processes of land reclamation from the sea along the northwest coast of bohai bay during 1974 to 2010," *Journal of Geographical Sciences*, vol. 32, pp. 1006–1012, 2012.

[19] A. A. Alesheikh, A. Ghorbanali, and N. Nouri, "Coastline change detection using remote sensing," *International Journal of Environmental Science & Technology*, vol. 4, no. 1, pp. 61–66, 2007.

[20] L. J. Moore, "Shoreline mapping techniques," *Journal of Coastal Research*, vol. 16, pp. 111–124, 2000.

[21] A. Saleem and J. L. Awange, "Coastline shift analysis in data deficient regions: exploiting the high spatio-temporal resolution Sentinel-2 products," *Catena*, vol. 179, pp. 6–19, 2019.

[22] K. Nassar, H. Fath, W. E. Mahmod, A. Masria, K. Nadaoka, and A. Negm, "Automatic detection of shoreline change: case of North Sinai coast, Egypt," *Journal of Coastal Conservation*, vol. 22, no. 6, pp. 1057–1083, 2018.

[23] B. Pradhan, H. Rizeei, and A. Abdulle, "Quantitative assessment for detection and monitoring of coastline dynamics with temporal RADARSAT images," *Remote Sensing*, vol. 10, no. 11, Article ID 1705, 2018.

[24] H. Liu and K. C. Jezek, "Automated extraction of coastline from satellite imagery by integrating canny edge detection and locally adaptive thresholding methods," *International Journal of Remote Sensing*, vol. 25, no. 5, pp. 937–958, 2004.

[25] B.-J. He, Z.-Q. Zhao, L.-D. Shen, H.-B. Wang, L.-G. Li, and B.-J. He, "An approach to examining performances of cool/hot sources in mitigating/enhancing land surface temperature under different temperature backgrounds based on Landsat 8 image," *Sustainable Cities and Society*, vol. 44, pp. 416–427, 2019.

[26] X. Li and M. C. J. Damen, "Coastline change detection with satellite remote sensing for environmental management of the Pearl River Estuary, China," *Journal of Marine Systems*, vol. 82, pp. S54–S61, 2010.

[27] M. Mishra, P. Chand, N. Pattnaik et al., "Response of long- to short-term changes of the Puri coastline of Odisha (India) to natural and anthropogenic factors: a remote sensing and statistical assessment," *Environmental Earth Sciences*, vol. 78, p. 338, 2019.

[28] M. M. Kale, M. Ataol, and I. S. Tekkanat, "Assessment of shoreline alterations using a Digital Shoreline Analysis System: a case study of changes in the Ye¸silırmak Delta in northern Turkey from 1953 to 2017," *Environmental Monitoring and Assessment*, vol. 191, p. 398, 2019.

[29] D. T. Thoai, A. N. Dang, and N. T. Kim Oanh, "Analysis of coastline change in relation to meteorological conditions and human activities in Ca mau cape, Viet Nam," *Ocean & Coastal Management*, vol. 171, pp. 56–65, 2019.

[30] X. Wu, C. Liu, and G. Wu, "Spatial-temporal analysis and stability investigation of coastline changes: a case study in shenzhen, China," *Institute of Electrical and Electronics Engineers Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 1, pp. 45–56, 2018.

[31] S. R. Ahmad and V. C. Lakhan, "GIS-based analysis and modeling of coastline Advance and retreat along the coast of Guyana," *Marine Geodesy*, vol. 35, Article ID 1e15, 2012.

[32] R. Dolan, B. Hayden, and J. Heywoode, "A new photogrammetric method for determining coastline erosion," *Coastal Engineering*, vol. 2, Article ID 21e39, 1978.

[33] T. Kuleli, "Quantitative analysis of coastline changes at the Mediterranean Coast in Turkey," *Environmental Monitoring and Assessment*, vol. 167, Article ID 387e397, 2010.

[34] T. Kuleli, A. Guneroglu, K. Fevzi, and M. Dihkan, "Automatic detection of coastline change on coastal Ramsar wetlands of Turkey," *Ocean Engineering*, vol. 38, Article ID 1141e1149, 2011.

[35] J. M. Kusimi and J. L. Dika, "Sea erosion at Ada Foah: assessment of impacts and proposed mitigation measures," *Natural Hazards*, vol. 64, Article ID 983e997, 2012.

[36] M. L. Yates, G. L. Cozannet, M. Garcin, E. Salai, and P. Walker, "Multidecadal atoll coastline change on manihi and manuae, French polynesia," *Journal of Coastal Research*, vol. 29, Article ID 870e882, 2013.

[37] S. S. Durduran, "Coastline change assessment on water reservoirs located in the Konya Basin Area, Turkey, using multitemporal landsat imagery," *Environmental Monitoring and Assessment*, vol. 164, Article ID 453e461, 2010.

[38] S. Quan, R. G. Kvitek, D. P. Smith, and G. B. Griggs, "Using vessel-based LIDAR to quantify coastal erosion during El Ni~no and inter-El Ni~no periods in Monterey Bay,

California," *Journal of Coastal Research*, vol. 29, Article ID 555e565, 2013.

[39] S. M. Solomon, "Spatial and temporal variability of coastline change in the Beaufort-Mackenzie region, northwest territories, Canada," *Geo-Marine Letters*, vol. 25, Article ID 127e137, 2005.

[40] D. C. Twichell, J. G. Flocks, E. A. Pendleton, and W. E. Baldwin, "Geologic controls on regional and local erosion rates of three northern Gulf of Mexico Barrier-Island systems," *Journal of Coastal Research*, vol. 63, Article ID 32e45, 2013.

[41] J. S. Lee and I. Jurkevich, "Coastline detection and tracing in SAR images," *Institute of Electrical and Electronics Engineers Transactions on Geoscience & Remote*, vol. 28, no. 4, pp. 662–668, 1990.

[42] J. Ryu, J. Won, and K. Min, "Waterline extraction from Landsat TM data in a tidal flatA case study in Gomso Bay, Korea," *Remote Sensing of Environment*, vol. 83, no. 3, pp. 442–456, 2002.

[43] D.-J. Kim, W. M. Moon, S.-E. Park, J.-E. Kim, and H.-S. Lee, "Dependence of waterline mapping on radar frequency used for sar images in intertidal areas," *Institute of Electrical and Electronics Engineers Geoscience and Remote Sensing Letters*, vol. 4, no. 2, pp. 269–273, 2007.

[44] R. Li, R. Ma, and K. Di, "Digital tide-coordinated shoreline," *Marine Geodesy*, vol. 25, no. 1-2, pp. 27–36, 2002.

[45] A. M. Muslim and G. M. Foody, "DEM and bathymetry estimation for mapping a tide-coordinated shoreline from fine spatial resolution satellite sensor imagery," *International Journal of Remote Sensing*, vol. 29, no. 15, pp. 4515–4536, 2008.

[46] E. H. Boak and I. L. Turner, "Shoreline definition and detection: a review," *Journal of Coastal Research*, vol. 214, no. 4, pp. 688–703, 2005.

[47] X. F. Ma, D. Z. Zhao, X. G. Xing, F. S. Zhang, S. Y. Wen, and F. Yang, "Means of withdrawing coastline by remote sensing," *Marine Environmental Research*, vol. 26, no. 2, pp. 185–189, 2007, in Chinese.

[48] J. Xu, Z. Zhang, X. Zhao et al., "Spatial and temporal variations of coastlines in northern China (2000-2012)," *Journal of Geographical Sciences*, vol. 24, no. 1, pp. 18–32, 2014.

[49] R. Gens, "Remote sensing of coastlines: detection, extraction and monitoring," *International Journal of Remote Sensing*, vol. 31, no. 7, pp. 1819–1836, 2010.

[50] M. Yasir, H. Sheng, H. Fan et al., "Automatic coastline extraction and changes analysis using remote sensing and GIS technology," *Institute of Electrical and Electronics Engineers Access*, vol. 8, pp. 180156–180170, 2020.

[51] X. Ma, "An overview of means of withdrawing coastline by remote sensing," *Remote Sensing Technology and Application*, vol. 22, no. 4, pp. 575–580, 2007.

[52] M. Zhou, M. Wu, G. Zhang, L. Zhao, X. Hou, and Y. Yang, "Analysis of coastal zone data of northern Yantai collected by remote sensing from 1990 to 2018," *Applied Sciences*, vol. 9, no. 20, p. 4466, 2019.

[53] S. Deng, *ENVI Remote Sensing Image Processing Method*, Higher Education Press, Beijing, China, 2014.

[54] S. K. McFeeters, "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, no. 7, pp. 1425–1432, 1996.

[55] N. Otsu, "A threshold selection method from gray-level histograms," *Institute of Electrical and Electronics Engineers Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[56] X. Ge, X. Sun, and Z. Liu, "Objectriented coastline classification and extraction from remote sensing imagery," Q. Tong, J. Shan, and B. Zhu, Eds., in *Proceedings of the Remote Sensing of the Environment: 18th National Symposium on Remote Sensing of China*, vol. 9158, Spie-Int Soc Optical Engineering, Bellingham, WA, USA, May 2014.

[57] B. Liu, W. Meng, J. Zhao, B. Hu, L. Liu, and F. Zhang, "Variation of coastline resources utilization in China from 1990 to 2013," *Journal of Natural Resources*, vol. 30, pp. 2033–2044, 2015.

[58] W. Sun, Y. Ma, J. Zhang, S. Liu, and G. Ren, "Study of remote sensing interpretation keys and extraction technique of different types of shoreline," *Bulletin Survey Mapping*, vol. 3, pp. 41–44, 2011.

[59] X. Hou, T. Wu, W. Hou, Q. Chen, Y. Wang, and L. Yu, "Characteristics of coastline changes in mainland China since the early 1940s," *Science China Earth Sciences*, vol. 59, no. 9, pp. 1791–1802, 2016.

[60] N. Xu, *Research on Spatial and Temporal Variation of China Mainland Coastline and Coastal Engineering*, Yantai Insititute of Coastal Zone Research CAS, Yantai, China, 2016.

[61] H. Tian, K. Xu, J. I. Goes, Q. Liu, H. D. R. Gomes, and M. Yang, "Shoreline changes along the coast of mainland China-time to pause and reflect?" *ISPRS International Journal of Geo-Information*, vol. 9, no. 10, p. 572, 2020.

[62] W. D. Thornbury, *Principles of Geomorphology*, John Wiley & Sons, New York, NY, USA, 1969.

[63] M. Mishra, "Geomorphic regionalization of coastal zone using geospatial technology," *International Journal of Environment and Geoinformatics*, vol. 3, no. 2, pp. 11–23, 2016.

[64] C. Wen-Hung, H. Bin-Chen, and C. Piao-Tsai, "Simulation of shoreline change behind a submerged permeable breakwater," in *Proceedings of the Taiwan-Polish Joint Seminar on Coastal Protection*, Tainan, Taiwan, November 2008.

[65] D. B. King, "The dynamics of inlets and bays," Technical Report No. 2, Coastal and Oceanographic Engineering Laboratory, University of Florida, Gainesville, FL, USA, 1974.

[66] Q. Fan, L. Liang, F. Liang, and X. Sun, "Research progress on coastline change in China," *Journal of Coastal Research*, vol. 99, no. sp1, pp. 289–295, 2020.

[67] B. Ai, R. Zhang, H. Zhang, C. Ma, and F. Gu, "Dynamic process and artificial mechanism of coastline change in the Pearl River Estuary," *Regional Studies in Marine Science*, vol. 30, Article ID 100715, 2019.

[68] X. Ding, X. Shan, Y. Chen, X. Jin, and F. R. Muhammed, "Dynamics of shoreline and land reclamation from 1985 to 2015 in the Bohai Sea, China," *Journal of Geographical Sciences*, vol. 29, no. 12, pp. 2031–2046, 2019.

[69] M. S. Zhu, T. Sun, and D. D. Shao, "Impact of land reclamation on the evolution of shoreline change and nearshore vegetation distribution in Yangtze River Estuary," *Wetlands*, vol. 36, no. S1, pp. 11–17, 2016.

[70] X. Zhang, D. Pan, J. Chen, J. Zhao, Q. Zhu, and H. Huang, "Evalu0ation of coastline changes under human intervention using multi-temporal high-resolution images: a case study of the Zhoushan Islands, China," *Remote Sensing*, vol. 6, no. 10, pp. 9930–9950, 2014.

Hindawi
Scientific Programming
Volume 2021, Article ID 5556011, 14 pages
https://doi.org/10.1155/2021/5556011

Research Article

Premature Ventricular Contractions' Detection Based on Active Learning

Xianrong Zhang,[1] **Muhammad Shafiq**,[2] **Guijun Zheng**,[3] **Junping Wan**,[1] **and Zhe Sun**[1]

[1]*Cyberspace Institute of Advanced Technology (CIAT), Guangzhou University, Guangzhou 510006, China*
[2]*School of Computer Science & Technology, Harbin Institute of Technology, Harbin 150001, China*
[3]*School of Software Engineering, University of Science and Technology of China, Hefei 230051, China*

Correspondence should be addressed to Zhe Sun; sunzhe@gzhu.edu.cn

Received 12 January 2021; Revised 12 February 2021; Accepted 25 February 2021; Published 8 March 2021

Academic Editor: Shaukat Ali

Premature ventricular contractions (PVCs) are one of the most common cardiovascular diseases with high risk to a large population of patients. It has been shown that supervised learning algorithms can detect PVCs from beat-level ECG data. However, a huge human effort is needed in order to achieve an accurate detection rate. A convolutional autoencoder was trained in this work in an unsupervised fashion to extract features automatically with zero prior specialized knowledge. Random forest was adopted as a supervised algorithm trained on the features generated by the autoencoder. Various active learning selection strategies, uncertainty-based and diversity-based, were studied on top of the random forest. In each iteration of active learning, the training data are updated with newly selected samples and fed into the classifier. The performance on an independent validation set is recorded in each iteration. As a result, among the different uncertainty sampling strategies, the least confidence score shows a better F_1 score of 0.85 than other methods. In between the two diversity-based strategies, the representative clustering sample had the best F_1 score than the k-center-greedy algorithm. By comparing the performance of different active learning methods trained on half of the original data size with the same classifier trained on the full set, the F_1 score of least confidence is still better than the full set. This study demonstrates that active learning could help reduce human annotation effort by achieving the same level of performance as the classifier trained on the fully annotated training data.

1. Introduction

Premature ventricular contractions (PVCs) are one of the most common arrhythmias that occur in a large population of patients [1]. In a more serious case, when PVCs happen with other cardiac risk factors concurrently, it could lead to many extreme situations, like cardiac death or heart attack. Electrocardiograms (ECG) are recognized as the most useful and noninvasive technique for monitoring cardiac activity, within which the monitoring of various arrhythmias are the main tasks [2]. Also, it would be difficult for clinicians to recognize PVCs from ECG if only a short period of ECG is provided since the evaluation of PVCs needs a reference from neighbor ECG segments. Therefore, it is desirable if one approach can automatically detect PVCs from just one beat of ECG.

Applying traditional machine learning algorithms on detection of PVCs is one mainstream category, in which feature engineering is the most important process in order to get better performance. ECG morphological based features have been proved to be useful. Geddes [3] applied the QRS complexity length interval gap between the two R peaks, and the first-order signal derivative of the signal as the manually selected features, and then fed them into a tree-based classification algorithm. However, to follow the same approach, in addition to the annotation of PVCs, the PQRS annotation for each beat of the ECG is also needed, which causes much more work for human clinicians. In work [4], the author came up with a combination of features from a different aspect. Besides the temporal features, sparse signal decomposition is also adopted for each segment of ECG. At the same time, human-design rules were

also proposed as an additional filter for features. The results show that PVCs can be detected in a robust way compared to just applying any signal aspect of the feature.

Within all the work above, domain knowledge is required to find the most plausible features. In recent years, the deep learning algorithms have become the focus of the research community. Due to their powerful classification capabilities and end-to-end methods, deep learning does not require tedious feature engineering to process the data. Researchers only need to feed the raw data into a deep neural network, and the network can extract the most important features and achieve a better performance than traditional algorithms. In [5], they used a convolutional autoencoder to extract features and then fed the extracted feature list into the random forest for the classification task. The results achieved over 90% accuracy over the whole patient cohort.

Despite traditional machine learning or deep neural network, the success of one classifier highly depends on the massive and accurate annotated training set, which requires tremendous human effort.

Active learning has been brought into much research as a novel idea due to its ability to achieve similar or better performance using only half of the original data [6–8], where the core idea behind active learning is to seek out the most informative data samples to annotate. The general process of active learning starts with an initial well-labeled training set and a data pool with no annotation. Classifiers are then trained on a baseline set of training and each data sample from the data pool will be evaluated by a pre-trained classifier. The output probability for each sample in the data pool will be used as the input for the active learning selection strategy. Finally, selected samples from the data pool will be transmitted to annotators to obtain the exact label and then appended to the training set for the second iteration of the training process. It is evident throughout the whole workflow that the selection strategy is the most important part of the active learning process [9].

Some of the main contributions of this study are listed as follows:

(1) Overall framework: An active learning framework is proposed to detect premature ventricular beats, which reduced the workload and the cost of manual labeling data. The advantages of artificial intelligence are creatively applied to biomedicine to help clinicians improve the accuracy of PVCs detection.

(2) Feature engineering: Convinced autoencoders are designed to automatically extract features from data without the prior knowledge of medical experts, providing novel insights for feature engineering of physiological data. After the extracted features are input into the classifier, the results show that the traditional machine learning methods have advantages. Using convolutional autoencoders to extract features is more convenient and fast than incorporating human annotation efforts.

(3) Data distribution: Initial training data distributions were investigated, and in most cases, random sampling of initial training data may not be sufficient to represent the entire data set. We propose an alternative approach to initial training data and test the impact of each approach.

(4) Selection strategy: In active learning, the selection strategy is crucial. By comparing the selection strategies, we put forward the selection strategies suitable for PVC testing.

This paper is organized as follows. Additional related work is reported in Section 2. Methods and the entire design of the work are reported in Section 3. A comprehensive illustration of the work is given in Section 4. Finally, the discussion section of this paper is in Section 5 and the conclusion is presented in Section 6.

2. Related Work

2.1. Traditional Machine Learning. Active learning has attracted a lot of spotlight in the machine learning community. Different selection strategies have proven to be useful in many tasks, which were designed from different perspectives. When considering uncertainty, the uncertainty is calculated based on output probability from the initiation classifier. When it comes to diversity, many geometry-based approaches are proposed. The core idea behind them is to calculate the distance between samples in the unlabeled pool and then select the data points that represent the entire distribution of the data pool. Many state-of-art methods have been proposed. In [10], the author took the selection strategy as the k-cover problem, which is the solution to find the best k data point out of the whole data pool. While the k-cover problem is NP-hard and has been proven, the authors tried a k-greedy algorithm to simulate the k-cover problem. Reference [11] proposed a novel way by building an auxiliary model to estimate the loss for each input, by which those samples with higher loss will be selected for annotation in each iteration. These works examine the data distribution of active learning, but these advanced active learning techniques have not been explored in physiological research.

2.2. Traditional Selection Strategies. Traditional selection strategies for active learning in ECG data have been pursued recently. Reference [12] applied active learning as an effective approach for finding the most relevant signals with motion artifacts in order to accurately classify the human activity, in which only 16% of the original training data is used. In [13], active learning is used mainly to produce more generalizable training data among the patient cohort rather than reducing the human annotation effort. The authors adopted a global recurrent neural network that captures the time order of the input signal. The selection strategy is based on a combination of entropy index, model output, and Premature-or-Escape-Flag index, which is temporal information learned from the embedding layer of the model. At the same time, the comparison between different selection strategies [14–17] has not been investigated in the field of physiological measurement research. In the biomedical area, active learning was adopted as the strategy combined with SVM trying to discriminate an ECG-based classification task

[18]. In [19], a novel selection strategy called AIFT was created. The results show that the proposed method can help improve the performance of classifying three biomedical image classification tasks, with less human effort involved. These works had contributed to the development of active learning, and selection strategies suitable for PVCs have not been studied.

2.3. Deep Learning Algorithms. Deep learning algorithms are also being used to create active learning classifiers for ECG-based classification tasks [20, 21]. In [22], a global recurrent neural network was adopted for the purpose of ECG beat classification. Morphological and temporal features for ECG beats were investigated, and active learning was utilized to select the most representative samples for training. As for the work [23], a convolutional neural network was applied for the wearable ECG classification. Breaking-ties and modified breaking-ties algorithms were used with active learning simultaneously to improve model performance. ECG abnormalities with the convolutional neural network were studied in [24]. Despite the noisy part, additional six arrhythmia events within beat ECG were also detected. Active learning was also planned in the procedure in order to deal with the unseen pattern inside the original training data. Manually set decision rules are used for PVCs detection in [25]. By identifying statistical and rhythm rules, the PVC beats could be detected with high accuracy. Electrocardiograms (ECG) data are a powerful tool for reflecting cardiovascular events. Different arrhythmias can be automatically detected through machine learning algorithms. In the most recent work [26], the authors proposed a deep learning method that can detect PVCs without any human annotation effort by localizing the PVC beats via deep learning algorithms.

However, active learning for PVCs detection is not well explored. The above methods are not particularly good for PVC detection. Firstly, there is a lack of research on initialization training data. Secondly, the above work is still lacking in the extraction of data features. Moreover, most of the work requires manual annotation of data, and sufficient sample features cannot be extracted from small samples. How to improve the accuracy with fewer data remains to be studied. Finally, there was less research on active learning selection strategies for those works. In particular, there are few studies on the classification of active learning in biomedicine.

3. Materials and Methods

Active learning algorithms are applied in PVCs for detection in this paper. The algorithm model comprehensively considers data initialization, data feature extraction, and sampling strategies. In order to improve the performance of the original classifier, this paper uses k-means++ for initialization. In order to extract features better, this paper designs a convolutional autoencoder. To fully study the impact of sampling strategies on active learning, uncertain sampling and diversity sampling were also studied.

3.1. Overview. Compared to related work, we hope to learn from small samples, to reduce the workload of manual annotation. In addition, in order to improve training accuracy, we proposed an initialization method for training data. To fully examine the data characteristics, we propose a convolutional autoencoder. Finally, we comprehensively examined selection strategies to improve the effectiveness of active learning. The overall flow of the framework is as follows:

(1) Initial training data: we propose a training data initialization method that is more suitable for PVC detection, and the method is k-means++.

(2) Feature engineering: in this framework, we design a convolutional autoencoder for feature engineering. Self-convolutional encoders can learn the characteristics of data by themselves to solve the problem of low efficiency of manual annotation data sets.

(3) Data pool selection strategy: in this work, two main aspects of the selection strategy, uncertainty and diversity, were well explored in the task of PVCs detection on beat-level ECG data. Uncertainty sampling is discussed as follows: ① Least confidence sampling. ② Margin sampling. ③ Shannon entropy sampling. Diversity sampling is discussed as follows: ① K-center-greedy sampling. ② Representative cluster sampling.

The process can be summarized as follows.

A convolutional autoencoder was trained to extract the features automatically. These data-driven features are then fed into a random forest classifier. During each iteration of active learning, the same algorithm is trained on the updated training data. The overall framework is shown in Figure 1.

We first downloaded PVCs detection data sets from the website, and these data sets have not been annotated by experts. In the initial phase, we performed random initialization and initialization of the K-means++ algorithm. This was done to study the impact of data initialization on the effect of active learning. After the initialization algorithm is complete, the active learning algorithm can select a portion of the representative data. This small part of the initialized data is integrated into the initial subset. We hope that the initial subset of the initialization algorithm can represent the overall data distribution. A selected initial subset was manually marked by oracle. After manually marking the initial subset, we constructed a convolutional autoencoder to extract features from the initial subset. In addition, these features are trained using a random forest classification algorithm. Before the initial data subset is trained, the weights of the convolutional autoencoder and the random forest classifier are randomly distributed.

Because we assume that the meaning of the initial data subset is that a portion of the data represents the total data, after training on the initial data subset, our convolutional autoencoder and classifier may fully learn the characteristics of the data set. In the next iteration of the algorithm, we need to reassign a representative part of the data set to the oracle for annotation. Therefore, in an iterative process, we first

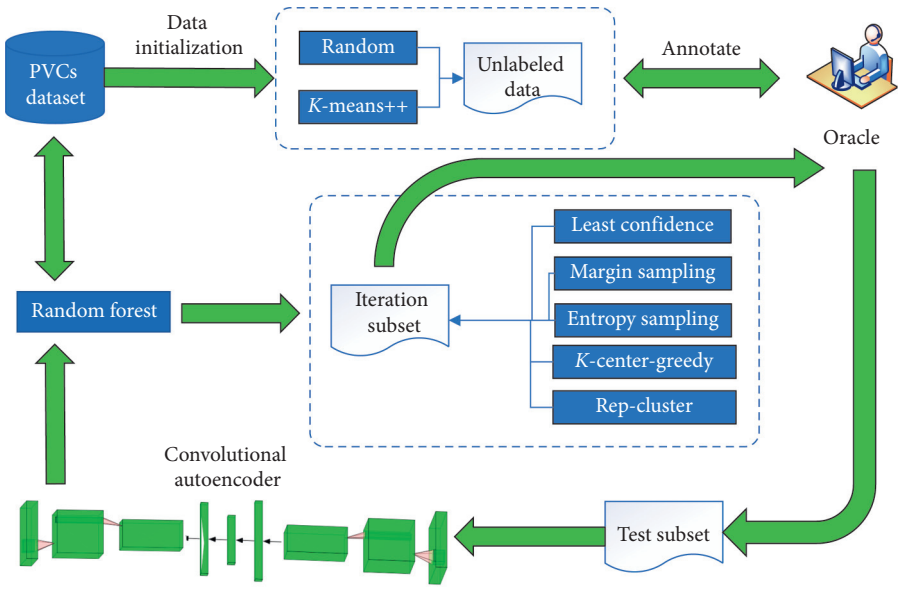


Figure 1: The overview of the workflow.

input the remaining unlabeled data into a convolutional autoencoder and classifier for predictive classification. We selected the data after classification using the uncertainty sampling strategy and the diversity sampling strategy. We call this data set an iterative subset. We resubmit data that are difficult to classify to Oracle for labeling. After multiple iterations, until the labeled data reaches half of the data pool, the iteration ends.

This paper presents a model from three aspects: training data initialization, iterative process, and selection strategy classifier.

3.2. Training Data Initialization. In most of the work, the researcher tended to select the initial training set randomly. At the same time, random selection might lead to the skewness of model parameters due to a lack of the big picture of all the training data. In this paper, we also propose a data initiation approach, k-means++, to investigate whether it is important to consider initiating the training set by design versus selecting randomly. The ultimate goal of k-means++ is to find the best subset that covers the whole big picture of the original training data. A detailed description of the significance of the results is presented in Table 1.

After data initialization is completed, feature engineering is required for the data. Feature selection directly affects the performance of the algorithm [27, 28]. Common deep learning algorithms often require a large number of classification tags, which need manual labeling. Manual labeling of data is a huge workload, which is also a disadvantage of deep learning. In order to reduce manual labeling data and fully extract features of small samples, we designed a convolutional autoencoder. An autoencoder is one of the most popular unsupervised neural network methods that take the input data also as output data. The fundamental architecture of the autoencoder is a three-layer neural network that includes an input layer, hidden layer, and output layer. The restrictions for the architecture are as follows: (1) The whole architecture is symmetric, being connected by the hidden layer. (2) The number of perceptrons of input and output are identical. (3) The number of perceptrons of the hidden layer must be smaller than the number of perceptrons of the input layer. By the "learning myself" philosophy, as shown in equation (1), the autoencoder managed to get a usable representation of the original input in a lower dimension from the hidden layer, and minimize the distance between the input x and output z as follows:

Table 1: The algorithm of k-means++.

Algorithm K-means ++
Input: Training data $Z = (x_i, y_i)$, $i = 1, \ldots, N$,
Number of selected centroids K.
initiation centroids = [one random selected points]
For j **from** 1 **to** **K − 1**
Distance ← { }
For t **from** 1 **to** length (centroids)
compute distances of "point" in **Z** to centroids
and store the point **P** with minimum distance
End
centroids = centroids + P
End

$$J(x, z) = \|x - z\|^2. \quad (1)$$

However, the ability of the autoencoder to handle complex real-world data is highly restricted if only the traditional three-layer architecture is adopted. The PVCs data set is sequence information and the potential connection between the data before and after cannot be ignored. Traditional autoencoders use ordinary neural networks to train data, which causes data stacking, and the context or spatial information of the data is also ignored. Therefore, it is

critical to design an autoencoder that can consider contextual information. As shown in the figure, the convolutional autoencoder improves the traditional autoencoder. Following the convolutional layer, linear rectification layer, and pooling layer, the convolutional autoencoder retains the spatial information of the PVCs data set. At the same time, more variants of autoencoders were proposed [29, 30]. Among these, convolutional autoencoders are widely used. Instead of importing raw data, the input data first go through several layers of convolutional kernels for feature extraction. Then, the extracted feature list is fed into a fully connected hidden layer. As shown in Figure 2, in our study, the input ECG recordings will go through a batch of one-dimensional convolutional layers, and then the length of the latent layer is set at 25. After that, the convolutional autoencoder makes the difference between the input and output as small as possible.

In this paper, each beat was extracted with the length of 250 data points, in which 89 samples were before the position of each R peak annotated by humans, and 160 after the position of each R peak annotated. In the model aspect, the dimension of latent space in the hidden layer is selected at 25, which has been proven to be sufficient to represent the raw 250-length signal [5]. We use the mean squared error as our loss function and Adam as the optimizer for the training process, and a more detailed description of the model is described in Table 2.

After the hyperparameters are set, the CAE is trained for 50 epochs at a batch size of 200. The final weight is selected in the epoch with the least loss during training, and this weight is used for inference purposes only except in the experiment in which a different approach of updating the weight of CAE is applied. In Table 1, the output size column indicates the number of samples.

3.3. Iterative Process and Selection Strategy. After the data are initialized with K-means++, we also design a convolutional autoencoder to extract features from the data. We will now discuss the iterative process of active learning for each round and the selection strategy for the iterative process.

As illustrated in Figure 3, active learning begins with initial training data to obtain the intermediate model. Then for each iteration, a batch of unlabeled data is selected according to the output probability of the intermediate model across the entire data pool, as illustrated in equation.

$$\min_{s^1:|s^1|\leq b} E_{x,y\sim p_z}\left[l\left(x,y;A_{s^0\cup s^1}\right)\right], \tag{2}$$

where s^0 and s^1 are denoted as labeled data sets and unlabeled data candidates, respectively, and b represents the selected samples during each iteration. The math behind equation (2) is trying to select b samples from s^1 which results in a minimal loss based on the current intermediate model.

Newly selected samples will be appended to the training set and fed into the next iteration. An independent validation set is used to evaluate the performance of the model in each iteration. The entire active learning process stops when the performance on the validation set is satisfied.

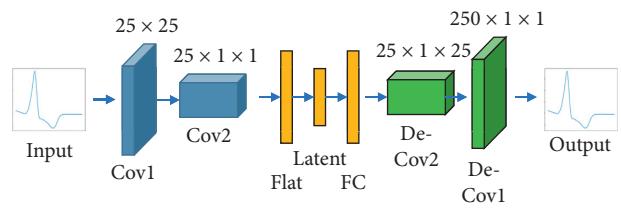


Figure 2: The architecture of the convolutional autoencoder.

Table 2: A detailed description of the architecture of the convolutional autoencoder model.

Layer number	Layer name	Kernel size	Output size
0	Input		(−1, 250)
1	Reshape		(−1, 250, 1)
2	Convolution1D	20	(−1, 25, 25)
3	Reshape		(−1, 25, 25, 1)
4	Convolution2D	(1, 20)	(−1, 25, 1, 1)
5	Convolution2D transpose	(1, 20)	(−1, 25, 1, 25)
6	Convolution2D transpose	(20, 1)	(−1, 250, 1, 1)
7	Reshape/output		(−1, 250)

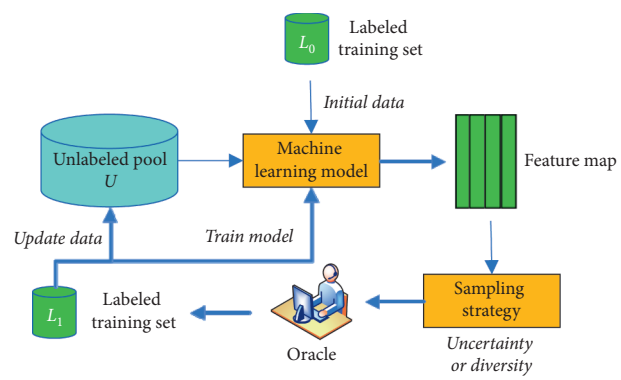


Figure 3: General workflow of active learning.

In each iteration of active learning, the selection strategy determines the quality of the results of active learning. Therefore, we discuss two types of sampling strategies, namely, uncertainty sampling and diversity sampling. Specific uncertainty sampling is discussed as follows: ① Least confidence sampling. ② Margin sampling. ③ Shannon entropy sampling. Diversity sampling is discussed as follows: ① K-center-greedy sampling. ② Representative cluster sampling. The discussion of different selection strategies enables to design a framework that is more suitable for the detection of PVCs.

As pool-based sampling strategies are the most widely used, and when there is only one unlabeled sample in the sample pool, it is equivalent to a stream-based sampling

strategy. Therefore, this article mainly studies this type of sampling strategy, especially uncertainty sampling strategies and differential sampling strategies.

A sampling strategy based on uncertainty is the most widely applicable type of sampling strategy. This sampling strategy selects the classifier to mark the samples whose predicted value of p is closest to 0.5. The sampling strategy is not only suitable for most classifiers; it effectively reduces the workload of human experts and greatly improves classifier accuracy and generalization abilities.

The sampling strategy selects one or a batch of samples in each iteration. We certainly hope that the information provided by the sample inquired is comprehensive, and the information provided by each sample is not repeated or redundant, that is, there are certain differences between the samples. In the case of extracting a single sample with the largest amount of information in each iteration and adding it to the training set, the model is retrained in each iteration, so that the acquired knowledge is used in the evaluation of sample uncertainty and can effectively avoid data redundancy. But if you query a batch of samples for each iteration, you should find ways to ensure the diversity of the samples and avoid data redundancy. This is the differential selection strategy.

By using the least confidence strategy, in each iteration, the learner will select samples which the intermediate model is most unconfident about, as shown in equation (3), in which X_{new} represents all the data from the unlabeled pool. For example, in a binary classification task class A and class B, there are two different unlabeled data samples s_1 and s_2. The intermediate model predicts samples s_1 with label A at a probability of 0.9 and samples s_2 with label A at a probability of 0.5. The least confidence strategy will select s_2 and transport it to the annotators for its actual label.

$$\phi_{\text{LC}}(x) = 1 - P_\theta(y^* | x_{\text{new}}). \tag{3}$$

Although the least confidence has been proven to be useful, it still has disadvantages when the model is only unconfident in one class, which will lead to a data imbalance problem. At the same time, the parameters in the intermediate model will skew toward one class. Margin sampling is capable of solving this problem. Instead of only focusing on the probability of one class, margin sampling also calculates the difference of probabilities between the first possible label (y_{first} in equation (4)) and the second possible label (y_{second} in equation (4)), as shown in equation (4). The sample with the least difference means the model is also confused as to which label this sample truly belongs to. Those samples will be selected under this strategy.

$$\text{Sample}_{\text{selected}} = \text{argmin}_x \left(P(y_{\text{first}} | x_{\text{new}}) - P(y_{\text{second}} | x_{\text{new}})\right). \tag{4}$$

Moving the margin sampling one step further, Shannon entropy allows us to consider the probabilities from all the possible classes in a classification task. In the field of information theory, entropy is a popular measurement of the randomness of a system. In the study of active learning, for each iteration, the Shannon entropy is calculated over all of the predicted label probabilities, as shown in equation (5). The higher the entropy value, the more uncertainty there will be. Under entropy sampling, samples with the highest entropy value will be selected for the annotation process.

$$\text{Sample}_{\text{selected}} = \text{argmax}_x - \sum_i p(y_i | x_{\text{new}}) \log p(y_i | x_{\text{new}}). \tag{5}$$

The main idea behind K-center-greedy is to select K points, which can represent the whole distribution of the unlabeled data pool. The K-center-greedy method starts with initiating the centroid with one randomly selected data point. For each iteration, the centroid is updated by adding the data point with the longest distance to the original centroid. The distance between one point and the centroid is calculated based on its distance from the nearest center point. Formally, we can denote existing pool s^1, labeled set s^0, and a budget b for each iteration, then the idea of K-center-greedy can be defined as follows :

$$\min_{s^1 : |s^1| \le b} \max_i \min_{j \in s^1 \cup s^0} \Delta(x_i, x_j). \tag{6}$$

Representative cluster sampling improves the K-center-greedy by selecting new points from the margin of the classes instead of from the whole data set. K-center-greedy has been proven to be successful, but its performance can be contaminated if there are too many outliers in the data. By only selecting new samples from the margin of each class, this situation can be greatly alleviated.

3.4. Random Forest. A good classifier can greatly improve the effectiveness of active learning. Therefore, a classifier suitable for PVCs data is the core component of the entire algorithm.

Random forest classifier is more commonly used for medical data because of its remarkable ability to resist training overfitting, which makes it a perfect choice for downstream active learning approaches. Since the output probability of the classifier is one major criterion for selection strategies, bini impurity is used as the criteria for splitting the nodes when the tree is building in this work.

The overall algorithm flow of this paper is shown in Table 3:

4. Results and Discussion

In this paper, we aimed to investigate the influence of active learning in terms of PVCs detection in three aspects: Initiation training data creation, different selection strategies, and different choices for updating the weights for CAE. This section presents the results of those experiments.

4.1. Experiment Data. To evaluate the proposed approach, we adopted the MIT-BIH arrhythmia database [26]. In this database, 48 30-minute two-channel ECG recordings were collected from 47 patients from 1975 to 1979. Those recordings were sampled at 360 Hz. In the annotation aspect,

Table 3: Overall algorithm flow of PVC detection.

```
Algorithm ALPVCsD
Input: Unlabeled data set Z = (x_i, y_i), i = 1, ..., N,
       Number of initial selected samples: K,
       Number of samples selected during iteration: i.
Output: Labeled data set L = (a_j, b_j), j = 1, ..., M (0 < M < N),
       Trained classifier model: random forest
Initialize: centroids = [one random selected points]
       Z_len ⟵ length (Z)
       L_len ⟵ length (L)
       Feature extraction from data by self convolution encoder
For j from 1 to K − 1
    Distance ⟵ { }
    For t from 1 to length (centroids)
       compute distance of "point" in Z to centroids and store the point P with minimum distance
    End centroids = centroids + P
End
L ⟵ L ∪ centroids (annotated the K data by Oracal)
Z ⟵ Z\centroids
Z_len ⟵ Z_len − K
L_len ⟵ L_len + K
Repeat:
    Training random forest with centroids.
    If the number of unlabeled data <1/2 N:
       Break
  For t in Z
       calculate the minimum confidence using equation (3) and store it in Q
    End
    P_data ⟵ Sort Q and select the smallest 50 data
    L ⟵ L ∪ P_data
    Z ⟵ Z\ P_data train random forest use the 50 labeled data
    Z_len ⟵ Z_len − 50
    L_len ⟵ L_len + 50
End
```

both the QRS complexes are annotated automatically first and then reviewed by a human expert, and beat-level arrhythmia types are also annotated by human experts into ten categories, in which normal beats and PVCs are the two populated classes. In this paper, we used the first 20 recordings, indexed from 100 to 125, as the training set and the remaining 24 records, indexed from 200–234 as the testing set. Figure 4(a) represents a normal ECG, and Figure 4(b) represents an ECG of PVCs beat. Ventricular premature beats occur before the sinus node impulse reaches the ventricle, at any part of the ventricle or ectopic rhythm point of the ventricular septum, and an electric pulse is sent out in advance, causing ventricular depolarization, called ventricular premature contraction, or PVC for short. The summary is: heart attacks resulting from untimely impulses originating in the ventricles are the most common arrhythmia.

The data preprocessing procedure follows the same method as in [5]. Each recording is split into beat-levels according to the position of the labeled R peak. The segment is then constructed using 89 data points before the R peak and 160 data points after. In this way, the input shape for the downstream classifier is 250, in which the position of the R peak is located at 90. A standard normalization is then applied on each input of length 250 to obtain the new segment with all data points valued between 0 and 1.

4.2. Evaluation Index. The problems studied in this paper belong to classification problems. The common evaluation indexes of classification problems include accuracy P, recall R, and F_1 score. The confusion matrix is needed to calculate the above indexes, and the confusion matrix [31] is shown in Table 4:

Accuracy is the score of the classifier correctly predicted in all samples [32]. F_1 value is a comprehensive consideration of precision and recall, as shown in equation (7). It can reflect the classification performance more comprehensively, and so it is the main evaluation index to measure the experimental effect in this paper.

$$\begin{aligned} \text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F_1 &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \end{aligned} \tag{7}$$

In the metrics presented above, although the accuracy rate can judge the overall correct situation, it cannot be used as a good indicator to measure the result when the sample is

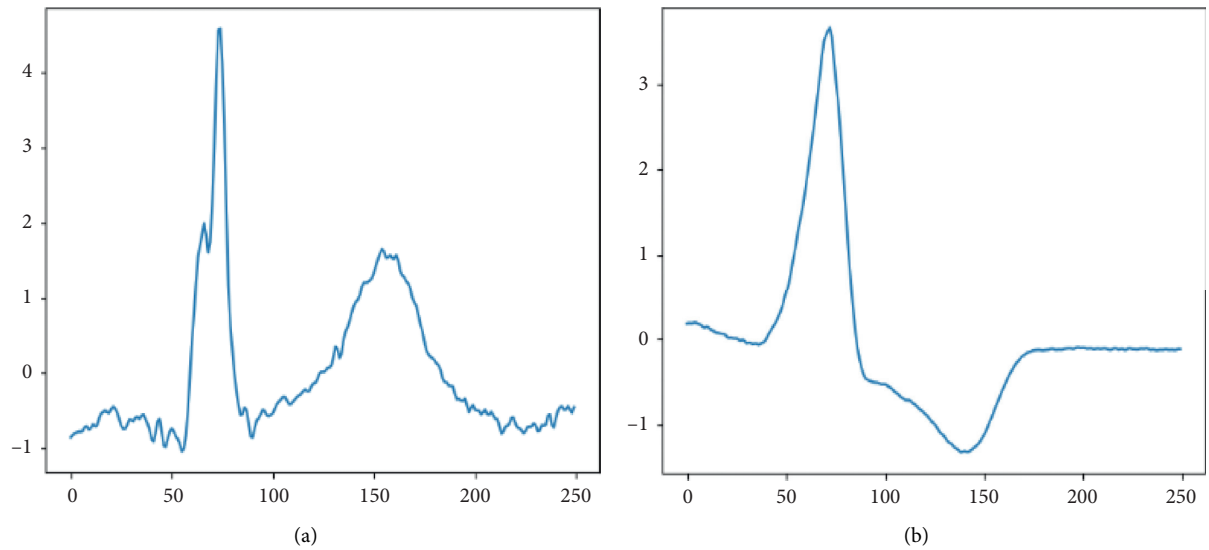


Figure 4: G Data preview of normal sinus rhythm and PVCs within one beat. (a) Normal beat. (b) PVC beat.

Table 4: Confusion matrix of the classification problem.

	Test positive	Test negative
Condition positive	True positive (TP)	False negative (FN)
Condition negative	False positive (FP)	True negative (TN)

not balanced. Precision and accuracy may seem similar, but they are two entirely different concepts. Precision represents the precision of the prediction in the positive sample results, but the accuracy rate represents the overall correctness of the prediction, including positive samples and negative samples. The recall rate is for our original sample, and it indicates how many positive examples in the sample are predicted correctly. F_1 scores are a harmonic measure of accuracy and recall.

4.3. Experiment Setup. As the first step in active learning, preparing the initial training set is crucial. In most of the work, the initial training set is selected randomly. One reason is that they ignored the importance of the training data at the start point. Another realistic reason is that, in a real-world situation, the chance of selecting the initial training data is not always available in many tasks.

The benchmark classifier maintained in the learning module must have a certain classification accuracy. Consequently, the benchmark classifier must be trained initially before active learning. The key to solving the problem is how to construct a high-performance initial training sample set. Overall, the initial training set selected at random is not representative, and the initial training set composed of representative samples is a prerequisite for training a high-precision benchmark classifier, and it can also speed up the active learning process more effectively.

A method based on clustering or distance similarity measurements is a common method for selecting representative examples. K-medoids form an initial training set; hierarchical clustering by example selection, K-means, and other measures have accelerated the process of active learning to varying degrees. The classification surface of the benchmark classifier is not far from the real classification surface from the beginning, avoiding the situation wherein the classification surface stays in the wrong direction for a long time. However, K-medoids construct the initial training set and hierarchical clustering sample selection is more suitable for image processing. The K-means method requires a manual setting of K, which leads to inaccurate algorithms.

As described in the method section, in this paper, we aimed to compare the difference between random initiation and selection using k-means++.

Active learning starts with training the classifier on initial training data. As with most work, the initial training data are generated by random selection. In the paper, we introduced an additional method called k-means++. The k-means++ algorithm can select the first K to initiate data points, which can represent the distribution of the whole data set, as described in the method section. The results are shown in Figure 5, where we observe that there is a difference between random and k-means++. The initial training data size, which is 3000 in our study, is not large enough, which makes the difference between the two distributions not as huge as we expected. Another rational reason could be that only the least confidence was incorporated, which may introduce bias into this experiment. For the next steps of active learning in this paper, k-means++ is adapted to initiate the training data. However, in general, the accuracy, recall rate, and F_1 score of data initialization with the K-means ++ method are better than random initialization. As a processing feature, in our convolutional autoencoder, the dimensionality of the hidden space in the hidden layer is assumed to be 25, which was shown to be sufficient to represent the original signal of 250 length [5, 33]. We have already introduced it in the section on convolutional autoencoders.

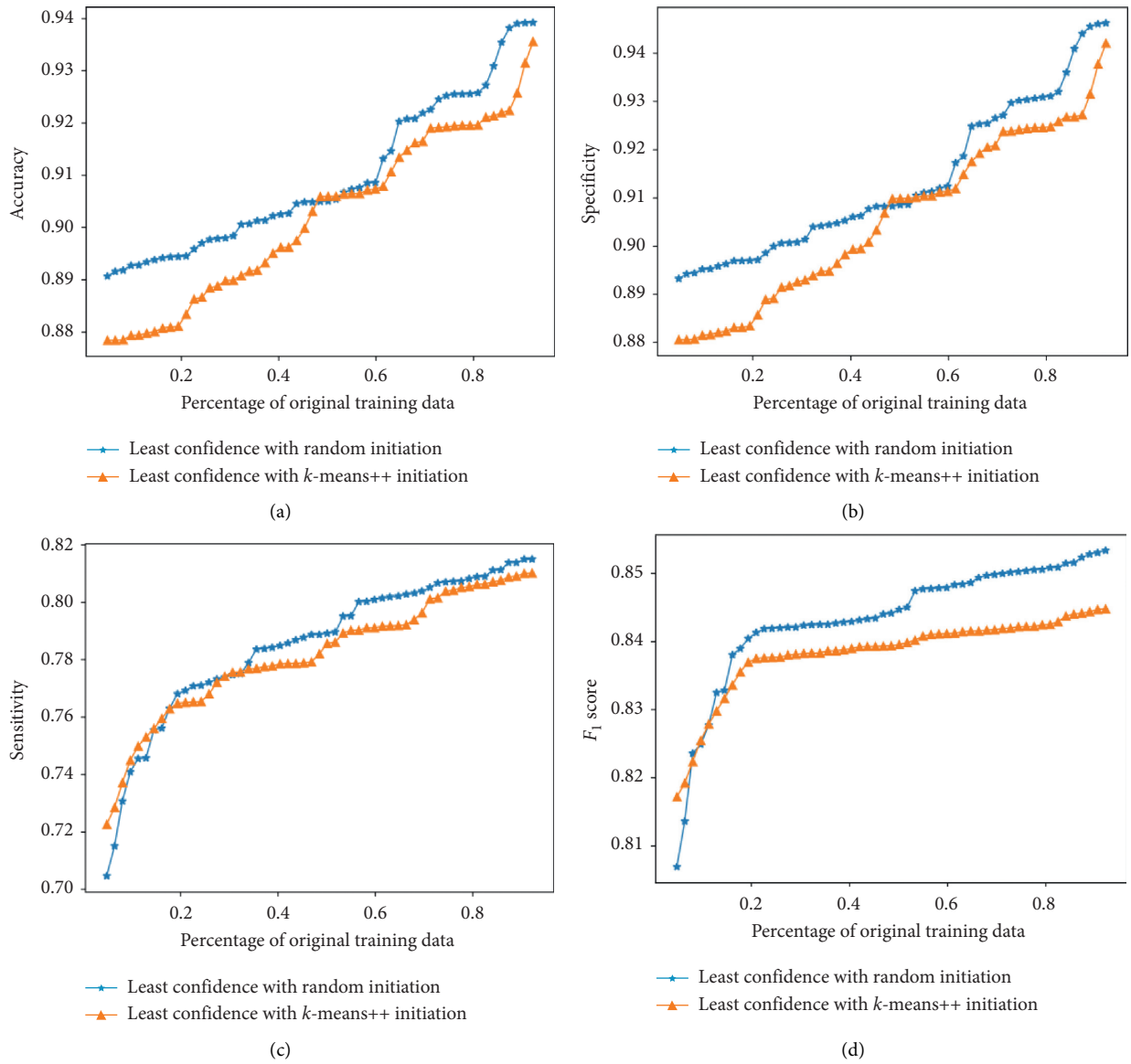


Figure 5: Performance comparison of different data initialization methods.

As the most important part of active learning, different selection strategies are well investigated in this experiment. As described in the method section, the three most classical strategies based on uncertainty sampling have been tested in this experiment.

Uncertainty sampling is one important aspect in the design of selection strategies. In this paper, we investigated three classic uncertainty sampling methods: least confidence sampling, margin sampling, and entropy sampling. Besides, random sampling was also included as a control approach. The performance of each method is reported in Figure 6. One thing worth noting is that there is an imbalance problem in both training and testing data, in which normal sinus rhythm has a higher prevalence. Instead of being measured by accuracy, the F_1 score is more appropriate in this situation. We can observe that the least confidence and margin sampling have similar performance across all metrics. In terms of F_1 score and sensitivity, the least confidence and margin sampling have the best performance than random sampling and entropy sampling. This phenomenon demonstrates that selecting a more informative sample can help achieve a better performance than random selection. Among all the four selection strategies, entropy sampling has the least performance and is even worse than random selection. The most plausible reason is that entropy sampling is more vulnerable to multiclass classification. From a theoretical perspective, the least confidence and margin sampling only focus on the classes with the best or the second-best prediction probability. However, the entropy sampling takes care of all the possible classes, which is a huge imbalance in our data set. This problem can be further alleviated by narrowing down the task from multiclass classification to binary because these three strategies were proven to be theoretically equal in the binary tasks.

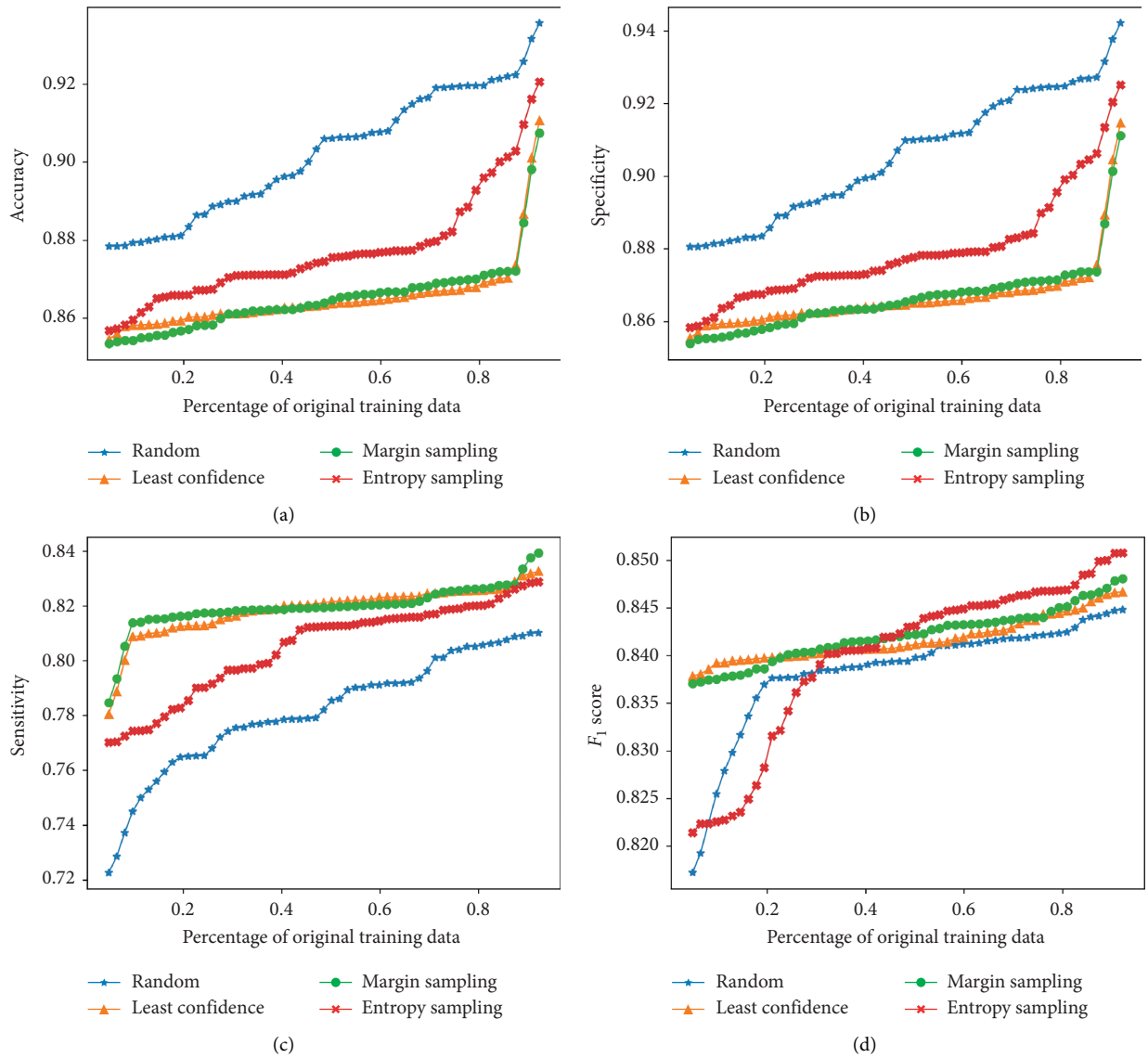


Figure 6: Performance comparison between different uncertainty selection strategies.

In addition to the uncertainty sampling, two geometry-based selection strategies were also investigated in this paper, k-center-greedy and representative cluster sampling.

Another important aspect of the active learning selection strategy is diversity. We studied two different approaches, k-center-greedy and its advanced version, representative clustering sampling, as described in the method section. The results are reported in Figure 7. We can observe that the representative clustering sampling outperforms k-center-greedy in all the metrics, which is expected. However, both methods are sampling K points that can represent the whole original distribution. The k-center-greedy selects points from all the data sets, which makes it more vulnerable to the outlier points. However, the representative clustering sampling would calculate the margin of each class and only select samples near the margins, by which the outlier problem could be heavily alleviated.

Uncertainty and diversity are two important aspects that active learning methods are trying to capture. In this paper, we studied five different selection strategies that cover both uncertainty and diversity. In Table 5, we list the performance of each strategy at the level of half of the original data being trained.

As discussed before, the ultimate goal of active learning is trying to reduce human annotation effort by keeping a similar performance at the same time. In this paper, we compared all the applied strategies trained on half the size of the original data set with the performance of the same classifier trained on the full data set. The results are displayed in Table 5. The best F_1 score is achieved with the least confidence, which is even better than training on the full data set, which demonstrates that active learning can help reduce human annotation without compromising on the performance. In terms of sensitivity, the classifier trained on the full set has the

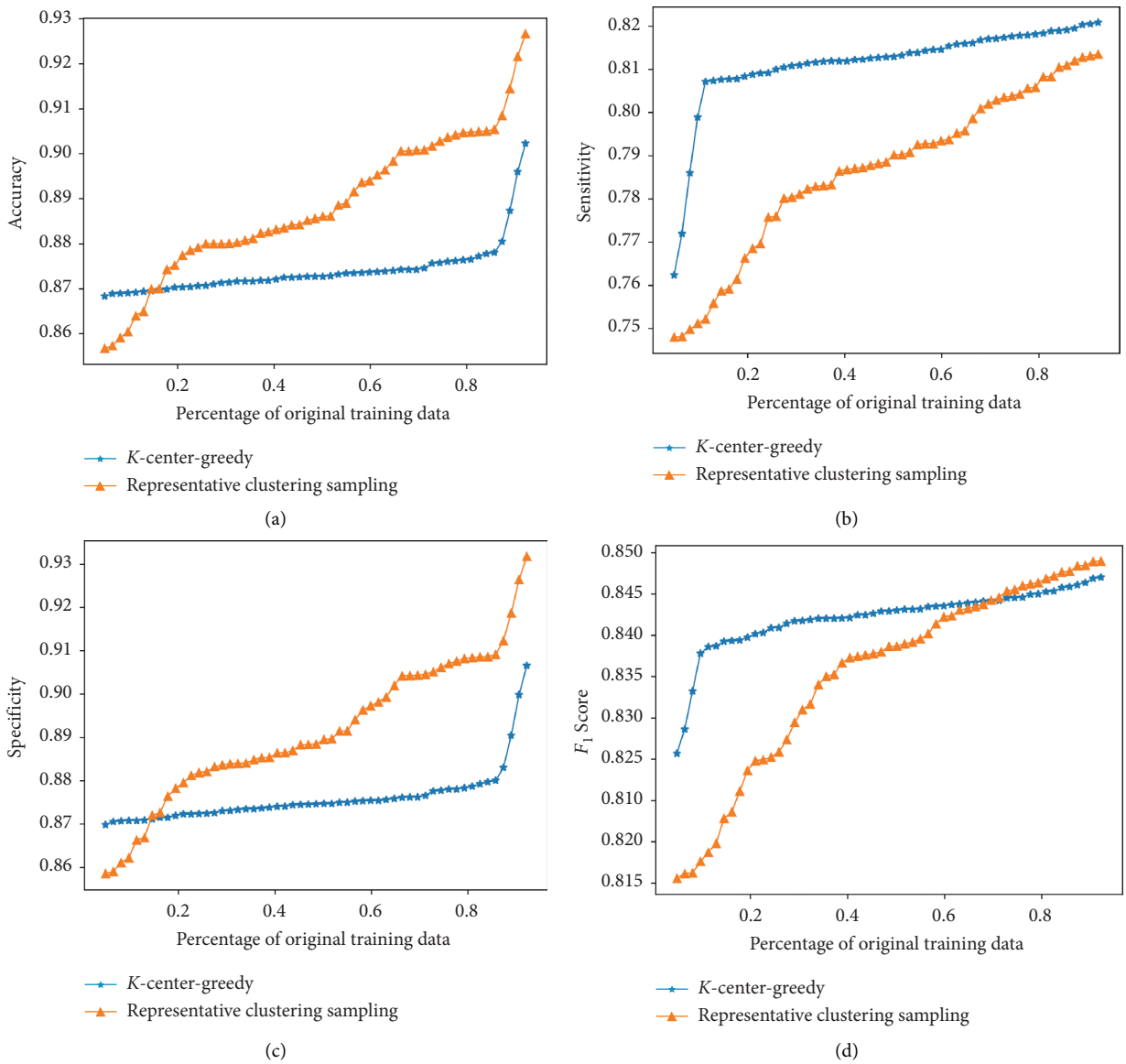


Figure 7: Performance comparison between different diversity selection strategies.

Table 5: Each strategy was trained to perform at the level of half of the original data.

	Random	Least confidence	Margin sampling	Entropy sampling	K-center-greedy	Rep-cluster	Train on the full data set
Accuracy	0.9068	0.8621	0.8485	0.8738	0.8648	0.8852	**0.9385**
Sensitivity	0.786	**0.8470**	0.8232	0.77267	0.81279	0.786	0.7704
Specificity	0.9106	0.8626	0.8493	0.8769	0.8663	0.8883	**0.9562**
F_1 score	0.8237	**0.8547**	0.8360	0.8215	0.8387	0.8341	0.8533

lowest sensitivity value. One possible reason is that there are too manypositive samples in the data set and too few negative samples. This reason is also proven by the highest specificity achieved by the full set training. This phenomenon indicates that the classifier is better at detecting normal sinus rhythms than PVCs.

The fundamental idea behind active learning is to find out the most informative samples from the original training data, which can help the classifier achieve similar performance with much less human effort involved. In the study, as the comparison factor, we first train our model on the whole training data set to be the reference for the downstream experiment. Following the same strategy described above, we can get similar results as [5]. The results are shown in Tables 6 and 7. Table 6 shows the representation of the two methods in the whole data set, and Table 7 shows the confounding matrix of the two methods on different data sets. As shown in Table 6, our accuracy and

Table 6: The performance of our method in the whole training data.

		Accuracy	PVCs sensitivity	PVCs specificity	PVCs F_1 score
Full database	Work at [5]	98.43	85.64	98.90	0.9179
	Our approach	96.49	81.50	97.56	0.8881
Test set	Work at [5]	87.80	86.65	88.09	0.8736
	Our approach	93.85	77.04	95.62	0.8533

Table 7: Confusion matrix of two methods on different data sets.

	Full data set				Test data set			
	Work at [5]		Our approach		Work at [5]		Our approach	
	True normal	True PVC	True normal	True PVC	True normal	True PVC	True normal	True PVC
Detected normal	50483	299	93734	1266	50542	810	49943	1266
Detected PVC	559	3334	2283	5595	501	2823	2283	4249

specificity are both higher than the existing algorithms, and the F_1 score is not far from the existing algorithm. The experimental results show that our algorithm can indeed play a significant role in PVCs detection of small samples.

5. Discussion

Since active learning engages with a large number of unlabeled samples and a small number of labeled samples, its advantages are evident in the advantages of traditional supervised learning. This paper conducts research on MIT's ECG data set. Research findings show that active learning techniques can effectively reduce the number of high-quality training samples required to build a classifier. On the basis of not affecting the generalization performance of the classifier, it can effectively reduce the burden of human experts.

Nonetheless, the results of this research have certain limitations. There are still many problems to be solved in the design of sampling strategies, algorithm theory, and practical applications. First, from the perspective of the algorithm architecture of this article, for new tasks how to select new instances and label instances and which features are selected to be labeled by human experts, in order for a highly versatile selection strategy.

These problems are worthy of further study, and secondly, whether the classifier proposed in this article can be replaced by a deep learning algorithm remains to be further studied. An in-depth study of feature selection [34] and classification algorithms will effectively improve the accuracy of recognition. Consequently, improvements to feature algorithms are also our next research focus.

6. Conclusions

In the medical research area, massive human annotation efforts are necessary in order to achieve higher detection performance from supervised machine learning algorithms. Active learning is a promising technique that utilizes the output of a trivial classifier to select the most informative samples for the request of annotation. By active learning, similar performance could be achieved with much less human annotation.

In this work, two main aspects of selection strategy, uncertainty and diversity, were well explored in the task of PVCs detection on beat-level ECG data. One convolutional autoencoder was trained to extract the features automatically. These data-driven features are then fed into a random forest classifier. During each iteration of active learning, the same algorithm is trained on the updated training data. It can be seen from the experimental resultsthat the F1 value of the least confidence sampling algorithm is better thanother algorithms. In addition, by comparing different active learning methods trained on half of the original data size with the same classifier trained on the full set, the performance of least confidence is still better than the full set one, which demonstrates that active learning works perfectly for the task.

Overall, the experimental effect and the sensitivity of our method are higher than the existing algorithms, demonstrating the superiority of active learning in PVC-detecting techniques. In F_1, active learning and existing methods can fundamentally improve the efficiency of relevant work. We will continue to improve our feature engineering in the coming period, deeply studying the distribution of data as we design our self-encoder, and working to ensure that the impact of active learning exceeds that of the existing work.

Data Availability

The data sets used in this paper to produce the experimental results are publicly available. ECG recordings can be downloaded from https://www.physionet.org/content/mitdb/1.0.0/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant no. 62002077), in part by the China Postdoctoral Science Foundation (Grant no. 2020M682657), and in part by Guangdong Basic and Applied Basic Research Foundation (Grant no. 2020A1515110385).

References

[1] C. L. Stanfield and W. J. Germann, *Principles of Human Physiology*, Pearson Benjamin Cummings, London, UK, 2008.

[2] H. J. L. Marriott, G. S. Wagner, and D. G. Strauss, *Marriott's Practical Electrocardiography*, Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, PA, USA, 2014.

[3] J. S. Geddes and H. R. Warner, "A PVC detection program," *Computers and Biomedical Research*, vol. 4, no. 5, p. 493, 1971.

[4] M. S. Manikandan, B. Ramkumar, P. S. Deshpande, and T. Choudhary, "Robust detection of premature ventricular contractions using sparse signal decomposition and temporal features," *Healthcare Technology Letters*, vol. 2, no. 6, pp. 141–148, 2015.

[5] M. Gordon and C. Williams, "PVC detection using a convolutional autoencoder and random forest classifier," in *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, pp. 42–53, Fairmont Orchid, Hawaii, January 2019.

[6] P. Bachman, A. Sordoni, and A. Trischler, "Learning algorithms for active learning," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 477–486, Sydney, Australia, August 2017.

[7] B. Settles, *Active Learning Literature Survey*, University of Wisconsin-Madison Department of Computer Sciences, Madison, WI, USA, 2009.

[8] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proceedings of the International Conference on Machine Learning (PMLR)*, Sydney, Australia, August 2017.

[9] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5972–5981, Seoul, Korea, October 2019.

[10] L. Yang, "Suggestive annotation: a deep active learning framework for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Quebec City, QC, Canada, September 2017.

[11] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 93–102, Long Beach, CA, USA, May 2019.

[12] V. Xia, N. Jaques, S. Taylor, S. Fedor, and R. Picard, "Active learning for electrodermal activity classification," in *Proceedings of the 2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–6, IEEE, Philadelphia, PA, USA, December 2015.

[13] G. Wang, C. Zhang, Y. liu et al., "A global and updatable ECG beat classification system based on recurrent neural networks and active learning," *Information Sciences*, vol. 501, 2019.

[14] W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler, *The Power of Ensembles for Active Learning in Image Classification*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (VBPR)*, pp. 9368–9377, Salt Lake City, UT, USA, June 2018.

[15] H. Ranganathan, H. Venkateswara, S. Chakraborty, and S. Panchanathan, "Deep active learning for image classification," in *Proceedings of the 2017 IEEE International Conference on Image Processing*, Beijing, China, September 2017.

[16] G. Yang, Q. Luo, Y. Yang, and Y. Zhuang, "Deep Learning and Machine Learning for Object Detection in Remote Sensing Images," in *Proceedings of Signal and Information Processing, Networking and Computers. (ICSINC 2017)*, pp. 249–256, Chongqing, China, December 2017.

[17] S. Priya, S. Singh, S. Kumar Dandapat, K. Ghosh, and J. Chandra, "Identifying Infrastructure Damage during Earthquake Using Deep Active Learning," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19)* pp. 551-552, New York, NY, USA, August 2019.

[18] S. Sun, P. Zhong, H. Xiao, and R. Wang, "An MRF model-based active learning framework for the spectral-spatial classification of hyperspectral imagery," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 1074–1088, 2015.

[19] P. H. Gosselin and M. Cord, "Active learning methods for interactive image retrieval," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1200–1211, 2008.

[20] B. Hou, J. Yang, P. Wang, and R. Yan, "LSTM-based autoencoder model for ECG arrhythmias classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1232–1240, 2019.

[21] M. Jangra, S. K. Dhull, and K. K. Singh, "ECG arrhythmia classification using modified visual geometry group network (mVGGNet)," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 3, pp. 3151–3165, 2020.

[22] G. Wang, C. Zhang, Y. Liu et al., "A global and updatable ECG beat classification system based on recurrent neural networks and active learning," *Information Sciences*, vol. 501, pp. 523–542, 2019.

[23] Y. Xia and Y. Xie, "A novel wearable electrocardiogram classification system using convolutional neural networks and active learning," *IEEE Access*, vol. 7, pp. 7989–8001, 2019.

[24] J. Pestana, D. Belo, and H. Gamboa, "Detection of abnormalities in electrocardiogram (ECG) using deep Learning," in *Proceedings of 13th International Conference on Bio-inspired Systems and Signal Processing(BIOSTEC 2020)*, pp. 236–243, Valletta, Malta, February 2020.

[25] D. L. Lustgarten, G. Rajagopal, J. Reiland, J. Koehler, and S. Sarkar, "Premature ventricular contraction detection for long-term monitoring in an implantable cardiac monitor," *Pacing and Clinical Electrophysiology*, vol. 43, no. 5, pp. 462–470, 2020.

[26] P. Novotna, T. Vicar, M. Ronzhina, J. Hejc, and J. Kolarova, "Deep-Learning Premature Contraction Localization in 12-Lead ECG from Whole Signal Annotations," in *Proceedings of the 2020 Computing in Cardiology Conference*, pp. 1–4, Rimini, Italy, September 2020.

[27] M. Shafiq, Z. Tian, M. Bashir, and X. Du, "Corrauc: a malicious Bot-IoT traffic detection method in iot network using machine learning techniques," *IEEE Internet of Things Journal*, vol. 99, p. 1, 2020.

[28] M. Shafiq, Z. Tian, Y. Sun, X. Du, and M. Guizani, "Selection of effective machine learning algorithm and Bot-IoT attacks traffic identification for internet of things in smart city," *Future Generation Computer Systems*, vol. 107, pp. 443–442, 2020.

[29] G. Dizaji, Kamran, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional

autoencoder embedding and relative entropy minimization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5736–5745, Venice, Italy, October 2017.

[30] S. Pan, "Adversarially regularized graph autoencoder for graph embedding," in *Proceedings of the IJCAI International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, July 2018.

[31] Y.-X. Li, C. L. Tan, X. Ding, and C. Liu, "Contextual post-processing based on the confusion matrix in offline handwritten Chinese script recognition," *Pattern Recognition*, vol. 37, no. 9, pp. 1901–1912, 2004.

[32] T. Xia and X. Chen, "A discrete hidden Markov model for SMS spam detection," *Applied Sciences*, vol. 10, no. 14, p. 5011, 2020.

[33] B. M. Mathunjwa, Y.-T. Lin, C.-H. Lin, M. F. Abbod, and J.-S. Shieh, "ECG arrhythmia classification by using a recurrence plot and convolutional neural network," *Biomedical Signal Processing and Control*, vol. 64, Article ID 102262, 2021.

[34] M. Shafiq, "Data mining and machine learning method for sustainable cities traffic classification," *Sustainable Cities and Society*, vol. 60, Article ID 102177, 2020.

Hindawi
Scientific Programming
Volume 2021, Article ID 5542852, 8 pages
https://doi.org/10.1155/2021/5542852

Research Article

TAME^{C}: Trusted Augmented Mobile Execution on Cloud

Syed Luqman Shah,[1] **Irshad Ahmed Abbasi**,[2] **Alwalid Bashier Gism Elseed**,[2] **Sikandar Ali**,[3,4] **Zahid Anwar**,[5] **Qasim Rajpoot**,[5] **and Maria Riaz**[5]

[1]*Department of Information Technology, University of Haripur, Haripur, Pakistan*
[2]*Department of Computer Science, Faculty of Science and Arts at Belgarn, University of Bisha, P.O. Box 60, Sabt Al-Alaya, Bisha 61985, Saudi Arabia*
[3]*Department of Computer Science Technology, China University of Petroleum-Beijing, Beijing 102249, China*
[4]*Beijing Key Lab of Petroleum Data Mining, China University of Petroleum-Beijing, Beijing 102249, China*
[5]*School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan*

Correspondence should be addressed to Irshad Ahmed Abbasi; aabasy@ub.edu.sa and Sikandar Ali; sikandar@cup.edu.cn

Received 21 January 2021; Revised 8 February 2021; Accepted 17 February 2021; Published 8 March 2021

Academic Editor: Shah Nazir

Cloud computing has emerged as an attractive platform for individuals and businesses to augment their basic processing capabilities. Mobile devices with access to Internet are also turning towards clouds for resource-intensive tasks by working out a trade-off between resources required for performing computation on-device against those required for off-loading task to the cloud. However, as with desktop clients, mobile clients face significant concerns related to confidentiality and integrity of data and applications moved to and from the cloud. Cloud-related security solutions proposed for desktop clients could not be readily ported to mobile clients owing to the obvious limitation in their processing capabilities and restrained battery life. We address this problem by proposing architecture for secure exchange and trusted execution between mobile devices and cloud hosts. We establish a symmetric-key-based secure communication channel between mobile and cloud, backed by a trusted coordinator. We also employee a Trusted Platform Module- (TPM-) based attestation of the cloud nodes on which the data and applications of mobile device will be hosted. This gives a comprehensive solution for end-to-end secure and trusted interaction of the mobile device with cloud hosts.

1. Introduction

The enhanced capabilities, improved connectivity, and increasing accessibility of smart phones have triggered a significant growth in the development of mobile applications. Mobile devices are equipped with a wide array of sensors and gadgets which constantly challenge the developers to come up with innovative products to capture the attention of mobile users as well as the market share. Application stores [1,2] are filled with applications that not only appeal to the entertainment-savvy users, but also cater to the information-sensitive domains of finances and health management.

Mobile devices have successfully captured a major chunk of users' attention away from the desktop machines and offer competitive Internet connectivity to add to their appeal. However the catch lies in the constrained processing capabilities and limited battery life of these devices. This is where mobile devices have to look towards their more grounded and plugged computing counterparts for assistance. Since the mobile devices can connect to more powerful machines via GPRS, Wi-Fi, and related networking technologies, it is a natural consequence to try augmenting the processing capabilities of mobile devices by off-loading resource-intensive tasks to resource-rich machines. Cloud computing [3] offers an attractive platform of choice for hosting resource-intensive applications on behalf of mobile devices. Solutions have already been proposed for augmented execution of mobile applications on clouds [4]. Resource-intensive tasks can be off-loaded to the cloud at

various levels of granularity by working out a trade-off between resources required for performing computation on-device against those required for off-loading task to the cloud.

Mobile phone users may utilize the software services hosted on cloud such as Google Apps [5] or even deploy a clone of the mobile phone on the cloud such as Amazon EC2 [6] garnering the benefits of closed-box execution of mobile applications. However, as with other cloud clients, the concerns of privacy and security associated with sending data/applications to the cloud prevail [7]. Even for an average user, the mobile applications and the data on which they operate are of very personal and private nature. A root-of-trust and associated chain-of-trust needs to be established to ensure the integrity and confidentiality of mobile users' data and applications running on the cloud. In case of desktop clients, standardized solutions based on remote hardware-based attestation in the form of Trusted Platform Modules (TPMs) are available [8] and may be extended to set up trusted clouds [9]. However, for mobile devices, these solutions are not applicable due to lack of hardware-based attestation support.

In this paper, we propose a model for securely hosting and executing mobile applications on clouds. We have extended the notion of secure clouds to mobile clients by providing secure communication and execution channels while keeping in view the practical limitations of mobile devices. In Sections 2 and 3, we explain the related work and background on which we build our solution. Section 4 describes our proposed security architecture in detail giving insight to our approach for ensuring secure communication and privacy-protecting data processing. Section 5 gives the security evaluation of the proposed system and summarizes the current progress. Section 6 concludes our work highlighting the future directions.

2. Related Work

The need to augment computing capabilities of a mobile device arises not only to host resource-heavy user applications but also to provide integrity measures to keep the device itself in a secure state. Such measures include running periodic virus scans and performing checks on the integrity of installed mobile applications [10]. The significance of these integrity measures has increased greatly with the widespread use of downloaded mobile applications [11] that may compromise mobile users' data and privacy. Keeping data synchronized between mobile and other devices of a user is another long-standing usage of mobile applications that augments mobile phone capabilities. CloneCloud [12] is a prominent effort in this regard for porting the execution of applications from mobile phones to their "clones" in the clouds, any time during execution. The concept of "elastic execution" proposed by the authors builds on the provision for migrating code running on a host virtual machine (VM) to a target VM at runtime. Since most mobile phones run applications on an underlying application-layer VM [13], this solution has wide-scale applicability. Their solution essentially opens up the Infrastructure as a Service (IaaS) capabilities of the cloud to mobile users. However, it does not focus on the security aspects of hosting data or complete clone of the mobile device on the cloud.

For desktop clients of clouds, a model for trusted clouds has been proposed [9] which ensures confidential execution of guest VMs on clouds. The proposed model extends the notion of remote attestation via TPMs to cloud computing paradigm. A cloud consists of a number of nodes on which the client's VM could be hosted. Every time a VM has to be launched or migrated from one node to the other in the cloud, it is ensured that the node is trustworthy. The focus is on enabling the IaaS providers such as Amazon EC2 to provide closed-box execution environment to their clients. However, the same approach cannot be readily applied to mobile phone clients due to their limited resources and lack of support for hardware-based attestation. In [14], the authors propose an architecture exploiting sealed storage mechanism that attests the cloud node prior to launching customer data and ensures confidentiality of customers' data even if the cloud node becomes compromised. In [15], the authors propose trusted attestation architecture for Infrastructure as a Service. The architecture ensures the attestation of virtual machine, hypervisor, and host operating system using the Trusted Platform Module (TPM) and Virtual Trusted Module (vTPM).

Standards for Mobile Trusted Modules (MTMs) [16] have been proposed for hardware-based attestation of mobile devices, and standardized hardware implementations are expected to be available soon. In [17], the authors have developed an emulator based on a subset of MTM standards for remote attestation of Android platform [13]. The work focuses on establishing trustworthiness of the applications running on Android platform including the Android Dalvik VM. Additionally, the remote attestation process is able to attest the classes loaded by the VM to establish the validity of applications running on top of the virtual machine. It is a significant improvement in the attestation process which is often limited to applications launched directly by the operating system. Their architecture can be readily ported to work with actual hardware MTMs once available. However, the confidentiality and integrity of key-pairs and hashes stored on disk cannot be ascertained. The trustworthiness of a mobile device is not guaranteed any further than the trustworthiness established via maintaining traditional key-stores at the mobile device.

In this paper, we combine the concepts presented in [9, 12] and add in the missing ingredient of trusted execution of mobile applications on clouds to provide a comprehensive end-to-end security architecture for mobile clients. In Section 4, we discuss our approach for addressing the security and trust establishment issues faced by mobile clients when hosting data and applications on cloud. Most of the work carried out in porting mobile applications to clouds focuses on Android-based mobile phones. This is due to the open platform that Android offers as well as the popularity and increasing market share of Android phones [11]. In this context, the focus of our proposed architecture is on securely augmenting capabilities of Android-based mobile devices by employing trusted cloud backend.

3. Background

3.1. Virtualization. Virtualization is a mechanism of dividing hardware or some subset of hardware of a computer system among several virtual machines (VMs), which bear resemblance to the physical system. Virtualization is carried out by a virtual machine monitor (VMM), a layer that divides the hardware. XEN [18] provides a virtualization layer that ensures the separation of VMs and executes instructions on their behalf. At the same time, XEN manages a domain (Dom0), which controls the access of VMs to physical hardware and facilitates communication among VMs hosted on the shared hardware. The virtualized system attempts to provide virtualized network and storage devices, thus providing an environment similar to that of a real system. The frontend drivers communicate with backend drivers via sockets. Normally, backend drivers are provided within Dom0 of XEN, but these drivers can also run within other virtual machines. In order to ensure the security, the applications running in one virtual machine should not be interfered by the applications executing on another VM; it is achieved by isolating the VMs by virtual machine monitor. This separation is maintained through different in-built privileged levels in the CPU. This separation can only be trusted if the whole software stack loaded into the system, right from BIOS to the VMM and over the Dom0, is relied upon. Trusted Computing Base (TCB) for a VM is formed only if the software running on the system is correct and works as expected. This issue has been addressed by measurement techniques provided by Trusted Computing Group (TCG).

3.2. TCG. The specifications published by TCG [19, 20] provide a mechanism to attest the computing platforms based on a Trusted Platform Module (TPM), a root of trust, and a co-processor mounted on the motherboard of the computer. As per latest specifications [21] TPM contains 24 Platform Configuration Registers (PCRs) that get filled with cryptographic hashes of the software stack loaded and running on the system. These PCRs are automatically established when the computer goes through the boot cycle; values in these registers can be extended later on but cannot be replaced with chosen values. The BIOS code forms the Core Root of Trust for Measurement (CRTM), and then every link in the chain is measured by the prior one. CRTM first calculates the BIOS, extends values in relevant register, and then transfer controls to the BIOS. Then, BIOS measures ROM configurations/data and extends values to the TPM registers. Similarly, all the components are measured right from BIOS to the operating system level and measurements are recorded in the PCRs. Even, up and above the OS, the applications can also be measured [22], stored in the kernel held Stored Measurement Log (SML), and then extended in the related register. This extension process ensures that the values in the PCRs cannot be manipulated and reached via any other route. PCR values are signed by Attestation Identity Key (AIK) residing within TPM, which allows the state of computer to be communicated with other parties, who then validates if the values and platform can be trusted. The validity of TPM is verified by these signing key certified by certification authority during remote attestation procedure.

3.3. Trusted Virtualization. A virtualized system enriched with additional components gets complex to measure. The conventional TCG scheme for measurement can be extended [23] to virtualized systems. It allows trust to be communicated over the full software stack of the system, which includes measurement of BIOS, bootloader, VMM, image for Dom0, additional kernel modules, and associated configuration files. The attestation of the system up to the base level can be achieved through this process; however, the end user may also want to attest the guest operating system, which can be achieved by further extending PCR measurements.

The already attested measurement agent in the Dom0 can give directions for the measurement and communication of guest operating systems. Virtual TPM (vTPM) [24] is just an extension of this idea where measurement services along with other TPM functions can be provided by each domain. These vTPMs can be made available from within Dom0 or from within a separate virtual machine linked to the VM.

4. The Architecture

The main objective of this work is to provide a mechanism for secure transfer and trusted execution of mobile users' data and applications on the cloud. In order to guarantee end-to-end security, the proposed system establishes the trustworthiness of the cloud node before giving access to the data via a secure channel. The three main participating entities are (a) the Android-based mobile client, (b) a number of cloud nodes ($N_1, \ldots, N_n$), managed by the Cloud Manager (CM), on which the processing is off-loaded, and (c) a Trusted Coordinator (TC) that acts as the trusted third party and attests the cloud nodes. TC is similar to certificate authority in the TCG adopted Direct Anonymous Attestation protocol [25].

We assume that the cloud nodes, as well as the trusted coordinator, are TPM-enabled platforms and can participate in carrying out remote attestation of the nodes. We also assume that a key-store containing symmetric-keys associated with the mobile client is available at the mobile device. Figure 1 provides an overview of our proposed system architecture. The selection of encryption and attestation technologies is made in keeping with the actual capabilities of each participating entity while ensuring a high level of trust. The Android client can generate and store symmetric keys (AES_{key}) in its key store along with the trusted *key* ($TK_{TC}{}^{public}$) of the TC. The details regarding message exchange between each set of communicating party and the role of different keys are discussed in the following sections.

The configuration files at the mobile client provide basic information related to off-loading data and applications to the cloud backend. A detailed model of elastic execution between mobile client and cloud host is given in [5, 12].

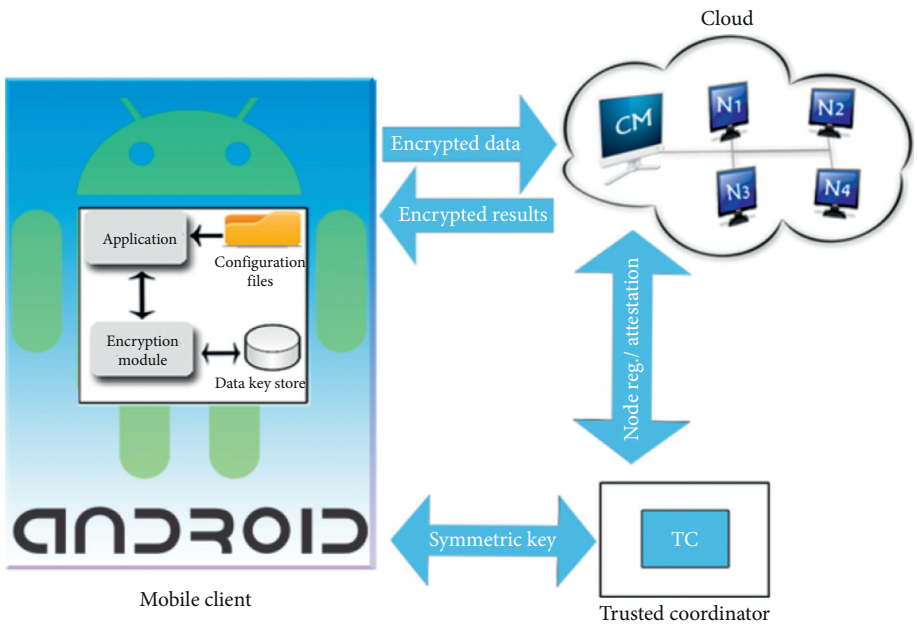


Figure 1: Overview of the system architecture.

4.1. Establishing Secure Channel. The data sent from the mobile client to the cloud nodes are encrypted using symmetric key cryptography. Owing to the security, fast speed, and low RAM requirements of AES, it is a suitable encryption algorithm for mobile clients; hence, we use AES key for encryption of the communication channel between mobile client and cloud. Moreover, AES supports encryption of large chunks of data without significant performance overhead at the client. The AES key (AES_{key}) used by the mobile client for encryption, however, is not readily made available to the cloud node receiving the encrypted message. The key is itself encrypted and a node can only gain access to it after it has successfully attested itself to the TC (explained in Section 4.3).

It is important to mention that we are interested not only in establishing a secure channel between the mobile client and the cloud node but also in restricting untrusted nodes from gaining access to the data. As the client does not know which node will be selected to process its data, we have not used SSL [26] which enables secure data communication between already trusting parties.

4.2. Registration of Cloud Nodes. In the proposed solution, registration of each cloud node is a required step that is performed before the node can be used in the cloud to perform any computations on behalf of the mobile clients. At the end of registration, the public part of an RSA key pair ($TK_{Ni}{}^{public}$, i.e., specific to a particular node) generated by a node and stored in its volatile memory is known to TC that is later used to exchange the key between node and TC as explained in next section. The reasons to store this key pair in volatile memory are (1) to require the node to get itself re-registered whenever its state is changed or is rebooted and (2) to prohibit the key to be used once a node's state is modified which is achieved because only the trusted applications are running on each node.

In our solution, both the cloud nodes ($N_1, \ldots, N_n$) and the trusted coordinator (TC) contain TPM chip for hardware attestation and hence can utilize the TCG remote attestation mechanism [19] along with registration of the cloud nodes. Instead of encrypting the complete Stored Measurement Log (SML) of the node (SML_{Ni}, i.e., specific to a particular node) and the TC (SML_{TC}), we encrypt the value of PCR-10 register containing the aggregate hash of the platform's state. This enables the receiving party to verify the SML of the sending party against the value of PCR-10 to check if a node maintains a trusted configuration [27]. Figure 2 shows the messages exchanged between the nodes and TC during the registration process.

Attestation Identity Key (AIK) is used for authentication purpose by the TPM to sign the PCR values and its corresponding public part is used by the other party to verify signatures. It is assumed that the public part of the AIK (AIK^{public}) of each node is already known to the TC and vice versa. Moreover, each node is also aware of a Trusted Key ($TK_{TC}{}^{public}$) which will be used to encrypt any data to be sent to TC that could be decrypted outside TPM, since the data encrypted with AIK can only be decrypted within the TPM.

When a node wishes to register itself with the TC, (1) it generates a nonce (n_{Ni}) and sends to the TC. Upon receiving the nonce, (2) TC signs the nonce along with the value of PCR-10 register using its AIK ($AIK_{TC}{}^{private}$). It also sends SML_{TC} and a nonce (n_{TC}) along with the encrypted content to the node. The node verifies signature of TC using AIK of TC ($AIK_{TC}{}^{public}$) and then verifies the SML of TC against the signed PCR-10 value. Since the nonce generated by the node (n_{Ni}) is also signed along with the PCR-10 contents by the TC, node can be sure that it is a fresh value reflecting the current configuration of the TC.

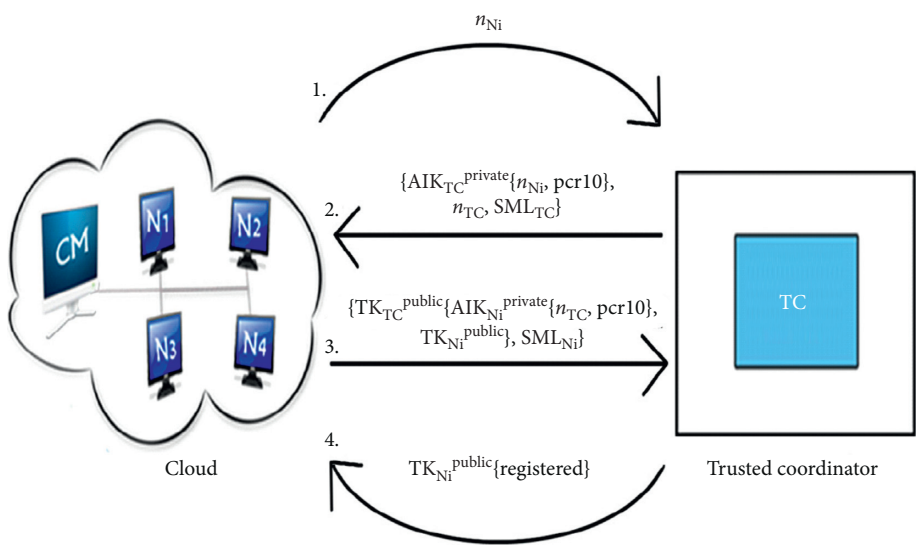


Figure 2: Registration of cloud nodes with the TC.

Afterwards, (3) the node sends the value of its PCR-10 and the nonce sent by TC (n_{TC}) signed by its AIK ($AIK_N^{private}$). It also sends the SML_{Ni} so that TC can verify the SML_{Ni} against PCR-10. The node also sends the public part of its Trusted Key (TK_{Ni}^{public}) to the TC which is stored with the TC and is used during node attestation as explained in the next section. All the message content, except for the SML_{Ni}, is encrypted using the TK_{TC}^{public} so that it is only accessible to the TC.

Upon receiving message (3), TC decrypts the contents using its $TK_{TC}^{private}$, verifies the signature on PCR-10, and matches the SML_{Ni} against the PCR-10. TC also compares the signed nonce sent back by the node to ensure that the SML_{Ni} are current. If all are verified, (4) TC sends the registration message signed by the newly received TK_{Ni}^{public} of the node. This signals the successful registration of the node with the TC along with the acceptance of the node's TK_{Ni}^{public}.

The node registration procedure is independent of the fact that the nodes host data/applications sent by a mobile client or a desktop client. Therefore, the notion of Trusted Clouds [19] can be utilized by the mobile clients in the same manner as other clients of cloud computing.

4.3. Trusted Computation on Cloud. This section explains the mechanism for carrying out trusted computation on the cloud on behalf of the mobile client. The mobile client, along with the cloud nodes, has access to the trusted key of TC, TK_{TC}^{public}. The TC also has the trusted keys of the nodes, TK_{Ni}^{public}. As mentioned in Section 4.1, the symmetric key (AES_{key})—used by the mobile client for encrypting the data sent to cloud—is made available to the node after successful attestation by the TC. This enables the mobile client to off-load its data and applications to the cloud nodes in a secure and trusted manner.

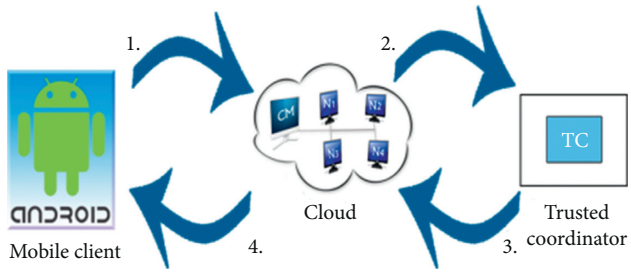
When a mobile client wishes to off-load the computation of a resource-intensive task to the cloud, (1) it encrypts the associated data using an AES_{key} before sending it to the cloud. The mobile client also needs to communicate the AES_{key} so that the receiving party can decrypt the information. The AES_{key} is encrypted using the TK_{TC}^{public} restricting the cloud node to gain access to the AES_{key} and hence the encrypted data, without first contacting the TC and proving that it is in trusted state. The mobile client also sends a nonce (n_U, i.e., nonce generated by the user) encrypted alongside the AES_{key} to ensure the freshness of the communication.

Upon receipt of the encrypted message, the node must prove to the TC that it is trustworthy so that the TC may decrypt and send the AES_{key} to the node. The various messages exchanged between communicating parties to achieve trusted computation are shown in Figure 3.

To access the AES_{key} encrypted with the TK_{TC}^{pbulic}, (2) the cloud node passes on the encrypted key and nonce sent by the mobile user n_U to the TC. It also sends a nonce n_N and node information N to the TC using TK_{TC}^{pbulic}. In order to demonstrate the fact that the node is already registered with the TC, it computes the hash of the above parameters and signs it using node's trusted key $TK_{Ni}^{private}$. If the node is registered with the TC, TC is able to decrypt this part using TK_{Ni}^{public} of the node - provided at the time of registration (explained in Section 4.2)-decrypts the AES_{key} and (3) sends back to the node the AES_{key} along with n_U and n_N encrypted with the TK_{Ni}^{public} of the node. In order to ensure that the message is from the TC, TC also computes the hash of the contents and signs it using its $TK_{TC}^{private}$. In case the node is not already registered, it is first required to register itself with the TC as explained in Section 4.2.

Once the AES_{key} is revealed to the node, it decrypts the message, performs the necessary computations on behalf of the mobile client, and (4) sends back the results encrypted by the same symmetric key (AES_{key}). The node also sends back the original nonce n_U to validate that the results are from a current computation.

It should be noted that the mobile device uses different AES_{keys} for setting up secure communication channel for



1. $\{AES_{key}\{Data\}, TK_{TC}^{public}\{n_U, AES_{key}\}\}$
2. $\{TK_{Ni}^{private}\{SHA1\{TK_{TC}^{public}\{n_U, AES_{key}\}, n_{Ni}, N\}\}, TK_{TC}^{public}\{n_U, AES_{key}\}, TK_{TC}^{public}\{n_{Ni}, N\}\}$
3. $\{TK_{Ni}^{private}\{SHA1\{n_{Ni}, n_U, AES_{key}\}\}\}, \{TK_{Ni}^{public}\{n_{Ni}, n_U, AES_{key}\}\}$
4. $AES_{key}\{n_U, N, Result\}$

Figure 3: Secure data exchange between mobile client and cloud nodes.

Table 1: Threats vs. security mechanisms.

Threat	Catered	Mechanism
Threats on cloud node/TC		
Masquerading	Yes	Each node's AIK is already known to TC
Eavesdropping	Yes	Use of only trusted software ensures that no information can be manipulated or tempered with
Message tempering	Yes	
Malware	Yes	Malware cannot remain undetected since TC checks status of each node through remote attestation process
Message replay	Yes	The nonces used ensure freshness
Threats to mobile client		
Man in the middle	Yes	The processed data are being received encrypted which can only be accessed by the relevant mobile client
Result tempering	Yes	Since the data are encrypted, no one can alter the data undetectably

off-loading different computations. Therefore, if a node having access to a particular AES key is compromised, it will still not be able to decrypt the mobile devices' data for subsequent interaction.

5. Discussion

In the proposed architecture, establishment of secure communication channel ensures the confidentiality of customer data while it is under processing or in transit. Each software component loaded on cloud node is measured and extended in the TPM registers ascertaining that cloud node cannot hide or mislead about the software or code running on it. In our solution, the cloud node gets encrypted data while the key to decrypt it is encrypted with the trusted key of TC (TC_{TK}^{public}). In order to get the key to decrypt data, cloud node has to contact TC with another trusted key (TK_{Ni}^{public}, specific to each client). TK_{Ni}^{public} is stored in the volatile memory at each node and is automatically destroyed whenever the configurations of a node changes or if the node is rebooted, requiring the node to re-register itself with TC. The use of nonce in the above protocols ensures freshness. Prior to revealing the key, TC on behalf of mobile client checks the cloud node, through remote attestation procedure, that it is in trustworthy state and is not executing any malicious software which can store or spy on mobile client data. Trusted software running at each node guarantees confidentiality of client data during computation.

The prototype implementation of the proposed system provides encouraging results; however, we believe that the performance can be further improved; hence, we are yet carrying out the experiments. From the initial results, we are convinced of the validity and usefulness of our proposed solution. The overhead caused is perfectly tolerable for large sized data, as the time taken for data processing is directly proportional to the size of the data to be processed, at the cost of security and processing provided for resource-starved devices. These factors provide strong basis for the practical implication of our proposed architecture. We provide detailed security analysis of the system by evaluating the system against the known threats at each stage in Table 1.

We have catered to the major threats on cloud node and on TC and the threats to mobile client by using the standard security mechanisms while exploiting trusted computing. The proposed solution provides end-to-end secure communication between mobile client and cloud by employing TPM-based remote attestation of cloud nodes ensuring the confidentiality and integrity of the computations performed by the cloud nodes. This demonstrates that the mobile devices can securely augment their capabilities by off-

loading computation to cloud. Moreover, our approach is readily useful for desktop clients lacking support for hardware-based attestation.

6. Conclusions and Future Direction

We have defined a mechanism for secure exchange of data between mobile devices and cloud hosts. It guarantees that only trusted cloud nodes are allowed access to the mobile device's data addressing the concerns related to confidentiality and integrity of the computations performed by the cloud node. By utilizing the concepts of elastic execution [9] and trusted clouds [12] and applying our model for secure communication and trust establishment, a comprehensive end-to-end security infrastructure has been provided for mobile clients of cloud computing.

The security architecture is tailored to the capabilities of the participating entities. At the mobile device's end, we have employed the AES symmetric key encryption to minimize the encryption overhead. Cloud nodes, on the other hand, are attested by a trusted coordinator TC to establish that they are in a secure and trusted configuration before gaining access to the symmetric key. Thus, only trusted nodes are selected to carry out the required computation on behalf of the mobile device.

The probability of sending any malicious data by rogue mobile clients is high, which can be harmful to the cloud and at the same time can steal the data of other customers, hence compromising the privacy and breaching the security of cloud [28]. Thus, future researchers would find it interesting to probe the dimensions of attesting the validity and reliability of the mobile clients. This will open a vast field of discussion for researchers to take into account protocols that enable mobile device verification prior to granting them access to cloud resources.

Recently, a new usage scenario that encompasses mobile applications and the cloud is emerging, that is, "testdriving" a potential application on the cloud prior to purchasing it. Consider the case of Amazon testdrive [29, 30] that allows users to preview any Android application in the App Store directly from their browsers by launching an emulated instance of Android on its EC2 cloud, allowing direct control from the browser (using Flash). The reader can easily imagine the usefulness of TAME^c in securing licensing and customized features in such a scenario. In fact, there are reports [31] that one of the primary reasons testdrive was disabled as soon as it was launched was fear of insecure data exchange.

In addition to implementing a complete prototype of our proposed system, a future direction could be identifying the failure models of the trust establishment mechanism itself. An important question arises that if a nontrusted node receives the encrypted data and is not able to attest itself to the TC, how will the mobile client be notified. In this case, there should be an upper bound on the time after which the mobile device should stop expecting to receive results from the cloud node. The TC may also notify the mobile device of a failed attestation attempt by a node. However, the issue still remains if the node receiving the data never contacts the TC. This and similar concerns are not necessarily specific to the mobile clients and offer an interesting area to explore. We hope that research in this direction will give useful insight for further enhancing the proposed security infrastructure.

Data Availability

The data used to support the findings of this study are provided within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the China University of Petroleum-Beijing and Fundamental Research Funds for Central Universities under grant no. 2462020YJRC001.

References

[1] H. Yih-Chun and A. Perrig, "A survey of secure wireless ad hoc routing," *IEEE Security & Privacy*, vol. 2, no. 3, pp. 28–39, 2004.
[2] Android market, 2021, https://play.google.com/store/apps.
[3] M. Armbrust, A. Fox, R. Griffith et al., "Above the clouds: a berkeley view of cloud computing," Technical Report No. UCB/EECS-2009-28, University of California, Berkeley, CA, USA, 2009.
[4] B. G. Chun and P. Maniatis, "Augmented smartphone applications through clone cloud execution," in *Proceedings of HotOS'09: 12th Workshop on Hot Topics in Operating Systems*, Monte Verità, Switzerland, May 2009.
[5] Google Apps, 2021, https://play.google.com/store/apps/.
[6] Amazon Elastic Compute Cloud (Amazon EC2), 2021, https://aws.amazon.com/ec2.
[7] M. B. Mollah, M. A. Kalam Azad, and A. Vasilakos, "Security and privacy challenges in mobile cloud computing: survey and way ahead," *Journal of Network and Computer Applications*, vol. 84, pp. 38–54, 2017.
[8] Trusted platform module, 2021, https://trustedcomputinggroup.org/resource/trusted-platform-module-tpm-summary.
[9] N. Santos, K. P. Gummadi, and R. Rodrigues, "Towards trusted cloud computing," in *Proceedings of the 2009 Conference on Hot topics in Cloud Computing: HotCloud'09*, Berkeley, CA, USA, June 2009.
[10] Mcafee, 2021, https://www.mcafee.com/en-us/index.html.
[11] Android market statistics, 2021, http://www.androlib.com/appstats.aspx.
[12] B. G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "CloneCloud: elastic execution between mobile device and cloud," in *Proceedings of the 6th European Conference on Computer Systems (EuroSys 2011)*, Salzburg, Austria, April 2011.
[13] Android DalvikVM specs, 2021, https://source.android.com/devices/tech/dalvik.
[14] G. Cheng and A. K. Ohoussou, "Sealed storage for cloud computing," in *Proceedings of the International Conference on Computer Design and Applications (ICCDA)*, Qinhuangdao, China, June 2010.

[15] X. Jin, X. Chen, C. Zhao, and D. Zhao, "Trusted attestation architecture on an infrastructure-as-a-service," *Tsinghua Science and Technology*, vol. 22, no. 5, pp. 469–477, 2017.
[16] Mobile Trusted Module, 2021, https://trustedcomputinggroup.org/resource/mobile-phone-work-group-mobile-trusted-module-specification/.
[17] M. Nauman, S. Khan, X. Zhang, and J. P. Seifert, "Beyond kernel-level integrity measurement: enabling remote attestation for the android platform," in *Proceedings of the 3rd International Conference on Trust and Trustworthy Computing (Trust 2010)*, Berlin, Germany, June 2010.
[18] P. Barham, B. Dragovic, K. Fraser et al., "Xen and the art of virtualization," in *Proceedings of 19th ACM Symposium on Operating Systems Principals (SOSP '03)*, Bolton Landing, NY, USA, October 2003.
[19] Trusted Computing Group, *TCG Pc Specific Implementation Specification*, Trusted Computing Group, Beaverton, OR, USA, 2003.
[20] C. Mitchell, *Trusted Computing (Professional Applications of Computing)*, IEEE Press, New York, NY, USA, 2005.
[21] TPM Main Specification Level 2 Version 1.2, 2021, http://www.trustedcomputinggroup.org/resources/tpm_main_specification.
[22] R. Sailer, T. Jaeger, X. Zhang, and L. van Doorn, "Attestation-based policy enforcement for remote access," in *Proceedings of the 11th ACM Conference on Computer and Communications Security (CCS 2004)*, Washington, DC, USA, October 2004.
[23] T. Garfinkel, B. Pfaff, J. Chow, M. Rosenblum, and D. Boneh, "Terra," *ACM SIGOPS Operating Systems Review*, vol. 37, no. 5, pp. 193–206, 2003.
[24] S. Berger, K. G. Goldman, R. Caceres, R. Perez, R. Sailer, and L. van Doorn, "vTPM: virtualizing the platform module," Technical Report RC23879, IBM Research, Yorktown Heights, NY, USA, 2006.
[25] E. Brickell, C. Jan, and L. Chen, "Direct anonymous attestation," in *Proceedings of the 11th ACM Conference on Computer and Communications Security (CCS '04)*, Washington, DC, USA, October 2004.
[26] Guide to SSL VPNs, 2021, http://csrc.nist.gov/publications/nistpubs/800-113/SP800-113.pdf.
[27] R. Sailer, X. Zhang, T. Jaeger, and L. V. Doorn, "Design and implementation of a TCG-based integrity measurement architecture," in *Proceedings of the 13th Conference on USENIX Security Symposium (SSYM 04)*, Berkeley, CA, USA, August 2004.
[28] E. T. Ristenpart, H. Shacham, and S. Savage, "Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds," in *Proceedings of the 16th ACM Conference on Computer and Communications Security CCS'09*, Chicago, IL, USA, November 2009.
[29] D. Etherington, Amazon's TestDrive is the real strength of appstore, 2021, https://gigaom.com/2011/03/28/amazons-testdrive-is-the-real-strength-of-appstore/.
[30] J. Kincaid, Amazon's android App store launches: test drive Apps directly from your browser, 2021, https://techcrunch.com/2011/03/22/amazon-android-app-store-3/.
[31] S. Anthony, Amazon Appstore for Android Test Drive Hands on: Surprisingly Cool, but Still US-Only, 2021, http://downloadsquad.switched.com/2011/03/28/amazon-appstore-for-android-test-drive-hands-on-surprisingly-co.

Hindawi
Scientific Programming
Volume 2021, Article ID 6618135, 7 pages
https://doi.org/10.1155/2021/6618135

Research Article

Estimation of Sea Level Change in the South China Sea from Satellite Altimetry Data

Shanwei Liu,[1] **Yue Jiao**,[1] **Qinting Sun**,[2] **and Jinghui Jiang**[1]

[1]*College of Oceanography and Space Informatics, China University of Petroleum Qingdao, Qingdao 266580, China*
[2]*School of Geosciences, China University of Petroleum Qingdao, Qingdao 266580, China*

Correspondence should be addressed to Shanwei Liu; shanweiliu@163.com and Qinting Sun; 1136613517@qq.com

Received 19 November 2020; Revised 11 January 2021; Accepted 27 January 2021; Published 16 February 2021

Academic Editor: Shaukat Ali

The South China Sea is China's largest marginal sea area, and it is rich in oil and gas mineral resources; thus, estimating its sea level changes is of practical significance. Based on linear and nonlinear sea level change characteristics, this paper decomposes 1992–2019 monthly mean sea level anomaly time series in the South China Sea into trend, seasonal, and random terms. This paper compares the seasonal autoregressive integrated moving average (SARIMA) and Prophet models for estimating the trend and seasonal terms and the long short-term memory (LSTM) and radial basis function (RBF) models for estimating random terms, and the more suitable models were selected. A Prophet-LSTM combined model was developed based on the accuracy results. This paper uses the combined model to study the effect of known data length on the experimental results and determines the best prediction duration. The results show that the combined model is suitable for short-term and medium-term estimations of 12–36 months. The accuracy at 36 months is 0.962 cm, which proves that the combined model has high application value for estimating sea level changes in the South China Sea.

1. Introduction

The South China Sea is China's largest marginal sea and is a transportation hub for maritime energy transportation; it also has abundant reserves of oil and gas resources [1]. In recent years, the sea level of the South China Sea has continued to rise [2], and, over the next 30 years, the sea level along the coast will also rise by 50~180 mm [3], which not only will have serious impacts on the natural environment, ecosystems, and social economy of the coastal areas [4] but will also pose challenges to maritime transport and energy extraction. Research on these changes can assist in measuring regional climate change, contribute disaster warning information, and provide a scientific basis for the coordination of maritime traffic and the rational exploitation of energy. Therefore, research on the estimation of sea level change trends in the South China Sea is necessary.

In terms of sea level estimation, predecessors tried different methods based on mathematical statistics, physical mechanisms, or combined model predictions according to different regions and data types. Mathematical statistics methods are mainly based on the mathematical law of sea level change time series, which is used to fit and extrapolate data [5]. Early mathematical statistical methods included simple linear regression [6], multivariate stepwise regression, maximum entropy spectrum analysis [7], and Kalman filtering [8]. The physical mechanism method considers climate change and sea temperature and salinity. Chen and others used the CCSM3 climate system model to simulate sea level changes, and the results showed that the global sea level will rise by 30 cm in the 21st century [9]. Zhang and others used the sea-atmosphere coupled model to predict the trends and spatial distribution of sea level in the South China Sea in the 21st century [10]. The use of these methods is relatively simple and does not take into account the nonstationary and uncertain characteristics of sea level changes. With the emergence of various neural network models, the estimation of nonstationary characteristics of sea level changes has become more possible; moreover, the combination of mathematical methods and neural networks can

improve accuracy. Zhao and others used a model that combined least squares and radial basis function neural network to predict sea level anomaly series in offshore China, and the reliability of the model for short-term predictions was demonstrated, and the accuracy reached 0.65 cm [11]. Among the methods selected in this article, SARIMA is widely used in epidemiological prediction [12] and the Prophet model has good performance in user traffic prediction [13]. As neural network models, LSTM and RBF have certain application value in rainfall and river flow predictions [14, 15].

Currently, few studies have focused on the estimation of sea level changes in the South China Sea. To more accurately estimate the change trends of the South China Sea, this paper divides the monthly average sea level abnormal time series according to the sea level change characteristics in the South China Sea as follows:

(i) Trend term and seasonal term combination series

(ii) Random term series

In terms of model selection, the SARIMA [12] and Prophet [16] models are suitable for fitting stationary and trending series and the LSTM [17] and RBF [18] models perform well when fitting nonlinear and random series. Therefore, the SARIMA and Prophet models are selected to fit the combination series, the RBF and LSTM models are selected to fit the random term series, and a combination of estimation models which is more suitable for the study area is determined by accuracy evaluation standards, such as RMSE and R^2. The experimental steps are shown in Figure 1.

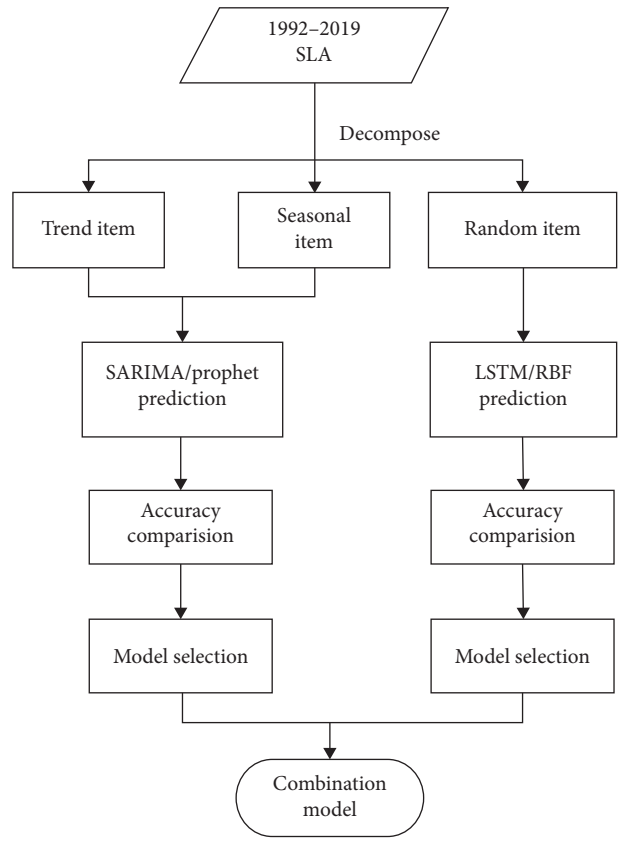


Figure 1: Model selection experiment steps.

2. Material and Methods

2.1. Data and Study Area. The satellite altimetry data used in the experiment are from the GDR data sets of T/P, Jason-1, Jason-2, and Jason-3, which were all released by the French Space Centre (CNES). This paper processes the data in accordance with the steps in [19]. The data obtained after processing are the mean sea level anomaly (SLA) time series data from October 1992 to December 2019 in units of months, as shown in Figure 2.

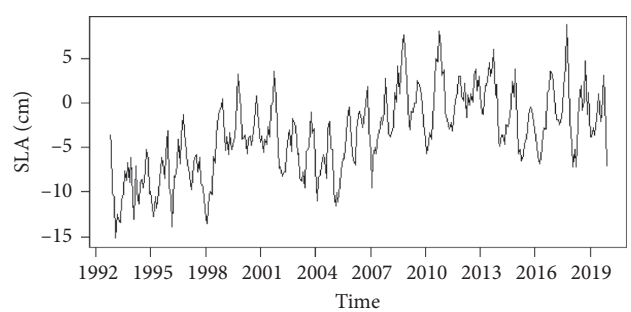


Figure 2: Time series of monthly mean sea level anomalies in the study area.

The research area of this paper is the South China Sea and it is shown in Figure 3. The coordinates are 110°–119°E, 14°–23°N, and the total area is approximately 1.19 million square kilometers. The area is located between the Pacific Ocean and the Indian Ocean, and it includes many important shipping lanes for material transportation. In addition, there are abundant oil and gas resources.

2.2. Seasonal Autoregressive Integrated Moving Average Model. The SARIMA model evolved based on the ARIMA model, which takes into account the seasonal factors of the time series [20–22]. It adopts the method of seasonal difference to estimate parameters and can effectively predict time series with seasonality, trend, and periodicity. The SARIMA model has performed well in industrial and medical research in recent decades [23]. The general form of the SARIMA model is SARIMA (p, d, q) (P, D, Q) $[S]$, where p is the autoregressive order, P is the seasonal autoregressive order, q is the moving average order, Q is the seasonal moving average order, d is the difference order, D is the seasonal difference order, and S is the seasonal period. The SARIMA model is expressed in the following equation:

$$\phi p(B)\Phi_p\left(B^S\right)(1-B)^d\left(1-B^S\right)^D y_t = \theta_q(B)\Theta_Q\left(B^S\right)\mu_t, \tag{1}$$

where y_t is the time series, μ_t is the random term, $\Phi p(B)$ is the nonseason AR(p) part, $\Phi_p(B^S)$ is the season AR(P) part, $(1-B^S)^D$ is the d-order progressive difference, $\theta_q(B)$ is the nonseason MA(q) part, and $\Theta_Q(B^S)$ is the season MA (Q) part.

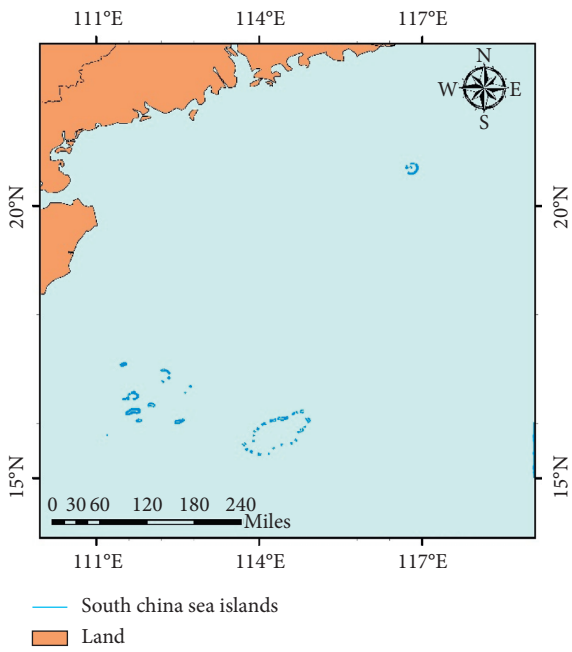


Figure 3: The study area.

2.3. Prophet Model. The Prophet model is a time series curve-fitting tool developed by Taylor and Letham [24]. This model is suitable for fitting time series with strong seasonal effects and has strong robustness against missing values, abnormal values, and trend changes. The basic form of the model is shown in the following equation:

$$P(t) = g(t) + s(t) + h(t) + \varepsilon_t, \tag{2}$$

where $g(t)$ is the trend term used to fit aperiodic changes in the time series, $s(t)$ is the periodic term and it uses a Fourier series to approximate the periodic component, $h(t)$ is the holiday term, and ε_t is the error term. Prophet is robust to missing data and trend changes and usually handles abnormal values well.

2.4. Long Short-Term Memory Model. Hochreiter and Schmidhuber proposed the LSTM model [25–28], which is considered a special recurrent neural network. It can solve the problems of gradient disappearance and gradient explosion and automatically learn the sequence features; therefore, it performs better in the prediction of longer time series and is more suitable for the prediction of sea level changes. The historical update information in LSTM model is controlled by the input gate, forget gate, and output gate as shown in the following equation:

$$\begin{cases} \tilde{c}_t = \tan\mathrm{h}\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right), \\ i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right), \\ f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right), \\ c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\ o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o\right), \\ h_t = o_t \odot \tan\mathrm{h}\left(c_t\right). \end{cases} \tag{3}$$

The steps are given follows:

(1) Calculate the candidate memory unit value $\tilde{c}_t$ at the current moment, where W_{xc} and W_{hc} represent the input data and the unit output weight at the previous moment, respectively

(2) Calculate the input gate value, where x_t represents the current input data, h_{t-1} represents the unit output value at the previous time, and C_{t-1} represents the memory unit value at the previous time

(3) Calculate the forget gate value f and control the influence of historical information on the current state

(4) Calculate the state value C_t of the memory unit at the current moment, where $\odot$ represents the point-by-point product

(5) Calculate the output gate value O_t

(6) Calculate the output of the final LSTM unit

2.5. Radial Basis Function Neural Network Model. The RBF neural network model is a feedforward neural network, and it has strong nonlinear mapping ability and can approximate a nonlinear function with arbitrary precision [29, 30]. Therefore, it is more suitable for the prediction of random terms in time series. The network consists of an input layer, a hidden layer, and an output layer. The radial basis function acts on the high-dimensional mapping between the input layer and the hidden layer, and the linear least square method is used to calculate the weight between the hidden layer and the output layer. The model is generally expressed as shown in the following equation:

$$y_t = \hat{y}_t + e_t = \sum_{i=1}^{M} \varphi_i\left(x_t\right)\omega_t\left(i\right) + e_t, \tag{4}$$

where M is the number of hidden layers, $\varphi_i(x_t)$ is the radial basis function of the ith node, $\omega_t(i)$ is the corresponding regression coefficient, and $i = 1, 2, \ldots, M$.

3. Results and Discussion

The curve in Figure 2 approximately reflects the characteristics of sea level changes in the study area. The peak value corresponds to summer and autumn, and the valley value corresponds to winter and spring. The sea level changes in this area are seasonal and cyclical. To consider both linear and nonlinear features when predicting sea level changes, the monthly mean SLA time series in the study area is decomposed into "season," "trend," and "random" terms based on the principle of addition (Figure 4).

In Figure 4(a), the trend term indicates that as the global climate warms, the sea level of the South China Sea will gradually rise, and the year corresponding to the falling part of the curve will correspond to time periods with strong El Niño phenomena, which fully reflects the trend of a slow rise and occasional decline in sea level observed from 1992–2019. In Figure 2(b), the seasonal term presents the same changes every year, thus reflecting the seasonal characteristics of sea

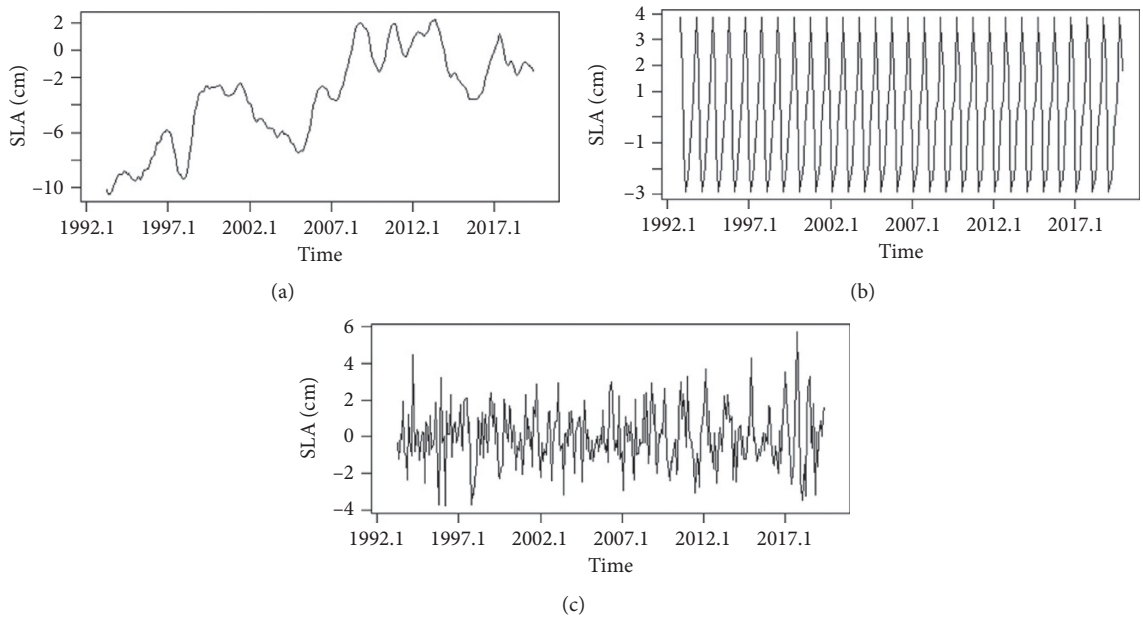


Figure 4: Decomposition results of the monthly mean sea level anomaly time series: (a) trend term, (b) seasonal term, and (c) random term.

level changes. In Figure 2(c), the noise generated by uncertain factors during sea level change is presented. This paper chooses 1992–2017 SLA data as the training set and 2018–2019 SLA data as the test set. Then, according to both the stationary and nonstationary characteristics of sea level change, the above decomposition results are divided into two groups: trend-seasonal series and random term series. The SARIMA and Prophet methods are used to fit the seasonal trend series, and the LSTM and RBF models are used to fit random term series. The root mean square error and the coefficient of determination R^2 were used as the criteria for evaluating the estimation results.

3.1. Estimation of Trend-Seasonal Series

3.1.1. Trend-Seasonal Series Estimated Using the SARIMA Model. The selection and fitting of the SARIMA model mainly include the following steps:

(i) Determine the main structure of SARIMA (p, d, q) (P, D, Q) $[S]$ through experience and autocorrelation and partial autocorrelation function graphs.

(ii) Experiment to obtain other unknown parameters.

(iii) Evaluate the degree of fit through a residual test.

(iv) Perform predictions based on known data [23].

From the above steps, $D = 1$, $d = 0$, $P = 0$, and $Q = 2$. To determine other parameters of the SARIMA model, this paper adopts an experimental method that takes the autoregressive order p and the moving average order q from 0 to 5. The Akaike information criterion (AIC), RMSE, and R^2 were used to evaluate the degree of model fit (Table 1).

The AIC is a measure of the fitting effect of a statistical model. The smaller the AIC value, the better the model fitting effect of the following equation:

Table 1: SARIMA model parameters and fitting results.

Groups	SARIMA model	AIC	RMSE (cm)	R^2
1	SARIMA (2, 0, 2) (0, 1, 2) [12]	−526.94	1.681	0.396
2	SARIMA (3, 0, 1) (0, 1, 2) [12]	−527.96	1.809	0.301
3	SARIMA (4, 0, 1) (0, 1, 2) [12]	−526.27	1.758	0.339
4	SARIMA (5, 0, 3) (0, 1, 2) [12]	−529.05	1.723	0.365

$$\mathrm{AIC} = 2k - 2\ln(L), \tag{5}$$

where k is the number of parameters and L is the likelihood function.

The RMSE can reflect the deviation between the predicted value and the true value and is shown in the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2}, \tag{6}$$

where m is the number of predicted values, y_i is the actual value, and $\hat{y}_i$ is the predictive value.

The coefficient of determination R^2 reflects the degree to which the model fits the test data. If the R^2 value is close to 1, the model fitting effect result will be better as shown in the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2}, \tag{7}$$

where n is the number of predicted values, y_i is the actual value, $\overline{y}_i$ is the mean of true values, and $\widehat{y}_i$ is the predictive value.

The R^2 values of the 4 groups in Table 1 are all small, indicating that the fitting effect is poor and the SARIMA model is not applicable, which may be related to the unobvious trend of sea level rise in 2018 and 2019. Thus, the advantages of the SARIMA model cannot be used.

3.1.2. Trend-Seasonal Series Estimated Using the Prophet Model. The Prophet model transforms a time series into a combination pattern of different time dimensions and then adds the overall trend. The model has a high degree of packaging and few adjustable parameters. According to the sea level change time series characteristics, the estimation frequency is set to "month," the holiday term is empty, and the seasonal mode is set to "Multiplicative" and "Additive." The accuracy of the final estimation results is shown in Table 2.

Obviously, the RMSE of the Additive model is small and the degree of fit is relatively satisfactory; thus, it is more suitable for estimating the time series of the sea level changes in this area.

3.2. Estimation of Random Term Series

3.2.1. Random Term Series Estimated Using the LSTM Model. The LSTM model can learn and remember long-term series information and perform selective forgetting. It is suitable for fitting and estimating the random terms of the time series in our study area. This paper uses 327 months of time series data from 1992 to 2019 as input data, and the number of output data sets is 24. The optimal LSTM estimation model was selected by adjusting the number of hidden layers. According to the RMSE and R^2 results in Table 3, when the hidden layer is 2, the LSTM model fits the best, and the RMSE value is 0.937 cm.

3.2.2. Random Term Series Estimated Using the RBF Model. The RBF neural network model has a strong nonlinear mapping ability and is suitable for fitting of random terms. It has 3 main parameters: the radial basis expansion speed S, the maximum number of neurons MN, and the network parameter DF, which are added each time. Generally, these parameters are determined based on experience. This paper attempted multiple parameter combinations and achieved good results. Although the RMSE of the model in the table is small, R^2 is also too small, indicating that the fitting effect is not good. Therefore, the RBF model is not suitable for the estimation of random terms in this time series (Table 4).

Table 2: Prophet model parameters and fitting results.

Groups	Prophet model	RMSE (cm)	R^2
1	Multiplicative	2.355	−0.185
2	Additive	0.979	0.795

Table 3: LSTM model parameters and fitting results.

Groups	LSTM model	RMSE (cm)	R^2
1	LSTM ($298 \times 1 \times 24$)	1.069	0.691
2	LSTM ($298 \times 2 \times 24$)	0.937	0.763
3	LSTM ($298 \times 3 \times 24$)	1.121	0.661

Table 4: RBF model parameters and fitting results.

Groups	RBF model ($S \times \mathrm{MN} \times \mathrm{DF}$)	RMSE (cm)	R^2
1	RBF ($5 \times 30 \times 4$)	1.145	0.421
2	RBF ($5 \times 50 \times 4$)	1.324	0.217
3	RBF ($5 \times 60 \times 4$)	1.113	0.270

3.3. Prediction by the Combination of Prophet and LSTM Models. According to the results of the model selection experiment, the Prophet model predicts the trend-seasonal series better and the LSTM model predicts the random term series better. This paper chooses the combination of Prophet and LSTM models to predict sea level changes in this area. To explore the influence of the known series length on the results and determine the best prediction duration, this paper sets up training samples and test samples of different lengths to test the prediction effect of the combined model. The experimental results are shown in Table 5. The mean absolute error (MAE) represents the average value of the absolute error between the predicted value and the observed value, which can avoid the problem of mutual cancellation of errors and accurately reflect the actual prediction error. The calculation formula is as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{y}_i - y_i\right|, \tag{8}$$

where n is the number of predicted values, $\widehat{y}_i$ is the predictive value, and y_i is the actual value.

According to the results in Table 5, when the prediction duration is 12–36 months, the accuracy indicators, such as the RMSE, are relatively ideal, indicating that the combined Prophet-LSTM model is suitable for medium- and short-term estimations of the study area for 12–36 months, and the best RMSE of 0.962 cm is obtained. Therefore, this paper uses all known monthly time series of sea level anomalies in the South China Sea from 1992 to 2019 as training samples to estimate sea level changes from 2020 to 2022. The results are shown in Figure 5.

Table 5: Combination model fitting results of training samples of different lengths.

Type of model	Prediction duration (months)	RMSE (cm)	MAE	R^2
Prophet-LSTM	12	1.243	0.744	0.831
Prophet-LSTM	24	1.334	1.083	0.814
Prophet-LSTM	36	0.962	0.718	0.824
Prophet-LSTM	48	2.261	1.849	0.567
Prophet-LSTM	60	4.241	3.967	−0.623

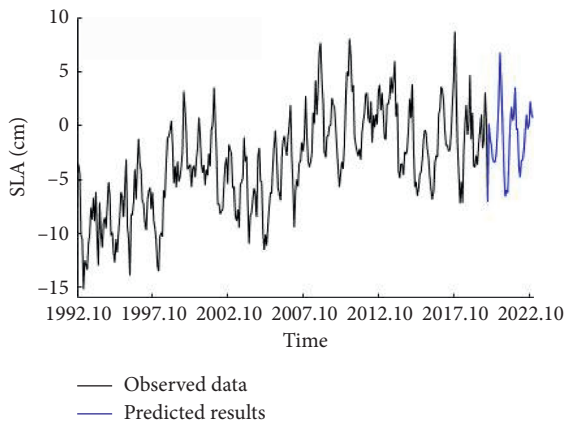


Figure 5: Estimation results of the combined Prophet-LSTM model (2020–2022).

4. Conclusions

The current research aims to evaluate the capability of different models in estimating sea level variability in the South China Sea. Based on satellite altimetry data from 1992 to 2019, this paper compares the estimation effects of the SARIMA, Prophet, LSTM, and RBF models by grouping, and the combination of Prophet and LSTM models was selected. The detailed results are as follows:

(1) A comparison of the estimation accuracy of the SARIMA and Prophet models shows that the Prophet model can better predict the trend-seasonal series

(2) A comparison of the estimation accuracy of the LSTM and RBF models shows that the LSTM model can better predict random term series

(3) The combined model has high accuracy and good performance for 12–36-month short- and medium-term sea level change predictions

The estimation of the time series in this paper simply considers the changing characteristics of the time series. If the temperature, salinity, tides, ocean currents, and climate anomalies of the sea can be included as reference parameters, then the accuracy will be further improved.

Data Availability

The excel 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 are no conflicts of interest in this paper.

Acknowledgments

The authors acknowledge the AVISO website of the French Space Center (CNES) for providing the satellite data. This research has been supported by the Fundamental Research Funds for the Central Universities (17CX02071), NSFC (61571009), and the Key R&D Program of Shandong Province (2018GHY115046).

References

[1] J. Chen, N. Wen, and X. Li, "The status of the resource potential and petroleum exploration of the South China Sea," *Progress in Geophysics*, vol. 8, no. 22, pp. 1285–1294, 2007.

[2] G. A. Milne, W. R. Gehrels, C. W. Hughes, and M. E. Tamisiea, "Identifying the causes of sea-level change," *Nature Geoscience*, vol. 2, no. 7, pp. 471–478, 2009.

[3] State Oceanic Administration, *2019 China Sea Level Bulletin*, State Oceanic Administration, Beijing, China, 2020.

[4] Y. Fu, X. Zhou, D. Zhou, J. Li, and W. Zhang, "Estimation of sea level variability in the South China Sea from satellite altimetry and tide gauge data," *Advances in Space Research*, vol. 7, 2019.

[5] B. M. Ayyub, H. G. Braileanu, and N. Qureshi, "Prediction and impact of sea level rise on properties and infrastructure of Washington, DC," *Risk Analysis*, vol. 32, no. 11, p. 1901, 2012.

[6] Z. Huang, W. Zhang, and H. Wu, "Prediction of sea level rising amplitude in 2030 and defensive countermeasures inthe Pearl River delta," *Science in China*, vol. 30, no. 2, pp. 202–208, 2000.

[7] Z. Qin and Y. Li, "A preliminary study on the law of sea level change in shanghai and its long-term forecast," *Acta Oceanologica Sinca*, vol. 19, no. 6, pp. 1–7, 1997.

[8] P.-H. Yen, C.-D. Jan, Y.-P. Lee, and H.-F. Lee, "Application of kalman filter to short-term tide level prediction," *Journal of Waterway, Port, Coastal, and Ocean Engineering*, vol. 122, no. 5, pp. 226–231, 1996.

[9] C. Chen, J. Zuo, and L. Du, "Long term trends in global sea level under IPCC SRES A2 scenario," *Acta Ocean-Ologica Sinaca*, vol. 34, no. 1, pp. 29–38, 2012, in Chinese.

[10] J. Zhang, J. Zuo, and J. Li, "Sea level variations in the South China Sea during the 21st century under RCP4.5," *Acta Oceanologica Sinica*, vol. 36, no. 11, pp. 21–29, 2014, in Chinese.

[11] J. Zhao, Y. Fan, and N. Ding, "Sea level anomaly forecasting using least square and the radial basis function neural network," *Marine Sciences*, vol. 42, no. 5, pp. 92–97, 2018.

[12] F. V. F. Nobre, A. B. S. Monteiro, P. R. Telles, and G. D. Williamson, "Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology," *Statistics in Medicine*, vol. 20, no. 20, pp. 3051–3069, 2001.
[13] Y. Li, Z. Ma, Z. Pan, and N. Liu, "Prophet model and Gaussian process regression based user traffic prediction in wireless networks," *Science China Information Sciences*, vol. 63, no. 4, 2020.
[14] Z. Xiang, J. Yan, and I. Demir, "A rainfall-runoff model with LSTM-based sequence-to-sequence learning," *Water Resources Research*, vol. 56, no. 1, 2020.
[15] M. A. Ghorbani, H. A. Zadeh, and M. Isazadeh, "A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction," *Environmental Earth Sciences*, vol. 75, no. 6, p. 476, 2016.
[16] N. Ge, L. Sun, and X. Shi, "Research on sales forcast of prophet-LSTM combination model," *Computer Science*, vol. 46, no. 6, pp. 446–450, 2019.
[17] S. Dhakal, Y. Gautam, and A. Bhattarai, "Exploring a deep LSTM neural network to forecast daily PM2.5 concentration using meteorological parameters in Kathmandu Valley, Nepal," *Air Quality, Atmosphere & Health*, vol. 14, no. 1, p. 83, 2020.
[18] M. Abbasi, M. N. Rastgoo, and B. Nakisa, "Monthly and seasonal modeling of municipal waste generation using radial basis function neural network," *Environmental Progress & Sustainable Energy*, vol. 38, no. 3, Article ID e13033, 2018.
[19] J. Wan, Q. Sun, S. Liu, and Y. Li, "Sea-level change over the China sea and its vicinity derived from 25-year T/P series altimeter data," *Journal of the Indian Society of Remote Sensing*, vol. 46, no. 12, pp. 1939–1947, 2018.
[20] M. Valipour, "Long-term runoff study using SARIMA and ARIMA models in the United States," *Meteorological Applications*, vol. 22, no. 3, pp. 592–598, 2015.
[21] W. Hu, S. Tong, K. Mengersen, and D. Connell, "Weather variability and the incidence of cryptosporidiosis: comparison of time series Poisson regression and SARIMA models," *Annals of Epidemiology*, vol. 17, no. 9, pp. 679–688, 2007.
[22] Q. Sun, J. Wan, and S. Liu, "Estimation of sea level variability in the China sea and its vicinity using the SARIMA and LSTM models," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, p. 3317, 2020.
[23] T.-M. Choi, Y. Yu, and K.-F. Au, "A hybrid SARIMA wavelet transform method for sales forecasting," *Decision Support Systems*, vol. 51, no. 1, pp. 130–140, 2011.
[24] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.
[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[26] K.-C. Miao, T.-T. Han, Y.-Q. Yao et al., "Application of LSTM for short term fog forecasting based on meteorological elements," *Neurocomputing*, vol. 408, no. 30, pp. 285–291, 2020.
[27] Z. Karevan and J. A. K. Suykens, "Transductive LSTM for time-series prediction: an application to weather forecasting," *Neural Networks*, vol. 125, p. 1, 2020.
[28] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons & Fractals*, vol. 135, Article ID 109864, 2020.
[29] G. Wang, L. Liu, and A. Xu, "The application of radial basis function neural network in the GPS satellite clock bias prediction," *Acta Geodaetica et Cartographica Sinica*, vol. 43, no. 8, pp. 803–807, 2014.
[30] P. Cai, Y. Wang, and G. Lu, "Tunable and transferable RBF model for short-term traffic forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4134–4144, 2019.

Hindawi
Scientific Programming
Volume 2021, Article ID 5553566, 10 pages
https://doi.org/10.1155/2021/5553566

Research Article

Correlation between Triadic Closure and Homophily Formed over Location-Based Social Networks

Nauman Ali Khan,[1] **Wuyang Zhou**,[1] **Mudassar Ali Khan**,[2] **Ahmad Almogren**,[3] **and Ikram Ud Din**[2]

[1]*Key Laboratory of Wireless-Optical Communication, University of Science and Technology of China, Hefei 230027, China*
[2]*Department of Information Technology, The University of Haripur, Haripur 22620, Pakistan*
[3]*Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11633, Saudi Arabia*

Correspondence should be addressed to Wuyang Zhou; wyzhou@ustc.edu.cn

Received 15 January 2021; Revised 25 January 2021; Accepted 29 January 2021; Published 15 February 2021

Academic Editor: Habib Ullah Khan

Social Internet of Things (SIoT) is a variation of social networks that adopt the property of peer-to-peer networks, in which connections between the things and social actors are automatically established. SIoT is a part of various organizations that inherit the social interaction, and these organizations include industries, institutions, and other establishments. Triadic closure and homophily are the most commonly used measures to investigate social networks' formation and nature, where both measures are used exclusively or with statistical models. The triadic closure patterns are mapped for actors' communication behavior over a location-based social network, affecting the homophily. In this study, we investigate triads emergence in homophilic social networks. This evaluation is based on the empirical review of triads within social networks (SNs) formed on Big Data. We utilized a large location-based dataset for an in-depth analysis, the Chinese telecommunication-based anonymized call detail records (CDRs). Two other openly available datasets, Brightkite and Gowalla, were also studied. We identified and proposed three social triad classes in a homophilic network to feature the correlation between social triads and homophily. The study opened a promising research direction that relates the variation of homophily based on closure triads nature. The homophilic triads are further categorized into transitive and intransitive groups. As our concluding research objective, we examined the relative triadic throughput within a location-based social network for the given datasets. The research study attains significant results highlighting the positive connection between homophily and a specific social triad class.

1. Introduction

Homophily identifies the groups of individuals who are socially connected based on shared interests or behaviors. In the past decades, numerous sociologists premeditated clusters of people based on various sociocommunity parameters, including gender, religion, race, place of living, and work. These parameters were used to infer various relations like close friends, coworkers, life-partner, and other social associations. Based on these social parameters and their similarities, few broad applications include user mobility, influencing, and segregation. With the rapid growth of communication networks, quantifying accurate homophily analysis is one of the most critical social network analysis (SNA) problems that is further subcategorized as triadic closure analysis and home location detection analysis. One of the fundamental challenges in detecting homophily is when a person with versatile personality features tends to change his behavioral pattern dynamically. Traditional techniques commonly use the clustering method to exploit and predict the reasons for a homophilic nature. For the scenario mentioned above, these techniques lack accuracy and precision when a social network accommodates diverse multiprofessional users having a dense structure.

Regarding detection applications of triadic closures and homophily, scientists also contributed to various application areas besides automation and network traffic management. These include refinement of recommendation systems, fake user identification, analysis of micro blogging, detection of natural disasters using real-time Twitter Big Data, business decision making, and healthcare systems [1–5]. Companies and businesses increase revenues and improve goodwill by maintaining their micro blogging systems. Machine learning algorithms extract meaningful information and help fetch the most related information, which helps in decision making [6]. In the literature, a great effort was made to gather the information related to a particular category of people on Facebook [7–9]. Aral and Walker identified the group of people on Facebook which were easier to influence. Their principal findings involve that young people are easier to influence in contrast to older generation people. Likewise, males have a more influential nature as compared to females. Similarly, other influential patterns were recognized in cross-gender comparisons. However, married people were categorized as the category which can get influenced [10].

A triadic closure in social networks can be interpreted as a communication group of precisely three individuals. Trio/triangle/triad is considered to be the necessary foundation of a social network. In literature, some modern research studies political campaigns, religious activities, organizational professionalism, web mining, and many more social networks based on such three-people subgraph [11]. Listing and counting of triads in a social network are considered triad census using the subgraph method of graph theory [7, 12, 13]. The clustering coefficient, a robust graph theory method, highlights the degree of nodes likely to be part of a cluster. A higher degree of the coefficient indicates a higher ratio of triads in a social network. One research also highlighted the positive correlation between the triads and community structures. Research findings reflect that community structures were coherent where the number of triads is remarkably high [14–17].

Social triad analysis in a multicluster environment helps to overcome the mentioned problem. Origins of dyads and triads in the social network encourage exploiting the homophilic nature further, specifically when the triad nodes belong to two different groups [18]. Generically, a triad is a group of three socially connected individuals in a social network, also referred to as the smallest group of that social network.

Triadic closure and homophily are two separate social network analysis evaluation measures. Applications of triadic closure and homophily involve friend recommendation systems, online social blogging services, community influence systems, and structural and informational construction systems. It further enhances learning systems, improves competition, and also increases work performance [19]. Previously, these evaluation measures were used individually to assist the above-mentioned issues and areas. In this research, we found a strong association between these measures and proposed a technique which uses these measures together.

In our research, we also explored the patterns of homophily in the multidomain social network. We took a sample of the call detail records (CDRs) dataset and constructed a social network graph. In large-scale dataset of CDRs, each record is represented in the following format.

Caller ID	Call type	Callee ID	Time

Duration	LAC ID	CELL ID

We constructed a social network using telecommunication-based anonymized call detail records and two openly available location-based social network datasets, similar to the work of Brightkite and Gowalla represented in [20]. Distinct caller ID is considered a distinct social network's user, and communication between two callers is considered a social tie. For every user, one home location is selected from various locations depending upon the maximum number of incoming and outgoing calls. Furthermore, we have identified the users' triads, in which all users belong to a shared home location. Figure 1 illustrates social triads' formation by variant home location of individuals in a social network. According to Figure 1, a standard social network is illustrated; each node is represented with v_v while each home location is represented with $\mathrm{HL}_{\mathrm{hl}}$. There is a scenario in which several triad nodes belong to a shared home location, such as v_2, v_3, and v_4 triad belonging to HL_2. Our research identifies the origins of triadic closure in a homophilic network and proposes a classification model that creates subclasses into three groups.

In this study, our contribution relates to the proper classification of the triads, which is discussed as follows:

(i) We first studied the user mobility patterns and their diversity by observing the entropy. We developed a social network graph of users and identified home location using home detection algorithm from the datasets.

(ii) Based on home locations, we grouped users and critically observe their interconnections. Furthermore, we identified the homophilic patterns formed inside the social network.

(iii) We investigated the origins of social triads in detail and examined the formation of triads. Based on the analysis, we categorized social triads and compared their behaviors within the homophilic social network. Interestingly, we found positive correlations between the homophily coefficient and a subset of social triads discussed in the relevant section.

(iv) In the later part of the research, we organized homophilic triads into transitive and intransitive groups, and we examined the effect of categorized triads with the network's throughput.

The rest of the article is organized as follows: Section 2 describes the literature review. Section 3 presents the problem formulation and evaluation measures. Section 4 introduces the triadic closure in the homophilic environment and its effect on homophily. Section 5 describes the datasets and observations. Section 6 explains the results and their discussion. Section 7 concludes with future recommendations.

2. Literature Review

A social network is generally composed of three artifacts, i.e., user description, social connection direction, and communication contents exchanged over the social network [21]. The user-based artifact study explores the user's behavior in different scenarios and environments [22]. Individual personal networks are the social network subgraphs that identify all the communication behavior of a single entity [23]. Individual personal networks have a transitive tendency, i.e., a friend of a friend is also a friend, as discussed by [24]. Transitivity is the propensity that two people, who are not direct friends to each other but have a familiar mutual friend, may also become friends over time [16, 25]. Researchers analyze the reason for triads' formation, why a dyad converts to a triad with time, and how, in a three-person small network, all the users want to reduce the hesitation discrepancies [18, 26]. In an unbalanced triad social network, where two different people like one person, but these two people do not like each other, this creates emotional tension between them, forcing the relationship to be complete and consistent, or discourages the triad formation [27]. According to a comprehensive survey, it was consistently observed that transitivity exists in about 70% to 80% of various small groups [28–30]. In another research study, the effect of gender was highlighted, and it was revealed that the formation of triads in boys is more common than in girls [31]. One other study compared homogeneous behavior of users with heterogeneous environment actors, and it was concluded that heterogeneous actors are less transitive concerning religion, race, and education than homogeneous actors [32, 33]. A study highlights the baseline of triads forming; trust plays a vital role in making the relationships more robust and balanced [34]. While establishing and building new ties, people may have hidden or apparent interests such as knowledge sharing and a social relationship like friendship, educational purpose, and scientific collaboration [35]. Moreover, an existing study shows the positive correlation between authorship sharing and research-based relationship building that spreads over time [36].

Online location-based social networking applications enable the users to build social ties based on location [37–39]. In addition to social connection details, a social network formed over a location-based application may have extra attached information such as location ID [35]. Similar to location-based social networks, CDRs (call detail records) datasets are the log files of users reordered over time. These logs include the details of user communications and the attached information of location ID. As per our literature exploration, many researchers used this location ID to draw the homophily of the social networks [37, 40, 41]. A study examined existing location-based human mobility trend evaluation techniques and categorized them into mainly three classes, i.e., user, place, and trajectory-based modeling [42–44].

Homophily refers to a social grouping concept where people with common interests tend to morph into a single group [45]. In literature, homophily is broadly based on two approaches, i.e., *induced* and *choice* homophily [46]. The combined effect of social triads is observed with homophily, and it is determined that choice homophily plays a vital role in building observed homophily [47]. Research findings also illustrated that making triads within homophilic regions is statically higher [47].

To summarize, triad creation and critical exploration in a social network help to understand social relationships that further assist in many applied areas already discussed. In literature, many research contributions have been conducted to exploit social triads for various aspects, though there is a need to further understand how location information can affect social triads and homophily.

3. Problem Formulation and Evaluation Measures

The formulation of the problem is stated as follows. Let $G = (V, E)$ be a graph representing a static social network of users and their communication links, where $V = \{v_1, v_2, \ldots, v_{|V|}\}$ is a set of actors/users in a social network and $E \subset V \times V$ is a set of social links between users. $e_{ij} \in E$ shows the existence of a communication link between v_i and v_j users. Let $T = \{\Delta = (v_i, v_j, v_k) | v_i, v_j, v_k \in V\}$ be a set of triads.

Definition 1. (CT: closed triads). Let $\mathrm{CT} = \{\Delta = (v_i, v_j, v_k) | \Delta \in T \wedge e_{ij}, e_{ik}, e_{jk} \in E\}$ be the set of closed triads.

Definition 2. (OT: open triads). Let $\mathrm{OT} = \{\Delta = (v_i, v_j, v_k) | \Delta \in T \wedge e_{ij}, e_{ik} \in E \wedge e_{jk} \notin E\}$ be the set of open triads.

Definition 3. (HL: user home location). Let $L = \{l_1, l_2, \ldots, l_{|L|}\}$ is a set of locations, where l_n denotes a distinct location. Let $\mathrm{HL} = \{h_1, h_2, \ldots, h_{|V|}\}$ be a set of user home locations, where h_n denotes a home location for user v_n. $h_n = \mathrm{Home\,Location}(v_n, L)$

According to the location-based social network, every user forms a social connection at a specific location. For v_n, the function $\mathrm{Home\,Location}(v_n, L)$ identifies one location from L as home location h_n based on home location algorithm stated in [48].

Definition 4. ($\Theta A, \Theta B, \Theta C$: types of triads).

For $\Delta = (v_i, v_j, v_k)$, let

$$\Theta A = \{\Delta | \Delta \in \mathrm{CT} \wedge h_i = h_j = h_k\},$$
$$\Theta B = \{\Delta | \Delta \in \mathrm{CT} \wedge h_i = h_j \wedge h_i \neq h_k\},$$
$$\Theta C = \{\Delta | \Delta \in \mathrm{CT} \wedge h_i \neq h_j \neq h_k\}.$$

Definition 5. (ψ: homophily coefficient).

$$\psi = \{\psi_{xy} | \psi_{xy} = \mathrm{Homophily}(v(h_x), v(h_y))\}, \quad (1)$$

where $\begin{cases} h_x, h_y \in \mathrm{HL} \\ h_x \neq h_y \end{cases}$.

Let $\psi =$ be a set of homophily, where ψ_{xy} denotes homophily of graph for two sets of vertices. $v(h_n)$ denotes a set of all the vertices belonging to h_n home location. Function

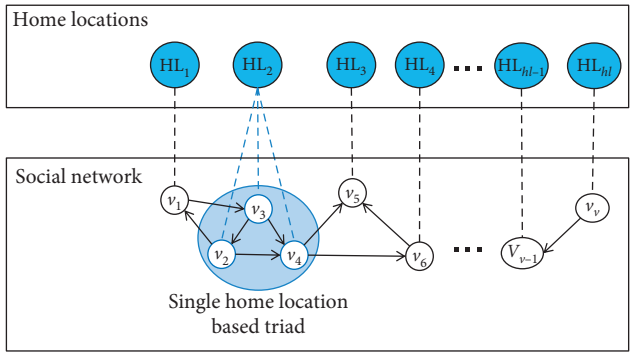


Figure 1: Formation of triad closure based on home locations.

Homophily$(v(h_x), v(h_y))$ takes two sets of vertices, i.e., $v(h_x)$ and $v(h_y)$, and initially counts the cross-home location edges $e_{v(h_x),v(h_y)} \vee e_{v(h_y),v(h_x)}$ as p and non-cross-home location edges $e_{v(h_x),v(h_x)} \vee e_{v(h_y),v(h_y)}$ as q. Then, it finds the expected cross-home location edges as $\xi = ((p+q)/2)$. After that, the homophily coefficient is calculated using the following equation [49].

$$\psi_{xy} = 1 - \frac{\xi}{p}. \tag{2}$$

Correlation Coefficient. Correlation coefficient among $\Theta A, \Theta B, \Theta C$ types of triads and homophily is defined in

$$r(\psi, \Theta) = \frac{\sum(\psi - \overline{\psi})(\Theta - \overline{\Theta})}{\sqrt{\sum(\psi - \overline{\psi})^2}\sqrt{(\Theta - \overline{\Theta})^2}}. \tag{3}$$

4. Social Triads in Location-Based Social Networks

A social network is the communication graph among many users. Datasets such as telecom call logs or location-based social network data have the details of the user's interaction and a hint of location information. Each record of the datasets represents a time-stamped location-based social link between two users in communication.

4.1. Triadic Closure Property in Homophilic Environment. Triadic closure refers to the communication of three nodes. Every closed triad can be either transitive or intransitive, depending upon the type of communication occurring [50]. Each node of the triads belongs to one specific location, treated as its home location. The location of home for each user or node is identified using the *home detection algorithm* [48]. While critically examining the formation of the closed triad, we identified and hence proposed three cases of triads, listed as follows:

(1) All users of the triad belong to the same home locations

(2) Any two triad users belong to one home location, and the remaining user belongs to any other home location

(3) All users of the triad belong to three different home locations

Figure 2 states an example of a social network based on a CDRs subdataset. In this figure, each hexagon shows a region of the telecommunication signal cell. A social network over the cellular signal region represents a communication graph, and each cell is considered as a home location of inside nodes. The green-colored hexagon is taken as a reference cellular signal region in the stated example, and other red-colored hexagons are considered out location cellular signal region. As described before, these three triad classes are also illustrated in Figure 2.

We named the three possible triads as Class A, Class B, and Class C for differentiation and further exploration. Our research first investigates each class, classifies it into transitive triads or intransitive triads, and then examines all possible combinations of social triads in a directed graph. Figure 3 illustrates a detail overview of all possible triads and defines them into three classes. Code underneath each triad represents the category, and the naming convention of the social triad is explained in [51]. However, we improvise the category and naming convention by adding an alphabet at the start of the code as a class name and by adding an extra digit as its variant. In the code $B210A1$, B is the class name, $210A$ is the existing naming convention, and 1 is the variation number.

4.2. Accumulative Homophily in Triadic Closure. Call detail records (CDRs) and online location-based social networks have extra associated information, i.e., location ID. In our research, we incorporated the location ID into identified homophily in a network. We utilize the existing home detection algorithm to identify the home location for each user [48]. In location-based social networks, by home location, we mean the most visited and stayed at place. The algorithm identifies one location out of all visited places as a home location. Further, we measure the correlation between the three classes of triads and homophily.

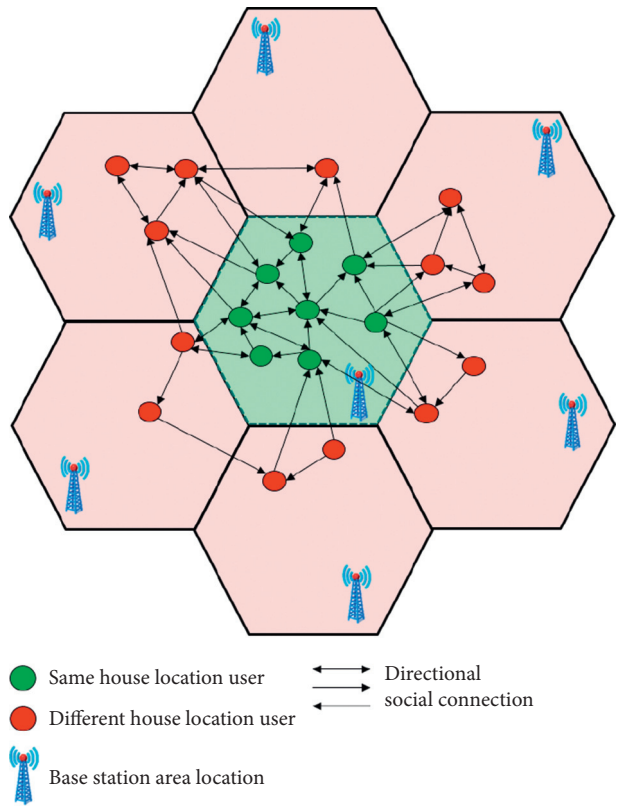


Figure 2: Illustration of social triads in location-based networks.

A triad is a group of three nodes, in which each node belongs to specific home locations. However, homophily is calculated based on only two groups. Initially, we calculate homophily using (2), and then we averaged them. For three home locations, e.g., h_x, h_y, and h_z, accumulative homophily is measured, as stated in

$$\mathrm{Acc}_{(\psi_{xyz})} = \frac{\psi_{xy} + \psi_{xz} + \psi_{yz}}{3}. \quad (4)$$

5. Datasets Characteristics and Observations

5.1. Data Description. In support of research, we incorporated one large call detail record (CDR) and two online location-based datasets, i.e., Gowalla and Brightkite [20]. The CDR dataset used in this study is provided by a Chinese mobile telecommunication company. The dataset contains 702,000 subscribers along with user demographic information. The data is logged over the period of one year, which has more than half a billion social ties.

Brightkite and Gowalla are openly available location-based social network datasets [20, 52]. Both datasets are gathered using the online social networking website. Websites maintain user check-in data by fetching mobile GPS location data. These services create an environment that enables people to build a social connection with nearby people. The Brightkite dataset contains 58,228 nodes and 214,078 edges, and Gowalla contains 196,591 nodes and 950,327 edges. In the data cleaning phase, we removed missing or wrong data types and empty rows. In the CDRs dataset, each record is represented as in the following column format.

Caller ID	Call type	Callee ID	Time

Duration	Call type	LAC ID	CELL ID

5.2. Observations. Call duration is one of the key attributes of the calling dataset. While mining the CDR dataset and investigating the social networks, we observed some interesting call duration facts. Figure 4 shows the relation of call duration and number of calls. We found two big spikes in the number of calls according to the call duration. We have found that the maximum number of calls has call duration in the range of either 10 to 30 seconds or 1 min to 2 min. This observation infers that people mostly prefer to have a short duration communication to convey their message. One research shows that direct calls are a kind of strong communication and are considered the baseline for the strong ties [53].

CDR logs contain another important item, i.e., the location ID attribute, which identifies the area from which the call was made. Initially, we applied the home location algorithm and inferred the home location based on the call logs, and then we segregated all users according to the location ID. Figure 5 shows the distribution of users based on location ID.

We carefully monitored the communication behavior of the people within each location.

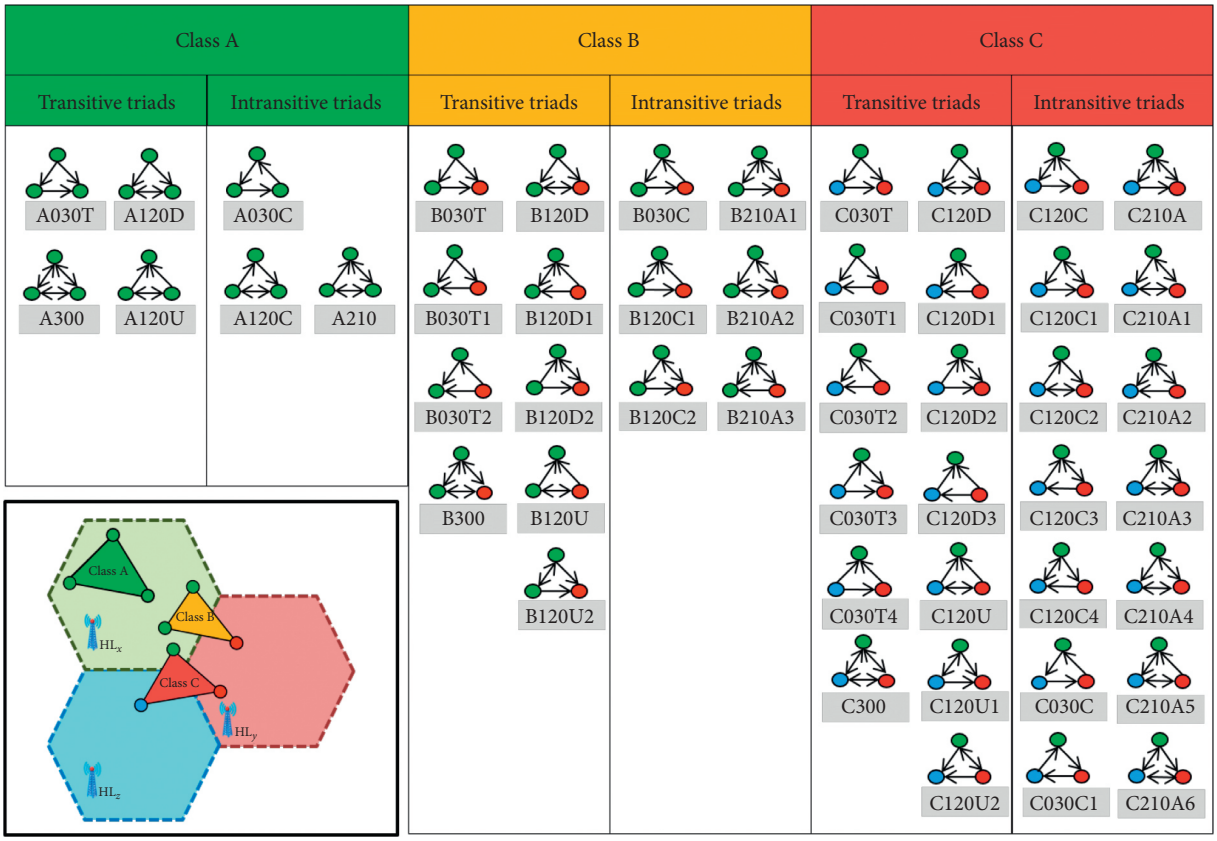


Figure 3: A fine-grained classification of social triads in location-based homophilic networks with all variations.

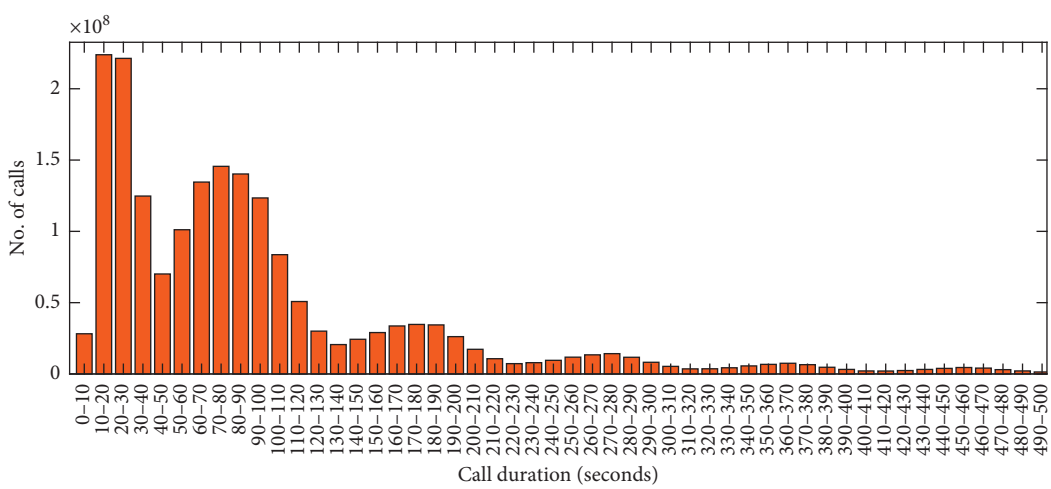


Figure 4: Segregation of number of calls in comparison to call duration.

During fact extraction, we found a high ratio of calls between people at the same location in comparison to that of different localities. Figure 6 is a preview of communications taking place for different locations or within the same location. Location-based cross-communication infers homophily which is based on location, which is the key motivation aspect for this study. Figure 6 shows that the interaction taking place between people from the same location is more than that between people from different locations, which further indicates the existence of location-based homophily. This

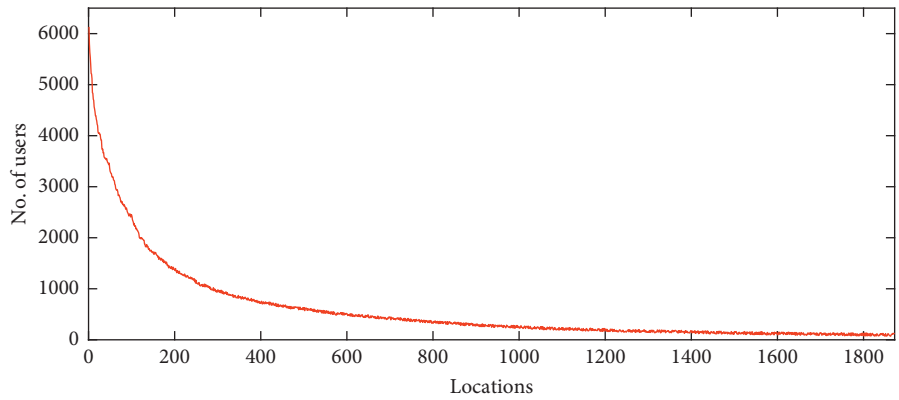


Figure 5: Location-based user density compiled through home location detection algorithm.

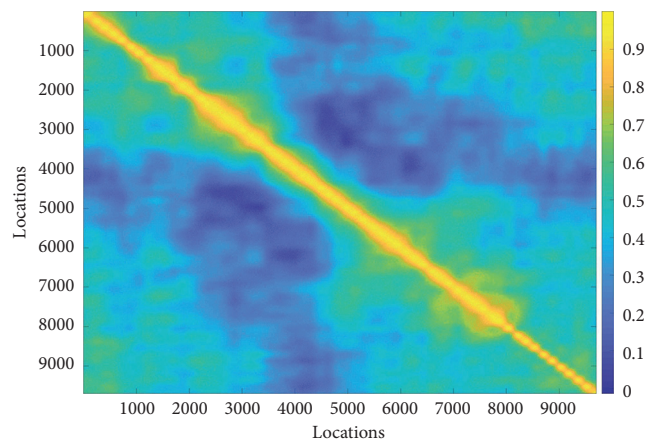


Figure 6: Visualization of homophily for intralocation-based user communication.

further adheres to the fact that there is a strong connection between location-based homophily and triadic social closure.

6. Results and Discussion

Our research evaluation results classify the empirical social triads into three groups based on the strong correlation between homophilic networks and social triads. We found a positive correlation between the homophily and a specific class of triads. Our findings indicate that people having the same home location are more likely to form a triad.

In this study, we incorporated two location-based large datasets and one close source CDR dataset. Figure 7 illustrates nine correlation comparisons, three for CDR, Brightkite, and Gowalla datasets. Results show the correlation between homophily and classes of triads. The y-axis shows the percentage of homophily, and the x-axis refers to the number of triads in percentage. Results shown in Figure 7 reveal that the accumulative homophily between the groups has a positive correlation with Class A triads. Simultaneously, Class A refers to a group of users triad having a common home location.

We initially measured the number of triads for all the three classes of the datasets and observed that the minimum quantity for a triads can be individually calculated from each category. A sum of 2,200 triads was found for Class C. For the understanding of results and normalization, we randomly selected 2,000 triads for the three classes. Results show that higher homophily corresponds to a higher number of social triads from Class A. However, the impact of homophily related to Class B and Class C is comparatively unspecific. A consistency of positive correlation was observed in all the three datasets between homophily percentage and triads of Class A.

The regression coefficient r of the correlation was examined using (3). From the comparisons between all datasets and Class A, we found the highest value for the regression coefficient of r. Besides high regression coefficient values r and consistency, our research also discovers all results' closeness, especially for the CDR dataset.

In the analysis, we found the maximum observations of homophily within the range of 25% to 80%, and the cross-relation between Class A and homophily highlights the maximum observation of triads in the range of 5% to 70%. All the three datasets produce symmetric and positive regression trend results. The regression coefficient $r = 0.61$, $r = 0.65$, and $r = 0.55$ is measured for CDR, Brightkite, and Gowalla dataset, respectively. The r value denotes the

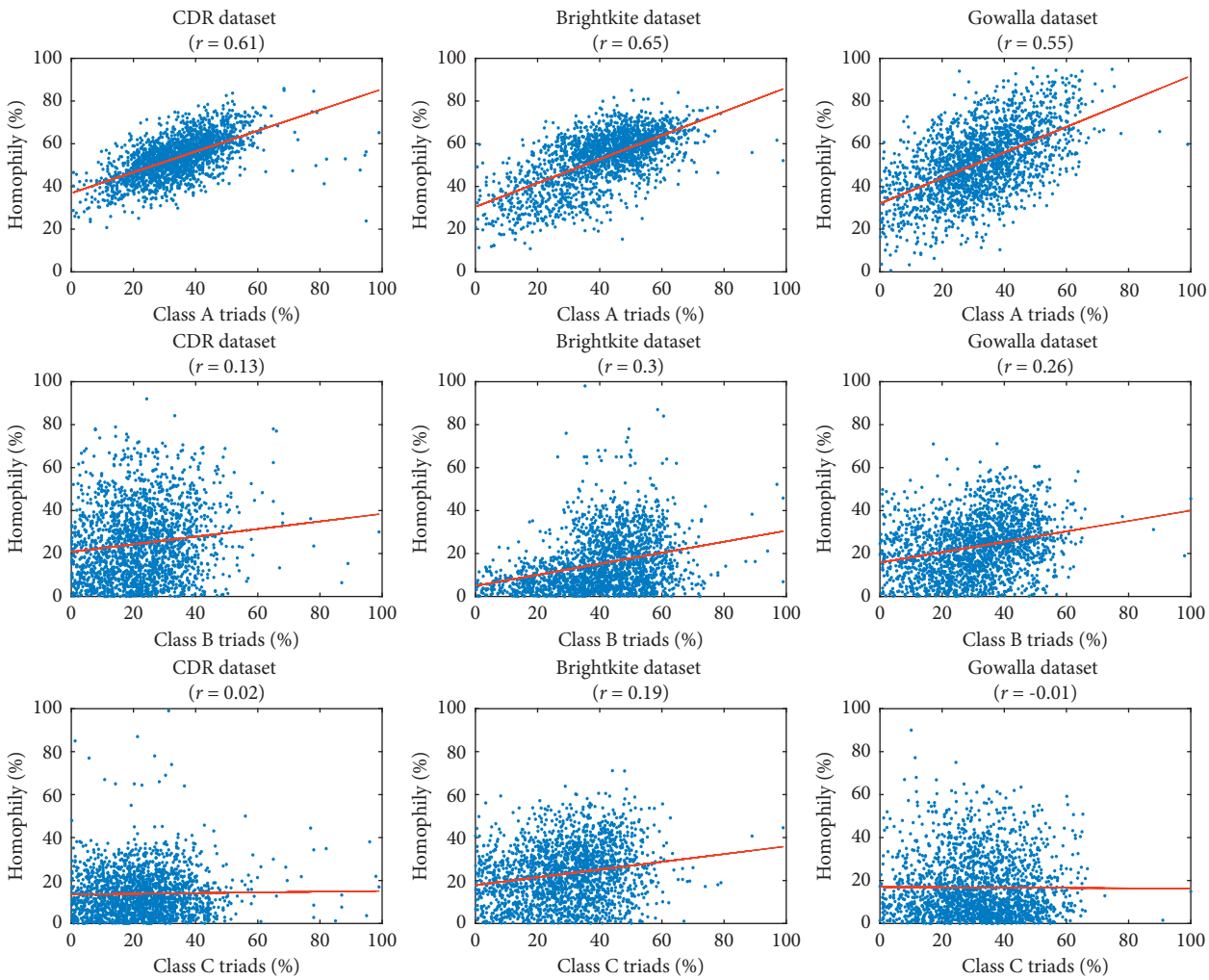


Figure 7: Correlation between the three classes of triadic closure and homophily using location-based datasets.

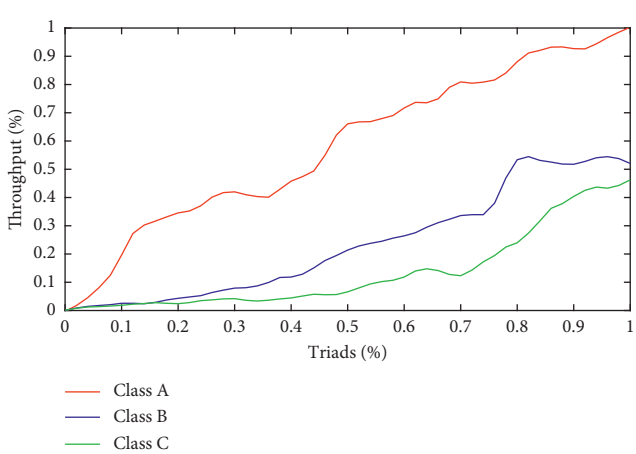


Figure 8: The relative throughput of triadic closure for the three classes.

existence of cause and effect relationship between the triadic closure and homophily, especially between Class A and homophily.

In the second phase of evaluation, we measured the accumulative throughput for Class A, B, and C in all the datasets. Figure 8 shows the overall throughput for the three datasets; the y-axis shows throughput percentage and the x-axis shows the number of triads in percentage. The throughput (T) is measured using (5). We used a relative throughput measure to cross-relate the results. The lowest and the highest values of the throughput were taken as reference values, and then accordingly the rest of the graph was plotted.

$$T = \frac{\text{calls made by triads}}{\text{total calls made by triads}} \times 100. \quad (5)$$

In this study, we observed that Class A triads consume the maximum amount of bandwidth. We encountered a significant rise in the throughput for Class A after 40%, which shows that people with a higher number of triads of the same home location also exchange a higher number of calls as shown in Figure 8. However, we came across the least throughput for Class B and Class C within the range of 1% to 50%. The lower values of throughput indicate the least communication among the triad users.

The throughput of Class C is comparatively less than that of Class B because all the three users of Class C were in different home locations. However, Class B, having any two users from a common home location, explains the slight increase in its throughput. This study highlights the higher throughput of Class A as compared to the rest of the classes. The results indicate that triads formed between people from the same home location have more communication rates than triads formed at different home locations.

7. Conclusion and Future Work

Triadic closure and homophily coefficient are the two mutually exclusive merits required to understand the behavior of the social network. In this study, we found the cause and effect relationship between the homophily and triad closure for the social networks formed based on the location. We have closely observed social triads' formation in a homophilic social network and found interesting relationships between them. Our study used Chinese telecommunication-based anonymized call detail records (CDRs) and two openly available location-based social network datasets, Brightkite and Gowalla. This research identifies three sets of social triad classes in a homophilic network and expresses the correlation between social triads and homophily. Examination findings opened a novel direction of measuring homophily based on multiple types of social triads. Based on the communication directions, we further organized homophilic triads into a transitive and intransitive group. In the last part of the research, we also examined the effect of a specific triadic class on a network's throughput. We will investigate the reasons for the formation of transitive and intransitive classes in homophilic networks in the future.

Data Availability

The data used can be found at http://snap.stanford.edu/data/index.html#locnet.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this work.

Acknowledgments

This work was supported by King Saud University, Saudi Arabia, through research supporting project number RSP-2021/184. Nauman Ali Khan acknowledges the support of the Chinese Government and Chinese Scholarship Council (CSC) for his Ph.D. studies at the University of Science and Technology, China. This research work was partially supported by Key Program of National Natural Science Foundation of China (Grant number 61631018).

References

[1] X. Luo, C. Jiang, W. Wang, Y. Xu, J.-H. Wang, and W. Zhao, "User behavior prediction in social networks using weighted extreme learning machine with distribution optimization," *Future Generation Computer Systems*, vol. 93, pp. 1023–1035, 2019.

[2] S. Nazir, S. Khan, H. U. Khan et al., "A comprehensive analysis of healthcare big data management, analytics and scientific programming," *IEEE Access*, vol. 8, pp. 95714–95733, 2020.

[3] F. Masood, A. Almogren, A. Abbas et al., "Spammer detection and fake user identification on social networks," *IEEE Access*, vol. 7, pp. 68140–68152, 2019.

[4] B. Amina and T. Azim, "Scancpeclens: a framework for automatic lexicon generation and sentiment analysis of micro blogging data on China pakistan economic corridor," *IEEE Access*, vol. 7, pp. 133876–133887, 2019.

[5] V. Gupta and R. Hewett, "Real-time tweet analytics using hybrid hashtags on twitter big data streams," *Information*, vol. 11, no. 7, p. 341, 2020.

[6] S. Pouyanfar, S. Sadiq, Y. Yan et al., "A survey on deep learning: algorithms, techniques, and applications," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–36, 2018.

[7] Q. Gong, Y. Chen, X. He et al., "DeepScan: exploiting deep learning for malicious account detection in location-based social networks," *IEEE Communications Magazine*, vol. 56, no. 11, pp. 21–27, 2018.

[8] A. Zrnec and D. Lavbič, "The role of social connections in plagiarism detection," in *Proceedings of the International Workshop on Learning Technology for Education in Cloud*, pp. 54–63, Springer, Maribor, Slovenia, August 2015.

[9] S. Ali, N. Islam, A. Rauf, I. Din, M. Guizani, and J. Rodrigues, "Privacy and security issues in online social networks," *Future Internet*, vol. 10, no. 12, p. 114, 2018.

[10] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Science*, vol. 337, no. 6092, pp. 337–341, 2012.

[11] L. Becchetti, P. Boldi, C. Castillo, and A. Gionis, "Efficient algorithms for large-scale local triangle counting," *ACM Transactions on Knowledge Discovery from Data*, vol. 4, no. 3, pp. 1–28, 2010.

[12] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, vol. 8, Cambridge University Press, Cambridge, UK, 1994.

[13] Z. Ali, M. A. Shah, A. Almogren, I. Ud Din, C. Maple, and H. A. Khattak, "Named data networking for efficient iot-based disaster management in a smart campus," *Sustainability*, vol. 12, no. 8, p. 3088, 2020.

[14] M. E. Newman, "Properties of highly clustered networks," *Physical Review E*, vol. 68, no. 2, Article ID 026121, 2003.

[15] K. Warren, B. Campbell, S. Cranmer et al., "Building the community: endogenous network formation, homophily and prosocial sorting among therapeutic community residents," *Drug and Alcohol Dependence*, vol. 207, Article ID 107773, 2020.

[16] N. Muyinda, J. M. Baetens, B. De Baets, and S. Rao, "Using intransitive triads to determine final species richness of competition networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 540, Article ID 123249, 2020.

[17] D. V. Foster, J. G. Foster, P. Grassberger, and M. Paczuski, "Clustering drives assortativity and community structure in ensembles of networks," *Physical Review E*, vol. 84, no. 6, Article ID 066117, 2011.

[18] H. Huang, Y. Dong, J. Tang, H. Yang, N. V. Chawla, and X. Fu, "Will triadic closure strengthen ties in social networks?" *ACM Transactions on Knowledge Discovery from Data*, vol. 12, no. 3, pp. 1–25, 2018.

[19] S. Khan, S. Nazir, and H. Khan, "Smart object detection and home appliances control system in smart cities," *Computers, Materials and Continua*, vol. 67, pp. 895–915, 01 2021.

[20] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1082–1090, San Diego, CA, USA, August 2011.

[21] J. Tang, Y. Chang, and H. Liu, "Mining social media with social theories," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 2, pp. 20–29, 2014.

[22] F. Amato, A. Castiglione, A. De Santo et al., "Recognizing human behaviours in online social networks," *Computers & Security*, vol. 74, pp. 355–370, 2018.

[23] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Personality predictions based on user behavior on the facebook social media platform," *IEEE Access*, vol. 6, pp. 61959–961969, 2018.

[24] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[25] S. Peng, Y. Zhou, L. Cao, S. Yu, J. Niu, and W. Jia, "Influence analysis in social networks: a survey," *Journal of Network and Computer Applications*, vol. 106, pp. 17–32, 2018.

[26] I. U. Din, A. Almogren, M. Guizani, and M. Zuair, "A decade of internet of things: analysis in the light of healthcare applications," *IEEE Access*, vol. 7, pp. 89967–89979, 2019.

[27] D. Krackhardt and M. Kilduff, "Whether close or far: social distance effects on perceived balance in friendship networks," *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 770–782, 1999.

[28] C. McMillan and D. Felmlee, "Beyond dyads and triads: a comparison of tetrads in twenty social networks," *Social Psychology Quarterly*, vol. 83, no. 4, pp. 383–404, Article ID 0190272520944151, 2020.

[29] D. T. Robinson and J. W. Balkwell, "Density, transitivity, and diffuse status in task-oriented groups," *Social Psychology Quarterly*, vol. 58, no. 4, pp. 241–254, 1995.

[30] D. Kretschmer, L. Leszczensky, and S. Pink, "Selection and influence processes in academic achievement-more pronounced for girls?" *Social Networks*, vol. 52, pp. 251–260, 2018.

[31] G. Kossinets and D. J. Watts, "Empirical analysis of an evolving social network," *Science*, vol. 311, no. 5757, pp. 88–90, 2006.

[32] H. Schäfer, "Relationality and heterogeneity: transitive methodology in practice theory and actor-network theory," in *Methodological Reflections on Practice Oriented Theories*, pp. 35–46, Springer, Berlin, Germany, 2017.

[33] X. Han, S. Cao, Z. Shen et al., "Emergence of communities and diversity in social networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 11, pp. 2887–2891, 2017.

[34] K. D. Doekhie, M. M. H. Strating, M. Buljac-Samardzic, and J. Paauwe, "Trust in older persons: a quantitative analysis of alignment in triads of older persons, informal carers and home care nurses," *Health & Social Care in the Community*, vol. 27, no. 6, pp. 1490–1506, 2019.

[35] N. A. Khan, S. Zhang, W. Zhou, A. Almogren, I. Ud Din, and M. Asif, "Inferring ties in social iot using location-based networks and identification of hidden suspicious ties," *Scientific Programming*, vol. 2020, Article ID 6667610, 16 pages, 2020.

[36] Z. Hu, A. Lin, and P. Willett, "Identification of research communities in cited and uncited publications using a co-authorship network," *Scientometrics*, vol. 118, no. 1, pp. 1–19, 2019.

[37] J. Luo, A. P. Sinha, and H. Zhao, "Location-sensitive friend recommendations in online social networks," in *Proceedings of the 2020 Pacific Asia Conference on Information Systems*, p. 155, Dubai, UAE, June 2020.

[38] N. Bibi, M. Sikandar, I. Ud Din, A. Almogren, and S. Ali, "IoMT-based automated detection and classification of leukemia using deep learning," *Journal of Healthcare Engineering*, vol. 2020, Article ID 6648574, 12 pages, 2020.

[39] B. Liao, Y. Ali, S. Nazir, L. He, and H. U. Khan, "Security analysis of iot devices by using mobile computing: a systematic literature review," *IEEE Access*, vol. 8, pp. 120 331–120 350, 2020.

[40] K. M. Kumar and B. Srinivasan, "Point-of-interest based classification of similar users by using support vector machine and status homophily," *International Journal of Machine Learning and Computing*, vol. 9, no. 5, pp. 615–620, 2019.

[41] S. Guha and S. B. Wicker, "Do birds of a feather watch each other? homophily and social surveillance in location based social networks," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 1010–1020, Vancouver, BC, Canada, March 2015.

[42] E. Toch, B. Lerner, E. Ben-Zion, and I. Ben-Gal, "Analyzing large-scale human mobility data: a survey of machine learning methods and applications," *Knowledge and Information Systems*, vol. 58, no. 3, pp. 501–523, 2019.

[43] G. Jadoon, I. Ud Din, A. Almogren, and H. Almajed, "Smart and agile manufacturing framework, a case study for automotive industry," *Energies*, vol. 13, no. 21, p. 5766, 2020.

[44] M. A. Khan, S. Israr, A. S. Almogren, I. U. Din, A. Almogren, and J. J. Rodrigues, "Using augmented reality and deep learning to enhance taxila museum experience," *Journal of Real-Time Image Processing*, pp. 1–12, 2020.

[45] M. Yohsuke, J. Hang-Hyun, T. János, K. János, and K. Kimmo, "Structural transition in social networks: the role of homophily," *Scientific Reports (Nature Publisher Group)*, vol. 9, no. 1, 2019.

[46] D. Cepić and Ž. Tonković, "How social ties transcend class boundaries? Network variability as tool for exploring occupational homophily," *Social Networks*, vol. 62, pp. 33–42, 2020.

[47] A. Asikainen, G. Iñiguez, J. Ureña-Carrión, K. Kaski, and M. Kivelä, "Cumulative effects of triadic closure and homophily in social networks," *Science Advances*, vol. 6, no. 19, Article ID eaax7310, 2020.

[48] Y. Gu, Y. Yao, W. Liu, and J. Song, "We know where you are: home location identification in location-based social networks," in *Proceedings of the 2016 25th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–9, IEEE, Waikoloa, HI, USA, August 2016.

[49] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[50] D. Doran, H. Alhazmi, and S. S. Gokhale, "Triads, transitivity, and social effects in user interactions on facebook," in *Proceedings of the 2013 Fifth International Conference on Computational Aspects of Social Networks*, pp. 68–73, IEEE, Fargo, ND, USA, August 2013.

[51] P. W. Holland and S. Leinhardt, "Local structure in social networks," *Sociological Methodology*, vol. 7, pp. 1–45, 1976.

[52] J. Leskovec and A. Krevl, "SNAP datasets: stanford large network dataset collection," 2014, http://snap.stanford.edu/data.

[53] J.-P. Onnela, J. Saramäki, J. Hyvönen et al., "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

Hindawi
Scientific Programming
Volume 2021, Article ID 6662097, 10 pages
https://doi.org/10.1155/2021/6662097

Research Article

Monitoring, Analyzing, and Modeling for Single Subsidence Basin in Coal Mining Areas Based on SAR Interferometry with L-Band Data

Zhiyong Wang,[1] Jingzhao Zhang,[1,2] Yaran Yu,[1] Jian Liu,[1] Wei Liu,[1] Na Jiang,[3] and Donge Guo[3]

[1]*College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China*
[2]*Shandong Institute of Geological Surveying and Mapping, Jinan 250011, China*
[3]*Shandong Institute of Land Surveying and Mapping, Jinan 250013, China*

Correspondence should be addressed to Jingzhao Zhang; 564158441@qq.com

Received 16 December 2020; Revised 15 January 2021; Accepted 24 January 2021; Published 9 February 2021

Academic Editor: Habib Ullah Khan

Excessive exploitation of underground mine resources has caused serious land subsidence in China. This paper focused on monitoring and modeling the single subsidence basin in coal mining area based on SAR interferometry (InSAR). The optimum InSAR processing strategy to monitor the mining subsidence was built to obtain the land subsidence with large deformation. And a method of three-dimensional mathematical modeling of single subsidence basin based on InSAR measurements was presented. Using Jining Coalfield (China) as the study area, we acquired 7 L-band PALSAR images from January 2008 to February 2010 to monitor the land subsidence in Jining Coalfield. The deformation maps in Jining Coalfield in different periods were obtained. Taking the Geting Coal Mine within the Jining coalfield as an example, we finely analyzed and interpreted the deformation maps. Compared with the simultaneous filed measurements, the precision of deformation measurement using D-InSAR in mining area was analyzed. The root mean square error was 1.37 cm. The method of fine interpretation and analysis for a single subsidence basin was established. The experiments have proved that InSAR technique with L-band InSAR data is suitable for monitoring mining subsidence with large deformation. And the 3D mathematical modeling method could be used for the single subsidence basin in coal mining area.

1. Introduction

Monitoring the land subsidence over mining regions is one of the most important tasks in monitoring the geographical conditions. Excessive exploitation of underground mine resources has caused serious land subsidence and ruined farmland and some water pit collapse in China [1, 2]. It has become one of the most serious problems in restricting the environmental, social, and economic sustainable development in coal mining area. So, it is urgent to obtain the information about the land subsidence in coal mining area. The traditional monitoring methods mainly include the leveling, Global Positioning System (GPS), and total station [2, 3]. These methods have some limitations such as needing much field work, being time consuming and laborious, and having high cost [1], and the observation points are difficult to preserve. In addition, the update period is too long and the measuring data is discrete. So, we should seek a new monitoring method with low cost, short production period, and continuous data in monitoring the land subsidence over mining regions.

The space-borne Interferometric Synthetic Aperture Radar (InSAR) is a new technique for Earth observation in the late 1900s [4–7]. It provides a new method to monitor the Earth surface deformation. It can quickly get the large-area surface deformation with high precision. And InSAR can identify some previously unknown land subsidence areas. It has been turned out to be an effective technique for

land subsidence measurement due to its precision, spatial coverage, and resolution [8, 9]. The capability of InSAR for surface deformation mapping has been demonstrated in many applications, such as earthquake activity [10], volcanic activity [11, 12], the land subsidence in the city caused by groundwater over exploitation [13, 14], landslide [15], and glacier movement [16].

InSAR technique has been applied to monitor the land subsidence in coal mining area [17–30]. Ji et al. [18] demonstrated InSAR's ability to cost-effectively monitor illegal mining activities. A DInSAR-based illegal-mining detection system (DIMDS) was proposed to exploit the geometric, spatial, and temporal characteristics of those subsidence patterns [19]. Zheng et al. [20] analyzed land subsidence induced by coal mining in a 200 km^2 area in the Ordos Basin for the time period 2006–2015 using SBAS InSAR and D-InSAR. Hayman et al. [21] investigated the performance of the three satellite missions (Radarsat-2, Sentinel-1, and ALOS-2) with different imaging modes for mapping longwall mine subsidence. Yang et al. [22] presented a novel space-based method for locating and defining the underground goaf caused by coal extraction using Interferometric Synthetic Aperture Radar (InSAR) techniques. Xia and Wang [23] proposed a method that relied on the principle of the probability integration method (PIM) and on synthetic aperture radar interferometry (InSAR) to retrieve the location of an underground goaf. Du et al. [24] proposed a feature-points-based method for the efficient location of mining goafs based on D-InSAR. Chen et al. [25] employed the small baseline subset interferometry synthetic aperture radar (SBAS-InSAR) technology to obtain the time-series residual surface deformation based on the 40 Sentinel-1A images acquired from 14th February 2017 to 17th May 2020.

Although InSAR technique has been applied to monitor the land subsidence over mining regions, the special surface environment in mining area and the characteristics of mining subsidence restrict the application of InSAR technique in coal mining area on a large scale. There have been some problems and difficulties in monitoring the land subsidence in mine area. For example, too large deformation will exceed the maximum deformation gradient [31] that InSAR can measure; the coherence caused by high vegetation land cover is poor; and the reliability and accuracy of InSAR monitoring are low. These increase the difficulty in monitoring the land subsidence with InSAR technique.

According to the problems about obtaining the mining subsidence information, we carried out some studies to obtain the land subsidence based on InSAR technique. We will explore a suitable and feasible method and technical process related to the InSAR data processing. In this paper, the major objective is to provide an effective solution to obtain accurate and critical information on the land subsidence in coal mining area.

This paper is organized as follows. The study area and SAR data are presented in Section 2. Section 3 describes the method and data processing strategy for monitoring the surface deformation with InSAR technique. And a method of three-dimensional mathematical modeling of subsidence basin based on InSAR measurements is presented. In Section 4, taking the Jining Coalfield (China) as study area, we obtain the land subsidence using InSAR technique with PALSAR data. And we analyze and interpret the results of mining subsidence based on InSAR technique. Finally, some valuable conclusions drawn from this study are given in Section 5.

2. Data and Materials

2.1. Study Area. The study area is located in the Jining City, which is the west-south part of the Shandong Province, North China. The region extends from 116.36°E to 116.94°E and from 35.32°N to 35.54°N (see Figure 1). There have been more than 20 coal mines, such as Liangbaosi Coal Mine, Geting Coal Mine, Tangkou Coal Mine, Daizhuang Coal Mine, Xuchang Coal Mine, Nantun Coal Mine, Dongtan Coal Mine, and Gucheng Coal Mine. It caused serious surface collapse because of long-term and high-intensitive coal mining.

2.2. Data. In order to monitor the mining-induced land subsidence with large deformation in coal mining area, the ALOS PALSAR data, which are L band (the wavelength is 23.6 cm), were used. Its central frequency is 1270 MHz. This means it has greater penetration.

We used a total of 7 L-band ALOS PALSAR images acquired from January 2008 to February 2010 over the Jining coalfield from an ascending orbit, as listed in Table 1. All PALSAR data were Fine Beam Single Polarization (FBS) imaging mode (single-look complex images, CEOS (Committee for Earth Observing Satellites) standard format and Level 1.1 products) and in HH polarization with the 34.3° incidence angle. The ALOS PALSAR data has a swath of about 70 km and a spatial resolution of about 7 m. The satellite repeat period of ALOS is 46 days.

From the optical remote sensing image, we can see that there are mostly farmlands in the study area and the vegetation is rich and well grown. This increases the difficulty in monitoring the land subsidence with InSAR technique.

In addition, the SRTM DEM in this region was used to remove the flat Earth phase in InSAR data processing and to geocode some products. It can also be used to do SAR simulating processing to remove the phase due to the topography [6, 7, 26].

3. Methods

SAR interferometry can provide the mining subsidence information and spatial-temporal evolution about the surface deformation based on time series radar data. The following is a brief introduction of the basic principle and process of InSAR about monitoring the mining deformation.

3.1. The Basic Principle for Monitoring the Land Subsidence Using InSAR Technique. In fact, the phase of interferogram consists of 5 parts as follows [6, 7]:

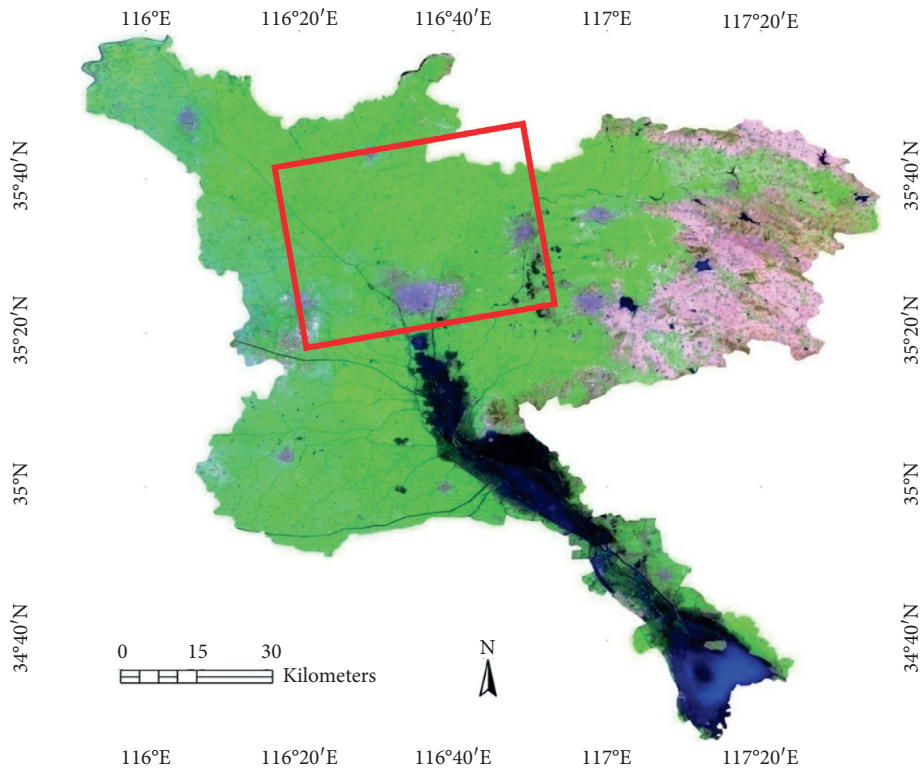


Figure 1: The location of the Jining coalfield is shown. The coverage of SAR data in Jining coalfield, Shandong province, China, is shown. The base map is the mosaic image product of Jining district from multiple Landsat5 TM data from September 2005 to March 2007. The red box is the coverage of PALSAR (path: 449, frame: 700).

Table 1: The SAR data used in this study.

No.	Sensor	Acquisition date	Orbit	Path/track	Frame	Incidence angle (°)
1	PALSAR	08/01/2008*	10,423	449	700	34.3
2	PALSAR	23/02/2008	11,094	449	700	34.3
3	PALSAR	09/04/2008	11,765	449	700	34.3
4	PALSAR	10/01/2009	15,791	449	700	34.3
5	PALSAR	25/02/2009	16,462	449	700	34.3
6	PALSAR	13/01/2010	21,159	449	700	34.3
7	PALSAR	28/02/2010	21,830	449	700	34.3

*date/month/year.

$$\varphi_{\text{int}} = \varphi_{\text{topography}} + \varphi_{\text{displacement}} + \varphi_{\text{atmosphere}} + \varphi_{\text{flat}} + \varphi_{\text{noise}}, \quad (1)$$

where $\varphi_{\text{topography}}$ is the phase due to the topography; $\varphi_{\text{displacement}}$ is the phase due to the surface deformation at line of sight (LOS) of radar; $\varphi_{\text{atmosphere}}$ is the phase due to the atmospheric effects; φ_{flat} is the flat Earth phase due to the special imaging geometry, side looking imaging; φ_{noise} is the phase noises from the speckle due to coherence imaging, system noise, and radar shadow.

For monitoring the surface deformation with InSAR technique, there are three methods, named as two-pass method, three-pass method, and four-pass method [4, 6, 7, 10]. The basic principle of InSAR can be found in the literatures [6, 7, 10]. Two-pass approach differential interferometry is more suitable for monitoring the land subsidence with large deformation [23, 26]. It needs two sets of radar data acquired from the similar orbit and the DEM with high precision.

3.2. The Optimum InSAR Processing Strategy to Monitor the Mining Subsidence. The procedures of two-pass D-InSAR include the interferogram generation, the SAR simulation based on DEM, the differential processing between the real interferogram and the simulated interferogram, the phase unwrapping, the transformation from the phase to deformation, the geocoding, and so on [6, 7]. The methods and flowchart of data processing can be seen in [6, 7, 10]. Figure 2 is the technical flowchart for monitoring the mining subsidence with InSAR technique in our study.

When the coherence of InSAR pair is low, for example, in the densely vegetated area, a prefilter (including the

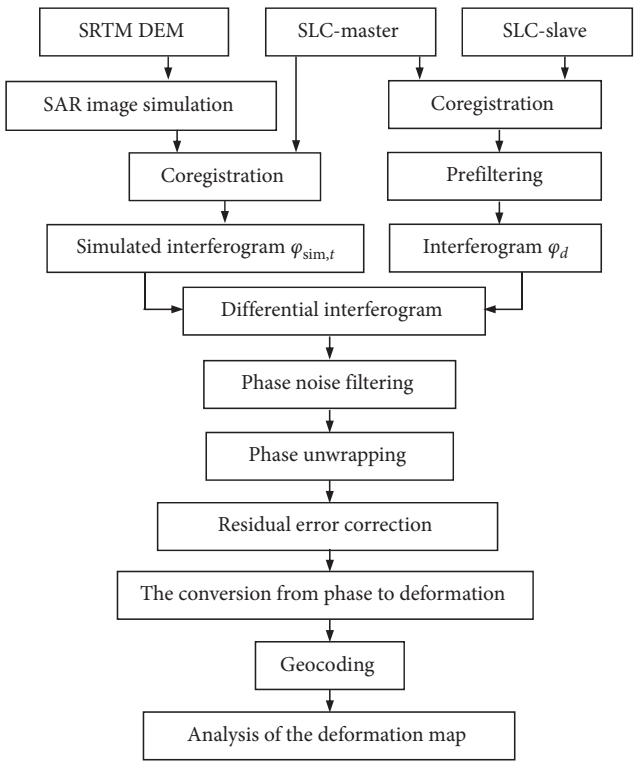


Figure 2: The block diagram of the approach implemented in our study.

spectral shift filter and Doppler filter) is necessary in InSAR data processing [6]. The spectra of master and slave acquisitions are not completely overlapping. The spectral shift filter is intended to remove the part of the master and slave spectra which are not overlapping. The Doppler filter can remove the portion of the azimuth spectra, which are not common between master and slave image. The prefilter can obviously improve the coherence of interferogram and therefore improve the reliability and accuracy of InSAR measuring.

In addition, we also proposed an optimum strategy from coarseness to fine in InSAR processing. The specific method was as follows. Firstly, it carries out the coarse differential interferometric processing for the whole image. The number of looks in range direction was selected bigger in multilooking processing now. For example, the multilook of ALOS PALSAR data is 4 : 10 in the range direction and azimuth direction, respectively. The differential interferogram should not carry out the subsequent data processing, such as the phase unwrapping and the transformation from phase to deformation. We can find several settlement regions according to the differential interferogram. Then, we subset the radar data to several parts according to the locations of settlement regions, which include one or two coal mines. At last, it carries out the fine differential interferometric processing and subsequent data processing for every part. The number of looks in the range direction should be as small as possible in multilooking processing now. For example, the multilook of ALOS PALSAR data is 1 : 2 in the range direction and azimuth direction, respectively. This strategy can not only accelerate the speed of data processing, but also ensure the accuracy of monitoring results. Especially for the phase unwrapping, it can obtain more reliable result in small region.

3.3. The Method of Three-Dimensional Modeling for Subsidence Basin Based on InSAR Measurements. After underground coal mining, a series of subsidence basins will form in the mining area. Based on analyzing a large number of interferograms using SAR interferometry in coal mine area, we found that they are usually manifested as a series of concentric circles or concentric ellipses with similar shapes for the single subsidence basin in the InSAR interferograms [32]. In order to conduct quantitative analysis of single subsidence basin, the InSAR monitoring results can be used to establish the mathematical model of the single subsidence basin. Through a series of experimental verifications, especially the analysis of the morphology of the horizontal section and vertical section of the subsidence basin, the mathematical model of the subsidence basin in the mining area can be established:

$$h = -ae^{-\left(\left(\left(x-x_0\right)^2\right)/b^2\right)+\left(\left(\left(y-y_0\right)^2\right)/c^2\right)\right)}, \tag{2}$$

where h is the settlement; (x, y) is the plane coordinates of settlement points; (x_0, y_0) is the position of maximum subsidence center; a is the influence of the subsidence factor; b and c are the semimajor axis and semiminor axis of an elliptic equation, respectively. That is to say, 5 parameters of the mathematical model should be needed to solve: x_0, y_0, a, b, c. The parameters x_0 and y_0 determine subsidence basin the location of the maximum settlement. The parameter a determines the size of the ground settlement shape. The parameters b and c determine the geometric shapes of subsidence basin. These five parameters will determine the position and form of subsidence basin in space.

Parameter a is obtained directly according to the maximum settlement amount of the subsidence basin monitored by InSAR. Parameters x_0 and y_0 are determined by the position of the maximum settlement amount of the subsidence basin. The maximum settlement amount and its position are detected and recorded through two-dimensional search in the deformation map. The other two parameters, b and c, determine the shape of the ellipse. And the solutions can be obtained by means of least square fitting based on some InSAR measurements at settlement points.

4. Results and Discussion

4.1. The Differential Interferogram in Jining Coalfield. In order to monitor the land subsidence in Jining coalfield in detail, we carried out the differential InSAR processing for the 7 PALSAR radar data. We built 6 optimum interferometric pairs according to the parameters of the time of data acquisition and the baselines. The information about the InSAR pairs can be seen in Table 2.

We carried out the interferometric data processing for all InSAR pairs according to the processing flowchart in Figure 2. The InSAR complex data registration adopted the

Table 2: The information about compositions and baselines of interferometric pairs.

No.	InSAR pair	Temporal baseline (d)	Perpendicular baseline $B_{\perp}$ (m)	Parallel baseline $B_{\parallel}$ (m)
1	08/01/2008–23/02/2008	46	679.3	294.6
2	23/02/2008–09/04/2008	46	443.4	397.5
3	09/04/2008–10/01/2009	6 * 46	−3914.2	−2833.0
4	10/01/2009–25/02/2009	46	194.9	154.8
5	25/02/2009–13/02/2010	7 * 46	1713.3	1216.4
6	13/02/2010–28/02/2010	46	604.1	451.1

automatic search technique based on window. The phase noise was filtered using the modified Goldstein Radar Interferogram Filter [33, 34]. The phase unwrapping process becomes difficult due to the presence of large areas of low coherence. In this case, the minimum cost flow (MCF) algorithm [35] enables obtaining better results than other methods. The ratio of multilooking is 1 : 2. The pixel size in range direction is 7.49 m, and the pixel size in azimuth direction is 6.15 m for the differential interferograms. It is necessary to carry out the processing of resampling because the pixel size is not the same in range direction and in azimuth direction. In order to further analyze the subsidence of deformation map, geocoding the differential interferograms is in need. They are the results of geocoded differential interferograms as shown in Figure 3.

4.2. The Mining-Induced Land Deformation Fields in Jining CoalField. Then, we can obtain the land deformation maps for every InSAR pairs after the conversion from the phase to deformation. The deformation maps have carried out some data processing procedures, such as the residual phase correction, the conversion from phase to deformation, and geocoding the products. They have absolute geographical coordinates. According to the amount of deformation, the settlement is classified with different colors (see Figure 4).

Through experiment, we also found that the InSAR pairs with too long interval or with too long perpendicular baseline cannot generate distinct interferometric fringes.

We also calculated the area of the land subsidence for the several important coal mines in different time intervals. They include Liaobaosi Coal Mine, Geting Coal Mine, Yunhe Coal Mine, Tangkou Coal Mine, and Daizhuang Coal Mine. The areas of land subsidence are listed in Table 3. From Table 3, we can find that the land subsidence is very serious due to excessive exploitation of underground mine resources. The land subsidence of these 6 coal mines within Jining coalfield exceeds 6 km^2.

A magnitude of 94.4 cm was firstly monitored by L-band InSAR in Jining coalfield. It appeared around the Dongtan Coal Mine in time interval from 10th January to 25th February, 2009. The radius of this subsidence basin is about 350 m and the major influence radius of this subsidence basin is about 256 m.

4.3. Accuracy Verification. In order to verify and evaluate the accuracy of settlement monitoring in InSAR mining area, precise leveling observation was carried out simultaneously in the study area.

The comparison of monitoring results between InSAR technique and leveling is shown in Figure 5. According to Figure 5, the deformation trend of InSAR monitoring results and leveling monitoring results is basically consistent, and the root mean square error was 1.37 cm. The root mean square error was relatively large. This was mainly caused by the inconsistency of observation time and space. Firstly, the time period of InSAR technique and leveling monitoring was not completely consistent. The time period of InSAR monitoring was from 13th January to 28th February, 2010, while the time period of leveling was from 11th January to 26th February, 2010. They were not completely consistent. Secondly, leveling observation monitors the deformation information on a "point," while InSAR technique deals with the deformation information on a "surface" (i.e., pixel). In fact, it is the comparison between "point" and "surface," so they cannot correspond exactly.

4.4. Fine Analysis and Interpretation for Single Subsidence Basin. It is very important to analyze the subsidence conditions of single coal mine. In the following, we take the interferometric pair with the time interval from 8th January to 23rd February, 2008 in Geting Coal Mine within Jining coalfield as an example to illustrate the fine analysis for the mining subsidence. We can generate the geocoded deformation map, the deformation contours, the subsidence profile, and three-dimensional deformation map for the single coal mine. Figure 6 is the results of fine analysis for single coal mine, taking the Geting Coal Mine as an example. Then, the subsidence area can be counted based on the deformation map. Table 4 is the statistics of subsidence area of Geting Coal Mine.

The maximum land subsidence in this time interval is 39.3 cm. The area which the settlement exceeds 5 cm reached 0.24 km^2. And the remaining settlement areas statistics are shown in Table 4.

4.5. Modeling of Subsidence Basin Based on InSAR Measurements. Taking Geting Coal Mine as an example, 46 points are selected to participate in the calculation. And the fitting model parameters can be obtained by calculating according to the least square method. In the image coordinate system, $b = 36.13$ and $c = 34.17$.

Thus, the three-dimensional model of the subsidence basin of Geting Coal Mine can be established as follows:

$$h = -0.393e^{-\left(\left((x-175)^2/36.13^2\right)+\left((y-159)^2/34.17^2\right)\right)}. \quad (3)$$

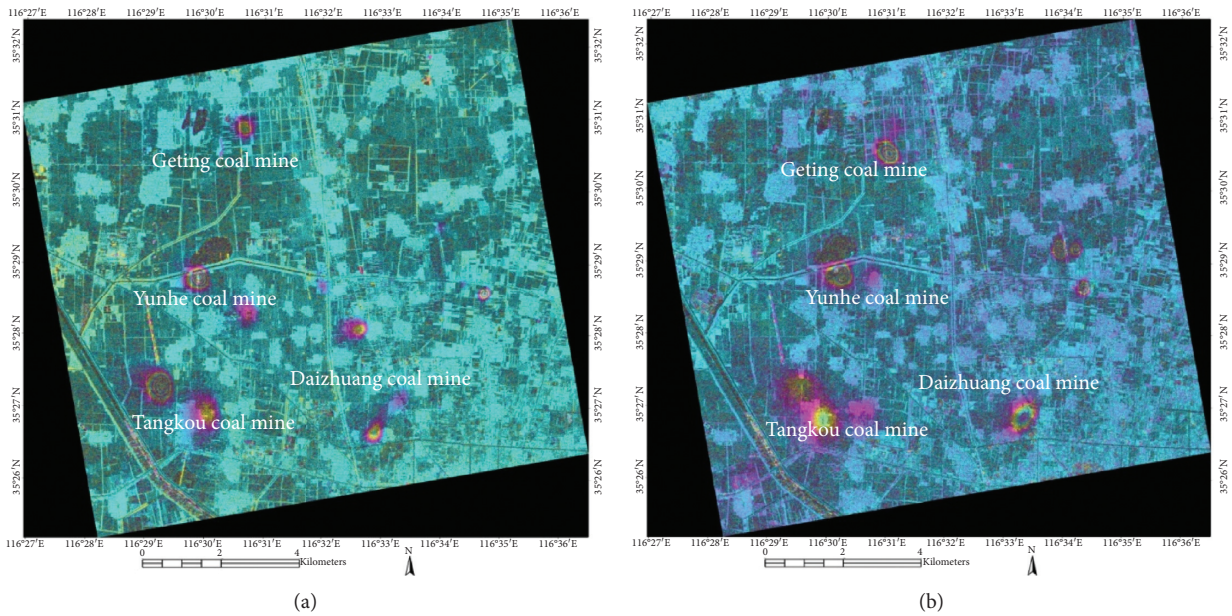


Figure 3: The geocoded differential interferogram in experimentation area. Two of the differential interferograms are only shown here. There are several distinct settlement regions in the differential interferograms. They are Geting Coal Mine, Yunhe Coal Mine, Tangkou Coal Mine, Daizhuang Coal Mine, and Xuchang Coal Mine. (a) is the geocoded differential interferogram in time interval from 8th January to 23rd February 2008; (b) is the geocoded differential interferogram in time interval from 10th January to 25th February, 2009.

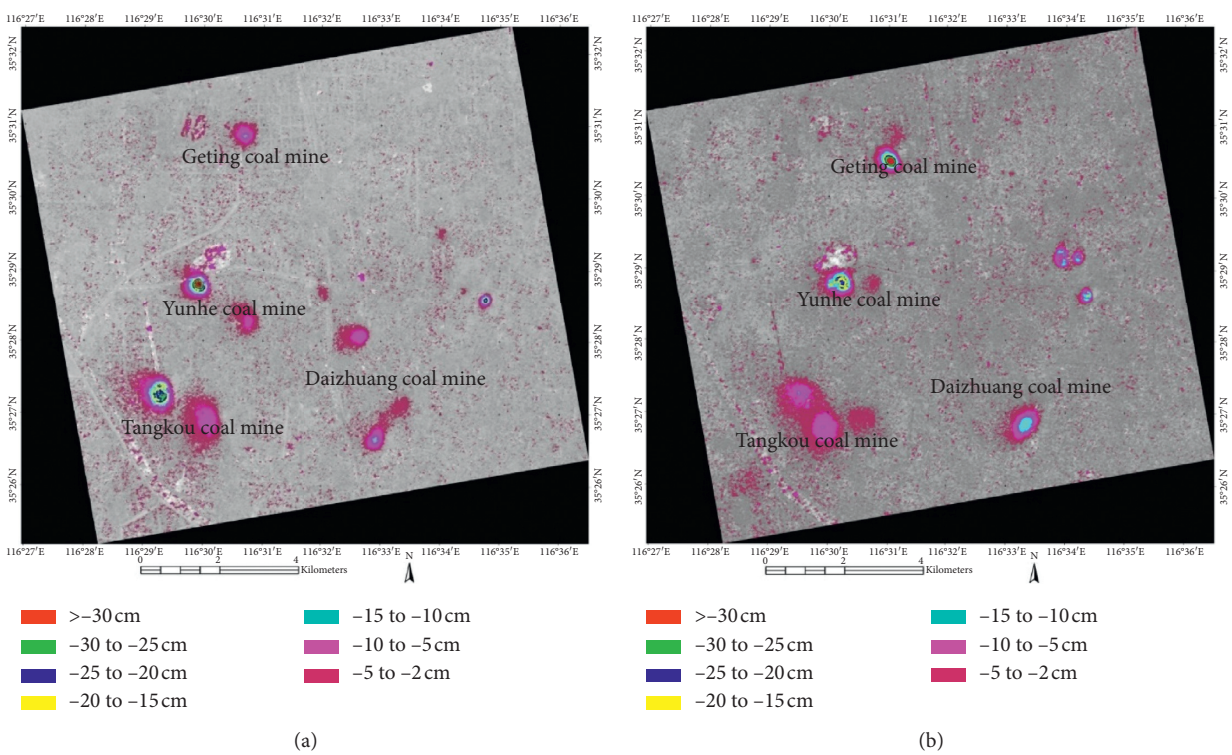


Figure 4: The deformation maps in Jining coalfield in two periods. (a) is the deformation map in time interval from 8th January to 23rd February, 2008; (b) is the deformation map in time interval from 10th January to 25th February, 2009.

In view of the transformation of image coordinate system to geographic coordinate system, pixel size and geographic coordinate system need to be considered.

Figure 7 is the 3D display of the established model. From Figure 7, it can be seen that the 3D model is very consistent with the subsidence basin monitored by InSAR. Therefore,

Table 3: The area statistics of the land subsidence of several important coal mines within Jining coalfield.

Coal Mine	The area of land subsidence (km^2)			
	08/01/2008–23/02/2008	23/02/2008–09/04/2008	10/01/2009–25/02/2009	13/01/2010–28/02/2010
Liaobaosi Coal Mine	0.672	0.778	0.868	0.418
Geting Coal Mine	0.187	0.070	0.416	0.209
Yunhe Coal Mine	0.360	0.227	0.503	—
Tangkou Coal Mine	0.755	0.902	1.757	1.664
Daizhuang Coal Mine	0.299	0.315	0.572	0.468
Dongtan Coal Mine	0.882	0.752	1.871	1.564

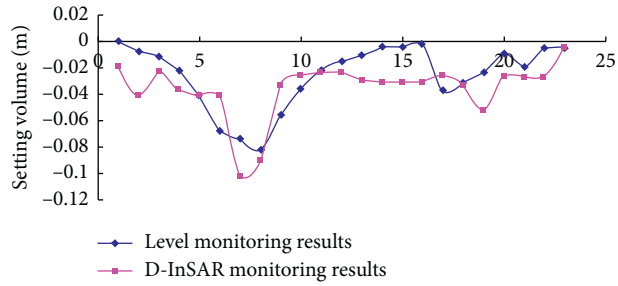


Figure 5: Comparison of monitoring results between D-InSAR and leveling.

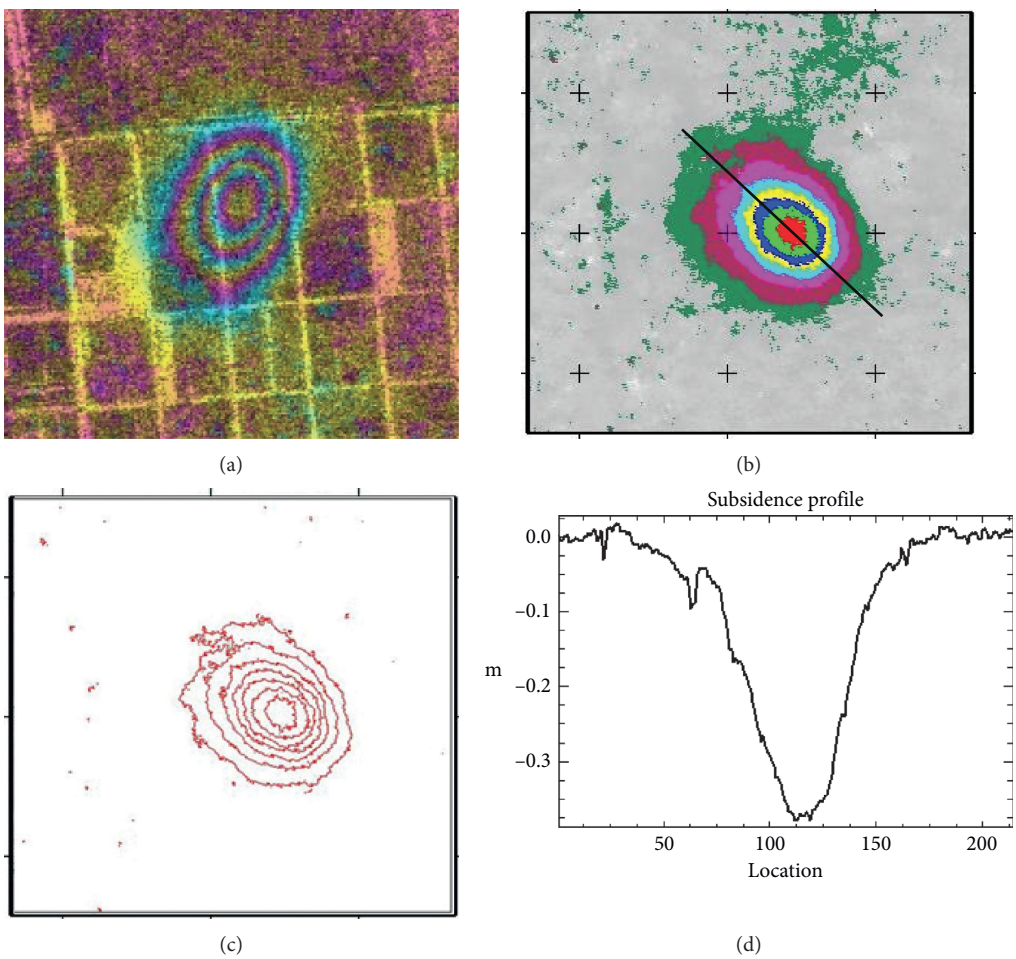


Figure 6: The fine analysis and interpretation for the deformation map in Geting Coal Mine in the interval from 10th January to 25th February, 2010. (a) is the original differential interferogram over the Geting Coal Mine in range direction and azimuth direction coordinate system. It has carried out the interferometric processing, removal of the flat-earth phase, and phase filtering. There are 3 fringes in the interferogram. That is to say, the maximum subsidence is about 35 cm. (b) is the geocoded deformation map in the Geting Coal Mine. It has carried out the phase unwrapping, the conversion from the deformation phase to land subsidence, and geocoding. The UTM (Universal Transverse Mercator) projection is selected in the map projection processing. The pixel size of geocoded map is 5 m. (c) The deformation contours in the Geting Coal Mine are shown. The deformation contours from outer to inner are −2 cm, −5 cm, −10 cm, −15 cm, −20 cm, −25 cm, and −30 cm, respectively. (d) The subsidence profile along the coal mining working face is shown.

Table 4: The area statistics of the land subsidence in Geting Coal Mine (10/01/2009–25/02/2009).

Land subsidence	>5 cm	>10 cm	>15 cm	>20 cm	>25 cm	>30 cm	>35 cm	>40 cm
Pixels	9,549	5,863	3,772	2,586	1,746	1,012	348	0
Area (m^2)	238,725	146,575	94,300	64,650	43,650	25,300	8,700	0

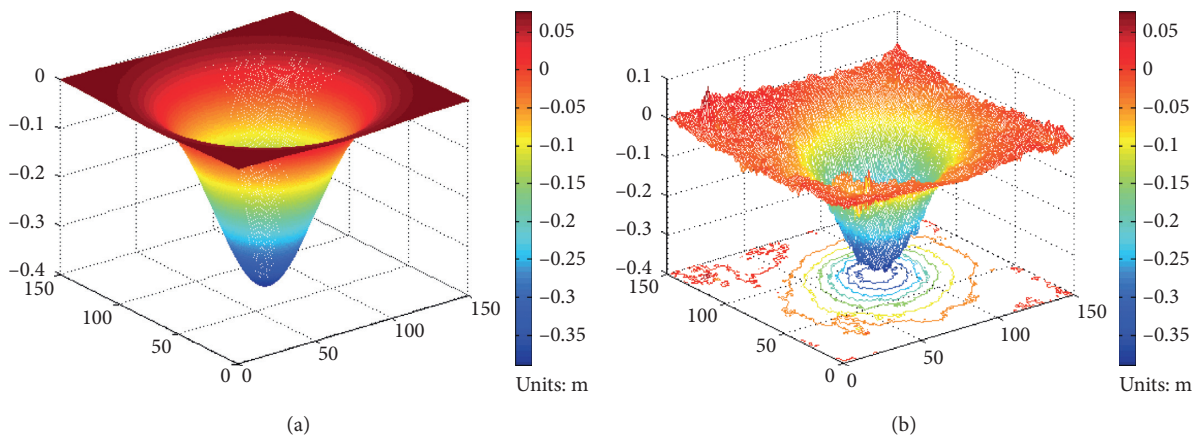


Figure 7: The 3D display of the established model in Geting Coal Mine. (a) is the 3D display of mathematical model. (b) is 3D display of the subsidence basin based on InSAR measurements.

based on the above mathematical model, quantitative analysis and simulation and early warning can be conducted on the mining settlement based on the monitoring results of InSAR.

5. Conclusions

In this paper, the land subsidence in coal mining area was monitored using InSAR technique. Using 7 scenes of L-band PALSAR data from January 2008 to February 2010, we successfully obtained the mining subsidence deformation maps in the Jining Coal Mine during different periods based on an optimum InSAR data processing flowchart and strategy.

Through this study, we got some valuable conclusions in monitoring the land subsidence in coal mining area with InSAR technique.

(1) In the Jining coalfield, some subsidence basins with the radius of tens of meters to one hundred or several hundreds were formed. Generally, the maximum deformation of the subsidence basin ranges from 30 cm to 50 cm. The land subsidence of 6 coal mines within Jining coalfield exceeds 6 km^2.

(2) The magnitude of the land subsidence in the coal mine is larger. For the larger deformation, it is easier to monitor the land subsidence using SAR interferometry with L-band data. Therefore, SAR interferometry with L-band data is an effective technique for mapping the land subsidence in mining area. In particular, SAR interferometry can detect some unknown subsidence basins.

(3) Compared with the simultaneous filed measurements, the precision of deformation measurement using D-InSAR in mining area was analyzed. The root mean square error was 1.37 cm. It can meet the needs of monitoring the mining subsidence.

(4) The method of three-dimensional mathematical modeling based on InSAR measurements is suitable for the single subsidence basin in the coal mine. The mathematical model can be used to quantitative analysis and simulation and early warning in the coal mine. In addition, for some interrupted or confused interferometric fringes caused by phase noise, the three-dimensional model of subsidence basin constructed in this paper can also be used to solve these problems, so that InSAR technology can be better applied to monitor the land subsidence in mining areas with large deformations.

When adequate SAR data are available, InSAR can partially replace the traditional leveling method for monitoring mining-induced subsidence. Therefore, InSAR technique can provide an efficient technique in monitoring the land subsidence in coal mining area.

With regard to monitoring the land subsidence in mining area, it should be noted that radar data can be used to obtain not only the quantitative ground subsidence but also the information about the land cover and the land change. In the future, we will focus on mining the surface coverage and surface changes using multitemporal SAR data.

Data Availability

The SRTM DEM data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The SRTM DEM data was provided by the National Imagery and Mapping Agency. This work was funded by the Major Science and Technology Innovation Projects of Shandong Province (2019JZZY020103). This research was supported in part by the National Natural Science Foundation of China (no. 41876202) and the Shandong Province Natural Science Foundation (no. ZR2017MD020).

References

[1] H.-d. Fan, W. Gu, Y. Qin, J.-q. Xue, and B.-q. Chen, "A model for extracting large deformation mining subsidence using D-InSAR technique and probability integral method," *Transactions of Nonferrous Metals Society of China*, vol. 24, no. 4, pp. 1242–1247, 2014.

[2] Z. Li, D. Zhang, J. Wang, Q. Zhao, and L. Pan, "Mapping land subsidence related to underground coal fires in the Wuda coalfield (northern China) using a small stack of ALOS PALSAR differential interferograms," *Remote Sensing*, vol. 5, pp. 1152–1176, 2013.

[3] L. Jiang, H. Lin, J. Ma, B. Kong, and Y. Wang, "Potential of small-baseline SAR interferometry for monitoring land subsidence related to underground coal fires: Wuda (Northern China) case study," *Remote Sensing of Environment*, vol. 115, no. 2, pp. 257–268, 2011.

[4] H. A. Zebker and R. M. Goldstein, "Topographic mapping from interferometric synthetic aperture radar observations," *Journal of Geophysical Research*, vol. 91, no. B5, pp. 4993–4999, 1986.

[5] A. K. Gabriel, R. M. Goldstein, and H. A. Zebker, "Mapping small elevation changes over large areas: differential radar interferometry," *Journal of Geophysical Research*, vol. 94, no. B7, pp. 9183–9191, 1989.

[6] R. F. Hanssen, *Radar Interferometry-Data Interpretation and Error Analysis*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.

[7] D. Massonnet and K. L. Feigl, "Radar interferometry and its application to changes in the earth's surface," *Reviews of Geophysics*, vol. 36, no. 4, pp. 441–500, 1998.

[8] M. Crosetto, O. Monserrat, M. Cuevas, and B. Crippa, "Spaceborne differential SAR interferometry: data analysis tools for deformation measurement," *Remote Sensing*, vol. 3, no. 2, pp. 305–318, 2011.

[9] X. Zhou, N.-B. Chang, and S. Li, "Applications of SAR interferometry in earth and environmental science research," *Sensors*, vol. 9, no. 3, pp. 1876–1912, 2009.

[10] Z. Wang, L. Liu, and X. Zhou, "Monitoring co-seismic deformation filed of Bam earthquake in Iran using D-InSAR technique," *Northwest Seismological Journal*, vol. 30, pp. 310–316, 2008.

[11] Z. Lu, T. Masterlark, and D. Dzurisin, "Interferometric synthetic aperture radar study of Okmok volcano, Alaska, 1992–2003: magma supply dynamics and post-emplacement lava flow deformation," *Journal of Geophysical Research*, vol. 110, pp. 1–18, 2005.

[12] Z. Lu, "InSAR imaging of volcanic deformation over cloud-prone areas—Aleutian islands," *Photogrammetric Engineering & Remote Sensing*, vol. 73, no. 3, pp. 245–257, 2007.

[13] L. Zhang, Z. Lu, X. Ding, H.-s. Jung, G. Feng, and C.-W. Lee, "Mapping ground surface deformation using temporarily coherent point SAR interferometry: application to Los Angeles basin," *Remote Sensing of Environment*, vol. 117, pp. 429–439, 2012.

[14] Y. Zhang, J. Zhang, W. Gong et al., "Monitoring urban subsidence based on SAR interferometric point target analysis," *Acta Geodaetica et Cartographica Sinica*, vol. 38, pp. 482–487, 2009.

[15] P. Tantianuparp, X. Shi, L. Zhang, T. Balz, and M. Liao, "Characterization of landslide deformations in three gorges area using multiple InSAR data stacks," *Remote Sensing*, vol. 5, no. 6, pp. 2704–2719, 2013.

[16] Z. Wang, J. Liu, J. Wang et al., "Resolving and analyzing landfast ice deformation by InSAR technology combined with Sentinel-1A ascending and descending orbits data," *Sensors*, vol. 20, no. 22, p. 6561, 2020.

[17] H. Fan, K. Deng, C. Ju, C. Zhu, and J. Xue, "Land subsidence monitoring by D-InSAR technique," *Mining Science and Technology (China)*, vol. 21, no. 6, pp. 869–872, 2011.

[18] M. Ji, X. Li, S. Wu, Y. Gao, and L. Ge, "Use of SAR interferometry for monitoring illegal mining activities: a case study at Xishimen iron ore mine," *Mining Science and Technology (China)*, vol. 21, no. 6, pp. 781–786, 2011.

[19] Z. Hu, L. Ge, X. Li, K. Zhang, and L. Zhang, "An underground-mining detection system based on DInSAR," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 615–625, 2013.

[20] M. Zheng, K. Fukuyama, and K. Sanga-Ngoie, "Application of InSAR and GIS techniques to ground subsidence assessment in the Nobi plain, Central Japan," *Sensors*, vol. 14, no. 1, pp. 492–509, 2014.

[21] A. Hayman, L. Ge, Z. Du et al., "Satellite radar interferometry for monitoring subsidence induced by longwall mining activity using Radarsat-2, Sentinel-1 and ALOS-2 data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 61, pp. 92–103, 2017.

[22] Z. Yang, Z. Li, J. Zhu et al., "Locating and defining underground goaf caused by coal mining from space-borne SAR interferometry," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 112–126, 2018.

[23] Y. Xia and Y. Wang, "InSAR- and PIM-based inclined goaf determination for illegal mining detection," *Remote Sensing*, vol. 12, no. 23, p. 3884, 2020.

[24] S. Du, Y. Wang, M. Zheng, D. Zhou, and Y. Xia, "Goaf locating based on InSAR and probability integration method," *Remote Sensing*, vol. 11, no. 7, p. 812, 2019.

[25] D. Chen, H. Chen, W. Zhang et al., "Characteristics of the residual surface deformation of multiple abandoned mined-out areas based on a field investigation and SBAS-InSAR: a case study in Jilin, China," *Remote Sensing*, vol. 12, no. 22, p. 3752, 2020.

[26] Z. Wang, J. Zhang, and G. Huang, "Precise monitoring and analysis of the land subsidence in Jining coal mining area based on InSAR technique," *Journal of China University of Mining and Technology*, vol. 43, pp. 169–174, 2014.

[27] X. Xu, C. Ma, D. Lian, and D. Zhao, "Inversion and analysis of mining subsidence by integrating DInSAR, Offset tracking, and PIM technology," *Journal of Sensors*, vol. 2020, Article ID 4136837, 15 pages, 2020.

[28] J. Huang, K. Deng, and H. Fan, "An improved adaptive template size pixel-tracking method for monitoring large-gradient mining subsidence," *Journal of Sensors*, vol. 2017, Article ID 3059159, 11 pages, 2017.

[29] B. Pu, C. Li, and M. Liao, "An approach for estimating underground-goaf boundaries based on combining DInSAR with a graphical method," *Journal of Sensors*, vol. 2020, 13 pages, Article ID 9375056, 2020.

[30] S. Yun, Q. Li, and X. Meng, "On time-series InSAR by SA-SVR algorithm: prediction and analysis of mining subsidence," *Journal of Sensors*, vol. 2020, Article ID 8860225, 17 pages, 2020.

[31] M. Jiang, Z. W. Li, X. L. Ding, J. J. Zhu, and G. C. Feng, "Modeling minimum and maximum detectable deformation gradients of interferometric SAR measurements," *International Journal of Applied Earth Observation and Geoinformation*, vol. 13, no. 5, pp. 766–777, 2011.

[32] Z. Wang, S. Wang, Y. Sun et al., "A new phase unwrapping algorithm based on the GVF-Snake model of edge detection," *Journal of China University of Mining and Technology*, vol. 46, no. 6, pp. 1394–1401, 2018.

[33] R. M. Goldstein and C. L. Werner, "Radar interferogram filtering for geophysical applications," *Geophysical Research Letters*, vol. 25, no. 21, pp. 4035–4038, 1998.

[34] Z. W. Li, X. L. Ding, C. Huang, J. J. Zhu, and Y. L. Chen, "Improved filtering parameter determination for the Goldstein radar interferogram filter," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 63, no. 6, pp. 621–634, 2008.

[35] C. W. Chen and H. A. Zebker, "Phase unwrapping for large SAR interferograms: statistical segmentation and generalized network models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 8, pp. 1709–1719, 2002.

Hindawi
Scientific Programming
Volume 2021, Article ID 6633428, 10 pages
https://doi.org/10.1155/2021/6633428

Research Article

Estimating the Impact of Land Cover Change on Soil Erosion Using Remote Sensing and GIS Data by USLE Model and Scenario Design

Anmin Fu,[1] Yulin Cai,[2] Tao Sun,[1] and Feng Li[1]

[1]*Academy of Inventory and Planning, National Forestry and Grassland Administration, Beijing 100714, China*
[2]*College of Geodesy and Geomatics, Shandong University of Sciences and Technology, Qingdao, Shandong 266590, China*

Correspondence should be addressed to Yulin Cai; caiyl@sdust.edu.cn

Received 29 December 2020; Revised 22 January 2021; Accepted 29 January 2021; Published 9 February 2021

Academic Editor: Habib Ullah Khan

Great efforts have been made to curb soil erosion and restore the natural environment to Inner Mongolia in China. The purpose of this study is to evaluate the impact of returning farmland to the forest on soil erosion on a regional scale. Considering that rainfall erosivity also has an important impact on soil erosion, the effect of land use and land cover change (LUCC) on soil erosion was evaluated through scenario construction. Firstly, the universal soil loss equation (USLE) model was used to evaluate the actual soil erosion (2001 and 2010). Secondly, two scenarios (scenario 1 and scenario 2) were constructed by assuming that the land cover and rainfall-runoff erosivity are fixed, respectively, and soil erosion under different scenarios was estimated. Finally, the effect of LUCC on soil erosion was evaluated by comparing the soil erosion under actual situations with the hypothetical scenarios. The results show that both land use/cover change and rainfall-runoff erosivity change have significant effects on soil erosion. The land use and land cover change initiated by the ecological restoration projects have obviously reduced the soil erosion in this area. The results also reveal that the method proposed in this paper is helpful to clarify the influencing factors of soil erosion.

1. Introduction

Soil erosion is one of the major and most widespread types of soil degradation. The Inner Mongolia autonomous region has one of the most severe soil erosion problems among all of China's provinces [1–3]. The area experiencing soil erosion is about 79 million hectares (66.99% of the region's total area), with increasingly negative effects on agricultural productivity and on the sustainability of economic development [1, 2, 4]. Serious soil erosion leads to the deterioration of the ecological environment, low level of agricultural production, and personal poverty, which is the fundamental cause of social and economic development. To mitigate the impacts of erosion, the Natural Forest Protection Project (NFPP) and Green for Grain Project (GCP, also known as the Conversion of Cropland to Forest and Grassland Program), both of which incorporate Inner Mongolia, were launched in 1998 and 1999, respectively [5, 6]. These programs aimed to prevent soil erosion by converting farmland on steep slopes into forests. An assessment of soil erosion hazards before and after the implementation of measures can help to assess the extent of effectiveness of these recovery strategies and thus aids decision makers in determining appropriate practices and formulating conservation plans.

A number of models have proven to be effective in estimating soil erosion at different scales in previous studies. One of the most commonly used models is the universal soil loss equation (USLE) developed by Wischmeier and Smith [7]. Although the model was later modified to a new version known as RUSLE [8, 9], USLE is still widely used for its simplicity [10–12]. USLE and its revised models have been used increasingly more widely with the development and integration of RS (remote sensing) and GIS (geographical information systems) technologies because they solve the problem in which the input data of models are difficult to obtain [10–18]. The models can be successfully used to estimate soil erosion because they consider climate, topography, soil, and management practices. However, they cannot assess the impact of a single factor on soil erosion,

such as LUCC (land use and land cover change) or rainfall-runoff erosivity change.

Land cover is considered to be one of the most important factors affecting soil erosion and the investigation of soil losses due to differences or changes in land cover types is a popular research topic [19–26]. Therefore, field experiments are an ideal method for observing the differences in hydrological characteristics (e.g., runoff) between one experimental area treated with vegetation cover and other experimental areas to assess the effect of vegetation on soil erosion [22, 26, 27]. However, field experiments are time-consuming and costly. The scenario design proves to be a useful tool in investigating soil erosion under different hypothesized climate scenarios [28]. This method can also serve as a constructive method in evaluating the effect of LUCC on soil erosion. In this study, the USLE model is used to obtain the spatial and temporal patterns of soil erosion in Liangcheng County, Inner Mongolia, from 2001 to 2010, and the scenarios are constructed to evaluate the impact of LUCC on soil erosion.

2. Study Area and Data

2.1. Study Area. Liangcheng County is located in southern Ulanqab, in the Inner Mongolia Autonomous Region, North China, between 112°28′ and 112°30′ east longitude and 40°29′ and 40°32′ north latitude, with an area of 3456.12 km^2 (Figure 1). Because it is located in the transition zone between the Loess Plateau and Mongolia Plateau, there are many gullies throughout this area. As shown in Figure 1, the study area is surrounded by mountains, and the middle part is a trough basin. Located in the north and south of the study area are the Manhan Mountain and the Matou Mountain Range, and the middle of the study area is the Daihai Basin. The mountainous area covers 47.83% of the study area, comprised an approximately equal proportion (25%) of hills and basins. The elevation of the study area is between 1100 and 2300 m (Figure 1). According to the sequence from old to new, the strata in the working area include middle Archean, Mesozoic, Cenozoic, Pleistocene, and Holocene. The main part of the area is the late Archean acid intrusive area. The late K-feldspar granite has an obvious gneissic texture and few phenocrysts [29].

The area belongs to the semiarid temperate continental monsoon climate. The annual average temperature is 6.1°C, and the average annual rainfall (1989–2018) is 409.6 mm [30]. Forest covers approximately 51.7% of the study area [31] in which various trees (such as poplar, birch, and aspen) and shrubs (such as Caragana, Ostryopsis, and sea buckthorn) predominate.

2.2. Research Data. The materials used in this study are as follows: (1) monthly precipitation data of 20 years (1992–2011) from the China Meteorological Administration (CMA), used for calculating R factor in 2001 and 2010; (2) ASTER GDEM data with a resolution of 30 meters downloaded from http://reverb.echo.nasa.gov used to calculate slope gradient and slope length; (3) two Landsat5/TM scenes with a resolution of 30 meters dated August 20, 2001, and August 10, 2010, that were obtained from the USGS Glovis data archive, and terrain-corrected Level 1T scenes with geodetic accuracies of one-quarter to less than half a pixel (Figure 2), used to extract land use/cover and vegetation information; (4) available 1 : 1,000,000 soil map from the Resource and Environmental Science Data Center of the Chinese Academy of Sciences, used to calculate soil erodibility factor; and (5) woodland survey data in 2001 and 2010 from China's Liangcheng County government, auxiliary data used to extract land use/cover information.

3. Methodology

3.1. USLE Model and Parameters Estimation. The USLE model was used to assess soil erosion of the study area:

$$A = R^* K^* L^* S^* C^* P, \tag{1}$$

where A is the mean annual soil loss (t ha^{-1} year^{-1}); R is the rainfall erosivity factor (MJ mm ha^{-1}h^{-1} year^{-1}); K is the soil erodibility factor (t ha h ha^{-1}MJ^{-1} mm^{-1}); L is the slope length factor (dimensionless); S is the slope factor (dimensionless); C is the cover management factor (dimensionless); P is the erosion control practice factor (dimensionless). All parameter preparation methods are included in the subsequent content, and the calculation process is mainly completed on the Arcmap platform.

3.1.1. Rainfall and Runoff Erosivity Factor (R). R is the rainfall and runoff factor by geographic location. The greater the intensity and duration of the rain storm, the higher the erosion potential. However, for most meteorological stations, the intensity and duration are difficult to obtain, and R must be estimated based on the amount of rainfall. The R factor is calculated using the equation proposed by Wischmeier and Smith [7] and developed by Arnoldus [32]:

$$R = \sum_{1}^{12} \left(1.735 \times 10^{\left(1.5 \log_{10}\left(p_i^2/p\right) - 0.08188\right)} \right), \tag{2}$$

where p represents annual precipitation (mm) and pi represents monthly precipitation (mm).

The precipitation data from CMA used to calculate R factor is estimated by the Energy and Water Balance System (EWBMS) based on two sources of information: (1) point precipitation data from meteorological stations and (2) cloud frequency data derived from the FY2c meteorological geostationary satellite [33, 34].

3.1.2. Soil Erodibility Factor (K). "K" values represent the susceptibility of soil to erosion and the amount and rate of runoff, as measured under the standard unit plot condition. K is a measure of the susceptibility of soil particles to the

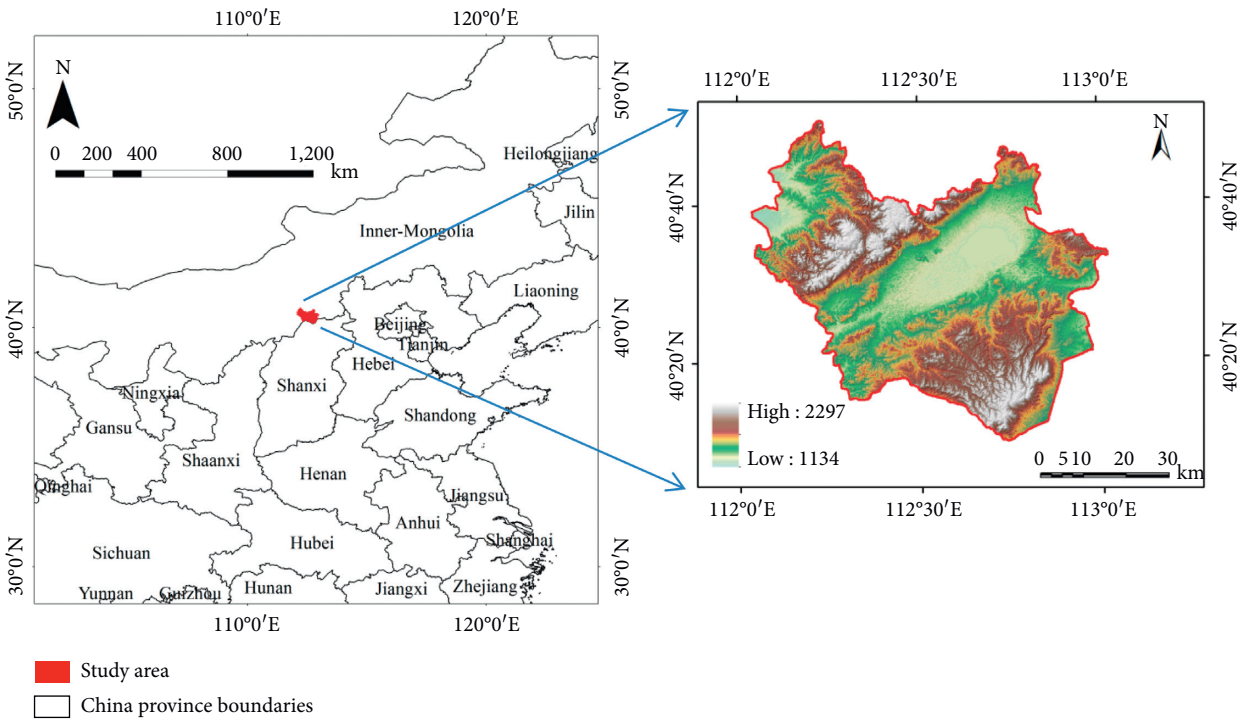


Figure 1: Location of the study area.

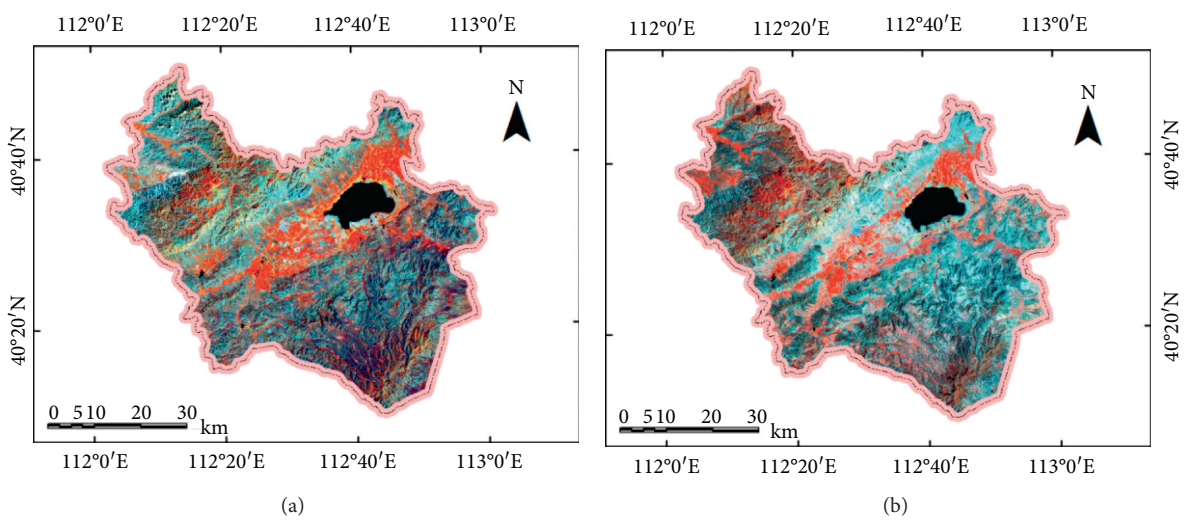


Figure 2: Landsat/TM false-color composite of the study area ((a) 2001; (b) 2010).

detachment and transport by rainfall and runoff [7]. Texture is the principal factor affecting K, but structure, organic matter, and permeability also contribute. The K factor was estimated based on the soil texture classes and organic matter content, which may be determined from the soil map with a 1 : 1,000,000 scale(Figure 3). This soil map scale is large, but it is the only available soil data in China, including soil type map and soil property data for each soil type. The K factor is calculated using the equation proposed by Williams [35]:

$$K = \left\{0.2 + 0.3 \times \exp\left[-0.0256 \times \mathrm{Sd} \times \left(\frac{1 - \mathrm{Si}}{100}\right)\right]\right\} \times \left[\frac{\mathrm{Si}}{(\mathrm{Cl} + \mathrm{Si})}\right] 0.3 \times \left\{\frac{1.0 - 0.25\mathrm{C}}{[\mathrm{C} + \exp(3.72 - 2.95\mathrm{C})]}\right\} \times \frac{[1.0 - 0.7 \times ((1 - \mathrm{Sd})/100)]}{\{((1 - \mathrm{Sd})/100) + \exp[-5.51 + 22.9 \times ((1 - \mathrm{Sd})/100)]\}}, \quad (3)$$

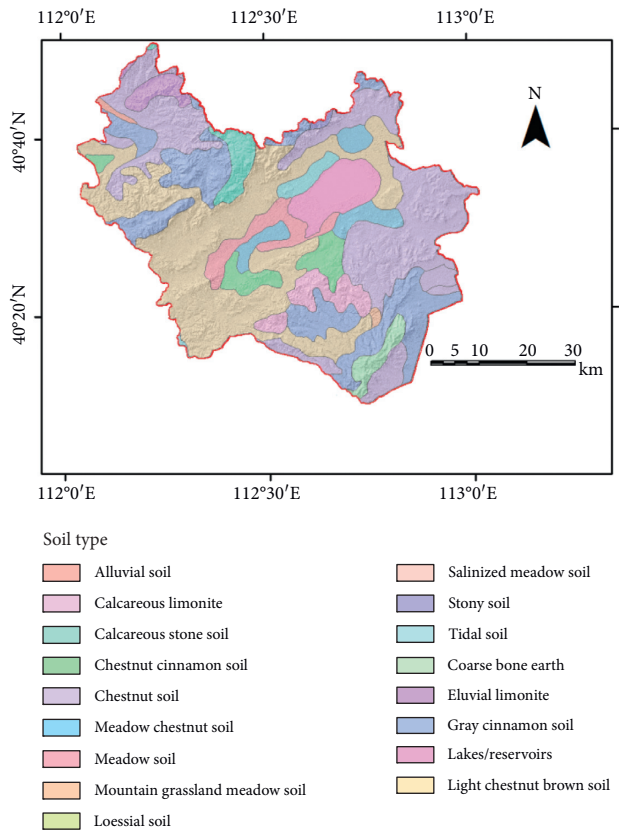


Figure 3: Soil type map.

where Sd represents the percentage of sand content; Si represents the percentage of silt content; Cl represents the percentage of clay content (%); C represents the percentage of organic matter content; the unit of k is t ha h $ha^{-1}MJ^{-1}$ mm^{-1}).

3.1.3. Slope Steepness Factor (S) and Slope Length Factor (L). The S and L factors are used to estimate the influence of topography on soil erosion in the ULSE. S and L represent the effect of slope steepness and slope length, respectively, on erosion. The computed soil erosion rates are more sensitive to slope steepness than to slope length—the steeper and longer the slope, the higher the risk for erosion.

The S factor is calculated using the equation proposed by Liu et al. [36, 37] and McCool et al. [38]:

$$S = \begin{cases} 10.8 \times \sin\theta + 0.03, & \theta < 5, \\ 16.8 \times \sin\theta - 0.5, & 5 \le \theta < 10, \\ 21.9 \times \sin\theta - 0.96, & \theta \ge 10, \end{cases} \quad (4)$$

where θ represents slope.

Slope length is the distance from the origin of overland flow along its flow path to the location of either concentrated flow or deposition, whereas L is the ratio of soil loss from the field slope length to a plot with a slope length of λ m under otherwise identical conditions.

The L factor is calculated using the equation proposed by Liu et al. [37, 39]:

$$L = \left(\frac{\lambda}{22.1}\right)^m, \quad (5)$$

where λ is the slope length and m is the slope index. λ is estimated by the equation [40, 41]: $\lambda = \text{Flowacc} * \text{constant grid size}$. Flowacc is runoff accumulation number obtained by using ArcGIS hydrological analysis module based on DEM. The m value can be assigned 0.2, 0.3, 0.4, and 0.5 for slope gradients <1%; 1%–3%; 3%–5%; and ≥5%.

3.1.4. Crop/Vegetation Management Factor (C). Soil and water conservation may be effectively improved by increasing vegetation cover. The crop/vegetation management factor represents the effect of plants, soil cover, soil biomass, and soil disturbance activities on soil erosion. It is used to determine the relative effectiveness of soil and crop management systems in terms of preventing soil loss. The C factor is a ratio that compares the soil loss from the land under a specific crop and management system to the corresponding loss from continuously fallow and tilled land. Therefore, it is dimensionless and its value is between 0 and 1.

The C factor is calculated using the equation proposed by Cai et al. [42]:

$$C = \begin{cases} 1, & \text{fc} = 0, \\ 0.6508 - 0.3436 \times \lg(\text{fc}), & 0 < \text{fc} < 78.3\%, \\ 0, & \text{fc} > 78.3\%, \end{cases} \quad (6)$$

where fc represents vegetation coverage, calculated based on NDVI derived from remotely sensed TM data, using a pixel dichotomy model [43].

3.1.5. Support Practice Factor (P). P represents the impact of support practices on erosion rates. In this study, the P factor value is assigned based on land use types derived from the land use map according to Liu et al. [44]. It is also dimensionless and its value is between 0 and 1. A 0.35 and 1.0 P factors are assigned to cultivated land and forested land, respectively, and the remaining land cover was assigned a P factor of 0.

3.2. Scenarios Construction. Land use and rainfall are the most important factors affecting soil erosion. Sometimes when studying the causes of soil and water loss, it is necessary to distinguish the effects of these two factors, that is, to study their effects separately. To do this, two scenarios (S1 and S2) are constructed according to different combinations of rainfall and land use factor:

S1: rainfall-runoff erosivity in 2001 + land use in 2010

S2: rainfall-runoff erosivity in 2010 + land use in 2001

S1 and S2 are designed to clarify the effect of rain-runoff erosivity and land use change on soil loss of the study area by comparing with the actual situation. The actual situation of 2001 can be defined as the combination of rainfall-runoff erosivity in 2001 and land use in 2001, and the actual

situation of 2010 can be defined as the combination of rainfall-runoff erosivity in 2010 and land use in 2010.

4. Results and Discussion

4.1. Land Use and Land Cover Change. Land use maps of the study area dated to 2001 and 2010 were derived from Landsat data, forest land protection, and utilization planning data from the Liangcheng County government. The land use and land cover results were obtained by combining visual interpretation of the standard false-color combination of Landsat data with field survey data from the local government forestry administration. Figure 4 illustrates land use and land cover in 2001 and 2010. Since this project focused on the effect of returning farmland to forest, only forest types were mapped in 2010. In the land use map of 2001, in addition to forest types, there were some farmlands that were later converted into forests. From 2001 to 2010, 8.17 square kilometres of cropland was converted to woodland, 280.42 square kilometres of cropland to shrubland and 5.72 square kilometres of cropland to sparse woodland.

Figure 5 shows the dynamic change of vegetation coverage in the study area, with the average coverage increasing from 51% to 63% between 2001 and 2010. As can be seen from the figure, the increase in vegetation coverage is obvious, especially in the northwest and southeast mountain areas of the study area.

4.2. Assessment of Factors. Figure 6 shows the spatial distribution of various USLE factors in the study area. As shown in Figures 6(a) and 6(b), there are significant differences in rainfall-runoff erosivity between 2001 and 2010. The statistical results show that the R value in 2010 is generally higher than that in 2001. The average R value in 2010 is 113.53 MJ mm ha^{-1}h^{-1} year^{-1}, compared with 98.62 MJ mm ha^{-1}h^{-1} year^{-1} in 2001. The R values between 120 and 170, 105 and 120, and 90 and 105 in 2010 are about 27%, 44%, and 29%, respectively, compared with 10%, 20%, and 40% in 2001.

Figures 6(c) and 6(d) show the K factor value and LS factor value. Because K factor and LS factor are relatively stable for a short period of time, it is assumed these values are the same for 2001 and 2010.

As depicted in Figure 6, the crop management factor (C) in 2010 changed significantly compared with that in 2001. The red and yellow areas in Figure 6(f), i.e., the region whose C value was lower than 0.72 in 2010, are much larger than those in 2001 in Figure 6(e). The C average of the study area decreased from 0.68 in 2001 to approximately 0.56 in 2010. This change should be attributed to a series of important ecological and water conservation projects, including the NFPP and GCP.

4.3. Soil Erosion Estimation in an Actual Situation. Figure 7 shows the spatial distribution of the soil erosion modulus in 2001 and 2010. According to the soil erosion grading standards established by China's Ministry of Water Resources, the soil erosion modulus of Liangcheng County

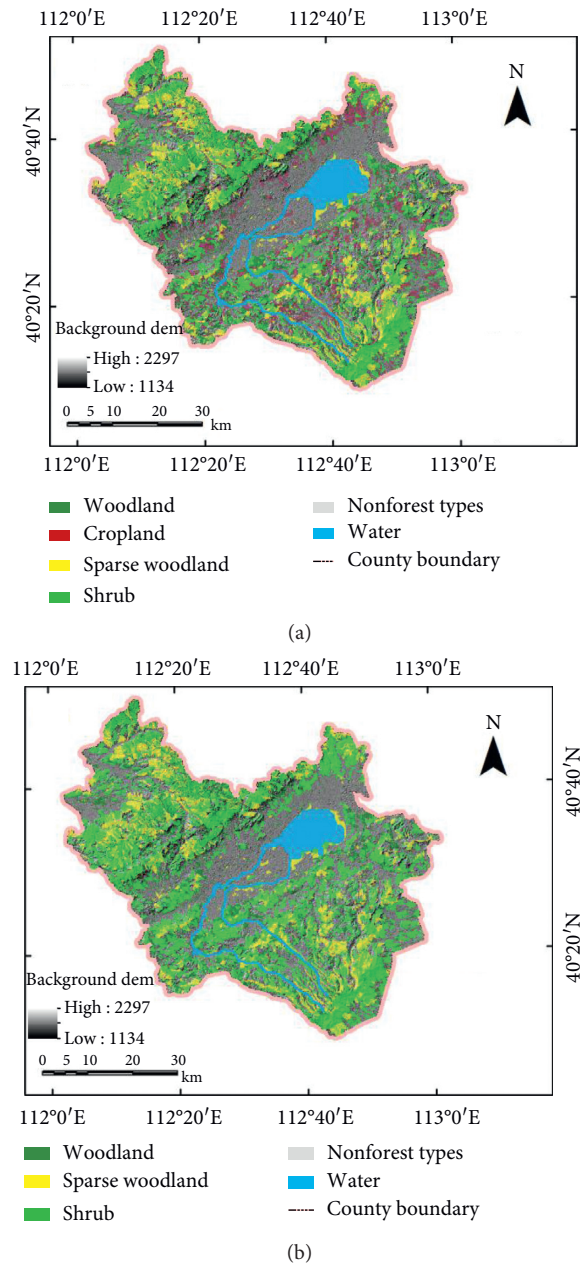


Figure 4: Land use map of the study area in 2001 (a) and 2010 (b).

in 2001 and 2010 may be divided into different grades. The mean soil erosion modulus was 14190 t/(km^2 a) in 2001 and 14012 t/(km^2 a) in 2010. Thus, compared with 2001, the mean soil erosion modulus in 2010 changed little.

4.4. Estimation of Soil Loss under Scenarios. The soil erosion modulus under different scenarios (S1 and S2) are calculated and presented in Figure 8. The mean soil erosion modulus under scenario 1 (S1) is 11906 t/(km^{2}a), while the mean soil erosion modulus under scenario 2 (S2) is 16270 t/(km^2 a).

For scenario 1 (S1), i.e., when the 2001 rainfall erosivity acts on the underlying surface of 2010, the simulated mean soil erosion modulus is much smaller than the actual soil

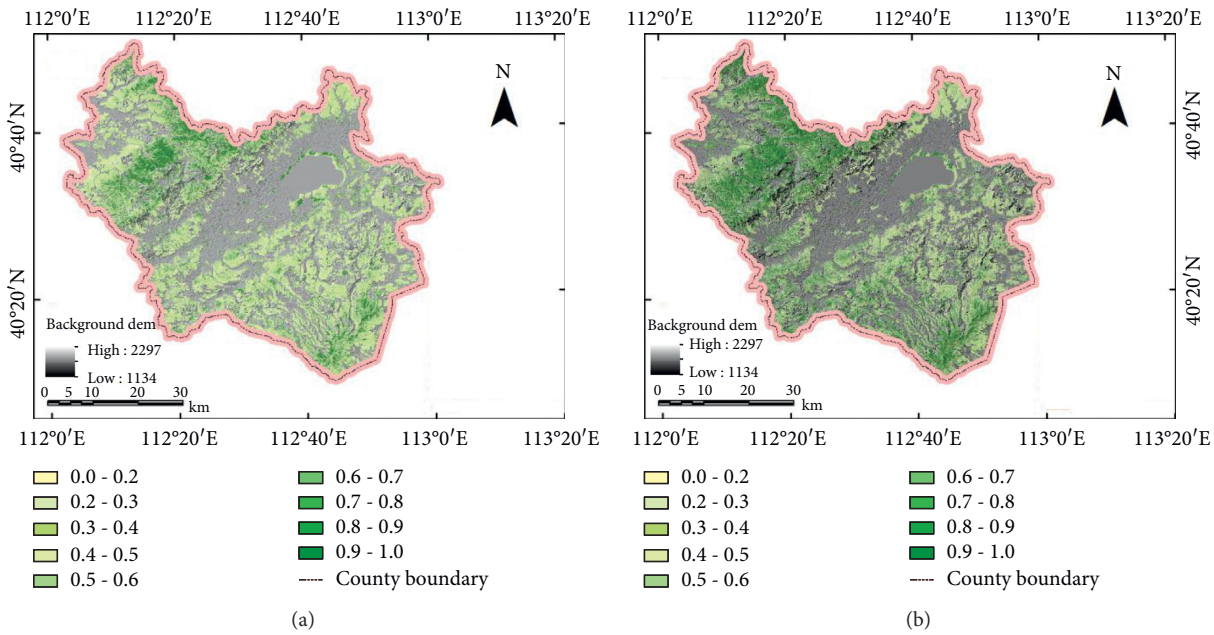


Figure 5: Vegetation coverage map in 2001 (a) and 2010 (b).

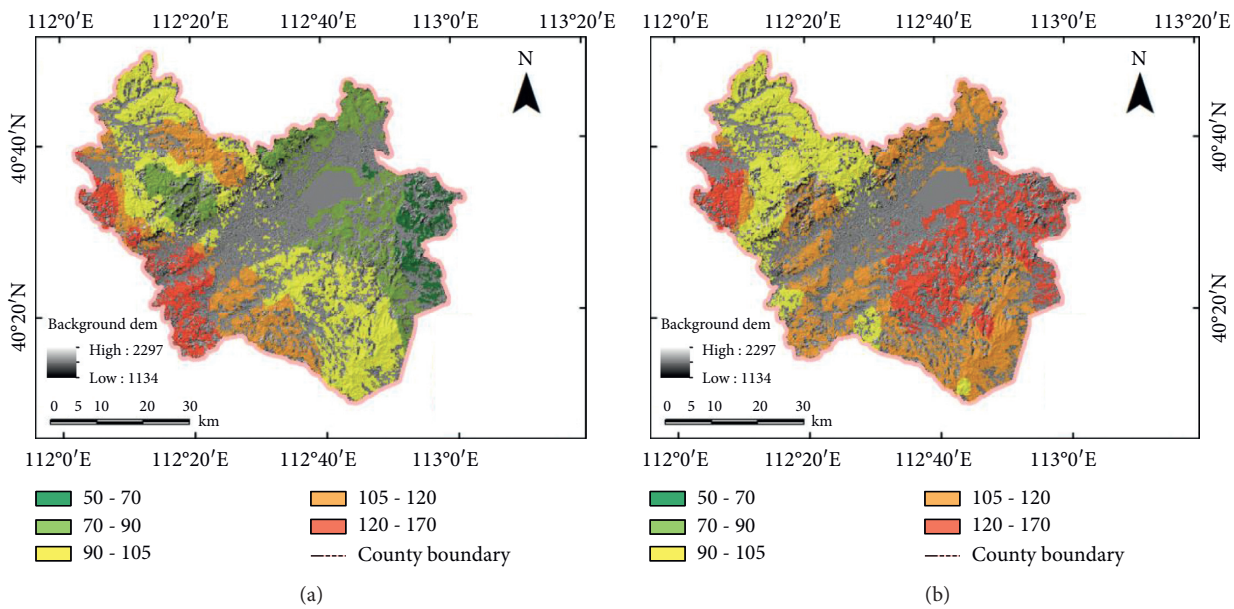


Figure 6: Continued.

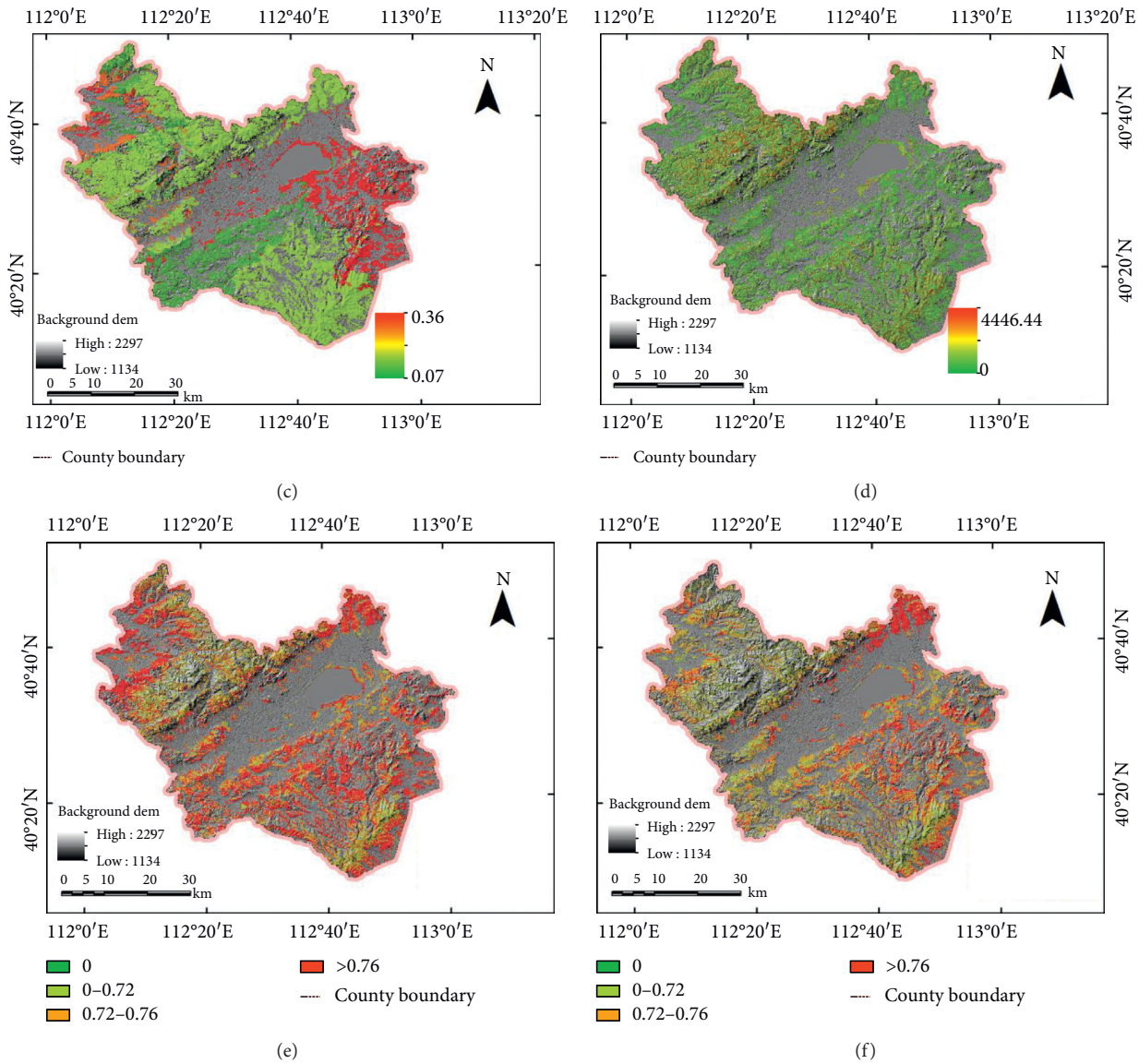


Figure 6: Factors superimposed on the elevation map of the study area. (a) R factor in 2001 (MJ mm ha^{-1}h^{-1} year^{-1}); (b) R factor in 2010 (MJ mm ha^{-1}h^{-1} year^{-1}); (c) K factor; (d) LS factor; (e) C factor in 2001; (f) C factor in 2010.

erosion modulus in 2001 and 2010 calculated above. The mean soil erosion modulus of S1 is approximately 16.10% smaller than the actual soil erosion modulus in 2001, indicating that the changes in land use and land cover between 2001 and 2010 effectively reduced soil erosion. The mean soil erosion modulus of S1 is approximately 15.67% smaller than the actual soil erosion modulus in 2010, indicating that the rainfall erosivity in 2010 is much higher than that in 2001.

The same conclusion can be obtained for the simulation of soil erosion under scenario 2 (S2). For scenario 2 (S2), i.e., when the 2010 rainfall erosivity acts on the underlying surface of 2001, the mean soil erosion modulus is as high as 16270 t/(km^2 a). Compared with the actual soil erosion modulus in 2001 and 2010, the soil erosion modulus under S2 is about 14.66% and 16.11% higher, respectively, indicating once again that the rainfall erosivity in 2010 is higher than that in 2001, and the land use and land cover change between 2001 and 2010 effectively reduced soil erosion.

The comparison presented in Table 1 illustrates the impact of rainfall erosivity and land use and land cover change more clearly. Compared with the actual situation in 2001, the percentage of area under S1 decreases by 2.87% and 4.06%, while that under S2 increases by 0.06% and 3.54%. Compared with the actual situation in 2010, the percentage of area under S1 decreases by 0.05% and 3.69%, while that under S2 increases by 2.78% and 3.91%.

Scenario construction allows us to understand that soil erosion in the region is significantly affected by rainfall erosivity in addition to land use and land cover change. This result is consistent with previous studies on the effects of rainfall erosivity on soil erosion [29, 45]. On the other hand, the conversion of cultivated land into forests will reduce soil

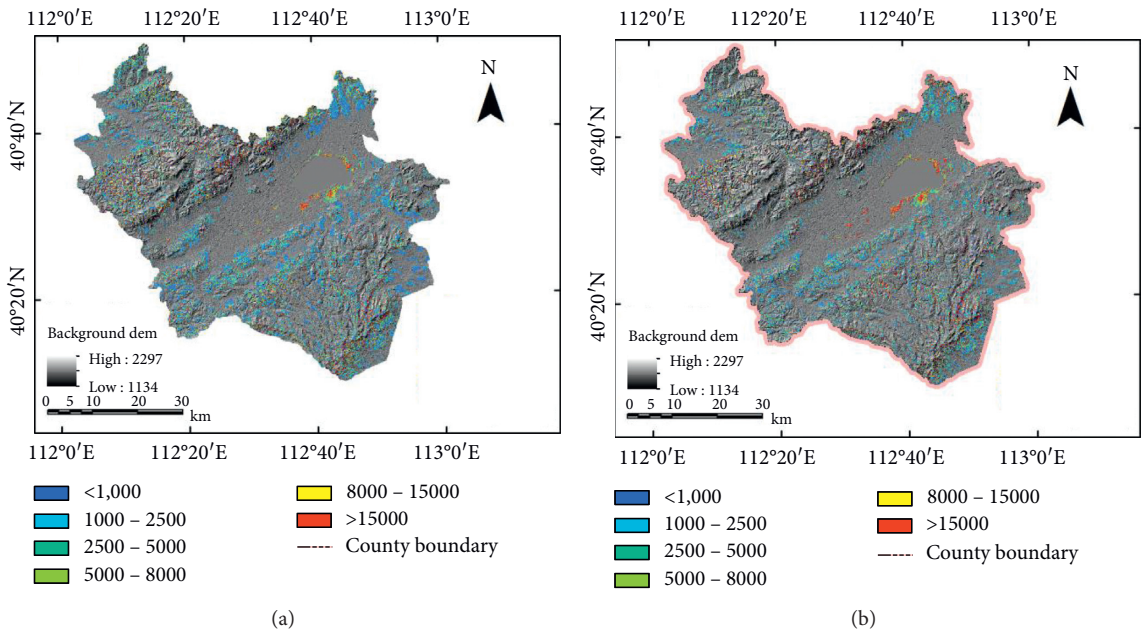


Figure 7: Soil erosion modulus map in 2001 (a) and 2010 (b).

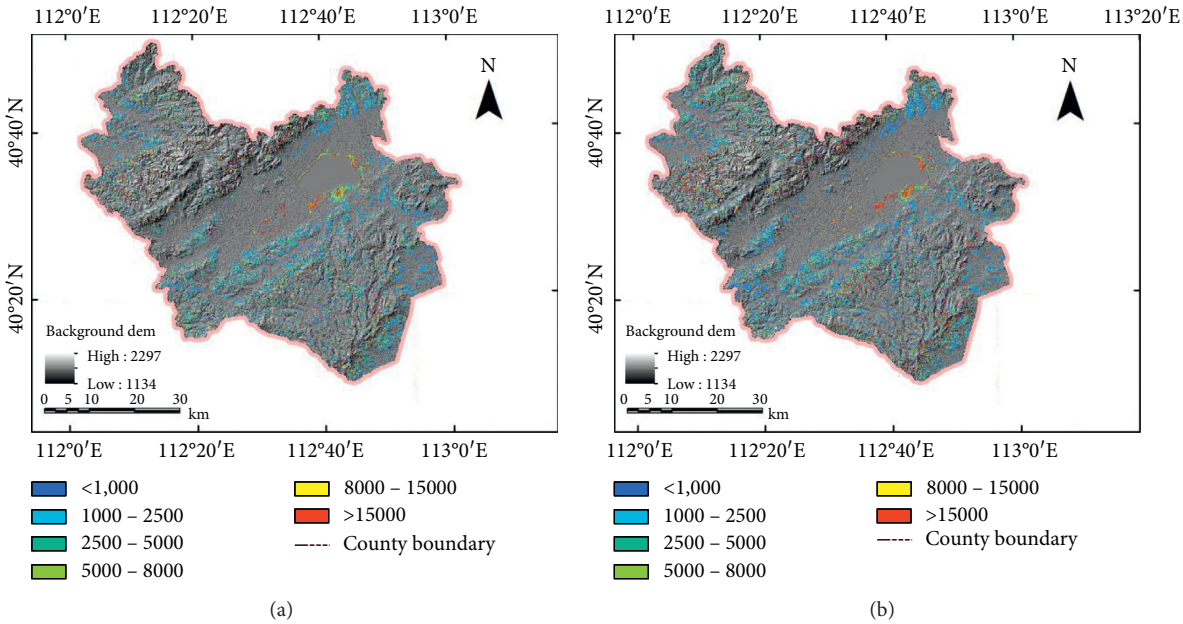


Figure 8: Soil erosion modulus under S1 (a) and S2 (b) superimposed on elevation map.

Table 1: Comparison of soil erosion intensity under different situations.

Soil erosion grades	2001 (%)	2010 (%)	S1 (%)	S2 (%)
Slight	8.48	22.16	21.61	8.6
Mild	14.43	9.64	11.93	12.69
Moderate	17.14	13.91	15.19	16.03
Strong	15.03	12.46	13.28	14.16
Intense	19.88	17.16	17.01	19.94
Severe	25.04	24.67	20.98	28.58

erosion. This is also confirmed by studies of other similar areas near the study area [46, 47], which shows that soil erosion in their study area is significantly reduced because of the large increase in forest land.

5. Conclusions

In this study, the influence of ecological restoration activities on soil erosion is assessed in the Inner Mongolia Autonomous Region, China. Due to the implementation of restoration activities, the forested area of the study region increased. The USLE model was applied to assess the soil erosion before and after the implementation of ecological projects. The scenario construction serves as a useful tool in investigating the causes of soil erosion. Actual soil erosion in the research area during the study period changed little. That is because rainfall erosivity increases soil erosion, while land use change reduces soil erosion. Measures to restore forests have significantly reduced soil erosion. The implementation of ecological projects, such as the Natural Forest Protection Project and Green for Grain Project, are constructive interventions. It is necessary to continue afforestation to offset the negative impacts of rainfall-runoff change. In the future, studying the impact of different land cover types on soil erosion will help determine the most appropriate vegetation to reduce soil erosion and maximize the benefits of environmental recovery efforts.

Data Availability

The materials used in this study are as follows: (1) monthly precipitation data of 20 years (1992–2011) from the China Meteorological Administration (CMA), used for calculating R factor in 2001 and 2010; (2) ASTER GDEM data with a resolution of 30 meters downloaded from http://reverb.echo.nasa.gov, used to calculate slope gradient and slope length; (3) two Landsat5/TM scenes with a resolution of 30 meters dated August 20, 2001, and August 10, 2010, that were obtained from the USGS Glovis data archive, and terrain-corrected Level 1T scenes with geodetic accuracies of one-quarter to less than half a pixel (Figure 2), used to extract land use/cover and vegetation information; (4) available $1:1{,}000{,}000$ soil map from the Resource and Environmental Science Data Center of the Chinese Academy of Sciences, used to calculate soil erodibility factor; (5) woodland survey data in 2001 and 2010 from China's Liangcheng County government, auxiliary data used to extract land use/cover information.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the China Major Special Project (grant no. 30-Y20A37-9003-15/17) and the National Natural Science Foundation of China (grant no. 41807170).

References

[1] Z. X. Zhang, "Situation of ecological environment and deteriorative causes in Inner Mongolia," *Inner Mongolia Environmental Protection*, vol. 12, no. 2, pp. 30–36, 2000.

[2] Y. Liu, J. Cao, J. Yao, H. Liang, and C. Zhao, "Present situation of water and soil loss and its countermeasures of management in Inner Mongolia," *Inner Mongolia Forestry Science and Technology*, vol. 1, pp. 39–45, 2002.

[3] W. Yang, Y. Bai, Z. Yang et al., "Study on technology of forest ecological network construction in ecological fragile zone in Inner Mongolia," *Journal of Arid Land Resources and Environment*, vol. 17, no. 5, pp. 86–90, 2003.

[4] Li Xu and Q. Zhang, "Summary of 30 years of soil and water conservation in Inner Mongolia autonomous region," *Soil and Water Conservation in China*, vol. 11, pp. 8–10, 2008.

[5] Z. Zhuang, "A great ecological project: China natural forest conservation programme," *World Forestry Research*, vol. 14, no. 3, pp. 47–54, 2001.

[6] G. X. Gao, G. L. Zhang, M. X. Liu, L. I. Wen-Zhong, G. Q. Liu, and H. U. Wen-Zhong, "Retrospect on the research and practice of the converting cropland to forest," *Journal of Northwest Forestry University*, vol. 22, no. 2, pp. 203–208, 2007.

[7] W. H. Wischmeier and D. D. Smith, *Predicting Rainfall Erosion Losses: A Guide to Conservation Planning*, Science, US Department of Agriculture Handbook, No. 537, Washington, DC, USA, 1978.

[8] K. G. Renard, G. R. Foster, G. A. Weesies, D. K. McCool, and D. C. Yoder, *Predicting Soil Erosion by Water: a Guide to Conservation Planning with the Revised Universal Soil Loss Equation (RUSLE)*, Agriculture Handbook N.703, U.S. Department of Agriculture Research Service, Washington, DC, USA, 1997.

[9] J. M. Van der Knijff, R. J. A. Jones, and L. Montanarella, *Soil Erosion Risk Assessment in Europe*, Office for Official Publications of the European Communities (EUR 19044 EN), Luxembourg, 2000.

[10] E. H. Erdogan, G. Erpul, and İ. Bayramin, "Use of USLE/GIS methodology for predicting soil loss in a semiarid agricultural watershed," *Environmental Monitoring and Assessment*, vol. 131, no. 1–3, pp. 153–161, 2007.

[11] J. Lu, X. Chen, Li Hui, H. Liu, J. Xiao, and J. Yin, "Soil erosion changes based on GIS/RS and USLE in Poyang Lake basin," *Transactions of the CSAE*, vol. 27, no. 2, pp. 337–344, 2011.

[12] M. Zhu, "Soil erosion assessment using USLE in the GIS environment: a case study in the Danjiangkou reservoir region, China," *Environmental Earth Sciences*, vol. 73, no. 12, pp. 7899–7908, 2015.

[13] H. Geng, B. Pan, C. Wang, and B. Huang, "Soil erosion of yuzhong county based on GIS and RS," *Journal of Lanzhou University*, vol. 45, no. 6, pp. 8–13, 2009.

[14] A. Adediji, A. M. Tukur, and K. A. Adepoju, "Assessment of revised universal soil loss equation (RUSLE) in katsina area, katsina state of Nigeria using remote sensing (RS) and geographic information system (GIS)," *Iranica Journal of Energy and Environment*, vol. 11, no. 3, pp. 255–264, 2010.

[15] X. Wang, B. Guo, and L. Jiang, "Research progress of watershed soil erosion based on USLE, GIS and RS," *Subtropical Soil and Water Conservation*, vol. 24, no. 1, pp. 42–47, 2012.

[16] S. A. Ali and H. Hagos, "Estimation of soil erosion using USLE and GIS in Awassa catchment, Rift valley, Central Ethiopia," *Geoderma Regional*, vol. 7, no. 2, pp. 159–166, 2016.

[17] A. U. Ozcan, O. Uzun, M. Baaran, G. Erpul, S. Aksit, and H. M. Palancioglu, "Soil erosion risk assessment for volcano cone of Alidagi mountain by using USLE/RUSLE, GIS and geostatistics," *Fresenius Environmental Bulletin*, vol. 24, no. 6, pp. 2090–2100, 2015.

[18] T. G. Pham, J. Degener, and M. Kappas, "Integrated universal soil loss equation (USLE) and geographical information system (GIS) for soil erosion estimation in a Sap basin: central Vietnam," *International Soil and Water Conservation Research*, vol. 6, no. 2, pp. 99–110, 2018.

[19] G. Mancino, A. Nolè, L. Salvati, and A. Ferrara, "In-between forest expansion and cropland decline: a revised USLE model for soil erosion risk under land-use change in a Mediterranean region," *Ecological Indicators*, vol. 71, pp. 544–550, 2016.

[20] V. Ferreira, A. Samora-Arvela, and T. Panagopoulos, "Soil erosion vulnerability under scenarios of climate land-use changes after the development of a large reservoir in a semi-arid area," *Journal of Environmental Planning and Management*, vol. 59, no. 7, pp. 1238–1256, 2016.

[21] C. A. Aguirre-Salado, L. Miranda-Aragón, M. Pompa-García et al., "Improving identification of areas for ecological restoration for conservation by integrating USLE and MCDA in a gis-environment: a pilot study in a priority region northern Mexico," *ISPRS International Journal of Geo-Information*, vol. 6, no. 9, 2017.

[22] W. Buytaert, V. Iñiguez, and B. D. Bièvre, "The effects of afforestation and cultivation on water yield in the Andean páramo," *Forest Ecology and Management*, vol. 251, no. 1-2, pp. 22–30, 2007.

[23] A. U. Ozcan, G. Erpul, M. Basaran, and H. E. Erdogan, "Use of USLE/GIS technology integrated with geostatistics to assess soil erosion risk in different land uses of Indagi Mountain Pass-Çankırı, Turkey," *Environmental Geology*, vol. 53, no. 8, pp. 1731–1741, 2008.

[24] A. N. Nunes, A. C. de Almeida, and C. O. A. Coelho, "Impacts of land use and cover type on runoff and soil erosion in a marginal area of Portugal," *Applied Geography*, vol. 31, no. 2, pp. 687–699, 2011.

[25] Y. Liu, B. Fu, Y. Lü, Z. Wang, and G. Gao, "Hydrological responses and soil erosion potential of abandoned cropland in the Loess Plateau, China," *Geomorphology*, vol. 138, no. 1, pp. 404–414, 2012.

[26] P. Panagos, P. Borrelli, J. Poesen et al., "The new assessment of soil loss by water erosion in Europe," *Environmental Science & Policy*, vol. 54, pp. 438–447, 2015a.

[27] D. F. Scott, D. C. Le Maitre, and D. H. K. Fairbanks, "Forestry and streamflow reductions in South Africa: a reference system for assessing extent and distribution," *Water SA*, vol. 24, no. 3, pp. 187–199, 1998.

[28] X. Wang, Y. Bu, Y. Li et al., "Geochemical characteristics of stream sediment survey and ore prospecting prediction in Liangcheng area, Inner Mongolia," *Science & Technology Information*, vol. 21, pp. 426-427, 2012.

[29] M. Zare, A. A. Nazari Samani, M. Mohammady, T. Teimurian, and J. Bazrafshan, "Simulation of soil erosion under the influence of climate change scenarios," *Environmental Earth Sciences*, vol. 75, no. 21, pp. 1–15, 2016.

[30] H. Chen, "Analysis on the characteristics of climate change in Liangcheng county in recent 30 years," *Science & Technology Information*, vol. 18, pp. 88–90, 2019.

[31] F. Jiang, "Conservation effect and measures of natural forest resources in Liangcheng county," *Inner Mongolia Forestry*, vol. 10, pp. 12-13, 2015.

[32] H. M. J. Arnoldus, "An approximation of the rainfall factor in the universal soil loss équation," in *Assessment of Erosion*, M. De Boodt and D. Gabriels, Eds., pp. 127–132, John Wiley & Sons, New York, NY, USA, 1980.

[33] Y. Hong, K. L. Hsu, S. Sorooshian, and X. G. Gao, "Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system," *Journal of Applied Meteorology*, vol. 43, no. 12, pp. pp1834–1853, 2004.

[34] Z. Gu, P. Shi, and C. Jin, "Precipitation interpolation research over regions with sparse meteorological stations:a case study in xilin gole league," *Journal of Beijing Normal University(Natural Science)*, vol. 42, no. 2, pp. 204–208, 2006.

[35] J. R. Williams and E. P. Renard, "A new method for assessing erosions effect on soil productivity," *Journal of Soil and Water Conservation*, vol. 38, no. 1, pp. 381–383, 1983.

[36] B. Y. Liu, M. A. Nearing, and L. M. Risse, "Slope gradient effects on soil loss for steep slopes," *Transactions of the ASAE*, vol. 37, pp. 1835–1840, 1994.

[37] B. Y. Liu, M. A. Nearing, P. J. Shi, and Z. W. Jia, "Slope length effects on soil loss for steep slopes," *Soil Science Society of America Journal*, vol. 64, no. 5, pp. 1759–1763, 2000.

[38] D. K. McCool, G. R. Foster, C. K. Mutchler et al., "Revise slope length factor for the universal soil loss equation," *Transactions of ASAE*, vol. 32, pp. 1571–1576, 1989.

[39] B. Liu, Y. Xie, and K. Zhang, *Soil Erosion Prediction Model*, China Science and Technology Press, Beijing, China, 2001.

[40] B. Xu, H. Lu, and D. Shao, "Study on regional soil erosion based on remote sensing and GIS technology," *Journal of Irrigation and Drainage*, vol. 33, no. Z1, pp. 291–294, 2014.

[41] B. Mahalingam, M. M. Malik, and M. Vinay, "Assessment of soil erosion using USLE technique: a case study of Mysore District, assessment of soil erosion using USLE technique: a case study of Mysore District, Karnataka, India, (November 2015)," 2016.

[42] C. Cai, S. Ding, Z. Shi et al., "Study of applying USLE and geographical information system IDRISI to predict soil erosion in small watershed," *Journal of Soil Water Conservation*, vol. 14, no. 2, pp. 19–24, 2000.

[43] F. Li, W. Chen, Y. Zeng, Q. Zhao, and B. Wu, "Improving estimates of grassland fractional vegetation cover based on a pixel dichotomy model: a case study in Inner Mongolia, China," *Remote Sensing*, vol. 6, no. 6, pp. 4705–4722, 2014.

[44] D. J. Liu, L. I. Run-Jie, W. Q. Wang, and G. J. Wei, "Completion of Xining city soil erosion monitoring based on GIS," *Research of Soil and Water Conservation*, vol. 13, no. 5, pp. 111–114, 2006.

[45] Z. Zhao, J. Wang, X. Wu et al., "Soil erosion pattern and change in Hinggan League, Inner Mongolia from 1990 to 2005 based on RUSLE," *Journal of Arid Land Resources and Environment*, vol. 28, no. 6, pp. pp124–129, 2014.

[46] X. M. Cui, *The Research on the Soil Erosion of Zhungeer County in Loess Hilly Region of Inner Mongolia*, Inner Mongolia Agricultural University, Hohhot, China, 2012.

[47] J. N. Zhou, *Spatio-temporal Coupling Relationship between Soil Erosion and LUCC in Duolun County, Inner Mongolia*, Inner Mongolia Agricultural University, Hohhot, China, 2017.

Hindawi
Scientific Programming
Volume 2021, Article ID 8841133, 15 pages
https://doi.org/10.1155/2021/8841133

Research Article

Collaborative Filtering Recommendation Using Nonnegative Matrix Factorization in GPU-Accelerated Spark Platform

Bing Tang,[1] **Linyao Kang**,[1] **Li Zhang**,[1] **Feiyan Guo**,[1] **and Haiwu He**[2]

[1]*School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China*
[2]*Shandong Computer Science Center (National Supercomputer Center in Jinan), Jinan 250014, China*

Correspondence should be addressed to Bing Tang; btang@hnust.edu.cn

Received 1 October 2020; Revised 16 December 2020; Accepted 21 December 2020; Published 5 January 2021

Academic Editor: Shah Nazir

Nonnegative matrix factorization (NMF) has been introduced as an efficient way to reduce the complexity of data compression and its capability of extracting highly interpretable parts from data sets, and it has also been applied to various fields, such as recommendations, image analysis, and text clustering. However, as the size of the matrix increases, the processing speed of nonnegative matrix factorization is very slow. To solve this problem, this paper proposes a parallel algorithm based on GPU for NMF in Spark platform, which makes full use of the advantages of in-memory computation mode and GPU acceleration. The new GPU-accelerated NMF on Spark platform is evaluated in a 4-node Spark heterogeneous cluster using Google Compute Engine by configuring each node a NVIDIA K80 CUDA device, and experimental results indicate that it is competitive in terms of computational time against the existing solutions on a variety of matrix orders. Furthermore, a GPU-accelerated NMF-based parallel collaborative filtering (CF) algorithm is also proposed, utilizing the advantages of data dimensionality reduction and feature extraction of NMF, as well as the multicore parallel computing mode of CUDA. Using real MovieLens data sets, experimental results have shown that the parallelization of NMF-based collaborative filtering on Spark platform effectively outperforms traditional user-based and item-based CF with a higher processing speed and higher recommendation accuracy.

1. Introduction

In recent years, the scale of data has grown exponentially. Globally, it is becoming a trend to research and develop big data technology, use big data to promote economic development, improve social governance, and improve government services and regulatory capabilities. How to effectively extract knowledge from big data, understand and analyze it, and finally make predictions are current popular research topics.

As an important mathematical tool for big data processing, nonnegative matrix factorization is a matrix decomposition approach that decomposes a nonnegative matrix into two low-rank matrices constrained to have nonnegative elements [1, 2]. This results in a reduced representation of the original data that can be seen either as a feature extraction or as a dimensionality reduction technique. The widespread usage of the NMF is due to its ability of providing new insights and relevant information about the complex latent relationships in experimental data sets. Since Lee and Seung's Nature paper [1], NMF has been extensively studied and has a great deal of applications in science and engineering. It has become an important mathematical method in machine learning and data mining and has been widely used in feature extraction, image analysis [3], audio processing [4], recommendation systems [5, 6], pattern recognition, data clustering [7], topic modeling [8], text mining [9], bioinformatics [10], and so forth. Unlike other factorization methods (e.g., PCA, ICA, SVD, VQ, etc.), NMF can be interpreted as a parts-based representation of the data because only additive combinations are allowed. In contrast to PCA and ICA, NMF strictly requires that the entries of both resulting matrices be nonnegative. Such a constraint is very meaningful in many applications, in which the data representation is purely additive; for instance, the user ratings of e-commerce websites are usually nonnegative values, and image pixels are nonnegative values.

The main problem of NMF is that the original matrix is usually high-order matrix, which makes the computational

complexity very high. Therefore, the parallel algorithm of NMF gradually attracts more attention, and some parallel NMF algorithms have been proposed. Although the parallelization of NMF can improve the computational efficiency to a certain extent, parallel algorithms should be matched to the machine hardware architecture and should have strong scalability, that is, the ability to effectively utilize increased processor resources.

Accelerating HPC applications is currently under extensive research using new hardware technologies such as the recent Central Processing Units (CPUs) that are getting multiple processor cores for parallel computing, Graphics Processing Units (GPUs) that process huge data blocks in parallel, and hybrid CPUs/GPUs computing which is a very common solution for HPC. GPUs are getting more attention than other HPC accelerators due to their high computation power, strong performance, functionality, and low price. The modern GPU is not only a powerful graphic engine but also a highly parallel programmable processor featuring peak arithmetic and memory bandwidth [11]. They are now used to accelerate graphics and some general applications with high data parallelism (GPGPU) due to the availability of Application Programming Interfaces (APIs), such as Compute Unified Device Architecture (CUDA) and Open Computing Language (OpenCL).

Spark is a distributed in-memory computation framework that was proposed by AMPLab of University of California, Berkeley, in 2009 and is based on a framework of processing large amounts of data in memory [12, 13]. It supports four programming languages, Scala, Java, Python, and R. Resilient Distributed Datasets (RDD) is a new concept proposed by Spark for data collections. RDD can support coarse-grained write operations [14]. Spark caches a particular RDD into memory, and the next operation can read directly from memory. The data is not written to disk, saving a lot of disk I/O overhead. Experimental performance evaluation confirmed that Spark's performance has increased by dozens or even 100 times compared to Hadoop, which relies on MapReduce model [15, 16] and data being stored in a distributed file system called HDFS rather than in memory.

Currently, some parallel approaches for nonnegative matrix factorization have been proposed, for example, high-performance approaches using message passing interface (MPI) [17], GPU-accelerated approaches [9, 18], and Hadoop-based MapReduce approaches [10, 19]. These approaches mainly utilize the multicore characteristics of the system, and there is still the potential to improve performance by utilizing memory, CPU, and GPU resources together.

Meanwhile, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences from group users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person. Through calculating the data similarity between two users, we can get the similarity between two users. Although traditional collaborative filtering is extremely successful in the recommendation system, as the data increases, the recommendation algorithms have been confronted with various problems, such as scalability problems, cold start problems, and matrix sparseness problems.

In order to solve the above problems, this paper proposes a combination of Spark-based and GPU-based acceleration model to develop scalable NMF parallel algorithm, which takes advantages of both GPU and in-memory computing, to obtain a highly scalable parallel NMF algorithm. Furthermore, this paper proposes a collaborative filtering method based on NMF and is parallelized and migrated to the Spark platform equipped with GPU to execute, which effectively improves the calculation efficiency, and thus can update the recommendation results faster and produce more accurate recommendation results. The experimental results show that the parallelization of NMF-based collaborative filtering effectively improves the calculation efficiency and accuracy.

The rest of the paper is organized as follows. Section 2 surveys related work. Section 3 introduces the mathematical fundamental of NMF and describes the general parallel principle of NMF. Section 4 describes the architecture of GPU-accelerated Spark platform and presents GPU-accelerated NMF on Spark. Section 5 introduces the collaborative filtering algorithm based on NMF. Section 6 presents performance evaluation results of GPU-accelerated NMF on Spark and collaborative filtering-based NMF, which is followed by conclusions in Section 7.

2. Related Work

2.1. Hybrid Big Data Processing Model. Due to the diversity of hardware equipment in high-performance computing system, in order to deal with the real-world complex applications, the mix of different computing modes has become a major direction, such as hybrid CPU/GPU and CPU/FPGA. From the model point of view, there are hybrid MapReduce/CUDA model [20], hybrid MapReduce/MIC model [21], hybrid OpenMP/MPI model [22], and so forth. In recent years, the practical experience of academia and industry has shown that computing platforms based on heterogeneous CPU/GPU system have great development potential and have attracted more and more attentions [11, 23].

2.2. Parallel Nonnegative Matrix Factorization. To handle large data sets through nonnegative matrix factorization, there are three main directions. The first class of algorithms is called online NMF algorithms [6, 24, 25], which are the oldest approach for dealing with high-dimensional data processing through NMF. The second class is known as distributed NMF algorithms, which distribute data over a network so that several small-scale data can be performed concurrently. The third class of algorithms is called compressed NMF algorithms [26, 27], which perform structured random compression to project the data onto the lower-dimensional manifolds. In this paper, we only focus on distributed NMF.

Nonnegative matrix factorization is usually solved through alternate iteration [2], which makes it suitable for parallelization. Three aspects that restrict the scalability of the parallel NMF algorithm are listed as follows: synchronization between computing processes, data loading and data transmission, and parallel granularity division. At present, some parallelization algorithms have been proposed to accelerate nonnegative matrix factorization.

Janecek et al. used linear algebra toolkits such as BLAS, LAPACK, and ARPACK to implement multithreaded programs on a single computer to perform efficient NMF [28]. Lopes and Ribeiro implemented a GPU-based machine learning library named GPUMLib, which contains an implementation of GPU-accelerated NMF [29]. Kysenko et al. also applied GPU-accelerated NMF to text mining [9]. Battenberg and Wessel implemented a parallel NMF, using the characteristics of a shared memory multicore system based on OpenMP and a many-core GPU based on CUDA technology, and applied it to audio signal processing, but it can only work for a single node [30]. A parallel NMF based on the combination of MPI and GPU is implemented in [18], and it is used for biological sequence comparison. Tang et al. proposed a hybrid parallel hierarchical NMF algorithm based on OpenMP and MPI [31].

Using some high-performance computing software packages, such as ParMETIS, ScaLAPACK, and HPSEPS, we can develop nonnegative matrix factorization parallel algorithms based on MPI/OpenMP/GPU using these software packages, which are ultimately not suitable for practical big data processing in Internet era. On the basis of open-source big data processing framework such as Hadoop and Spark, it is a more suitable idea to develop a parallel algorithm of NMF to make it suitable for Internet big data processing. In [10], Liao et al. realized the distributed NMF based on MapReduce for biological data processing. Sun et al. realized large-scale NMF based on MapReduce in [32], and Liu et al. also proposed a distributed NMF based on MapReduce for processing large-scale web data using Hadoop streaming method [19]. In our previous work [33], we proposed a parallel NMF algorithm in Spark platform, which makes full use of the advantages of in-memory computation mode.

2.3. Parallel and Distributed Collaborative Filtering. Many e-commerce companies have already incorporated recommendation systems with their services, for example, product recommendations by Amazon (http://www.amazon.com) and Taobao (http://www.taobao.com) and movie recommendations by Netflix (http://www.netflix.com). The implementations and algorithms of collaborative filtering for the applications of recommendation systems face several challenges. First is the size of processed datasets. The second one comes from the sparseness of rating matrix, which means for each user only a relatively small number of items are rated. With the increase of a large amount of data and the complexity of the data, it is confronted with the problem of low efficiency. Thus, highly efficient collaborative filtering algorithm is needed.

Recently, great interest has turned towards parallel and distributed implementations of collaborative filtering algorithms. They can be classified into several categories: (1) distributed memory implementations, such as the MPI-based parallel implementations proposed in [34, 35]; (2) shared memory implementations, such as MPI implementations in [36, 37]; (3) GPU-based implementations, such as [38, 39]; (4) platform-based implementations, such as the Hadoop platform-based collaborative filtering implemented by [40, 41]; and (5) heterogeneous implementation, such as the hybrid OpenMP + MPI implementation proposed by Narang et al. [42] and Karydi and Margaritis [43].

On the other hand, these challenges of collaborative filtering have been well taken care of by matrix factorization (MF). Matrix factorization methods have recently received greater exposure as unsupervised learning methods for latent variable decomposition and dimensionality reduction [44]. It is a powerful technique to find the hidden structure behind the data. There are several matrix factorization models that could be used for collaborative filtering recommendations, for example, Singular Value Decomposition (SVD) [45, 46], Principal Component Analysis (PCA) [47], Probabilistic Matrix Factorization (PMF) [48], and nonnegative matrix factorization [49]. However, efficient collaborative filtering algorithm also places demands on fast matrix factorization.

To sum up, high-dimensional NMF is time-consuming, and there is an urgent need for high-performance parallelization solutions. At present, there is no flexible distributed processing framework that considers both the memory computing mode and GPU technologies for NMF at the same time. Considering Spark distributed processing framework and combining the powerful computing advantages of GPU and large-capacity memory, large-scale NMF parallel algorithm would enable the algorithm to be easily adapted to the processing of Internet big data. To the best of our knowledge, it is the first NMF-based collaborative filtering implementation that is parallelized and migrated to the Spark platform equipped with GPU. Experimental results validate that the parallelization of NMF-based collaborative filtering on Spark platform effectively improves the calculation efficiency and accuracy.

3. Parallel Nonnegative Matrix Factorization

3.1. Nonnegative Matrix Factorization. Nonnegative matrix factorization seeks to approximate a nonnegative $n \times m$ matrix V (in this context, a matrix is called nonnegative if all of its elements are nonnegative) by a product $V \approx WH$ of nonnegative matrices W and H of dimensions $n \times r$ and $r \times m$, respectively, with a given and typically low maximal rank r. Usually, r is chosen to satisfy $r = \min\{m, n\}$ such that WH can be thought of as a compressed form of the original data. It forms the basis of unsupervised learning and data reduction algorithms with applications to image recognition, speech recognition, data mining and collaborative filtering, and so forth.

NMF is able to represent a large input data set as the linear combination of a reduced collection of elements named *factors*. In this way, W contains the reduced set of r factors, and H stores the coefficient of the linear combination of such factors which rebuilds V. NMF iteratively modifies W and H until their product approximates to V. Such modifications, composed by matrix products and other algebraic operations, are derived from minimizing a cost function that describes the distance between WH and V. Lee and Seung presented two NMF algorithms based on multiplicative update rules whose objective functions are *Square of Euclidean Distance* (SED) and *Generalized Kullback-Leibler Divergence* (GKLD), respectively [1, 2]:

$$E(V\|WtH) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(v_{ij} - \sum_{k=1}^{r} w_{ik}h_{kj}\right)^2, \quad (1)$$

$$D(V\|WtH) = \sum_{i,j}\left(v_{ij}\log\frac{v_{ij}}{(WH)_{ij}} - v_{ij} + (WH)_{ij}\right). \quad (2)$$

Then, the objective of NMF is converted to optimize the following: $\min_{W,H} E(V\|WtH)$ or $\min_{W,H} D(V\|WtH)$, and s.t. $W, H \geq 0$, $\sum_{i=1}^{n} w_{ij} = 1, 1 \leq j \leq r$.

In this paper, we define SED as the objective function, so we have $\min(\|V - WH\|_F^2)$, which leads to the updating rules for matrices H and W:

$$h_{ij} = h_{ij}\frac{\left(W^T V\right)_{ij}}{\left(W^T WH\right)_{ij}}, \quad (3)$$

$$w_{ij} = w_{ij}\frac{\left(VH^T\right)_{ij}}{\left(WHH^T\right)_{ij}}. \quad (4)$$

3.2. Parallel Nonnegative Matrix Factorization. Before describing our experimental study, we briefly introduce the main existing parallel techniques of NMF. By analyzing equations (1) and (2), we can get the basic principle of iteration calculation of NMF in parallel manner. Matrix operations are performed in blocks. The block-based parallel updating rules for matrices H and W over multiple processes are shown in Figure 1, and the size of b_m can be adjusted according to the hardware configurations. At the time of initialization, initial W and H are produced. Because SVD-based initialization has been proven to be more effective for iteration of H and W [50], we generate initial W and H by the method of SVD. As you see, the size of matrix W is $n \times r$, the size of the matrix block V_j is $n \times b_m$, and the size of the matrix block H_j is $r \times b_m$, and finally the updated matrix block H_j is obtained. As shown in Figure 1(b), the new matrix H is used to compute the new matrix block W_i and so on. Matrix H and W are updated alternatively.

It can be seen from the analysis that the original matrix V is equivalent to a read-only variable, which is shared among all processes. With the iteration, matrices W and H need to be synchronized among all processes. The algorithm works by iteratively all-gathering the entire matrix H or W to each processor and then performing the Local Update Computations to update W_i or H_j.

4. GPU-Accelerated NMF on Spark

4.1. Spark. Conceptually, Apache Spark is an open-source in-memory data analytics cluster computing framework [12, 13]. As a MapReduce-like cluster computing engine, Spark also possesses good characteristics such as scalability and fault tolerance as MapReduce does. The main abstraction of Spark is RDDs, which make Spark be well qualified to process iterative jobs, including PageRank algorithm and K-means algorithm. RDDs are unique to Spark and thus differentiate Spark from conventional MapReduce engines. In addition, on the basis of RDDs, applications on Spark can keep data in memory across queries and reconstruct automatically data lost during failures. RDD is a read-only data collection, which can be either a file stored in an external storage system, such as HDFS, or a derived data set generated by other RDDs. RDDs store much information, such as its partitions, and a set of dependencies on parent RDDs called lineage. With the help of the lineage, Spark recovers the lost data quickly and effectively. Spark shows great performance in processing iterative computation because it can reuse intermediate results and keep data in memory across multiple parallel operations.

4.2. GPU-Accelerated Spark Platform. Modern GPUs are now capable of general computing. Due to the popularity of the CUDA on Nvidia GPUs, which can be considered as a C/C++ extension, we will mostly follow CUDA terminologies to introduce GPU computing. Current generations of GPUs are used as accelerators of CPUs and data are transferred between CPUs and GPUs through PCI-E buses. NVIDIA GPU programming is generally supported by the NVIDIA CUDA environment. A program on the host (CPU) can call a GPU to execute CUDA functions called kernel.

GPU is a multicore processor designed to parallelizable computational intensive tasks. It has very high computational processing power and data throughput. In scientific research and practical applications, the parallelizable computing task modules with less logical processing in the system are often transplanted to the GPU for execution, and a large execution performance improvement can usually be achieved.

However, Spark cluster will slow down when processing extremely large-scale data sets, especially when the node number is not very high. At the same time, more and more developers use GPUs for parallel computing to obtain high throughput and performance. Combining Spark with GPU, the mixed architecture is quickly becoming an emerging technology, which embeds the GPU into Spark, implements CPU/GPU integration, and builds an efficient heterogeneous parallel system.

In the CPU/GPU heterogeneous parallel cluster, the CUDA-based GPU acceleration technology is used, and the Spark computing tasks are accelerated by GPU. The basic

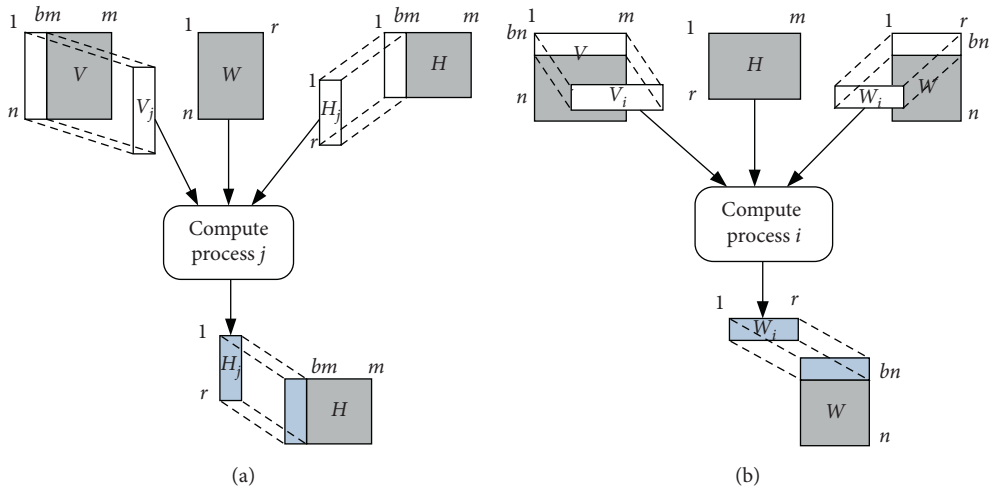


Figure 1: Block-based parallel updating rules for matrices H and W over multiple processes. (a) Iteration of matrix H and (b) iteration of matrix W.

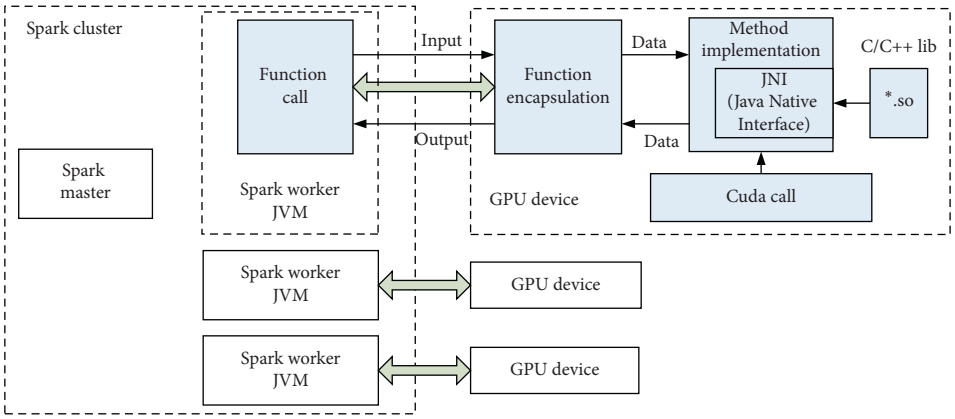


Figure 2: Architecture of GPU-accelerated Spark platform.

idea is that a part of operations of the Spark RDD are transferred to the GPU cores. GPU code execution flows are as follows: (1) copy data from main memory to GPU global memory; (2) GPU is driven by CPU instructions; (3) GPU parallel processing is in each core; and (4) GPU returns results to main memory. According to this idea and combined with Spark workflow, the GPU code is encapsulated, and then the data is transmitted between Spark Worker and GPU. The basic principle of Spark-GPU fusion is shown in Figure 2.

From the perspective of programming language, since the GPU program is usually developed in C/C++ language and the Spark platform uses Java language for program develop, Java's JNI (Java Native Interface) technology provides a solution to bridge the GPU and Spark through code encapsulation to implement interfaces for the Worker to call. Several JNI tools for GPU programming can be used. For example, JCuda (http://www.jcuda.org) is a development kit that provides bindings to the CUDA runtime, which currently includes multiple packages such as JCublas, JCufft, JCurand, JCusparse, JCusolver, JCudpp, JNpp, and JCudnn etc. It is convenient to write GPU programs in Java language. Other user-defined GPU programs written in C/C++ can also be called after being packaged into Java functions.

For the developers, a bidirectional transmission channel between the main memory and the GPU global memory should be established. If the operation of the RDD is transferred to the GPU core, high-speed data transmission between the main memory and the GPU global memory is required, which is also implemented by function encapsulations, as is demonstrated in Figure 2.

4.3. GPU-Accelerated NMF. As we demonstrated the matrix iterative process in equations (3) and (4) and Figure 1, the main principle of GPU-based parallel NMF is presented in Figure 3. The basic idea of GPU-based parallel NMF is to design several kernel functions to implement update rules for matrices H and W. H and W are blockwise-transferred. In Figure 3, circled operations denote CUDA kernels, and "$\cdot*$" and "$\cdot/$" denote point-wise matrix operations, multiplication, and division, respectively. Most of the matrix operations can be implemented using the libraries of Cublas and Cusparse, together with two self-defined operations, dot multiplication and dot division. In order to reduce the

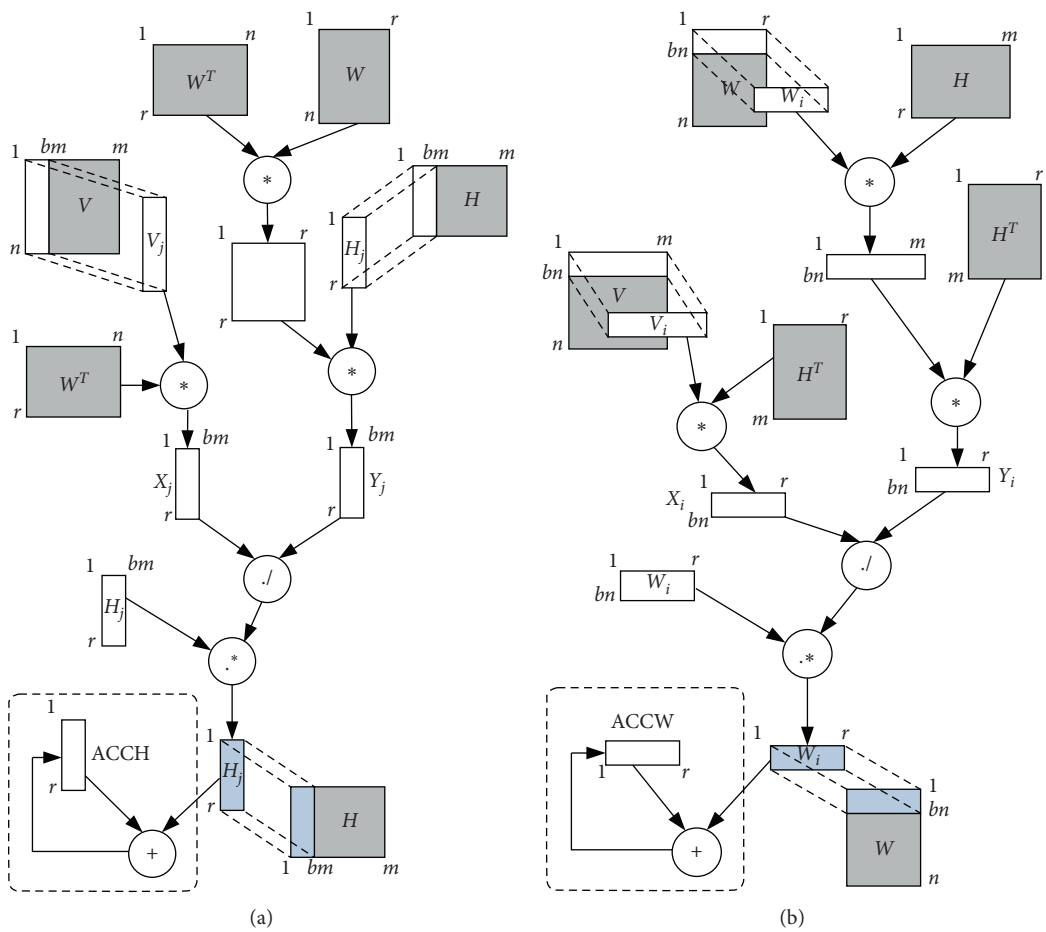


Figure 3: GPU implementation of iteration.

programming difficulty, JNI technology is used to transfer the CUDA programs to Java function encapsulations, which are called by Spark executors.

4.4. GPU-Accelerated NMF on Spark. Spark has advantages in iterative computing, and GPU has advantages in numerical calculation of vectors and matrices. In the Spark-GPU fusion platform, fast memory read and write, combined with GPU acceleration, can play their respective advantages to improve performance. NMF calculation is started and controlled by Spark driver. The Workers calculate the parallel tasks iteratively in a distributed manner. Workers are optimized with the highest speed using the GPU device and running the GPU kernel functions to complete the task. All intermediate results are written to the memory in each iteration and exchanged among the Workers and sent to the GPU global memory. Until the iterations are terminated, the tasks are completed and the results are written to HDFS.

The whole algorithm is described in Algorithm 1. Matrix V is broadcasted to all executors, and each Worker obtains the corresponding matrix block W_i or H_j from RDD. In the Spark platform, after the Action operator is triggered, all accumulated operators form a directed acyclic graph. Task is split into different stages based on different dependencies between RDDs. One stage consists of a series of function execution pipelines. The stages of GPU-accelerated NMF through RDD are listed as follows:

Stage 1: Read and convert matrices W and H; perform *mapPartition* function to update H blocks;

Stage 2: Splice all blocks of H after one iteration through performing a *collect* operation;

Stage 3: Read and convert matrices W and H; perform *mapPartition* function to update W blocks;

Stage 4: Splice all blocks of W after one iteration through performing a *collect* operation, and prepare for the next iteration.

Then, iteratively preform the above four stages. The method of caching data in memory is much faster than in file system for each iteration. When the convergence condition is reached, the matrices updating is terminated, and the results are then written to HDFS.

Input: Original matrix $V_{n\times m}$, low rank r and iteration times *iter*
Input: Context of Spark Environment *sc*
Input: Number of executors *en* and number of data partitions *pn*
Input: Data collection of matrix elements $dc\,M$ and $dc\,H$ for matrices M and H
Input: Data collection in the form of RDD $rdd\,M$ and $rdd\,H$ for matrices M and H
Output: Matrices $W_{n\times r}$ and $H_{r\times m}$ after decomposition
(1) generate initial W, H by random
(2) $dc\,W \leftarrow W, dc\,H \leftarrow H$
(3) broadcast V
(4) **for** $k = 1$: iter **do**
(5) $r\,dd\,H \leftarrow sc.parallelize(dc\,H, pn)$
(6) //update H
(7) call $r\,dd\,H.mapPartition(mapH)$
(8) **function** mapH(data, result)
(9) $X \leftarrow gpu\_multiply(W^T, V)$
(10) $WW \leftarrow gpu\_multiply(W^T, W)$
(11) $Y \leftarrow gpu\_multiply(WW, data)$
(12) $da\,ta \leftarrow gpu\_do\,t\_multiply(data, X)$
(13) $result \leftarrow gpu\_do\,t\_divide(data, Y)$
(14) return result
(15) **end function**
(16) $dc\,H \leftarrow rdd\,H.collect$
(17) $rdd\,W \leftarrow sc.parallelize(dc\,W, pn)$
(18) //update W
(19) call $r\,dd\,W.mapPartition(mapW)$
(20) **function** *mapW(data, result)*
(21) $X \leftarrow gpu\_multiply(V, H^T)$
(22) $WH \leftarrow gpu\_multiply(W, H)$
(23) $Y \leftarrow gpu\_multiply(WH, H^T)$
(24) $da\,ta \leftarrow gpu\_do\,t\_multiply(da\,ta, X)$
(25) $result \leftarrow gpu\_do\,t\_divide(da\,ta, Y)$
(26) $return\,result$
(27) **end function**
(28) $dc\,W \leftarrow rdd\,W.collect$
(29) **end for**
(30) $W \leftarrow dc\,W, H \leftarrow dc\,H$

Algorithm 1: GPU-accelerated NMF on Spark.

5. Collaborative Filtering Algorithm Based on NMF

5.1. Classic Collaborative Filtering Algorithm. Collaborative filtering recommendation algorithms can be divided into two categories: user-based CF and item-based CF. The recommendation process based on collaborative filtering can be described as three stages:

Stage 1: *Collect user preferences.* After preprocessing the user behavior data, according to different behavior analysis methods, you can choose grouping or weighting to obtain a "user-item" preference matrix V whose size is $n \times m$, where n is the number of users, m is the number of items, and matrix element v_{ij} denotes the i-th user's preference for the j-th item, which is generally a floating point number in the range [1, 5] or a binary value of 0 or 1. The value highly depends on the content of the item. If the item is a commodity in e-commerce, the value indicates whether the user purchased or not. Sometimes, it means whether the user watched or not, or the interest is like or dislike, or the interest is high or low.

Stage 2: *Discovery of similar users or items.* In the "user-item" preference matrix, a user's preference for all items is used as a vector to calculate the similarity between users to obtain the similarity matrix sim. For a specific user u, from the remaining $n-1$ users in the system, the similarity value corresponding to the user u is sorted in descending order; the k-nearest neighbor users with the largest similarity value are selected to form the nearest neighbor user set $N = \{n_1, n_2, \ldots, n_k\}$.

For the item-based CF, all users' preferences for an item are regarded as a vector to calculate the similarity between items. Generally, there are three common methods for calculating similarity: Euclidean distance, Pearson correlation coefficient, and Cosine similarity. This paper uses Pearson correlation coefficient as an example [44]. The reason why we choose Pearson correlation coefficient is that, different from the Euclidean distance, Pearson correlation coefficient is able

to reduce the grade inflation error, which is relevant in the recommendation domain. The formula for calculating the similarity using Pearson correlation coefficient is presented as follows:

$$\operatorname{sim}(u_1, u_2) = \frac{m\sum_{j=1}^{m} v_{u_1 j} v_{u_2 j} - \sum_{j=1}^{m} v_{u_1 j} \sum_{j=1}^{m} v_{u_2 j}}{\sqrt{m\sum_{j=1}^{m} v_{u_1 j}^2 - \left(\sum_{j=1}^{m} v_{u_1 j}\right)^2}\sqrt{m\sum_{j=1}^{m} v_{u_2 j}^2 - \left(\sum_{j=1}^{m} v_{u_2 j}\right)^2}}, \tag{5}$$

where u_1 and u_2 denote two users, $\operatorname{sim}(u_1, u_2)$ is the similarity of users u_1 and u_2, and v_{u1j} and v_{u2j} are the ratings of j-th items given by users u_1 and u_2.

Stage 3: *Generate the prediction matrix and Top-N recommendation results.* Using the score given by the nearest neighbor on the item, the user's score on the specific item is calculated through the weighted average of the similarities. Suppose that user u's nearest neighbor set $N = \{n_1, n_2, \ldots, n_k\}$; user u's prediction score for an item i is denoted as v'_{ui}, which is shown in the following equation:

$$v'_{ui} = \overline{u} + \frac{\sum_{r \in N} \operatorname{sim}(u, r) \times (v_{ri} - \overline{r})}{\sum_{v \in N} \operatorname{sim}(u, v)}, \tag{6}$$

where $\overline{u}$ is the average rating of items by user u, $\operatorname{sim}(u, r)$ is the similarity between user u and user r, v_{ri} is the rating of item i by user r, and $\overline{r}$ is the average rating of items by user r. Then, sort the items that user i did not score or purchase according to the predicted score, and obtain the Top-N items as the recommendation data set and recommend them to user i.

5.2. Collaborative Filtering Algorithm Based on NMF. The collaborative filtering algorithm based on NMF proposed in this paper can be divided into two processes: matrix factorization with dimensionality reduction and collaborative filtering.

(1) Matrix factorization and dimension reduction

Step 1: Using GPU-based NMF, the large-scale user preference matrix V is approximated by the product of two matrices W and H. The base matrix W stands for the item feature matrix, which contains the reduced set of r factors (r is the rank in NMF), and the projection matrix H stands for the user feature matrix, which stores the coefficient of the linear combination of the r factors.

Step 2: According to matrix W, the projection vector of the target user u_i's rating vector corresponding to the base matrix W can be calculated and denoted as h_i.

However, choosing a suitable number of latent factors will have an impact on the effect of NMF. In this paper, in order to improve the collaborative filtering-based recommendation, we need to select the optimal rank r for NMF. According to the cophenetic correlation coefficient [51], we repeat NMF several times per rank and calculate how similar the results are and, in other words, how stable the identified clusters are, given that the initial seed is random. We choose the highest r before the cophenetic coefficient drops.

(2) Collaborative filtering

Step 1: For the GPU-accelerated user similarity calculation, each user is assigned a thread in CUDA programming model, a kernel function is designed to calculate the Pearson correlation coefficient, and the similarity between h_i and each column of the projection matrix H is calculated in parallel to obtain the user similarity matrix *sim*.

Step 2: Top k users with the highest value of similarity form the nearest neighbor set N for user u_i.

Step 3: Use the neighbors of u_i in the nearest neighbor set N and the corresponding original scores in V to perform weighted calculation to generate the score prediction matrix p.

Step 4: Sort and get Top-N recommendation result using the prediction matrix p.

6. Performance Evaluations

6.1. Experiment Setups. For our experiments, we have used four n1-standard-4 instances of Google Compute Engine, and each instance is configured with 4 vCPU, 15 GB memory, and 100 GB SSD hard disk in asia-east1 district. Each instance is also configured with a NVIDIA K80 GPU with 2496 CUDA cores and 12 GB global memory. In the 4-node cluster, 64-bit Ubuntu 16.04 LTS is installed, and other software packages include Hadoop 2.7, Spark 2.3, JDK 1.8, and CUDA 9.0.

6.2. Data Set. MovieLens data set (https://grouplens.org/datasets/movielens/) provided by the GroupLens research group is used in the experiment. It contains the scores of 130,642 movies scored by 7,120 users. We randomly select data sets of different sizes for testing. Each user must rate at least 20 movies, the range of ratings is from 1 to 5, and the higher the rating is the more satisfied the user was. In the experiment, the movie ratings are converted into a scoring matrix. If a user does not rate a movie, the corresponding matrix element value is 0; thus the scoring matrix is a typical sparse matrix. In our experiment, the data set V we randomly selected needs to be further divided into a training set VT and a test set T through splitting the nonzero elements,

and 70% of the data set as the training set and the other 30% as the test set, making sure that matrix VT and matrix T have the same size as that of matrix V.

6.3. Baseline Algorithms

Serial NMF. Serial NMF algorithm is performed in a single thread using CPU only. According to equation (3) and (4), the method of alternately updating W and H is used to obtain the decomposition results by performing multiple iterations.

GPU-based NMF. GPU-based NMF algorithm is also performed in a single thread but with one GPU device support. As you see in Figure 3, alternately updating W and H is accelerated by GPU, implemented using the libraries of Cublas and Cusparse, together with two self-defined operations, dot multiplication and dot division.

Spark-based NMF without GPU support. For this algorithm, NMF is computed in a Spark cluster, and each node has no GPU device. Similar to Algorithm 1, in the two stages of $rddH.mapPartition$ and $rddW.mapPartition$, there is no GPU support for the updating of H and W and only CPU for matrix operations in each iteration.

6.4. Evaluation Metrics for Recommendation. In order to accurately measure the performance of algorithms, in addition to the running time, the accuracies of prediction scores and recommendation results are also considered. In this paper, we use root mean square error (RMSE) and mean absolute error (MAE) to measure the accuracy of prediction scores. For the measurement of the accuracy of recommended results, the accuracy rate (Precision) and the recall rate (Recall) are generally used for measuring, together with F-measure for comprehensive consideration of contradictions between the two indicators.

RMSE measures the accuracy of predictions based on the root mean square error between the predicted score and the actual score. The smaller the value of RMSE, the more accurate the prediction result and the higher the quality of the recommended algorithm. The prediction score set $p = \{p_1, p_2, \ldots, p_n\}$ is obtained through training, and the actual user preference score set $T = \{t_1, t_2, \ldots, t_n\}$ is in the test set. Therefore, RMSE is defined as

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(p_i - t_i)^2}{n}}. \tag{7}$$

MAE measures the accuracy of predictions based on the average deviation between the predicted score and the actual score, which is defined as

$$\text{MAE} = \frac{\sum_{i=1}^{n} |p_i - t_i|}{n}. \tag{8}$$

In our experiments, we use three evaluation metrics to evaluate the performance: *Precision*, *Recall*, and *F-measure*. Among the items that have never been purchased or rated, N items with the highest predicted ratings are selected to form the Top-N recommendation list. We define R_u as the set of items recommended for user u and define T_u as the set of items actually liked by user u in the test set. Accuracy means the proportion of related items in the recommended items. Simply speaking, it is recommendation hit rate (the hit means the recommended item has a score in the test set and the score exceeds a certain threshold). We define U as the set of all users, and the recommended accuracy is defined as Precision:

$$\text{Precision} = \frac{\sum_{u \in U} |R_u \cap T_u|}{\sum_{u \in U} |R_u|}. \tag{9}$$

The ratio of the correct recommended items to all items in the recommendation results is defined as Recall:

$$\text{Recall} = \frac{\sum_{u \in U} |R_u \cap T_u|}{\sum_{u \in U} |T_u|}. \tag{10}$$

F-measure is weighted harmonic average of Precision and Recall, which is defined as follows:

$$F - \text{measure} = \frac{\text{Precision} \times \text{Recall} \times 2}{\text{Precision} + \text{Recall}}. \tag{11}$$

In this paper, the training data VT is factorized by NMF, and then Top-N recommendation results are generated according to the algorithms in Subsection 5.2. The test data T is only used for calculating various recommendation evaluation metrics, such as RMSE, MAE, *Precision*, and *Recall*, without projecting the test data in the latent space created by the training data.

6.5. Result Analysis of NMF. In the experiments, we conducted performance evaluations using four algorithms: (i) Serial NMF, (ii) GPU-based NMF, (iii) Spark-based NMF without GPU support, and (iv) Spark-based NMF with GPU support which is proposed in this paper and developed on Spark-GPU fusion platform. We designed three performance comparisons to validate the new proposed algorithm. We select some typical matrix dimensions, and the number of iterations is 100.

6.5.1. Performance of GPU Speedup. We performed GPU-based NMF in a single node, and we varied the matrix dimensions as seen in Figure 4. We measured the computation time, and then we also performed the serial NMF in the same node so as to calculate the GPU speedup to validate the effectiveness of GPU acceleration. The speedup is defined as the ratio of the computation time of the single-node serial method to the computation time of the single-node GPU method; that is, Speedup = $(T_{\text{serial}}/T_{gpu\_\text{parallel}})$. The speedup varies with matrix dimensions, and we have obtained maximum speedup of 45x for GPU when compared with CPU.

6.5.2. Performance of NMF on Spark. In this evaluation, we started the Spark cluster, and the number of worker nodes is varied from 1, 2, and 3 to 4. We varied the matrix dimensions

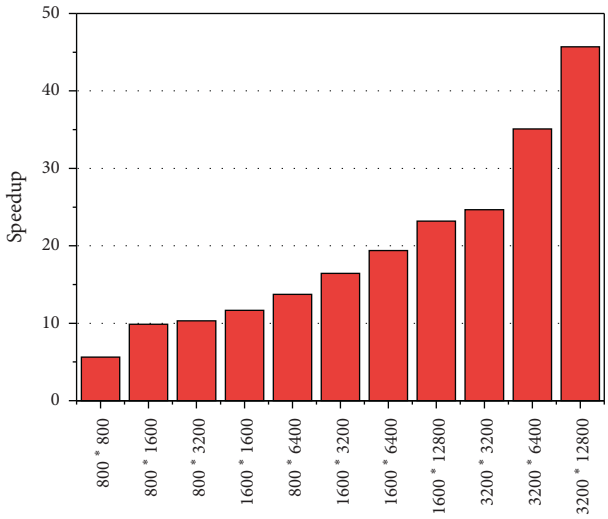


Figure 4: Performance of GPU speedup.

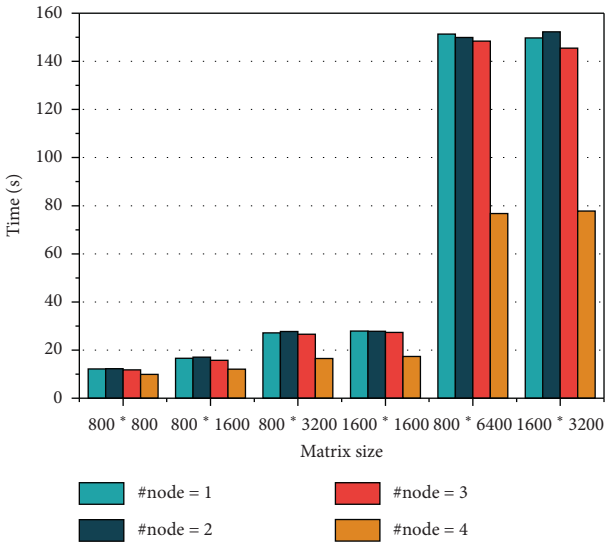


Figure 5: Performance of NMF on Spark.

from $800 * 800$, $800 * 1600$, $800 * 3200$, $1600 * 1600$, and $800 * 6400$ to $1600 * 3200$ and measured the computation time of NMF in Spark platform, and results are shown in Figure 5. When the number of nodes is 4, we set the number of Spark executors to 16, and, with the increase of the matrix dimensions, the advantages of 4 nodes are becoming more and more obvious. Compared with 3-node Spark platform, the computation time of 4 nodes saves about 50% of the time.

6.5.3. Performance of NMF on Spark with GPU Support. In the last evaluation, we started the Spark cluster, the number of nodes is 4, and we varied the matrix dimensions from $6400 * 6400$, $3200 * 25600$, and $6400 * 12800$ to $6400 * 25600$ and compared GPU support with non-GPU support. As can be seen from Figure 6, in the 4-node Spark platform, the computation time of NMF with GPU is smaller than that of NMF without GPU. When the size of matrix is $6400 * 25600$, NMF on Spark with GPU support saves about 10.8% of the time. NMF on GPU-accelerated Spark platform obviously shows execution efficiency.

Due to the mathematical fundamental of NMF and the blockwise-based parallel principle, there are frequent data distributions and data collections among all executors, and the communication cost is very high for the NMF on Spark. However, compared with data distributions and data collections, the execution of mapPartition function takes much less time due to the GPU acceleration. From the perspective of time analysis, communication and data exchange are the bottlenecks of NMF parallel algorithm. NMF on GPU-accelerated Spark platform still has great potential for improvement.

6.6. Result Analysis of Collaborative Filtering. In the experiments, we compared the performance of three algorithms: traditional user-based CF, traditional item-based CF, and the NMF-based CF proposed in this paper. The size of the matrices changes from $400 * 800$, $400 * 1600$, $800 * 1600$, and $800 * 3200$ to $1200 * 3200$ for testing. The number of iterations is 100, and we select 50 items for each user as the Top-50 recommendation list.

6.6.1. Comparison of Score Prediction Accuracy. In order to compare the prediction accuracy, we compared RMSE and MAE of the three algorithms, as shown in Figures 7(a) and 7(b), respectively. Under five different score matrix sizes, NMF-CF is significantly better than User-CF and Item-CF in terms of both RMSE and MAE. The results of RMSE and MAE for NMF-CF are the smallest, while the Item-CF algorithm obtains the largest prediction error and the worst prediction effect. When the size of the matrix is $400*800$, compared with the Item-CF algorithm, the result of RMSE for NMF-CF is reduced by 31.64%, and the MAE for NMF-CF is reduced by 28.5%.

6.6.2. Comparison of Recommendation Accuracy. The recommendation performances of Precision, Recall, and F-measure of the three algorithms are shown in Figures 8(a)–8(c), respectively. Under the five different score matrix sizes, as the size of the matrix increases, all the indicators for three algorithms have declined. NMF-CF is superior to User-CF and Item-CF algorithms in all three indicators, which explicitly shows that the quality of CF recommendations based on NMF is the best. However, with the increase of the matrix size $n \times m$, especially the increase of m, it means that the number of items increases, and we only recommend 50 items in collaborative filtering. When calculating the two indicators Precision and Recall, we compare with the 30% test set, and the hit rate of recommended items will be lower, and the advantage of NMF is getting smaller and smaller. When the matrix size is $1600 * 3200$, the results of NMF-CF and User-CF algorithm are almost the same. The recommended effect of the Item-CF algorithm has always been the worst.

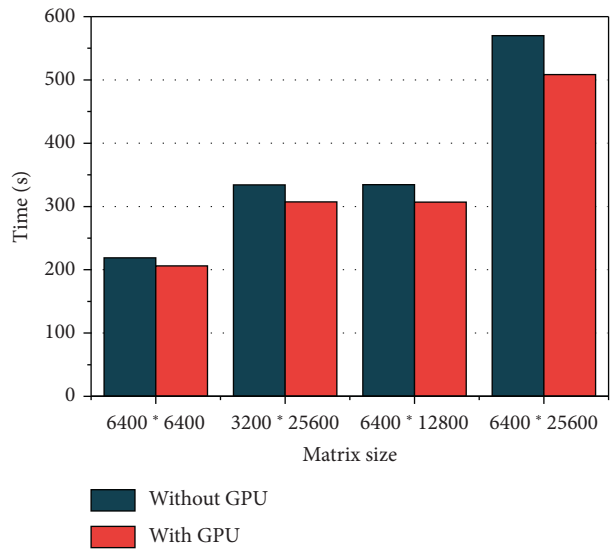


Figure 6: Performance of NMF on Spark with/without GPU support.

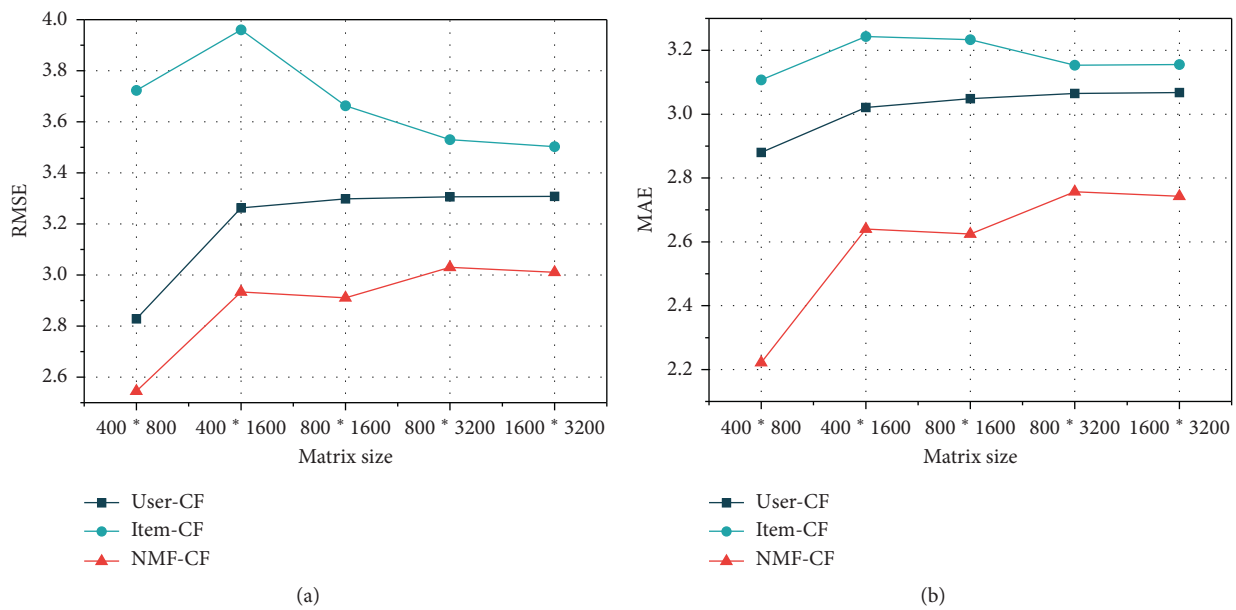


Figure 7: Score prediction accuracy results. (a) Comparison of RMSE and (b) comparison of MAE.

6.6.3. Comparison of Running Time for Recommendation. First, we only evaluated the running time of NMF-CF algorithm, and we considered two conditions in Spark platform: (i) CPU-based NMF-CF (only 1 node used) and (ii) CPU + GPU-based NMF-CF (4 nodes used). In the scenarios of five matrix sizes, the result of the running time is shown in Figure 9. It can be seen from the figure that when the GPU acceleration is adopted, the computation time for NMF is significantly reduced. As the matrix size becomes larger, the parallel efficiency is getting higher, and the acceleration effect is also getting better. When the matrix size is $1600 * 3200$, due to the utilization of GPU, CPU + GPU-based approach is reduced by 44.8% compared to CPU-based approach, which also proved the acceleration performance of GPU for NMF-CF.

Then, the running time comparison has been performed in the 4-node Spark platform with GPU support for the three algorithms. In all three algorithms, GPU is used to calculate Pearson correlation coefficient, and NMF-CF algorithm uses GPU to calculate NMF. The running time comparison results are shown in Figure 10. It can be seen from the figure

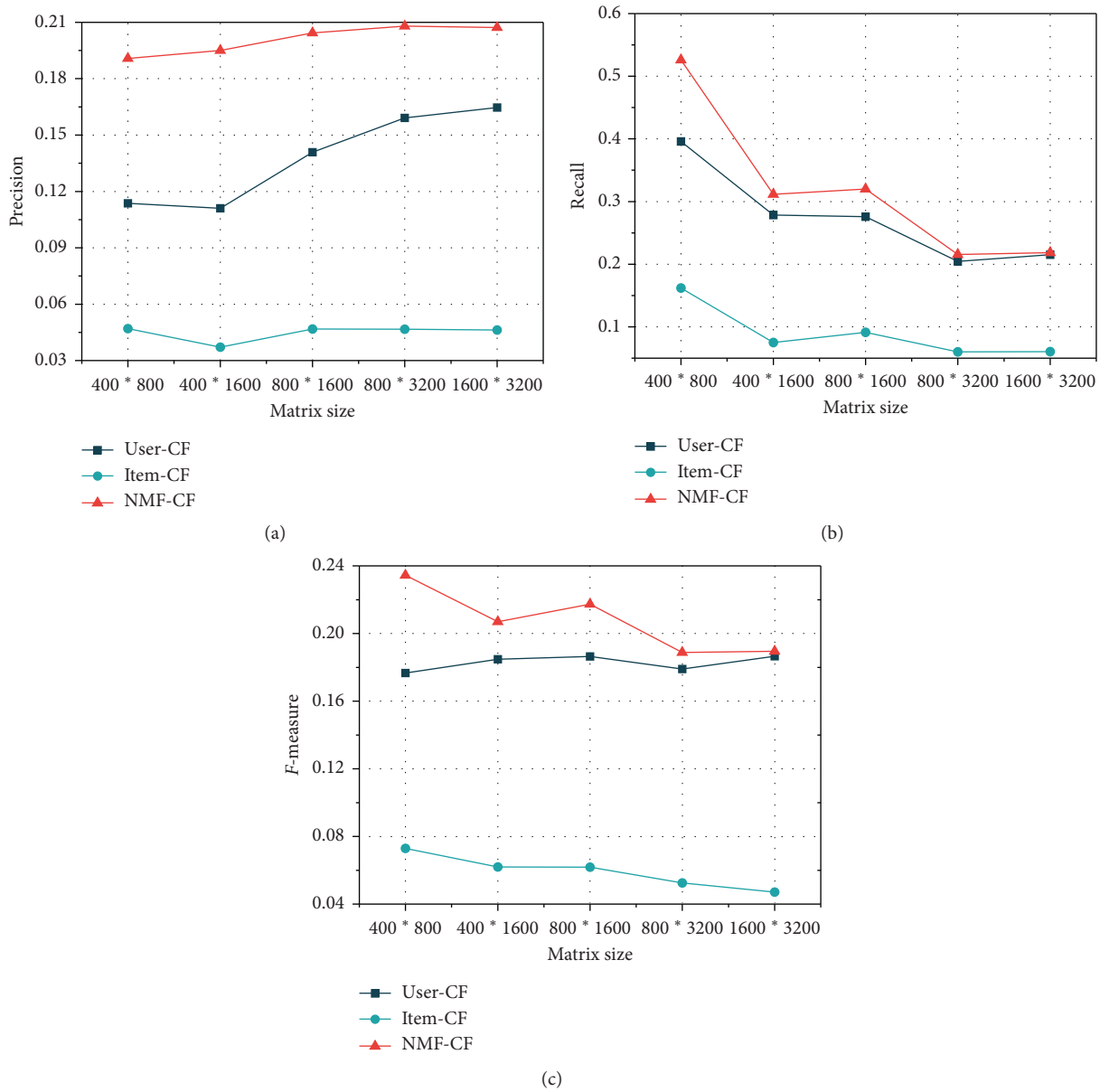


Figure 8: Recommendation accuracy results. (a) Comparison of Precision, (b) comparison of Recall, and (c) comparison of F-measure.

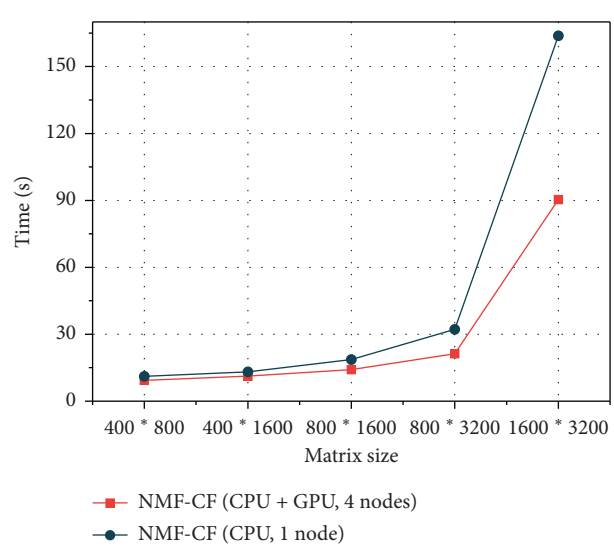


Figure 9: Running time comparison of NMF-CF under two conditions.

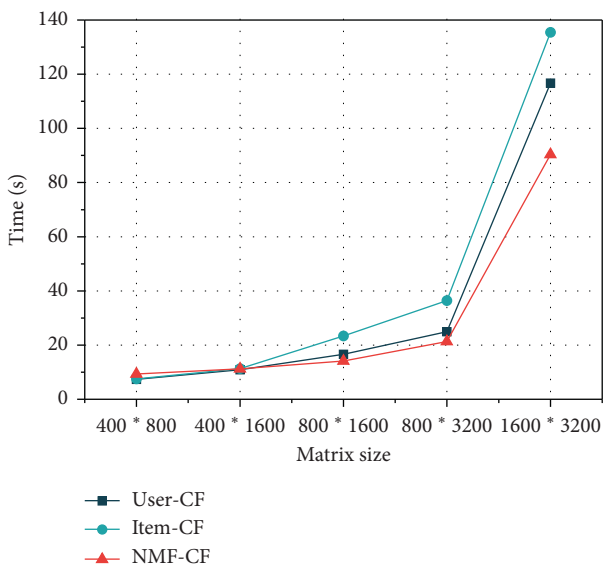


Figure 10: Running time comparison of three CF recommendation algorithms.

that, with the increase of the matrix size, the running time of each algorithm has increased. The running time of the NMF-CF recommendation algorithm significantly outperforms the User-CF and Item-CF algorithms. When the matrix size is $1600 * 3200$, the running time of the NMF-CF algorithm is reduced by 33.3% compared with the Item-CF algorithm and is reduced by 22.5% compared to the User-CF algorithm.

Overall, compared with the traditional CF, the NMF-CF recommendation algorithm contains the process of the decomposition of the scoring matrix V into W and H, which seems to be a time-consuming operation. In fact, when calculating the correlation coefficient later, it will save a lot of time for NMF-CF. Since the size of matrix W is $n \times r$, where the value of r reflects the number of features or topics, the value of r is usually very small (it generally takes a value of 2 to 10), and the size of matrix W is small, so through calculating the correlation coefficient to obtain k-nearest neighbors for each user takes much less time in NMF-CF algorithm than in the User-CF algorithm or Item-CF algorithm. In addition, except the increased accuracy, NMF-based CF recommendation algorithm uses GPUs to run in parallel and the elapsed computation time is still the shortest.

7. Conclusion

In the heterogeneous CPU/GPU cluster, nodes have large memory resources and GPU multicore resources, and the advantages of distributed storage between nodes and data sharing within nodes should be utilized. Heterogeneous parallel computing is an efficient and feasible parallel programming strategy. A GPU-accelerated NMF algorithm on Spark platform has been designed in this paper to solve the problem of low processing speed of NMF as the size of the matrix increases. Through the performance evaluations, experimental results have proved that the combination of Spark-based in-memory computing and GPU has higher execution efficiency. On the other hand, recommendation systems have been widely applied in many fields, but as the user number and item number increase, the computational speed also becomes slower and the accuracy of recommendation decreases. Although traditional collaborative filtering is extremely successful in the recommendation system, as the data increases, the recommendation algorithms have been confronted with various problems, such as scalability problems, cold start problems, and matrix sparseness problems. This paper implemented the NMF algorithm for collaborative filtering recommendation, which combines NMF with traditional collaborative filtering methods, decomposes the original score data into base matrix and projection matrix, and runs in parallel on Spark platform accelerated by GPU. Experiments on matrices with different size show that the parallel NMF collaborative filtering recommendation algorithm not only improves the prediction and recommendation accuracy but also greatly improves the calculation efficiency.

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 they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants nos. 61602169 and 61872138, the National Key R&D Program of China under Grant no. 2018YFB1402800, and the Natural Science Foundation of Hunan Province under Grant no. 2018JJ2135, as well as the Scientific Research Fund of Hunan Provincial Education Department under Grant no. 18A186.

References

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., pp. 556–562, MIT Press, Denver, CO, USA, 2000.

[3] A. Falini, G. Castellano, C. Tamborrino et al., "Saliency detection for hyperspectral images via sparse-non negative-matrix-factorization and novel distance measures," in *Proceedings of the 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems, EAIS 2020*, pp. 1–8, IEEE, Bari, Italy, 2020.

[4] V. Leplat, N. Gillis, and A. M. S. Ang, "Blind audio source separation with minimum-volume beta-divergence NMF,"

IEEE Transactions on Signal Processing, vol. 68, pp. 3400–3410, 2020.

[5] X. Luo, M. Zhou, Y. Xia, and Q. Zhu, "An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1273–1284, 2014.

[6] S. Rendle and L. Schmidt-Thieme, "Online-updating regularized kernel matrix factorization models for large-scale recommender systems," in *Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008*, P. Pu, D. G. Bridge, B. Mobasher, and F. Ricci, Eds., pp. 251–258, ACM, Lausanne, Switzerland, 2008.

[7] Y. Chen, M. Rege, M. Dong, and J. Hua, "Non-negative matrix factorization for semi-supervised data clustering," *Knowledge and Information Systems*, vol. 17, no. 3, pp. 355–379, 2008.

[8] J. Choo, C. Lee, C. K. Reddy, and H. Park, "UTOPIAN: user-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 1992–2001, 2013.

[9] V. Kysenko, K. Rupp, O. Marchenko, S. Selberherr, and A. Anisimov, "GPU-accelerated non-negative matrix factorization for text mining," in *Natural Language Processing and Information Systems—17th International Conference on Applications of Natural Language to Information Systems, NLDB 2012, Groningen, The Netherlands, June 26–28, 2012. Proceedings, Lecture Notes in Computer Science*, G. Bouma, A. Ittoo, E. Métais, and H. Wortmann, Eds., vol. 7337, pp. 158–163, Springer, Berlin, Germany, 2012.

[10] R. Liao, Y. Zhang, J. Guan, and S. Zhou, "CloudNMF: a mapreduce implementation of nonnegative matrix factorization for large-scale biological datasets," *Genomics, Proteomics & Bioinformatics*, vol. 12, no. 1, pp. 48–51, 2014.

[11] S. Mittal and J. S. Vetter, "A survey of CPU-GPU heterogeneous computing techniques," *ACM Comput. Surv.*vol. 47, no. 4, pp. 1–69, 2015.

[12] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: cluster computing with working sets," in *2nd USENIX Workshop on Hot Topics in Cloud Computing, HotCloud'10*, E. M. Nahum and D. Xu, Eds., USENIX Association, Boston, MA, USA, 2010.

[13] M. Zaharia, R. S. Xin, P. Wendell et al., "Apache spark," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[14] M. Zaharia, M. Chowdhury, T. Das et al., "Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing," in *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2012*, pp. 15–28, USENIX Association, San Jose, CA, USA, 2012.

[15] J. Dean and S. Ghemawat, "MapReduce," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[16] B. Tang, M. Tang, G. Fedak, and H. He, "Availability/network-aware mapreduce over the internet," *Information Sciences*, vol. 379, pp. 94–111, 2017.

[17] R. Kannan, G. Ballard, and H. Park, "A high-performance parallel algorithm for nonnegative matrix factorization," in *Proceedings of the 21st ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPoPP 2016*pp. 1–11, Barcelona, Spain, 2016.

[18] E. Mejía-Roa, D. Tabas-Madrid, J. Setoain, C. García, F. Tirado, and A. D. Pascual-Montano, "NMF-mGPU: non-negative matrix factorization on multi-GPU systems," *BMC Bioinformatics*, vol. 16, pp. 1–43, 2015.

[19] C. Liu, H. Yang, J. Fan, L. He, and Y. Wang, "Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce," in *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, Eds., pp. 681–690, ACM, Raleigh, NC, USA, 2010.

[20] H. Jiang, Y. Chen, Z. Qiao, K.-C. Li, W. Ro, and J.-L. Gaudiot, "Accelerating mapreduce framework on multi-GPU systems," *Cluster Computing*, vol. 17, no. 2, pp. 293–301, 2014.

[21] T. Gao, Y. Guo, B. Zhang et al., "Memory-efficient and skew-tolerant mapreduce over MPI for supercomputing systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 12, pp. 2734–2748, 2020.

[22] M. Sergent, M. Dagrada, P. Carribault, J. Jaeger, M. Pérache, and G. Papauré, "Efficient communication/computation overlap with MPI+OpenMP runtimes collaboration," in *Euro-Par 2018: Parallel Processing - 24th International Conference on Parallel and Distributed Computing, Turin, Italy, August 27-31, 2018, Proceedings, Lecture Notes in Computer Science*, M. Aldinucci, L. Padovani, and M. Torquati, Eds., vol. 11014, pp. 560–572, Springer, Berlin, Germany, 2018.

[23] J. A. Stuart and J. D. Owens, "Multi-GPU mapreduce on GPU clusters," in *25th IEEE International Symposium on Parallel and Distributed Processing, IPDPS 2011*, pp. 1068–1079, IEEE, Anchorage, AK, USA, 2011.

[24] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Online nonnegative matrix factorization with robust stochastic approximation," *IEEE Trans. Neural Networks Learn. Syst.*vol. 23, no. 7, pp. 1087–1099, 2012.

[25] D. Tu, L. Chen, M. Lv, H. Shi, and G. Chen, "Hierarchical online NMF for detecting and tracking topic hierarchies in a text stream," *Pattern Recognition*, vol. 76, pp. 203–214, 2018.

[26] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011.

[27] M. Tepper and G. Sapiro, "Compressed nonnegative matrix factorization is fast and accurate," *IEEE Transactions on Signal Processing*, vol. 64, no. 9, pp. 2269–2283, 2016.

[28] A. Janecek, S. S. Grotthoff, and W. N. Gansterer, "libNMF—a library for nonnegative matrix factorization," *Computing and Informatics*, vol. 30, no. 2, pp. 205–224, 2011.

[29] N. Lopes and B. Ribeiro, "Non-negative matrix factorization implementation using graphic processing units," in *Intelligent Data Engineering and Automated Learning - IDEAL 2010, 11th International Conference, Paisley, UK, September 1-3, 2010. Proceedings, Lecture Notes in Computer Science*, C. Fyfe, P. Tiño, D. Charles et al., Eds., vol. 6283, pp. 275–283, Springer, Berlin, Germany, 2010.

[30] E. Battenberg and D. Wessel, "Accelerating non-negative matrix factorization for audio source separation on multi-core and many-core architectures," in *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center*, K. Hirata, G. Tzanetakis, and K. Yoshii, Eds., pp. 501–506, International Society for Music Information Retrieval, Kobe, Japan, 2009.

[31] B. Tang, L. Bobelin, and H. He, "Parallel algorithm of non-negative matrix factorization based on hybrid MPI and OpenMP programming model," *Computer Science*, vol. 44, no. 3, pp. 51–54, 2017.

[32] Z. Sun, T. Li, and N. Rishe, "Large-scale matrix factorization using mapreduce," in *ICDMW 2010, The 10th IEEE International Conference on Data Mining Workshops*, W. Fan, W. Hsu, G. I. Webb et al., Eds., pp. 1242–1248, IEEE Computer Society, Sydney, Australia, 2010.

[33] B. Tang, L. Kang, Y. Xia, and L. Zhang, "GPU-accelerated large-scale non-negative matrix factorization using spark," in *Collaborative Computing: Networking, Applications and Worksharing—14th EAI International Conference, CollaborateCom 2018, Shanghai, China, December 1–3, 2018, Proceedings, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, H. Gao, X. Wang, Y. Yin, and M. Iqbal, Eds., vol. 268, pp. 189–201, Springer, Berlin, Germany, 2018.

[34] B. Kwon and H. Cho, "Scalable co-clustering algorithms," in *Algorithms and Architectures for Parallel Processing, 10th International Conference, ICA3PP 2010, Busan, Korea, May 21–23, 2010. Proceedings. Part I, Lecture Notes in Computer Science*, C. Hsu, L. T. Yang, J. H. Park, and S. Yeo, Eds., vol. 6081, pp. 32–43, Springer, Berlin, Germany, 2010.

[35] Z. Liu, Y. Zhang, E. Y. Chang, and M. Sun, "PLDA+: parallel latent dirichlet allocation with data placement and pipeline processing," *ACM Trans. Intell. Syst. Technol.*vol. 2, no. 3, pp. 1–26, 2011.

[36] E. Karydi and K. G. Margaritis, "Multithreaded implementation of the slope one algorithm for collaborative filtering," in *Artificial Intelligence Applications and Innovations - 8th IFIP WG 12.5 International Conference, AIAI 2012, Halkidiki, Greece, September 27–30, 2012, Proceedings, Part I, IFIP Advances in Information and Communication Technology*, L. S. Iliadis, I. Maglogiannis, and H. Papadopoulos, Eds., vol. 381, pp. 117–125, Springer, Berlin, Germany, 2012.

[37] H. Yu, C. Hsieh, S. Si, and I. S. Dhillon, "Scalable coordinate descent approaches to parallel matrix factorization for recommender systems," in *12th IEEE International Conference on Data Mining, ICDM 2012*, M. J. Zaki, A. Siebes, J. X. Yu, B. Goethals, G. I. Webb, and X. Wu, Eds., pp. 765–774, IEEE Computer Society, Brussels, Belgium, 2012.

[38] K. Kato and T. Hosino, "Solving k-nearest neighbor problem on multiple graphics processors," in *Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, CCGrid 2010, 17–20 May 2010*, pp. 769–773, IEEE Computer Society, Melbourne, Victoria, Australia, 2010.

[39] Z. Wang, Y. Liu, and S. Chiu, "An efficient parallel collaborative filtering algorithm on multi-GPU platform," *The Journal of Supercomputing*, vol. 72, no. 6, pp. 2080–2094, 2016.

[40] J. Jiang, J. Lu, G. Zhang, and G. Long, "Scaling-up item-based collaborative filtering recommendation algorithm based on hadoop," in *World Congress on Services, SERVICES 2011*, pp. 490–497, IEEE Computer Society, Washington, DC, USA, 2011.

[41] S. Schelter, C. Boden, and V. Markl, "Scalable similarity-based neighborhood methods with mapreduce," in *Sixth ACM Conference on Recommender Systems, RecSys '12*, P. Cunningham, N. J. Hurley, I. Guy, and S. S. Anand, Eds., pp. 163–170, ACM, Dublin, Ireland, 2012.

[42] A. Narang, A. Srivastava, and N. P. K. Katta, "Distributed scalable collaborative filtering algorithm," in *Euro-Par 2011 Parallel Processing—17th International Conference, Euro-Par 2011, Bordeaux, France, August 29—September 2, 2011, Proceedings, Part I, Lecture Notes in Computer Science*, E. Jeannot, R. Namyst, and J. Roman, Eds., vol. 6852, pp. 353–365, Springer, Berlin, Germany, 2011.

[43] E. Karydi and K. G. Margaritis, "Parallel implementation of the slope one algorithm for collaborative filtering," in *Proceedings of the 16th Panhellenic Conference on Informatics, PCI 2012*, pp. 174–179, IEEE Computer Society, Piraeus, Greece, 2012.

[44] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[45] S. Gong, H. Ye, and Y. Dai, "Combining singular value decomposition and item-based recommender in collaborative filtering," in *Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining, WKDD 2009*, pp. 769–772, IEEE Computer Society, Moscow, Russia, 2009.

[46] H. Polat and W. Du, "SVD-based collaborative filtering with privacy," in *Proceedings of the 2005 ACM Symposium on Applied Computing (SAC)*, H. Haddad, L. M. Liebrock, A. Omicini, and R. L. Wainwright, Eds., pp. 791–795, ACM, Santa Fe, NM, USA, 2005.

[47] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, "Eigentaste: a constant time collaborative filtering algorithm," *Information Retrieval*, vol. 4, no. 2, pp. 133–151, 2001.

[48] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds., pp. 1257–1264, Curran Associates, Inc., Vancouver, British Columbia, Canada, 2007.

[49] N. Thai-Nghe, L. Drumond, T. Horváth, A. Nanopoulos, and L. Schmidt-Thieme, "Matrix and tensor factorization for predicting student performance," in *CSEDU 2011—Proceedings of the 3rd International Conference on Computer Supported Education*, A. Verbraeck, M. Helfert, J. Cordeiro, and B. Shishkov, Eds., vol. 1, pp. 69–78, SciTePress, Noordwijkerhout, Netherlands, 2011.

[50] C. Boutsidis and E. Gallopoulos, "SVD based initialization: a head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, 2008.

[51] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

Hindawi
Scientific Programming
Volume 2020, Article ID 3671835, 1 page
https://doi.org/10.1155/2020/3671835

Corrigendum

Corrigendum to "Security Measurement in Industrial IoT with Cloud Computing Perspective: Taxonomy, Issues, and Future Directions"

Sahar Shah,[1] Mahnoor Khan,[2] Ahmad Almogren,[3] Ihsan Ali,[4] Lianwen Deng,[5] Heng Luo,[5] and Muazzam A. Khan[6]

[1]*Department of Electronics, Quaid-i-Azam University, Islamabad, Pakistan*
[2]*Department of Physics, Government Post Graduate College, Nowshera, Pakistan*
[3]*Chair of Cyber Security, Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia*
[4]*Faculty of Computer Science and IT, University of Malaya, Kuala Lampur, Malaysia*
[5]*School of Physics and Electronics, Central South University, Changsha, China*
[6]*Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan*

Correspondence should be addressed to Ahmad Almogren; ahalmogren@ksu.edu.sa and Ihsan Ali; ihsanalichd@siswa.um.edu.my

Received 16 December 2020; Accepted 16 December 2020; Published 30 December 2020

In the article titled "Security Measurement in Industrial IoT with Cloud Computing Perspective: Taxonomy, Issues, and Future Directions" [1], there was an error in the third and sixth affiliations. The corrected affiliation list is shown above. In addition, the Acknowledgements section should be updated as follows:

"The authors are grateful to the Deanship of Scientific Research, King Saud University for funding through Vice Deanship of Scientific Research Chairs and partially supported by the Faculty of Computer Science and Information Technology, University of Malaya, under Postgraduate Research Grant PG035-2016A."

References

[1] S. Shah, M. Khan, A. Ahmad et al., "Security measurement in industrial iot with cloud computing perspective: taxonomy, issues, and future directions," *Scientific Programming*, vol. 2020, Article ID 8871315, 31 pages, 2020.

Hindawi
Scientific Programming
Volume 2020, Article ID 6613263, 13 pages
https://doi.org/10.1155/2020/6613263

Research Article

AIoT-Based Smart Bin for Real-Time Monitoring and Management of Solid Waste

Aniqa Bano,[1] Ikram Ud Din,[1] and Asma A. Al-Huqail[2]

[1]*Department of Information Technology, The University of Haripur, Haripur, Pakistan*
[2]*Department of Botany and Microbiology, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia*

Correspondence should be addressed to Asma A. Al-Huqail; aalhuqail@ksu.edu.sa

Received 18 November 2020; Revised 8 December 2020; Accepted 17 December 2020; Published 29 December 2020

Academic Editor: Habib Ullah Khan

In the current time, the immense growth in population creates unhygienic environment for the citizen of a society with respect to waste generation. This rapid generation of waste leads to various infectious diseases in the environment. As followed by the traditional municipal system, in our surroundings, we can see over flooding of solid waste in the garbage bins. Solid waste management is a pivotal aspect in traditional systems and it is becoming dangerous in most populated areas. Arduous labor works and costs are required to manage and monitor garbage bins in real time. To maintain the cleanliness of a city and for real-time monitoring of trash bins, a smart bin mechanism (SBM) for smart cities is proposed in this paper, which is based on Artificial Intelligent of Things (AIoT). The SBM works on the 3R concept, that is, Reduce, Recycle, and Reuse. The SBM has the access to get real-time information about each bin and avoid overloading of these bins. The proposed framework reduces the labor cost and saves time and energy of the system. It also reduces the rate of disease infections by keeping the cities clean. Fuzzy logic is used for decision-making in selecting appropriate locations in the cities to install trash bins. The framework is implemented in the multiagent modeling environment, that is, NetLogo.

1. Introduction

The significant growth of the Internet is leading to the emergence of new technology, that is, Internet of Things (IoT) [1]. The term IoT was first used by Ashton in 1999 [2]. IoT is the hub of physical devices that are interlinked through the Internet. These physical devices, that is, sensors, RFID tags, and various intelligent nodes, can communicate at any time from anywhere. IoT is the backbone of future communication systems where everything will communicate and share information smartly without human instructions [3]. The interconnected devices are transformed as smart objects, which have computational skills that are used to monitor an environment leading to smart cities.

IoT promotes various application areas, such as smart health, [4–13] smart city, [14, 15], environment monitoring [16, 17] smart home, [18, 19], traffic management [20–24], smart education system, [25], smart farming, [26, 27], and many others [28–35]. In smart cities, various problems occur when devices communicate with each other; one of the important problems is waste management. The main causes for this problem are the rapid growth of the urban population, high demands for food, and various other factors that are influencing the environment in smart cities. With an increase in population, the management of waste or garbage is a very hectic job to do in the current time. Being a member of the society, every local house, industry, and factory generate some amount of waste on a daily basis. This waste is ultimately collected in waste bins and eventually collected by the municipal vehicles and moved to dumping areas for disposing or recycling processes.

To keep the environment green and clean, monitoring and disposing of waste is very important these days. Improper disposal and poor monitoring of collected waste and waste bins can cause serious damage to human lives. This waste can spread various life-threatening diseases that in turn harm the lives of a whole city and country as well.

Nowadays, cities are facing various problems, such as small parking spaces [36], waste management, communication barriers in traditional systems, and health issues to name a few [37, 38]. All these problems directly affect the living of humans in their daily routine lives [39]. To overcome and solve the existing problems, a new concept has emerged in the light of IoT, named smart city [14]. IoT provides various new services in a smart city and develops an intelligent society [40, 41]. In IoT-based smart cities, physical devices interact and provide ease to humans according to their own intelligence [42]. IoT is further divided into various fields presented in Figure 1 [43], which explains the web-based rankings of each area in percentage. Based on the literature study, smart homes and smart cities get the highest rankings, which shows the trends in the modern era of technologies.

The rapid growth in population and generation of daily routine garbage or waste make the environment unhygienic for the citizens. The waste is divided into two types, that is, wet and solid waste [44]. In this paper, the focus is on solid waste management. Therefore, a waste management mechanism is proposed for smart cities, named SBM (smart bin mechanism), in order to sanitize and clean the environment intelligently. It is designed for solid waste management and recycling of waste because waste is recyclable and can be reused. The waste management procedure comprises five steps that include collecting waste, transporting, analyzing and processing, recycling, and disposing [44]. In SBM, smart bins (SB) are installed in the urban areas at different points that store garbage. Primarily, the SBM is designed for real-time monitoring of the garbage collecting points. The proposed system will reduce the labor work, time, and cost that are very high in the traditional garbage collection system.

In SBM, different entities have been used which show their own roles in providing services to the citizens. It has trash bins, trash collecting vehicles, and a central database to keep records of the levels. All the entities and their roles are discussed in Section 3.

1.1. SBM Contributions. Waste management is an emerging era of most populated as well as less populated cities. In IoT, it also has emerged as a field in smart cities. Various prior research has been conducted for collecting waste, but most of the research is server-based or authority monitoring of garbage bins that are installed in public places. These systems act upon receiving requests from garbage bins and sending collecting vehicles toward the requesting bins. This mechanism consumes enough an amount of energy and time on fulfilling of a single request. Therefore, an intelligent edge-node based mechanism is necessary for collecting waste from requesting bins, which consumes less energy and time. The proposed framework is based on edge-nodes, that is, the trash bin. When a trash bin reaches the threshold level, it makes a request directly to the collecting vehicle instead of forwarding the request to any central authority. The proposed framework has a significant role in waste collecting procedures as it consumes low power because of its novel edge-based mechanism. Time consumption is also low in SBM because of the less iteration from the request-to-response procedure.

Some significant contributions of the proposed framework are as follows:

(i) Proposing a smart bin mechanism that is based on IoT technology and applications

(ii) Real-time monitoring of the trash bins in a smart city

(iii) Using trash bins in an effective way to facilitate the municipal department and citizens as well

(iv) Reducing labor cost and optimizing resources

(v) Improving environment goals and cleaning cities with limited resources

The proposed study is novel in terms of real-time monitoring and decision-making using fuzzy logic processing. Fuzzy logic provides the best suitable and less dense site of the city to install trash bins. Two fuzzy parameters are used in fuzzy inference systems, that is, distance from the collecting/dumping zone and access to the trash bins.

The proposed study is beneficial in the future era of modern technologies, where everything will be connected via the Internet and communicate without human interruption. The study supports the smart city concept by providing real-time monitoring for climate change. It provides a decision-making mechanism by using a fuzzy inference system.

The rest of the paper is divided into five sections; that is, literature study is discussed in Section 2, SBM methodology is presented in Section 3, results and discussion are provided in Section 4, the paper is summarized in Section 5, and challenges and future work are elaborated in Section 6.

2. Literature Study

For the last few years, many researchers are focusing on IoT-based applications, especially smart city [45]. According to [46], a smart city is an infrastructure where everything is interconnected and can interact with each other. In a smart city, everything is supposed to be smart and intelligent in decision-making ability [47]. A smart city leads to a smart environment [30], smart health [48], smart parking, smart economy, smart administration [49], and smart living of the people [50]. The smart city provides all the better facilities to citizens and assures that there is a clean and green environment for them [51]. To make the environment clean, there should be an effective system for collecting waste. In this section, various research about garbage or waste collection and a better management mechanism for the collected waste is reviewed.

2.1. SWMS. Waste management systems play a vital role in reducing the unhygienic objects from a particular area. To avoid these conditions, a smart waste management system (SWMS) [44] is proposed that is based on IoT technology. The SWMS consists of public garbage collectors with

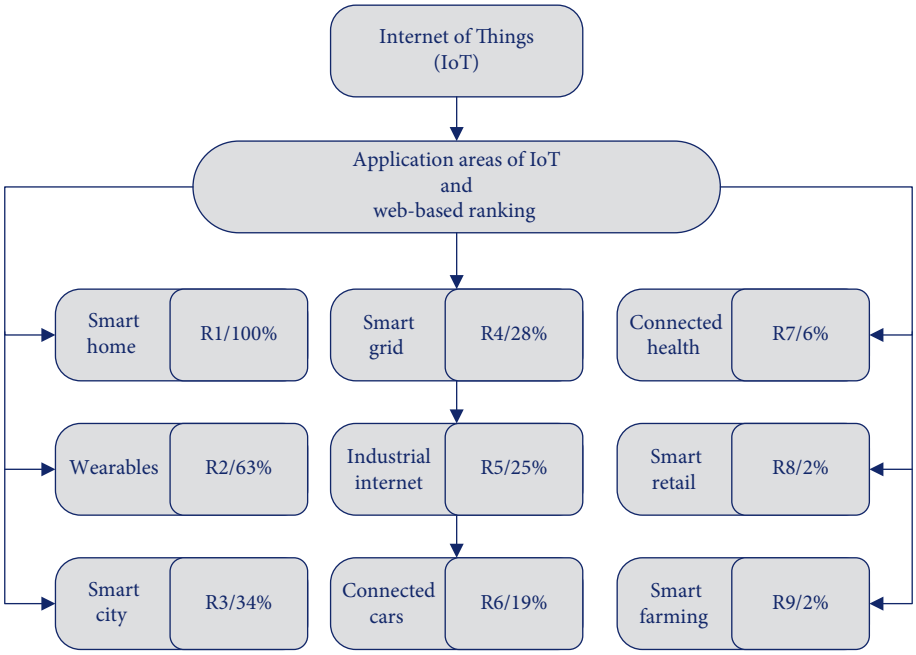


Figure 1: Application areas of Internet of Things.

embedded technology that is used to monitor real-time level of garbage bins in public places. Based on the level of garbage bins, an optimized path is selected for the garbage collecting van that eventually reduces fuel cost. The SWMS divides garbage bins into master and slave bins. Each garbage bin is composed of three sensors, such as level sensor, humidity sensor, and load sensor. The master bin is continuously transmitting its data to the cloud with the help of Wi-Fi. Through real-time monitoring, accurate reports can be generated, and, therefore, the efficiency of the system is enhanced.

2.2. GMS. Overflowing of dust bins at public places increases unhygienic environment for the people, especially, in developing countries; this creates serious health problems for the citizens. To cope with these types of situations, an IoT-based garbage monitoring system (GMS) is proposed in [52]. The system contains various dust bins that are distributed in the city. Dust bins transmit the data to concerned authorities in order to clean the garbage. The block diagram of the proposed model consists of two sections, that is, transmitter and receiver sections. The transmitter is installed in dustbins, which is used to transmit collected data from sensors to the receiver end. At the receiver end, the central system receives the data sent from the dustbin and processes it accordingly. The authors used Raspberry Pi, RF receiver, and a web browser to fulfil the requirements of the system. The proposed system has some limitations in terms of lacking in reliability of communications among different modules.

2.3. IoT-Based SWM. Waste management is an important service provided by smart cities and supported by IoT. An enhanced system for waste management is proposed in [53] by considering the growth of the population in urban areas. The proposed model mainly consists of four entities, such as smart bins, waste areas, management centers, and collecting trucks, as shown in Figure 2. Statistical analysis and decision-making are successfully done based on the data provided by the above-mentioned entities. The authors stated that the proposed model overcomes the existing issues in the waste collection process, that is, location issues, cleaning costs, health hazards, and many others related to waste management.

2.4. SWM by K-Query Scheduling. An IoT-based system is proposed in [54] that is used for waste management with the help of K-Query scheduling. The system is composed of microcontroller module, GPS module, and ultrasonic sensor. These modules are installed in trash cans. The sensors are used to monitor the trash cans. When a trash can reach an appropriate level, the sensor calculates the level and transmits it to the cloud through Internet. The K-Query scheduling is used to store threshold values in a table created in the MySQL database. The architecture of the system is shown in Figure 3. The K-Query is helpful to reduce unknown entries in the database. For a shorter path, a code with a map and location point is executed only one time. There is no need to execute the code in order to find the route for every event. This system is helpful in reducing manpower used in collecting waste from different locations using manual systems. However, the system has lacking with respect to power interruption.

2.5. SWC as a Service. An IoT-enabled solid waste management system is proposed in [55] for monitoring garbage bin and dynamic routing of the garbage collectors. The proposed system consists of an embedded device for real-

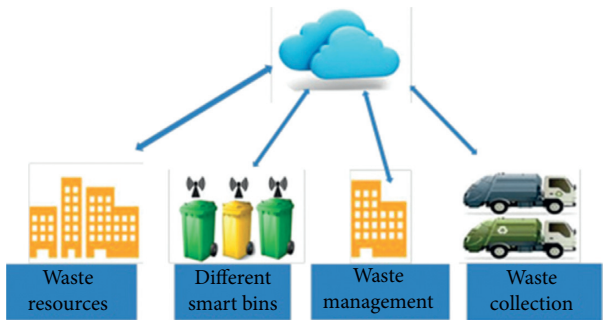


Figure 2: Architecture of IoT-based SWM.

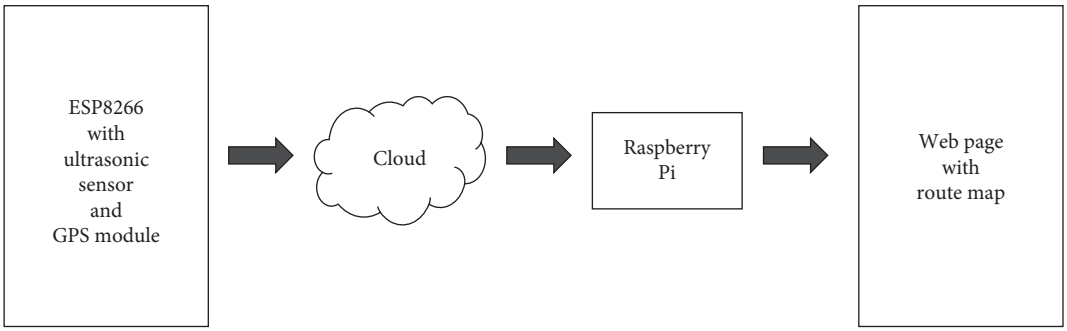


Figure 3: Architecture of the K-Query SWM.

time monitoring and scheduling of routes for garbage trucks. The architecture of the proposed system is presented in Figure 4. A mobile application is also designed for the truck driver to handle the data coming from garbage bins and further transmitting it to the cloud. In the proposed system, two garbage bins are installed in one place and solid waste is segregated from dry and wet garbage bins. The system is efficient in decision-making as it uses the GPS module and Google Map API for finding the optimal route to reach the garbage bins.

2.6. SCGCMS. In SCGCMS [56], a smart waste management and monitoring system is proposed for public waste collection that is based on IoT technology. The system consists of two phases where, in phase one, dustbins are installed in different locations and filled randomly while, in phase two, the route for collecting vans is decided on the basis of the dustbin filling ratio. The filling threshold is defined by the authors, which is 10 cm per dustbin. The system architecture of SCGCMS is shown in Figure 5. In this system, genetic algorithms are used for gathering waste. The dustbins are composed of a weight sensor and Raspberry pi Uno board that is connected with GSM modem and ultrasonic sensor for communications.

2.7. ML-Based WMS. With the rapid growth of IoT and its applications, various critical issues have appeared in todays' lifestyle. One of the most critical issues is waste management in urban areas. To reduce these types of issues, a waste management system [57] is designed for a campus of the Ton Duc Thang University in Vietnam, which is based on machine learning (ML) in the IoT environment. The authors used graph theory and ML that provide optimal path selection for waste collection on predicting the probability of garbage in trash bins. The proposed system is used for real-time monitoring through integrating multichoices, that is, ultrasound distance, E32 TTL-100 433 MHz with the LoRa spread-spectrum technology. Energy supply to each node in the network is provided by different sources, that is, solar and batteries. The proposed system for waste collection is better than the existing systems in terms of optimal path-finding and flexibility.

Table 1 presents the advantages and limitations of the surveyed schemes.

3. Proposed Methodology

In light of IoT technology, waste management is an important service that is supported by IoT. In today's time, waste management is a collective issue in most countries, which needs uninterrupted importance for management. In traditional waste management systems, the rapid growth of garbage leaves the public places unhygienic and dirty. The unhygienic environment can cause various deadly diseases. The prior research focused on the centralized system for waste management that is managed by a central authority. In this study, we are proposing a smart waste management system for real-time monitoring of "trash bins" in order to take timely actions for cleaning the bins and maintaining a disease-free environment for the people. The proposed system is based on edge-nodes, that is, trash bins. In this

Figure 4: Architecture of SWC as a service.

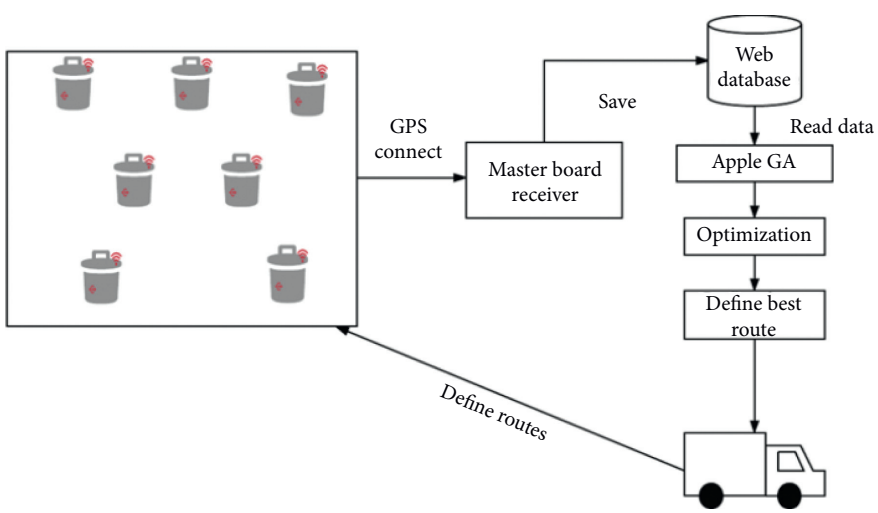


Figure 5: System architecture of SCGCMS.

system, a trash bin is working as an intelligent node in the entire processing of waste management. The smart bin mechanism is illustrated in the form of a block diagram in Figure 6.

3.1. System Design. The proposed SBM system is composed of three main entities, that is, trash bins (TB), trash collecting vehicle (TCV), and central database (CDB). These entities are part of SBM and their duties and characteristics are defined in the following subsections.

3.1.1. Trash Bin. In SBM, a trash bin is an intelligent edge-node and a waste storage point in public areas. The trash bin provides the following information to the central database of the system: level of the bin (TBL) in percentage, color of the bin (TBC), and weight of the bin (TBW). Each trash bin has

Table 1: Contributions and limitations of related schemes.

Scheme	Contribution	Limitation
SWMS [44]	Path optimization for garbage collecting van	Failure of sensors leads to system failure
GMS [52]	Fast transmitting mechanism for garbage collection	Lacking in the reliability of communications
IoT-based SWM [53]	Provision of statistical analysis	Lazy transportation affects all four entities
SWM by K-Query scheduling [54]	K-Query scheduling is used for database management	False monitoring can be a disadvantage for K-Query scheduling
SWC as a service [55]	Utilization of mobile app to facilitate van drivers	Bandwidth constraints of cloud can affect the mechanism
SCGCMS [56]	Uses genetic algorithms for collecting waste	Scalability issues can occur with complexity of the system
ML-based WMS [57]	Flexibility and optimal path selection using machine learning	Failure of batteries or interruption in solar provision

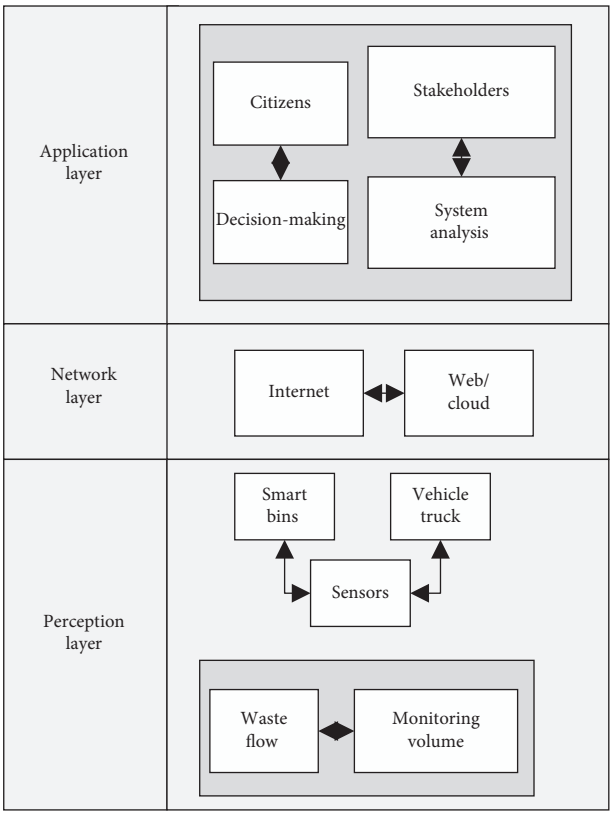


Figure 6: Block diagram of a smart bin mechanism.

its unique identification. Being a part of the IoT network, all trash bins are interconnected with each other through the Internet. Initially, each bin is green in color with *level* <*90%*. When the level of the bin increased to *level* >= *90%*, its color turns into red, which is the sign of a full bin.

3.1.2. Trash Collecting Vehicle. The vehicles that collect waste from the trash bins are dependent on the populated areas of smart cities. Mostly, smart cities are overpopulated which leads to difficulty in the collection of waste from densely populated areas using the same size of collecting vehicles. Each TCV is connected with the central database from where it gets information about the requesting TB. The TCV collects waste from the trash bins and brings it to the dumping zones for further treatment.

3.1.3. Central Database. The central database is used as an information center as well as a storage point, which contains each and every detail of the TBs, TB-IDs, TCVs, their locations, and every single detail about these entities. Whenever an event occurs in the system, the processing information and status of trash bins are stored in the database.

3.2. Transmission Pattern of the Proposed System. The whole mechanism of collecting waste is described in Figure 7, adopted from [56]. When a trash bin gets filled or reaches its threshold limit, it changes its color into red and transmits a notification to the TCV through a gateway. The TB notification consists of TBL, TBC, and TBW. The TCV receives the request from the TB and forwards the status of the requesting TB to the database for updating. The TCV collects waste from the filled bin for further treatment, such as disposing waste or recycling and reusing waste. The TCV updates the status of requesting TB in the central database after collecting waste from the bin. The whole processing of the system is supported by Reduce, Reuse, and Recycle mechanisms.

3.3. Processing of the Trash Bin. The trash bin checks the level of waste. If the *level* >= *90%*, the TB changes its color into red and forwards a request to TCV for the cleaning process. If the *level* <*90%*, the TB color remains green, and without forwarding request, it rechecks its level and so on.

Generally, the proposed system is a repetitive mechanism that consists of the following steps: collecting waste, planning and analysis, segregating waste at the waste plant, and recycling or disposing of waste. The hardware structure of SBM consists of TBs, which are installed at different locations of the city having their unique IDs. At the initial level, each TB is green in color, while the weight and level of TBs are recorded accordingly. Once a TB reaches its threshold level, the color of that particular TB will change into red with obvious measurements of weight and level of waste in the percentage form. The TCV is another significant

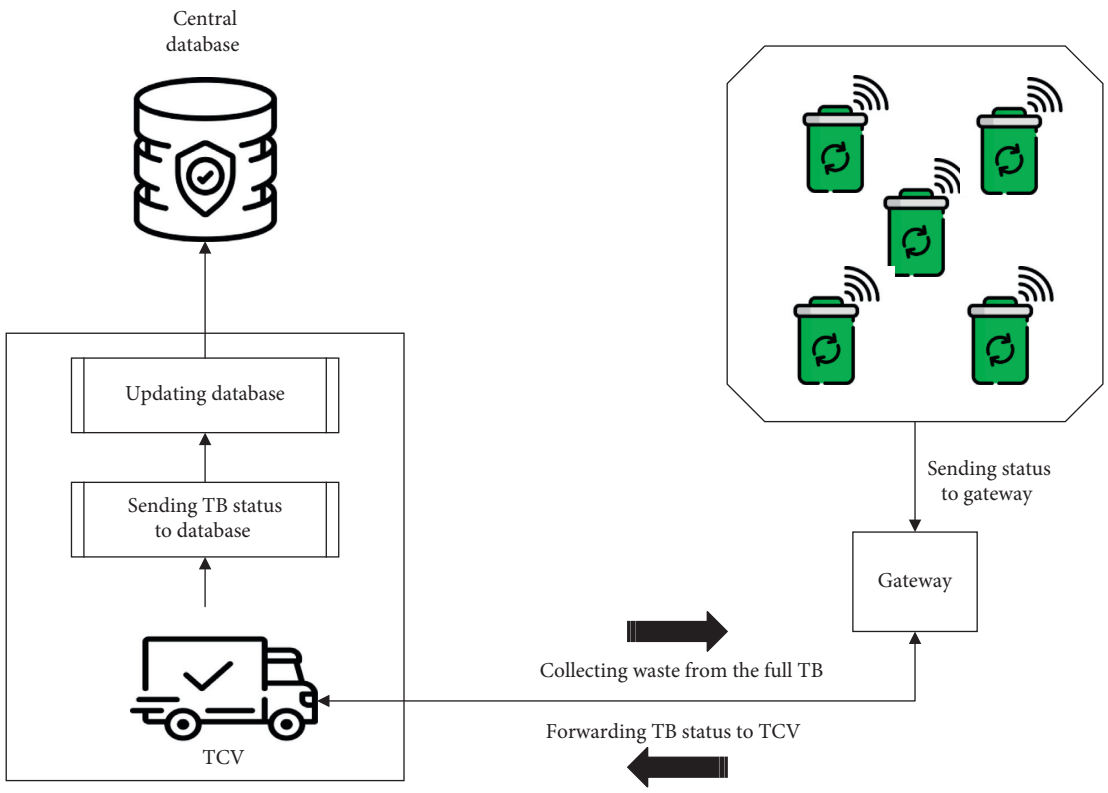


Figure 7: Transmission pattern of SBM.

factor in the proposed system. The TCV is informed by the filled TB to collect waste from it and updates the database with the new status of the requesting TB. The CDB is the central database in the SBM model. The duty of CDB is to store data regarding the location of the TB and TCV with the optimal route for waste collection. In SBM, for each event that occurred in the processing of waste collection and management, data for each process is also updated on the cloud. Eventually, the cloud contains information about each single event of the entities that are participating in the system.

Working of the SBM is described in the form of flow diagram, as shown in Figure 8. The main function of the proposed system is "trash to cash" that is based on three concepts, that is, Reduce, Recycle, and Reuse. "Reduce" refers to minimizing the amount of waste in the smart city as a particular situation of SBM. "Recycle" refers to recovering or reprocessing of dumped material that is extracted from the trash bins. "Reuse" refers to utilizing the dumped material after its recycling process. The processing of the system is started when an event is created in the trash bin. A threshold limit is set for each TB that helps easily accessing the bin and starts the cleaning process for quick and fast service provisioning.

3.4. Trash Bin Control Using Fuzzy Logic Processing. The fuzzy set theory plays a significant role in real-time scenarios to make decisions. Fuzzy logic was first introduced by Zadeh in 1965 [58]. The fuzzy logic is beneficial to deal with vagueness and uncertainty in real-time monitoring of the environment.

The fuzzy expert system (FES) is composed of three fundamental steps: fuzzification, inference rules, and defuzzification. A fuzzy expert system is the combined form of membership functions, if-then rules, and fuzzy operators. Moreover, FES is a mapping between the input and output values. Fuzzification converts crisp input values into fuzzy input, fuzzy rule base and/or knowledge base apply appropriate if-then rules, and defuzzification reconverts the fuzzy output into crisp output or in human readable form. The basic architecture of FES is presented in Figure 9.

In this study, fuzzy logic is utilized for decision-making in selecting appropriate locations to install trash bins. Fuzzy logic provides a better reading of the trash bin in real-time monitoring of the situation by using different levels of linguistic values. The linguistic variables have three levels of readings with different categories. We used three attributes for acquiring data of each trash bin, that is, TBL, TBC, and TBW, which indicate the trash bin level, color, and weight, respectively. Table 2 describes the input values of the trash bin level with categories low, medium, and high.

The input values for the attribute trash bin color are categorized into red, yellow, and green, and their ranges are defined in Table 3.

Table 4 shows the ranges of input values for trash bin weight with categories of light, medium, and heavy.

One attribute, that is, trash bin status (TBS), is used for the output readings of the trash bins. The TBS is classified

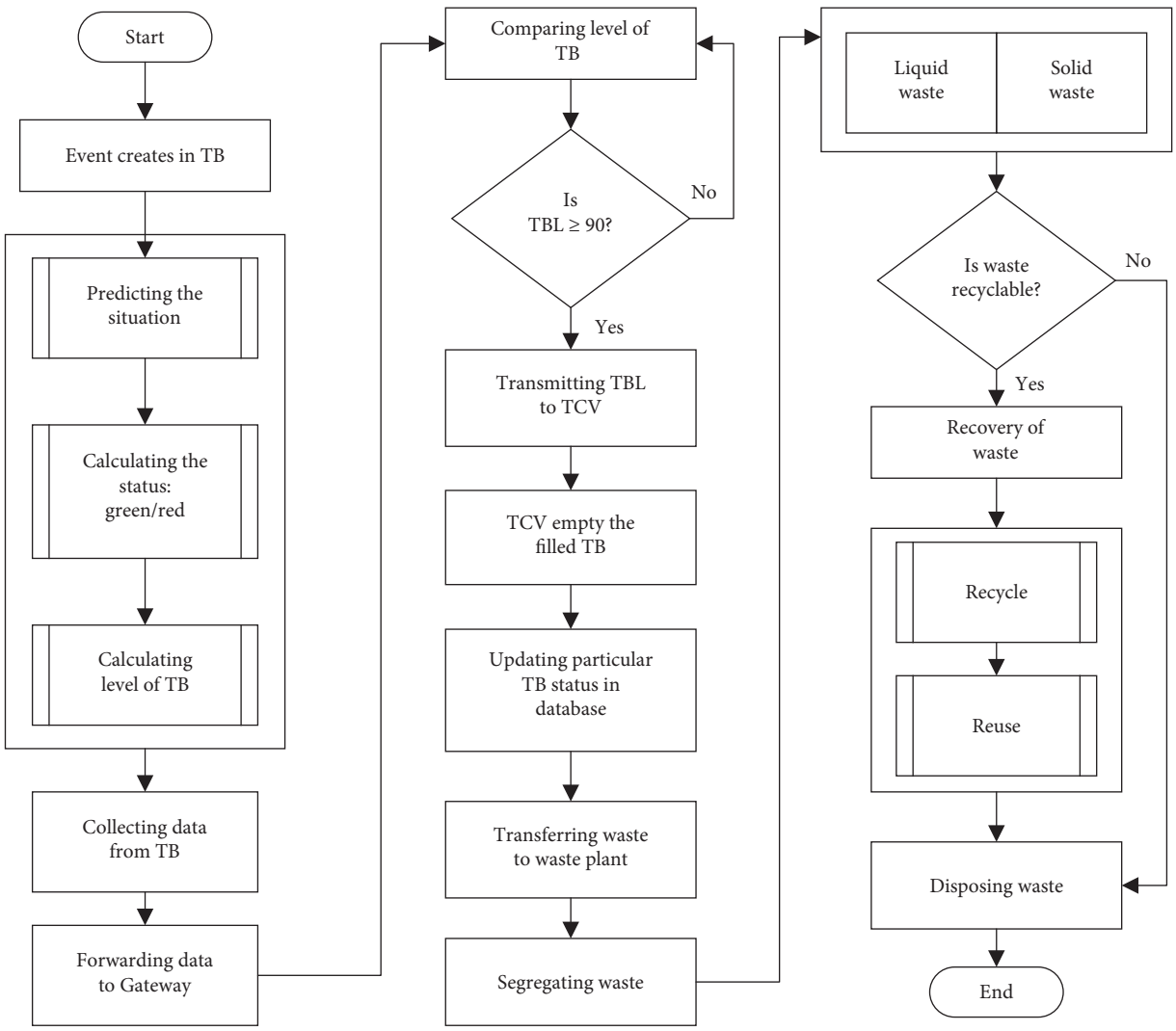


Figure 8: Functional flow diagram of the smart bin mechanism.

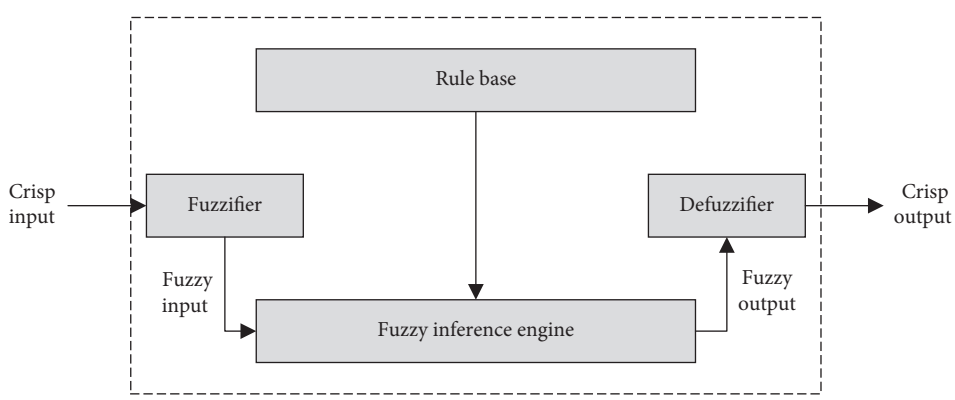


Figure 9: Fundamental architecture of FES.

into three stages, that is, bad, average, and good. The output ratings are described in Table 5.

Fuzzy logic helps in decision-making in such types of scenarios while selecting appropriate locations and size for installing trash bins. In this scenario, two attributes, that is, access to heavy vehicles and distance from collecting points, are used for choosing suitable places to install trash bins. Based on these two attributes, the system makes decisions and these are divided into different categories described in Table 6.

Table 2: Input ranges of the trash bin level (TBL).

Categories	TBL ratings	Symbols
Low	0.1 to 0.5	TBLL
Medium	0.5 to 1.0	TBLM
High	1.0 to 1.5	TBLH

Table 3: Input ranges of trash bin color (TBC).

Categories	TBC ratings	Symbols
Red	0.1 to 0.2	TBCR
Yellow	0.2 to 0.4	TBCY
Green	0.4 to 1.0	TBCG

Table 4: Input ranges of trash bin color (TBC).

Categories	TBC ratings	Symbols
Light	0.1 to 0.6	TBWL
Medium	0.6 to 1.2	TBWM
Heavy	1.2 to 2.0	TBWH

Table 5: Output ranges of TB status (TBS).

Categories	TBS ratings	Symbols
Bad	0% to 33%	TBSB
Average	34% to 66%	TBSA
Good	67% to 100%	TBSG

Table 6: Input values for selecting the site to install TB.

Criteria	Location 1	Location 2	Location 3
Access for heavy vehicles	Medium	Difficult	Easy
Distance from collecting points	Small	Medium	Large

These parameters are helpful in selecting the best suitable site in the city to install trash bins for the citizens. The proposed system is reliable in terms of providing consistent services to the municipal department for monitoring and cleaning the city in real-time scenarios. For decision-making, the discussed attributes are used for generating fuzzy rules, such as if-then rules. These rules are then integrated to form an output value, which is demonstrated in Figure 10. Two input variables are used to choose a suitable location for the TB installation, which gives a single output by applying inference rules.

4. Results and Discussion

We have performed simulations of the proposed framework in real-life experimental environment with different test runs of loading and unloading of trash bins. The NetLogo platform is used to implement the smart waste management mechanism by using different simulation cases in time T (minutes). At the initial stage, 20 to 25 trash bins are randomly distributed in a smart city with TBL = 0% and TBC = green. On tick 1, TBL = 10% and TBC remains green. When TBL = 90%, TBC turns into red that is an alarming situation from the bin to a vehicle and thus the bin makes a request to the nearby collecting vehicle for the unloading process. For tick 1, the simulation time $T = 0$. In the first case, 10 trash bins are distributed in time $T = 45$ minutes. Figure 11 shows the distribution of bins where the reading of each bin is recorded with different waste levels as per tick. The waste level in each trash bin is recorded against the total number of bins in the experimental environment. Each TB has some attributes, which are integrated to find the status/readings of each bin. The process is shown in the following equation:

$$\sum_{i=1}^{N} \mathrm{TB}_{(\mathrm{TB}_L+\mathrm{TB}_C+\mathrm{TB}_W)}. \quad (1)$$

For a test run, one TB is selected as a testing bin and with waste level or level of filling, different rounds are performed on that particular bin. In Figure 12, 8 different rounds for a TB are shown that are directly corresponding to the level of filling the bin. These records show different readings on throwing action of garbage in a bin. The bin-unload operation is performed when a bin is full to its level. In Figure 12, R denotes the readings of the selected testing bin.

The measuring weight level of a testing bin recorded with respect to time is described in Figure 13. It shows the weight of solid waste in a bin with regard to kg per time T. It shows the readings of throwing waste as per operation or interval. The weight of waste/garbage that is thrown in the TB is measured in kilograms against the measured time T in which a citizen throws it in the TB.

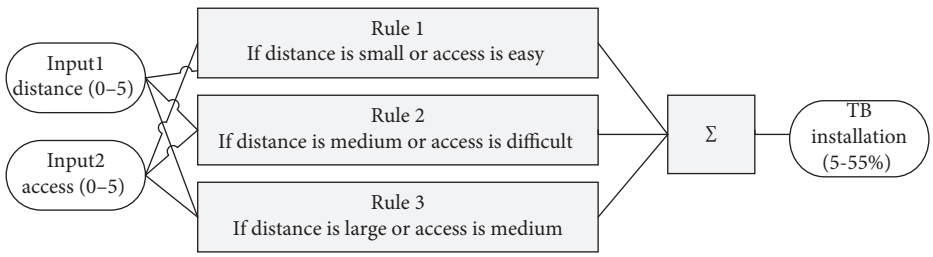


Figure 10: If-then rule to select appropriate locations for TB.

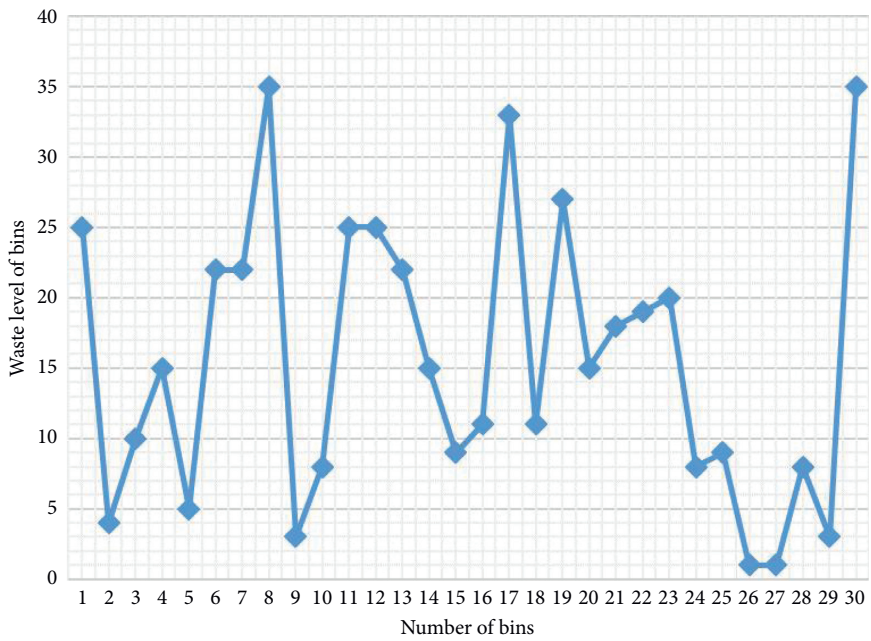


Figure 11: Readings of bins with different waste levels.

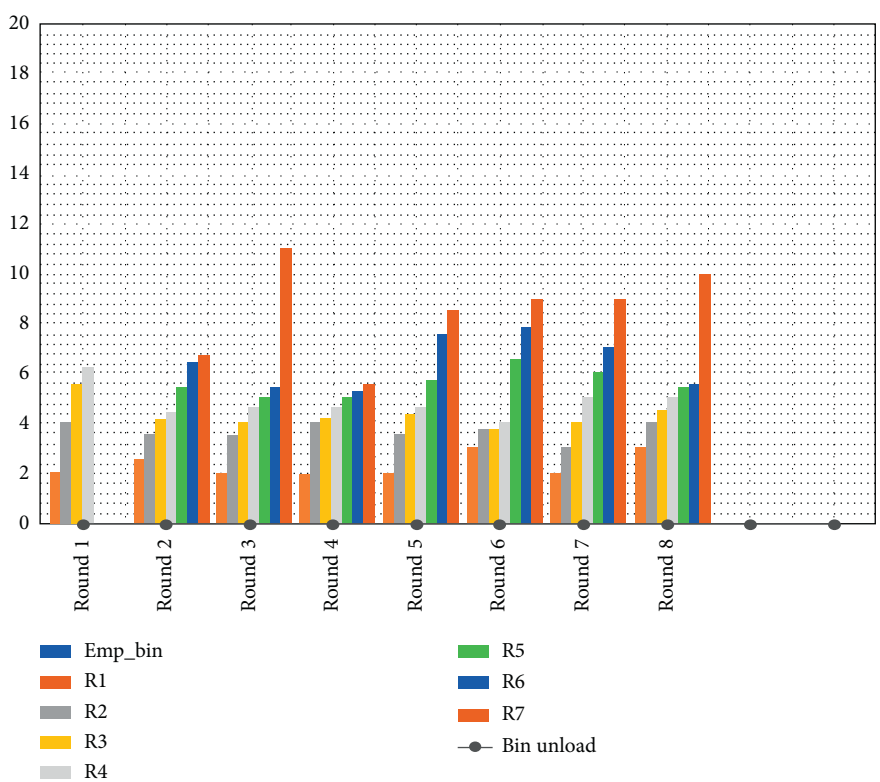


Figure 12: Readings of filling level with TB 1.

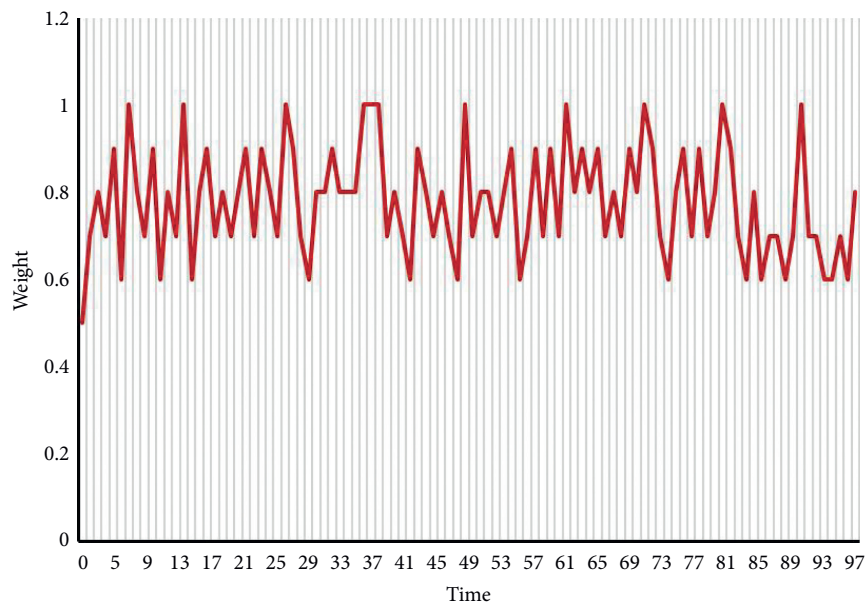


Figure 13: Measured weight of testing bin versus time T.

Two fuzzy parameters are used to decide an appropriate location for installing trash bins in the city. Fuzzy parameters are helpful in selecting the size of the trash bins depending on the density of the area, such as less, moderate, and high dense areas.

5. Conclusion

Conventional solid waste management systems have several shortcomings in terms of late unloading, hindrance in new techniques, lacking in throughput, less access to actual data, and many more. Therefore, an advanced approach is the need for the time to overcome all existing problems in the waste collecting process. Generally, waste collection has more consumption of cost from the municipality budget. In this paper, a real-time smart bin monitoring framework is proposed to get real-time access to data from the bins and implement the collecting procedure accordingly. The proposed framework is achieved by using a theoretical and architectural model. The model is implemented in a real-time environment of NetLogo and the experimental results show that the proposed framework is very responsive and effective for the environment. It is also effective for the economical aspects as it reduces the cost of labors and fuel cost of collecting vehicles by minimizing their extra visits in checking bins' status. Once a bin reaches its threshold limit, it informs the collecting vehicle for the cleaning process, which saves time, cost, and energy. The SBM is user-friendly as it obstructs the overflow of bins. It is useful for IoT-based smart cities, which helps to keep the environment clean and disease-free for the citizens. The SBM is supportive for real-time scenarios by using fuzzy logic processing in order to designate trash bins according to space and density of the environment in public areas. Fuzzy logic helps the system in selecting the best fitted site for each trash bin. Generally, fuzzy logic boosts the system for performing effectively in the environment.

6. Challenges and Future Work

Though the SBM framework designed for smart cities in the context of IoT has potentials, at the same time, it has the following challenges:

(i) Distribution of trash bins in the most populated areas where the amount of waste is unpredictable on daily basis

(ii) Disturbance in the Internet connectivity due to various causes, that is, weather disruption or defected connection

(iii) Lazy transportation: traffic jam could be a big challenge for vehicles to reach on time and collect garbage

(iv) Communication between two entities and damage of batteries could be severe challenges for the system

In the future, the model may be extended to an alternate and the shortest pathfinding for collecting vehicles in order to enhance transportation and remove collecting barriers. In addition, adding alternate sources for connectivity in case of power failure or weather hindrance may also be considered. Further, to facilitate the mechanism and save more energy, automated segregating TBs can be installed for dry, wet, and hazardous types of waste.

Data Availability

This research is based on simulations, which are performed in a simulator. Therefore, there is no dataset used in this research.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by King Saud University, Saudi Arabia, through Research Supporting Project no. RSP-2020/186.

References

[1] C. Srinivasan, B. Rajesh, P. Saikalyan, K. Premsagar, and E. S. Yadav, "A review on the different types of internet of things (IOT)," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 11, no. 1, pp. 154–158, 2019.

[2] K. Ashton, "That "internet of things" thing," *RFID Journal*, vol. 22, no. 7, pp. 97–114, 2009.

[3] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: a survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 414–454, 2014.

[4] L. Wang, Y. Ali, S. Nazir, and M. Niazi, "Isa evaluation framework for security of internet of health things system using ahp-topsis methods," *IEEE Access*, vol. 8, pp. 152316–152332, 2020.

[5] A. Tahir, F. Chen, H. U. Khan et al., "A systematic review on cloud storage mechanisms concerning e-healthcare systems," *Sensors*, vol. 20, no. 18, p. 5392, 2020.

[6] A. Ahmad, A. Ullah, C. Feng et al., "Towards an improved energy efficient and end-to-end secure protocol for iot healthcare applications," *Security and Communication Networks*, vol. 2020, Article ID 8867792, 10 pages, 2020.

[7] A. J. Jara, M. A. Zamora, and A. F. G. Skarmeta, "An internet of things-based personal device for diabetes therapy management in ambient assisted living (AAL)," *Personal and Ubiquitous Computing*, vol. 15, no. 4, pp. 431–440, 2011.

[8] M. Sikarndar, W. Anwar, A. Almogren, I. Ud Din, and N. Guizani, "Iomt-based association rule mining for the prediction of human protein complexes," *IEEE Access*, vol. 8, pp. 6226–6237, 2020.

[9] K. A. Awan, I. U. Din, A. Almogren, H. Almajed, I. Mohiuddin, and M. Guizani, "NeuroTrust-artificial neural network-based intelligent trust management mechanism for large-scale internet of medical things," *IEEE Internet of Things Journal*, In press.

[10] A. Almogren, I. Mohiuddin, I. U. Din, H. Al Majed, and N. Guizani, "FTM-IoMT: fuzzy-based trust management for preventing sybil attacks in internet of medical things," *IEEE Internet of Things Journal*, In press.

[11] R. Wirza, S. Nazir, H. U. Khan, I. García-Magariño, and R. Amin, "Augmented reality interface for complex anatomy learning in the central nervous system: a systematic review," *Journal of Healthcare Engineering*, vol. 2020, Article ID 8835544, 15 pages, 2020.

[12] S. R. Khan, M. Sikandar, A. Almogren, I. Ud Din, A. Guerrieri, and G. Fortino, "IoMT-based computational approach for detecting brain tumor," *Future Generation Computer Systems*, vol. 109, pp. 360–367, 2020.

[13] H. Chen, S. Khan, B. Kou, S. Nazir, W. Liu, and A. Hussain, "A smart machine learning model for the detection of brain hemorrhage diagnosis based internet of things in smart cities," *Complexity*, vol. 2020, Article ID 3047869, 10 pages, 2020.

[14] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, 2014.

[15] 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.

[16] S. D. T. Kelly, N. K. Suryadevara, and S. C. Mukhopadhyay, "Towards the implementation of iot for environmental condition monitoring in homes," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3846–3853, 2013.

[17] M. T. Lazarescu, "Design of a WSN platform for long-term environmental monitoring for iot applications," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 3, no. 1, pp. 45–54, 2013.

[18] S. Tozlu, M. Senel, W. Wei Mao, and A. Keshavarzian, "Wi-fi enabled sensors for internet of things: a practical approach," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 134–143, 2012.

[19] W. Ali, I. U. Din, A. Almogren, M. Guizani, and M. Zuair, "A lightweight privacy-aware iot-based metering scheme for smart industrial ecosystems," *IEEE Transactions on Industrial Informatics*, In presss.

[20] L. Foschini, T. Taleb, A. Corradi, and D. Bottazzi, "M2m-based metropolitan platform for IMS-enabled road traffic management in iot," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 50–57, 2011.

[21] K. A. Awan, I. U. Din, A. Almogren, M. Guizani, and S. Khan, "StabTrust—a stable and centralized trust-based clustering mechanism for iot enabled vehicular ad-hoc networks," *IEEE Access*, vol. 8, pp. 21159–21177, 2020.

[22] F. M. Malik, H. A. Khattak, A. Almogren, O. Bouachir, I. U. Din, and A. Altameem, "Performance evaluation of data dissemination protocols for connected autonomous vehicles," *IEEE Access*, vol. 8, pp. 126896–126906, 2020.

[23] S. K. Tayyaba, H. A. Khattak, A. Almogren et al., "5G vehicular network resource management for improving radio access through machine learning," *IEEE Access*, vol. 8, pp. 6792–6800, 2020.

[24] G. Jadoon, I. Ud Din, A. Almogren, and H. Almajed, "Smart and agile manufacturing framework, a case study for automotive industry," *Energies*, vol. 13, no. 21, p. 5766, 2020.

[25] Z. Ali, M. A. Shah, A. Almogren, I. Ud Din, C. Maple, and H. A. Khattak, "Named data networking for efficient iot-based disaster management in a smart campus," *Sustainability*, vol. 12, no. 8, p. 3088, 2020.

[26] K. Haseeb, I. Ud Din, A. Almogren, and N. Islam, "An energy efficient and secure IoT-based WSN framework: an application to smart agriculture," *Sensors*, vol. 20, no. 7, p. 2081, 2020.

[27] K. A. Awan, I. Ud Din, A. Almogren, and H. Almajed, "AgriTrust—a trust management approach for smart agriculture in cloud-based internet of agriculture things," *Sensors*, vol. 20, no. 21, p. 6174, 2020.

[28] W. Ali, I. U. Din, A. Almogren, and N. Kumar, "Alpha: an anonymous orthogonal code-based privacy preserving scheme for industrial cyber physical systems," *IEEE Transactions on Industrial Informatics*, In press.

[29] I. U. Din, M. Guizani, S. Hassan et al., "The internet of things: a review of enabled technologies and future challenges," *IEEE Access*, vol. 7, pp. 7606–7640, 2018.

[30] I. U. Din, M. Guizani, J. J. P. C. Rodrigues, S. Hassan, and V. V. Korotaev, "Machine learning in the internet of things: designed techniques for smart cities," *Future Generation Computer Systems*, vol. 100, pp. 826–843, 2019.

[31] K. Haseeb, A. Almogren, I. Ud Din, N. Islam, and A. Altameem, "SASC: secure and authentication-based sensor

cloud architecture for intelligent internet of things," *Sensors*, vol. 20, no. 9, p. 2468, 2020.

[32] K. Haseeb, I. Ud Din, A. Almogren, N. Islam, and A. Altameem, "RTS: a robust and trusted scheme for iot-based mobile wireless mesh networks," *IEEE Access*, vol. 8, pp. 68379–68390, 2020.

[33] B. Liao, Y. Ali, S. Nazir, L. He, and H. U. Khan, "Security analysis of iot devices by using mobile computing: a systematic literature review," *IEEE Access*, vol. 8, pp. 120331–120350, 2020.

[34] K. Haseeb, N. Islam, A. Almogren, and I. U. Din, "Intrusion prevention framework for secure routing in WSN-based mobile internet of things," *Ieee Access*, vol. 7, pp. 185496–185505, 2019.

[35] K. A. Awan, I. U. Din, A. Almogren, M. Guizani, A. Altameem, and S. U. Jadoon, "RobustTrust—a pro-privacy robust distributed trust management mechanism for internet of things," *IEEE Access*, vol. 7, pp. 62095–62106, 2019.

[36] K. S. Awaisi, A. Abbas, M. Zareei et al., "Towards a fog enabled efficient car parking architecture," *IEEE Access*, vol. 7, pp. 159100–159111, 2019.

[37] I. U. Din, A. Almogren, M. Guizani, and M. Zuair, "A decade of internet of things: analysis in the light of healthcare applications," *IEEE Access*, vol. 7, pp. 89967–89979, 2019.

[38] N. Islam, K. Haseeb, A. Almogren, I. U. Din, M. Guizani, and A. Altameem, "A framework for topological based map building: a solution to autonomous robot navigation in smart cities," *Future Generation Computer Systems*, vol. 111, pp. 644–653, 2020.

[39] B. Jan, H. Farman, M. Khan, M. Talha, and I. U. Din, "Designing a smart transportation system: an internet of things and big data approach," *IEEE Wireless Communications*, vol. 26, no. 4, pp. 73–79, 2019.

[40] C. Zhu, V. C. Leung, L. Shu, and E. C.-H. Ngai, "Green internet of things for smart world," *IEEE Access*, vol. 3, pp. 2151–2162, 2015.

[41] S. U. Islam, H. A. Khattak, J.-M. Pierson et al., "Leveraging utilization as performance metric for CDN enabled energy efficient internet of things," *Measurement*, vol. 147, p. 106814, 2019.

[42] K. Janjua, M. A. Shah, A. Almogren, H. A. Khattak, C. Maple, and I. U. Din, "Proactive forensics in IoT: privacy-aware log-preservation architecture in fog-enabled-cloud using holochain and containerization technologies," *Electronics*, vol. 9, no. 7, p. 1172, 2020.

[43] R. Liu and J. Wang, "Internet of things: application and prospect," in *Proceedings of the MATEC Web of Conferences*, p. 02034, 2017.

[44] S. Mahajan, A. Kokane, A. Shewale, M. Shinde, and S. Ingale, "Smart waste management system using iot," *International Journal of Advanced Engineering Research and Science*, vol. 4, no. 4, 2017.

[45] H. A. Khattak, H. Farman, B. Jan, and I. U. Din, "Toward integrating vehicular clouds with iot for smart city services," *IEEE Network*, vol. 33, no. 2, pp. 65–71, 2019.

[46] C. Balakrishna, "Enabling technologies for smart city services and applications," in *Proceedings of the 2012 Sixth International Conference on Next Generation Mobile Applications, Services and Technologies*, pp. 223–227, IEEE, Paris, France, September 2012.

[47] H. Khattak, Z. Ameer, U. Din, and M. Khan, "Cross-layer design and optimization techniques in wireless multimedia sensor networks for smart cities," *Computer Science and Information Systems*, vol. 16, no. 1, pp. 1–17, 2019.

[48] S. U. Khan, N. Islam, Z. Jan, I. U. Din, A. Khan, and Y. Faheem, "An E-health care services framework for the detection and classification of breast cancer in breast cytology images as an IoMT application," *Future Generation Computer Systems*, vol. 98, pp. 286–296, 2019.

[49] A. Siddiqua, M. A. Shah, H. A. Khattak, I. Ud Din, and M. Guizani, "ICAFE: intelligent congestion avoidance and fast emergency services," *Future Generation Computer Systems*, vol. 99, pp. 365–375, 2019.

[50] T. Anagnostopoulos, A. Zaslavsky, K. Kolomvatsos et al., "Challenges and opportunities of waste management in iot-enabled smart cities: a survey," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 3, pp. 275–289, 2017.

[51] O. A. Khan, M. A. Shah, I. U. Din et al., "Leveraging named data networking for fragmented networks in smart metropolitan cities," *IEEE Access*, vol. 6, pp. 75899–75911, 2018.

[52] M. H. A. Gawad, M. D. Katkoria, M. S. Kadam, and M. D. T. Jain, "Garbage monitoring system using IOT," *International Journal of Engineering Sciences & Research Technology*, vol. 7, pp. 319–324, 2018.

[53] N. Abdullah, O. A. Alwesabi, and R. Abdullah, "Iot-based smart waste management system in a smart city," in *Proceedings of the International Conference of Reliable Information and Communication Technology*, pp. 364–371, Springer, Kuala Lumpur, Malaysia, July 2018.

[54] F. M. Hadria, S. Jayanthy, A. Arunraja, and E. E. Vigneswaran, "Iot based smart waste management using top k-query scheduling," in *Proceedings of the 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 448–452, IEEE, Madurai, India, July 2018.

[55] S. S. Chaudhari and V. Y. Bhole, "Solid waste collection as a service using iot-solution for smart cities," in *Proceedings of the 2018 International Conference on Smart City and Emerging Technology (ICSCET)*, pp. 1–5, IEEE, Mumbai, India, January 2018.

[56] M. S. Chaudhari, B. Patil, and V. Raut, "Iot based waste collection management system for smart cities: an overview," in *Proceedings of the 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 802–805, IEEE, Erode, India, March 2019.

[57] T. Anh Khoa, C. H. Phuc, P. D. Lam et al., "Waste management system using iot-based machine learning in university," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 6138637, 13 pages, 2020.

[58] L. A. Zadeh, "A computational approach to fuzzy quantifiers in natural languages," *Computers & Mathematics with Applications*, vol. 9, no. 1, pp. 149–184, 1983.

Hindawi
Scientific Programming
Volume 2020, Article ID 8836927, 15 pages
https://doi.org/10.1155/2020/8836927

Research Article

An Intelligent IoT Based Healthcare System Using Fuzzy Neural Networks

Kashif Hameed,[1] **Imran Sarwar Bajwa**,[1] **Shabana Ramzan**,[2] **Waheed Anwar**,[1] **and Akmal Khan**[1]

[1]*Department of Computer Science & IT, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan*
[2]*Department of Computer Science & IT, The Govt. Sadiq College Women University, Bahawalpur, Pakistan*

Correspondence should be addressed to Kashif Hameed; kashif_hameed78@yahoo.com

Received 20 July 2020; Revised 10 November 2020; Accepted 10 December 2020; Published 28 December 2020

Academic Editor: Shaukat Ali

Healthcare facilities in modern age are key challenge especially in developing countries where remote areas face lack of high-quality hospitals and medical experts. As artificial intelligence has revolutionized various fields of life, health has also benefited from it. The existing architecture of store-and-forward method of conventional telemedicine is facing some problems, some of which are the need for a local health center with dedicated staff, need for medical equipment to prepare patient reports, time constraint of 24–48 hours in receiving diagnosis and medication details from a medical expert in a main hospital, cost of local health centers, and need for Wi-Fi connection. In this paper, we introduce a novel and intelligent healthcare system that is based on modern technologies like Internet of things (IoT) and machine learning. This system is intelligent enough to sense and process a patient's data through a medical decision support system. This system is low-cost solution for the people of remote areas; they can use it to find out whether they are suffering from a serious health issue and cure it accordingly by contacting near hospitals. The results of the experiments also show that the proposed system is efficient and intelligent enough to provide health facilities. The results presented in this paper are the proof of the concept.

1. Introduction

Internet of things (IoT) is a network in which many devices are connected, and these devices can communicate by computer network [1]. By this worldwide network, we can get information through sensors which relate to it. By using computer network, we can access this information anywhere in this world. Internet of things can connect physical objects to Internet and can provide opportunity of building systems which are based on various technologies such as near field communication (NFC) and wireless sensor network (WSN). In wireless sensor network, sensors sense the environment and send information to base station.

IoT has different methodologies such as smart dustbin, monitoring environment, IoT based irrigation system, smart healthcare system, and traffic control. In healthcare system, IoT brings gadget for monitoring health [2]. Health data can be accessed with the help of IoT by using sensors. Healthcare is a system which is used to improve health and help in treating diseases [3].

Health related issues/complications are increasing day by day, among which lung- and heart-related issues are top-listed. Health can be monitored by wireless technology, which is a modern concept. In wireless health monitoring systems, different technologies are used, including wearable sensors, portable remote health system, wireless communications, and expert systems. Life is precious; even a single life is also valuable, but due to lack of health facilities, awareness about diseases, and proper access to healthcare systems, people are dropping their lives. In all situations, Internet of things (IoT) helps in the indication of diseases and treatment of patients [4].

In IoT healthcare system, there exit wireless systems in which different applications and sensors are attached to

patients, information is obtained, and this information is forwarded to a doctor or specialist through an expert system [5]. Medical devices for Internet of things (MD-IoT) are remotely accessed, where devices are connected to the Internet and sensors, actuators, and other communication devices can monitor patient health [6]. Through these devices, the patient information and data are transmitted by the expert system via gateway onto a secured cloud based platform where the information is stored and can be analyzed.

In developing countries like Pakistan, telemedicine is used to handle health issues. Telemedicine refers to the practice of caring for patients remotely when the provider and patient are not physically present with each other. Telemedicine is simply defined as "the remote delivery of healthcare services." Although telemedicine brings with it many benefits, it has some downsides as well. Providers, payers, and policymakers alike know that there are some gray areas that are difficult to keep up with. While the field will grow exponentially over the next decade, it will bring with it both practical and technological challenges.

1.1. Unclear Policies. Because technology is growing at such a fast pace, it has been difficult for policymakers to keep up with the industry. There is great uncertainty regarding matters like reimbursement policies, privacy protection, and healthcare laws. In addition, telemedicine laws vary from state to state.

There are currently 29 states with telemedicine parity laws, which require private payers to reimburse telemedicine services in the same way they would reimburse in-person visits. As additional states adopt parity laws, private payers may institute more guidelines and restrictions for telemedicine services. Although it is a step in the right direction, there is still uncertainty regarding reimbursement rates, billing procedures, and more.

1.2. Fewer Face-to-Face Consultations. Several physicians and patients are finding it difficult to adapt to telemedicine, especially older adults. Physicians are very concerned about patient mismanagement. While advances in medicine have made it more efficient to use technology, there are times when system outages occur. There is also the potential for error as technology cannot always capture what the human touch can.

1.3. Technology Is Expensive. Healthcare systems that adopt telemedicine solutions can attest that they require a lot of time and money. Implementing a new system requires training, and sometimes staff members find it difficult to welcome this change. Practice managers, nurses, physicians, and more have to learn how to utilize the system so that practices can see the benefits. Although telemedicine is expensive in the beginning, healthcare systems should see a positive return on investment over time due to more patients and less staff.

The major components of healthcare systems are identification, location, sensing, and connectivity as shown in Figure 1. Smart healthcare is implemented through a wide range of systems: emergency services, smart computing, sensors, lab on chips, remote monitoring, wearable devices, connectivity devices, and big data.

The IoT based systems are equipped with body sensor networks within telemedicine systems. They include devices with special type of nodes that sense periodic difference of patient data; to check the ventilation conditions for the patients in rooms, sensors are used to collect data for different measures contributing to ventilation process of a room. These sensors are programmed to assess data of different ranges for temperature, pressure, humidity, and other significant environmental variables.

These arrangements help to monitor the patient conditions remotely. The system can send periodic reports to the hospital and maintain the patient history. The hospital staff can view the data and prepare the treatment plan for the patient under observation. The second type of devices used in IoT healthcare systems is based on wireless sensor networks. The situation is more complex than the above scenario in terms of remote area patient monitoring and management task. In some situations, IoT is the most reliable and cheapest solution, and the relationship between different devices and interactive communication systems also needs to be investigated with more formal objectives.

Technology makes it easier to monitor the patient health by sending information to healthcare teams such as a doctor, nurses, and specialists through IoT (Internet of things) and mobile technologies. It would be helpful for professionals to save and gather patient data using store-and-forward method so that it is accessible at any time. The role and services of IoT in modern healthcare are depicted in Figure 2.

Internet of things (IoT) has different methodologies: smart healthcare, traffic control, smart dustbin, and vehicle parking. The health of patient is monitored by screen, so it is difficult to examine the patient all the time. Therefore, here, patient's current status, i.e., pulse rate, temperature, position of body, blood glucose, and ECG, can be measured intensively by using sensors. The sensors are attached to Arduino UNO sensors that, when attached to the body Arduino board, get information and transmit it to the server. From this server, the information is forwarded to the doctor who advises for medicine.

Smart healthcare system is actually a technology in which treatments of patients are possible and can improve the standard of life [7]. In the concept of smart health, the e-health concept is also included which has commands on many technologies like electronic record management, smart home services, and intelligent and medical connected devices. Sensors, smart devices, and expert systems support the health practice for smart healthcare system.

Healthcare facilities in modern age are key challenge especially in developing countries where remote areas face lack of high-quality hospitals and medical experts. As artificial intelligence has revolutionized various fields of life, health has also benefited from it. The existing architecture of

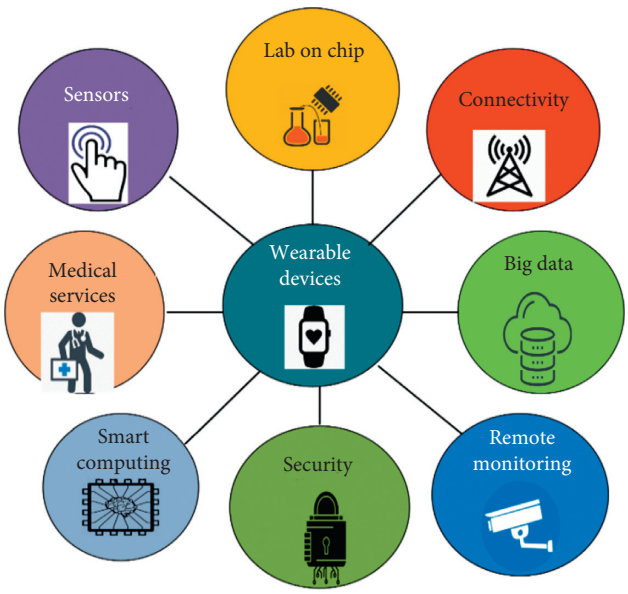


Figure 1: Typical components of IoT based smart healthcare.

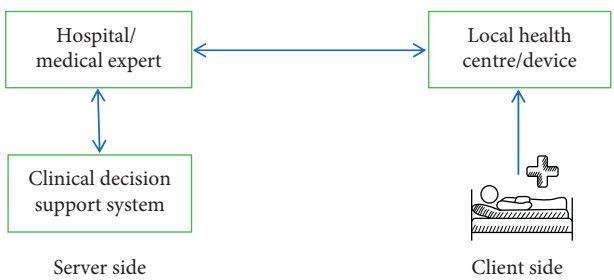


Figure 2: Structure of IoT in healthcare.

store-and-forward method of conventional telemedicine is facing some problems:

(i) Need for a local health center with dedicated staff.

(ii) Need for medical equipment to prepare patient reports.

(iii) Time constraint of 24–48 hours in receiving diagnosis and medication details from a medical expert in a main hospital.

(iv) Cost of local health centers.

(v) Need for Wi-Fi connection.

In this paper, a novel and intelligent healthcare system is proposed; it is based on modern technologies like Internet of things (IoT) and machine learning. This system is intelligent enough to sense and process a patient's data through a medical decision support system. This system is low-cost solution for the people of remote areas; they can use it to find out whether they are suffering from a serious health issue and cure it accordingly by contacting near hospitals.

2. Related Work

The IoT term was initially coined in 1999 and got attention of community as one of the advanced technologies as it is a combination of sensors [8], imbedded electronics, and software components for decision support systems [9]. The IoT also supports smart systems with the help of lightweight networking connections and sensors data [10]. The IoT covers almost all domains such as home based smart systems like security, entertainment, and health; transport systems like smart parking, traffic, logistics emergency services, and highway management; community based systems like smart metering, business intelligence, surveillance, environment, and retail systems [10].

Healthcare is one of the primary concerns that need to be improved by IoT and related technologies [11]. IoT based healthcare consists of three stages of automation, namely, data collection by sensors, analytics based on the collected data, and decision making based on the collected data [12]. Healthcare systems have a lot of potential to be improved by IoT based systems. There are many different types of healthcare applications proposed by the researchers like e-health [13], community health [14], home health monitoring [15], telemedicine [16], and clinical support for doctors [17]. The primary contribution of this research is to provide devices that help in monitoring, management, and communication between different stakeholders in the healthcare domain.

Table 1 shows the comparison of our work with related work. It shows that IoT based models provide much assistance to patients, but the time constraints can be reduced with the help of CDSS in the absence of medical staff.

The methodology of IoT for smart home, vehicle parking, and traffic control is different as compared to the health monitoring and management systems. The one obvious benefit is to monitor the patients 24/7 [26], which is almost impossible with manpower. The second goal achieved by IoT based solution is to monitor primary measures needs to determine the patient conditions, and treatment plan may include pulse rate, body temperature, respiratory rate, body position, blood pressure, ECG, and glucose level. These sensor networks are connected through Arduino board to collect the information through attached sensors. The collected information can be transmitted to the server and further refined for decision making or decision support systems.

Investigational experiments are made with the assistance of sensors, and the patient's health is traced with Internet. What remains is the keen observation of pulse rate, eco of heart, pressure level, temperature, etc. If there is any disturbance or change in pulse rate or temperature, the system alerts the person taking care of the patient. Through the Internet, the system shows the pulse rate and temperature of the patient.

The IoT with mobile technology provides smart and easier ways to look after the patients under observation, their body movements, and health conditions and provides intelligent mechanisms to handle and share the relevant information with relevant stakeholders. The study [11] designed a system which collects patient data and sends it to cloud for further utilization by people investigating health domain. The multipurpose application may also provide the families of patients with regular updates regarding patient

Table 1: Comparison between the related work and the proposed model.

Work	Technique	Local health center	CDSS	IoT	AI	24–48 hours
[18]	Cloud computing	✗	✗	✓	✓	✗
[19]	Automating design methodology	✗	✗	✓	✓	✗
[20]	Cloud computing	✗	✗	✓	✓	✓
[21]	Video streaming	✓	✗	✓	✓	✓
[22]	Cloud computing	✗	✗	✓	✓	✗
[23]	WSM	✗	✗	✓	✗	✗
[24]	Cloud computing	✗	✗	✓	✗	✗
[25]	Raspberry pi	✓	✗	✓	✗	✗

health. Ghosh et al. [26] demonstrated a system to automatically gather data from patients and store the gathered data into cloud for permanent use to help health professionals. The system also helps the guardians of the patients to know the health information.

The study [27] proposes a system to track the patient records with the measures of pulse rate, ECO, blood pressure, and body temperature and maintain the patient history. If the system detects any abnormal behavior in the measures observed, it immediately alerts the emergency team to handle the situation. The article [28] provides a survey on the smart healthcare. It discusses in detail the importance, application, requirements, and classification of healthcare along with the challenges, vulnerability, and security attacks. Healthcare system plays a vital role in increasing application by using connectivity technologies. The body sensor as a medical device is used to implement smart healthcare as shown in Figure 3. Smart telemedicine systems [15] are designed to monitor and manage the patient records by using sensors and microcontrollers. The system observes the body conditions and transmits the data to cloud servers. The patient condition is observed and stored on servers for further use and decision making.

The study [29] investigates the challenges and consequences of remote health systems. The system comprises wireless transmission system which collects ECG, body temperature, and pulse rate of the patients in remote locations for severe problems like cardiogenic shock. The patient is monitored, and data is sent to the doctor to analyze them and prepare the treatment plan. This data also helps the supporting staff to take the necessary actions [30].

The study [31] states that health monitoring system is essential for a good health because health problems are increasing day by day like cardiac failure, lungs failure, and heart related diseases. Nowadays, IoT became a platform for many services and applications in which sensor nodes are used. The monitoring of patients that is continuously done by doctors is the base of revealed consequences of generic health monitoring system.

The data analytics with big data enhanced the capacities of healthcare management system. The IoT healthcare is based on sensors, data collection devices, cloud services or connectivity provider devices, and mobile applications. The main concern of the physician is to separate the information of one patient from all other massive information of patients in the healthcare system. From such huge information, the physician makes critical decision about the patient health and suggests the treatment. He et al. concluded that altering patient information in real time is very important [32].

In order to build a smart system or application, the physical objects are connected by using IoT (Internet of things). The study used IoT for smart resource management system (SRMS) and intelligent chair system (ICS). An Arduino board is attached to the sensors, user ID is connected through RFID reader with the chairs, and chair allotment is managed and monitored by this system.

The study [33] investigates the influence of medical system with remote patients. The patient monitoring is the main purpose of the system with a prototype application. The main service provided by this prototyping system is the monitoring of vital signs of patient health in ICU. The system is more effective for patients undergoing surgical procedures or other treatments that need intensive care and monitoring. The major benefits claimed by the authors of the study are low power profile sensor, wireless communication, and gateway for cheap communication. The system is also available on web domain for the patient caretakers to observe and be informed about the patient status.

A smart health monitoring system [34] is designed and implemented for ambulance coupled with communication channels. The IoT was used with this smart healthcare system with the capability of low power sensors. The human body sensors are considered an efficient mode of communication for near field body sensor network application.

The study [35] presents a generic model for IoT based healthcare system. The model identifies key components with an end-to-end IoT healthcare system. The authors claim that there was no end-to-end IoT based remote patient monitoring system. The system consists of five sensors, three of which were for monitoring patient conditions like pulse rate, respiratory rate, and body temperature. The other two sensors were used to monitor blood pressure and blood oxygen. The paper also identifies technological challenges and potential opportunities for remote healthcare monitoring and management system with IoT.

The study [36] presents the applications and potential usage of IoT in healthcare systems. The major task performed by this experimental study was to monitor the patient conditions and make it possible to use more optimized and accurate medical equipment. The basic architecture of

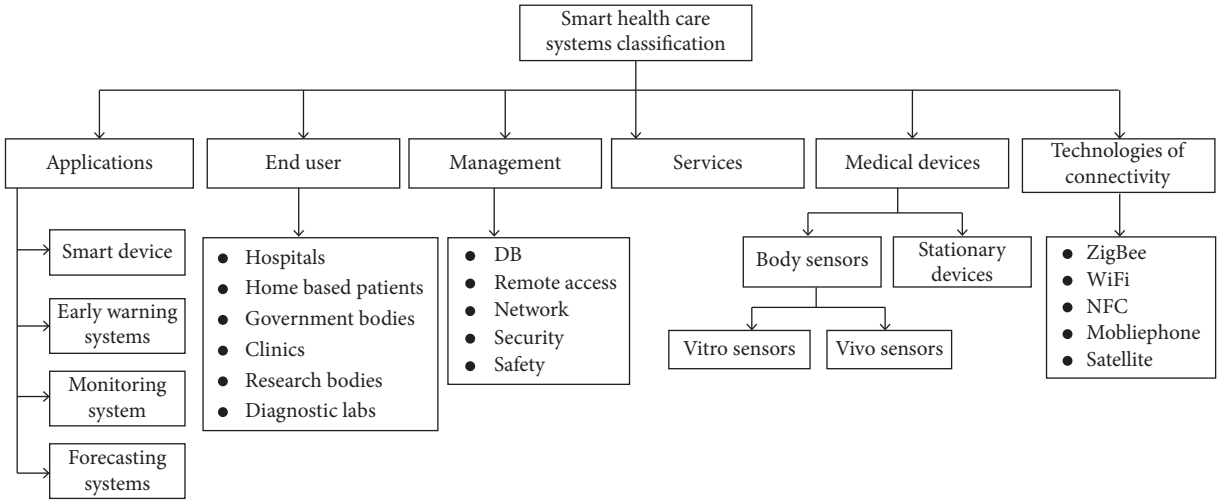


Figure 3: Classification of smart healthcare system.

the proposed system was based on sensor data and analyzes the patient data to make it possible to take basic decision for the purpose. The proposed system was collecting body temperature, respiration rate, and heartbeats and observing the patient body movements.

Wearable body devices [37] were used to design systems for monitoring and management of patients health. The study used body wearable sensors with IoT for smart patient monitoring and management. The focus of this study was to observe the patient health during surgical procedures, considering more useful and reliable data collection during such complex situations where human observational skills are not enough. The second major benefit was the reduction of equipment size for patient monitoring and wireless environment for patient care. Devices like Fitbit health monitor, Pebble smart watch, and Google glass are considered as modern devices for health monitoring and body care solutions. The important wearable devices are measuring the blood pressure which is used to assess the stress on a human mind. IoT imparts a valuable role to electronics and electrical devices to monitor and manage healthcare for humankind.

3. Architecture of Smart Healthcare System

The proposed smart healthcare system has the capability of decision making as per the observed conditions of the patient based on body temperature, pulse rate, and heartbeats. This architecture is also energy efficient solution because it does not turn on all the sensors all the time. The algorithm used in the system will handle the usage of the sensors and control their cost and lifetime. The proposed system addresses the issue of remote monitoring of patients and provides them with necessary treatment through experts in the hospital.

The smart healthcare monitoring and patient management system proposed in this study consists of communication channels, embedded internal and external sensors, IoT server, and cloud storage and is supported by a gateway. These activities are performed at different levels of refinement named application layer, management layer, network layer, and device layer. The architecture of the proposed system is presented in Figure 4.

The architecture shown in Figure 4 is revised to show more details. The use of sensors and decision support system in telemedicine is a novel idea that improves working performance of telemedicine in rural areas.

3.1. Data Collection through Sensors for Smart Healthcare System. With the help of IoT (Internet of things), the proposed system will be designed to implement a device in remote clinic. The device will take data of patient's heartbeats, body temperature, and blood pressure as input and will send it to the doctor concerned in the hospital. With the help of the data, the doctor will analyze the condition of the patient and will inform the remote area clinic crew about the necessary steps for patient's best treatment.

The architecture presented in Figure 5 shows physical view with necessary components of the proposed system. The system consists of three sensors: body temperature sensor, pulse rate sensor, and heartbeat sensors. These three sensors are connected through Arduino board to collect and classify the patient data. The data transmission is managed by communication and networking devices. The data analytics provides the decision-making facilities, and the fuzzy logic system is used in this arrangement to provide decision making. The doctor view provides the facility to hospital staff to monitor and communicate with the patient at remote place.

The next subsection explains the fuzzy logic system implemented in this smart patient monitoring and management system for decision making. The fuzzy system is placed at the server and it will order the decisions regarding patient conditions and treatment and alert the doctor about the situation of the patient. The system is fully automated. The last subsection gives the technical details and description of the proposed system.

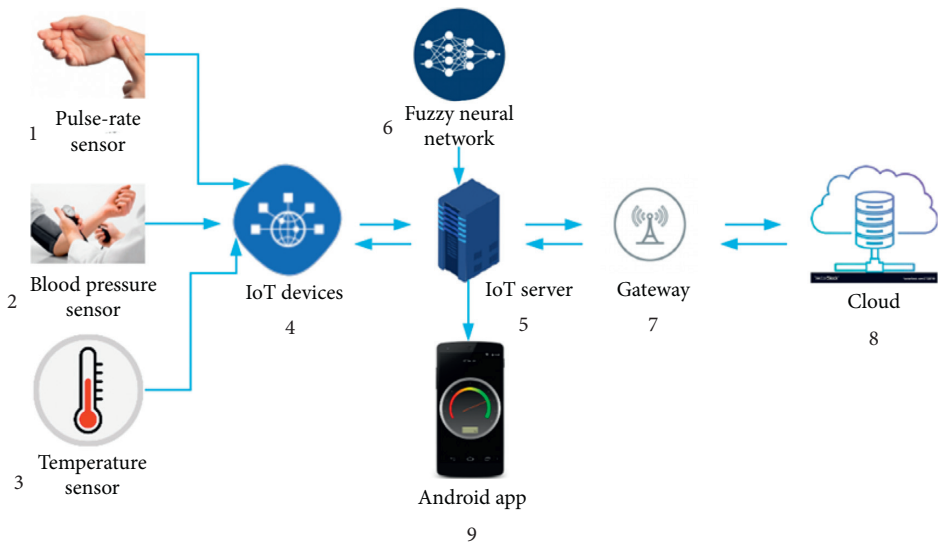


Figure 4: Smart healthcare system architecture.

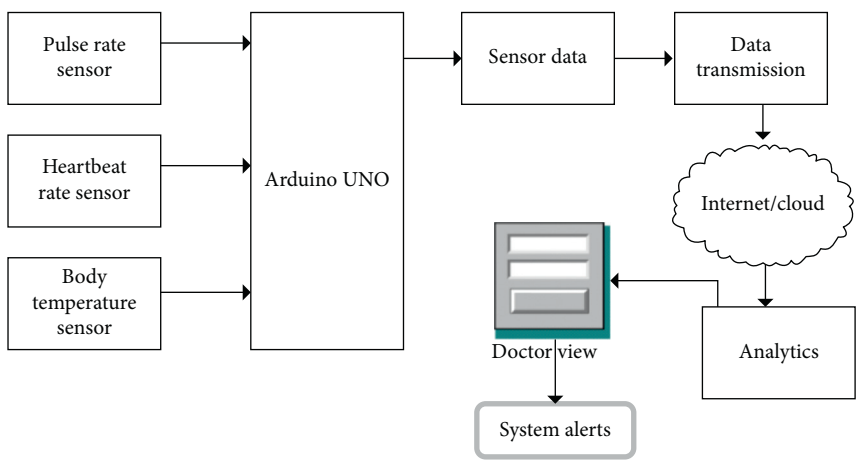


Figure 5: Diagram of monitoring patients in remote area clinics.

3.2. Fuzzy Logic-Based Smart Healthcare Monitoring and Management. There are the following problems: a single model is not enough, so two or more models are combined to solve that problem [38]. When different models are combined, they provide an effective solution to the problem, referred to as a hybrid system. A hybrid system is used to obtain indoor air quality using the fuzzy logic system and neural networks represented as the fuzzy neural network (FNN).

Neural networks focus on perceiving patterns, not on the logic of how the decision is made [39]. The fuzzy logic systems are good at explaining how the decision is made, but the inference rules are difficult task as prior knowledge is required [40]. These limitations lead to the fuzzy neural network. Rules of fuzzy systems are acquired from the neural networks patterns [41].

This process begins with a "fuzzy neuron," and the process of the fuzzy neuron is divided into two steps as follows [42]:

(i) Evolution of a fuzzy neuron model.

(ii) Development of the model and its algorithm that consolidate fuzziness into the neural system.

Figure 6 indicates that neural inputs are provided for neural network that provides neural outputs. Neural outputs are the inference rules for the fuzzy interface that are stored in the system as a database and used for decision making and provide learning algorithms for the neural network as prior knowledge. Data of neural networks is gathered by propagation algorithm, so the procedure is slow. Including specific data into the neural network to clarify learning techniques is a difficult task. Fuzzy rules are explained, and they provide better performance, so fuzzy systems are used in restricted systems and knowledge acquisition is a difficult task. To solve these problems in solution design, the fuzzy rules are designed from numerical data.

The neural network model named Approximate Reasoning Intelligent Control (ARIC) (see Figure 7) uses fuzzy neuron system. This fuzzy neuron system is trained by physical system forecast. It applies a fine-tuning refreshing data method to control the information base.

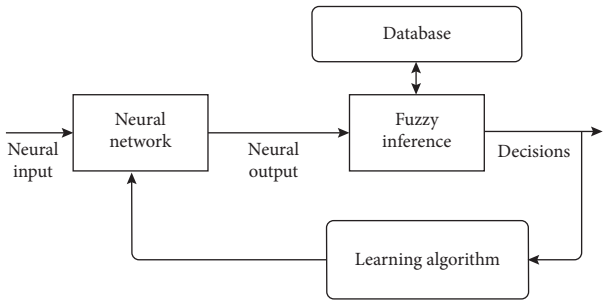


Figure 6: Model of fuzzy neural network.

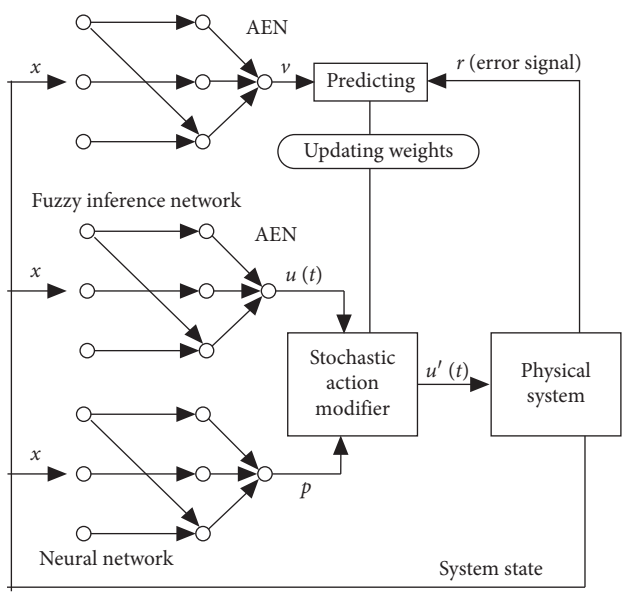


Figure 7: Berenji's ARIC architecture [43].

This is a perfect combination of fuzzy system with neural networks which boost the advantages of both decision-making methods. The framework can learn, and information utilized in the framework has the type of if-then fuzzy system. The rules are defined in advance, and the system can start without outside help, so it adapts quicker than a standard neural system. The framework named ARIC consists of AEN action position which is used to evaluate the network constructed through information base. The ARIC also contains ASN operation used for network selection. It is a multilayer neural network technique with fuzzy control system. The ASN component has two separate fuzzy interfaces in the first layer of the proposed system. The neural network is placed in the second layer. The neural system finds $f\ [a,\ a+1]$ operation, a part of confidence acquired through fuzzy inference. It should gain $p\ (a+1)$ using the amount of time denoted by t and the condition of framework $t+1$. A modification module which is stochastic in nature improves the control with p(t) of fluffy part and the expected likelihood regrading decisions and produces the final output.

$$u'(t) = r(u(t), \mathrm{d}[t, t+1]. \tag{1}$$

The unit c_i of fuzzy inference is organized to assess the fuzzy guideline. The unit for information a_j is a standard ancestor and acquires a unit u. The unit u communicates with the activity control. It is called defuzzified mixture which finishes the process. The information layer in this framework is fuzzified; it is monotonic in nature and has the capacities to utilize its components in ARIC model. The fuzzy tag is used in rules to balance local standards. The ancestor's enrolment is estimated by these standards and then duplicated and uses load joining with the association of information component. The qualities base of this system produces the final input. Each unit which was obscured is exceptional monotonic work communicating with final standard. The monotonicity of this function yields the output. The process is effortlessly determined by the opposite capacity. This esteem is produced with the function of heaviness and with the association of hidden unit. The yield value is finally determined by weighted average method.

The action operators used to evaluate the network, which tries to forecast the model activity. The neural netowrk method used in this system is a typical feedforward neural network system. This feedforward neural network system is based on shrouded layer which collects the model states as information. It uses the blunder flag r from the physical model as a piece of helping data. The process gets $v[t, t]$ of the proposed system produced as a forecast for future. This system relies on load of time t and the model constraints. The t is either t or $t+1$. The conditions in this system are portrayed by fortification of higher values for information collected for decision making. The change in load is managed by support method that uses the output of the system states of the network and action state evaluation method. The engineering of ARIC was connected to postadjustment. It is also demonstrated that the model with the comprehension and its assignments.

The signs and weights are real numbers with input neurons. The information does not change these signs. The yield is very much equivalent to the information. The signal a_i may collaborate with the load to w_i to construct these items.

$$d = w_i a_i, \quad i = 1, 2. \tag{2}$$

The input data d_i is collected, by addition, to deliver the information,

$$\text{net} = d_1 + d_2 = w_1 a_1 + w_2 a_2, \tag{3}$$

to the neuron. The neuron utilizes its exchange work f, which could be a sigmoidal function $f(x) = (1 + e^{-x})^{-1}$, to figure out the output:

$$y = f(\text{net}) = f(w_1 a_1 + w_2 a_2). \tag{4}$$

This basic neural net, which utilizes duplication, addition, and sigmoidal f, will be called an ordinary neural net as shown in Figure 8.

In this event, we utilize different activities like a t-norm, or a t-conorm, to join the approaching information to a neuron; we get what we call hybrid neural net. These modifications lead to fuzzy neural engineering dependent on fuzzy arithmetic tasks. This gives us a chance to express the sources of info a_1, a_2 and the weights w_1, w_2 over the unit intervals [0, 1]. The immediate fuzzification of regular neural

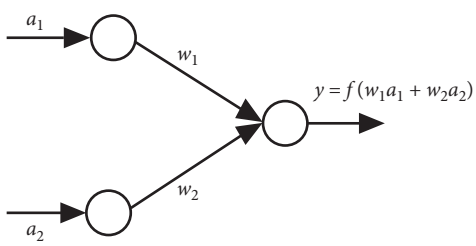


Figure 8: Neural net.

systems is to broaden association weights and additional inputs to fuzzy numbers as shown in Table 2.

A set of fuzzy rules were defined for the clinical decision support system used for IoT based telemedicine. These rules are based on the facts and fuzzy data shown in Table 2. Following are a few examples of fuzzy rules defined.

IF (Temperature == High) AND (Pulse_Rate == Low) AND (Blood_Pressure == Very_High)

THEN Decision = High

IF (Temperature == High) AND (Pulse_Rate == Low) AND (Blood_Pressure == High)

THEN Decision = High

IF (Temperature == Normal) AND (Pulse_Rate == High) AND (Blood_Pressure == Medium)

THEN Decision = Low

IF (Temperature == Low) AND (Pulse_Rate == High) AND (Blood_Pressure == Medium)

THEN Decision = Low

IF (Temperature == Normal) AND (Pulse_Rate == Normal) AND (Blood_Pressure == Low)

THEN Decision = High

3.3. Implementation Details. A microcontroller board (Arduino) (see Figure 9), which has model number ATmega328, has 4 digital pins for input and output sources. The six i/o pins are PWM output. The microprocessor has 16 MHz with a power jack, USB connection. The other components on this microcontroller chip are analog input and reset button with ICSP header. The power is supplied by a USB interface, and Arduino is designed as open electronic platform. The basic settings on Arduino board are input/output, set/reset button, sensor lights, and activating motor with output LED.

HC-05 Bluetooth module: To add wireless functionality of two ways (full duplex) to your project, HC-05 is very cool module. If communication is required between two microcontrollers, Bluetooth module is used as Arduino and can communicate with any device with the functionality of Bluetooth like a laptop or a phone. Bluetooth SSP (serial port protocol) module is designed for wireless transport. HC-05 can be used in a master or slave configuration that will be great solution for wireless communication.

Temperature sensor (see Figure 10) is used to detect heat stroke, body temperature, and fever. In wearable healthcare

Table 2: Direct fuzzification of neural network.

Fuzzy neural net	Weights	Inputs	Target
Type 1	Crisp	Crisp	Fuzzy
Type 2	Fuzzy	Crisp	Crisp
Type 3	Fuzzy	Fuzzy	Crisp
Type 4	Crisp	Fuzzy	Crisp
Type 5	Crisp	Fuzzy	Fuzzy
Type 6	Fuzzy	Fuzzy	Fuzzy
Type 7	Fuzzy	Crisp	Fuzzy
Type 8	Crisp	Crisp	Fuzzy
Type 9	Fuzzy	Fuzzy	Crisp

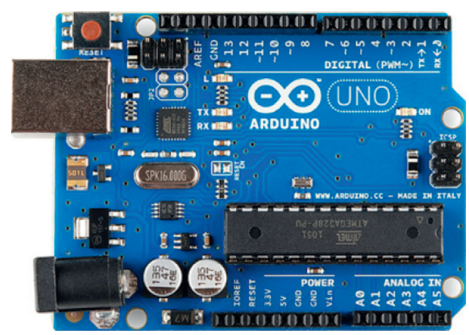


Figure 9: Arduino board.

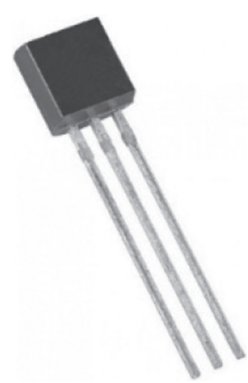


Figure 10: LM-35, analog temperature sensor.

system, body temperature is used as a diagnostic tool. For the measurement of body temperature, thermistor type sensors are used. Temperature sensing accuracy is limited.

The temperature sensor (see Figure 11) is integrated circuit which is used to measure the body temperature in centigrade. The temperature is shown as voltage output. The model number of this sensor is LM35. This model of body temperature sensor is considered better in performance than linear temperature sensor. The reason is that user need not convert Kelvin scale to centigrade scale by using this model. The sensor under this setup is very useful for remote sensing and calibrates Celsius scale.

The emergency conditions are measured through cardiac arrest, pulmonary embolism, vasovagal syncope, and pulse sensor. The pulse rate is primary measure for critical medical conditions and body fitness conditions. The pulse rate sensor is the most used and researched sensor in patient care and management domain. It is used to assess heartbeats and

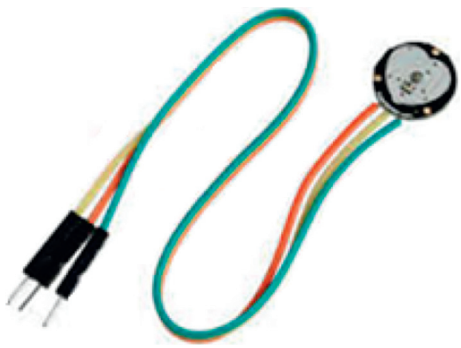


Figure 11: Pulse rate sensor.

Figure 12: Blood pressure sensor.

complex diseases like heart attack. The sensor works when the object places finger on input panel. The output is detected on output panel. The power required for this sensor is 5 volt direct current. The working principle of this module is based on blood flow rate through finger. The heartbeat sensor normal reading was 60–100 bpm. Figure 12 shows the used blood pressure sensor to measure the blood pressure of the patient and record it in an Excel sheet for further processing.

4. Experimental Results

4.1. Experiment Setup. The system is tested under the supervision of medical staff. Samples are collected from different areas of South Punjab using the proposed device. The data collected through sensors was forwarded to the server. The results are presented at the Arduino application and web browser. Table 3 shows the information about locations that we selected to test the proposed model. Almost eight different locations are selected for testing. The distance from BVH and testing period of selected locations are different.

4.2. Dataset. Table 4 shows the report sample of the patient that is generated on the server after receiving data collected through sensors and forwarded through smart device. The report has three sections: patient's data, sensor data, and symptoms of the patient.

Table 5 shows the comparison of response time of the queries that are responded to by CDSS and by physician. Almost 270 queries were received on the server from selective areas. Most of the queries were treated by CDSS. Table 5 clearly shows that average response time of the queries that are responded to by CDSS is quite short as compared to the response time of the queries responded to by physician. The proposed system is low-cost and efficient solution for the people of remote areas; they can use it to find out whether they are suffering from a serious health issue and cure it accordingly by contacting near hospitals.

Using sensors and decision support system in telemedicine is a new idea, and Table 5 shows how it minimizes time constraint in comparison to the classical telemedicine method.

4.3. Used Tools and Data Analysis. The use of analytics potentially improves the accuracy and permits early disease detection, personalization, and cost reduction in medical facilities. The following set of tools and libraries were used to process and interpret patient's symptoms and health data.

(i) Fuzzy neural networks based clinical decision support systems
(ii) A set of three sensors to gauge patient's health data
(iii) A GUI for recording input of patient's symptoms
(iv) An Android mobile application for user interface

The lab measurements and calculations are the primary concern and are important for current medical practice. On the other hand, wearable sensors have many advantages over lab and office measurements due to radial incorporation of multiple physiological measurements. This flexibility makes it possible to gather data. It is required with greater temporal

Table 3: Experiment setup details and data collection.

Serial #	Location ID	Distance from hospital in km (BVH)	Selective sample	Testing period
1	Loc-3	15	3	Nov 2019
2	Loc-1	35	4	Oct 2019
3	Loc-6	47	3	Jan 2020
4	Loc-7	25	2	Feb 2020
5	Loc-4	45	1	Dec 2019
6	Loc-2	60	3	Jan 2010
7	Loc-8	53	1	Aug 2019
8	Loc-5	22	2	Dec 2019

Table 4: Patient report.

Patient data
Patient name: ABC Patient CNIC: 3120245627438 Patient address: XYZ
Sensor data
Body temperature: 99°F Pulse rate: 76 BPM Blood pressure: 90/130
Symptoms
(i) Headache (ii) Shortness of breath (iii) Dry mouth (iv) Weight loss (v) Fever

Table 5: Comparison of response time.

Patient ID	Queries	Average response time (in hours)	No. of queries (CDSS response)	No. of queries (physician response)
P-1	30	3	27	3
P-2	43	13	11	32
P-3	61	7	50	11
P-4	38	10	25	13
P-5	42	6	28	14
P-6	55	5	50	5

sampling and longer longitudinal time scales. This arrangement provides vast and valuable opportunity for data analytics and machine learning methods. The machine learning algorithms identify correlations between data and clinical diagnoses trends.

5. Results and Discussion

The smart healthcare patient monitoring and management system is designed as intelligent system. The proposed system benefited from fuzzy logic system which is easy to use and implement for decision making. The organization of the proposed system is quite new by using sensors data and fuzzy based decision making. The implementation details are already presented in a previous section with the hardware used for this proposed system. The data collected through sensors was forwarded to the server. The results are presented at the Arduino application and web browser. The user may perform some actions against the information presented by the system. The three types of sensor data received from the sensors are further processed into output by fuzzy logic system. The classification is shown in Table 6; there were four classes of temperature measure (no fever, fever, high fever, very high fever) detected by different temperature ranges from 100°F to 105°F.

Table 7 represents the three classes of pulse rate for a normal human being which are low, normal, and high. The pulse rate less than 60 per minute is considered as low pulse rate. The pulse rate between 60 and 100 is considered as normal pulse rate. The pulse rate greater than 100 is considered as high pulse rate.

Table 8 represents the normal to abnormal range for blood pressure. The blood pressure 120/80 BP is considered as normal blood pressure. The blood pressure 129–140/81–89 BP is considered as high blood pressure. The blood pressure greater than 141/91 BP is considered as very high blood pressure.

Table 9 represents the data collected through sensors at different intervals of a patient. The data ranges are also calibrated by Tables 6–8

Figure 13 represents the variation in data collected through temperature sensor, pulse rate sensor, and blood pressure sensor. The data ranges are also calibrated from Tables 6–9.

The input data is collected and calibrated; in the second step, fuzzy logic is applied for the decision making for the determination of patient condition. Table 10 represents the calibrated output values for the inputted data.

The fuzzy logic system takes the decision, and accuracy of the decision is measured (Figure 14). It is shown in Table 10 that accuracy of the system is from 94% to 100% for the proposed system. It shows that the proposed system is working as per the rules defined for the decision making of patient care and management system. The accuracy of the proposed system is measured by the formulae in

$$\text{Accuracy} = \sum \frac{\mu(\text{ai})}{n}. \tag{5}$$

The accuracy of the proposed system is calculated by (5). In (5), $\mu(ai)$ is the accuracy in the percentage for the data in experiment, and n is the number of experiments. The average accuracy achieved in this dataset is 97%.

The experimental results show that intelligent and smart decision making makes the sensor based IoT system convenient and feasible. The methodology used with IoT improves the performance and throughput of the system. The percent error of the results is calculated by using the formula shown in (6). Here, the accepted value is required accuracy, and experimental accuracy is achieved accuracy of the experiments.

$$\text{percent\_error} = \frac{\text{accepted\_value} - \text{experiemtnal\_value}}{\text{total\_value}} \times 100. \tag{6}$$

The data depicted in Figures 15 and 16 show the accuracy and reliability of the achieved results.

Section 2 discussed many systems consisting of sensors with IoT. The proposed system is the first patient care monitoring and management system which uses fuzzy logic system to determine the patient conditions and decide the possible treatments. The results in Tables 9 and 10 show that our approach performs better with the help of sensors and decision support systems. The fuzzy logic system decision making enhances the usefulness and the accuracy of the proposed system. This system is novel in terms of using intelligent decision making with sensor and IoT based system. The results show that the proposed approach is more accurate, time saving, cheap, and easy to use. The proposed system has the following contributions. This is the first approach that presents a smart irrigation system for tunnel farming.

Table 6: Temperature levels.

Temperature (°F)	Class
<99	No fever
99–101	Fever
101.1–103	High fever
>103.1	Very high fever

Table 7: Classes of pulse rate.

Pulse rate (BPM)	Class
>100	High/tachycardia
61 to 100	Normal
<60	Low/bradycardia

Table 8: Classes of blood pressure.

BP (HG)	Class
<110/<70	Very low
120–110/80–70	Low
120/80	Normal
130–139/80–89	High
>140/>90	Very high

Table 9: Sensor data for the experiments.

Sr. no.	Temperature (°F)	Pulse rate (%)	Blood pressure (BP-low)	Blood pressure (BP-high)
1	100	61–100	100	180
2	103	60	89	139
3	100	110	91	141
4	102	107	80	133
5	98	106	80	122

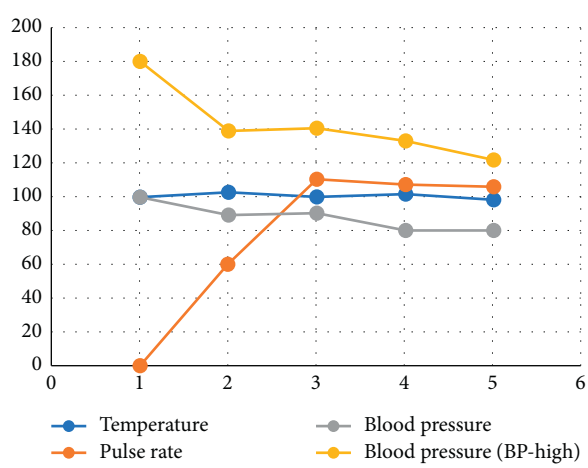


Figure 13: Data variation collected through sensors.

Table 10: Calibration of sensor data with fuzzy logic system.

Sr. no.	Temperature	Pulse rate	Blood pressure	Fuzzy logic decision	Accuracy (%)	Percent error (%)
1	High	Low	Very high	High	97.8	2.2
2	High	Low	High	High	94.6	5.4
3	Normal	High	Medium	Low	87.8	12.2
4	Low	High	Medium	Low	89.1	10.9
5	Normal	Normal	Low	High	93.6	6.4
6	High	High	Medium	High	97.6	2.4
7	High	Medium	Medium	Medium	94.5	5.5
8	Medium	Low	Medium	Medium	91.5	8.5
9	Very high	High	High	High	95.6	4.4
10	Medium	High	Medium	Low	87.9	12.1

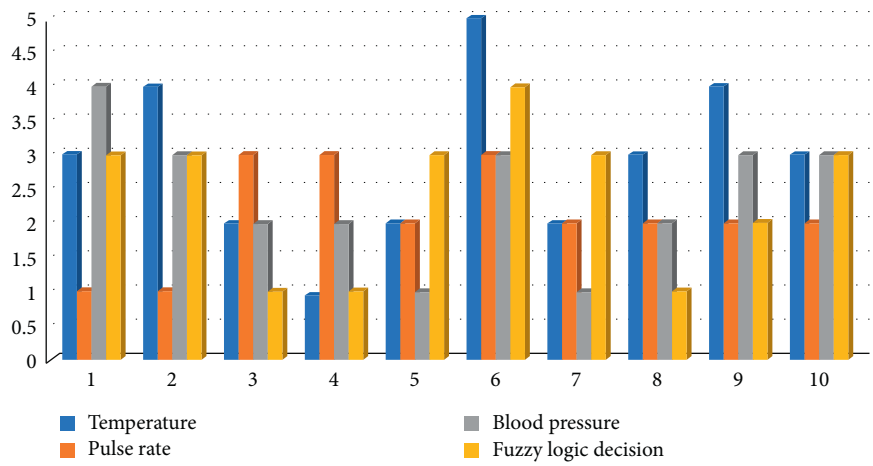


Figure 14: Results of fuzzy logic based decision making.

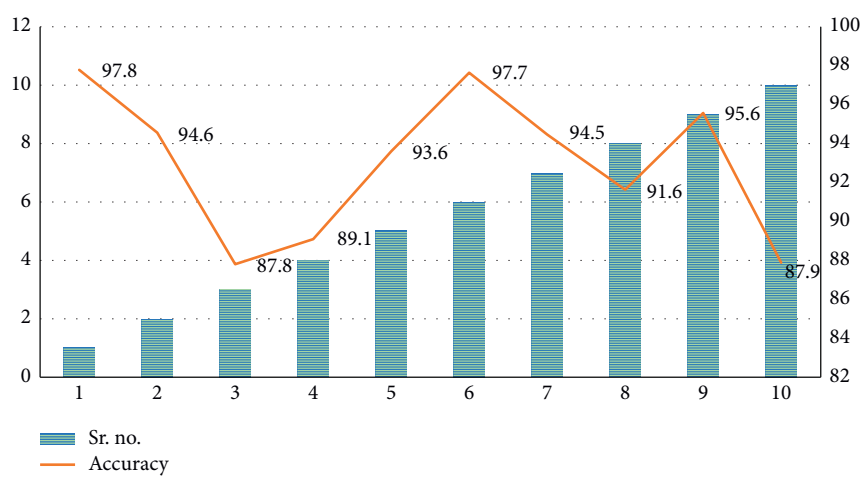


Figure 15: Accuracy of results of fuzzy logic based decision making.

(i) The previous approaches for patient care and remote monitoring were using simplistic decision making while the proposed method is using fuzzy logic system for decision making.

(ii) The proposed method uses sensors to collect data while most of the previous systems were using video data for monitoring and communication.

(iii) Previous methods were also using manual patient treatment with doctors for determination of patient conditions, while the proposed system uses intelligent decision-making approach for this purpose.

(iv) A knowledge base is also established to determine patient conditions.

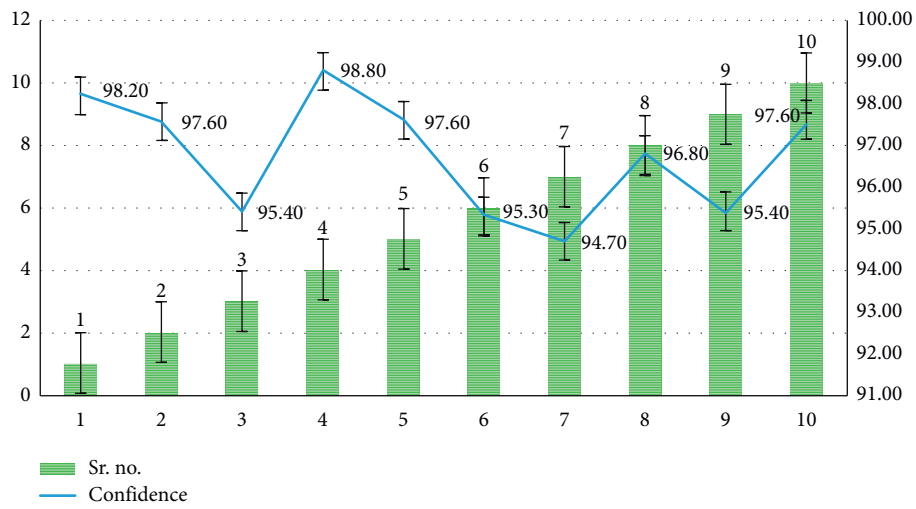


Figure 16: Reliability of results of fuzzy logic based decision making.

6. Conclusions and Future Work

The proposed method consists of sensors for body temperature, pulse rate, and blood pressure to assess the condition of the patient under observation. For determining the possible conditions and cure, the system used a knowledge base and fuzzy logic system for intelligent decision making for patient care, monitoring, and management. The proposed method also tries to improve the effectiveness of the system for patient care and monitoring in terms of time, cost, and manpower utilization. The proposed approach addresses the patient monitoring with sensors and shows reasonable accuracy and cost savings with respect to the systems in use. The study was tested on a small sample of the population and found to be effective, accurate, and efficient for the purpose. The proposed approach is generalized so far, and it is possible to customize it for more critical conditions like operation theatre, intensive care unit patients, newborn babies, and more complex patients. There are three contributions of this work, summarized in Section 5:

(1) The novel idea of using sensors with conventional telemedicine

(2) The new and improved way of diagnosis using fuzzy neural networks based approach

(3) The use of decision support system to minimize time constraint of conventional store-and-forward method of telemedicine in rural areas

The results also show that fuzzy logic system is good choice for intelligent decision-making systems and it also provides a lightweight solution in terms of its devices and software components. In the future, we propose the use of more sensors to get more patient data for better and improved diagnosis.

Data Availability

The datasets used in the experiments and discussed in the paper are available from the corresponding author on reasonable request.

Conflicts of Interest

None of the authors have conflicts of interest related to the research and results presented in this paper.

References

[1] A. Whitmore, A. Agarwal, and L. Da Xu, "The internet of things-a survey of topics and trends," *Information Systems Frontiers*, vol. 17, no. 2, pp. 261–274, 2015.

[2] P. P. Ray, "Home health hub internet of things (H^3 IoT): an architectural framework for monitoring health of elderly people," in *Proceedings of the 2014 International Conference on Science Engineering and Management Research (ICSEMR)*, pp. 1–3, Chennai, India, November 2014.

[3] K. K. Goyal, A. Garg, A. Rastogi, and S. Singhal, "A literature survey on internet of things (IOT)," *International Journal of Advanced Networking and Applications*, vol. 9, no. 6, pp. 3663–3668, 2018.

[4] B. K. Chae, "The internet of things (IoT): a survey of topics and trends using twitter data and topic modeling," in *Proceedings of the 22nd ITS Biennial Conference of the International Telecommunications Society (ITS): Beyond the Boundaries: Challenges for Business, Policy and Society*, Seoul, South Korea, June 2018.

[5] A. Ahmed, R. Latif, S. Latif, H. Abbas, and F. A. Khan, "Malicious insiders attack in IoT based multi-cloud e-healthcare environment: a systematic literature review," *Multimedia Tools and Applications*, vol. 77, no. 9, pp. 1–19, 2018.

[6] P. V. Krishna, S. Gurumoorthy, and M. S. Obaidat, *Internet of Things and Personalized Healthcare Systems*, Springer, Berlin, Germany, 2019.

[7] J. H. Abawajy and M. M. Hassan, "Federated internet of things and cloud computing pervasive patient health monitoring system," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 48–53, 2017.

[8] S. Madakam, R. Ramaswamy, and S. Tripathi, "Internet of things (IoT): a literature review," *Journal of Computer and Communications*, vol. 3, no. 5, Article ID 164, 2015.

[9] Y. Liao, F. Deschamps, E. d. F. R. Loures, and L. F. P. Ramos, "Past, present and future of industry 4.0—a systematic literature review and research agenda proposal," *International*

Journal of Production Research, vol. 55, no. 12, pp. 3609–3629, 2017.

[10] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (IoT): a vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[11] S. Amendola, R. Lodato, S. Manzari, C. Occhiuzzi, and G. Marrocco, "RFID technology for IoT-based personal healthcare in smart spaces," *IEEE Internet of Things Journal*, vol. 1, no. 2, pp. 144–152, 2014.

[12] P. Gope and T. Hwang, "BSN-care: a secure IoT-based modern healthcare system using body sensor network," *IEEE Sensors Journal*, vol. 16, no. 5, pp. 1368–1376, 2015.

[13] A.-M. Rahmani, N. K. Thanigaivelan, T. N. Gia et al., "Smart e-health gateway: bringing intelligence to internet-of-things based ubiquitous healthcare systems," in *Proceedings of the 2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 826–834, Las Vegas, NV, USA, January 2015.

[14] Y. Liu, J. Niu, L. Yang, and L. Shu, "EBPlatform: an IoT-based system for NCD patients homecare in China," in *Proceedings of the 2014 IEEE Global Communications Conference*, pp. 2448–2453, Austin, TX, USA, December 2014.

[15] H. N. Saha, N. F. Raun, and M. Saha, "Monitoring patient's health with smart ambulance system using internet of things (IOTs)," in *Proceedings of the 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pp. 91–95, Bangkok, Thailand, August 2017.

[16] X. M. Zhang and N. Zhang, "An open, secure and flexible platform based on internet of things and cloud computing for ambient aiding living and telemedicine," in *Proceedings of the 2011 International Conference on Computer and Management (CAMAN)*, pp. 1–4, Wuhan, China, May 2011.

[17] M. Hassanalieragh, A. Page, T. Soyata et al., "Health monitoring and management using internet-of-things (IoT) sensing with cloud-based processing: opportunities and challenges," in *Proceedings of the 2015 IEEE International Conference on Services Computing*, pp. 285–292, New York City, NY, USA, June 2015.

[18] H. Sattar, I. S. Bajwa, R. U. Amin, and U. Shafi, "Smart wound hydration monitoring using biosensors and fuzzy inference system," *Wireless Communication and Mobile Computing*, vol. 2019, Article ID 8059629, 15 pages, 2019.

[19] K. Ullah, M. A. Shah, and S. Zhang, "Effective ways to use Internet of Things in the field of medical and smart health care," in *Proceedings of the 2016 International Conference on Intelligent Systems Engineering (ICISE)*, pp. 372–379, Islamabad, Pakistan, January 2016.

[20] M. R. Ruman, B. Amit, W. Rahman, K. R. Jahan, M. J. Roni, and M. F. Rahman, "IoT based emergency health monitoring system," in *Proceedings of the 2020 International Conference on Industry 4.0 Technology (I4Tech)*, pp. 159–162, Pune, India, February 2020.

[21] C. Raj, C. Jain, and W. Arif, "HEMAN: health monitoring and nous: an IoT based e-health care system for remote telemedicine," in *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 2115–2119, Chennai, India, March 2017.

[22] V. Tripathi and F. Shakeel, "Monitoring health care system using internet of things—an immaculate pairing," in *Proceedings of the 2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS)*, pp. 153–158, Jammu and Kashmir, India, December 2017.

[23] R. Nawaz Bashir, I. Sarwar Bajwa, M. Malik, and S. Ali, "Internet of things (IoT) and machine learning based leaching requirements estimation for saline soils," *IEEE Internet of Things*, vol. 7, no. 5, pp. 4464–4472, 2020.

[24] J. K. Reena and R. Parameswari, "A smart health care monitor system in IoT based human activities of daily living: a review," in *Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 446–448, Faridabad, India, February 2019.

[25] K. Saleem, I. Sarwar Bajwa, N. Sarwar, W. Anwar, and A. Ashraf, "IoT healthcare: design of smart and cost-effective sleep quality monitoring system," *Journal of Sensors*, vol. 2020, Article ID 8882378, 17 pages, 2020.

[26] A. M. Ghosh, D. Halder, and S. A. Hossain, "Remote health monitoring system through IoT," in *Proceedings of the 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pp. 921–926, Dhaka, Bangladesh, May 2016.

[27] H. N. Saha, S. Auddy, S. Pal et al., "Health monitoring using internet of things (IoT)," in *Proceedings of the 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pp. 69–73, Thailand, Bangkok, August 2017.

[28] P. Sundaravadivel, E. Kougianos, S. P. Mohanty, and M. K. Ganapathiraju, "Everything you wanted to know about smart health care: evaluating the different technologies and components of the internet of things for better health," *IEEE Consumer Electronics Magazine*, vol. 7, no. 1, pp. 18–28, 2018.

[29] K. Suma, S. Sandeep, S. Vikram, K. Hanjar, and S. Sudharshan, "Cardiogenic shock monitoring system for ambulance," in *Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1357–1360, Kochi, India, August 2015.

[30] V. Pardeshi, S. Sagar, S. Murmurwar, and P. Hage, "Health monitoring systems using IoT and Raspberry Pi—a review," in *Proceedings of the 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pp. 134–137, Bangalore, India, February 2017.

[31] P. Dineshkumar, R. SenthilKumar, K. Sujatha, R. Ponmagal, and V. Rajavarman, "Big data analytics of IoT based health care monitoring system," in *Proceedings of the 2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering (UPCON)*, pp. 55–60, Varanasi, India, December 2016.

[32] J. He, A. Atabekov, and H. M. Haddad, "Internet-of-things based smart resource management system: a case study intelligent chair system," in *Proceedings of the 2016 25th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–6, Waikoloa, HI, USA, August 2016.

[33] A. Archip, N. Botezatu, E. Şerban, P.-C. Herghelegiu, and A. Zală, "An IoT based system for remote patient monitoring," in *Proceedings of the 2016 17th International Carpathian Control Conference (ICCC)*, pp. 1–6, High Tatras, Slovakia, May 2016.

[34] P. Sundaravadivel, S. P. Mohanty, E. Kougianos, V. P. Yanambaka, and H. Thapliyal, "Exploring human body communications for IoT enabled ambulatory health monitoring systems," in *Proceedings of the 2016 IEEE International Symposium on Nanoelectronic and Information Systems (iNIS)*, pp. 17–22, Gwalior, India, December 2016.

[35] S. B. Baker, W. Xiang, and I. Atkinson, "Internet of things for smart healthcare: technologies, challenges, and opportunities," *IEEE Access*, vol. 5, pp. 26521–26544, 2017.

[36] M. Safdar Malik, I. Sarwar Bajwa, and S. Munawar, "An intelligent and secure IoT based smart watering system using fuzzy logic and blockchain," *Computers and Electrical Engineering*, vol. 77, no. 1, pp. 109–119, 2018.

[37] D. Metcalf, S. T. J. Milliard, M. Gomez, and M. Schwartz, "Wearables and the internet of things for health: wearable, interconnected devices promise more efficient and comprehensive health care," *IEEE Pulse*, vol. 7, no. 5, pp. 35–39, 2016.

[38] H. Sattar, I. S. Bajwa, and U. F. Shafi, "An intelligent air quality sensing system for open-skin wound monitoring," *Electronics*, vol. 8, no. 7, Article ID 801, 2019.

[39] C.-T. Lin and C. S. G. Lee, "Neural-network-based fuzzy logic control and decision system," *IEEE Transactions on Computers*, vol. 40, no. 12, pp. 1320–1336, 1991.

[40] O. Nelles, *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*, Springer Science & Business Media, Berlin, Germany, 2013.

[41] J.-S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[42] H. Sattar, I. S. Bajwa, R. Ul-Amin et al., "An intelligent and smart environment monitoring system for healthcare," *Applied Sciences*, vol. 9, no. 19, Article ID 4172, 2019.

[43] H. Sattar, I. S. Bajwa, R. U. Amin et al., "An IoT-based intelligent wound monitoring system," *IEEE Access*, vol. 7, no. 1, pp. 144500–144515, 2019.

Hindawi
Scientific Programming
Volume 2020, Article ID 8831611, 21 pages
https://doi.org/10.1155/2020/8831611

Research Article

A Privacy-Preserving Attack-Resistant Trust Model for Internet of Vehicles Ad Hoc Networks

Muhammad Haleem Junejo,[1] **Ab Al-Hadi Ab Rahman**,[1] **Riaz Ahmed Shaikh**,[2] **Kamaludin Mohamad Yusof**,[1] **Imran Memon**,[3] **Hadiqua Fazal**,[3] **and Dileep Kumar**[4]

[1]*Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia*
[2]*Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*
[3]*Department of Computer Science, Bahria University, Karachi Campus, Sindh, Pakistan*
[4]*State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China*

Correspondence should be addressed to Dileep Kumar; dk2kes21@gmail.com

Received 3 September 2020; Revised 6 November 2020; Accepted 24 November 2020; Published 11 December 2020

Academic Editor: Shah Nazir

The Internet of things (IoT) and advancements of wireless technology have evolved intelligent transport systems to integrate billion of smart objects ready to connect to the Internet. The modern era of the Internet of things (IoT) has brought significant development in vehicular ad hoc networks (VANETs) which transformed the conventional VANET into the Internet of Vehicle (IoV) to improve road safety and diminished road congestion. However, security threats are increasing due to dependency on infrastructure, computing, dynamic nature, and control technologies of VANET. The security threats of VANETs could be addressed comprehensively by increasing trustworthiness on the message received and transmitting node. Conversely, the presence of dishonest vehicles, for instance, Man in the Middle (MiTM) attackers, in the network sharing malicious content could be posed as a severe threat to VANET. Thus, increasing trustworthiness among nodes can lead to increased authenticity, privacy, accuracy, security, and trusted information sharing in the VANET. In this paper, a lightweight trust model is proposed, presented model identifying dishonest nodes and revoking its credential in the MiTM attack scenario. Furthermore, addressing the privacy and security requirement, the pseudonym scheme is used. All nodes in the VANET established trust provided by initially RSU, which is a trusted source in the network. Extensive experiments are conducted based on a variety of network scenarios to evaluate the accuracy and performance of the presented lightweight trust model. In terms of recall, precision, and *F*-score, our presented model significantly outperformed compared to MARINE. The simulation results have validated that the proposed lightweight model realized a high trust level with 40% of MiTM attackers and in terms of *F*-score 95%, whereas the MARINE model has 90%, which leads to the model to attain high detection accuracy.

1. Introduction

In our daily lives, the transport system plays an undeniable role. It is projected that this increasing number of vehicles on roads reaches up to 2 billion or more in the coming decades [1]. As a consequence, we encounter an unfortunate rise in accidents, traffic jams, congestion, pollution, and so forth. The World Health Organization (WHO) has released a report of 1.35 million deaths due to road accidents [2, 3]. To improve transport efficiency and security, the vehicular ad hoc networks (VANETs) present the foundation of the smart city paradigm and Intelligent Transportation Systems (ITSs) [4–6]. The Internet of things (IoT) is a novel concept in the current era that is evolved to integrate billion of smart objects ready to connect to the Internet [7]. The newest technologies have enabled smart object, remote devices, and wireless and wired networks to be part of IoT. The IoT combines all electronic, mechanical, and computing devices

to part of the Internet. The vehicular ad hoc networks (VANETs) connected to IoT bring the concept of the Internet of Vehicles (IoV) [8–10]. Internet of VANET is application of IoT to improve the urban transport system, reduce accidents, and enhance the traffic monitoring system [10]. The main features of IoV are high creditability, controllability, manageability, and operationalization efficiency [7, 11]. The VANETs are considered as subclasses of Mobile Ad Hoc Networks (MANETs) [12–14]. Under the umbrella of VANET, vehicles are capable of communicating with other vehicles and the roadside units by dedicated short-range communication (DSRC) radiofrequency. Particularly, in VANET, two types of communication are established. Firstly, it is among vehicle-to-vehicle (V2V) communication and secondly, in between vehicles-to-infrastructure (V2I) communication. The primary nodes in VANETs are smart vehicles and Road Side Units (RSU) that are communicating among each other to exchange safety, security, and infotainment information.

In a situation where the exchanged information is incorrect, it leads to some counterproductive; consequently, accidents and traffic congestion would increase. Over the last decade, promising advancements have been made in the field of VANET [15]. Accordingly, the scientific community has contributed a lot to overcome the challenges in the scope of security, safety, and engineering design. In the context of effectively using the VANETs, the most important aspect is to deal with the safety, security, and privacy parameters.

To that respect, several solutions have been proposed by the research community to foster security in VANETs [16–20]. In those solutions, the authors suggested, as a solution, the use of traditional cryptography which utilizes the Public Key Infrastructure (PKI) and certificates to achieve security in the network. However, these solutions suffer from several factors that reduce network efficiency in VANETs. These factors include

(i) The mobility of vehicles randomly dispersed throughout a network with low- and high-speed vehicles

(ii) The presence of a roadside unit or network infrastructure in a rural area is not assured all the time

(iii) The propagation of untrusted messages in VANET in case of an inside attack is a result of a compromised cryptographic solution

The cryptographic-based solution can protect VANETs from outside attacks. However, it is incapable of assuring message reliability and quality, which may lead to undesirable consequences. This leads to the emergence of trust-based solutions, which aim to protect VANETs from inside attacks [21–24].

Trust, in the VANET, is described as the confidence of one (vehicle) to the other for performing a requirement or a set of conditions [19, 20]. In VANET, the trust is created between two or more vehicles based on the intercommunication. Once the message is received, the assessor node computes the trust based on numerous factors, which are the vehicle's previous communication, reputation in the network, and neighbors' recommendations regarding a specific vehicle.

It must be noted that, due to extremely mobile and randomly distributed vehicles, the trust was established for a short duration [12, 25, 26]. Therefore, it is challenging to creating, calculating, quantifying, and assessing the trust in received messages based on varied factors in a limited time. The trust, as a method to attain security in VANETs, is in its early stages of development. The trust models (TM) are fixed within vehicles to assess the reliability, accuracy, and authenticity of received messages. The TMs confirm the broadcast of trusted information in the network by retracting both dishonest nodes and malicious messages.

These challenges are imposed because of the ephemeral nature of VANETs [27]. In the literature, most of the existing trust models did not properly address security control to countermeasure the security vulnerability and attacks in VANET. To cover this gap, the trust metric value should be taken into account for multiple factors and for protection against the attacks. The recently designed architecture of VANET trust models encompasses the key new features to reduce the effect of security attacks, which are the ability to configure, control, and combine security services. Vehicles, RSU, and other node parts of the VANET network should be trusted and reliable. To identify the malicious, misbehaving, and compromised node in the VANET network is challenging due to the aforementioned points. Furthermore, it is an open issue to evaluate the trustworthiness of a node. The safety of human lives can be lost in case of any sort of miscommunication in VANET. Several parameters need to be considered before trusting the received message from another node based on the following questions:

(1) What is the reliability of a node before transmitting a critical message?

(2) How criteria are defined on the basis of that the trustworthiness of the node?

(3) How to detect the misbehavior in calculating the trustworthiness of a vehicle?

To address the security challenges required by VANETs are availability, authentication, confidentiality, integrity, privacy, nonrepudiation, and others. The security threats of VANETs could be addressed comprehensively by increasing trustworthiness on the message received and transmitting node. In this paper, we propose a trust management model for vehicular ad hoc networks. The presented model consists of two main blocks: Trust Estimation Model and Decision Model.

(i) The trust estimation in the proposed model is based on five parameters, namely, Location Closeness, Data Integrity, Authentication, Time Stamp Verification, and Peer Alert Message. The trust estimation part calculates the threshold value on the data received from all of five parameters.

(ii) The decision model received the trust value from the trust estimation block to decide whether to process the message or discard it on the basis of the threshold

value. If the trust value is less than the threshold value a TRUE message is generated, and the decision box accepts the value send an update to a database and takes an application-specific decision. In case, if the threshold value exceeds, the threshold value message is discarded and the FALSE message is generated. On the basis of false generated message value, invoke/revoke procedure decide to invoke or revoke the message.

The main contribution of our presented model is as follows:

(1) An attack-resistant trust model for VANETs that efficiently addresses the privacy issue by using the pseudonym scheme

(2) Propose a trust model, identifying dishonest nodes and revoking its credential in a MiTM attack scenario

(3) The RSU is a trusted source in the network, RSU assigns an initial trusted value in the coverage area and based on the presented scenario generates a peer alert message to inform vehicle in the coverage area about the presence of a malicious vehicle.

This paper is organized as follows. In Section 2, related work is presented. Section 3 discusses the architecture of VANET and security threats. Section 4 represents the trust model in detail, and in Section 5, we present the evaluation of the presented trust model in the presence of four variant of MiTM attacks scenario. In Section 6, conclusion of the paper is demonstrated.

2. Related Work

The trust established between the nodes can be classified into two:

(1) Infrastructure based

(2) Self-organizing

Infrastructure trust is based on the certificates carried by each vehicle in the network, while the self-organizing as the term is quite self-explanatory. Meaning that the self-organizing is based on the trust that is directly between two nodes, indirect between the nodes, and a combination of direct and indirect is termed as a hybrid. In VANET, the trust is calculated on a node or the received message. The trust calculation can be centralized or distributed based on the environment and the infrastructure used.

In VANET, the TMs are divided into three distinct classes:

(1) Data-oriented

(2) Entity-oriented

(3) Hybrid

The purpose of entity-oriented (EO) is to remove dishonest vehicles by assessing the reliability of the node. The data-oriented (DO) evaluates the trust in the received messages (data). And, finally, the hybrid trust models (HTM) calculation is based on both vehicle and data for the trust creation.

2.1. Data-Oriented Trust Model. In recent studies, few trust models are proposed for data-oriented trust calculation. In DO, the calculation of trust is performed on the trustworthiness of the received messages.

A framework proposed by [28] on data-centric trust creation is based on location and time. The authors' approach is based on the evaluator node (EV) that initially receives data from vehicles in the area and then allocates weights to each received data based on two factors: location and time. The proposed frame is not well suited to dynamic and sparse environments as trust is computed all the way, and data is received at a node. In his approach, the author utilized several decision logics, specifically weighted voting, Bayesian inference, and Dempster–Shafer Theory. He concludes that Bayesian inference achieved better results the Dempster–Shafer based on multiple events. The shortcoming of the proposed scheme that it is appropriate only in a condition when there is adequate evidence is available in favor or against a given scenario for a particular event [29].

Gurung et al. [30], in their proposed trust model, evaluate the trustworthiness of the message based on multiple factors such as context similarity, content conflict, and routine similarity. In their conclusion of the paper, the author concluded that the proposed trust model meets the requirements of the dynamic of VANETs nature. The shortcoming of work proposed by the author is that the model contains real-time confirmation of received messages which is not possible in high mobility and scant situation.

Shaikh and Alzahrani [21], in their work, proposed a trust model based on the timing and fake location attacks. The trust model is decentralized and suitable for real-time application in VANETs as it introduces linear time complexity and simple. Moreover, the trust model proposed method detects the false location, time, and robustness. The computation of the trustworthiness of the message is based on previous information on node holds. Furthermore, the trust value of the event decides to accept or reject the value.

Mármol and Pérez [22] proposed a trust model, namely, TRIP. In the work introduced by authors, computation of trust of node is based on three factors. First, direct experiences based on previous interaction with node; second, interactive communication with surrounding nodes and their recommendations; and third, the communication between RSU and central authority and central authority send recommendations. Computation of reputation score map all three values received from conditions 1, 2, and three based on fuzzy sets that are ((One) trust; (Two) not trust; and (Three) +/– trust)). In three conditions of trust to accept or discarded, first, if the score is placed in "not trust," discard the message, and the presence of the dishonest node is sent to infrastructure. In other cases, if the score is placed in "trust," then the message is accepted and forwarded to other vehicles in the network. In the last condition, if the reputation score is computed as "+/– trust," the message is processed as reliable with the condition of tunable

probability; furthermore, it is not forwarded to nodes in the network. We find that the proposed assumption is not realistic. In addition to this, to build a history and reputation of the received message of vehicles, in this scenario, the actual identities of vehicles should be known.

Patwardhan et al. [31] proposed the Data Intensive Reputation Management model. The protocol integrates reputation and agreement to guarantee the reliability of data and kindle proactive collaboration. Furthermore, in their model, they exercise multiple factors such as frequency of encounters, persistent identities, and a known set of trustworthy sources for creating trust relationships among existing unknown devices. The trustworthiness of data depends upon majority consensus among peers or in case it is received from trustworthy sources. In addition to this, the authors supposed that each node must have a unique persistent identity, and this assumption violates identity privacy.

Chen et al. [32] proposed a trust model framework for evaluation and message propagation. In their trust model, the authors used experience-based trust, trust opinions, and role-based trust models to model the quality of information shared between nodes. The model is based on a binary operation that is either to (trust) or (not trust) information. This binary condition limits the situation based on incomplete information or in other cases are in uncertain situations. Moreover, in their work, the key important features such as privacy and robustness are not widely addressed.

Lo and Tsai [33] have proposed a trust modeling framework based on Traffic Safety Event. In their method, specifically, the event-based Reputation System (ERS) is used to stop the nodes to broadcast compromised, untrustworthy, and malicious warning messages. Furthermore, the method uses a cooperative-event observation and reputation adaptation schemes, with two types of thresholds, event confidence and event reputation, to calculate the event intensity and event reliability simultaneously. The major shortcoming of the proposed model is the time taken to share the trusted information with peers in time.

Liu et al. [24] have proposed a trust model, namely, LSOT in VANETs, based on two types of evaluation methods: certificate-based and recommendation-based trust. In their work, authors address the high mobility and random distribution dynamics of VANETs. Furthermore, the LSOT model operates in a fully distributed environment. To the calculation of trust, the three weight factors were used, which are number, time decay, and context to accurately determine overall trust. The main drawbacks of this model are that the authors failed to differentiate between the message and trust of the node.

2.2. Entity-Oriented Trust Model. The entity-oriented (EO) aims to remove dishonest vehicles by assessing the reliability of the node. The EO evaluates the trust on the node and identifies the presence of a malicious vehicle in the network. There is a considerable amount of literature work carried out by several authors on data-oriented trustworthiness.

Mármol and Pérez [22] have presented a trust scheme based on reputation infrastructure, for vehicular ad hoc networks. In their work, the authors are considering three different types of information to calculating the reputation score for every node in the network. The three estimating parameters are direct interaction with the previous vehicle, suggestions, and recommendations from nearby vehicles in the network and central authority recommendations. To accept or reject them based on the three conditions after the trust score is generated if the generated trust score is found as "not trust," the message is dropped and the presence of the dishonest node is sent to infrastructure. In other cases, if the trust value is calculated as "trust," then the message is accepted. In the last condition, if the trust value is calculated as "+/– trust," the message is accepted, and it is not forwarded to nodes in the network. Furthermore, in their model, trust establishment is connected to the node verification of trustworthiness of the node. The main shortcoming of the proposed trust model is that multiple senders will send the reputation of the sender, and this will generate additional overhead.

Khan et al. [34] have proposed a trust model DMN in Vehicular Ad Hoc Networks based on cluster-based mechanisms. The Cluster Head (CH) is responsible to calculate the trust and forward it to a Trusted Authority (TA). Furthermore, the TA is responsible to remove a malicious node from a network based on information received from CH. The main drawback of the proposed approach is that this approach is high overgenerated due to continuous reporting, which reduces network efficiency. Moreover, the network communication detail between CH, TA, and vehicles is missing.

Gerlach [35] developed a preliminary method in which each vehicle builds a profile of another vehicle when other vehicles come in contact. The proposed TM is sociological trust and based on the principle of confidence tagging and trust. The evaluation of trustworthiness is based on the interaction between vehicle profile histories. The EO model approach has serious weaknesses. First, VANET is highly dynamic, and interaction between the vehicles is for a limited time; this leads to difficulty in collecting enough evidence to calculate trust. Second, in case the vehicle itself is trustworthy, however, the message sent by the vehicle is either correct or not. In conclusion, the author presents a method of trust tagging exercising probabilities for representing trust and a trust model for vehicular applications for trust and applications. The shortcoming of Gerlach's proposed trust model is that it does not include the formalization of the architecture. Furthermore, their work failed to address a combination of the different types of trust together.

Minhas et al. [23], in their work, proposed role-based trust and experience-based trust as the evaluation method metric for the integrated reliability of nodes. This model also permits a vehicular entity to vigorously investigate about an event by sending requests to other entities but restricts the received number of reports. The multifaceted trust management model of the author has combined role based and experience based that are incorporated into the priority-based model, the two factors used to choose proper advisers.

The advisors are using the majority-opinion method to receive feedback. Furthermore, based on feedback aggregation received from advisors, two more factors were also considered: time and location closeness. The authors, further in their work, suppose that authorities predefined the roles and are assumed to behave in a certain way. The shortcoming of the work is that the robustness has not been addressed widely.

Yang [36] proposed a trust model based on Reputation Management for VANETs. The author used a similarity mining approach to calculate the trustworthiness of the vehicle. Furthermore, the reputations of recommenders are exercised as weights for calculating a full reputation for the message generator. The main drawback of the approach used by the author is that it proposed TM based on Euclidean distance between two vehicles as this contrast global information on the similarity of the generated message.

Jesudoss et al. [37] proposed a trust model scheme based on the reputation and election of CH. Authors in the scheme utilize the truth-telling approach to propagate true content to receive a better reputation. Moreover, the election is held among nodes to elect as CH. Furthermore, in election, nodes assign incentives in the form of weights. Higher the weight is, the more trusted the node by CH. Although this approach is interesting, it suffers from a rural scenario and highly mobile where only a few numbers of vehicle participates in election.

Haddadou et al. [38] proposed an economic incentive-based trust model. The authors used a distinct approach in which the credit value is assigned in a distributed manner. The credit value can be increased and decreased based on node behavior in the network. Furthermore, the credit value decreases each time in case an attack occurs in a network.

Zhang et al. [39] have proposed a trust scheme based on the Chinese remainder theorem (CRT). The authors work based on securing nodes privacy and offer authentication. Their scheme is based on tamper-proof device (TPD) identity, RSUs, and TAs. The shortcoming of the proposed scheme is that it is fully centralized, depends on RSUs and TAs, and is not applicable in rural areas where the VANET infrastructure is not available.

Guleng et al. [40] have proposed a trust scheme based on fuzzy logic to evaluate direct trust on the node. The author utilized honestness, cooperativeness, and responsibility factors in their approach based on fuzzy logic. The main shortcoming of this approach is the limitation of coverage area as the scheme is fully decentralized.

2.3. Hybrid Trust Models. Hybrid rust models combined the properties of both entity-centric and data-centric trust model scheme. Recently, in the literature, several studies have been conducted on the trust established based on hybrid trust models. A hybrid trust model has evaluated the trustworthiness of peers and utilizing modeling outcomes to calculate the reliability and trustworthiness of data.

Sedjelmaci and Senouci [41] have proposed a trust model based on the mobility and accuracy of VANET. The author claims that the trust model addresses the basic characteristics of a network for instance node's mobility and rapid topology change. The authors claim that the proposed lightweight model will adversary address the most dangerous attacks such as a black-hole attack, wormhole attack, and Sybil attacks by using a watchdog mechanism. Furthermore, the proposed solution is divided into two level intrusion detection systems. The first part is based on collaborative detection, whereas the second part of the framework deals with a global detection system that was processed by RSU. The main shortcoming of the proposed solution is that time to elect the cluster head will pose a delay in the network and time-consuming process.

Dhurandher et al. [42] proposed a framework, a Reputation and Plausibility Checks-based approach, by transmitting safety and security-related messages. The authors work based on the Vehicular Security throw reputation and plausibility check (VSRP) mechanism which utilizes three terminologies in the algorithms, and they are event modification message, data grouping, and false event generation. The main drawback of the proposed solution is that the detection range is very short that is 50 meters. Furthermore, detection is based on the vehicles' embedded sensors.

Abdelaziz et al. [43] proposed a trust-based scheme for VANETs, namely, Trust Model with Delayed Verification for Message Relay. The authors divided data traffic into four distinct classes specifically based on the priority given to safety-related messages from high to low as follows: (1) background traffic, (2) best-effort traffic, (3) video traffic, and (4) voice traffic. The main drawback of the proposed trust model is that author assumes that a dishonest vehicle will behave constantly all over to their journey in the network; this approach is invalid in the VANET.

Dotzeret et al. [44] proposed a trust scheme that is based on a distributed reputation model piggybacking opinion approach. In this approach, every forwarding node adds its own opinion regarding the trustworthiness of data. The trustworthiness algorithm is based on multiple trust factors that include direct and indirect trust, sender base reputation, and geo-situation orientation.

The main drawbacks of the scheme provider by authors are that they failed to provide sufficient and complete details about the approach. Moreover, the author mentioned that, in algorithm, sender-based information is managed; however, it failed to provide about how reputed information in TM will be updated.

Chen et al. [25] proposed a framework based on the message propagation and evaluation framework. The framework is based on trustworthiness message propagation in a distributed and collaborative fashion. The authors in their model address basic characteristics of VANETs; they are network scalability and system effectiveness. Moreover, those two characteristics include the addition of information evaluation based on the pervasive presence of false information in a network.

Rai et al. [3] proposed a hybrid VANET-based trust scheme, namely, a hybrid dual-mode trust management scheme for vehicular networks. The author's scheme is dual applicable for urban and rural based. Their scheme is based on the crediting technique. The credit value is obtained by

looking at sender node history and validation of the message received. The main shortcoming of the approach is the missing of central authority and infrastructure of VANET.

2.4. Authentication Schemes Based on a Pseudonym. The key requirements of privacy in VANETs are the unlinkability and the secrecy of the message. The safety-related beacons are broadcasted every 300 ms in a vehicle-to-vehicle communication. This phenomenon can lead to potential endangers the privacy of drivers by tracking the mobility pattern of the targeted driver. The main motive of attacks on privacy is to get sensitive information about vehicles and drivers [45]. A pseudonym scheme facilitates hiding the identity of a vehicle and addresses the privacy and security requirements of the system [46]. Furthermore, a pseudonym is a temporary certificate assigned to a vehicle to hide its real identity [14, 46–50]. In the literature, a range of pseudonym schemes is proposed to provide privacy protection and changing pseudonyms periodically.

Buttyán et al. [47] proposed a scheme, namely, SLOW: a practical pseudonym changing scheme for location privacy in VANETs. In their scheme, the authors proposed that the vehicle must not send beacons in case its speed is reduced below a given threshold. Furthermore, a vehicle must change a pseudonym for the duration of such a silent period.

In [49, 51, 52], the pseudonym of vehicle changes in case it enters the social spot and mix zone. In [53], the authors put forward a cooperative pseudonym change method among its neighbors. In [47, 54], once a pseudonym is changed then the vehicle will keep communication silent. The assure legitimacy and integrity of message authentication are indispensable. Various approaches have been proposed in [49, 55], and with these approaches, the authors developed the methods of verifying the certificate and message. These approaches can authenticate the legitimacy of the sender and validity of a message without revealing the vehicle identity. The main weaknesses in their approaches are the trustworthiness of received messages.

3. Architecture of Internet of VANETs

The main components of VANETs are vehicles embedded with OBU, RSU a communication component consist of RF antennas and process unit, and telecommunication network, for example, satellite communication. There are mainly three types of communication modes:

(1) Intervehicle communication (V2V)
(2) Vehicle-to-roadside communication (V2I)
(3) Interroadside communication (I2I)

(a) Intervehicle communication (V2V): in this mode of communication, vehicles communicate with another vehicle with the help of OBU embedded in every vehicle. In this communication mode, vehicle to vehicle communicates with each other with wireless technology. Furthermore, the message transmitted among the vehicle is broadcast so all vehicles in the coverage area received the transmitted information, as shown in Figure 1.
(b) Vehicle-to-roadside communication (V2I): in this mode of communication, vehicles will communicate with roadsides communication equipment Roadside Unit (RSU). Furthermore, in this mode, a direct wireless communication link is established between vehicle and infrastructure units located around the road [56].
(c) Interroadside communication (I2I): in this mode, communication RSU communicates with another RSU and core network, for example, 5G, satellite, or wired telecommunication system.
(d) Trusted authority: trusted authority (TA) is the heart of the VANET system. The primary responsibility is registering the RSUs, OBUs, and vehicles. The secondary responsibility includes assuring security management by verifying authentication of vehicle, user identification, and OBU identification to secure the vehicle from attack.
(e) Roadside units (RSU): these are communication based units installed near highways, which transmit useful information to vehicles that came in the radio range of RSU. They are connected to a central network with means of wired or wireless.
(f) Vehicles: vehicles are the basic units of VANET; they are equipped with the computing device installed on it called the On-Board Unit (OBU). The main responsibility of OBU to communicate with neighboring OBU installed on the vehicle as well as RSU. TA sends multiple pseudonyms to registered vehicles in the network.
(g) Legitimate nodes trust variation: in this section, legitimacy and dishonesties of TM will be measured in the presence of an attacker. Furthermore, compromised messages and trust rating was shared by an attacker.
(h) Malicious nodes trust variation: in this section, we define the ability of TM to implement the lowest level for the attackers.
(i) Centralized reputation serve (CRS): it assigns an initial reputation value for each registered vehicle in the network. CRS is responsible for managing and updating reputation. In case the reputation value is less than the threshold, CRS revokes the vehicle from the network.
(j) Pseudonyms: these are identities that are assigned to nodes in the network and only once used. The basic functionality is to maintain the privacy of nodes. Central authority keeps changing assigned pseudonyms periodically. A pseudonym is a temporary certificate assigned to a vehicle to hide its real identity.
(k) Mix zone: this is the coverage area in the VANET that is not under the surveillance range of the dishonest attacker. This is suitable for a node to change their pseudonym to prevent tracking.

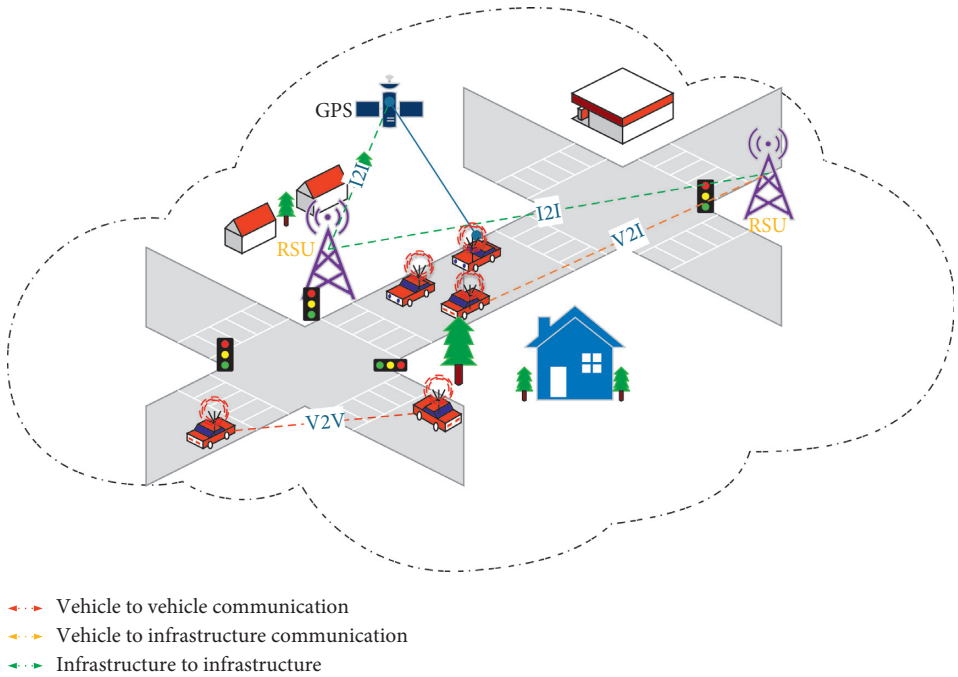


Figure 1: Internet of vehicular ad hoc network architecture.

Furthermore, in this coverage area, multiple nodes exist simultaneously, and this makes the attacker difficult to track the node.

3.1. VANET Challenges in a Road Network. Modeling of VANET trustworthiness peers in road network faces enormous challenges [14]. The key challenges that encounter by VANET can be categorized into two major conditions. Firstly, vehicles are continuously moving in the network and are extremely dynamic [57–59]. The speed of a vehicle is on the highway typically between 80 and 120 km/h. Furthermore, at this high speed on the road, to respond to a forthcoming event is critical in real time, and peers must be able to validate incoming information [12, 29]. Secondly, it may be expected that the number of vehicles in the network can increase in any instance. For example, in urban areas, for all the time, ten thousand peers are always in the network, and during peak rush hours, it will increase dramatically to a higher number. This leads to congestions in the network which poses several issues. Moreover, the VANET is a shared channel network; during rush hour, peers received a lot of information from other peers in a network; this results in information overload [60]. Consequently, there is a great need for intelligent vehicle communication systems that potentially respond to hazardous conditions by efficiently deciding with which peers communicate in a network [61, 62].

A third key challenge is related to modeling trust in the VANET environment as it is a decentralized and open system; this means that no centralized infrastructure exists in the VANET [63]. Furthermore, the vehicle at any time joins or leaves the network, and it may be not guaranteed that, in future, interaction with the same vehicle will happen. Therefore, practically it is not worth to rely on a mechanism which utilizes a centralized system, for example, using Central Certificate Authority and Trusted Third Party or to create long-term relationship depends on a social network.

3.2. VANET Security. It is well known that VANET security is a complex issue with several challenges. These challenges are given, in detail, below. To address these challenges, different requirements must be taken into account. These requirements, for simplicity, can be classified into six main categories, i.e., Availability, Authentication, Confidentiality, Integrity, Privacy, and Nonrepudiation. These requirements, for simplicity, can be classified into six main categories, i.e.,

(1) Availability
(2) Authentication
(3) Confidentiality
(4) Integrity
(5) Privacy
(6) Nonrepudiation

3.2.1. Availability. This requirement is quite self-explanatory, meaning that the VANET ad hoc network must be available all the time to ensure the safety of vehicles. The unavailability could be possible by the DOS attack, as mentioned in [64]. To ensure availability, high connectivity and bandwidth must be disposable. That is, the network must be available all the time, and, at times, it must have a fast response time to some specific applications. A delay, or even milliseconds, could make the message futile, as highlighted in [65–67]. In addition to the aforementioned safety

aspects, security is also highly linked with the availability of the network. In a way, availability is a prerequisite to the overall security of the system [68].

3.2.2. Authenticity. Authentication is one of the major security aspects and plays an important role in VANETs. It is crucial for verifying the claim of authenticity, that is, verifying the identity of a vehicle, and differentiates the legitimate vehicles from the malicious ones. Otherwise, it may lead to serious safety issues, such as human injuries, traffic disruptions, and, in some extreme cases, it may lead to human loss. The process of authentication in VANETs includes three major parameters, i.e., identification, access control, and authentication. This can be achieved by acquiring security certificates and signatures. Specifically, cryptographic mechanisms are used to achieve authentication in VANETs, as it represents the first line of defense against any sort of external danger.

3.2.3. Confidentiality. The confidentiality in VANET plays an important role in maintaining users' privacy by safeguarding the content of information transmitted between two users. Confidentiality is achieved by using shared public keys and certificates in peer-to-peer communication. The cryptography mechanism is exercised to persuade confidentiality in VANET.

3.2.4. Integrity. In VANET, the Integrity assures that the message communicated between two nodes has not been altered, modified, and/or changed during the transmission. The Integrity in VANET could be achieved by cryptography as well as by the Trust. In cryptography, the public key and revocation methods are used to ensure Integrity [68, 69]. The received message, at the end node, could be trusted if it is free from alteration, modification, and change [68, 70–72].

3.2.5. Nonrepudiation. In VANET, the nonrepudiation requirement ensures that the sending node cannot deny a send message. The nonrepudiation matches the nodes' identification with the messages received. This is achieved by utilizing cryptographic approaches to meet certain requirements of nonrepudiation in VANET [69].

3.2.6. Privacy. Privacy is the foremost key requirement in VANET. The major sensitive information related to the nodes is Vehicle location, Identification of vehicle, identification of the driver, and details of the traffic route to be followed by the vehicle. While the communication in VANET is broadcasted, the attacker could take advantage of tacking the vehicle identity and location. Therefore, to ensure the privacy of the vehicle, cryptography and the Trust methods can be exercised in VANET.

3.3. Attacks in VANET. This section lists the common threats faced by VANET [29, 32, 68, 70–72].

(1) Certificate Replication Attack: in this attack, the certificate is replicated multiple times.

(2) Eavesdropping Attack: attacker intercept transmitted the communication to gain access or password.

(3) Tracking Tracing Attack: trace or track the correct position of device and vehicle.

(4) Denial of Service Attack (DoS): this attack is caused by preventing accessing the network from functioning properly and timely manner. This causes a legitimate vehicle not to access the application or services.

(5) Jamming Attack: this attack is almost the same as a DoS attack, but this time the shared bandwidth among the nodes or network is jammed.

(6) Coalition and Platooning Attack: this attack works in a group, where multiple dishonest vehicles collaborate to perform malicious activities such as bandwidth usage or stopping any services.

(7) Betrayal Attack: this attack occurs when an honest vehicle becomes dishonest during transmission.

(8) Replayed, Altered, and Injected Message Attack: this attack altered or modified the information during messages transmission. This will cause to send multiple erroneous messages.

(9) Illusion Attack: typically, this attack is related to hardware component, for example, wrong sensor reading, and incorrect messages are sent to other vehicles.

(10) Masquerading Attack: this attack is caused by a dishonest vehicle wearing a legitimate certificate by disturbing and doing malicious activities.

(11) Impersonation Attack: a dishonest node assumes to be another node by using the wrong identity.

(12) Sybil Attack: a dishonest node transmits multiple fabricated message IDs to the legitimate node, where the legitimate nodes assume that they are dealing with multiple devices.

(13) GPS Position Faking Attack: falsified positioning based on geographical coordinates.

(14) Timing Attack: the dishonest node adds the delay between the packets, which cause unforeseen incidents.

(15) Blackhole Attack: a dishonest node transmits a false reply message to the other vehicle that the dishonest host is optimal route information to the destination.

(16) Grayhole Attack: a dishonest host drops the packet of the particular vehicle in the network and transmits other packets to its destination.

3.4. Identity and Location Privacy Protection in VANET. In VANET, through a continuous exchange of Safety Beacon Messages (SBMs), all peers in the network would receive safety-related information in well time and help peers to be

aware of incidents happening in the surrounding, for example, traffic congestion, accidents on road, and updated traffic flows. The SBM includes major information is speed, location, vehicular identity content of a request, and others. In VANET, information regarding location and identity is most important [51, 73, 74]. Moreover, vehicular identity information is usually protected by utilizing a pseudonym, which is produced by the Central Authority (CA) in the traditional approach used in VANET. In the case, if CA is compromised, this leads to threatening the privacy of the vehicle. SBMs are produced according to location information. The traditional encryption process is used to protect location information which helps that location information during transmission will not be leaked or stolen. However, this approach does not assure that the information in CA or another related server in a centralized structure will not be lost or leaked. The users nowadays are more curious about their private information, so the system must be robust to protect the vehicular location and identity. Privacy protection is the utmost basic requirement of VANETs. Moreover, to protect the information of users, pseudonym technique is used commonly. This strategy helps vehicles to amend pseudonym periodically to avoid being tracked in the system [49]. As a result, the attack on the privacy of the vehicle is the motive of an attacker to get access to sensitive data of the vehicle. Pseudonym schemes are developed to address privacy, security, and system requirement in VANET. To protect the real identity of a vehicle, a temporary certificate is issued, and this terminology is termed as a pseudonym. The authors in [50, 75, 76], when vehicles enter a range of mix-zone or social spot, amend its pseudonym.

4. System Model

This section describes the proposed Lightweight Trust Model (TM), in terms of lightweight, fewer arithmetic operations are used to reduce the complexity, such as square root log and complex geometry of the model. Trustworthiness involves several steps to calculate trust from received information from the sender. Our proposed model is hybrid, which calculates trust in data and node based on V2V and V2I communication. The proposed model comprises of the following two key components:

(1) Trust Estimation Model

(2) Decision Model

4.1. Trust Estimation Model. The trust estimation is performed based on five parameters: Location closeness, Data Integrity, Authentication, Time Stamp verification, and Peer Alert Message. The trust value is calculated based on the value generated by each of the five parameters. The vehicle received a message from another vehicle V2V or Roadside unit V2I.

The TM, in the initial following parameter, can be used, and the parameters may be changed depending on the simulation results and performance of TM, as shown in Figure 2.

4.1.1. Vehicle Location. A vehicle may provide incorrect location information during network interaction. Thus, the trust model should be able to detect the correct location. This parameter is either calculated or assumed to be shared between peers. When the model detects false location information of a vehicle, it will be discarded.

Vehicular Network System comprises of several vehicles. Every vehicle can communicate with other vehicles by using short radio signals dedicated to short-range communication DSRC (5.9 GHz), within a 1-kilometer range area. The communication between each vehicle is an Ad Hoc communication that means each connected node can move freely; usually, in a VANET, each node is supposed to have an onboard unit (OBU). The OBU enables vehicles to share messages with another vehicle in a prescribed coverage area. The coverage area is based on multiple factors, and they are the position and height of the transmitting antenna. Based on coverage, we present validation mechanisms to provide location closeness in VANET. In our approach, we use four different methods to calculate the location closeness. The trusted zone consists of the Road Side Trust Zone coverage area $\mathrm{RSU}_{\mathrm{TZ}}$, vehicle trust zone coverage area V_{TZ}, and vehicle zone coverage area $V_Z(V_r, V_s)$ for the sender and receiver vehicle:

$$L_C = \left\{ \begin{array}{ll} 1 & \text{if } V_L \in \left|\mathrm{RSU}_{\mathrm{TZ}}\right| \cap \left|V_{\mathrm{TZ}}\right| \\ \dfrac{2}{3} + \dfrac{1}{\left|\mathrm{Send}_{\mathrm{loc}} - \mathrm{Recv}_{\mathrm{loc}}\right|} & \text{if } V_L \in \dfrac{|\mathrm{RSU}|}{\left|V_{\mathrm{TZ}}\right|} \\ \dfrac{1}{3} + \dfrac{1+\gamma}{\left|\mathrm{Send}_{\mathrm{loc}} - \mathrm{Recv}_{\mathrm{loc}}\right|} & \text{if } V_L \in \dfrac{\left|V_Z\right|}{|\mathrm{RSU}|} \\ 0 & \text{if } V_L \notin \dfrac{\left|V_{\mathrm{TZ}}\right|}{\left|\mathrm{RSU}_{\mathrm{TZ}}\right|} \end{array} \right\}. \tag{1}$$

The equation shows that the vehicle received a message from several sources, and based on the received message, calculate L_C to trust the message or discard it. We assume four different cases to calculate location closeness, the distance between the two nodes, the distance between the sender and RSU, and location closeness based on L_C. Here, in our scenario, we assume the coverage area of RSU is (50, 50), whereas the radius is 25.

4.1.2. Integrity. Data integrity ensures the genuineness of a message in terms of modification. The message exchange is one of the essential services of VANET applications. A message should be delivered timely, and accurate information for drivers should be provided to assure safety and enhance travel experiences. Due to the distributed, wireless, and open nature of the vehicular network, it faces serious security challenges. This may lead to a need for common security metrics to quantify the efficacy of VANET security measures. Here, we are using the WAVE (Wireless Access in

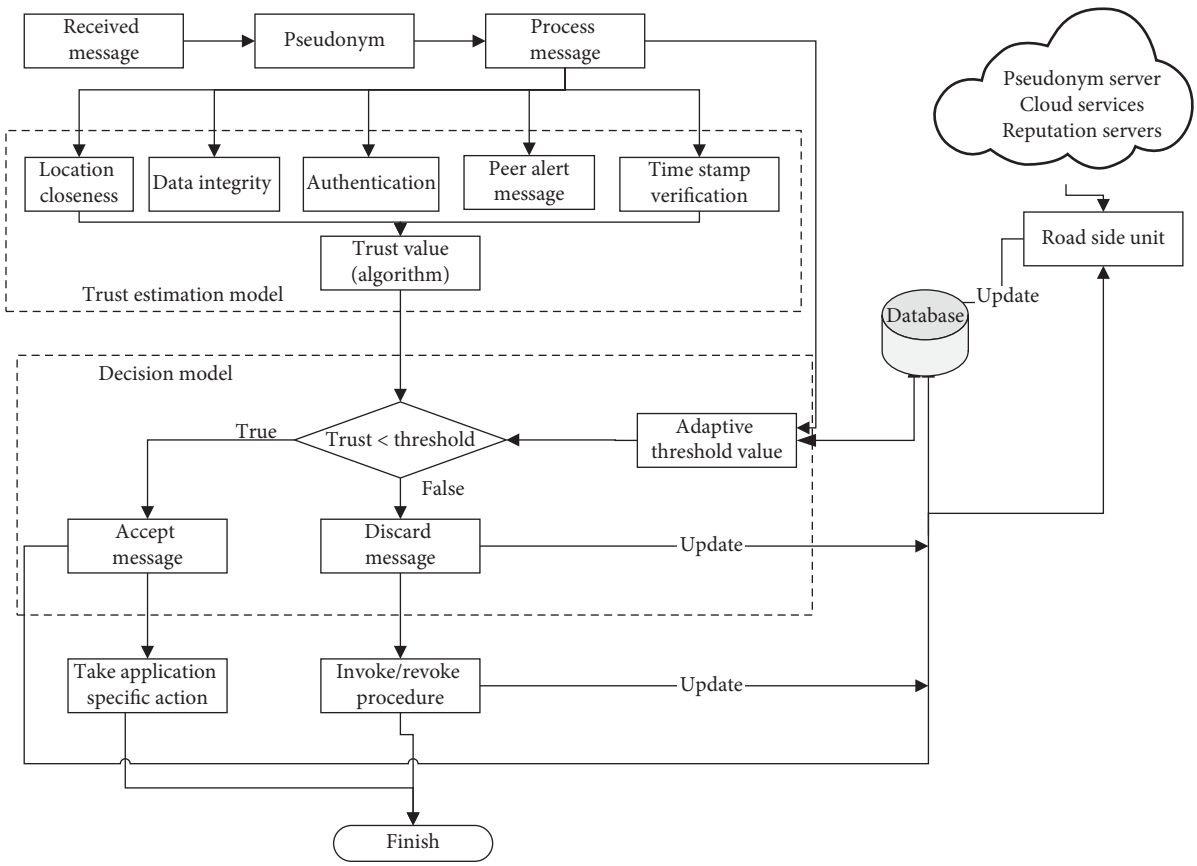


Figure 2: Proposed trust model.

Vehicular Environments) application to secure the content. WAVE specifications can assist V2V and V2I wireless communications, and these functionalities can be utilized to improve vehicle operational safety. Integrity prevents the unauthorized modification of messages in the transmission of the message between V2V. The integrity of considered applications is violated when the correctness and appropriateness of the content of a message are modified, destroyed, or deleted. Data integrity is assured that the message from a sender is protected by the hashing algorithm. To address any security limitations which are inherent mostly in wireless communications, the WAVE standard aims to enhance vehicle safety, to lessen traffic congestion, to activate services for vehicle maintenance, and to provide the potential for new commercial services. The hash algorithm SHA-256 will be used for integrity, as shown in Figure 3.

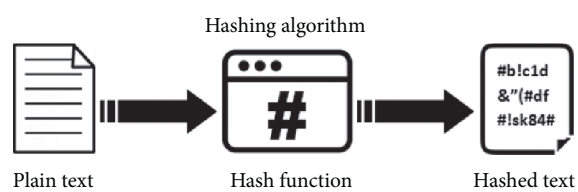


Figure 3: Data integrity hashing algorithm.

4.1.3. Authentication. Authentication is the process of proving something to be true, genuine, or valid. It is compulsory to identify a vehicle that sorts out the genuine sender and receiver. This ensures the identity first to kick out intruders and lower the chance of information loss [77]. The receiver vehicle must be able to verify whether the message is transmitted by a true sender vehicle [78].

WAVE security approaches use a Public Key Infrastructure (PKI) [79]. Here, in our TM, we use Public Key Infrastructure (PKI) scheme for authentication. The V2V and V2I communication in both cases is authenticated. The PKI scheme is scalable [8]. The nature of VANET is that some vehicles are moving quickly and changing the coverage area of one TM to another TM [80, 81]. A huge number of keys are required so that if the numbers of vehicles are increased in the TM area, all vehicles will receive keys. When certificates become invalid for any reason, that certificate will be revoked and updated information will be sent to the database. The revoked information is communicated by RSU in the trust area to other vehicles by generating Peer Alert Message. In our TM public key, cryptography confirms authenticity, integrity, confidentiality, and nonrepudiation [82]. In the scheme, if both vehicles wanted to interact, they have to exchange their public keys authentically, and the process requires the preliminary distribution of public keys. On the contrary, the private key is held only by other vehicles. Here, in our scenario, Public keys are generated by the

RSU and distributed through a secured channel to the vehicles. The Distribution of key comprises the initialization process, registration process, certification, and key updating in the case required [83]. All keys used have a validity date which is updated based on usage. RSU is hosting the public keys as well as the Certificate Revocation List (CRL). Furthermore, it is connected with a centralized database, in distributed manners, as shown in Figure 2.

4.1.4. Peer Alert Message. Message received from peers shares information regarding road condition or safety, and other options regarding the information can be trusted [32].

Peers (vehicles) in a VANET interact with each other by sharing road condition and safety information, to improve passenger and road safety and to effectively route traffic through dense urban areas. These systems concentrate primarily on ensuring the reliable delivery of messages among peers. Here, in our scenario, RSU generates peer alert message to inform vehicle in the coverage area about the safety and untrusted vehicle in the coverage area of the RSU trust zone.

The peer message is generated to inform about the critical condition of the situation regarding safety and security. Figure 4 shows if the peer alert message generated our TM model, we assign the higher wait regarding other parameters used in the TM.

4.1.5. Time Stamp Verification. VANET applications are time critical, and the safety messages are received from the neighboring nodes. Disseminating incorrect time information in the safety message has a severe impact on the security of VANET applications, time verification, and correctness in the VANET.

VANET applications are time critical, and the safety messages are received from the neighboring nodes. Disseminating incorrect time information in the safety message has a severe impact on the security of VANET applications.

4.1.6. Time Stamp Verification Algorithm (Packet)

t_{pkt} = fetch – timestamp (pkt)

Calculate $t_e = t_r - ((\text{dist}(V_1; V_2))/C)$

C = Speed of light
V_1 = Vehicle one
V_2 = Vehicle second
t_e = Event time
t_r = Received time

if $(t_e == t_{\text{pkt}})$
return 1
else
return 0;

The algorithm of Time Stamp Verification explains that time is verified by comparing the Event time received in a packet and current time. The Event time we are calculating here is the current time minus distance of Vehicle 1 to Vehicle 2 divided by the speed of light [21].

4.2. Decision Model Process. The decision model in our model received as the trust value from TM to decide whether to process the message or discard it based on a threshold value. If the trust value is less than the threshold value, a TRUE message is generated, and the decision box accepts the value, sends an update to a database, and takes an application-specific decision. Our TM is for two types of applications that are safety and traffic efficiency. If the threshold value exceeds, the threshold value message is discarded and the FALSE message is generated. False generated message is sent to discard, and update is sent to the database. On the basis of false generated message value, invoke/revoke procedure decides to invoke or revoke the message. Road Side Unit (RSU) is the trusted unit in the model. RSU will provide the initial trust value to all vehicles in the region of interest. All vehicles will have a unique ID in the region. RSU generated an alert message to inform about the malicious vehicle in the region of interest, and this alert message helps vehicles in the region not trust the information received from the malicious node. The decision model in our model received a trust value from TM to decide whether to process the message or discard it depending on the threshold value. If the trust value is less than the threshold value, a TRUE message is generated, and the decision box accepts the value, sends an update to a database, and takes an application-specific decision. Our TM is for two types of applications, which are safety and traffic efficiency.

(1) If the threshold value exceeds, the threshold value message is discarded and the FALSE message is generated.
(2) False generated message is sendt to discard and update is sent to the database. On the basis of false generated message value, invoke/revoke procedure decides to invoke or revoke the message.
(3) Road Side Unit (RSU) is the trusted unit in the model. RSU will provide the initial trust value to all vehicles in the region of interest. All vehicles will have a unique ID in the region. RSU generated an alert message to inform about a malicious vehicle in the region of interest, and this alert message helps vehicles in the region not trust the information received from the malicious node.

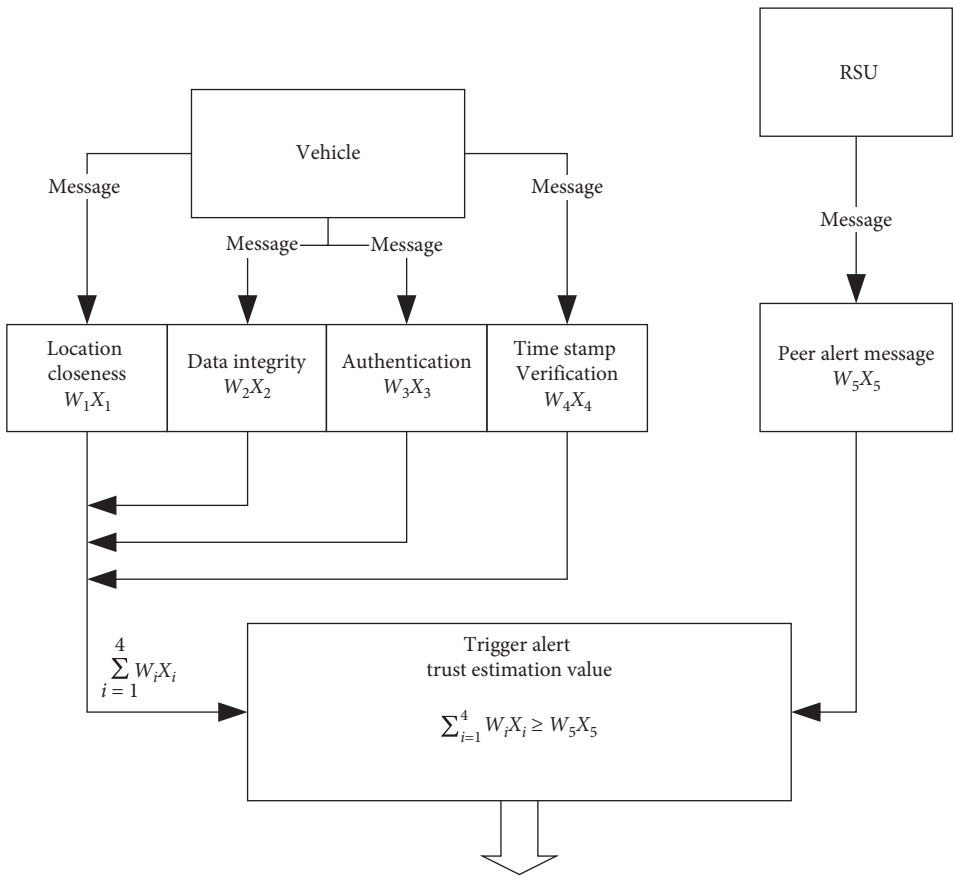


Figure 4: Peer alert message flow diagram.

4.2.1. Trust Calculation Algorithms

w_5P_R = Initial Trust value provided by RSU

```
        #   Trust model
(1)     If L_C = 0 || T_S = 0 || AU = 0 ||DI = 0        # P_1 = L_C (Location closeness)
                                                        # P_2 = T_S (Timestamp)
                                                        # P_3 = AU (Authentication)
                                                        # P_4 = DI (Data integrity)
(2)     Trust = 0
(3)     Else
(4)         If Peer Recommendation available
(5)
        Trust = ((Σ_{i=0}^{4} w_iP_i + w_5P_R)/5)
(6)     else
(7)
        Trust = ((Σ_{i=0}^{4} w_iP_i)/4)
(8)     Endif
(9)         Endif
        # Decision model
(10)    ğ = calculate threshold (s, r, A_U)          #s = sender
                                                      #r = receiver
                                                      #Au = authentication
(11)    if      trust < ğ
(12)    accept message
(13)    else
(14)    discard message
(15)    Invoke-revoke procedure (s, r, T, ğ)          T = Trust, ğ = threshold
(16)    endif
```

5. Evaluations

In this section, the proposed lightweight trust model is evaluated based on the IEEE 802.11p standard. To evaluate the performance of the lightweight TM, the weighted voting method is used which is universally used in trust management schemes for wireless and vehicular networks [84–86]. The performance lightweight TM is evaluated against the MARINE [87] trust management scheme. Furthermore, the performance of our trust model is evaluated in the presence of four variants of Man in the Middle (MiTM) attacker. Moreover, the efficiency of the model is compared to a MARINE trust model based on the weighted voting method. In scientific research, the facility to use computer models and simulator programs to simulate a nearly real-world scenario facilitates a rapid and comparatively inexpensive study of complex real-time issues. Furthermore, than time and cost, simulation using computing resources can enable a view into experimentation. VANET research computer simulation permits research to build up applications and models for utilizing in real life before applying to cars and drivers. To deliver practical usable and realistic scenario-based results, the simulated system must be an accurate representation of real-road infrastructure. To present the real-world scenario-based simulation, in this study, we use UTM as the reference map to simulate the traffic pattern. Figure 5 shows the selected

Figure 5: Traffic map based on simulation.

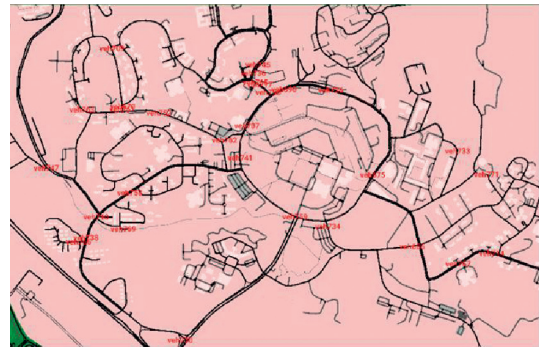


Figure 6: Traffic (vehicles) movement.

area on which we will run the different simulations by changing and varying different traffic-based patterns.

Map 1, shown in Figure 5, will import in SUMO to simulate the traffic patterns. The map has some roads: one-way, single line, double line, two ways, number of signals, speed breakers, and bridges.

Figure 6 shows the movement of the vehicle inside the area, and every vehicle has a unique vehicle ID.

Table 1 provides details of the simulation values, which we will use in our simulation scenario. Road traffic simulation is performed by SUMO such as road length, several lanes, and speed of vehicles, and other details are listed in Table 1. Physical network communication of vehicles and RSU will be performed by using OMNET++ such as Frequency, packet size, and transmission rate, and transmission power. VEINS will integrate the physical and network structure scenarios. According to [88], most of the vehicles in the VANET are legitimate and behave honestly in the network. Consequently, to investigate the behavior of TM, the number of malicious nodes in the different network simulation scenarios will be varied from 10% to 50% in OpenStreetMap [89–91]. To evaluate and assess the TM the well-known machine learning evaluation parameters are used are Precision (P), Recall (R), and F-score. The Precision (P), Recall (R), and F-score are defined as follows:

Precision (P): the term Precision (P) is defined as the ability of TM to precisely forecast the trustworthiness of an event. Let P_M = number of real malicious nodes caught probability and P_U = total number of untrustworthy nodes caught probability. So,

$$\text{Precision } (P)\text{: } \frac{P_M}{P_U}. \tag{2}$$

Recall (R): the term Recall (R) is described as the capability of TM to predict absolute malicious content disseminating by the nodes. Let P_M = number of real malicious nodes caught probability and P_T = total number of truly malicious nodes:

$$\text{Recall } (R)\text{: } \frac{P_M}{P_T}. \tag{3}$$

F-Score: the term F-Score is described as the weighted average of Precision (P) and Recall (R). Moreover, accuracy of TM depends on F-Score. The higher F-Score values correspond more accurately TM. F-Score is defined as

$$F - \text{Score} = 2^* \frac{(P)^* (R)}{(P) + (R)}. \tag{4}$$

Trust variation metrics: in the paper, trust-related metrics is also considered, which illustrates the capability of TM and its efficiency to forego real events in a vehicular network [92]. In particular, to check, given three terms are described.

5.1. Attacker Scenario 1: Identity and Content Tempering. The graph in Figure 7 depicts the accuracy of the trust model on the base of attacker model 1. The two important considerations here are that the attacker is changing the content of safety messages and his identity; furthermore, the adversary tempering trust rating is within the coverage area. The precision and recall of the trust model are illustrated in Figures 8 and 9, and it can be drawn that the smaller number of MiTM attackers achieved high precision as well as recall. Moreover, in this case, if the number of MiTM attackers is increased holding tempering capability, this will result in decreasing corresponding precision and recall. Increasing MiTM attackers in the coverage area generates a high volume of compromised messages, resulting in limiting the ability of a vehicle to distinguish between legitimate and malicious messages. The presented trust model achieved high accuracy in term of F-score by comparing with MARINE, as shown in Figure 7. Furthermore, in terms of tempering ability, our model is more accurate compared to MARINE and assures accuracy around 88% with 11% MiTM attackers, whereas MARINE has 77% accuracy in terms of F-score.

Table 1: Simulation detailed parameters.

Parameters	Value approximation	Values used in simulation
Road length	1300–1400 meters	1400 meter
Number of road lanes	According to map	According to map
A frequency of vehicles entering per hour	0 to 4000 per hour	3000 per hour
Desired speed	10–70 km/hour	40 km/hour
Number of signals	0 to 2	1
Speed at signal	0 to 10 km/hr	0 when signal red 5 km/hr when a signal is orange 10 km/hr signal is green
Frequency	5.9 GHz for V2V	5.9 GHz for V2V
Transmission propagation vehicle	0 to 25 meters	0 to 25 meters
Transmission propagation RSU	0 to 50 meters	0 to 50 meters
Packet size	44 to 1000 bytes	200 bytes
Transmission rate	4–6 Mbps	6 Mbps
Transmission power	17–20 dBm EIRP	18 dBm EIRP
MAC protocol	IEEE 802.11p	IEEE 802.11p
Network protocol	IEEE 1609.4	IEEE 1609.4

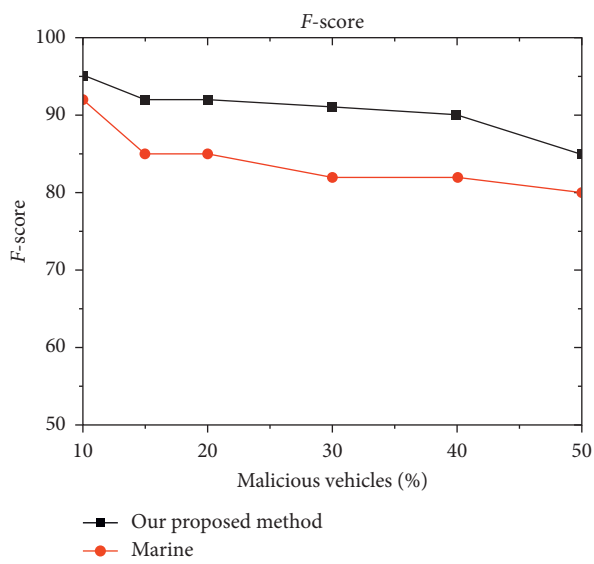


Figure 7: F-score attacker-1.

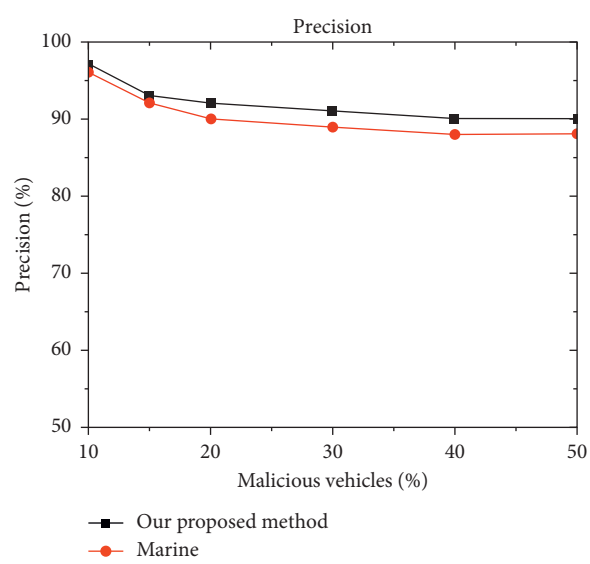


Figure 8: Precision based on attack-1.

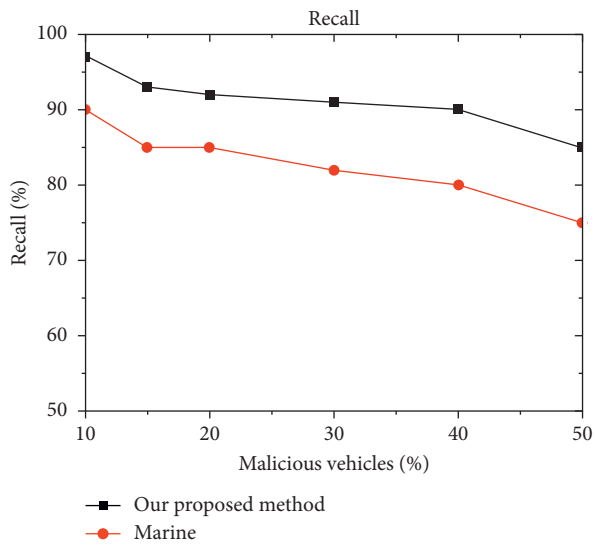


Figure 9: Recall based on attack 1.

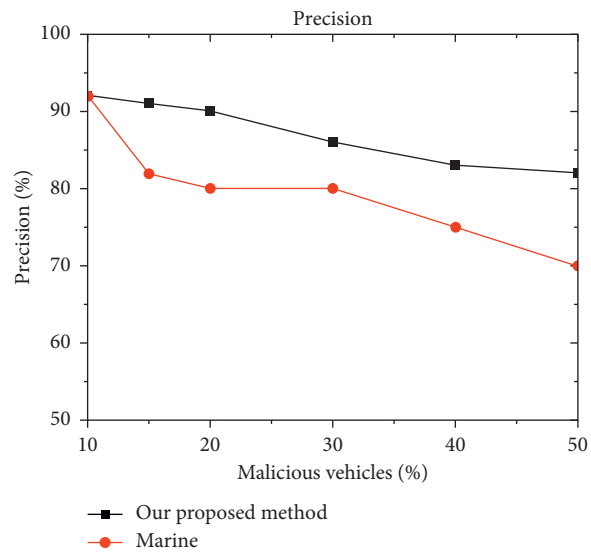


Figure 10: Precision attack 2.

5.2. Attacker Model-2 Is Based on Dropping and Delaying Messages. In this attacker model, the accuracy is measured based on precision, recall, and *F*-score. The malicious node in the coverage area is deliberately delaying and dropping the messages. Delaying and dropping of messages will delay the significant information received in time. Figures 10–12 highlighted the impact of delaying and dropping messages as the number increased the precision and *F*-score and recall decreased. Our trust model is efficient in finding such an attacker reason that the lower layer of the node detects the vehicle applying MiTM attacks. In the case of coverage area based on the high volume of MiTM attackers, 25% our proposed model assured around 82% of recall and 87% of precision values. This significantly concludes that our model is efficient in terms of identifying a malicious node in a network.

On the contrary, Figure 12 shows the accuracy in terms of *F*-score, and our trust model has high accuracy compared to MARINE. The percentage of malicious vehicles increased from 5 to 50 than the accuracy, which also decreased from 92% to 83% almost as compare to MARINE. The MARINE accuracy ranges in the same scenario from 91.5% to 72%. In this attack scenario, it concluded that our trust model is attack resistant to MiTM attacks. Furthermore, the trust model assures to disseminate trusted messages in the case high volume of malicious nodes.

5.3. Advance Zig-Zag Attack. To test the trust model at high-efficiency, several experiments are conducted which are reflected in Figure 13. MiTM attackers perform intelligently throughout the network to deceive legitimate nodes with disseminate tempered information and propagate compromised messages. The introduction of advanced zig-zag attack patterns has a drastic impact and reduced significantly precision and recall, and this is given in Figures 14 and 15. This advanced zig-zag attack will

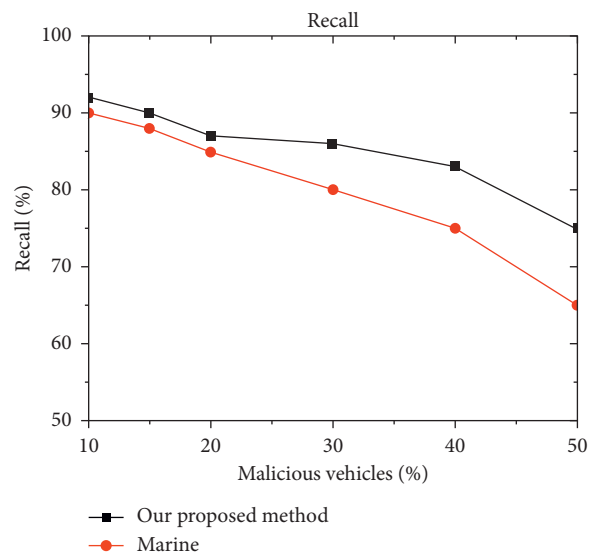


Figure 11: Recall attack 2.

help attackers to deceive legitimate vehicle to identify the attacker. In this case, our trust model helps the node to detect such an attacker with the zig-zag attack pattern. The following reasons explained trust model detection capability. Primarily, the trust establishment at lower layers helps early detection of MiTM attackers in a node-centric scenario. Secondary, Figure 13 reflects the *F*-score of the trust model used to measure the accuracy of the proposed method in identifying malicious content and detecting MiTM attackers. Finally, the overall performance notably decreased by varying the MiTM attackers' pattern in the network. Furthermore, changing the attack pattern will considerably reduce recall and precision. In this attack scenario, the accuracy of the trust model is more accurate as compared to MARINE in terms of *F*-score. Furthermore, in an attack scenario where 35% of vehicles are

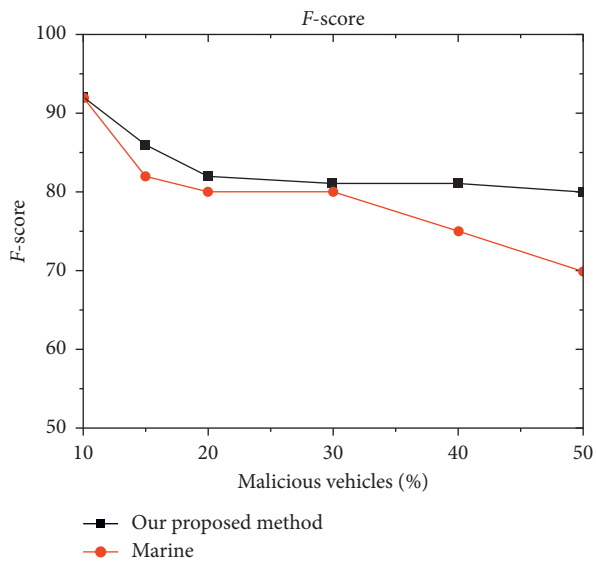


Figure 12: *F*-score attack-2.

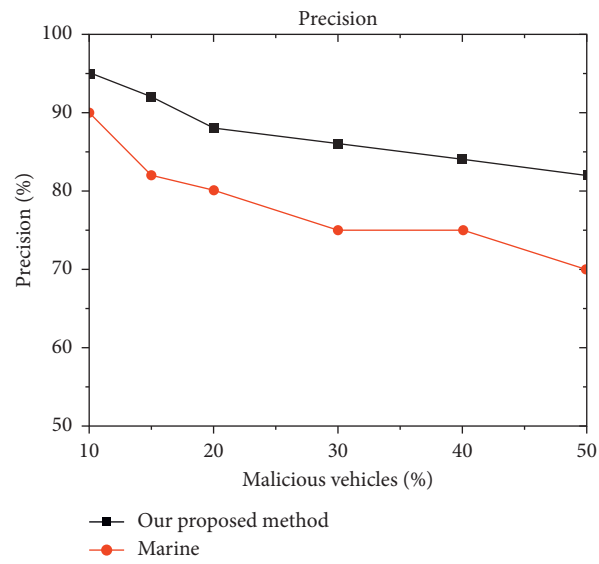


Figure 14: Precision attack-3.

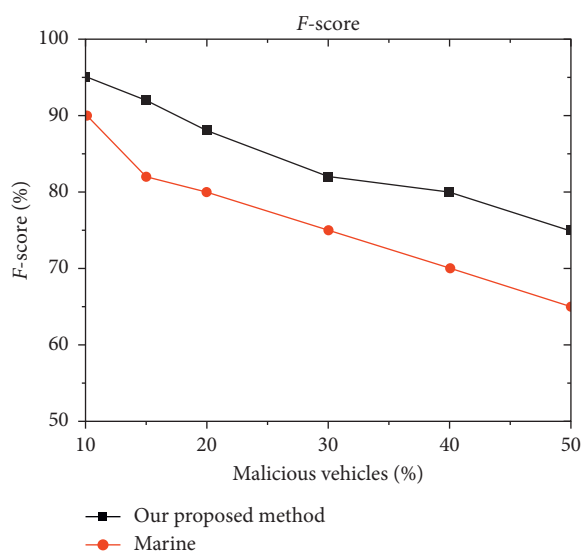


Figure 13: *F*-score attack-3.

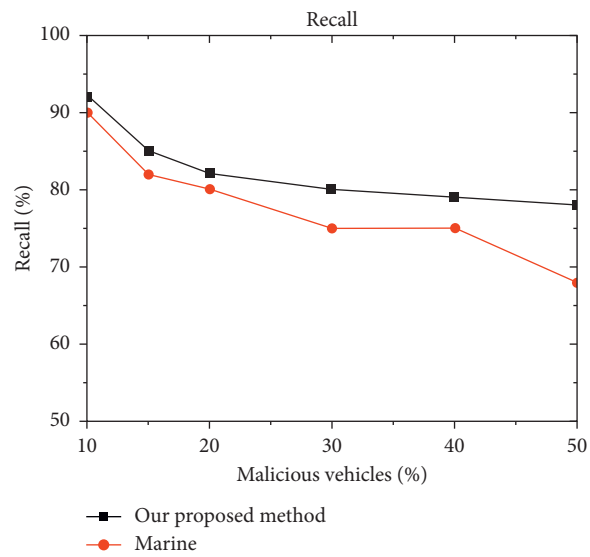


Figure 15: Recall attack 3.

malicious, the accuracy of the model is around 82%, whereas the MARINE was approximately 66%.

5.4. Trust Perspective Measurement. The trust establishment is a key parameter which enhances the security of nodes against inside attackers. The presented trust model in Figure 16 shows efficiency to classify and identify malicious content concerning the trust. In this scenario, a MiTM attack is generated in the VANET, and the trust of the network is decreased by increasing the malicious content in the network. The main reason behind trust decrease is that increasing malicious content in the network limits the ability of a legitimate vehicle to classify legitimate messages received. The presented trust model is capable of classifying and identifying legitimate nodes in the presence of attackers. The key factors for this are, primarily, presented the trust model which intelligently classify malicious messages as well as identify malicious nodes at lower layers. Secondarily, the presence of a role-oriented evaluator node in the network is to help the legitimate vehicle to process and true events. Finally, the evaluator node based on the abovementioned points will distinguish between an attacker and a legitimate node. In the current scenario, with 40% of MiTM attackers, the presented trust model achieves 86% of the trust level, now, as compared to MARINE, and it was 83%. Figures 16 and 17 describe the trust for both legitimate and untrustworthy nodes correspondingly. The given metrics are foremost important as they play an important role in the measurement of efficiency of the presented trust model

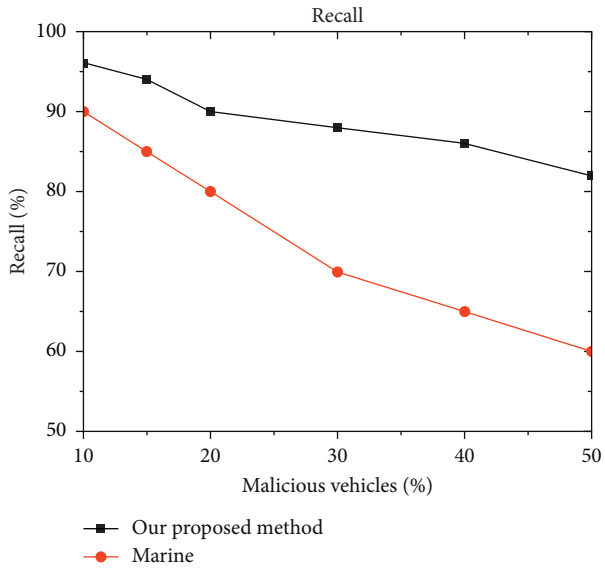


Figure 16: Recall attack 4.

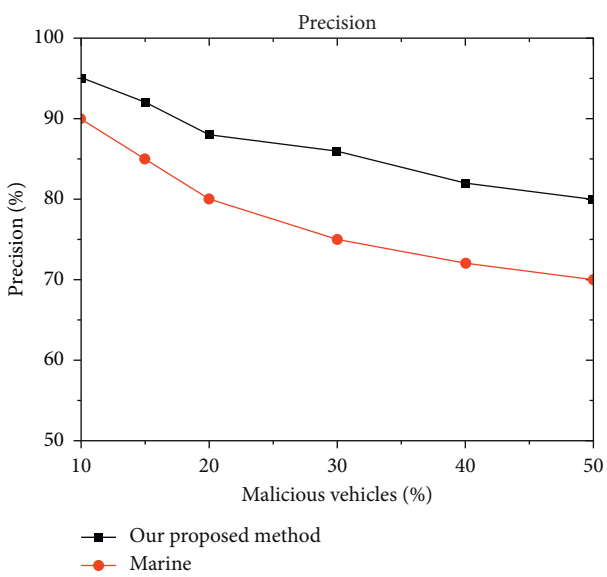


Figure 17: Precision attack 4.

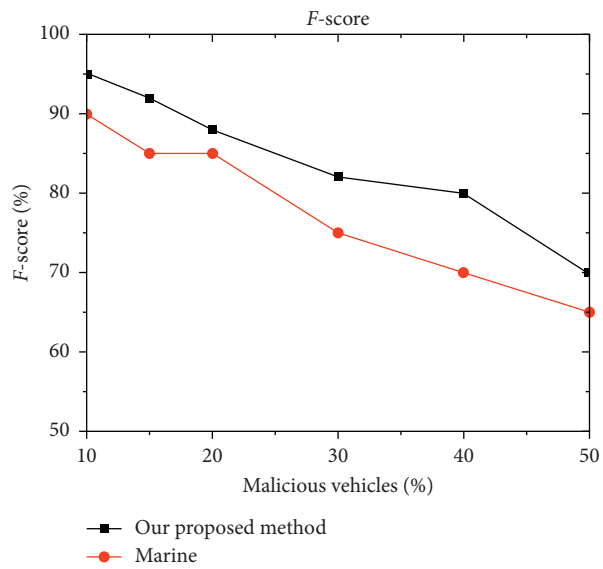


Figure 18: F-score attack-4.

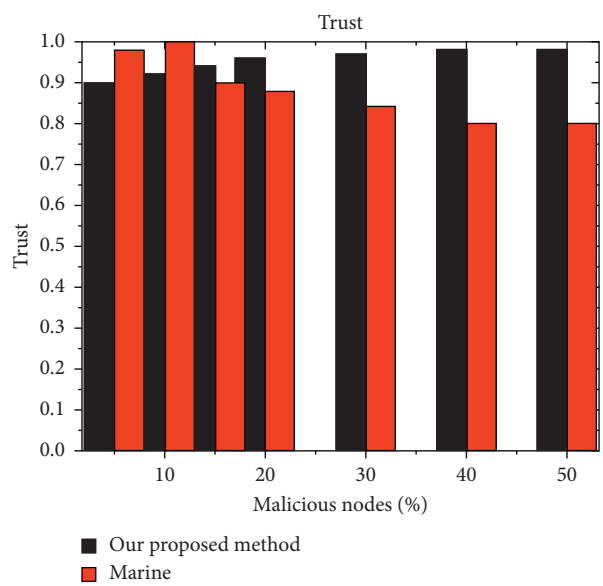


Figure 19: Impact of trust metric.

for evaluating trust in the received content. The presented trust model is a higher trust comparing MARINE, and the mentioned points below further elaborate it. The information in Figure 18 describes that the trust value within the legitimate vehicle is always more than the threshold value; here, in the presented trust model case, it is 0.5 even though if a high volume of MiTM attackers are present in the network. In this regard, it can be said that the presented trust model experiences a small number of false positive in the VANET. Furthermore, in Figure 19, trust between the MiTM vehicles is considerably lower than a threshold value, and this is because a very small number of false negative are generated by the presented model.

6. Conclusion

A privacy-preserving attack-resistant lightweight trust model is proposed to increase Internet of vehicles (IoV) security by promptly identifying dishonest nodes and revoking its credential in the MiTM attack scenario. Besides, for the trust model in terms of lightweight, fewer arithmetic operations are used to reduce the complexity, such as square root log and complex geometry of the model. The performance of the trust model is measured in the presence of four variants of Man in the Middle (MiTM) attacker and compared with a MARINE trust model based on the weighted voting method. Furthermore, for addressing the privacy and security requirement, pseudonym scheme is used. All nodes in the VANET established trust provided by RSU initially,

which is a trusted source in the network; once the trustworthiness of the sender is verified then the content can be processed. The results have validated the lightweight features of the trust model such as less arithmetic complexity, and low memory consumption leads the model to attain high detection accuracy in MiTM attacks. It has also manifested that the proposed model outperforms in terms of *F*-score, recall, and precision as compared to the MARINE model. Moreover, the proposed model has achieved a high trust level with 40% of MiTM attackers, and in terms of F-score 95%, whereas the MARINE model has 90%, which leads to the model to attain high detection accuracy. Despite the fact, the privacy-preserving attack-resistant trust model due to lightweight enables the participating nodes to hastily identify dishonest nodes and prevent them to poison the network from malicious content, and it also remains stable even when the number of malicious vehicles is increasing.

Data Availability

The data used to support this study are available at https://www.openstreetmap.org/#map=15/1.5645/103.6403.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. Latif, S. Mahfooz, B. Jan et al., "Multicriteria based next forwarder selection for data dissemination in vehicular ad hoc networks using analytical network process," *Mathematical Problems in Engineering*, vol. 2017, Article ID 4671892, 18 pages, 2017.

[2] World Health Organization, "Global status report on road safety 2018: summary," World Health Organization, Geneva, Switzerland, 2018.

[3] I. A. Rai, R. A. Shaikh, and S. R. Hassan, "A hybrid dual-mode trust management scheme for vehicular networks," *International Journal of Distributed Sensor Networks*, vol. 16, no. 7, 2020.

[4] N. Gupta, R. Manaswini, B. Saikrishna, F. Silva, and A. Teles, "Authentication-based secure data dissemination protocol and framework for 5G-enabled VANET," *Future Internet*, vol. 12, no. 4, p. 63, 2020.

[5] F. Al-Turjman and J. P. Lemayian, "Intelligence, security, and vehicular sensor networks in internet of things (IoT)-enabled smart-cities: an overview," *Computers & Electrical Engineering*, vol. 87, p. 106776, 2020.

[6] M. Balta and İ. Özçelik, "A 3-stage fuzzy-decision tree model for traffic signal optimization in urban city via a SDN based VANET architecture," *Future Generation Computer Systems*, vol. 104, pp. 142–158, 2020.

[7] A. Bhargava, S. Verma, B. K. Chaurasia, and G. S. Tomar, "Computational trust model for internet of vehicles," in *Proceedings of the 2017 Conference on Information and Communication Technology (CICT)*, pp. 1–5, Gwalior, India, November 2017.

[8] R. Iqbal, T. A. Butt, M. Afzaal, and K. Salah, "Trust management in social Internet of vehicles: factors, challenges, blockchain, and fog solutions," *International Journal of Distributed Sensor Networks*, vol. 15, no. 1, 2019.

[9] I. A. Sumra and A. N. Akhtar, "Applications of internet of vehicle (IoV): a survey," *LGURJCSIT*, vol. 4, no. 2, pp. 59–70, 2020.

[10] M. K. Priyan and G. U. Devi, "A survey on internet of vehicles: applications, technologies, challenges and opportunities," *International Journal of Advanced Intelligence Paradigms*, vol. 12, no. 1-2, pp. 98–119, 2019.

[11] T. Halabi and M. Zulkernine, "Trust-based cooperative game model for secure collaboration in the internet of vehicles," in *Proceedings of the ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pp. 1–6, Shanghai, China, May 2019.

[12] R. A. Shaikh and A. S. Alzahrani, "Trust management method for vehicular ad hoc networks," in *Proceedings of the International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, pp. 801–815, Noida, India, January 2013.

[13] R. A. Shaikh, H. Jameel, S. Lee, S. Rajput, and Y. J. Song, "Trust management problem in distributed wireless sensor networks," in *Proceedings of the 12th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA'06)*, pp. 411–414, Sydney, Australia, August 2006.

[14] I. Memon and H. T. Mirza, "MADPTM: mix zones and dynamic pseudonym trust management system for location privacy," *International Journal of Communication Systems*, vol. 31, no. 17, p. e3795, 2018.

[15] E. R. Cavalcanti, J. A. R. de Souza, M. A. Spohn, R. C. d. M. Gomes, and A. F. B. F. d. Costa, "VANETs' research over the past decade," *ACM SIGCOMM Computer Communication Review*, vol. 48, no. 2, pp. 31–39, 2018.

[16] M. Asghar, R. R. M. Doss, and L. Pan, "A scalable and efficient PKI based authentication protocol for VANETs," in *Proceedings of the 2018 28th International Telecommunication Networks and Applications Conference (ITNAC)*, pp. 1–3, Sydney, Australia, November 2018.

[17] S. Sakhreliya and N. Pandya, "Public key infrastructure (PKI) using symmetric key cryptography (SC) in VANETs," *International Journal of Computer Science and Information Technologies*, vol. 5, pp. 3556–3561, 2014.

[18] A. H. Salem, A. Abdel-Hamid, and M. A. El-Nasr, "The case for dynamic key distribution for PKI-based VANETS," 2016, https://arxiv.org/abs/1605.04696.

[19] A. Hesham, A. Abdel-Hamid, and M. Abou El-Nasr, "A dynamic key distribution protocol for PKI-based VANETs," in *Proceedings of the 2011 IFIP Wireless Days (WD)*, pp. 1–3, Niagara Falls, ON, USA, 2011.

[20] S. A. Thileeban, C. S. Narayan, J. Bhuvana, and V. Balasubramanian, "PKI model optimisation in VANET with clustering and polling," in *Proceedings of the International Conference on Innovations In Bio-Inspired Computing and Applications*, pp. 321–329, Kochi, India, December 2018.

[21] R. A. Shaikh and A. S. Alzahrani, "Intrusion-aware trust model for vehicular ad hoc networks," *Security and Communication Networks*, vol. 7, no. 11, pp. 1652–1669, 2014.

[22] F. G. Mármol and G. M. Pérez, "TRIP, a trust and reputation infrastructure-based proposal for vehicular ad hoc networks," *Journal of Network and Computer Applications*, vol. 35, no. 3, pp. 934–941, 2012.

[23] U. F. Minhas, J. Zhang, T. Tran, and R. Cohen, "A multifaceted approach to modeling agent trust for effective communication in the application of mobile ad hoc vehicular networks," *IEEE Transactions on Systems, Man, and*

Cybernetics, Part C (Applications and Reviews), vol. 41, no. 3, pp. 407–420, 2011.

[24] Z. Liu, J. Ma, Z. Jiang, H. Zhu, and Y. Miao, "LSOT: a lightweight self-organized trust model in VANETs," *Mobile Information Systems*, vol. 2016, Article ID 7628231, 15 pages, 2016.

[25] C. Chen, J. Zhang, R. Cohen, and P.-H. Ho, "A trust-based message propagation and evaluation framework in Vanets," in *Proceedings of the 2010 2nd International Conference on Information Technology Convergence and Services*, Cebu, Phillippines, August 2010.

[26] Z. Lu, W. Liu, Q. Wang, G. Qu, and Z. Liu, "A privacy-preserving trust model based on blockchain for vanets," *IEEE Access*, vol. 6, pp. 45655–45664, 2018.

[27] A. Tajeddine, A. Kayssi, and A. Chehab, "A privacy-preserving trust model for VANETs," in *Proceedings of the 2010 10th IEEE International Conference on Computer and Information Technology*, pp. 832–837, Bradford, UK, June 2010.

[28] M. Raya, P. Papadimitratos, V. D. Gligor, and J.-P. Hubaux, "On data-centric trust establishment in ephemeral ad hoc networks," in *Proceedings of the 2008 Proceedings IEEE INFOCOM-The 27th Conference on Computer Communications*, pp. 1238–1246, IEEE, Phoenix, AZ, USA, April 2008.

[29] J. Zhang, "A survey on trust management for Vanets," in *Proceedings of the 2011 IEEE International Conference on Advanced Information Networking and Applications*, pp. 105–112, Singapore, March 2011.

[30] S. Gurung, D. Lin, A. Squicciarini, and E. Bertino, "Information-oriented trustworthiness evaluation in vehicular ad-hoc networks," in *Proceedings of the International Conference on Network and System Security*, pp. 94–108, Madrid, Spain, June 2013.

[31] A. Patwardhan, A. Joshi, T. Finin, and Y. Yesha, "A data intensive reputation management scheme for vehicular ad hoc networks," in *Proceedings of the 2006 3rd Annual International Conference on Mobile and Ubiquitous Systems-Workshops*, pp. 1–8, San Jose, CA, USA, July 2006.

[32] C. Chen, J. Zhang, R. Cohen, and P.-H. Ho, "A trust modeling framework for message propagation and evaluation in VANETs," in *Proceedings of the 2010 2nd International Conference on Information Technology Convergence and Services*, pp. 1–8, Cebu, Phillippines, August 2010.

[33] N.-W. Lo and H.-C. Tsai, "A reputation system for traffic safety event on vehicular ad hoc networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, no. 1, Article ID 125348, 2009.

[34] U. Khan, S. Agrawal, and S. Silakari, "Detection of malicious nodes (DMN) in vehicular ad-hoc networks," *Procedia Computer Science*, vol. 46, pp. 965–972, 2015.

[35] M. Gerlach, "Trust for vehicular applications," in *Proceedings of the Eighth International Symposium on Autonomous Decentralized Systems (ISADS'07)*, pp. 295–304, Sedona, AZ, USA, March 2007.

[36] N. Yang, "A similarity based trust and reputation management framework for vanets," *International Journal of Future Generation Communication and Networking*, vol. 6, no. 2, pp. 25–34, 2013.

[37] A. Jesudoss, S. V. Kasmir Raja, and A. Sulaiman, "Stimulating truth-telling and cooperation among nodes in VANETs through payment and punishment scheme," *Ad Hoc Networks*, vol. 24, pp. 250–263, 2015.

[38] N. Haddadou, A. Rachedi, and Y. Ghamri-Doudane, "Trust and exclusion in vehicular ad hoc networks: an economic incentive model based approach," in *Proceedings of the ComComAP'2013*, pp. 13–18, Hong Kong, China, April 2013.

[39] J. Zhang, J. Cui, H. Zhong, Z. Chen, and L. Liu, "PA-CRT: Chinese remainder theorem based conditional privacy-preserving authentication scheme in vehicular ad-hoc networks," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[40] S. Guleng, C. Wu, X. Chen, X. Wang, T. Yoshinaga, and Y. Ji, "Decentralized trust evaluation in vehicular internet of things," *IEEE Access*, vol. 7, pp. 15980–15988, 2019.

[41] H. Sedjelmaci and S. M. Senouci, "An accurate and efficient collaborative intrusion detection framework to secure vehicular networks," *Computers & Electrical Engineering*, vol. 43, pp. 33–47, 2015.

[42] S. K. Dhurandher, M. S. Obaidat, A. Jaiswal, A. Tiwari, and A. Tyagi, "Securing vehicular networks: a reputation and plausibility checks-based approach," in *Proceedings of the GLOBECOM Workshops (GC Wkshps)*, pp. 1550–1554, IEEE, Miami, FL, USA, December 2010.

[43] K. C. Abdelaziz, N. Lagraa, and A. Lakas, "Trust model with delayed verification for message relay in VANETs," in *Proceedings of the 2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 700–705, Nicosia, Cyprus, August 2014.

[44] F. Dotzer, L. Fischer, and P. Magiera, "Vars: a vehicle ad-hoc network reputation system," in *Proceedings of the Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks*, pp. 454–456, Taormina-Giardini Naxos, Italy, June 2005.

[45] J. M. De Fuentes, A. I. González-Tablas, and A. Ribagorda, "Overview of security issues in vehicular ad-hoc networks," in *Handbook of Research on Mobility and Computing: Evolving Technologies and Ubiquitous Impacts*, pp. 894–911, IGI Global, Hershey, PA, USA, 2011.

[46] J. Petit, F. Schaub, M. Feiri, and F. Kargl, "Pseudonym schemes in vehicular networks: a survey," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 1, pp. 228–255, 2014.

[47] L. Buttyán, T. Holczer, A. Weimerskirch, and W. Whyte, "Slow: a practical pseudonym changing scheme for location privacy in VANETS," in *Proceedings of the 2009 IEEE Vehicular Networking Conference (VNC)*, pp. 1–8, Tokyo, Japan, October 2009.

[48] D. Förster, F. Kargl, and H. Löhr, "PUCA: a pseudonym scheme with strong privacy guarantees for vehicular ad-hoc networks," *Ad Hoc Networks*, vol. 37, pp. 122–132, 2016.

[49] S. Wang and N. Yao, "A RSU-aided distributed trust framework for pseudonym-enabled privacy preservation in VANETs," *Wireless Networks*, vol. 25, no. 3, pp. 1099–1115, 2019.

[50] R. Lu, X. Lin, T. H. Luan, X. Liang, and X. Shen, "Pseudonym changing at social spots: an effective strategy for location privacy in vanets," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 1, pp. 86–96, 2011.

[51] D. Liao, G. Sun, M. Zhang, V. Chang, and H. Li, "Towards location and trajectory privacy preservation in 5G vehicular social network," in *Proceedings of the 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, vol. 2, pp. 63–69, Guangzhou, China, July 2017.

[52] J. Wang, Y. Zhang, Y. Wang, and X. Gu, "RPRep: a robust and privacy-preserving reputation management scheme for pseudonym-enabled VANETs," *International Journal of*

Distributed Sensor Networks, vol. 12, no. 3, Article ID 6138251, 2016.

[53] Y. Pan and J. Li, "Cooperative pseudonym change scheme based on the number of neighbors in VANETs," *Journal of Network and Computer Applications*, vol. 36, no. 6, pp. 1599–1609, 2013.

[54] K. Emara, W. Woerndl, and J. Schlichter, "Context-based pseudonym changing scheme for vehicular adhoc networks," 2016, https://arxiv.org/abs/1607.07656.

[55] B. Mishra, S. K. Panigrahy, T. C. Tripathy, D. Jena, and S. K. Jena, "A secure and efficient message authentication protocol for VANETs with privacy preservation," in *Proceedings of the 2011 World Congress on Information and Communication Technologies*, pp. 880–885, Mumbai, India, December 2011.

[56] J. A. Guerrero-Ibáñez, C. Flores-Cortés, and S. Zeadally, "Vehicular Ad-Hoc networks (Vanets): architecture, protocols and applications," in *Next-Generation Wireless Technologies*, pp. 49–70, Springer, Berlin, Germany, 2013.

[57] S. Kumari, M. Karuppiah, X. Li, F. Wu, A. K. Das, and V. Odelu, "An enhanced and secure trust-extended authentication mechanism for vehicular ad-hoc networks," *Security and Communication Networks*, vol. 9, no. 17, pp. 4255–4271, 2016.

[58] A. Agrawal, A. Garg, N. Chaudhiri, S. Gupta, D. Pandey, and T. Roy, "Security on vehicular ad hoc networks (VANET): a review paper," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 1, pp. 231–235, 2013.

[59] P. Fabian, A. Rachedi, and C. Guéguen, "Programmable objective function for data transportation in the Internet of Vehicles," *Technologies for Emerging Future Wireless Networks*, vol. 31, no. 5, p. e3882, 2020.

[60] W. Li, W. Song, Q. Lu, and C. Yue, "Reliable congestion control mechanism for safety applications in urban VANETs," *Ad Hoc Networks*, vol. 98, Article ID 102033, 2020.

[61] I. Memon, "A secure and efficient communication scheme with authenticated key establishment protocol for road networks," *Wireless Personal Communications*, vol. 85, no. 3, pp. 1167–1191, 2015.

[62] I. Memon and Q. A. Arain, "Dynamic path privacy protection framework for continuous query service over road networks," *World Wide Web*, vol. 20, no. 4, pp. 639–672, 2017.

[63] C. Wang, J. Shen, J.-F. Lai, and J. Liu, "B-TSCA: blockchain assisted trustworthiness scalable computation for V2I authentication in VANETs," *IEEE Transactions on Emerging Topics in Computing*, vol. 2020, 2020.

[64] J. S. Sengar, "SURVEY: reputation and trust management in VANETs," *International Journal of Grid and Distributed Computing*, vol. 8, no. 4, pp. 301–306, 2015.

[65] G. Samara and Y. Al-Raba'nah, "Security issues in vehicular ad hoc networks (VANET): a survey," 2017, https://arxiv.org/abs/1712.04263.

[66] M. A. H. Al Junaid, A. A. Syed, M. N. M. Warip, K. N. F. K. Azir, and N. H. Romli, "Classification of security attacks in VANET: a review of requirements and perspectives," *MATEC Web of Conferences*, vol. 150, p. 6038, 2018.

[67] M. B. Mansour, C. Salama, H. K. Mohamed, and S. A. Hammad, "VANET security and privacy-an overview," *International Journal of Network Security & Its Applications*, vol. 10, 2018.

[68] M. S. Sheikh and J. Liang, "A comprehensive survey on VANET security services in traffic management system," *Wireless Communications and Mobile Computing*, vol. 2019, Article ID 2423915, 23 pages, 2019.

[69] C. A. Kerrache, C. T. Calafate, J.-C. Cano, N. Lagraa, and P. Manzoni, "Trust management for vehicular networks: an adversary-oriented overview," *IEEE Access*, vol. 4, pp. 9293–9307, 2016.

[70] S. S. Tangade and S. S. Manvi, "A survey on attacks, security and trust management solutions in VANETs," in *Proceedings of the 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1–6, Tiruchengode, India, July 2013.

[71] M. S. Al-Kahtani, "Survey on security attacks in vehicular Ad Hoc networks (VANETs)," in *Proceedings of the 2012 6th International Conference on Signal Processing and Communication Systems*, pp. 1–9, Gold Coast, Australia, December 2012.

[72] B. Mokhtar and M. Azab, "Survey on security issues in vehicular ad hoc networks," *Alexandria Engineering Journal*, vol. 54, no. 4, pp. 1115–1126, 2015.

[73] D. Liao, H. Li, G. Sun, M. Zhang, and V. Chang, "Location and trajectory privacy preservation in 5G-Enabled vehicle social network services," *Journal of Network and Computer Applications*, vol. 110, pp. 108–118, 2018.

[74] H. Li, L. Pei, D. Liao, G. Sun, and D. Xu, "Blockchain meets VANET: an architecture for identity and location privacy protection in VANET," *Peer-to-Peer Networking and Applications*, vol. 12, no. 5, pp. 1178–1193, 2019.

[75] Y. Sun, B. Zhang, B. Zhao, X. Su, and J. Su, "Mix-zones optimal deployment for protecting location privacy in VANET," *Peer-to-Peer Networking and Applications*, vol. 8, no. 6, pp. 1108–1121, 2015.

[76] B. Ying, D. Makrakis, and Z. Hou, "Motivation for protecting selfish vehicles' location privacy in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5631–5641, 2015.

[77] S. Ahmed, M. U. Rehman, A. Ishtiaq, S. Khan, A. Ali, and S. Begum, "VANSec: attack-resistant VANET security algorithm in terms of trust computation error and normalized routing overhead," *Journal of Sensors*, vol. 2018, Article ID 6576841, 17 pages, 2018.

[78] A. Mondal and S. Mitra, "TDMAC: a timestamp defined message authentication code for secure data dissemination in VANET," in *Proceedings of the 2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pp. 1–6, Bangalore, India, November 2016.

[79] J.-H. Kang, S.-J. Ok, J. Y. Kim, and E.-G. Kim, "Software implementation of wave security algorithms," *Journal of the Korea Academia-Industrial Cooperation Society*, vol. 15, no. 3, pp. 1691–1699, 2014.

[80] H. Hartenstein and K. P. Laberteaux, "A tutorial survey on vehicular ad hoc networks," *IEEE Communications Magazine*, vol. 46, no. 6, pp. 164–171, 2008.

[81] S. Shrivastava, "V2V vehicle safety communication," in *Connected Vehicles*, pp. 117–155, Springer, Berlin, Germany, 2019.

[82] H. Zhao, D. Sun, H. Yue, M. Zhao, and S. Cheng, "Dynamic trust model for vehicular cyber-physical systems," *International Journal of Network Security*, vol. 20, no. 1, pp. 157–167, 2018.

[83] H. Hasrouny, A. E. Samhat, C. Bassil, and A. Laouiti, "Trust model for secure group leader-based communications in VANET," *Wireless Networks*, vol. 25, no. 8, pp. 4639–4661, 2018.

[84] J.-M. Chen, T.-T. Li, and J. Panneerselvam, "TMEC: a trust management based on evidence combination on attack-

resistant and collaborative internet of vehicles," *IEEE Access*, vol. 7, pp. 148913–148922, 2018.

[85] W. Li and H. Song, "ART: an attack-resistant trust management scheme for securing vehicular ad hoc networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 960–969, 2016.

[86] M. A. Azad, S. Bag, S. Parkinson, and F. Hao, "TrustVote: privacy-preserving node ranking in vehicular networks," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 5878–5891, 2018.

[87] F. Ahmad, F. Kurugollu, A. Adnane, R. Hussain, and F. Hussain, "MARINE: man-in-the-middle attack resistant trust model in connected vehicles," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3310–3322, 2020.

[88] B. Parno and A. Perrig, "Challenges in securing vehicular networks," in *Proceedings of the Workshop on Hot Topics in Networks (HotNets-IV)*, pp. 1–6, College Park, MD, USA, November 2005.

[89] D. Krajzewicz, G. Hertkorn, C. Rössel, and P. Wagner, "SUMO (simulation of urban mobility)-an open-source traffic simulation," in *Proceedings of the 4th Middle East Symposium on Simulation and Modelling (MESM20002)*, pp. 183–187, Sharjah, UAE, 2002.

[90] M. Haklay and P. Weber, "Openstreetmap: user-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[91] Veins, Vehicles in Network Simulation, The Open Source Vehicular Simulation Framework, http://veins.car2x.org.

[92] F. Ahmad, A. Adnane, F. Kurugollu, and R. Hussain, "A comparative analysis of trust models for safety applications in IOT-enabled vehicular networks," in *Proceedings of the 2019 Wireless Days (WD)*, pp. 1–8, Manchester, UK, February 2019.

Hindawi
Scientific Programming
Volume 2020, Article ID 8866041, 12 pages
https://doi.org/10.1155/2020/8866041

Research Article

An Efficient Skewed Line Segmentation Technique for Cursive Script OCR

Saud Malik,[1] Ahthasham Sajid,[2] Arshad Ahmad,[3] Ahmad Almogren,[4] Bashir Hayat,[5] Muhammad Awais,[6] and Kyong Hoon Kim[7]

[1]*COMSATS University Islamabad, Attock Campus, Islamabad 43600, Pakistan*
[2]*Department of Computer Science, Faculty of ICT, BUITEMS, Quetta, Baluchistan 87300, Pakistan*
[3]*Department of IT and Computer Science, Pak-Austria Fachhochschule: Institute of Applied Sciences & Technology, Khanpur Road, Mang, Haripur 22620, Pakistan*
[4]*Department of Computer Science, College of Computer & Information Sciences, King Saud University, Riyadh 11633, Saudi Arabia*
[5]*Institute of Management Sciences, Peshawar 25000, Pakistan*
[6]*School of Computing and Communications, Lancaster University, Bailrigg, Lancaster LA1 4YW, UK*
[7]*School of Computer Science & Engineering, Kyungpook National University, Daegu 41566, Republic of Korea*

Correspondence should be addressed to Arshad Ahmad; yaarshad@gmail.com

Received 27 August 2020; Revised 4 November 2020; Accepted 8 November 2020; Published 4 December 2020

Academic Editor: Shaukat Ali

Segmentation of cursive text remains the challenging phase in the recognition of text. In OCR systems, the recognition accuracy of text is directly dependent on the quality of segmentation. In cursive text OCR systems, the segmentation of handwritten Urdu language text is a complex task because of the context sensitivity and diagonality of the text. This paper presents a line segmentation algorithm for Urdu handwritten and printed text and subsequently to ligatures. In the proposed technique, the counting pixel approach is employed for modified header and baseline detection, in which the system first removes the skewness of the text page, and then the page is converted into lines and ligatures. The algorithm is evaluated on manually generated Urdu printed and handwritten dataset. The proposed algorithm is tested separately on handwritten and printed text, showing 96.7% and 98.3% line accuracy, respectively. Furthermore, the proposed line segmentation algorithm correctly extracts the lines when tested on Arabic text.

1. Introduction

The OCR systems have standard measures, different types, and rich history. Urdu language is widely spoken and understood in mainly South Asian countries. In contrast to its vast usage, little to no improvement has been made for recognition of its script [1]. The script recognition system for Urdu printed and handwritten text has been approached recently as compared to the OCR systems for other scripts. This void of research is mainly due to the lack of benchmark datasets, dictionaries, and other necessary factors.

In recent years, the OCR system has attracted a lot of researchers' attention towards cursive text scripts, while the text of languages like Urdu and Arabic is still far behind in attaining convincing accuracy [2]. The writing pattern of any language is the main reason for the high or low accuracy rate of text recognition. In this context, cursive text (Arabic, Urdu, etc.) is having more accuracy concerns. Line segmentation and recognition of line in the cursive script is a hard task because of their words shape and composition. Some of the cursive scripts are mentioned in Table 1. OCRs of cursive languages are not so mature that there is still room for improving accuracy. Urdu handwritten OCR is very beneficial, but we cannot attain its benefits until we overcome problems in segmentation.

In the field of pattern recognition and image processing, English text OCR systems attain good accuracy. In the OCR system, segmentation is the most error generating

Table 1: Different cursive scripts.

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.
English cursive script (printed)
The quick brown fox jumps over the lazy dog.
English cursive script (handwritten)
یہ گھر کس کا ہے
Arabic Naskh script
یہ گھر کس کا ہے
Urdu Nastalik script

part in cursive scripts like Urdu and Arabic. That is why segmentation-free approaches are used to handle this problem to some extent. But in large datasets, this approach also failed to give satisfactory accuracy. As an alternative, we have segmentation-based approaches.

Besides the aforementioned motivations, the aim of this study is also to enhance the segmentation accuracy and progress in the state-of-the-art methodologies. The major objective is to solve segmentation accuracy for Urdu OCR, starting from text line segmentation up to ligature segmentation. In cursive scripts, the skewness of a text image is also affecting the segmentation stage. Segmentation of overlapped and skewed text lines is a hurdle in attaining high segmentation accuracy [3]. So, in this research work, skew detection and baseline detection are also taken to uplift the segmentation method. Finally, the ligature segmentation approach is also discussed to fully cover the segmentation stage and achieve high recognition accuracy through the proposed method. In cursive scripts, beside character segmentation, line segmentation is also the error generating part because of the following problems.

Overlapped text lines: in cursive scripts, words of adjacent lines overlap each other. This problem occurs in both handwritten and printed text. Some words are connected with words of adjacent lines and partly segmented during the segmentation process. Due to this, there is a high probability of loss of information.

Unequal line height: this issue occurs in only handwritten text. As there is no standard writing throughout the document, the height of words varies in line, there is no uniform position of words, and characters spread in both vertical and horizontal directions; because of this, a line may not remain straight.

Inconsistent space between lines: in Urdu's handwritten text, there is no standard baseline on which text is written. The space between text lines varies throughout the document. In such cases, baseline is not horizontally straight; it may be in a curve or oscillatory shape. For example, in some cases, at the beginning, start space between lines is minor and at the end of adjacent lines, there is significant space between them which causes skewness problems in segmenting text lines.

Dot/diacritics overlapping: in Urdu, dot/diacritics are mostly spread in white spaces between lines. In these cases, always there is a chance that a dot/diacritic of a line may segment with the adjacent line.

This research contributed to the following directions:

(i) Firstly, a printed and handwritten Urdu text dataset is generated along with ground truth (by removing the skewness of the next page). Afterward, the page is converted to lines and ligatures.

(ii) Secondly, the proposed scheme is evaluated on a manually generated Urdu printed and handwritten dataset, which consists of 80 Urdu handwritten pages (687 lines) and 48 printed pages (495 lines).

(iii) Thirdly, the presented framework addressed existing problems in text line segmentation which are variable text size, the inconsistency of gap between lines, and skewness of text lines.

(iv) In the end, the proposed algorithm is tested separately on handwritten and printed text, showing 96.7% and 98.3% line accuracy, respectively.

Results validated that the proposed line segmentation algorithm correctly extracts the lines when tested on Arabic text.

Till now, we have discussed the OCR system, different types of text, and their problem. Section 2 consists of the related work of the study. Section 3 contains the proposed methods of the study. Then, the results and discussion are addressed in Section 4, and finally, the paper is concluded in Section 5.

2. Related Work

The modes for an OCR system can be categorized according to input text image acquisition that can be offline or online, writing mode that can be printed or handwritten, and finally the font variations leading to font constraints, handled by a recognition system with the support of segmentation process. In this section, we will review existing techniques for text segmentation of printed and handwritten cursive scripts.

Projection profile is the widely used segmentation approach for line segmentation. This method automatically identifies line regions from an image. In this technique, information is used for text line segmentation, while this technique does not give good results for an image having inconsistent line spacing and skewness also affects the accuracy. The skewed image develops small peaks which make the projection profile confused about proper text lines; this problem is solved by using local or piece-wise projection profiles [4]. A modified projection profile is used as an adaptive technique in which partial line fractions are used to develop projection [5]. In [6], projection profile is used to segment ligatures during the line segmentation stage while Y-histogram projection is used for line segmentation. The projection profile based approach is also employed in [7] for baseline detection and division of connected components and then words are segmented into ligatures through vertical

projection. Our approach also used the vertical projection technique for ligature segmentation from lines.

Muna Ayesh et al. [8] presented an algorithm to segment the Arabic text lines. In Arabic, line segmentation problem is mainly due to the placement of diacritics. As diacritics do not follow any baseline, they merge with neighboring lines. The authors presented a line and diacritics segmentation algorithm tested on 43000 lines, giving 99.5% accuracy. The projection profile approach is used with a profile amplitude filter in [9] for Arabic text line, words, and character segmentation. This algorithm is for printed documents irrespective of word size and fonts. Line segmentation algorithms work in two steps. Firstly, rough segmentation is done using a horizontal profile method; after that, the proposed technique having multiple rules is applied to rough segmentation to get final segmentation results.

In [2], the authors presented a segmentation approach using chain code. In this approach, the start point is detected after thinning. Through Freeman chain code, segmented points are marked, while the HMM model is used to recognize each segment as a character or diacritic. Line segmentation (overlapping) and ligature segmentation (primary and secondary) algorithms are presented in [10]. The hybrid top-down approach (projection profile) is used for line segmentation, while the bottom-up approach is for ligature segmentation, in which connected components are extracted and collected for ligature extraction. After that, diacritics and ligatures are categorized as primary and secondary components. This technique gives an accuracy of 99.02% in ligature segmentation and 99.11% in line segmentation.

Moysset et al. [11] used one of the recent models of recurrent neural network for line segmentation. A six-layer deep neural network is used, in which four LSTM layers and the remaining two are a convolutional layer. This approach addresses the location of the text line. Novel line segmentation technique is used in which information energy of pixels is used to segment text lines. The characters are also segmented by using an artificial neural network. The method gives 95% accuracy for line segmentation and 94% for character segmentation.

To segment palm leaf manuscripts of Dai, Ge Peng et al. [12] used algorithm based on HMM, to evaluate all segmentation paths. In another study, Quang Nhat and Lee [13] proposed multilingual text line segmentation approach, in which trained fully conventional network (FCN) is used to figure out text lines pattern. Through FCN, line map is extracted through which initial segmentation is done and after that line, adjacency graph is used to handle overlapping words between lines. This gives 98.6% accuracy on ICDAR-2013.

In water flow approaches, angles of a document (left-right and top-bottom) are used for hypothetical water flow. This algorithm works on this hypothetically assumed situation. For line extraction, strip of un-wetted areas is used. It is assumed that spaces between lines are to be filled and form wet areas after labeling images divided into two parts, wetted and non-wetted areas, where the wetted areas contain spaces and the remaining un-wetted areas contain text lines. Darko Brodi'c [14] modified a linear water flow algorithm, by changing its linear function by power function. Waterflow deals with linearly straight lines; in this modified algorithm, bounding boxes are added to handle angular dimensions of the text.

The smearing approach is considered as one of the earliest approaches used for line segmentation. The smearing approaches refer to the concept of smearing all consecutive dark pixels along the horizontal direction between lines. Then, white spaces within dark pixels are filled with black pixels. Through this, black pixels cover a large area along with the text. This black pixel growing area enclosed a separated text line [15]. In recent times, the RLSA algorithm has been introduced which is based on smearing techniques. Novel painting approach is presented to smear the foreground portion of the image; this way, foreground is separated from background pixels this method is used for text line segmentation.

Probabilistic algorithms are used in stochastic approaches. This method accomplishes the nonlinear path in between overlapping text lines. Then, HMM is used to extract these text lines and the image is divided into small units. In the case of touching components, a high probability path crosses the touching component with minimum dark pixels, while this method drops accuracy in text having a large number of the black pixels contact point of text [16, 17].

Kumar and Choudhary [18] proposed an algorithm for English cursive script. This algorithm vertically segments the ligatures of connected words. First, get the single-pixel stroke width of the scanned image by skew angle correction and thinning. The algorithm segments characters on the base of their geometrical shape. The proposed method is tested on the local dataset. In [19], researchers compared different segmentation algorithms. Experiments are carried out on the CCC benchmark dataset. The horizontal and vertical projection method gives 95.65% segmentation accuracy while Hough transform technique shows 98.9% accurate line segmentation.

In [20], Naz et al. used implicit segmentation for Urdu text segmentation. The horizontal projection profile is integrated into the segment page into lines. Different features of the text, like zone based statistical measures and chain codes, etc., are computed and then neural network is trained for recognition. The method is evaluated on the UPTI dataset. Line and ligature segmentation algorithms are presented in [21].

In [22], the author presents a set of rules that are derived heuristically to search character boundaries of the cursive script that is validated by using an ensemble of neural confidence. Rehman [23] introduced a new concept of core-zone for segmenting such difficult slanted handwritten words. Also in [24], Qaroush et al. used the baseline detection method for the segmentation of characters and CNN for recognition of characters.

Mullick et al. [25] presented a novel approach to segment lines from handwritten Bangla document image. In this approach, first blur the white spaces between words (so that after blurring the white spaces in between words, only the remaining white spaces are in between lines). This way, the most prominent pixels that remained are points of separation.

In [26], multilingual novel baseline approach is used for handwritten text line segmentation, in which significant contours are extracted for making the curve. Orientation invariant features of this curve are used to determine whether the extracted region is a baseline of the text line or not. SVM is trained using the orientation invariant features of the curves and then trained SVM is used to figure outlines from the text. This approach gives 89.6% baseline accuracy.

In [27], Surinta et al. proposed an algorithm to segment lines from historical documents. The novelty of this approach is its artificial agent which handles the abnormality of historical text. In this approach, both ends of each line are detected using the smoothed horizontal ink density histogram. The proposed algorithm uses different cost functions to keep a distance from ink pixels and computes the shortest distance between them to detect lines. This method gives 99.9% line segmentation accuracy on the Saint Gall dataset.

3. Proposed Methods

For Urdu cursive script text, in recent decades a lot of research has been carried out for Urdu printed (Nastaliq) script. But in the case of Urdu handwritten text, a lot of efforts are needed to develop an algorithm that leads to an ideal OCR system. This paper proposes a method for skew correction and line segmentation of printed and handwritten documents. The presented methodology works in the following sequential steps:

(1) Preprocessing

(2) Skew correction

(3) Text line segmentation

(4) Ligature\word segmentation

The focus of this paper is mainly on preprocessing and segmentation of Arabic/Urdu scripted OCR. Before segmenting text into its basic shape, the system preprocesses the noisy and skewed images. This paper reflects a segmentation technique along with the projection profile technique. The paper reflects an enhanced pixel counting based robust algorithm which is independent of any script-specific knowledge. Moreover, a modified header and baseline detection technique [28] is used for the line segmentation algorithm.

Pixel based method mainly lacks in detecting noisy and skew text images. We overcome both the drawbacks of the header and baseline method in this proposed method by employing the adaptive threshold method for noise detection and a novel algorithm for skew detection. The graphical representation of the method is shown in Figure 1.

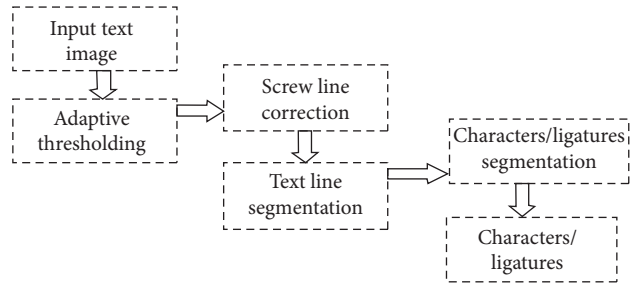


Figure 1: Proposed methodology.

3.1. Preprocessing. Preprocessing is a very essential phase for better segmentation results. For preprocessing, adaptive thresholding approach is used, in which noisy image is given as input. Adaptive thresholding is applied to the input image which is based on the intensity of the image. Grayscale image contains image pixel intensity. For the RGB image, the image is converted into YCbCr color space, where Y contains (black and white pixels) intensity, 8-bit depth (grayscale), and 24-bit depth (RGB) images can be used in this method. After converting the image into grayscale, adaptive thresholding is applied.(Algorithm 1).

An input image is given as grayscale or RGB format, for processing convert it into a binary image (white and black image). Text image has mainly two colors: (1) text color and (2) background color. By converting into binary images, the contrast of the image increases and is easy for global thresholding using Otsu's method [29].

An adaptive local thresholding algorithm separates the background from the foreground of the image with non-uniform brightness. Apply adaptive (mean/median) filter to highlight image features and then apply the Otsu threshold to generate a binary image. It has been found that Otsu's thresholding has produced good results on text data. For better-preprocessing results, the binarization technique is used, in which pixels having more than specified intensity will be converted into white pixels; otherwise, they will be converted into black pixels; therefore, the image is converted into black and white pixels. The equation is as follows:

$$\sigma_T^2 = \sum_{i=0}^{L-1} (i - \mu T)^2 Pi, \tag{1}$$

where P is probability and I is the intensity of L number of bins. After that, a single-pixel thick edge boundary (in the whole image) is created for each character from a canny edge detector (auto-thresholding). As per our knowledge, the canny edge detector is used for preprocessing, as it removes dot noise with a low-pass filter. At that point, apply a Sobel filter and then use nonmaximal suppression to choose the finest pixel for edges when there are different local possibilities.

3.2. Skew Detection. The proposed line segmentation is dependent on Algorithm 2, skew correction algorithm. A line segmentation algorithm requires skewing fewer images for good performance. The proposed skew correction algorithm works on the pixels intensity information of the image. The main idea of this algorithm is to extract the area between text lines and fit it to a straight line.

Find the white spaces (in the case if the background is white) that define the zones between the lines with letters on them. Rather than finding the text, we will find the white spaces between the lines. Then, we will go from the center of the first spaceline to the center of the next one, till the end of the page. In this method, skew correction is considered at the page level. At last, find the tilted angle of lines and rotate the image around its center point.

```
Input: Grayscale image or RGB.
Output: Clean image.
//Begin
Step 1. //If the image is RGB, convert it into YCbCr color space.
        YCbCr ← RGB
Step 2. //Eliminate CBCR. And the image becomes grayscale.
Step 3. //Apply adaptive thresholding.
Step 4. Apply the global image threshold using Otsu's method.
Step 5. Adaptive (mean\median) filter to highlight image features.
Step 6. Then, apply the Otsu threshold to segment and generate a binary image.
//End
```

Algorithm 1: Preprocessing.

```
Input: Noiseless skewed image.
Output: Skew-less image.
//Begin
Step 1. Convert input text image into a grayscale image.
Step 2. Scan the document and extract ROIs. // ROI = area between text lines.
Step 3. Find all pixels between text lines (other than text pixels).
Step 4. Join these pixels to fit on a line (having the same slope as text).
Step 5. Go for the center of each fitted line.
Step 6. Find the angle between the fitted line and the horizontal line of the page.
Step 7. Rotate area.
//End
```

Algorithm 2: Skew correction algorithm.

3.3. Line Segmentation. Noiseless and skew-less images are fed to a system for further processing. The input binarized image is inverted for getting an image I'HXW with black text on a white background. Then, pixel strength (P) is calculated for the black text in a document.

$$P = \sum_{y=1}^{W} l'(x, y)[1, H]. \tag{2}$$

This pixel strength (P) determines the threshold of text image; it varies from image to image. Rows having dark pixels greater than the standard deviation value of the document are extracted. The text line is the combination of consecutive rows, which are extracted as pixels (P) lying between header and footer.

$$\begin{aligned} \text{For header}: \ & Px_i\text{:}\ x_{i-1} < x_i > x_{i+1}, \\ \text{For baseline}: \ & Px_I'\text{:}\ x_{i-1}' > x_i' < x_{i+1}'. \end{aligned} \tag{3}$$

The above equations represent the criteria for assigning header and baseline to a particular text line. Two Lines in a text are separated when black pixels in a row are less than the adaptive threshold in rows of the entire document, shown by "white spaces" in Figure 2. This white pixel acts as a border between two text lines.

As this technique focuses on the start and end of a text line, this is why it is referred to as the "header and baseline detection" technique. The text line is extracted by using two parameters of the text line, the starting point of the line and the baseline (last row of the line). Header and baseline are determined by the number of black and white pixels in rows of text images. Exceeding threshold consecutive black pixel rows are labeled as text line whereas repeated white pixel rows are considered as the separation area between two lines (shown in Figure 2). Pictorial representation of the line segmentation algorithm is shown in Figure 3.

In the proposed method, the adaptive threshold of the page is set by calculating the standard deviation of text pixels (black for white background), which determines the diversity of text pixels on a page. It works on the idea that higher value of standard deviation means greater distance between lines. This adaptive threshold determines the minimum number of consecutive text rows in a line. As in the case of Arabic/Urdu scripted text, diacritics of text appear above the line and contain fewer pixels. In some cases, those rows which have black pixels less than the minimum threshold value affect recognizing text pixels by not detecting dot/diacritics in a line. Algorithm 3 addresses the proposed methodology with all abbreviations in Table 2.

This technique purely depends on the counting pixels approach. The main idea of this technique is that lines of page contain larger number of black pixels than the spaces between lines. Height threshold is used for the height of segmented line. Threshold is set according to the page in consideration. If the spaces between the lines are constant, then the algorithm is tuned to fix the threshold value according to spaces between lines, while if the line distance

between the adjacent lines is changing constantly, then the algorithm must be capable of adaptive thresholding.

For ligature segmentation, the projection profile method is used for text line segmentation and lines are transformed into words through the vertical profile method. In the proposed method, Urdu handwrote, or printed text page, is inserted as an input. Firstly, the page is segmented into several text lines and further fed into a word/ligature segmentation algorithm which divided lines into the smallest possible ligatures. As the algorithm segments words in sequential order, ligatures are automatically arranged in sequence.

3.4. Dataset Generation. Urdu's handwritten dataset is composed of a collaboration of 24 writers. Each participant writes a different number of words, lines, and pages having 687 lines which combine to form 80 pages. The number of words in a line and the number of lines on a page vary throughout the dataset. Sample data is versatile with almost all types of writing problems that make the dataset complex and reduce the accuracy of the algorithm; for example, each writer has their own writing style having different difficulty levels of recognition, and each participant is having a different text size. These handwritten text images were captured using a high-resolution digital camera and stored in jpg format after scanning. Samples of handwritten dataset are shown in Figure 4. Details of the dataset are presented in Table 3 and available online on GitHub (https://github.com/saud00/Urdu-text-dataset).

To evaluate the algorithm on Urdu OCR, we present a diverse and comprehensive Urdu handwritten dataset. The dataset is written with a blue and black ball pen and also pointer pen is used so that the dataset contains all types of writing intensity. Then, this dataset is scanned and converted to a binary image (white and black). Eighteen teachers and ten students (both male and female) contributed to the dataset. They were told to write paragraphs (without any restriction of content). They wrote different numbers of lines in different writing with different pens. Samples are shown in Figure 4. The ground truth is also manually created for a handwritten dataset. They were not trained so that this database reflects the true essence of challenging real databases.

Urdu Nastaliq printed dataset is also generated and the data is collected from three different sources.

(i) 27 pages are collected from online books (by taking a screenshot and then cropping through Paint)

(ii) 10 pages of newspapers (scanned from a camera)

(iii) 11 pages of digests (scanned from a camera)

To maintain diversity in the dataset, data is collected from three different sources. Firstly, 27 pages of the online book (Shahab Nama) are collected (Figure 4(c)), by taking screenshots of twenty-seven pages and then cropping them through Paint. Twenty-seven pages contain a total of 275 lines. Secondly, ten paragraphs of newspapers are collected through a digital camera and then scanned for further use (Figure 4(a)). Ten paragraphs of newspapers contain 86 lines. Finally, randomly 10 pages are scanned from a digest (Figure 5(b)), which contains 131 lines.

4. Results Analysis and Discussion

The above-generated dataset is tested for evaluating the proposed method. It is ensured in the dataset that it contains images of possibly all renowned formats: jpg, png, and grayscale, etc. For compiling results, MATLAB 2017a is used. The code of this project is available on GitHub (https://github.com/saud00/Line_and_Word_Segmentation_URDU) along with dataset and output images.

The accuracy of the proposed line segmentation is dependent on the accuracy of skew correction. Line segmentation algorithm requires skew-less image for good performance. The proposed skew algorithm works on the pixels intensity information of image.

Results are generated at the end of segmentation stage, later used for recognition rate. Firstly, preprocessing technique is applied to remove noise so as to handle false line detection; after that, if the image is tilted in either way, it is removed by applying skew detection algorithm. De-skewed image is fed into line segmentation algorithm for segmenting lines.

4.1. Results Analysis. The accuracy of the proposed framework is tested with labeled ground truth text line images. The dataset contains 495 printed and 681 handwritten line images snapshots along with manually created ground truth. As part of a framework, the dataset is developed and labeled manually. The number of characters in labeled ligature images is counted and compared with several recognized characters to find recognition accuracy using

$$\text{accuracy} = \frac{\sum_{i=1}^{n} L_i - \left(\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} R_i\right)}{\sum_{i=1}^{n} L_i} \times 100, \tag{4}$$

where R represents correctly recognized lines and L is input labeled line image. The segmented lines from the input page are compared with ground truth images to find line recognition accuracy. For evaluating results, we use precision, recall, and F-measure matrices which are defined as follows:

$$\begin{aligned} \text{recall} &= \frac{\text{correctly segmented lines by the algorithm}}{\text{actual total number of lines}}, \\ \text{precision} &= \frac{\text{correctly segmented lines by algorithm}}{\text{total number of lines segmented by the algorithm}}, \\ \text{F} - \text{measure} &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \end{aligned} \tag{5}$$

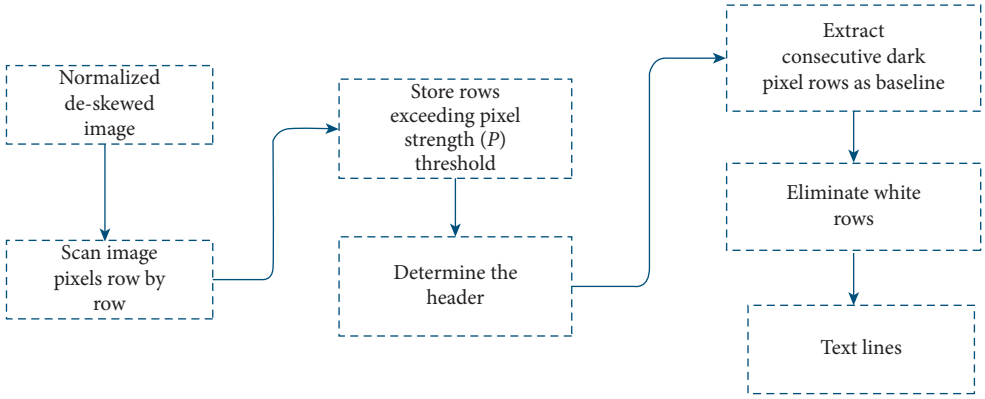


Figure 2: Header and baseline in the text image.

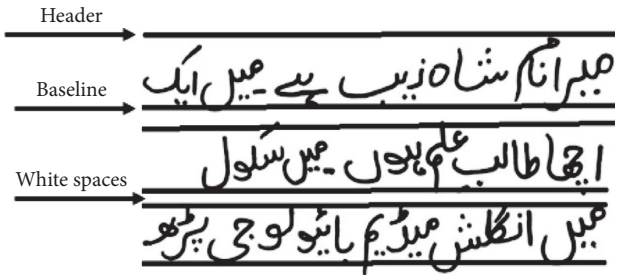


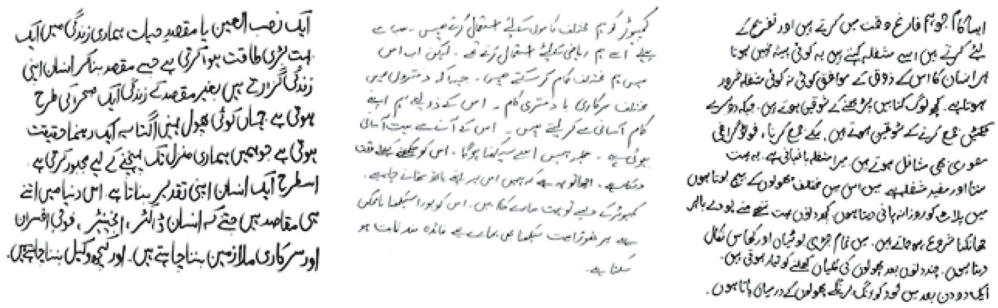
Figure 3: Flow chart of line segmentation methodology.

```
Input: Normalized de-skewed image
Output: Segmented lines.
//Begin.
Step 1. //Preprocessing: image binarization (using adaptive threshold).
Step 2. //De-skew the image (if needed).
Step 3. //Scan Image row by row.
Identify the intensity for each pixel (0 or 1).
Step 4. //Calculate the standard deviation of the image (use as minimum black pixels in a text row).
Step 5. If(Black_Pixels > Std)
Black_Row = row
Step 6. Else
Space_Row = row
Step 7. For (Start from 1st_Black_Row: till Last_space_Row)
Step 8. If (Height_Row > Min_ Height_Row)
//Consider these consecutive text rows as text line until any white_row occurs.
Step 9. If (Space_Row occur)
Step 10. If (Space_Row > Min_Height_Row)
Break text_line and go to Step 11.
Step 11. else
Go to Step 7
Step 12. else
Search for next black_text_row
Step 13. Else
Go to Step 10
//End
```

Algorithm 3: Text line segmentation algorithm.

Table 2: Abbreviations used in Algorithm 3.

Abbreviations	Description
Std	The standard deviation of the image
Black_Pixels	Number of black pixels in a row
Black_Row	Row having black text pixels greater than the threshold
Space_Row	Space between text lines
Space_Row	Space between text lines
1st_Black_Row	Last space_row of page
Height_Row	Number of consecutive black_text_rows
Min_ Height_Row	Minimum threshold of consecutive black rows



(a) (b) (c)

Figure 4: Handwritten dataset samples.

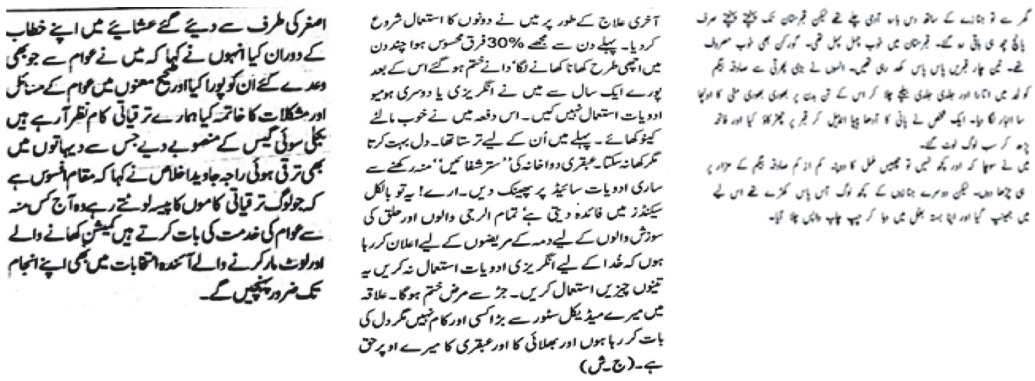
Table 3: Details of handwritten dataset.

Dataset details	Statistics
Total number of handwritten pages	80
Total number of writers	28
Number of pages written by a writer	3 (average)
Number of text lines per page	9 text lines (average)
Number of skewed pages	12
Total number of lines	687
Number of skewed lines	97
Approx. number of words per page	103
Approx. number of lines by a writer	27
Approx. number of lines by a writer	306
Approx. number of words per line	12
Total number of words	8,208

Precision is the fraction of relevant segmented lines among the retrieved lines, while recall is the fraction of relevant lines that have been retrieved over the total amount of relevant lines. The skew detection algorithm detects the skew of the text image by using the proposed skew correction algorithm. Figure 6(b) shows the output of the skew correction algorithm. The skew correction algorithm is evaluated based on true line segmentation. A total of 13 skewed images are tested on the algorithm, from which 11 images are truly de-skewed.

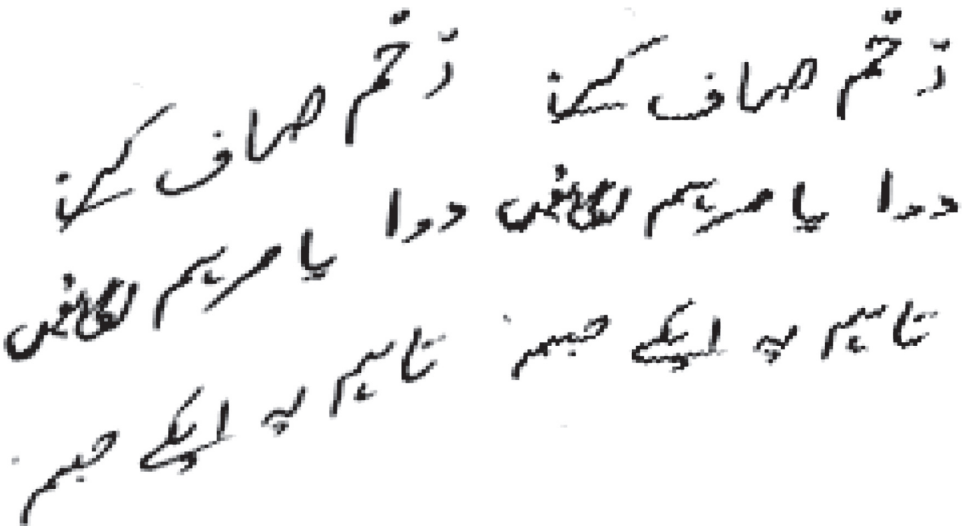
4.1.1. Handwritten Documents. The framework is tested on both handwritten and printed Urdu text. Dataset of 80 pages (687 lines) is evaluated on the text line segmentation algorithm. Line segmentation algorithm correctly segments 687 lines, from which 8 lines are under-segmented and incorrect 18 lines are formed while over-segmentation issue affects 4 text lines. Line segmentation algorithm gives 96.7% line accuracy by correctly spotting 665 handwritten text lines. The result is shown in Table 4.

4.1.2. Printed Documents. The algorithm is tested on the Urdu Nastaliq printed dataset which is collected from 3 different mediums online books, newspapers, and digest. From a total of 48 pages, 27 pages are taken from an online book, 10 pages from a newspaper, and 11 scanned pages collected from digest shown in Table 5. For printed data, a total of 48 pages (495 lines) are tested. The result of the printed dataset is given in Table 5 which shows 98.38% accuracy by detecting 487 lines from a total of 495 lines.



(a) (b) (c)

Figure 5: Printed dataset samples. (a) Newspaper sample. (b) Digest sample. (c) Online page sample.



(a) (b)

Figure 6: (a) Skewed image. (b) Output image.

Table 4: Evaluation matrices.

Text type	No. of lines	Detected lines	Correctly detected lines	Precision	Recall	F-measure
Handwritten	687	681	665	97.6	96.79	97.3
Printed	495	491	487	99.18	98.38	98.74

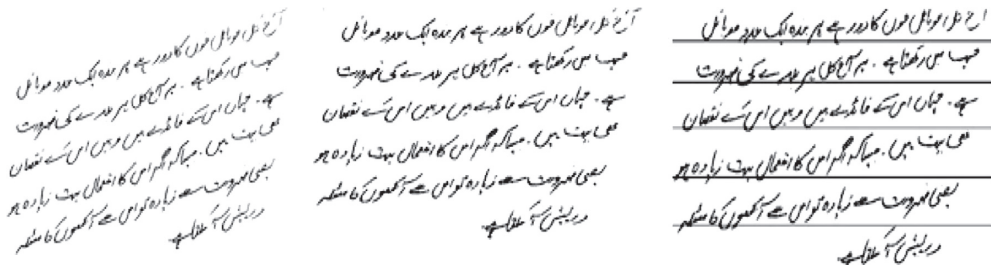
Table 5: Details of Urdu printed dataset and its corresponding results.

Source	Pages	Lines	Correctly detected lines	Accuracy
Online book	27	275	272	98.9
Newspaper	10	86	84	97.7
Digest	11	131	131	100
Overall	48	495	487	98.38

Among 3 types of data type, the algorithm detects lines of digest with ease by showing 100% accuracy. We fed 11 pages of digest having 131 lines; these all are correctly recognized. From 275 lines of an online book, only 3 misled the proficiency, showing a 98.9% recognition rate.

From the overall 48 printed pages, the newspaper shares 11 pages with 86 lines. The algorithm detects 84 lines. But, there are a lot of differences between paragraphs of the newspaper (Figure 5(a)) and paragraphs of digest (Figure 5(b)).

The number of pages in the newspaper and digest is nearly the same but having a large difference in the number of lines. The algorithm correctly detects all the 131 lines in digest and attains 100% accuracy. Figure 7 shows the whole



(a) (b) (c)

Figure 7: (a) Original image, (b) de-skewed image, (c) output of proposed line segmentation algorithm.

(a)

(b)

Figure 8: (a) Original image, (b) ligature segmented image.

Table 6: Comparison of the proposed technique with the existing methods.

Source	Number of pages	Total lines	Detected lines	Text type	Accuracy (%)
Younes et al. [30]	90	1000	940	Handwritten	94
Din et al. [31]	30	310	306	Printed	98.7
Ahmad et al. [21]	47	607	602	Printed	99.17
Proposed method	80	687	674	Handwritten	96.7
Proposed method	48	495	491	Printed	98.3

sequential process (left to right), in which the algorithm first removes the skewness of an image (Figure 7(b)) and then segments it into lines. After segmentation of page into lines, the projection profile method is used for text line segmentation and lines are transformed into words through the vertical profile method. The boundary of each connected dark region in the profile is extracted as a separation region. Ligatures/subwords are segmented as an output of the algorithm, which is shown in Figure 8(b).

4.2. Comparison with Previous Work. Despite prevalent contributions in Urdu OCR, there is no such versatile dataset available for text line segmentation which covers the handwritten and printed text.

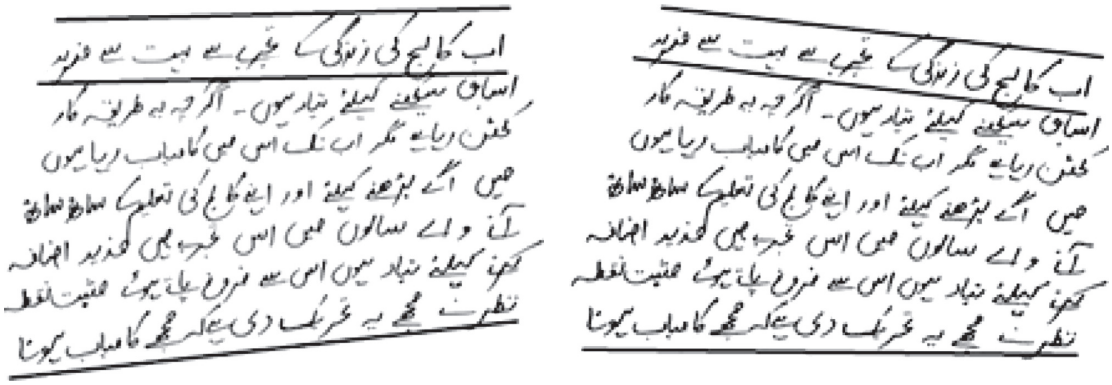
Table 6 shows the previous relevant works with their accuracies. Most works in the OCR field have not used any available dataset; as mentioned in [21, 32, 33], they used their datasets. Therefore, the accuracy of the algorithm depends on their dataset. Mainly in Urdu text, dot/diacritics allocation and skew detection are two issues. The algorithm presented in [21] handles the dot/diacritics allocation problem but does not work well for skew documents. The proposed algorithm overcomes this issue, by using Algorithm 2.

4.3. Results Discussion. Skew detection algorithm easily corrects the skewness of the image if the angle between lines is the same or has a little variation. The result of segmentation suffers when the angle between lines is changing throughout the image as shown in Figure 9(a). As the algorithm is not rotating each line separately, the whole image is rotated as shown in Figure 9(b).

Mainly, two types of issues have occurred during line segmentation, which are over-segmentation and missed/under-segmentation. Partially segmented lines are considered as false detection. Accuracy of segmentation is calculated as either "detected lines" or "not detected lines," because when one line is wrongly segmented into two lines (Figure 10), then it is not recognized in the latter stages.

Inter-line skewness causes under-segmentation. If the paragraph has multi-skewed lines, the algorithm bypasses these lines because the algorithm is unable to detect multi-skewed lines and segments both lines as one line. This issue causes false segmentation of lines as shown in Figure 11.

Over-segmentation is mainly due to dot and diacritics presence. In the proposed algorithm, when the number of dark pixels exceeds the threshold limit, then it is considered as a text line. As given in Figure 10, one line is over-segmented; these lines are false detected lines and considered as incorrect



(a) (b)

Figure 9: (a) Original multi-skewed text image, (b) image after skew correction.

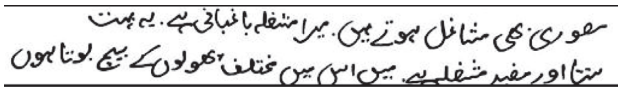


Figure 10: over-segmentation.

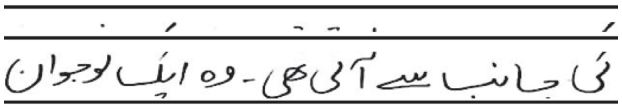


Figure 11: Under-segmentation.

segmented lines. In this way, one line is segmented into two or more lines. This wrong segmentation of the line is known as over-segmentation.

In this research, we focus on handling handwritten text (with and without skewness). This technique has a limitation in segmenting process over dot and diacritics properly.

5. Conclusion

Recently, many script-dependent algorithms for line and ligature have been proposed. But in this research, we put efforts to step forward to propose an efficient algorithm that deals with both printed and handwritten text. The proposed algorithm works well on both printed and handwritten Urdu documents. In the handwritten text, lines are not straight and have a variable size. Especially for handwritten text, the algorithm deals with skewed pages and variable text line size. In the proposed method, the page is preprocessed using an adaptive thresholding technique. Then, the image is rotated if it is skewed and the lines of the de-skewed image are segmented. The proposed line segmentation algorithm is based on counting pixel density (black and white) in a row. The header line and baseline of the text line are determined and segmented. After line segmentation, projection profile technique is used to segment ligatures from the segmented line. The proposed line segmentation algorithm shows promising results on handwritten and printed Urdu text. We make this approach more flexible for handwritten text so that dot/diacritics will remain in the concerned line and not be part of adjacent lines. In this paper, we mainly deal with Urdu text. In the future, we will expand this work as a general technique so that this approach will be applicable to all OCR systems.

Data Availability

All the data used to support the findings of this study are included within the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to acknowledge the support by King Saud University, Saudi Arabia, through Researchers Supporting Project number RSP-2020/184.

References

[1] S. A. Malik, M. Muazzam, A. Farhan et al., "An efficient segmentation technique for Urdu optical character recognizer (OCR)," in *Future of Information and Communication Conference*Springer, Berlin, Germany, 2019.

[2] A. F. Ganai and F. R. Lone, "Character segmentation for Nastaleeq Urdu OCR: a review," in *Proceedings of the International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, IEEE, Chennai, TN, India, March 2016.

[3] K. Keisham and S. Dixit, "Recognition of handwritten English text U minimisation," in *Information Systems Design and Intelligent Applications*, pp. 607–614, Springer, Berlin, Germany, 2016.

[4] M. Arivazhagan, H. Srinivasan, and S. Srihari, "A Statistical approach to handwritten line segmentation. document recognition and retrieval XIV," in *Proceedings of the SPIE*, pp. 6500T-6501T, San Jose, CA, USA, January 2007.

[5] I. Bar-Yosef et al., "Line segmentation for degraded handwritten historical documents," in *Proceedings of the 10th International Conference on Document Analysis and Recognition ICDAR'09*, IEEE, Barcelona, Spain, July 2009.

[6] C. I. Patel, R. Patel, and P. Patel, "Handwritten character recognition using neural network," *International Journal of Scientific & Engineering Research*, vol. 2, no. 5, pp. 1–6, 2011.

[7] Z. A. Shah, "Ligature based optical character recognition of Urdu-Nastaleeq font," in *Proceedings of the International Multi Topic Conference Abstracts. INMIC 2002*, IEEE, Karachi, Pakistan, February 2002.

[8] M. Ayesh, K. Mohammad, A. Qaroush, S. Agaian, and M. Washha, "A robust line segmentation algorithm for Arabic printed text with diacritics," *Electronic Imaging*, vol. 2017, no. 13, pp. 42–47, 2017.

[9] M. A. Mousa, M. S. Sayed, and M. I. Abdalla, "Arabic character segmentation using projection based approach with profile's amplitude filter," 2017, http://arxiv.org/abs/1707.00800.

[10] G. S. Lehal, "Ligature segmentation for Urdu OCR," in *Proceedings of the 12th International Conference on.Document Analysis and Recognition (ICDAR), 2013*, IEEE, Washington, DC, USA, August 2013.

[11] B. Moysset, K. Christopher, W. Christian et al., "Paragraph text segmentation into lines with recurrent neural networks," in *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), 2015*, IEEE, Tunis, Tunisia, August 2015.

[12] G. Peng, P. Yu, H. Li et al., "Text line segmentation using Viterbi algorithm for the palm leaf manuscripts of Dai," in *Proceedings of the 2016 International Conference on Audio, Language and Image Processing (ICALIP)*, IEEE, Shanghai, China, February 2016.

[13] Q. N. Vo and G. Lee, "Dense prediction for text line segmentation in handwritten document images," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, IEEE, Phoenix, AZ, USA, August 2016.

[14] D. Brodić, "Text line segmentation with water flow algorithm based on power function," *Journal of Electrical Engineering*, vol. 66, no. 3, pp. 132–141, 2015.

[15] S. Marinai and P. Nesi, "Projection based segmentation of musical sheets," in *Proceedings of the Fifth International Conference on Document Analysis and Recognition ICDAR'99*, IEEE, Bangalore, India, October 1999.

[16] D. Brodić and Z. Milivojević, "A new approach to water flow algorithm for text line segmentation," *Journal of Universal Computer Science*, vol. 17, no. 1, pp. 30–47, 2011.

[17] R. Student, "Off-line handwritten Kannada text recognition using support vector machine using zernike moments," *IJCSNS*, vol. 11, no. 7, p. 128, 2011.

[18] A. Choudhary and V. Kumar, "A robust technique for handwritten words segmentation into individual characters," in *Speech and Language Processing for Human-Machine Communications*, pp. 99–106, Springer, Berlin, Germany, 2018.

[19] P. Dhande and R. Kharat, "Segmentation and feature extraction for cursive English handwriting recognition," *IJETT*, vol. 1, no. 2, 2017.

[20] S. Naz, R. Imran, S. Imran et al., "An Ocr system for printed Nasta'liq script: a segmentation based approach," in *Proceedings of the IEEE 17th International Multi-Topic Conference (INMIC)*, IEEE, Karachi, Pakistan, December 2014.

[21] I. Ahmad, X. Wang, R. Li, M. Ahmed, and R. Ullah, "Line and ligature segmentation of Urdu Nastaleeq text," *IEEE Access*, vol. 5, pp. 10924–10940, 2017.

[22] A. Rehman, "An ensemble of neural networks for nonlinear segmentation of overlapped cursive script," *International Journal of Computational Vision and Robotics*, vol. 10, no. 4, pp. 275–288, 2020.

[23] A. Rehman, "Cursive overlapped character segmentation: an enhanced approach," 2019, http://arxiv.org/abs/1904.00792.

[24] A. Qaroush, A. Abdalkarim, M. Mohammad, and Z. Malik, "Segmentation-based, omnifont printed Arabic character recognition without font identification," *Journal of King Saud University-Computer and Information Sciences*, 2020.

[25] K. Mullick, S. Banerjee, and U. Bhattacharya, "An efficient line segmentation approach for handwritten Bangla document image," in *Proceedings of the 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, IEEE, Kolkata, India, March 2015.

[26] D. Chakraborty and U. Pal, "Baseline detection of multilingual unconstrained handwritten text lines," *Pattern Recognition Letters*, vol. 74, pp. 74–81, 2016.

[27] O. Surinta, L. Schomaker, M. Wiering et al., "A path planning for line segmentation of handwritten documents," in *Proceedings of the 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, IEEE, Heraklion, Greece, December 2014.

[28] S. Palakollu, R. Dhir, and R. Rani, "A new technique for line segmentation of handwritten Hindi text," *Special Issue of International Journal of Computer Applications*, vol. 5, pp. 0975–8887, 2011.

[29] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[30] M. Younes and Y. Abdellah, "Segmentation of Arabic handwritten text to lines," *Procedia Computer Science*, vol. 73, pp. 115–121, 2015.

[31] I. U. Din, Z. Malik, I. Siddiqi, and S. Khalid, "Line and ligature segmentation in printed Urdu document images," *Journal of Applied Environmental and Biological Sciences*, vol. 6, no. 3, pp. 114–120, 2016.

[32] F. Shafait, D. Keysers, and T. M. Breuel, "Layout analysis of Urdu document images," in *Proceedings of the Multitopic Conference INMIC'06*, IEEE, Islamabad, Pakistan, December 2006.

[33] S. S. Bukhari, F. Shafait, and T. M. Breuel, "High performance layout analysis of Arabic and Urdu document images," in *Proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, Beijing, China, September 2011.

Hindawi
Scientific Programming
Volume 2020, Article ID 6667610, 16 pages
https://doi.org/10.1155/2020/6667610

Research Article

Inferring Ties in Social IoT Using Location-Based Networks and Identification of Hidden Suspicious Ties

Nauman Ali Khan,[1] **Sihai Zhang**,[1] **Wuyang Zhou**,[1] **Ahmad Almogren**,[2] **Ikram Ud Din**,[3] **and Muhammad Asif**[1]

[1]*Key Laboratory of Wireless-Optical Communication, University of Science and Technology China, Hefei 230027, China*
[2]*Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11633, Saudi Arabia*
[3]*Department of Information Technology, The University of Haripur, Haripur 22620, Pakistan*

Correspondence should be addressed to Wuyang Zhou; wyzhou@ustc.edu.cn

Received 21 October 2020; Revised 5 November 2020; Accepted 16 November 2020; Published 1 December 2020

Academic Editor: Habib Ullah Khan

Stochastic Internet of Things (IoT)-based communication behavior of the progressing world is tremendously impacting social networks. The growth of social networks helps to quantify the effect on the Social Internet of Things (SIoT). Multiple existences of two persons at several geographical locations in different time frames hint to predict the social connection. We investigate the extent to which social ties between people can be inferred by critically reviewing the social networks. Our study used Chinese telecommunication-based anonymized caller data records (CDRs) and two openly available location-based social network data sets, Brightkite and Gowalla. Our research identified social ties based on mobile communication data and further exploits communication reasons based on geographical location. This paper presents an inference framework that predicts the missing ties as suspicious social connections using pipe and filter architecture-based inference framework. It highlights the secret relationship of users, which does not exist in real data. The proposed framework consists of two major parts. Firstly, users' cooccurrence based on the mutual location in a specific time frame is computed and inferred as social ties. Results are investigated based upon the cooccurrence count, the gap time threshold values, and mutual friend count values. Secondly, the detail about direct connections is collected and cross-related to the inferred results using Precision and Recall evaluation measures. In the later part of the research, we examine the false-positive results methodically by studying the human cooccurrence patterns to identify hidden relationships using a social activity. The outcomes indicate that the proposed approach achieves comprehensive results that further support the theory of suspicious ties.

1. Introduction

A social network is a web of social ties among individuals. Social ties are the kind of one-to-one communication links among humans or Social Internet of Things [1, 2]. Formation of social ties depends upon many attributes, such as the location of living, personality, age, gender, workplace, activities, and many more [3, 4]. These ties are built based on some needs or relationships. People use many mediums for communication such as calls, chatting on social networking websites, reading and writing comments to person, and reviewing and suggesting some purchasing mobile applications [5]. Investigating human behavior and machine performance, how they react to and participate in social networks remained a center of attention for researchers [2, 6]. Social network analysis is the computer science field that quantifies, evaluates, and analyzes human behavior [7, 8].

A concept of social communication links was proposed by Granovetter [9]. According to their research finding, the communication link between two people is considered as the social tie. Also, each communication link's strength was

further classified as strong or weak depending upon the frequency of communication, number of times, emotional attachment, number of mutual ties, relationship actions, and a combination of these mentioned parameters [5]. Equation (1) quantifies the strength of social ties, which is denoted by weight, such that higher weight tells stronger ties and vice versa [10]. wAB represents the weight of social tie between Node A and Node B, while CA and CB represent the degree of Node A and Node B; cAB is the number of mutual nodes between A and B. Community structures were also found one of the main reasons for social tie strength [11]; it was found that people from the same communities have strong ties as compared to different communities [5]:

$$wAB = \frac{cAB}{CA - 1 + CB - 1 - cAB}. \quad (1)$$

Various models and techniques are developed to infer the social network based on inadequate aspects [7, 12]. One of the specific categories belonging to such inference determines the cooccurrence based on time and location. Despite encountering many measures, there remains a deficiency in acquiring precise and accurate inferences. In our research, we consider several threshold parameters to quantify more precise inferences. We also develop a framework that infers existing social ties and the hidden relationships in a social network.

Initially, in our research, we present the inference of social ties among people by correlating to their physical presence at several sites and their direct connections. We define a social connection if two individuals X and Y cooccur in a s cell within t hr time frame, such that X calls to person R while connected to a $b1$ base station and in the same time frame Y calls a person S from the same $b1$ base station. Furthermore, we counted the number of cooccurrence of X and Y. Firstly, we find social ties depending upon the number of direct calls between two people. To ensure the correct social connection, we state a threshold, such that the count of direct calls is more than the threshold. Secondly, we evaluate the social relationship between two people by counting the number of calls by X and Y in a specific time frame. Figure 1 states an example that explains the procedure of quantification. Each hexagon represents an area of a single base station. X and Y are together 6 times in various base stations, and there is a variation in the gap of calls. In the first part of research, we use the CDR data set provided by the telecommunication company and two openly available location-based social network data sets, i.e., Brightkite and Gowalla [13]. All data sets resemble to the stated example in Figure 1. We counted the number of concurrences based on multiple gap time frame thresholds and mutual friends. Furthermore, we correlate results with the direct calls based on social connection using Precision and Recall evaluation measures.

In the second phase of research, we explore the false-positive results formed by the CDR-based social tie inference model. We state a missing tie as a suspicious tie between two people if they do not have any direct calls but are found together numerous times. Also, they have a certain number of mutual friends. In the literature, missing ties are defined as either nonresponse or absent ties [12]. In an activity, an actor does not give any information about a tie considered as nonresponse [14], while an absent tie means when an actor does not give any indication about the tie detail. A survey was conducted to monitor the social behavior of the boys' and girls' liking pattern. That was limited to binary data, such that one represents a tie while zero represents no tie. Figure 2 shows visual representation (block modeling) of adjacency matrix made according to survey data [14]. In Figure 2, green-filled slots represent the existence of tie, regardless of its strength, while white slots indicate either absent ties or nonresponse ties. In our research, we explore and classify a subset of missing ties as suspect ties. We conduct a social activity and simulation that generates a data set the same as the CDR data supporting this concept. Furthermore, we correlate the CDR-based social tie inference model's false-positive results with the activity and simulation results.

The contributions of this paper are as follows:

(1) We developed an inference model and a classifier that identifies location-based social ties. The inference model is tested on the CDR-based social network, Brightkite, and Gowalla, using Precision and Recall measures.

(2) We identify a class of suspect ties by examining the social tie inference model's false-positive results.

(3) We conducted an activity-based survey and a simulation that demonstrates and evaluates the suspects' ties.

The rest of this article is organized as follows: Section 2 describes the literature review. Section 3 presents the descriptions of cooccurrence count normalization, inference algorithm, and social tie inference. Brief concepts about the hidden relationship and suspicious links are described in Section 4. The proposed framework and an algorithm to infer suspicious relationships are given in Section 5. Section 6 describes the data set description, results, and analysis. Finally, the conclusion of this article is presented in Section 7.

2. Related Work

The physical world social network is represented as a graph, where nodes are treated as people, and edges are represented as the social tie between two people [15]. In the literature, edge weight is represented as the strength of that particular social tie [10, 16]. A social network such as Twitter forms a bidirectional graph, e.g., a fan follows a celebrity but the celebrity hardly ever follows back. Usage of bidirectional graphs investigates influential networks and most inflectional people [17–19]. Recommendation and targeted marketing are some of the essential objectives of exploring social ties [20, 21]. Theme-based model adopts dynamic programming to explore critical factors, for example, playing and dating are the kind of themes [22]. Social ties coupling and predicting the mobility of users were researched by

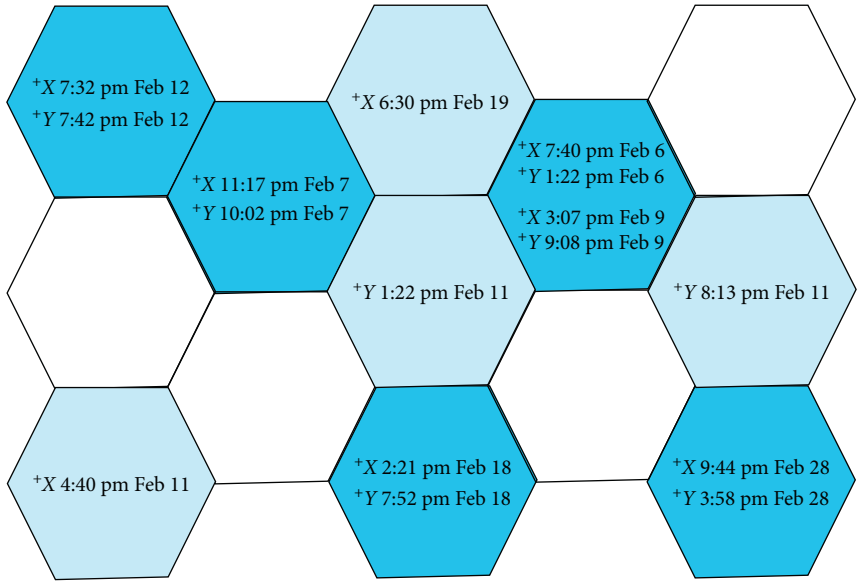


Figure 1: Illustration of social tie example for two individuals that have cooccurrence while connected to various base stations of different times.

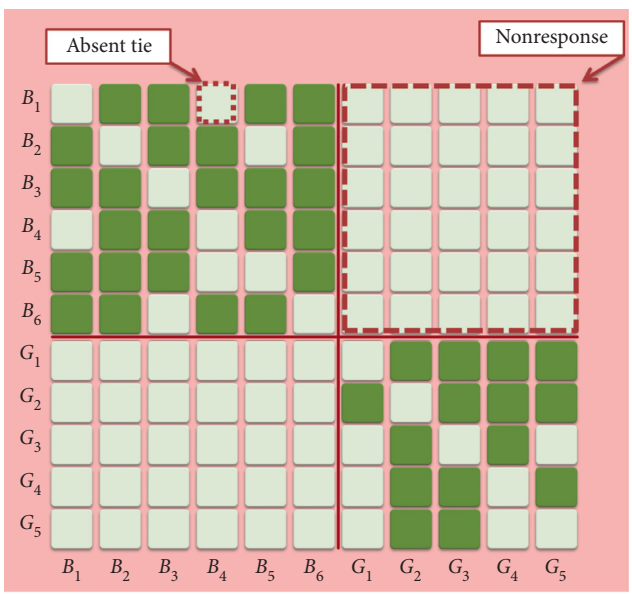


Figure 2: Boys and girls liking social network (absent tie and nonresponse).

seeing the physical and network properties (geosocial properties) [23, 24]. An effective prediction technique was proposed to find the typical patterns of two users by comparing the check-in details [24, 25]. Area significance is measured using a weight-assigning method by incorporating two users' cooccurrence for a specific area. A coffee shop is a more significant area as compared to an ordinary place [26]. The scoring mechanism helps in categorizing and labelling of social ties [10]. Inference about the any social network is incomplete if associated features are neglected. The baseline of any social network is the single social connection between two people. In the state of the art, social ties are generally categorized as (1) strong ties, (2) weak ties, and (3) absent ties [9, 17], whereas the strength of tie depends upon (a) amount of time, (b) emotional intensity, (c) intimacy variables, and (d) social distance [3, 10]. The repeated presence of two individuals in a specific geographical location within a limited time also infers a social connection [7]. The strength of the social tie is directly proportional to the happening of such high cooccurred events.

IoT has emerged as one of the most powerful and impressive technological research domains [2]. IoT presents a novel connectivity concept, where machines can equally collaborate with humans based on actuators and sensors [27, 28]. One research forecasts that smart devices such as electronic medical kits and smart watches will reach up to the worth of USD 160 billion by the end of 2026 [28]. The communication network between smart devices and human forms Social Internet of Things (SIoT) and further opens up new research challenges for researchers. Managing problems such as data scalability, velocity, and variety are few of the emerging issues in SIoT [29]. Understanding social ties among human-to-human, human-to-machines, and machines-to-machines helps to quantify the network performance issue [30].

Social ties are the backbone of any social network [31]. Formation and deformation of social ties affect communities in a network [32]. Besides social tie strength: factors such as location, emotion, situation, age, gender, religion, personality, and many more have a substantial impact on the social connection [10, 33]. Granovetter highlighted the strong connection between weak social connections and finding jobs [34]. In the literature, sources of data commonly used for social analysis are call logs [35, 36], emails, and social-networking websites [5]. In the literature, extensive challenges associated with the integration of visible and invisible networks are highlighted [37]. Investigating criminal social networks using limited clues is one of the emerging research areas of social network analysis (SNA) [38, 39].

Statistically, there are always some hidden or visible associated parameters among cooccurred events. Social network analysis is performed to explore such intriguing knowledge. In the physical world, social network analysis is utilized in job searching [4], studying urban life psychology, investigation of guilt association [12], finding communities [40], spreading of news [41], and influential networks [18, 42]. In the recent era of information and technologies, massive logs are generating for each person, e.g., call records, bank transactions, online purchase records, daily emails, CCTV cameras, and much more mediums [7, 43, 44]. In contrast to the physical world, such mediums further concise the accuracy of results by highlighting such associated features. Despite numerous data sources, there is no optimal procedure to quantify stochastic human nature and social network evolution [45].

The grouping method identifies hidden social groups, which further explores the friend circles and focuses under high privacy settings [46]. Another research explores the hidden social ties using respondent sampling [47]. In the literature, hidden social ties refer to that population, which is hard to access. The population that tries to hide from the social network is hidden in a network [47]. In our research, suspect ties mean actors in a social network that are present and accessible, but they try to hide their social connections. Our second part of the research explores the suspicious ties within the existing network instead of a hidden population in a social network.

3. Data Set Characteristics and Evaluation Measure

3.1. Data Set Descriptions. In our research, we incorporated three large location-based data sets, i.e., CDR, Brightkite, and Gowalla [13]. The CDR large data set used in this study was provided by one of the Chinese mobile telecommunication operator companies. The data set contains 202,000 subscribes along with user demographic information. Calling detailed records contain six months (June 2014—December 2014), and calling detailed records contain these 202, 000 subscribes, which have 221, 451, 169 records. Each record of the data set is represented in the following format.

Caller ID	Call Type	Callee ID	Time

Duration	LAC ID	CELL ID

Brightkite and Gowalla are openly available location-based social network data sets [13, 48]. Both data sets are gathered using the online social-networking websites. Websites maintain user check-in data by fetching mobile GPS location data. These services use to help people in finding the nearby users and to build social connection. Brightkite contains 58,228 nodes and 214,078 edges, and Gowalla contains 196,591 nodes and 950,327 edges. Other than social network data, both data sets also contain direct social tie data.

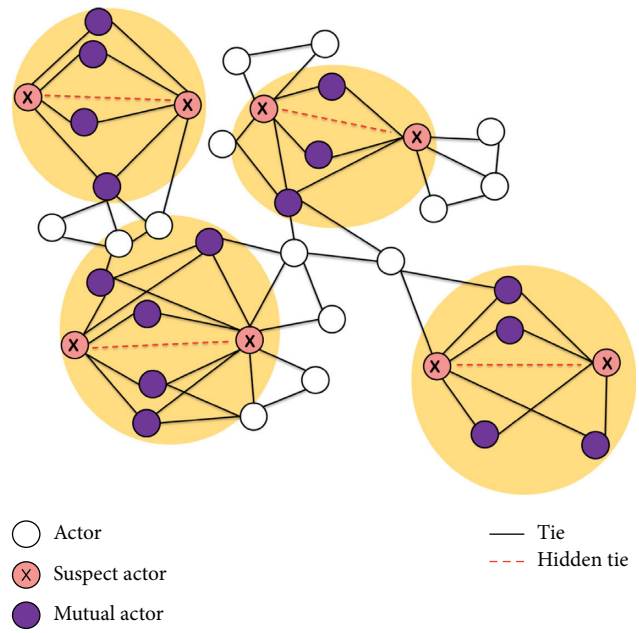


Figure 3: Hidden social ties referring suspects.

3.2. Abbreviations and Evaluation Measures. Figure 3 states the example of the social network, having a case of suspect actors and their hidden ties. Actors with several mutual friends but do not have direct connection may have a secret connection. This information helps in identifying them as a suspect tie. The social network evolves, and new connections expand the scope of the social network. One social network is a combination of multiple social networks involving different individuals [31]. A social network can be sliced based on starting and ending time. Social networks can also be divided into subsocial networks monthwise if it has been developed over one year [49, 50]. Social network slicing helps our research further to explore the missing ties between friends of friend relationships.

The following list of abbreviations is used for the quantification processes of Precision and Recall, which will also be used in several parts of the paper:

SK = calling record

The calling records represent the actual number of direct calls that occur between two users. The value of SK is counted to identify the social tie between two users.

CK = times of cooccurrence

Time cooccurrence represents the presence of two users in the range of a common base station. We counted CK when two users were connected to a common base station, and they called any other user.

G = time − frame gap value

The time-frame gap value represents the time interval between two users' calls while connected to a specific base station. For example, user X calls someone at 2 pm and user Y calls someone else at 4 pm; in this case, the gap between the calls is 2 hr. To quantify the results and

evaluate Precision and Recall values, we experimented on the following set of gap values: $G = 30$ mins, 1 hr, 2 hr, 6 hr, 12 hr, and 24 hr.

SNK = threshold for direct calls

The threshold for direct calls represents the set of threshold values for assessing direct calls between two users. We evaluated Precision and Recall curves on the bases of the following set of threshold values SNK = 2, 5, 10, and 15.

CTK = threshold for cooccurrence

The threshold for cooccurrence represents the set of threshold values for the two users' presence in a specific base station. We tested the performance on the CTK threshold values ranges from 1 to 40.

MuF = threshold for mutual friend

The threshold for mutual friend represents the set of threshold values for the two users' mutual friend. We tested the performance on the MuF threshold values ranges from 1 to 100:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

$$\text{where as} \begin{cases} \text{TP} = (\text{SNK} \geq \text{SK and CTK} \geq \text{CK}), \\ \text{FP} = (\text{SNK} < \text{SK and CTK} \geq \text{CK}), \\ \text{FN} = (\text{SNK} < \text{SK and CTK} < \text{CK}), \\ \text{TN} = (\text{SNK} \geq \text{SK and CTK} < \text{CK}), \end{cases} \tag{4}$$

$$F1\ \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5}$$

3.3. Cooccurrence Count Normalization Measure. Cooccurrence count value CV tells the presence of two users in the region of one base station. An issue related to CV counting is explained and resolved using an example for the two users X and Y, shown in Figure 4. We counted CV when two users were connected to a common base station, and they called any other user in a specific time frame. The example is shown in Figure 4 states the call log details of users X and Y gathered in a time frame T.

Let $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and $Y = \{y_1, y_2, y_3\}$, then

$$\text{CV} = \{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_i\}, \tag{6}$$

where $\alpha_i = \text{Time Gap}(x_i, Y)$, $i = 1$ to 6, $j = 1$ to 3,

$$\text{Time Gap}(x_i, Y) = \min\{(x_i, y_1), (x_i, y_2), \ldots, (x_i, y_j)\}. \tag{7}$$

In Figure 4, x_1 and x_2 call times have the closest call time to y_1 call. In this case, a count value of CV can be calculated as 2. However, such counting may lead to a wrong inference. It is the same as if one person calls once, and another person calls n-times within a specific time frame, equals n as the count value. To resolve this issue, we propose a normalization equation that decreases the count value periodically. We introduce Beta (β) value as a periodic normalizing factor.

Let X denotes a set of calls by user X and Y denotes a set of calls by user Y;if $(X \geq Y)$, then $X \times Y$ else $Y \times X$.

According to the example stated in Figure 4, we assumed β for set $Y = (y_1\beta, y_2\beta, y_3\beta)$.

For first match value of $\beta = 1$.

For the second match values of $\beta = \beta/2$.

Likewise, for the nth match value of $\beta = \beta/n$,

$$\begin{aligned} &= \left[((x_1, y_1) \times (y_1\beta)) + \left((x_2, y_1) \times \left(\frac{y_1\beta}{2}\right)\right)\right] \\ &\quad + \left[((x_3, y_2) \times (y_2\beta)) + \left((x_4, y_2) \times \left(\frac{y_2\beta}{2}\right)\right) + \left((x_5, y_2) \times \left(\frac{y_2\beta}{3}\right)\right)\right] \\ &\quad + [(x_6, y_3) \times (y_3\beta)], \end{aligned} \tag{8}$$

$$= \sum_{m=1}^{mk} \sum_{n=1}^{nk} \left((x_n, y_m) \times \left(\frac{y_m\beta}{n}\right)\right). \tag{9}$$

In equation (9), mk refers to the total number of calls made by user X to Y, while nk refers to the total number of calls made by user Y to X. equation (9) finds intermediate value for CV, i.e., 4.33, instead of maximum 6 or minimum 3 values.

4. Social Tie Inference

We initially investigated direct social ties formed by CDR data sets and compared them to the indirect social ties formed based on common location using Algorithm 1. By

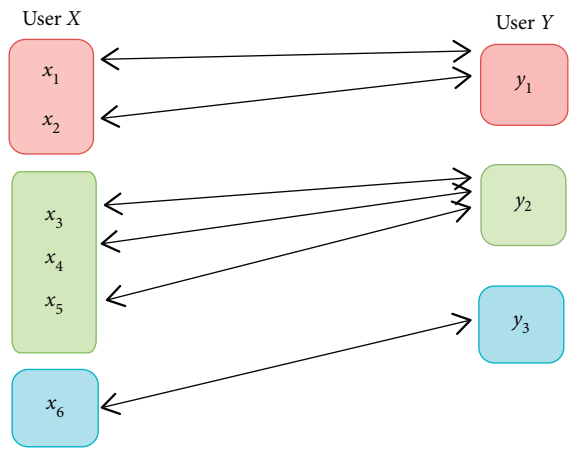


Figure 4: User X and user Y cooccurrence count comparison case.

Require:
SN = Social Network
GV = Time Gap Threshold Value
SNK = 2, 5, 10, 15
Ensure:
DT = Direct Social Ties
IST = Infer Social Ties
(1) **while** SN ≠ 0 **do**
(2) DT = **Direct Ties** (SNK)
(3) **end while**
(4) **while** SN ≠ 0 **do**
(5) CTK = **Calculate Cooccurrence Count** (DT, GV)
(6) IST = **Infer Social Ties** (CTK, DT, SNK)
(7) **end while**

Algorithm 1: Social tie inference.

direct ties, we mean calling or direct connection. For example, person A calls person B refers to a direct tie between A and B. Algorithm 1 takes GV, SNK, and CDR data sets (social network) as inputs. Furthermore, the algorithm has two parts; initially, it finds the direct ties between two individuals depending upon the SNK threshold value. Secondly, it counts the presence of two individuals based on several parameters. The **Calculate Cooccurrence Count()** function finds the number of cooccurrences using equation (9), explained in the previous section. **Infer Social Ties()** function finds the social connections depending upon CTK, DT, and SNK and inferred them as the social ties.

4.1. CDR-Based Social Tie Inference. A social tie is inferred between two persons if they are found together at several sites numerous times. The inference algorithm identifies two sets of results, i.e., direct social ties and inferred social ties. For the cross-validation of results, we correlate the direct tie results with the inferred ones. Precision and Recall evaluation measures are used to examine the results. We tested all records based on threshold values, K is the direct calls, M is the times of cooccurrence, and G is the time frame gap value. While N as direct call count shows the degree of friendship, more value of N indicates the friendship strength. Figure 5 shows the Precision graph, which contains four sets, Figures 5(a)–5(d). The whole data set is examined based on K and the value of M and v.

In Figure 5(a), the value of K is 15 which represents the users with direct calls between each other equals to or greater than 15. The value of M is the number of cooccurrence for two different users. The Precision values are comparatively significantly less for M in the range of 0 to 10. In contrast, the value of Precision increases exponentially for the value of M in the range of 10 to 30. The higher value of M indicates higher cooccurrence of users. A positive correlation can be observed between the values of M and Precision. It infers that cooccurrence is a significant attribute that affects positively in identifying social ties. All graphs in Figure 5 have six different lines; each line represents the different time gap ranges. It can also be seen that the values of gap value 30 minutes are having more Precision while the rest lines of 1 hour, 2 hours, 6 hours, 12 hours, and 24 hours are having less Precision. This also clues that the strength of ties has a specific effect on Precision. Users having strong social connections, most of the time, are found together in certain areas. This pattern is explicitly observed by looking cooccurrence value $M = (20 \text{ to } 40)$ and gap time frame $G = 30$

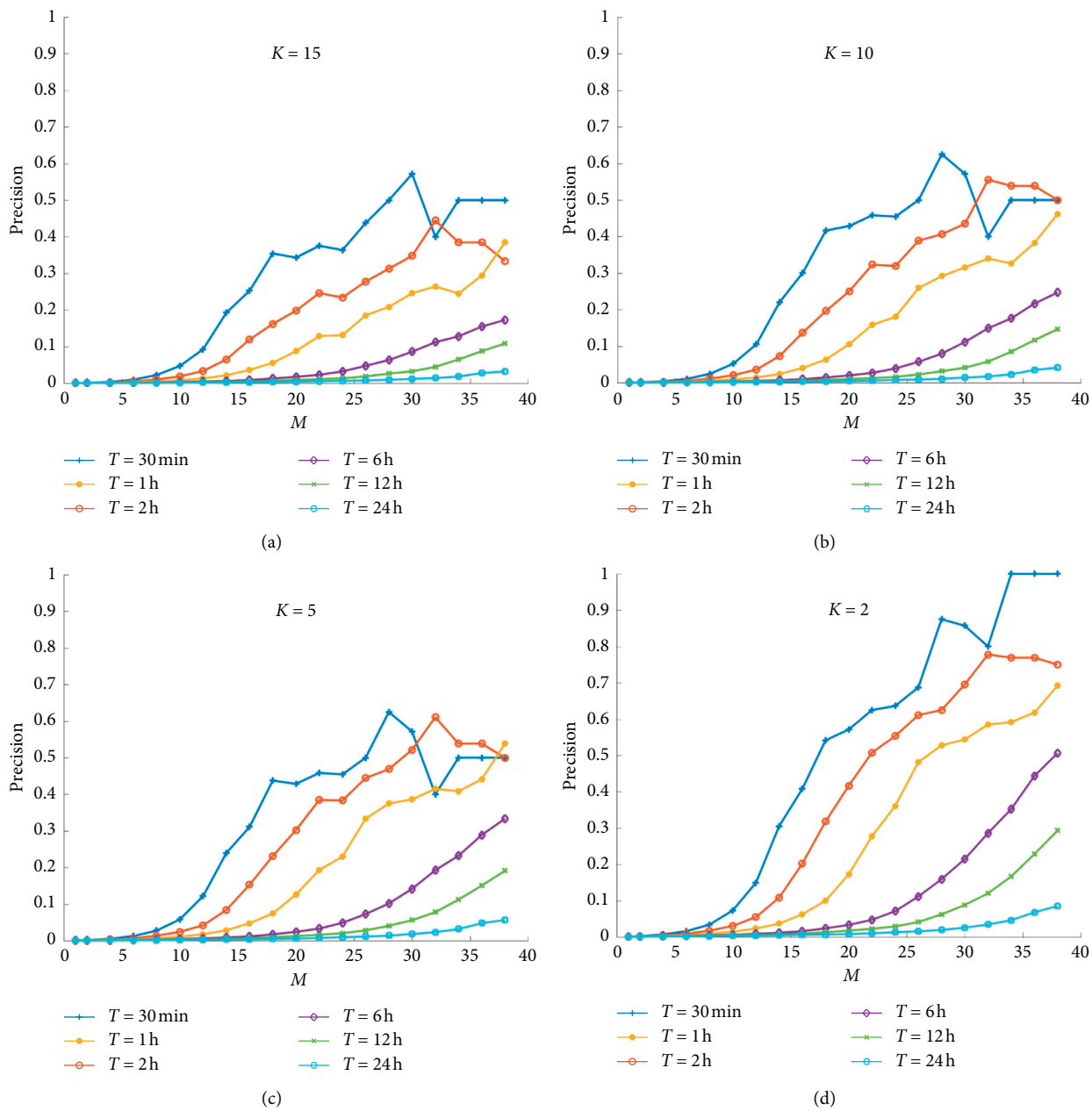


Figure 5: The Precision formed between two callers due to social contact depending on the number of times they have connected through the same base station within a specific time frame. Figure 5(a) shows the Precision for the six different time frame lines, whereas M is the number of cooccurrences and K is the number of direct calls. (a) $K = 15$, (b) $K = 10$, (c) $K = 5$, and (d) $K = 2$.

minutes. Another positive correlation is found between the degree of friendship and physical presence at a specific place.

To see the effect of friendship strength, we evaluated results for the four different K values, i.e., 2, 5, 10, and 15. A typical pattern is found in all the graphs shown in Figure 5. It shows that Precision is less for people whose mutual presence is less at different sites. Also, people with strong social ties spent less than 1 hr time together at a specific location. To understand the graph's actual meaning, we quantify and reconcile with the actual direct social ties. It is observed that a positive correlation in results infers that people with strong social connections often visit places together.

Figure 6(a) represents the Recall results. We tested and evaluated Recall based on the same measures as Precision, i.e., direct calls, cooccurrence, and gap time frame. Figure 6(a) states that the value of Recall is at a maximum when we consider a less number of commonplaces. An inverse trend is observed between the values of Recall and M, specifically for the M value ranging between 0 and 3. The same as Precision, Recall is also evaluated based on six different time gap values.

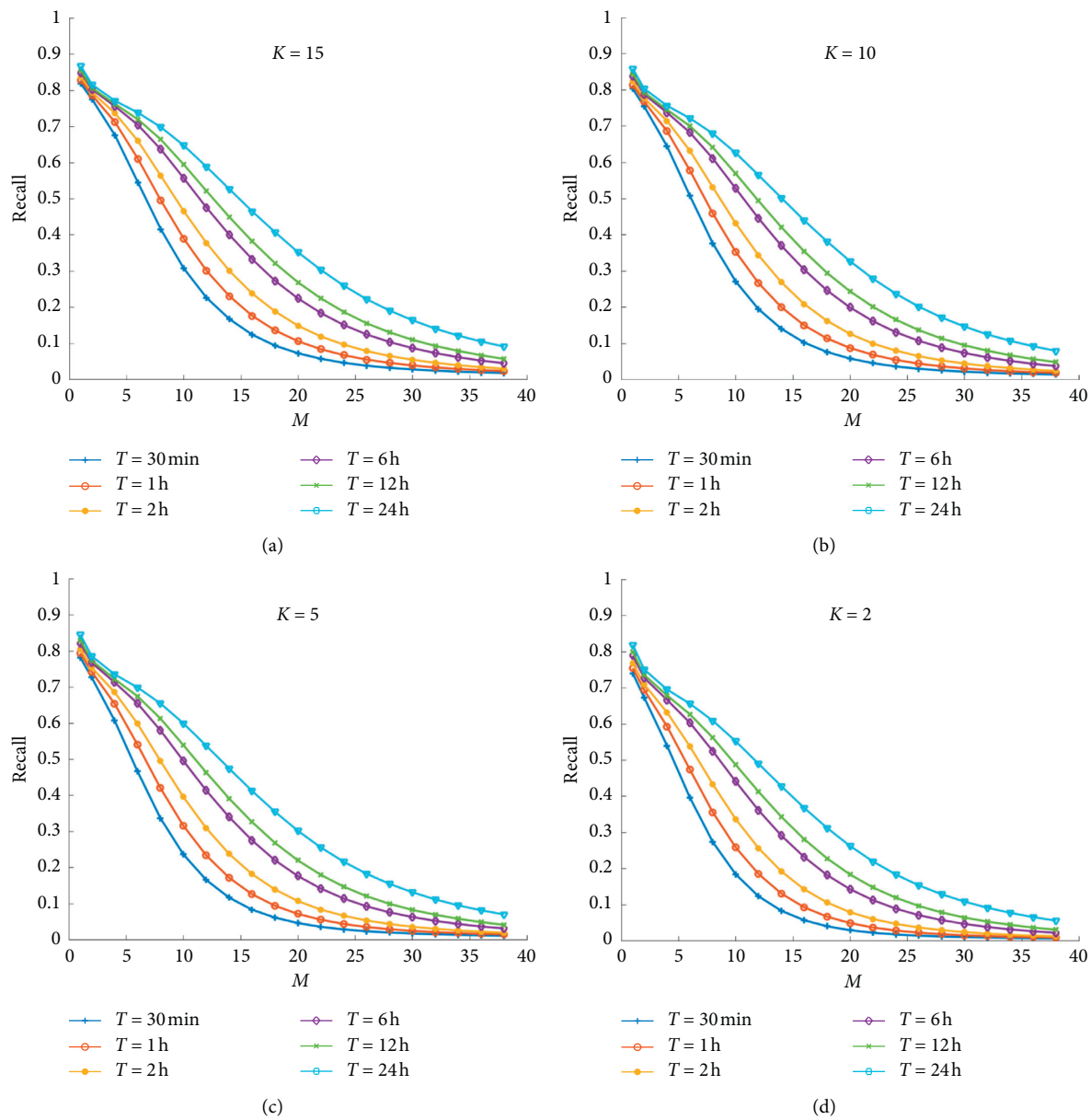


Figure 6: The Recall formed between two callers due to social contact depending upon the number of times they have connected through the same base station within a specific time frame. Figure 6(a) shows the Recall for the six different time frame lines, whereas M is the number of cooccurrences and K is the number of direct calls. (a) $K = 15$, (b) $K = 10$, (c) $K = 5$, and (d) $K = 2$.

This part of the research finds people's cooccurrence based on the same base station connectivity in a specific time frame and infers them as social ties. Furthermore, it cross relates the inferred results with direct call results.

4.2. Brightkite- and Gowalla-Based Social Tie Inference. Brightkite and Gowalla data sets contain direct social ties as well as the check-in information of each user. In our study, we investigated both data sets based on several dimensions and found some of the very interesting facts. During the analysis, we observed positive correlations between mutual friend count values MuF and user check-in details. In Brightkite, MuF ranging 6 to 40 and in Gowalla, MuF ranging 25 to 90 shows the positive corelation with Precision. We also measure the effect of gap time value on the Precision and Recall and evaluated results based on several gap time values.

Figures 7(a) and 7(c) of line graph show the relation between mutual friend count values MuF and Precision, while Figures 7(b) and 7(d) show the relation between mutual friend count values MuF and Recall. In Figure 7, results shows that people having a certain number of mutual friends use to visit place together or with little gap of interval.

The social tie inference framework infers some of the absent ties as social ties in the form of false-positive results. To further extract the actual meaning of such incorrect

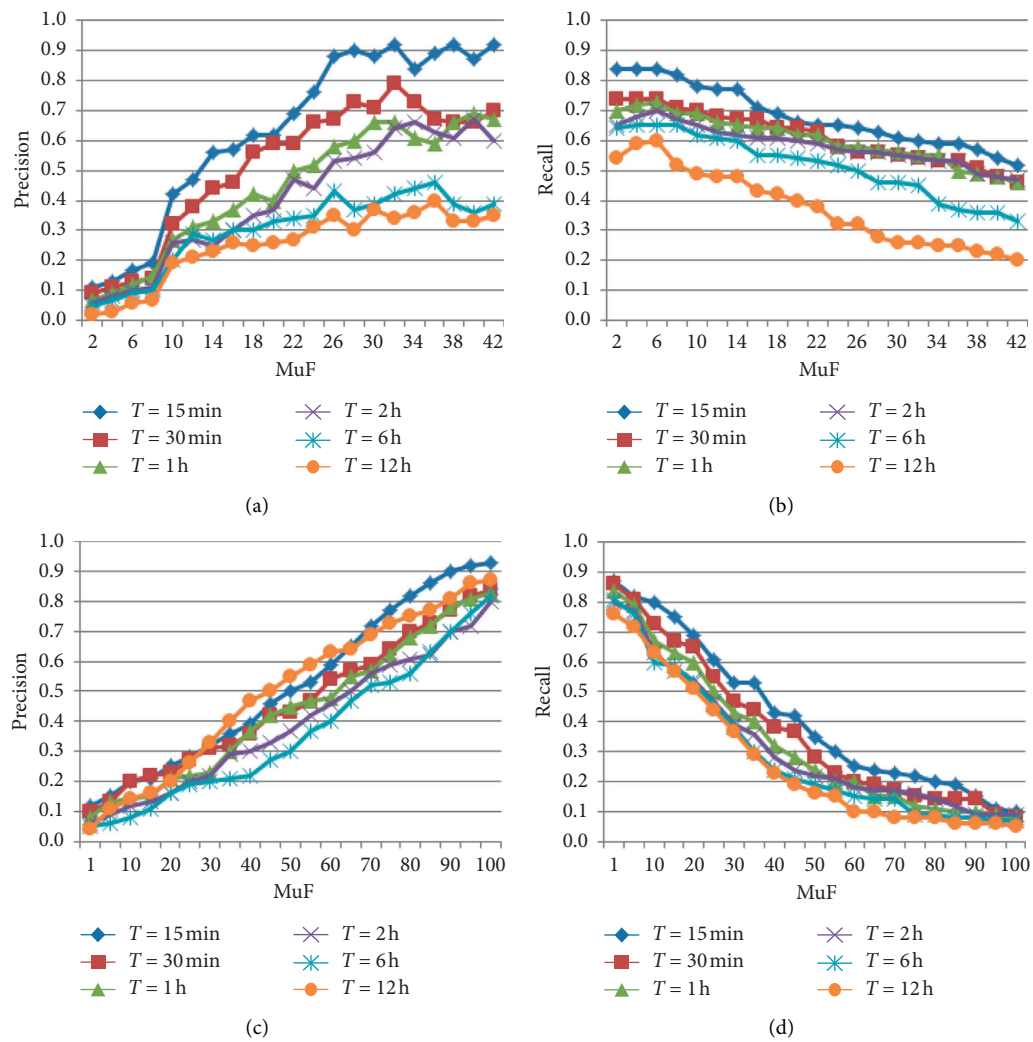


Figure 7: The Precision and Recall formed between two users due to social contact depending on the number of times they have checked in to certain location within a specific time frame. Figure 7(a) shows the Precision for the six different time frame lines, whereas MuF is the number of mutual friend count. (a) Brightkite Precision, (b) Brightkite Recall, (c) Gowalla Precision, and (d) Gowalla Recall.

inferences, we conducted a social activity and simulation. The false-positive results of the first part of the research serve as the foundation for the second part. Activity under the first part data set is conducted, and the false-positive results are examined by studying the human cooccurrence patterns, described in the next section Suspicious Ties. This stage of research gave us a clue to further exploit the category of missing ties.

5. Suspicious Ties

An absent tie can be inferred as a suspect tie, if it satisfies the following properties:

(1) M, number of mutual friends

(2) C, number of cooccurrence on different sites

In the CDR data set, each cell is treated as a single cell of the base station. Our model adopts the following four features for evaluation:

(1) Base station

(2) User ID

(3) Gap time threshold

(4) Call time stamp

The following mathematical model explains the problem and its formulation:

Let BS denotes a set $\{bs_i\}$ of bs_n points, and bs_i is called as the distinct base station cell

Let DU denotes a set $\{du_i\}$ of du_n points, and du_i is called as the distinct user

Let TH denotes a set $\{\mathrm{th}_i\}$ of th_n points, and th_i is called as the threshold value for timeframe

Let C denotes a set $\{c_i\}$ of c_n points, and c_i is called as the call information,

$$c_i = (\mathrm{id}, \mathrm{ct}), \quad \text{where} \begin{cases} \text{id as call id,} \\ \text{ct as outgoing call time.} \end{cases} \tag{10}$$

Let R denotes a set $\{r_i\}$ of r_n points, then

$$r_i = [(\mathrm{d}u_i, c_i(\mathrm{id})) \times (\mathrm{d}u_i, c_i(\mathrm{id}))] \wedge \mathrm{d}u_i \neq \mathrm{d}u_i. \tag{11}$$

Let RT denotes a set $\{rt_i\}$ of rt_n points, then

$$rt_i = [(\mathrm{d}u_i, c_i(\mathrm{ct})) - (\mathrm{d}u_i, \mathrm{ct}_i(\mathrm{ct}))] \wedge \mathrm{d}u_i \neq \mathrm{d}u_i. \tag{12}$$

Let S denotes a set $\{s_i\}$ of s_n points, then

$$s_i = [(r_i \in bs_i) \wedge (rt_i \leq th_i)], \tag{13}$$

wheres_i is a set of elements that identifies distinct callers based on the same base station connectivity and a definite number of calls in a specific time frame.

6. Suspect Inference Framework

We studied the pattern of exceptional cases belong to the false-positive set and described a subset of the false-positive set as suspect ties. Physical activity was designed and conducted to investigate the formation of suspect social ties. Activity consisting of 50 people, and a data set was generated within almost 4-5 hours. A basketball court was utilized for the activity. Nine circles were drawn physically on the basketball court, assumed as the base station cell. Out of 50 people, nine were directed to act as a base station. The boundary of each circle was considered as a range of the base station cell. Rests of the 41 persons were directed to perform the following two steps.

6.1. Selection Step. Initially, each person from 41 people was asked to choose two sets of friends. One set as obvious friends and the second set of hidden friends such that the size of the hidden friend set should be at most 1/5 of the obvious friend set size, e.g., if one person has five people in apparent friend set, he can have no more than one hidden friend. After the selection of both sets by each individual, information was shared with one of our representatives.

6.2. Operation Step. In this phase, each person was directed to follow the following rules:

(1) You should not call your hidden friend

(2) You should call all of your obvious friends at least once

(3) You should conduct a maximum number of calls to your closest obvious friend and second maximum to a second level obvious friend and likewise to the least friend

(4) You must try to meet your hidden friend as much as possible physically

The method of calling is like, if person A wants to call person B from base station $B1$, the person has to go to the base station $B1$ and register a call with a person acting and standing in base station $B1$. Respective base station person will write and make an entry with five parameters, i.e., *Caller Name*, *Callee Name*, *Time*, *From Base Station Name*, and *To Base Station Name*. The data set was gathered in the following format. An example is given below.

Caller Name	Callee Name	Time Stamp	From Base Station	To Base Station
User 1	User 12	02/03/2019 10:15:08	BS 7	BS 2
..	..	..	..	..

For the understanding of the variations and patterns, the same activity was also designed using simulation. The whole simulation followed the same conditions, and another data set was generated using a random function. Based on the activity, a framework is designed to separate a class of suspect ties. Proposed inference framework work is designed and implemented using pipe and filter architecture, shown in Figure 8. Algorithm 2 shows the implementation of suspicious tie inference framework, explained in Figure 8. The framework takes the social network matrix, count threshold value, gap time value, and mutual friend count value as inputs and filters the result. Initially, **Calculate Levels()** function finds the five level depth information for each distinct user. Let us say if A calls B, B calls C, C calls D, and D calls E, it implies that $A = 0$, $B = 1$, $C = 2$, $D = 3$, and $E = 4$ represent five levels. This step ensures that all the levels have distinct users. Secondly, **Find Suspects()** function selects only those sets of users from level 1 and level 3 that do not have any direct calls and M number of mutual friends. Furthermore, **Calculate Subsocial Network()** function generates subgraphs using level 1 and level 3 details depending upon the gap time value. Results are filtered on these bases of time gap value, e.g., two users called some other user while connected to the same base station within the given time frame, explained in the previous section. After that, **Infer Hidden Ties()** function uses the proposed normalization method to find the number of the count, defined in equation (9), and then all results are filtered according to the cooccurrence count CV and the mutual friend threshold value M. Based on mentioned parameters and thresholds, Algorithm 2 significantly identifies the subclass of missing ties as suspicious ties.

The results of the activity and simulation are computed and evaluated using Precision and Recall measures.

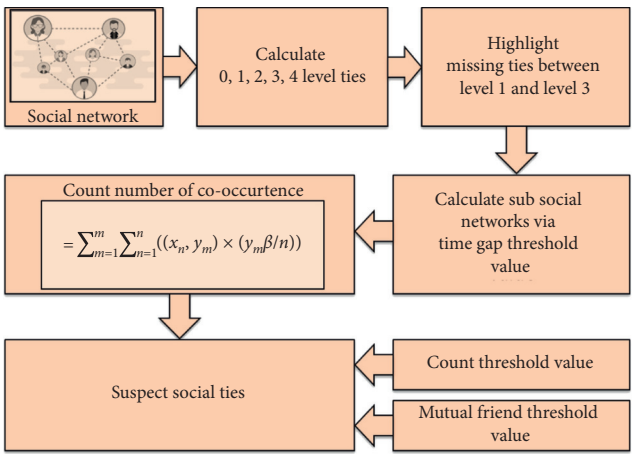


Figure 8: Suspect social tie inference framework.

Require:
SN = Social Network
GV = Time GapThreshold Value
CV = Cooccurrence Count Threshold Value
M = Mutual Friend Count
Ensure:
ST = Inferred Social Suspect Ties
(1) **while** SN ≠ 0 **do**
(2) Levels [5, n] = **Calculate Levels()**
(3) Suspects = **Find Suspects** (Level 1, Level 3)
(4) SG = **Calculate Subsocial Network** (Suspects, GV)
(5) ST = **Infer Suspect Ties** (SG, CV, M) using equation (9)
(6) **end while**

Algorithm 2: Suspicious hidden social tie inference.

Table 1: Evaluation of simulation and social activity.

	Gap time	$M \geq 2$						$M \geq 5$					
		Simulation			Social activity			Simulation			Social activity		
		P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
CV ≥ 5	GV < 10	0.12	0.68	0.20	0.12	1.00	0.21	0.13	0.62	0.21	0.12	0.79	0.21
	GV < 20	0.10	0.42	0.16	0.15	0.79	0.25	0.15	0.31	0.20	0.13	0.42	0.20
	GV < 30	0.21	0.22	0.21	0.71	0.59	0.64	0.25	0.18	0.21	0.23	0.28	0.25
	GV ≥ 30	0.34	0.23	0.27	1.00	0.12	0.21	0.30	0.10	0.15	0.32	0.19	0.24
CV ≥ 10	GV < 10	0.10	0.33	0.15	0.12	0.48	0.19	0.09	0.30	0.14	0.11	0.61	0.19
	GV < 20	0.11	0.28	0.16	0.27	0.50	0.21	0.09	0.18	0.12	0.10	0.39	0.16
	GV < 30	0.13	0.18	0.15	0.29	0.30	0.29	0.09	0.10	0.09	0.09	0.20	0.12
	GV ≥ 30	0.13	0.08	0.10	0.26	0.05	0.08	0.10	0.04	0.06	0.13	0.07	0.09
CV ≥ 20	GV < 10	0.04	0.18	0.07	0.08	0.23	0.12	0.05	0.29	0.09	0.11	0.39	0.17
	GV < 20	0.07	0.08	0.07	0.05	0.24	0.08	0.04	0.11	0.06	0.10	0.28	0.15
	GV < 30	0.04	0.06	0.05	0.09	0.09	0.09	0.03	0.05	0.04	0.09	0.19	0.12
	GV ≥ 30	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.05	0.02

*P = Precision, R = Recall, and $F1$ = $F1$ Score.

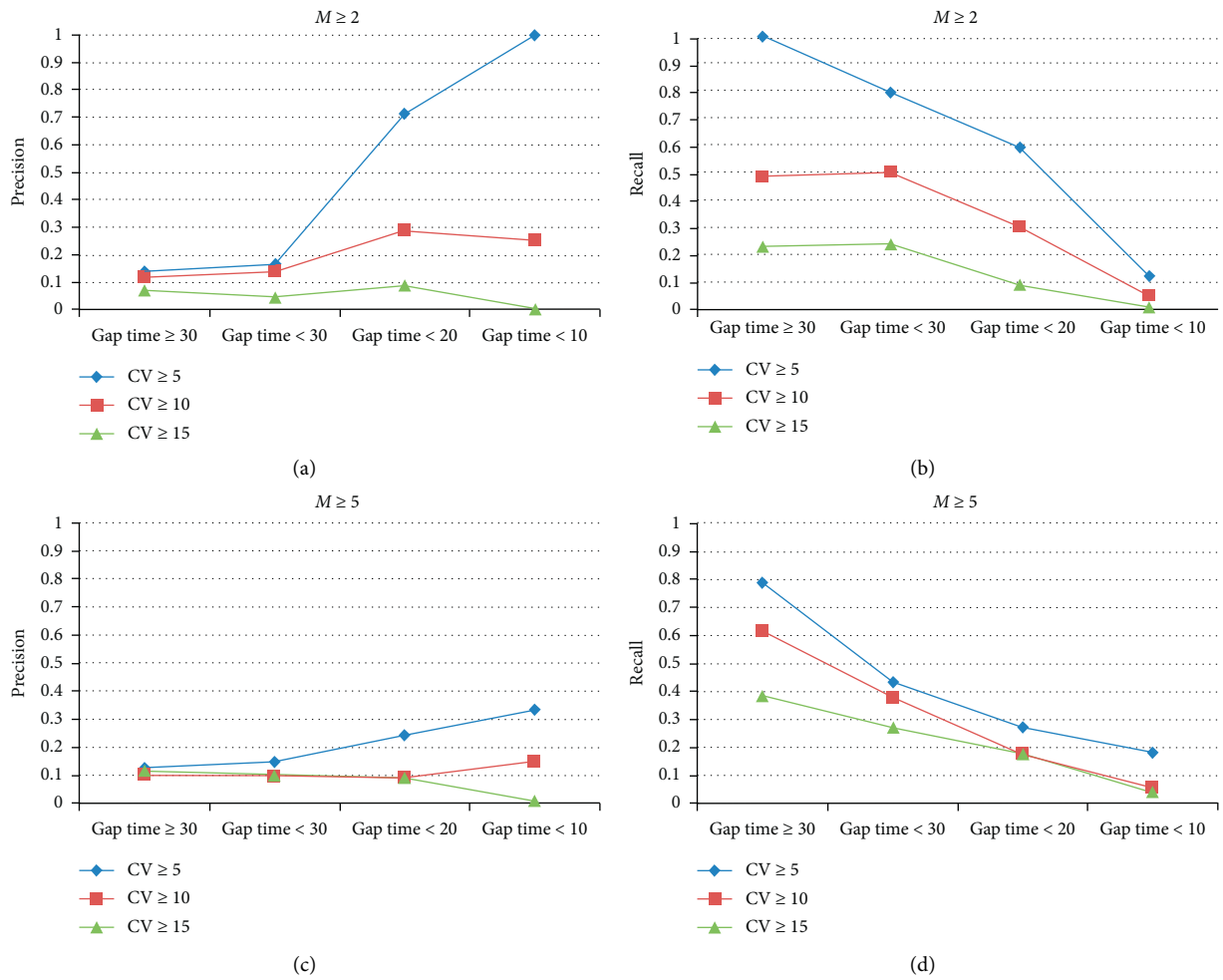


Figure 9: Evaluation of social activity using Precision and Recall.

Evaluation results of simulation and social activity conducted are shown in Table 1. Precision, Recall, and $F1$ Score measures are used to evaluate the framework that is further calculated based on the cooccurrence count CV, mutual friend count M, and gap time GV parameter setting values. $F1$ Score is calculated using equation (5). Definitions of the related parameters are given as follows.

TP	Which were hidden friends and system infer as hidden friends
TN	Which were not hidden friends and system infer as not hidden friends
FP	Which were not hidden friends and system infer as hidden friends
FN	Which were hidden friends and system infer as not hidden friends

6.3. Results and Discussion. We tested and evaluated all records based on the cooccurrence count value CV, mutual friend M, and gap time value GV as the gap time. Figures 9(a)–9(d) show the evaluation of results for the activity, based on three different values of the cooccurrence count value CV, i.e., CV = 5, 10, and 15, and two different values of the mutual friend M, i.e., $M \geq 2$ and ≥ 5. Likewise, Figures 10(a)–10(d) show the evaluation of results for the simulation data set along with mentioned CV and M values. Results were generated on four values of gap time, i.e., ≥ 30, < 30, < 20, and < 10.

In Figures 9(a) and 9(b), Recall is maximum and Precision is less where the value of GV ≥ 30, CV ≥ 5, and $M \geq 2$. The system obtains a maximum number of relevant hidden ties along with false-positive results. By GV ≥ 30, CV ≥ 5, and $M \geq 2$, it means that the gap between the two calls is 30 mins or more while the count of cooccurred events is kept minimum five and mutual friend count as two or less. The system's performance drops when gap time is reduced to < 30, < 20, and < 10. Even though there is a drop in true-positive values, a significant drop in the false-positive values can be seen. Results become concise, with the variation in both values of GV and M. The limited data set collected using activity highlights the occurrences of hidden ties. Whole activity and simulation were designed to get similar fields of data as CDR so that the proposed framework is compatible with the CDR data set.

Results shown in Figures 9 and 10 exhibit the existence of hidden relationships. The parameter, such as the gap time value GV, helps to identify the time frame selection such that

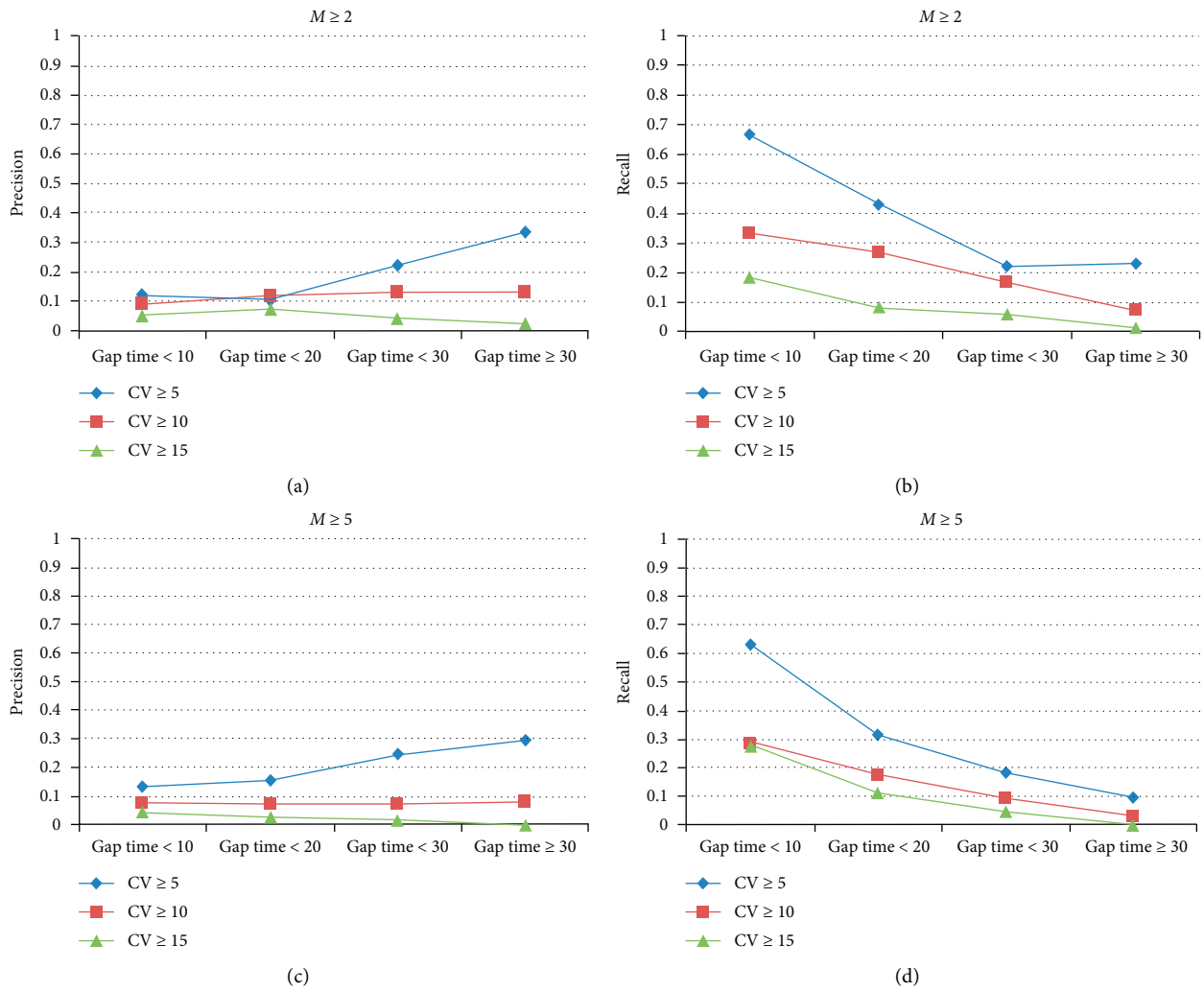


Figure 10: Evaluation of simulation using Precision and Recall.

what gap value is suitable to get optimal performance? Likewise, Figures 10(a) to 10(d) represent the Precision and Recall curves for the simulation data set. Results are gathered by setting the same values for the threshold parameters CV, GV, and M. Figures 10(a) and 10(d) give the least values for Precision and Recall, which tells that if threshold values are tightened, so does the count relevant retrieve results get low. Out of all the results, Precision and Recall values for both simulation and activity data set show maximum, if threshold values GV ≥ 30, CV ≥ 5, and $M \geq 2$ are set. We found some exciting dissimilarities in the results between simulation and activity data set results during the complete evaluation process. The simulation results do not give higher Recall and Precision value compared to activity data results. We concluded that the data set of simulation is generated using random function while the activity had the human hiding patterns. These results also help to infer human psychology in building hidden ties. The simulation data set generator works based on the same constraints mentioned as rules of the activity. However, the key difference between simulation and activity is the selection of friends, hidden friend, and the pattern of calling. While conducting the simulation, trivial variation in results was observed as the random selection has random patterns. According to our findings, simulation and activity both exhibit patterns of hidden ties. However, activity results are more pronounced and significantly identifying the hidden relationships.

It is essential to highlight some critical questions and dependent variables that help to find hidden social ties between two people, for example, why two people hide their social ties? Is this deliberate or unintentional action? What if they are deliberately hiding their social tie for a purpose? In such a scenario, extracting a social tie for two people is a kind of intense problem. It is a kind of investigation process which explores a clue to draw some relationship between two people. Investigation designates if two people are posting a picture on social media to be more cautious about not identifying their social ties. While if they are doing some private activity, they will be less careful, for example, posting a picture on Facebook or Flickr compared to calling to a person from a specific location. Although extracting hidden social ties include various privacy issues. We designed

activity and simulation that generated the data set by the caller data record (CDR) data set. We thoroughly investigated the patterns of connectivity and established a framework to infer social ties. This research has opened up a new direction further to explore connectivity in the Social Internet of Things (SIoT), specifically machine-to-machine direct communication and machine-to-human or human-to-machine hidden relationships. In many cases, several machines work together but they rarely have a direct connection.

7. Conclusion and Future Work

In this research, we have examined the developments of social connection patterns based on physical gathering. In the first part of our research, we explored the correlation between direct communication and two individuals' physical presence. To check the system's performance and evaluation, we examined all results by utilizing Precision and Recall evaluation measures. We also present a periodic normalization equation for the cooccurrence count. In the second phase, we propose the suspect tie inference framework. False-positive results of the first part of the research are the ground to the second part of the study. The proposed framework adopts pipe and filter architecture, where the threshold values control each filter. The framework's fundamental objective is to take the data set such as CDR (caller data record) and infer suspect social ties, depending upon the specified threshold values. Analyzing the results critically, we propose a theory that identifies suspect social ties. Besides this, for comparison and evaluation, we conducted real-time human-based activity and simulation. Keeping in mind the structure of the actual CDR data set, the whole activity was designed and evaluated. In contrast to existing work, our research focus is on hidden ties instead of the hidden actors. In the future, we are aiming to explore the homophilic nature of suspect ties within the Social Internet of Things (SIoT).

Data Availability

The data used can be found at http://snap.stanford.edu/data/index.html#locnet.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this work.

Acknowledgments

This work was supported by King Saud University, Saudi Arabia, through research supporting project number RSP-2020/184. Nauman Ali Khan acknowledges the support of the Chinese Government and Chinese Scholarship Council (CSC) for his PhD studies at the University of Science and Technology, China. This research work was partially supported by Key Program of the National Natural Science Foundation of China (grant number. 61631018).

References

[1] L. Weng, M. Karsai, N. Perra, F. Menczer, and A. Flammini, "Attention on weak ties in social and communication networks," in *Complex Spreading Phenomena in Social Systems*, pp. 213–228, Springer, Berlin, Germany, 2018.

[2] A. M. Ortiz, D. Hussein, S. Park, S. N. Han, and N. Crespi, "The cluster between internet of things and social networks: review and research challenges," *IEEE Internet of Things Journal*, vol. 1, no. 3, pp. 206–215, 2014.

[3] P. Luarn and Y.-P. Chiu, "Key variables to predict tie strength on social network sites," *Internet Research*, vol. 25, no. 2, pp. 218–238, 2015.

[4] D. Goel and K. Lang, "Social ties and the job search of recent immigrants," *ILR Review*, vol. 72, no. 2, pp. 355–381, 2019.

[5] H. Kizgin, A. Jamal, N. Rana, Y. Dwivedi, and V. Weerakkody, "The impact of social networking sites on socialization and political engagement: role of acculturation," *Technological Forecasting and Social Change*, vol. 145, pp. 503–512, 2019.

[6] P. Suebvises, "Social capital, citizen participation in public administration, and public sector performance in Thailand," *World Development*, vol. 109, pp. 236–248, 2018.

[7] Z. Liu, Y. Qiao, S. Tao, W. Lin, and J. Yang, "Analyzing human mobility and social relationships from cellular network data," in *Proceedings of the 2017 13th International Conference on Network and Service Management (CNSM)*, pp. 1–6, Tokyo, Japan, November 2017.

[8] L. Matosas-López and A. Romero-Ania, "The efficiency of social network services management in organizations. an in-depth analysis applying machine learning algorithms and multiple linear regressions," *Applied Sciences*, vol. 10, no. 15, p. 5167, 2020.

[9] M. S. Granovetter, "The strength of weak ties," in *Social Networks*, pp. 347–367, Elsevier, Amsterdam, Netherlands, 1977.

[10] T. Nguyen, L. Zhang, and A. Culotta, "Estimating tie strength in follower networks to measure brand perceptions," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 779–786, Vancouver, Canada, August 2019.

[11] H. Mattie, K. Engø-Monsen, R. Ling, and J.-P. Onnela, "Understanding tie strength in social networks using a local bow tie framework," *Scientific Reports*, vol. 8, no. 1, p. 9349, 2018.

[12] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences*, vol. 107, no. 52, pp. 22436–24441, 2010.

[13] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1082–1090, San Diego, CA, USA, August 2011.

[14] A. Žnidaršič, P. Doreian, and A. Ferligoj, "Absent ties in social networks, their treatments, and blockmodeling outcomes," *Advances in Methodology & Statistics/Metodoloski Zvezki*, vol. 9, no. 2, 2012.

[15] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network?: the structure of the twitter follow graph," in *Proceedings of the 23rd International Conference on World Wide Web*, pp. 493–498, Seoul, Republic of Korea, April 2014.

[16] X. Zuo, J. Blackburn, N. Kourtellis, J. Skvoretz, and A. Iamnitchi, "The power of indirect social ties," 2014, http://arxiv.org/abs/1401.4234.

[17] M. A. Brandão and M. M. Moro, "Tie strength in co-authorship social networks: analyses, metrics and a new computational model," in *Proceedings of the Anais Estendidos do XXV Simpósio Brasileiro de Sistemas Multimídia e Web*, pp. 17–20, Rio de Janeiro, Brazil, 2019.

[18] H. A. Khanday, A. H. Ganai, and R. Hashmy, "A comparative analysis of identifying influential users in online social networks," in *Proceedings of the 2018 International Conference on Soft-Computing and Network Security (ICSNS)*, pp. 1–6, Coimbatore, India, February 2018.

[19] R. Garcia-Guzman, Y. A. Andrade-Ambriz, M.-A. Ibarra-Manzano, S. Ledesma, J. C. Gomez, and D.-L. Almanza-Ojeda, "Trend-based categories recommendations and age-gender prediction for pinterest and twitter users," *Applied Sciences*, vol. 10, no. 17, p. 5957, 2020.

[20] A. Talukder, M. G. R. Alam, N. H. Tran, D. Niyato, and C. S. Hong, "Knapsack-based reverse influence maximization for target marketing in social networks," *IEEE Access*, vol. 7, pp. 44182–44198, 2019.

[21] N. Ampazis, T. Emmanouilidis, and F. Sakketou, "A matrix factorization algorithm for efficient recommendations in social rating networks using constrained optimization," *Machine Learning and Knowledge Extraction*, vol. 1, no. 3, p. 928, 2019.

[22] N. Zhou, X. Zhang, and S. Wang, "Theme-aware social strength inference from spatiotemporal data," in *Proceedings of the International Conference on Web-Age Information Management*, pp. 498–509, Macau, China, June 2014.

[23] P. A. Grabowicz, J. J. Ramasco, B. Gonçalves, and V. M. Eguíluz, "Entangling mobility and interactions in social media," *PLoS One*, vol. 9, no. 3, Article ID e92196, 2014.

[24] D. Yang, B. Qu, J. Yang, and P. Cudre-Mauroux, "Revisiting user mobility and social relationships in LBSNs: a hypergraph embedding approach," in *Proceedings of the World Wide Web Conference*, pp. 2147–2157, San Francisco, NY, USA, May 2019.

[25] R.-H. Li, J. Liu, J. X. Yu, H. Chen, and H. Kitagawa, "Co-occurrence prediction in a large location-based social network," *Frontiers of Computer Science*, vol. 7, no. 2, pp. 185–194, 2013.

[26] V. Jayadevan, K. Bharadwaj, A. Kumar, and P. Khandelwal, "Discovering local social groups using mobility data," *International Journal of Computer Applications*, vol. 120, no. 21, 2015.

[27] A. J. Jara, Y. Bocchi, and D. Genoud, "Social internet of things: the potential of the internet of things for defining human behaviours," in *Proceedings of the 2014 International Conference on Intelligent Networking and Collaborative Systems*, pp. 581–585, Salerno, Italy, September 2014.

[28] W. A. D. M. Jayathilaka, K. Qi, Y. Qin et al., "Significance of nanomaterials in wearables: a review on wearable actuators and sensors," *Advanced Materials*, vol. 31, no. 7, Article ID 1805921, 2019.

[29] J. Ren, H. Guo, C. Xu, and Y. Zhang, "Serving at the edge: a scalable iot architecture based on transparent computing," *IEEE Network*, vol. 31, no. 5, pp. 96–105, 2017.

[30] I. Seeber, E. Bittner, R. O. Briggs et al., "Machines as teammates: a research agenda on ai in team collaboration," *Information & Management*, vol. 57, no. 2, Article ID 103174, 2020.

[31] T. Pi, L. Cao, P. Lv, Z. Ye, and H. Wang, "Inferring implicit social ties in mobile social networks,," in *Proceedings of the 2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, Barcelona, Spain, April 2018.

[32] N. Tahir, A. Hassan, M. Asif, and S. Ahmad, "MCD: mutually connected community detection using clustering coefficient approach in social networks," in *Proceedings of the 2019 2nd International Conference on Communication, Computing and Digital Systems (C-CODE)*, pp. 160–165, Islamabad, Pakistan, March 2019.

[33] V. A. Lewis, C. A. MacGregor, and R. D. Putnam, "Religion, networks, and neighborliness: the impact of religious social networks on civic engagement," *Social Science Research*, vol. 42, no. 2, pp. 331–346, 2013.

[34] M. S. Granovetter, "The strength of weak ties: a network theory revisited," *Sociological Theory*, vol. 1, pp. 201–233, 1983.

[35] J. Gui, Z. Zheng, X. Zhao, and Z. Qin, "Statistical properties and temporal properties of calling behavior," in *Proceedings of the 2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, pp. 1940–1944, Chengdu, China, December 2019.

[36] I. H. Sarker, "A machine learning based robust prediction model for real-life mobile phone data," *Internet of Things*, vol. 5, pp. 180–193, 2019.

[37] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer, "Quantifying the invisible audience in social networks," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 21–30, Paris France, April 2013.

[38] R. Montasari, R. Hill, V. Carpenter, and F. Montaseri, "Digital forensic investigation of social media, acquisition and analysis of digital evidence," *International Journal of Strategic Engineering*, vol. 2, no. 1, pp. 52–60, 2019.

[39] M. Yip, N. Shadbolt, and C. Webber, "Structural analysis of online criminal social networks," in *Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics*, pp. 60–65, Arlington, VA, USA, June 2012.

[40] Z. Zhao, C. Li, X. Zhang, F. Chiclana, and E. H. Viedma, "An incremental method to detect communities in dynamic evolving social networks," *Knowledge-Based Systems*, vol. 163, pp. 404–415, 2019.

[41] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[42] Y. Fang, J. Gao, Z. Liu, and C. Huang, "Detecting cyber threat event from twitter using IDCNN and BILSTM," *Applied Sciences*, vol. 10, no. 17, p. 5922, 2020.

[43] P. Lu, L. Zhang, X. Liu, J. Yao, and Z. Zhu, "Highly efficient data migration and backup for big data applications in elastic optical inter-data-center networks," *IEEE Network*, vol. 29, no. 5, pp. 36–42, 2015.

[44] O. Sharif, M. Hoque, A. Kayes, R. Nowrozy, and I. Sarker, "Detecting suspicious texts using machine learning techniques," *Applied Sciences*, vol. 10, no. 18, p. 6527, 2020.

[45] K. C. Roy, M. Cebrian, and S. Hasan, "Quantifying human mobility resilience to extreme events using geo-located social media data," *EPJ Data Science*, vol. 8, no. 1, p. 18, 2019.

[46] A. Squicciarini, S. Karumanchi, D. Lin, and N. Desisto, "Identifying hidden social circles for advanced privacy configuration," *Computers & Security*, vol. 41, pp. 40–51, 2014.

[47] L. Chen, F. W. Crawford, and A. Karbasi, "Seeing the unseen network: Inferring hidden social ties from respondent-driven sampling," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 1174–1180, Phoenix, AZ, USA, February 2016.

[48] J. Leskovec and A. Krevl, "SNAP datasets: stanford large network dataset collection," 2014, http://snap.stanford.edu/data.

[49] X. Zheng, Y. Wang, and M. A. Orgun, "Contextual subnetwork extraction in contextual social networks," in *Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA*, pp. 119–126, Helsinki, Finland, August 2015.

[50] A. Madani and M. Marjan, "Mining social networks to discover ego sub-networks," in *Proceedings of the 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, pp. 1–5, Muscat, Oman, March 2016.

Hindawi
Scientific Programming
Volume 2020, Article ID 8844083, 11 pages
https://doi.org/10.1155/2020/8844083

Research Article

Particle Swarm Optimization in the Presence of Malicious Users in Cognitive IoT Networks with Data

Noor Gul,[1] Muhammad Sajjad Khan,[1,2] Su Min Kim,[2] Marc St-Hilaire,[3] Ihsan Ullah,[4] and Junsu Kim[2]

[1]*Department of Electrical Engineering, Faculty of Engineering and Technology, International Islamic University, Islamabad 44000, Pakistan*
[2]*Department of Electronics Engineering, Korea Polytechnic University, 237 Sangidaehak-ro, Siheung-si, Gyeonggi-do 15073, Republic of Korea*
[3]*School of Information Technology and Department of Systems and Computer Engineering Carleton University, Ottawa, ON, Canada*
[4]*Department of Computer Science Engineering, Korea University of Technology, Cheonan, Republic of Korea*

Correspondence should be addressed to Junsu Kim; junsukim@kpu.ac.kr

Received 28 September 2020; Revised 28 October 2020; Accepted 30 October 2020; Published 11 November 2020

Academic Editor: Habib Ullah Khan

With the increasing applications in the domains of ubiquitous and context-aware computing, Internet of Things (IoT) is gaining importance. The study to efficiently exploit and manage a spectrum resources for industrial IoT (IIoT) applications is currently in the interest of research community. As increasing number of IIoT devices is heading towards the future-connected society with the cost of high system complexity, to meet the growing demands of wireless communication in future, cognitive IoT (CIoT) technology is considered as a choice. Reliable detection of the vacant spectrum holes is a vital task in the CIoT network with data. However, the performance of spectrum sensing severely degraded with the existence of malicious users (MUs) which falsifies the sensing results by reporting false data to the fusion center (FC). In this paper, we focus on the use of particle swarm optimization (PSO) to safeguard the cooperative spectrum sensing (CSS) from the negative effects caused by the MUs. The effectiveness of the proposed scheme is verified numerically in various scenarios with different types of MUs through analysis and simulations.

1. Introduction

Wireless communication networks have gained tremendous progress in the last decade to meet the growth in application devices from 1G to 4G Long-Term Evolution (LTE) advanced wireless networks [1]. These generations have played their roles in order to achieve improved data rate, high reliability, minimum latency and more things on the way. Wireless communication is facing the challenge of how to connect wireless devices with each other at anytime and anywhere. In the evolution process, 5G is expected to offer significant contributions towards spectrum management, public safety consideration, energy utilization efficiency, improved data rate, and low latency [2–4]. Since the 5G wireless communication technology is on the horizon in combination with IoT consideration as its center stage, IoT devices will perform a central role in the formation of a 5G network paradigm [5].

The term IoT introduced and mentioned by Ashton for the first time is a technological revolution that brings heterogeneous networks under the common umbrella of IoT [6]. This technological revolution represents the future of connectivity and reachability. Unlike the traditional networks of embedded systems, IoT is capable of interconnecting heterogeneous devices, having diverse functionalities, produced by different manufacturers [7]. IoT has changed the landscape of numerous industries tremendously ever since it has been introduced [8]. It is a promising subject of the social, technical, and economic implications that will hold a strong and meaningful impact

on our daily life in the near future. IoT is going to help in the improvised logistic learning, automation, e-health-care units, and intelligent transportation systems [8, 9]. The functionality of IoT is extended using mobile computing in the healthcare environment to bring massive healthcare in the form of mobile healthcare in [10]. Similarly, the fog-based IoT healthcare framework proposed to minimize energy consumption of the fog sensing nodes along with network delay in [11]. The major focus is this paradigm from a technological perspective is to enhance computation, communication, and connectivity procedures. However, connectivity and radio spectrum management are more crucial and challenging responsibilities in front of the research community. In the near future, over 50 billion wireless devices have to be interconnected that may demand for a large number of spectrum resources [12]. In [13], the authors argued on the importance and employment of cognitive capabilities in IoT with the objective that without implication of the cognitive capability, it is similar to an awkward stegosaurus with all brawn and no brains. A dozen of wireless communication technologies are already in use such as WiFi, Bluetooth, LTE, earlier 3G standards, ZigBee, Near Field Communication (NFC), and different satellite services. Therefore, the rapid growth in wireless communications demands for the new wireless services in both the used and unused part of the radio spectrum [14]. Spectrum sharing in 5.4 GHz band has already been legalized by the federal communication commission (FCC), where devices sense the existence of military radars before accessing the channel [15]. Cognitive radio (CR) is an intelligent wireless communication technology with efficient radio spectrum utilization abilities trying to learn and adjust its internal states according to environment [16]. The primary users (PUs) are able to transmit any time with no restrictions while the secondary users (SUs) gain the benefit of the spectrum access only when it is declared to be free [17].

In CR networks (CRNs), an incorrect detection of the PU results in false alarm and reduces the SUs' opportunities to access the spectrum. Similarly, any misdetection of the occupied PU channel produces interference to the PU by the secondary access. Masking of the optimal interference subcarrier is obtained based on genetic algorithm (GA) to suppress intercarrier interference caused by the SUs to the PU channel [18]. In [19], a side-lobe reduction scheme using a generalized side-lobe canceller combined with GA and differential evolution is proposed to minimize the impact of the interference.

Spectrum sensing with single SU is facing a number of limitations such as the limitations with the energy constraints, shadowing, fading, and hidden terminal problems [20]. On the other hand, in the cooperative spectrum sensing (CSS), the sensing problems faced by the single user is mitigated by allowing the cooperation among the multiple SUs to share their sensing results to make a global decision on the existence of PU [21]. In the CSS, the SUs forward their local decisions to the fusion center (FC) to make a global decision to infer the absence or presence of the PU [21].

However, the existence of malicious users (MUs) in the CSS severely reduces effectiveness of cooperation. Therefore, proper detection and exclusion of the MUs' information are extremely critical [22]. Significant investigations have been carried out to make the CSS robust to the attack of MUs. The MU transfers erroneous sensing reports to the FC, in order to create confusion about the spectrum conditions. Such attacks are referred to as spectrum sensing data falsification (SSDF) attack [23]. A systematic review is conducted in [24] to analyze the security problem of IoT devices and to counter various security challenges using mobile computing.

Boosted trees algorithm (BTA) is proposed in [25] that uses the AdaBoost ensemble method to make results of the cooperative decision at the FC reliablity in the presence of abnormal sensing data. The work in [26] suggested the use of differential evolution (DE) to identify the weighting coefficient vector against user sensing reports. This strengthens the reports of normal sensing users with high weights compared to the abnormal sensing users. An enhanced CSS scheme is determined at the FC using flower pollination algorithm (FPA) in [27]. Similarly, performance comparison is made at the FC between different hard combination schemes in the presence of abnormal reports of the lazy MUs in [28]. The work in [29] reduces the effect of the false sensing reports before making the final decision at the FC using modified double-sided neighbor distance algorithm with a GA optimization scheme. A machine learning scheme such as support vector machine (SVM) in [30] effectively classified the normal sensing users and different categories of MUs to help FC decision. In [31], malicious sensing nodes with false sensing reports are quantified in the simulation environment of the Poisson point process. As the MUs do not share honest sensing reports with the FC, a contract theory approach with incentive design scheme is proposed in [32] to reward honest SUs and to strengthen their cooperation. The normal SUs discussed in [33] follow the FC recommendation as a final decision of the PU channel and use their local sensing decisions to guarantee the CSS reliability. A Bayesian-inference scheme is proposed in [34] to identify and countereffects of the individual and collaborative SSDF attackers using a sliding window trust model.

A robust scheme which deals with always yes malicious users (AYMUs) is implemented in [35]. An extended sequential cooperation scheme with reduced sensing reports and improved sensing performance is investigated in [36]. The soft fusion schemes such as maximum gain combining (MGC) and equal gain combining (EGC) combine the energy statistics reported from the SUs to make a decision [37–39]. All cooperating SUs in the hard decision scheme forward the binary values denoting the local decisions to the FC to make a global decision [40–42]. The works in [43, 44] utilize GA for optimizing the detection and false alarm probabilities to minimize the error probability. A novel evolution-based CSS mechanism is discussed in [45] to select and optimize the weight coefficients of the SUs' sensing result. The binary GA- (BGA-) based soft fusion scheme proposed in [46] is used to improve the detection performance and bandwidth utilization. Particle swarm optimization (PSO) is utilized as a tool for the optimization of the threshold point to enhance the spectral efficiency and detect the potential spectrum [47, 48]. An energy-efficient PSO

which provides high protection to the legitimate user is proposed in [49]. In our previous study in [50], FC determines the Kullback–Leibler (KL) divergence score based on the users' soft energy sensing reports. The KL divergence score is acknowledged to the users and stored in the FC local database to improve future decision. Similarly, in the proposed method in [51, 52], a double-sided neighbor distance (DSND) and outlier detection schemes followed by the majority voting decision in GA is used to reduce global decision error probability of the centralized CSS.

In this paper, the PSO algorithm has been employed to search for the spectrum information representing the actual status of the PU's activity on behalf of all cooperative SUs. The spectrum selection of the PSO results in overcoming the effect of MUs in the CSS. In the proposed scheme, the SUs forward their sensing results to the FC at certain sensing intervals. The FC utilizes the PSO algorithm to determine the most suitable energy statistics among the information received from the SUs including the MUs. Please note that the MUs pretend to be normal SUs. The one-to-many Hamming distances and z-score is used as a composite outlier score and fitness function of the PSO algorithm. Out of the PSO population, the sensing report with minimum outlying is selected as the PU channel status on behalf of all SUs for a global decision. The global decision of the PU channel is made with EGC, MGC, and majority voting hard fusion schemes based on the selections of the PSO algorithm. The PSO algorithm selection contains less harmful effects from any MUs; thus, the FC's decision becomes more reliable which improves the overall CSS performance.

The proposed scheme is verified in the false sensing of always no MU (ANMUs), AYMU, opposite MU (OMU), and random opposite MU (ROMU) in a cooperative environment. The AYMU sends an always high-energy statistics of the channel irrespective of the actual status; thus, it increases false alarm probability and reduces throughput of the SUs. The ANMU category of the MU forwards always low-energy statistics that result in misdetection and induce interference to the PU. The OMU, which is the most harmful type of the MU, forwards the opposite values of energy statistics against its actual sensing result. Finally, the ROMU acts like the OMU with probability P and like normal SU with probability $1 - P$.

The rest of the paper is organized as follows. In Section 2, the system model considered through this paper is presented. Section 3 describes the details that how the PSO algorithm is utilized to reduce the effects of abnormal SUs by identifying accurate sensing results before using EGC, MGC, and majority voting decision at the FC. Numerical evaluations and analysis are presented in Section 4. Finally, Section 5 concludes the paper.

2. System Model

As the probability of experiencing deep fading at all SUs is extremely low, the shared sensing results of the users to cooperatively decide the PU activity can reduce the sensing problems which may occur with single SU's sensing.

The objective is to minimize the error probability $P_e = P_f + P_m$, where P_f and P_m denote the false alarm and misdetection probabilities. Therefore, in order to reduce P_e, the detrimental effects of the misdetection $P_m = 1 - P_d$ and false alarm P_f probabilities must be minimized.

As in Figure 1, the SUs cooperate to sense the activity of the PU channel and inform the FC about their sensing information. The received information from the AYMU is an always high-energy signal representing busy status of the channel. Similarly, the ANMU provides with a low-energy signal to the FC. The OMU negates the actually sensed status of the PU and the ROMU acts like an OMU or a normal SU probabilistically. Hence, the ROMU's nature is more difficult to predict. Based on the received reports from the SUs, the FC makes global decision of the channel availability.

The binary hypothesis test at the lth time slot with the jth SU received signal is as follows [35]:

$$y_j(l) = \begin{cases} H_0, & n_j(l), \\ H_1, & h_j s(l) + n_j(l), \end{cases} \tag{1}$$

where the hypotheses H_0 denotes the idle status of the PU channel and H_1 represents the channel occupation by the PU, $y_j(l)$ is the received signal by the jth SU in the lth time slot, $n_j(l)$ is the additive white Gaussian noise (AWGN) at j^{th} SU, h_j is the channel gain between the PU channel and the jth SU, and $s(l)$ is the signal transmitted by the PU in the lth time slot, respectively.

The received signal energy of the PU channel by the jth SU at the ith sensing interval is

$$E_j(i) = \begin{cases} \sum_{l=l_i}^{l_i+K-1} \left|n_j(l)\right|^2, & H_0, \\ \sum_{l=l_i}^{l_i+K-1} \left|h_j s(l) + n_j(l)\right|^2, & H_1, \end{cases} \tag{2}$$

where K is the number of samples in the ith sensing interval. According to the central limit theorem, the number of samples needs to be large enough so that the energy reported by each SU becomes similar to a Gaussian random variable under both H_0 and H_1 as [35, 53].

$$E_j \sim \begin{cases} N\left(\mu_0 = K,\ \sigma_0^2 = 2K\right), & H_0, \\ N\left(\mu_1 = K\left(\eta_j + 1\right),\ \sigma_1^2 = 2K\left(\eta_j + 1\right)\right), & H_1. \end{cases} \tag{3}$$

In (3), η_j is the signal to noise ratio (SNR) between the PU and j^{th} SU. Similarly, (μ_0, σ_0^2) and (μ_1, σ_1^2) are the mean and variance values of the energies reported under the H_0 and H_1 hypotheses.

3. Proposed Particle Swarm Optimization Process at FC

PSO is derived from the bird flocking or fish swarming introduced by Eberhart and Kenedy in 1952 [54]. In PSO, individual intelligence, as well as collective intelligence, plays an important role in finding an enhanced solution. In GA, it

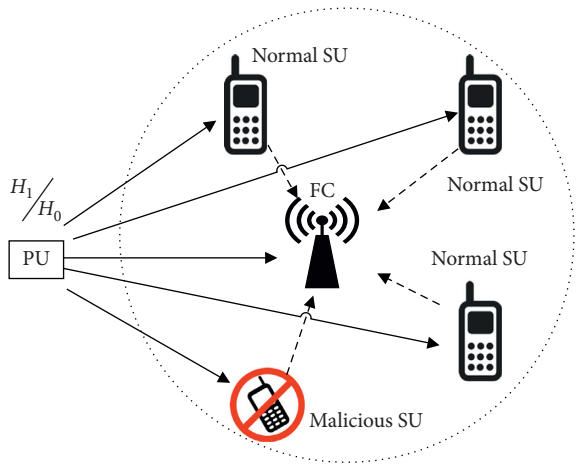


Figure 1: System model: CSS in CRN.

is likely that every novel group is flourishing better than the previous generations. Similarly, in the PSO, the same group is likely to become better and better. The individuals establish their intelligence and improve it with the passage of time. The whole group is expected to improve its group intelligence. Particles in the PSO algorithm utilize its own and neighbor knowledge to update their position and velocity. The PSO particle exchanges information about their best position among each other during a number of iterations.

The proposed model of the CSS using the PSO is shown in Figure 2. In this model, the SUs sense the PU channel and forward their energy statistics to the FC for a number of observations to form the PSO population. Then, the FC applies the PSO technique in identifying sensing reports which are closer to the actual status of the PU channel. The FC measures the fitness score under all sensing iterations and declares the minimum outlying particle as the actual channel information for a final decision. Fusion schemes are applied by the FC, based on the selected global best particle of the population to generate a more accurate and reliable final decision of the PU channel.

In the PSO algorithm, a particle represents a row of the population matrix and each particle element (soft energy report) has a certain position and velocity. Initially, we assume that the positions and velocities of the particles are set to zero. The overall process of the PSO algorithm to determine the sensing reports, on the basis of which global decision is taken by the FC, is by proceeding the following steps:

Step 1: local spectrum decisions

The FC receives the soft energy reports from the SUs to form a history reporting matrix consisting soft energy statistical observations in the N_0 sensing intervals representing all SUs such as

$$E = \left[E_{ij}\right] = \begin{bmatrix} E_{11} & E_{12} & \cdots & E_{1M} \\ E_{21} & E_{22} & \cdots & E_{2M} \\ E_{31} & E_{32} & \cdots & E_{3M} \\ \vdots & \vdots & \ddots & \vdots \\ E_{N_0 1} & E_{N_0 2} & \cdots & E_{N_0 M} \end{bmatrix}, \quad i \in 1, 2, \ldots, N_0,\ j \in 1, 2, \ldots, M, \tag{4}$$

where E_{ij} denotes the energy information of the jth SU in the ith sensing interval. Spectrum sensing information is gathered at the FC database for the M SUs including the MUs in the N_0 intervals as in (3). The SSDF effect caused by the MUs can be minimized by utilizing the following steps.

Step 2: finding the fitness of the particles

After the collection of energy information as in (4), the FC modifies the particle positions to observe the differences in each individual sensing report with the reports provided by the other SUs. A new population is

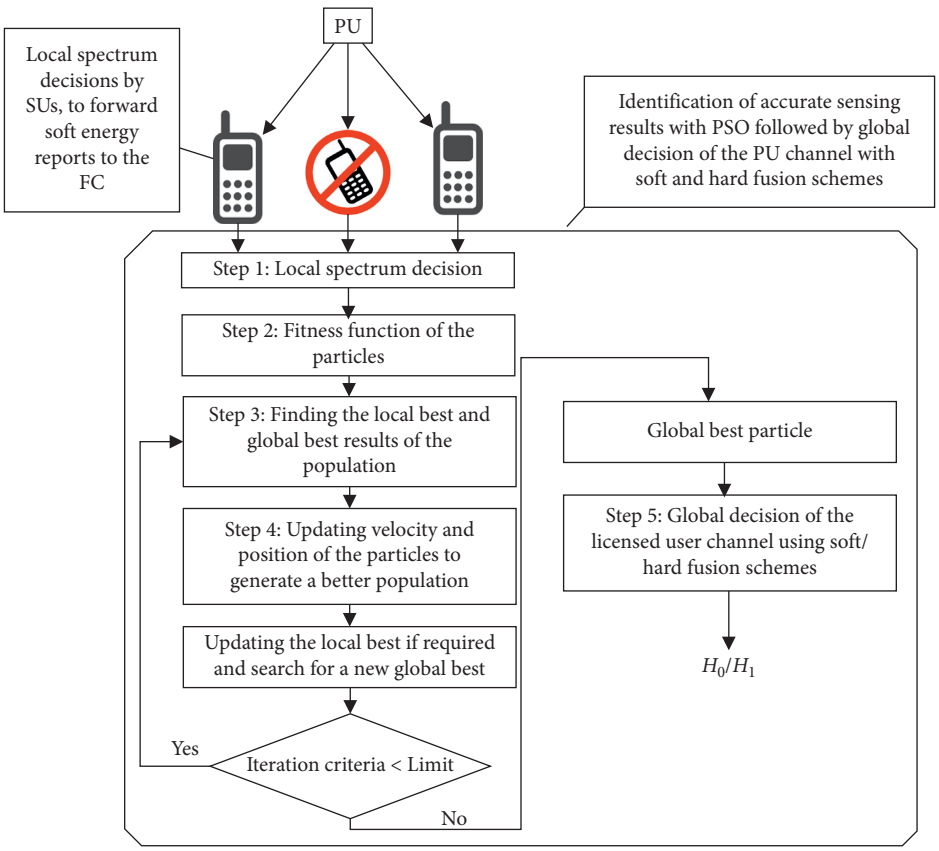


Figure 2: Proposed cooperative spectrum sensing scheme using PSO.

formed for all SUs based on the information already collected in (4) as

$$E' = \left[E'_{ij}\right] = \begin{bmatrix} E'_{11} & E'_{12} & \cdots & E'_{1M} \\ E'_{21} & E'_{22} & \cdots & E'_{2M} \\ E'_{31} & E'_{32} & \cdots & E'_{3M} \\ \vdots & \vdots & \ddots & \vdots \\ E'_{N_0 1} & E'_{N_0 2} & \cdots & E'_{N_0 M} \end{bmatrix}, \quad i \in 1, 2, \ldots, N_0,\ j \in 1, 2, \ldots, M, \tag{5}$$

where $E'_{ij} = |(\sum_{j=1}^{M} E_{ij} - E_{ij})/(M-1)|$, which denotes the average of individual soft energy reports provided by all other SUs while taking out the report of the jth user in this averaging.

Step 2.1: outlying using one-to-many sensing distances

Outlying factors are determined for the sensing reports from the SUs based on the one-to-many sensing distances $\mathbf{d}_j(i)$ for the jth SU in the ith sensing particle as

$$\mathbf{d}_j(i) = \left|E_{ij} - E'_{ij}\right|, \quad i \in 1. \ldots, N_0,\ j \in 1, \ldots, M. \tag{6}$$

Based on the results in (6), the outlier score $\mathbf{d}_j(i)$ of the normal SUs and MUs is added to discover the total one-to-many Hamming distance score under each sensing interval:

$$\mathbf{d}_i = \sum_{j=1}^{M} \left(\mathbf{d}_j(i)\right), \quad i \in 1, \ldots, N_o,\ j \in 1, \ldots, M, \tag{7}$$

where $\mathbf{d}_i$ is the total outlier score representing absolute sum of the Hamming distances of the individual reports E_{ij} with the average reports E'_{ij} of all other SUs in the ith sensing interval.

The measurement in (7) is made for the N_0 intervals, and the results are collected as

$$\mathbf{d} = \left[\, \mathbf{d}_1 \;\; \mathbf{d}_2 \;\; \mathbf{d}_3 \;\; \cdots \;\; \mathbf{d}_{N_0} \,\right]^T, \tag{8}$$

where $\mathbf{d}$ is the outlier score results of the N_0 sensing intervals. This score is a measurement of how far the report of each SU is away from the average sensing reports provided by all other SUs by separating those sensing intervals during which the MUs and the imperfection of the normal SU were misguiding the FC final decision about the PU channel.

Step 2.2: outlying using z-score

Similarly, the other outlier score measurement is made with the z-score measurement in comparison with the sensing reports received from each SU as

$$\mathbf{o}_j(i) = \left| \frac{\left(E_{ij} - \mu(i)\right)}{\sigma(i)} \right|, \quad i \in 1, \ldots, N_0,\ j \in 1, \ldots, M, \tag{9}$$

where $\mu(i) = (\sum_{j=1}^{M} E_{ij})/M$ is the mean and $\sigma(i)$ is the standard deviation of the ith particle in the PSO population. $\mathbf{o}_j(i)$ is the z-score outlying of the jth report in the ith interval of the history log. The result of $\mathbf{o}_j(i)$ in (9) shows how much the local sensing observation of the jth user is detached away from the group observations provided by all other SUs.

Now, for guaranteeing the authenticity of each of the ith reports, the sum of z-score measurements for all particles is made as

$$\mathbf{o}_i = \sum_{j=1}^{M} \left(\mathbf{o}_j(i)\right), \quad i \in 1, \ldots, N_o, j \in 1, \ldots, M. \tag{10}$$

The total z-score of the N_0 particles of PSO population is collected as

$$\mathbf{o} = \left[\, \mathbf{o}_1 \;\; \mathbf{o}_2 \;\; \mathbf{o}_3 \;\; \cdots \;\; \mathbf{o}_{N_0} \,\right]^T. \tag{11}$$

As the fitness function is the representation for the suitability of each sensing reports, the final selection of the fitness of each sensing reports from both the normal SUs and MUs is determined, and the best selection of the sensing results having less abnormal behavior is calculated.

The criteria for selection of the particles according to their fitness values are declared according to (6) and (9) as

$$\mathbf{f}(i) = \mathbf{d}_i + \mathbf{o}_i. \tag{12}$$

The result in (12) declares the minimum score for sensing reports with fewer abnormalities in comparison to those that are badly affected due to the abnormal behavior of the MUs.

Step 3: updating population

The global best position $\mathbf{g}$ is the particle that results in minimum outlying score among all particles in $\mathbf{E}$ according to (12). Each particle may improve its own if its new version is better compared to the previous one. Local best particles of the population are selected as $P = E$.

The positions and velocities are initially set to zero. The particle velocities are updated with the individual and collective intelligences as

$$\begin{aligned} V_{(i+1)j} = V_{ij} &+ C_1 \times R_1 \times \left(P_{ij} - E_{ij}\right) \\ &+ C_2 \times R_2 \times \left(\mathbf{g}_j - E_{ij}\right), \end{aligned} \tag{13}$$

where C_1 and C_2 are the learning acceleration coefficients that describe the particles' individual and social contributions. Similarly,R_1 and R_2 are the uniformly distributed random numbers in the range 0 to 1 to present stochastic contribution to the algorithm.

Next to the measurements of particles' velocities with the local and global intelligence, these velocities are rounded to the two extremes as

$$V_{(i+1)j} = \begin{cases} \max(V), & V_{ij} \succ \max(V), \\ \min(V), & V_{ij} \prec \max(V). \end{cases} \tag{14}$$

The jth particle's position representing soft energy information at the $(i+1)^{\text{th}}$ iteration is updated with the measured velocities as

$$E_{(i+1)j} = E_{ij} + V_{(i+1)j}, \tag{15}$$

where $E_{(i+1)j}$ are the reports of the modified population, E_{ij} is the initial report of the jth SU in the ith interval, and $V_{(i+1)j}$ are the velocities as in (14).

Step 4: updating local best and global best

Fitness measurements of the new population in (15) are determined by following the same procedure as in (12).

The novel particle fitness is compared with the earlier population fitness to search for any improvements in the local and global best positions in comparison with the earlier energy reports. Similarly, the local best positions of the population are updated as

$$P_i = \begin{cases} E_i, & f(E_i) \prec f(P_i), \\ P_i, & \text{otherwise}, \end{cases} \quad i \in 1, \ldots, N_0. \tag{16}$$

In (16), the results of the local best particles are updated by comparing the fitness of the new population (15) to that of the local best particles P fitness. The local best particles are updated and take values of the new particles if their outlying results in (12) are higher compared to the newly created population.

Similarly, a search is made to identify new global best particle in the entire population by cross analysis of the fittest. Fitness of the updated local best particles in (16) is placed for comparison to search for any improvement in the selection of the global best particle as

$$\mathbf{g} = \begin{cases} P_i, & f(P_i) \prec f(\mathbf{g}), \\ \mathbf{g}, & \text{otherwise}, \\ \forall i \in 1, \ldots, N_0. \end{cases} \tag{17}$$

In (17), outlying score of each particle of the local best population is compared with the global best particle determined earlier. If any particle of the local best population has a fitness function found to be optimum in comparison with the global best particle with the minimum outlying score in (12), then the global best particle is replaced.

Here, the new global best particle is selected as $\mathbf{g}$ representing the particle with the best fitness function having minimum outlying results in the current and previous PSO population.

The PSO production of the new population and search for the global best results continues until the stopping criterion is met. At the end of the desired number of iterations, the final global best particle containing reliable and trusted soft energy reports against M cooperating SUs is elected for a final decision by the FC.

Step 5: global decision combination schemes

Based on the final selection of the global best particle $\mathbf{g}$ as the soft energy reports on behalf of all M cooperative SUs, FC utilizes soft and hard combination schemes in Section 2 for declaring a unanimous decision about the PU channel. The EGC, MGC, and majority voting hard fusion combination schemes are used as decision criteria in this section.

The EGC is combining the individual statistical information of all SUs by giving equal weight to each individual SU decision and summed coherently. The combination is compared with the threshold by the EGC as

$$\text{EGC} = \begin{cases} H_1, & \dfrac{\left(\sum_{j=1}^{M} \mathbf{g}_j\right)}{M} \geq \gamma, \\ H_0, & \text{otherwise}. \end{cases} \tag{18}$$

The cooperative detection and false alarm probabilities P_{d_EGC} and P_{f_EGC} made by the EGC scheme based on the global decision made about the PU spectrum are

$$\begin{aligned} P_{d\_\text{EGC}} &= \Pr\left\{\frac{\left(\sum_{j=1}^{M} \mathbf{g}_j\right)}{M} \geq \gamma | H_1\right\}, \\ P_{f\_\text{EGC}} &= \Pr\left\{\frac{\left(\sum_{j=1}^{M} \mathbf{g}_j\right)}{M} \geq \gamma | H_0\right\}. \end{aligned} \tag{19}$$

In the MGC scheme, each receiving signal branch is multiplied with a weighed function proportional to the branch gain. The branches with a strong signal in the MGC are amplified more, while the weak signal components receive attenuations with the weights. The idea to boost the strong signal component and attenuate weak signal component in the MGC diversity is exactly the same as that of filtering and signal weighting in the matched filter receiver. Similarly, the MGC scheme at the FC is giving higher weights to the decision of the SUs with higher SNR values and low weights to the decisions of the SUs with low SNR values as

$$\text{MGC} = \left\{ H_1, \quad \sum_{j=1}^{M} \left(w_j \times \mathbf{g}_j\right) \geq \gamma, H_0, \text{otherwise}, \right. \tag{20}$$

where $w_j = \eta(j) / \sum_{j=1}^{M} \eta(j)$. The cooperative detection and false alarm probabilities of the MGC scheme are measured based on the received soft energy statistics as

$$\begin{aligned} P_{d\_\text{MGC}} &= \left\{\left(\sum_{j=1}^{M}\left(w_j \times \mathbf{g}_j\right) \geq \gamma\right) | H_1\right\}, \\ P_{f\_\text{MGC}} &= \left\{\left(\sum_{j=1}^{M}\left(w_j \times \mathbf{g}_j\right) \geq \gamma\right) | H_0\right\}. \end{aligned} \tag{21}$$

In the majority voting schemes, the FC counts the total number of the SUs with their energy value greater than the threshold as

$$\text{MV} = \left\{ H_1, \quad \sum_{j=1}^{M} \left(\mathbf{g}_j \geq \gamma_j\right) \geq k, H_0, \text{otherwise}. \right. \tag{22}$$

The three commonly used hard combination schemes are the majority voting, OR, and AND fusion combination schemes. In the count hard decision, a global decision on the PU existence is made if k out of total M cooperative users provide PU detection information with their energies larger than a threshold. The FC concludes a final decision H_1 if k users' reports validate the PU existence. Similarly, the total number of cooperative users with PU detection information less than k lead the FC to conclude in favor of H_0 to declare an idle condition of the PU channel. The counting score k is taken as 1 for the OR fusion rule and M for the AND rule. In the proposed work, the majority voting scheme is selected with $k = M/2$. In case of majority voting, if half cooperative

SUs energies are passing the threshold, a global decision is made as H_1; otherwise, the decision is made in favor of H_0.

The detection and false alarm probabilities measurement of the majority voting hard decision schemes based on the best selection of the PSO at the FC are as follows:

$$P_{d\_\mathrm{MV}} = \Pr\left\{\sum_{j=1}^{M} \mathbf{g}_j \geq \frac{M}{2} | H_1\right\},$$
$$P_{f\_\mathrm{MV}} = \Pr\left\{\sum_{j=1}^{M} \mathbf{g}_j \geq \frac{M}{2} | H_0\right\}, \quad (23)$$

where P_{d_MV} and P_{f_MV} are cooperative detection and false alarm probabilities of the majority voting schemes when PSO is used as a detection mechanism at the FC.

4. Numerical Evaluation

For simulation purposes, parameter adjustment is made for the CRN with $M = 11$ SUs. Among the total SUs, 7 SUs are selected as normal and 4 SUs are randomly selected as AYMUs, ANMUs, OMUs, and ROMUs. The sensing time is kept as 1 ms which contains $K = 270$ sensing samples. The total number of sensing iterations N is selected as 100. The sensing interval in which the ROMU performs a malicious act is adjusted randomly from 1 to N. The system performance is verified in the presence of equal distributions of OMU, ROMU, AYMU, and ANMU users. The sensing reports of the SUs formed the PSO population of size $N_0 \times M$ with N_0 particles representing the sensing information of the M cooperating SUs.

In this part of the simulation, the MUs are first selected as AYMU, and then, its nature is changed to ANMU. In Figure 3, results are drawn to compare the performances of EGC, MGC, and majority voting schemes. From the simulation results in Figure 3, it is obvious to show an improvement in the detection results of the PSO-based EGC, MGC, and majority voting schemes against the conventional combination schemes. The cooperative scheme performance under the considerations of AYMU and ANMU is more optimized for the PSO based soft and hard combinations. It is shown that the detection response in both the cases when only AYMU and the one with only ANMU users' considerations are identical. The equal consideration of AYMU and ANMU cases are similarly treated in the CSS with almost identical probability of detection P_d for a given false alarm P_f. Figure 3 also shows better receiver operating characteristics (ROC) results for the PSO based-MGC scheme which is followed by the EGC scheme. The majority voting hard fusion combination illustrates minimum detection results compared with the other two schemes. It is also obvious that the PSO-based soft and hard fusion combination schemes are able to outperform the simple MGC, EGC, and hard fusion combinations for any given false alarm.

In the second part, authenticity of the system is verified by comparing results of the proposed PSO-based soft and hard combinations with conventional schemes. In this case,

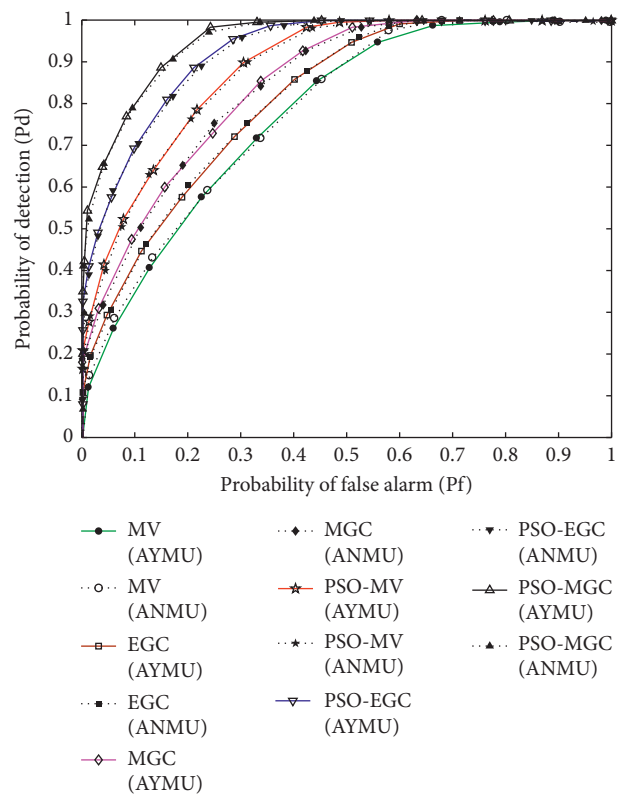


Figure 3: ROC curve, when AYMU and ANMU exist in the network.

the MUs are first selected as OMU, and then, their nature changes to ROMU. The result illustrates that the MGC scheme shows better detection compared with the EGC and majority voting counterparts. The ROC collection of the three schemes under the proposed and the conventional schemes show the reliabilities of the PSO-based combination techniques. In Figure 4, the ROMUs affect the sensing environment more hazardously than the OMU. The proposed scheme is superseding the traditional fusion schemes in both the OMU and ROMU cases.

In the third part of the simulation, the performances of the conventional and the proposed PSO-based fusion combinations schemes are tested, when the MUs are distributed equally as AYMU, ANMU, OMU, and ROMU in Figure 5.

The minimum ROC results in Figure 5 show the performance of the conventional fusion schemes under the consideration of all 4 MUs, while the upper three ROC curves show the PSO fusion combination scheme performance under the same parameter settings. This shows an improvement in the detection performance of the PSO-based fusion combination schemes compared with the conventional combination schemes. It is noticeable that the MGC fusion combination scheme provides more sophisticated detection performance compared to the other schemes.

The proposed PSO-based fusion combination scheme is further verified by illustrating the error probability P_e according to the SNR varying from −35 dB to 0 dB in

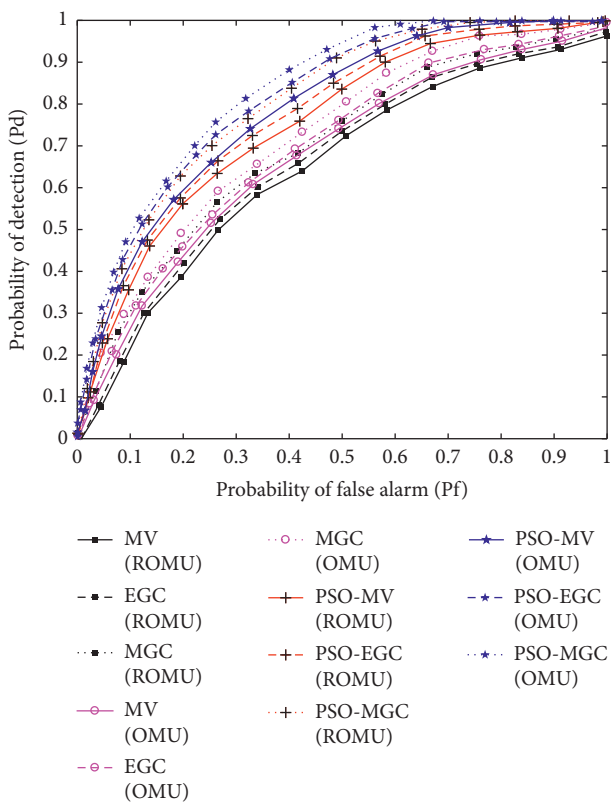


Figure 4: ROC curve, when OMU and ROMU exist in the network.

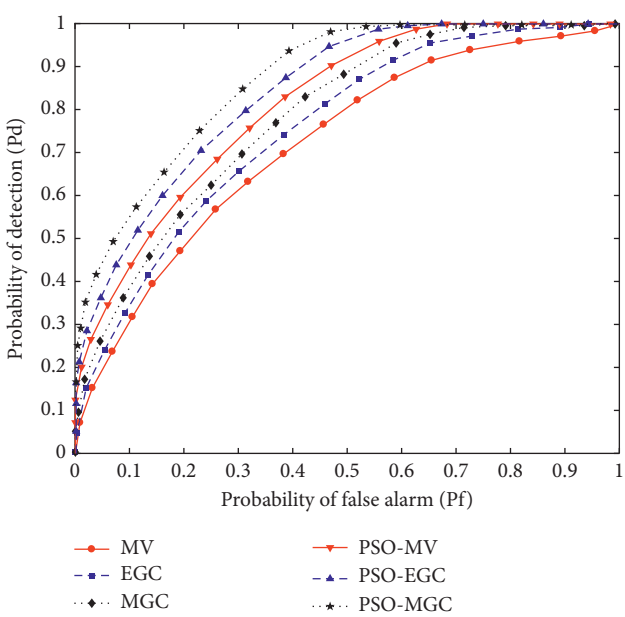


Figure 5: ROC curve, when AYMU, ANMU, OMU, and ROMU are equally distributed.

Figure 6. The error in sensing the PU channel by the proposed scheme is the minimum of all and with increasing

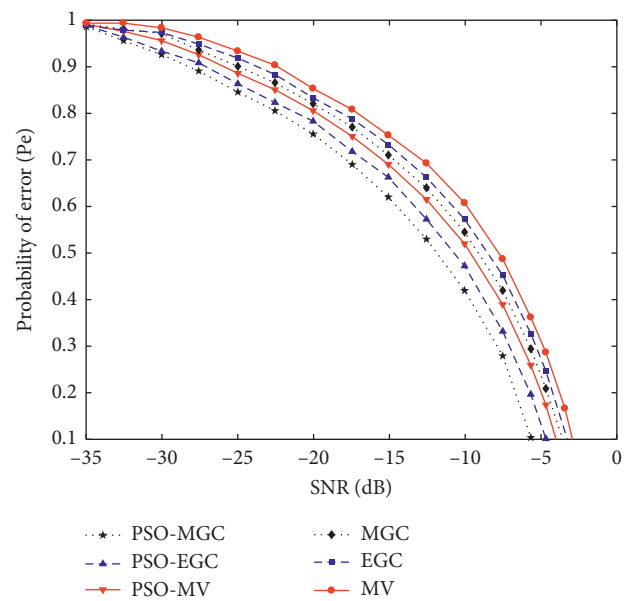


Figure 6: Probability of error vs. SNR.

SNR the proposed scheme error reduces more quickly as compared with the other traditional schemes.

It is clear that, by following the PSO-based algorithm, the proposed fusion combination schemes are more optimized and accurate in the presence of the MUs. The selections of the PSO following by the soft and hard fusion combinations make the CSS authentic and suitable against the MUs. The risk of considering the MUs in the CSS is significantly reduced with the proposed scheme. The results show that the SUs cooperation is more effective using the proposed scheme. The proposed scheme is able to eliminate the considerations of MUs in making global decision at the FC and produce reliable sensing results.

5. Conclusions

The impact of the MUs on the CSS reduces the effectiveness of cooperation in CIoT. Therefore, it is necessary to detect the MUs in order to avoid any confusion about the actual status of the PU channel. This paper focuses on improving the performance of the CSS by using the PSO algorithm. Based on the energy statistics reported by the SUs, the PSO is able to reduce the effect of the MUs in authenticating the global decision of the PU's existence. The FC combines the diversified sensing reports of the users using the proposed EGC, MGC, and majority voting decisions to acquire the global decision of the PU activity. The proposed PSO algorithm is able to overcome the effects of OMU, ROMU, AYMU, and ANMU categories of the MUs in the soft and hard combinations. Simulations verify the superiority and the authenticity of the proposed scheme in producing more accurate and reliable decisions for the soft and hard combination schemes at the FC.

Data Availability

The data used to support the finding of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2018-0-01426) supervised by the IITP (Institute for Information and Communication Technology Planning, and Evaluation) and in part by the National Research Foundation (NRF) funded by the Korea government (MSIT) (No. 2019R1F1A1059125).

References

[1] A. Agarwal, G. Mishra, and K. Agarwal, "The 5th generation mobile networks-key concepts, networks, architecture and challenges," *American Journal of Electrical & Electronics Engineering*, vol. 3, no. 2, pp. 22–28, 2015.

[2] T. Q. Duong and N. S. Vo, "Wireless communication and network for 5G and beyond," *Mobile Networks and Applications*, vol. 24, no. 2, pp. 443–446, 2019.

[3] B.-S. P. Lin, F. J. Lin, and L. P. Tung, "The role of 5G mobile broadband in the development of iot, big data, cloud and SDN," *Communication and Network*, vol. 8, no. 1, pp. 9–21, 2016.

[4] R. Chavez-Santiago, M. Szydeko, A. Kliks et al., "5G: the convergence of wireless communication," *Wireless Personal Communications*, vol. 83, no. 3, pp. 1617–1642, 2015.

[5] W. Ejaz, A. Anpalagan, M. A. Imran et al., "Internet of things (IoT) in 5G wireless communications," *IEEE Access*, vol. 4, pp. 10310–10314, 2016.

[6] K. Ashton, "Internet of things," in *The Real World, Things Matter More than Ideas*Springer, Berlin, Germany, 2009.

[7] F. A. Awin, Y. M. Alginahi, E. Abdel-Raheem, and K. Tepe, "Technical issues on cognitive radio-based internet of things systems: a survey," *IEEE Access*, vol. 7, pp. 97887–97908, 2019.

[8] R. Khan, S. U. Khan, R. Zaheer, and S. Khan, "Future internet: the internet of things architecture, possible applications and key challenges," in *Proceedings of the IEEE International Conference on Frontiers of Information Technology*, Islamabad, Pakistan, December 2012.

[9] A. A. Khan, M. H. Rehmani, and A. Rachedi, "When cognitive radio meets the internet of things," in *Proceedings of the IEEE International Wireless Communication & Computing Conference (IWCMC)*, Paphos, Cyprus, September 2016.

[10] S. Nazir, Y. Ali, N. Ullah, and I. García-Magariño, "Internet of things for healthcare using effects of mobile computing: a systematic literature review," *Wireless Communications and Mobile Computing*, vol. 2019, Article ID 5931315, 15 pages, 2019.

[11] C. Feng, M. Adnan, A. Ahmad, A. Ullah, H. U. Khan, and H. U. Khan, "Towards energy-efficient framework for iot big data healthcare solutions," *Scientific Programming*, vol. 2020, Article ID 7063681, 9 pages, 2020.

[12] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: a survey on enabling technologies, protocols, and applications," *IEEE Communications Survey & Tutorial*, vol. 17, no. 4, pp. 2347–2376, 2015.

[13] Q. Wu, G. Ding, Y. Xu et al., "Cognitive internet of things: a new paradigm beyond connection," *IEEE Internet of Things Journal*, vol. 1, no. 2, pp. 129–143, 2014.

[14] A. Ghasemi and E. S. Sousa, "Spectrum sensing in cognitive radio networks: requirements, challenges and design tradeoffs," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 32–39, 2008.

[15] S. Mishra, A. Sahai, and R. Brodersen, "Cooperative sensing among cognitive radio," in *Proceedings of the IEEE International Conference on Communications*, Istanbul, Turkey, 2006.

[16] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, 2005.

[17] L. Zhai, H. Wang, and C. Gao, "A spectrum access based on quality of service (QoS) in cognitive radio networks," *PLoS One*, vol. 11, no. 5, pp. 2005–2009, 2016.

[18] L. Miao, Z. Sun, and Z. Jie, "The parallel algorithm based on genetic algorithm for improving the performance of cognitive radio," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 5986482, 6 pages, 2018.

[19] A. Elahi, I. M. Qureshi, F. Zaman, N. Gul, and T. Saleem, "Suppression of mutual interference in noncontiguous orthogonal frequency division multiplexing based cognitive radio systems," *Wireless Communications and Mobile Computing*, vol. 2017, Article ID 1860134, 9 pages, 2017.

[20] E. Axell, G. Leus, E. G. Larsson, and H. V. Poor, "Spectrum sensing for cognitive radio : state-of-the-art and recent advances," *IEEE Signal Processing Magazine*, vol. 29, no. 3, pp. 101–116, 2012.

[21] Y. He, J. Xue, T. Ratnarajah, M. Sallaturai, and F. Khan, "On the performance of cooperative spectrum sensing in random cognitive radio networks," *IEEE Systems Journal*, vol. 12, pp. 1–12, 2016.

[22] M. S. Khan and I. Koo, "Mitigation of adverse effect of malicious users by hausdorff distance in cognitive radio networks," *Journal of Information and Communication Convergence Engineering*, vol. 13, no. 2, pp. 74–80, 2015.

[23] H. Li and Z. Han, "Catch me if you can: an abnormality detection approach for collaborative spectrum sensing in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, 2010.

[24] B. Liao, Y. Ali, S. Nazir, L. He, and H. U. Khan, "Security analysis of iot devices by using mobile computing: a systematic literature review," *IEEE Access*, vol. 8, pp. 120331–120350, 2020.

[25] N. Gul, M. S. Khan, S. M. Kim, J. Kim, A. Elahi, and Z. Khalil, "Boosted trees algorithm as reliable spectrum sensing scheme in the presence of malicious users," *Electronics*, vol. 9, no. 6, pp. 1038–1123, 2020.

[26] N. Gul, I. M. Qureshi, M. S. Khan, A. Elahi, and S. Akbar, "Differential evolution based reliable cooperative spectrum sensing in the presence of malicious users," *Wireless Personal Communications*, vol. 114, no. 1, pp. 123–147, 2020.

[27] H. Asfandyar, N. Gul, I. Rasool, and A. Elahi, "Enhanced cooperative spectrum sensing in cognitive radio network using flower pollination algorithm," in *Proceedings of the 1st International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pp. 24-25, Swat, Pakistan, July 2019.

[28] A. Ahmed, N. Gul, I. Rasool, and A. Elahi, "Performance comparison of hard decision schemes in the presence of malicious users," in *Proceedings of the 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pp. 24-25, Swat, Pakistan, July 2019.

[29] N. Gul, M. S. Khan, J. Kim, and S. M. Kim, "Robust spectrum sensing via double-sided neighbor distance based on genetic algorithm in cognitive radio networks," *Mobile Information Systems*, vol. 2020, Article ID 8876824, 10 pages, 2020.

[30] M. S. Khan, L. Khan, N. Gul, M. Amir, J. Kim, and S. M. Kim, "Support vector machine-based classification of malicious users in cognitive radio networks," *Wireless Communications and Mobile Computing*, vol. 202011 pages, Article ID 8846948, 2020.

[31] M. Fathy, A. Tammam, and A. Saafan, "Influence of relaying malicious node within cooperative sensing in cognitive radio network," *Wireless Networks*, vol. 25, no. 5, pp. 2449–2458, 2019.

[32] N. Gupta, S. K. Dhurandher, and A. Sehgal, "A contract theory approach-based scheme to encourage secondary users for cooperative sensing in cognitive radio networks," *IEEE Systems Journal*, vol. 14, pp. 1–11, 2019.

[33] Z. Sun, Z. Xu, M. Z. Hammad, X. Ning, Q. Wang, and L. Guo, "Defending against massive SSDF attacks from a novel perspective of honest secondary users," *IEEE Communications Letters*, vol. 23, no. 10, pp. 1696–1699, Oct, 2019.

[34] Y. Fu and Z. He, "Bayesian-inference-based sliding window trust model against probabilistic SSDF attack," *IEEE Systems Journal*, vol. 14, pp. 1–12, 2019.

[35] P. Kaligineedi, M. Khabbazian, and V. K. Bhargava, "Malicious user detection in a cognitive radio cooperative sensing system," *IEEE Transactions on Wireless Communications*, vol. 9, no. 8, pp. 2488–2497, 2010.

[36] V. V. Hiep and I. Koo, "A sequential cooperative spectrum sensing scheme based on cognitive user reputation," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 4, pp. 1147–1152, 2012.

[37] J. Ma, G. Zhao, and Y. Li, "Soft combination and detection for cooperative spectrum sensing in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 11, pp. 4502–4507, 2008.

[38] Y. L. Lee, W. K. Saad, A. Abd El-Saleh, and M. Ismail, "Improved detection performance of cognitive radio networks in AWGN and rayleigh fading environments," *Journal of Applied Research and Technology*, vol. 11, no. 3, pp. 437–446, 2013.

[39] D. Hamza, S. Aïssa, and G. Aniba, "Equal gain combining for cooperative spectrum sensing in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4334–4345, 2014.

[40] D. B. Teguig, B. Scheers, and V. Le Nir, "Data fusion schemes for cooperative spectrum sensing in cognitive radio networks," in *Proceedings of the Military Communications And Information Systems Conference*, MCC, Gdansk, Poland, 2012.

[41] N. Marchang, R. Rajkumari, S. B. Brahmachary, and A. Taggu, "Dynamic decision rule for cooperative spectrum," in *Proceedings of the International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, India, 2015.

[42] J. Unnikrishnan and V. V. Veeravalli, "Cooperative spectrum sensing and detection for cognitive radio," in *Proceedings of the IEEE Global Telecommunications Conference (Globecom)*, Washington, DC, USA, November 2007.

[43] S. Bhattacharjee, "Optimization of probability of false alarm and probability of detection in cognitive radio networks using GA," in *Proceedings of the 2nd IEEE International Conference on Recent Trends in Information Systems*, Kolkata, India, 2015.

[44] M. S. Khan, N. Gul, J. Kim, I. M. Qureshi, and S. M. Kim, "A genetic algorithm-based soft decision fusion scheme in cognitive iot networks with malicious users," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 2509081, 20 pages, 2020.

[45] M. Akbari and M. Ghanbarisabagh, "A novel evolutionary-based cooperative spectrum sensing mechanism for cognitive radio networks," *Wireless Personal Communications*, vol. 79, no. 2, pp. 1017–1030, 2014.

[46] A. A. EL-Saleh and K. Hussain, "Cognitive radio engine model utilizing soft fusion based genetic algorithm for cooperative spectrum optimization," *Proceedings of the International Journal of Computer Networks & Communications (IJCNC)*, vol. 2, no. 3, pp. 169–173, 2013.

[47] A. Rauniyar and S. Y. Shin, "Improved detection performance of energy detector by optimization of threshold using BPSO algorithm for cognitive radio networks," in *Proceedings of the 2nd International Conference on Industrial Application Engineering*, Vancouver, Canada, 2015.

[48] M. Taha and D. Alnadi, "Threshold adaptation in spectrum sensing for cognitive radio using particle swarm optimization," in *Proceedings of the International Conference on Control*, Suwon-si, South Korea, October 2014.

[49] R. A. Rashid, A. H. F. Bin Abdul Hamid, N. Fisal et al., "Efficient in-band spectrum sensing using swarm intelligence for cognitive radio network," *Canadian Journal of Electrical and Computer Engineering*, vol. 38, no. 2, pp. 106–115, 2015.

[50] N. Gul, I. M. Qureshi, A. Omar, A. Elahi, and M. S. Khan, "History based forward and feedback mechanism in cooperative spectrum sensing including malicious users in cognitive radio network," *PLoS One*, vol. 12, no. 8, 2017.

[51] N. Gul and A. Naveed, "A combination of double-sided neighbor distance and genetic algorithm in cooperative spectrum sensing against malicious users," in *Proceedings of the 14th International Bhurban Conference On Applied Sciences & Technology (IBCAST)*, Islamabad, Pakistan, 2017.

[52] N. Gul, I. M. Qureshi, A. Elahi, and I. Rasool, "Defense against malicious users in cooperative spectrum sensing using genetic algorithm," *International Journal of Antennas and Propagation*, vol. 2018, Article ID 2346317, 11 pages, 2018.

[53] M. Jun and Li. Ye, "Soft combination and detection for cooperative spectrum sensing in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, 2007.

[54] A. M. Vargas and A. G. Andrade, "Comparing particle swarm optimization variants for a cognitive radio network," *Elsevier Applied Soft Computing*, vol. 13, no. 2, pp. 1222–1234, 2013.

Hindawi
Scientific Programming
Volume 2020, Article ID 8871315, 31 pages
https://doi.org/10.1155/2020/8871315

Review Article

Security Measurement in Industrial IoT with Cloud Computing Perspective: Taxonomy, Issues, and Future Directions

Sahar Shah,[1] Mahnoor Khan,[2] Ahmad Almogren,[3] Ihsan Ali,[4] Lianwen Deng,[5] Heng Luo,[5] and Muazzam A. Khan[6]

[1]*Department of Electronics, Quaid-i-Azam University, Islamabad, Pakistan*
[2]*Department of Physics, Government Post Graduate College, Nowshera, Pakistan*
[3]*Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11633, Saudi Arabia*
[4]*Faculty of Computer Science and IT, University of Malaya, Kuala Lampur, Malaysia*
[5]*School of Physics and Electronics, Central South University, Changsha, China*
[6]*Department of Computer Sciences, Quaid-i-Azam University, Islamabad, Pakistan*

Correspondence should be addressed to Ahmad Almogren; ahalmogren@ksu.edu.sa and Ihsan Ali; ihsanalichd@siswa.um.edu.my

Received 6 September 2020; Revised 4 October 2020; Accepted 8 October 2020; Published 26 October 2020

Academic Editor: Shah Nazir

In recent years, cloud computing has gained massive popularity in information technology and the industrial Internet of things. It provides facilities to the users over the wireless channel. Many surveys have been carried out in cloud security and privacy. The existing survey papers do not specify the classifications on the basis of cloud computing components. Therefore, they fail to provide the techniques with their specialities as well as the previously available literature review is outdated. This paper presents the security for cloud computing models with a new aspect. Unlike the previously existing surveys, the literature review of this paper includes the latest research papers in the field of cloud security. Also, different classifications are made for cloud computing security on the basis of different cloud components that are used to secure the cloud models. Furthermore, a total of eleven (11) classifications are considered, which includes cloud components to secure the cloud systems. These classifications help the researchers to find out the desired technique used in a specific component to secure the cloud model. Moreover, the shortcoming of each component enables the researchers to design an optimal algorithm. Finally, future directions are given to highlight future research challenges that give paths to researchers.

1. Introduction

This survey paper is organized in such a manner in which Section 1 contains an introduction to cloud computing, applications, and security. Section 1.1 is the contribution of our survey paper to the field of cloud computing security. Section 2 is called the literature review. The whole literature review section is subdivided into eleven (11) classification components. Every classification includes different research articles and at the end of each classification component a table is drawn which summarizes the overall classification component. After Section 2, the next is Section 3, the conclusion section, which concludes the presented survey. Then, in Section 4 the future research directions are discussed. At last, all the acronyms used in this manuscript are listed in Table 1. Figure 1. shows the organization of the manuscript.

The word cloud is described in 2006 for the business models which provide services over the Internet and the availability of data is the main concern in cloud storage [1]. According to NIST, cloud computing is classified into two major categories. One is based on services and the other is based on deployment models [2]. The service cloud computing category is further classified into three types: Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS). The deployment category

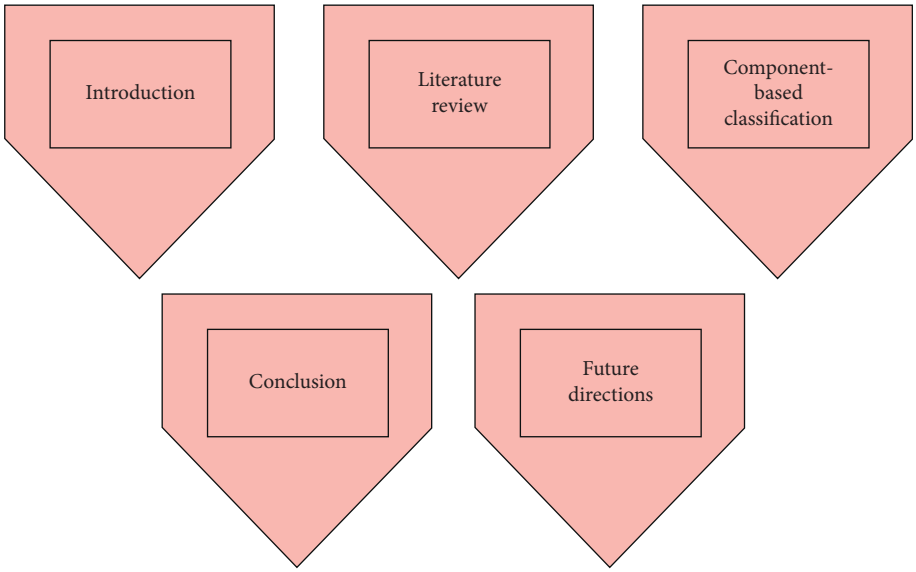


Figure 1: Organization of the manuscript.

is classified into the public cloud (P_b), private cloud (Pr_C), hybrid cloud (HC), and community cloud (CC). In the SaaS model, the overall software services are offered on the cloud. In the PaaS model, the development platforms are offered on the cloud. In the IaaS model, all the services related to hardware in a cloud are performed. While, in the P_bC type of cloud related to a specific organization and not connected to any other firm, such clouds have high costs and security. In P_rC type, the infrastructure is hosted by the traders of the cloud. The user has no control in this framework. The combination of P_bC and P_rC is called HC which is a scalable and cost-effective cloud. Mobile computing is used to encounter the security challenges in IoT [3]. In the CC model, infrastructure is shared between several organizations from a specific community. Furthermore, NIST also describes the most extremely important parameters. These parameters differentiate the cloud computing from the others, namely, on-demand self-service, broad network access, resource pooling, rapid elasticity, and measurement service.

The cloud computing field took huge attention of the researchers in information technology by having different powerful parameters. The parameters cost reduction, on-demand self-service, rapid elasticity, resource pooling, broad network access, high service scalability, flexibility, and high capacity of storage to afford the big data [4]. Many problems in the business domain have been solved by cloud computing and provide an efficient platform for the business community. These all advantages switched the many business communities to put their IT infrastructure to a cloud environment.

Obviously, cloud security and privacy are important parameters in every cloud system. If no proper security and privacy are provided to cloud models, then the cloud models will be no longer used. Cloud security and privacy are major barriers to cloud service adoption. The cloud security parameters attract the noncloud models toward the cloud framework. To provide security to the cloud computing components, different techniques are applied. The cloud security is responsible to provide a leakage-free, hijacker-free, and threat-free cloud system. The cloud environment is widely shared and aggressive [5]; hence, different blitz at a cloud network, framework, and approach to cloud duty can affect both the accessibility and security of the cloud computing. These risks and blitz can be inside and outside cloud. Inner risks and blitz are further classified into two divisions: one is malevolent insider working and the other is working inside for an organization. Fog-based IoT healthcare has an optimal response in terms of energy consumption and network delay of the fog nodes [6]. The cloud computing environment is distributed among companies and users. Therefore, cloud service distributors should maintain several basic mechanisms to strengthen cloud services in the security aspect. The industrial IoT merged the representative technologies such as machine-machine communication and 5G [7].

Many research articles, regarding cloud security and privacy, have been published. In [8], cloud security is classified into five different categories. The classification in this article is on the basis of security basics, structure, access control, cloud framework, and data. Numerous devices can be connected to IoT for industrial applications [9]. The basic analysis scheme called the Service Measurement Index (SMI) is developed in [10] for the industrial Internet of things. The SMI is further classified into seven subdivisions and one of them is cloud security. The subdivision of cloud security includes Access Control and Privilege Management, Information Privacy, and Loss. The industrial IoT is used for business purpose in China this business is in three aspects resources efficiency, sustainable energy, and transparency [11].

In this survey, cloud security and privacy are analyzed in a different aspect. Different techniques, tools, software, and algorithms are used in cloud components to strengthen the cloud system. This paper classifies cloud computing into eleven (11) categories, as shown in Figure 2. The

classifications involve different cloud components in which different methods, techniques, policies, models, approaches, software, and tools are used to enhance cloud security and privacy. Every classification is further subdivided into the overviews, advantages, the techniques used, and the paper publication year. This work provides the easiest way to the researchers in finding where, when, and how to use a cloud component for security purposes in any cloud model. The demerits of each classification provide an opportunity for scientists and researchers to design an efficient technique. Then, in future trends, all the deficiencies related to the specific cloud computing component observed can be enhanced and modified further.

1.1. Our Contributions. Shortly, the paper contributes in the following ways:

(i) This survey considers the classifications, based on different techniques used in cloud computing components, to secure the cloud model. The paper classifies the literature review into eleven (11) different classifications used to make sure the cloud security. Each classification includes various articles. The table is provided with each classification in which the overview, advantages, techniques, references, and the years are mentioned. This approach makes the easiest way for the researchers to find out the concerned cloud component related to any specific security issue in cloud computing. While the other surveys do not classify the cloud computing security into different classifications, which is a struggle for the researchers in finding the technique used, its advantages, and demerits.

(ii) Many survey articles in cloud computing security have been published. The shortcoming exists in their work. They include the oldest papers in the literature review, while, in this survey, the latest papers were included from 2015 to 2020 in the literature review section.

(iii) In this survey, all the possible future directions in cloud security are mentioned. The techniques, software, etc. used in cloud computing components have different flaws. This paper addresses them which helps to enhance them in future work. However, previously published surveys fail to discuss these flaws. Table 2 shows the differentiation key points which differentiate our survey from the existing surveys.

(iv) The bibliographic-based survey performed in our paper while the other surveys fail to do this. This bibliographic-based survey involved country-, author-, and year-based surveys in the field of cloud computing security. This shows the top authors, countries, and years which have the main role in cloud computing security.

(v) Due to the high demand for cloud computing models, the cloud setup enhances its storage capacity at low cost. Cloud models increase the number of tools and machine learning functions. These enforced the organizations to move on a cloud computing-based platform. Besides these points, the advancement of the cloud computing model from earlier is based on five points. (i) Earlier, from two decades, the cloud computing models more clarify and simplify the customer touch-points. (ii) Creating a definition of Internet protocol (IP) ownership related to artificial intelligence (AI). (iii) Minimizing risk and simplifying multicloud. (iv) Creating a more robust sales and partner strategy. (v) Delivering platform innovation.

Table 2 is for the sake of simplicity, which differentiates this survey from the already existing surveys on the basis of key parameters. The tick symbol shows the inclusion of the parameters such as parametric-based survey, contribution section, cloud security applications, component-based classifications, and future trends in the respective survey, while the cross symbol shows the parameters not included in the mentioned survey papers.

2. Literature Review

2.1. Classifications. The literature review section is divided into different classifications according to cloud computing components used to secure the cloud systems. Table 3 illustrated different classified components which are categorized for the cloud security purposes and contains different number of articles. Figure 3 shows the classification of the related work in the form of bar graph which includes number of years versus number of papers.

2.2. Storage-as-a-Service. Table 4 shows the references, overviews, algorithms, techniques used, and advantages along with the publication years for each paper included in this category of the classification. Data security needs serious attention as the cloud storage increases day by day. In [20], the cloud storage security policy is analyzed. The model permits the users to avail of the resources on-demand including computer and storage systems providing fast, efficient, and inexpensive computing. Such a specification of a cloud model makes it a large data cloud computing model from the tradition cloud model. There is no standardization, normalization, security policy, and security mechanism in the process of data transmission of cloud storage technology. In the network layer, data is not only leaked in the form of electromagnetic waves but also intercepted in network communication. So, by this the hackers can take advantage of vulnerabilities and technical errors of the network. By using the technology in an improper way, the data in the data link, network, and transport layers will be at risk; even SSL, SSH, IPSEC, and other VPN technologies are considered. The cloud storage system is a complex method and needs the security arrangements of different cloud providers have their security policies and technical solutions. However, there are two types of risk on data storage called hardware application strategy and cloud storage service providers management.,

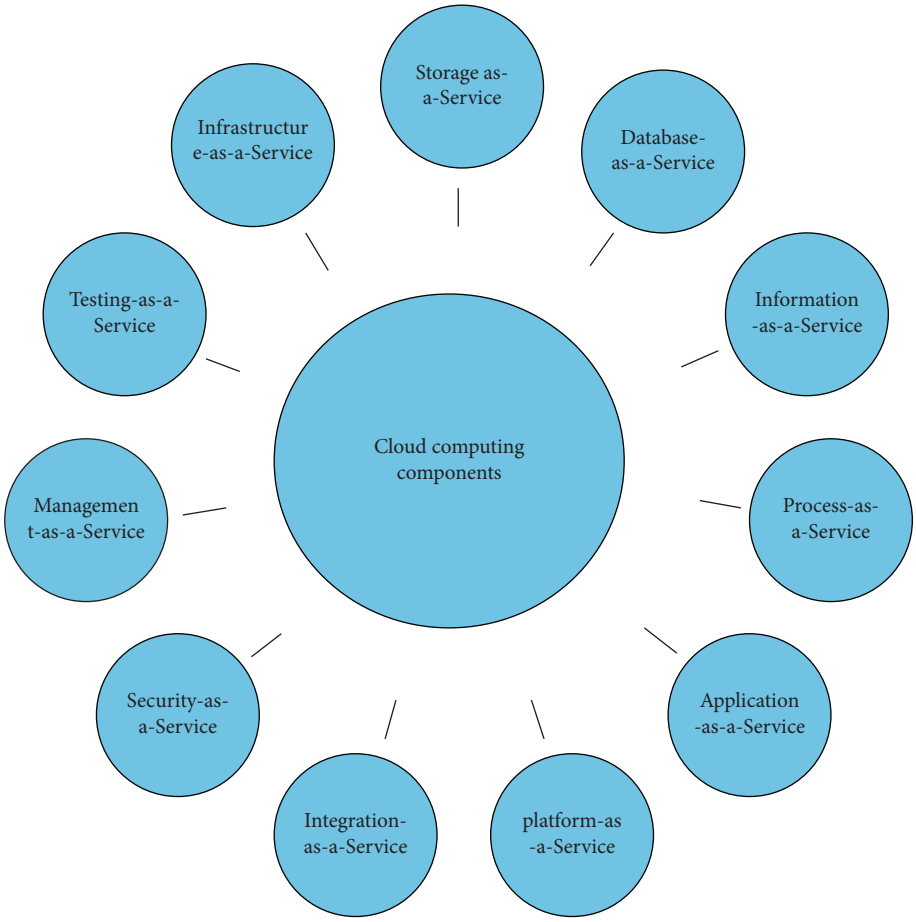


Figure 2: Cloud component-based classification.

In this framework, the online distributed storage system is presented, in which the identity authentication, access control, and encrypt transmission techniques are used to protect the cloud storage system from the security threats. In this model, firstly, confirmation of user identity has been done to create mutual authentication among the user and cloud storage server. Secondly, to protect data from threats, the end-to-end data transmission is achieved using encryption on it. Thirdly, to ensure that the data is not used by any other unauthentic person, access control is used. These all protect the data regarding integrity, availability, and confidentiality. The encryption method provides the confidentiality of the data. The data is fragmented in this model using Shamir's threshold secret sharing program. The data storage security assurance is obtained by using data scattered storage technology, and there is a direct relationship between the error rate of storage data and storage system capacity. The data store in the cloud storage model, and then it is distributed in different places through which distributed online cloud storage security is achieved. In short, in this model, some technical issues regarding security are discussed on layers and data storage cloud, which gives their protecting solutions.

In [21], the addressing of the data placement problem in a cloud system is proposed. To overcome the problem of high data retrieval time and threat level for data security, an intelligent method that achieves high performance along with the security satisfaction is discussed in the proposed framework. In this work, a novel approach that addresses the data placement in a cloud storage system is discussed. The whole work is divided into steps. In the first step, the linear data programming model is formulated for the data placement problem. This is responsible for reducing the data retrieval time and distributing over different storage nodes. In the second step, a heuristic algorithm called the SADP mechanism developed to solve the problem for the cloud storage system (Sedulous). Therefore, a novel approach is needed which guarantees data security by minimizing the overhead of the security service. The performance issue with the security requirement creates the problem of data placement. To solve this issue the entire data file is divided into multiple numbers and the data spread over the pieces of the file and each will store in a separate mode. These pieces of the file spread over storage nodes with a specific distance between any two pairs of the fragments. This mechanism ensures that if an attacker successfully enters a piece of file data, then it is unable to leak, reveal, and find the location of the meaningful data of the user. By solving the three subproblems, the data placement issue can be solved. The three problems are the decision of the number of chunks, the decision of the size of each chunk, and, the last but not the least, the selection of storage nodes. These three issues are

Table 1: Abbreviation used in the paper.

Abbreviations	Words
PaaS	Platform-as-a-service
SaaS	Software-as-a-service
IaaS	Infrastructure-as-a-service
Pb	Public cloud
Pr	Private cloud
HC	Hybrid cloud
NE	Nash equilibrium
CSU (s)	Cloud service user (s)
CSP (s)	Cloud service provider (s)
CI	Cloud infrastructure
SSG	Security Stackleberg game
FEBM	Feedback evaluation and Bayesian model
CSPM	Cloud security and privacy model
NIST	National Institute of Standards and Technology
ACPML	Access control and privilege management layer
CCAF	Cloud computing adoption framework
BPMN	Business process modeling Notation
CISL	Cloud infrastructure security layer
PESL	Physical environmental security layer
HTTP	Hypertext transfer protocol
DDoS	Distributed denial of service
SDLC	Software development life cycle
ABDS	Attribute-based data sharing
DAC	Data access control
SK (s)	Secret key (s)
PSSFP	Previous-selected-server-first policy
MOO	Multiobjective optimization
DFS	Distributed file system
CCTV	Closed circuit television
IM-SecaaS	Intrusion management Security-as-a-Service
OPEX	Operational expense model
IDPS	Intrusion detection/prevention system
CEPM	Compliance enforcement and policy management
SSDP	Simulation software development process
ADIRS	Attack detection and intrusion rejection system
CSA	Cloud security Alliance
SecSLA	Security service level agreement
QPT	Quantitative policy trees
QHP	Quantitative hierarchical process
QoS	Quality of service
DoS	Denial of service
PM	Privacy manager
GA	Genetic algorithm
SADP	Security-aware data placement
PSO	Particle Swarm optimization
ACO	Ant colony optimization
CAS	Chaotic ant Swarm
GA-CAS	Genetic algorithm-based chaotic ant Swarm
MHA (s)	Metaheuristic algorithms
NDTMSI	Nondeterministic task meta-Schedular integrated
AES	Advanced encryption standard
HEVC	High efficiency video coding
IEVS	Intraencoded video stream
TLS	Transport layer security
SED2	Secure efficient data distribution
EDcon	Efficient data conflation
MDMCO	Multidimensional and multiconstraint optimization
BC (s)	Blockchain (s)
BPDPP	Blockchain's public and distributed peer-to-peer
BWH	Block withholding
PoW	Proof-of-Work
PPLN	Pay-per-Last N
SBOS	Security benchmark for open stack
PKE	Public key encryption
SLA (s)	Service level agreement (s)

solved by introducing a linear programming model, a fast heuristic algorithm called Sedulous. The Sedulous follows a greedy approach in which the preference is given to the nodes which can transmit data with high speed. For the security addressing, the T-coloring approach used makes sure that the two adjacent nodes will never be with the same colors, in which they never store the simultaneous chunks of data. The simulation results show that the presented framework reduces the retrieval time up to 20 percent for the random network topology and 19 percent for the Internet topology systems, and these results were compared with the baseline methods and only the security parameter was noticed. The results also show that the model achieves minimum retrieval with the best performance and data security. However, the proposed work scarifies the rejection ratio and reduces the cost of the commercial cloud provided.

The huge data can be stored on cloud storage with no limitation of storing memory. In [22], to secure the cloud data in an open platform, an efficient scheme is proposed. The scheme encrypts and decrypts the files of data, multiple auditing processes, and uses dynamic operations with the integrity of data. From start to end, the whole algorithm is divided as the start, public audibility of data files, checking data dynamics, verifying integrity proof, multiple batch auditing processes, privacy-preserving public auditing scheme, process storage system, and end.

In [23], the authors target different security issues that cannot be ignored and which degrade the performance and trust of the CSPs. The proposed work then evaluates the security issues of the cloud by proposing techniques to make sure the security of the cloud. Different issues in the cloud are data loss or leakage, account, insecure interfaces, dos, and abuse and nefarious use. The article focuses to store the data in the cloud with strong security. To store the data in the cloud, it is first divided into small and large pieces and then put on the different media. Every small and large piece of data has its advantages and disadvantages, i.e., the large piece of data is easy to follow and read while the small piece is difficult in terms of detrimental performance. The hackers can hijack and target large piece of data because it is easy to follow and read. The proposed framework encrypts the piece of data through programming before putting them on the different media and decrypts it gain on the user side. The symmetric encryption, also called a public key cryptographic approach, is applied to the data to encrypt it due to the larger size of data. The symmetric encryption approach is preferred rather than the asymmetric approach due to the public key algorithm and to encrypt big data files especially in

Table 2: Key points which differentiate our survey from the existing surveys.

References	Years of publication	Parametric-based survey	Contributions	Cloud security applications	Component-based classifications	Future trends
[12]	2017	x	x	x	X	x
[13]	2018	x	x	x	X	x
[14]	2018	x	x	x	x	x
[15]	2015	x	x	x	x	x
[16]	2016	x	x	x	x	x
[17]	2020	x	x	x	x	x
[18]	2020	x	x	x	x	x
[19]	2019	x	x	x	x	x
Our survey		√	√	√	√	√

Table 3: Number of papers in each classification.

Sections	Classification title	Number of papers	References
2.2	Storage-as-a-Service	Seven	[20–26]
2.3	Database-as-a-Service	Three	[5, 27, 28]
2.4	Information-as-a-Service	Five	[29–33]
2.5	Process-as-a-Service	Five	[10, 34–37]
2.6	Application-as-a-Service	Three	[38–40]
2.7	Platform-as-a-Service	Four	[41–44]
2.8	Integration-as-a-Service	Four	[45–48]
2.9	Security-as-a-Service	Ten	[49–58]
2.10	Management-as-a-Service	Eight	[2, 8, 59–64]
2.11	Testing-as-a-Service	Three	[10, 65, 66]
2.12	Infrastructure-as-a-Service	Four	[67–70]

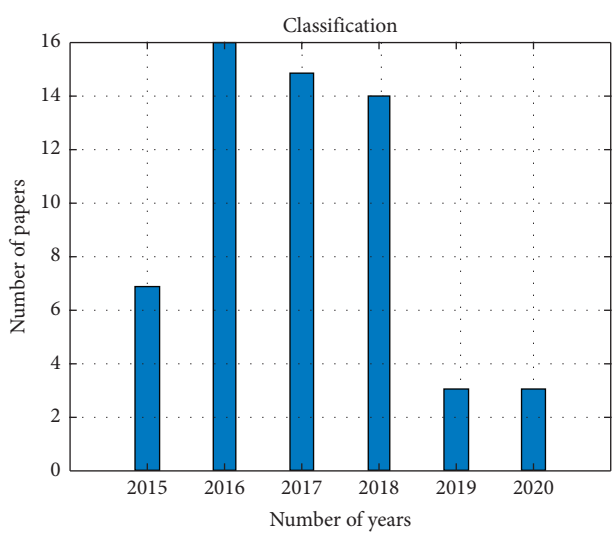


Figure 3: Classification of the related work.

gigabytes. The AES algorithm is used to encrypt the data. This work efficiently secures the big data. However, small data encryption is not discussed.

In [24], the data classification-based secure cloud computing model is proposed. The previous approaches regarding storage frameworks use the same key size for the data encryption without looking to the confidentiality level with an unnecessary overhead as an output which increases the processing time. In the proposed framework two important issues of mobile cloud computing have been solved: one is called security and the other one is called storage. For the confidentiality of the data, an efficient framework is proposed which provides strong confidentiality to data and integrity in the cloud storage system in both aspects, i.e., transmission and storing operations. Along with it, this framework reduces the processing time to encrypt the data and complexity. Internally or externally of a cloud system, all the threats are hacked in an efficient framework. The second is without taking the confidentiality degree when the encryption of the data is performed. To tackle this problem the proposed framework includes the algorithm which enables the user to encrypt its data with a key that is not available to any other person. Huge data, i.e., 100 gigabytes in a combined form is difficult to encrypt with the same key; therefore, a cloud storage model is proposed which includes three levels to encrypt the data with confidentiality. The three levels are basic, confidential, and high confidentiality. The idea of specifying in the confidentiality level of data follows in the proposed work which is called the manual classification. For the encryption of data, different cryptographic approaches are used called AES, TLS, and security hashing algorithm (SHA). The data with higher and lower critical values will be stored on faster and slower media, respectively. In short, an efficient confidentiality-based cloud storage framework is proposed, which enhances the data encryption process time and integrity. The high data confidentiality and integrity is achieved.

The cloud is mostly used to store and process big data. In [25], the authors propose a big data division mechanism. In this model, the big data is divided into small sequence parts

and then the sequenced parts are stored into different multiple CSPs. After the division of big data, the divided big data sequences are collectively combined with their sequence and stored on different cloud services. There are two types of storing the data on a cloud service: one is public data and the other is confidential data. The public data is accessible by any user and it is open to all, while the confidential data can be accessed by the relevant users. To make the big data of tenants secure, a secure cloud big data cryptographic-based virtual mapping storage scheme is proposed. In this scheme, the big dataset is divided into sequential data parts with a certain principle called the same data-type block or IP resembled. The big data of tenants divided into n sequenced numbers, where each divided sequence of big data stored on m different storage providers. A unique storage path for big data is formed when the big data is stored. The trapdoor functions are used to protect the mapping of various data. The trapdoor functions are widely used in cryptography. It is difficult and even impossible to secure or encrypt the big data as a whole, so we only need to encrypt the storage path of the big data, and according to this, we can obtain a cryptographic value called cryptographic virtual mapping of the big data. To more improve the robustness of the proposed work, the model stores multiple copies of each data on different indexes of the providers. If one data is lost or missed, then the cloud checks the missing data sequenced number in different indexes and tries to recover and store it on the relevant provider. The simulation results show that the theoretical proof analyzes the model in an efficient and secure way. Two different scenarios are considered, and comparing these two scenarios with the related approaches finally proves that the proposed scheme is more effective and feasible to protect big data for cloud tenants.

In [26], different security services are discussed using a sequence of game models amongst the cloud user and provider. The security assessment model is used with the help of which users can find the risks of their data privacy which is likely to be hacked by cloud. By taking into account the security of the data amongst cloud user and provider, this work investigates three types of scenarios, which are one user, multiuser model, and multiservice provider. Different series of models has been discussed to develop the security for CSPs and TSPs. In the first game model, a game is established between one user and a cloud provider. The user has the choice of whether to use the cloud services or not while on the hand the cloud provider has the choice to steal the data of the user or remaining host. In the second game model, there are many users and only one service provider. If the data of one user has been stolen by the service provider, then all the other users lose the trust of the cloud provider. In the third game model, many users and many cloud providers are discussed. The utility function is used to differentiate the incentive for each user and provider. In this, the relation of all users and providers is not the same and is classified into competitive, cooperation, and dependent. In the first one, different providers provide their services, terms, and conditions. In the second, the users can use the sources of all providers. In the last classification, the providers serve as the third party.

2.3. Database-as-a-Service. Table 5 presents the short summary of the Database-as-a-Service classification.

In [5], the pricing and investment problems between the cloud insurers and the users in a cloud market are proposed. The cyber threats are responsible to damage the cloud user's data. The market includes users, cloud providers, and cloud insurers. The cloud providers provide cloud services to users. The cloud insurance has a product which the users buy to protect their data from damage. When an attack happens to the cloud service, then the cloud insurer pays a claim. The users are dependent on each other in which they can take the benefit of the security effects which are produced by the other user's investments in security. In this model, it is assumed that the cloud provider and the cloud insurer are the business partners. Therefore, to improve the security levels, the cloud insurer charges the cloud platform, i.e., to enhance the quality of the cloud service and reduce the paying claim probability. The Stackelberg game is proposed in this model which has two stages. In the first stage, the price charging is set on the users by the cloud insurers and the improvement in the cloud security quality decides by the investment. In the second stage, the users decide on cloud insurances to purchase based on the observed prices and qualities. By applying the game-theoretic approach called the backward induction, the optimal pricing, security investment strategies, and optimal strategies of the users are presented. The NA of the game regarding the best responses of the users is found. The best response of each user can be found by taking the derivative of the utility function concerning for to function on demand. The simulation result shows that the Stackelberg model is proposed to maximize the utility of the users which correspondingly maximizes the profit of the cloud insurers.

In [27], the authors propose a pay-per-use IM-SecaaS model for cloud security. The architecture of the IM-SecaaS involves the main components: intrusion detection, intrusion response, reporting, and logging. Before the data reaches the users, the model checks and cleans the data by monitoring web traffic or the attackers attack the data. There should be no investment for anything on premise solutions. However, the client should pay based on the pay-per-use model. IM-SecaaS is provided as a service for the users, so, in this case, the users should pay in the form of OPEX. The model is proposed for Pb in which proof of concept prototype is implemented. The function of the IM-SecaaS is like the policy administrator. On one side the public Internet is input to IM-SecaaS where doubtful traffic comes while on the other side the client organizers are present. By applying all the policies on the input doubtful traffic, it becomes clean and then delivers it to the client. On a virtual machine, the model is implemented in a Pr. Both the IM-SecaaS core and IM-SecaaS managers have been implemented in different virtual machines. The IDPs from multiple vendors are more efficient than using it for a single vendor. The uninterrupted service is provided to the cloud by proof of concept (POC). The model enhances the flexibility, control, privacy, and cost for a Pr.

In [28], the authors trying to solve some security issues regarding errors of screening indicators lack validation

Table 4: Storage-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[20]	The model presents different security policies on the cloud storage and data; the data is not secure everywhere in the cloud data which is fragmented for security purposes; the identity authentication is used for security	The model analyses the cloud storage different methods are applied at different stages to achieve security	The data transmission encryption, access control, identity authentication, and Shamir's threshold secret sharing program are used in the presented model	2017
[21]	The data placement problem is addressed for the big data in the cloud storage system; the data placement problems are solved by the decision, size, and number of chunks	The renewal time minimizes up to 20 percent and 21 percent for the network and Internet topology, the linear programming model, heuristic algorithm, and T-coloring approach	The linear programming model, heuristic algorithm, and T-coloring approach	2016
[22]	As the cloud network is an open platform for the users to store and compute the data; the users store and compute the data in the cloud over a wireless remote network by using the Internet; security of the data is the concern in an open system for this purpose encryption and decryption of data outperformed to secure the data	Different steps involved in the proposed work which strongly secures the data from the start to end of the system	Encryption and decryption of data with integrity are applied	2017
[23]	Several different security issues are discussed in this article; after that it proposes its own work; the data is divided into small and large sizes; the symmetry encryption approach is applied to large size data	A strong encryption on data applied which fully secures the larger size data from threats	The AES symmetric encryption algorithm is applied to the system	2015
[24]	The proposed work targets the two main security issues called the confidentiality level of data and data encryption processing time; the key provided to the users which are not for to avail by any other user	The data encryption time and overhead reduced by achieving a reliable cloud network	The AES, TLS, and the secure hashing algorithm (SHA)	2016
[25]	The big data is divided into n sequences and stored in m different CSPs which is proposed to ensure the security of big data; the cryptographic virtual mapping and trapdoor techniques are proposed to achieve high feasible protection of the big data	Two different scenarios are analyzed in the proposed model, and the simulation results concluded that the proposed model provides more security for big data in an efficient way than the other traditional approaches	The cryptographic virtual mapping, data type block or IP resembles, and trapdoor function	2016
[26]	Three game models are discussed to analyze the security in the cloud system; the assessment model helps the users to find the risk of data privacy	This model ensures the privacy of the data and hence minimizes the influence of the third party in private data	Security assessment model, one user, multiuser, and multiprovider	2019

reputation scientifically. Based on these issues, a well-reputed security model is presented using S-Alex Net convolution neural network and dynamic game theory called SCNN-DGT. These algorithms are used to privatize the health data in the Internet of Things. The proposed model is classified in two different stages; firstly, the health data information of the user is arranged using S-Alex Net; secondly, game-theoretical approach is used. In other words, for the sake of data security from harmful identity game, a theoretical approach called security reputation model is used and hence data security is improved. For learning invalidation which is caused due to small matrix features, S-Alex Net neural convolution is used. Moreover, this paper presents the SCNN-DGT model which includes three stages: predispose stage, big data security stage, and early warning supervision.

2.4. Information-as-a-Service. As a summary of this classification component, Table 6 is shown.

In [29], the authors propose a new attribute-based sharing scheme used to resolve the security issue while the data sharing is performed in a cloud. The data sharing on a cloud becomes more popular due to its low cost. To share the data in a fine manner attribute-based encryption (ABE) gain more attention. The previous ABE

Table 5: Database-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[5]	The SSG is proposed for cloud security; the defender and attackers are compared by applying the model; the utility function, best strategy, and payoff are analyzed in the game model	The information on the attacker's behavior is collected by active and passive stages of the SSG; the efficient utility function is achieved by this model; the defender availability and cost are maximum and minimum, respectively	The security Stackelberg model, active and passive stages, utility function modeling, natural roles, and defense strategy	2018
[27]	This model works on a pay-per-use basis; doubtful traffic enters at one side of the model by applying different policies by the model components; a clean output data is provided to the clients	The model not only improves the security but also the flexibility, control, effectiveness, and performance as well	The IDPS and POC are used in this model	2016
[28]	This proposed work aims to manage and secure cloud data; incorrect screening and privacy of the cloud data is encountered	The outcomes of the experiments show that the proposed model solves the problem of low accuracy and reliability of the data in the cloud environment	The neural network model called S-Alex Net and game-theoretical approach SCNN-DGT is discussed	2019

presented by the researchers has the disadvantages of high consumption overhead and weak data security. The proposed work addresses these challenges by introducing a new ABE technique with additional features. By the addition of the system public parameters, the computation task has been eliminated with partial encryption consumption offline. Most of the computation overheads are eliminated by adding the public ciphertext before the decryption phase. Furthermore, for more securing the data, a chameleon hash function used to generate immediate ciphertext which is blinded by the offline to obtain online ciphertext. For the ABDS for mobile users, the new online/offline ABE scheme introduces to eliminate the moving encryption computation overhead on data at the owner side to the offline phase. The public ciphertext enables the user to check the cost of whether the equation holds a low cost or not before going to expensive decryption. The proposed framework is proven that the chosen ciphertext attacks (CCA2) are recognized as a standard security notion. Both experimental and theoretical results prove that the ABDS system outperformed resource-limited mobile users in cloud computing. The simulation results show the proposed work stands well regarding cost in the online and offline encryption time.

The data security is linked not only with the CSP but also with the user concerning data at rest, transferring data, and processing the data. In [30], the BPMN for the cloud data security is presented. The aim is to provide security to the cloud data; for this purpose, two types of CSP are used called CSP-1 and CSP-3, while the CSP-2 is disclosed and does not provide information for academic publishing. The BPMN can point out the section which is affected by the security attack and hence save the time and resources of the attack to perform recovery actions. In BPMN, large-scale penetration testing performed to test the power of the security system and service. The BPMN model is used effectively for the security business organization. For this purpose, the BPM is used for the security of the cloud. Bonita soft is software used for business process management (BPM) modeling and SSDP. The BPMN is used to simulate the data security for the three CSPs called CSP-1, CSP-2, and CSP-3. These service providers get the security data designs partially from the service providers and partially from the users. The CSP-1 is responsible for secure delivery and high-level services such as networking, storage, database services, infrastructure, and computing to the cloud system. It focuses on the data security model shared between the provider and the client. The several stages of data security and transfer security are performed by CSP-3. It incorporates the security policies, company structural security, data management, access control, personal, physical, environmental, and infrastructure securities. In short, the business security model proposes using CSP-1 and CSP-2 models. In the first case, the user requests the service provider after which it enters into the system and passes through different several management and security checks. In the second case, there is no request to enter into the system, but the request for logging into the client system. After the completion of these two cases, the whole data passes through many layers for the security purpose; if any layer misses the data, then it passes through the ADIRS. This type of model is applied in the health sector, national security services, banking, and many other companies which store confidential data.

The proposed framework achieved a minimum makespan and maximizes the reliability by assigning different tasks and data blocks in a cloud system [31]. The proposed framework tries to tackle the NP-hard problem; for this purpose, different intelligent computational algorithm such as GA, PSO, and ACO are proposed. The ACO is a probabilistic algorithm used to find the optimal path in a graph by the behavior of an ant colony. The ACO has many properties with the drawback of stagnation in the evaluation process. To overcome this problem, chaos factor is introduced. The CAS is a heuristic random search algorithm based on intelligence theory which affects the evolution of an ant colony

Table 6: Information-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[29]	To tackle the problems of computation efficiency and weak data security, ABDS scheme is proposed in cloud computing; the users are enabled in this approach to check the cost of the ciphertext in online/offline encryption mode before going to expensive full decryption	The users can get online and offline encryption with minimum cost and time; the proposed work gives strong security to weak data	The ABE, attribute-based data sharing (ABDS), chameleon hash function, and online/offline ciphertexts	2018
[30]	The framework used two types of CSPs to secure the cloud data while sharing the data and passes through several stages; the main focus is to provide security to the big data system	The model highly provides security to cloud data services such as storage, networking, data management, access control, and infrastructure security	Bonitasoft, business process management, CSP-1, and CSP-2	2016
[31]	To tackle the NP-hard problem, a series of MHAs, characteristics with a simple structure, fast confluent, and energy have been proposed; the MOO problem solved in terms of reliability, make span, and time	The proposed CAS solved the problems of the salesman, economic dispatch, and fuzzy system identification; the ACO finds the optimal path in a graph investigating the behavior of an ant colony, make span and flow time analyze the reliability factor	The ACO, chaos factor, make span, time flow, and GA-CAS	2016
[32]	The proposed work highly secures the cloud data; in this work, the data is split and then distributed by different cloud servers; the paper tries to resolve the abuse issue; the direct access of the cloud to the user's original data is prevented	The experiment proves the proposed work efficiently performs this task by consuming less time as compared to AES	The algorithms SED2 and EDcon	2016
[33]	The US government data is protected from advanced persistent threats (APTs); the cloud security assessment model is used for four multitenant IaaS cloud architecture	The CCS penetration probability is high if minimum security control sets are applied	The cloud security assessment model, APTs, and virtual machine	2017

from chaotic to individual self-organizing in a random search process. The CAS evolution involves two phases called the chaotic phase and the organization phase. Besides this, the Markov-based method is proposed for reliability. Furthermore, in the GA-CAS algorithm, four operators and natural selection are applied to solve the MOO problems. The four operators according to the characteristic of the cloud scheduling problem enable the CAS to solve the combinational optimization problems. The task scheduling problem in cloud computing is solved by the proposed MOO model. The user's task is taken into account in a multi-objective model and illustrated the scheduling performance in terms of make span, flow time, and reliability by applying queuing theory and Markov process. The make span and flow time encounter the reliability factor as follows: (a) only one task can be executed by each node at any time, (b) the links or communication links are independent of each other, (c) there is no discontinuity in the process of subtask execution, and (d) all nodes play their own role, and there is no useless node. even a series of tasks are inputted to the model. The simulation results concluded that when the proposed GA-CAS algorithm is compared with the other meta-heuristic approaches, it outperforms in solving the task scheduling problem of a cloud system.

A novel-based approach is presented in [32] to provide high security and an efficient mass distributed storage (MDS) service to the cloud data. The anxiety and adaptability of the users can be increased by cloud if the cloud operator directly reached to sensitive data. The paper mainly focuses on this issue; therefore, a novel approach is proposed to overcome this issue. The proposed algorithm efficiently splits the files and data which store separately on the distributed cloud service. In this phase, data is unable to reach the cloud service operators directly. Two main algorithms worked in this propose article called SED2 and EDcon. The framework targets the problem of abuse issue. All the data, encrypted and distributive, is stored on the different cloud servers without causing any big overhead and latency. The SED2 algorithm splits the data to prevent the data from leakage by using minimum cost. In executing the SED2 algorithm, two other supportive algorithms are used to encrypt and decrypt the data in a good way. In summary, the whole framework was proposed for the two main points. In the first, the proposed work prevents the cloud provider to reach directly to the original user data. In the second, a highly efficient mechanism is used to split the data which never produces big overheads but ensures data retrievability. The experimental results show the proposed work is compared with AES. The execution time for the proposed work is much lesser than the AES algorithm.

The vulnerabilities to advanced persistent threats (APTs) in cloud computing systems (CCSs) are important.

The reference model of the cloud covers and controls the security. In [33], the cloud trust assessment model is proposed to estimate the high-level security which is a high quality of confidentially and integrity by a CSP. The proposed model can access the security levels of four multi-tenant IaaS clouds which have alternatively equipped architecture for the cloud security model. The proposed CCS reference model and an assessment model ensure high security to IaaS, CCSs, and CSPs. Cloud tenants can be used to decide on which one CSP security features need to implement. The proposed CSS four architectures are designed to protect the government official data, and then they are practically implemented in the US. Whether these architectures successfully protect the US government official data has been analyzed. It is based on the BNM of the CCS. The spanning of the CCS attack is carried out by the APT. Each attack path needs the space, and the APT attack steps to implement. The CCS secrecy status is summarized by the two key security metrics: the first one is the chance when an APT can access high magnitude information. The second is the detecting chance of APT by the cloud tenant. In first security metric checks, high magnitude data called "Gold" information weather adjusts or is deleted from the CCS. The second metric assesses the analysis of cloud monitoring tenants, file approach, and alertness data to find intrusions into a tenant's cloud network, whether they contribute to intrusion detection or not. The results show the penetration probability of CCS is high if a minimal set of security controls are implemented. The CCS penetration probability drops substantially when the cloud protection in depth security is adopted which protects the virtual machine images.

2.5. Process-as-a-Service. In Table 7, the summary of this classification is shown. This summary is based on the references, overviews, techniques, advantages, and papers' publication year of each paper cited here. Cloud service becomes more popular due to highly upgrade of information technology and due to low cost, service on demand, high service scalability, and many more. However, security and privacy are not completely provided yet. In [34], the new security and privacy model for cloud service are provided. The proposed model is called CSPM. The NIST provides the main structures and buildings for the cloud called services, deployment models, and characteristics. The services for the cloud are security as a platform, infrastructure, and software. The deployment models include public, hybrid, community, and private. The characteristics are five in numbers and called on-demand self-sourced, broad network access, resource pooling, rapid elasticity, and measured service. The proposed models investigate different threats and give valid solutions for the corresponding threats. The security issues are ACPML, data (geographic, integrity, loss, privacy, and physical), environmental security, and proactive threat. The CSPM consists of five layers called physical and environmental security, CI, network securities, data and access control, and ACPML. These layers introduce the security policies, management, and monitoring steps for the cloud service. These layers are investigated at every stage of the cloud by the proposed model and make the difference between the attacks threaten availability and security along with the countermeasures which provide the security services to their clients. The model helps the CSPs in terms of security management and privacy. And the monitoring of cloud facilitation can be achieved. The threats and attacks in a cloud system can be insider attacks, cloud malware injection attacks, and cryptographic attacks. In short, the layered model presents the threats and attacks along with the countermeasures.

The novel scheme called attribute-based encryption (ABE) with the policy updating method is proposed in [35]. It mainly focuses on access control in an efficient manner with the updated dynamic policy for big data. In the proposed approach, the computational work minimizes due to avoiding the transmission of encrypted data and by using previously encrypted data with access policies. For different types of access policies, different updating algorithms are proposed in this work. The grand challenges in policy updating are correctness, completeness, and security. The correctness is the ability of the users who possess the attributes and are able to decrypt the encrypted data under the new access policy. The completeness is the method of policy updating which have the ability to update any type of access policy. The security of the access control system should not break by policy updating. The main goal of the proposed scheme is to solve the problems in policy updating using ABE systems. Firstly, the formulation has been done of the policy updating problem in ABE systems, according to which new methods are developed to outsource the policy update to the server. An expressive and efficient DAC scheme for the big data is presented through which dynamic efficient policy updating can be achieved. For different types of access policies, different policy updating algorithms are proposed such as Boolean formulas and access trees. The proposed algorithm not only satisfies all the above problems but also avoids the transferring of encrypted data in back and forth shape. The policy updating problem in big data is incorporated in this proposed scheme. Furthermore, a method is proposed which enables the data owners to check the correctness of the ciphertext updating. It also provides the safety in terms of the data owners, i.e., it cannot use their SKs to decrypt any ciphertext encrypted by other owners, although their SKs contain the components with the attributes. Although, in the designed policy, updating algorithms is based on water. In short, the simulation results prove that the proposed scheme is good in terms of cost, ciphertexts updating, and policy checking. Also, the proposed scheme provides the correctness, completeness, and security to the big data.

The demand for cloud service increases day by day according to the demand for cloud services; they provide enhance scaling, agility, availability, and flexibility. However, the cloud has some issues which are to be improved: load balancing, security, and fault tolerance. In [36], Rahul Rathore et al. presented a cluster on geographical-based dynamic distributed load balancing technique. The job assigned to each cloud provider is 100 in length and the distributed arrangement is chosen for all cloud servers. If the job number 101 has arrived at any service provider, then cluster applies its load balancing algorithm. In this model, a security mechanism is also introduced which secures the

Table 7: Process-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[34]	The model investigates different security and privacy issues; in this work, five layers are introduced which protect the cloud system at every layer from threats and attacks	This model provides a secure service with confidence; it provides the difference between attacks and security along with the countermeasure secure services	The model used five different layers to protect the cloud from threats called the CI, CISL, PESL, data layer (DL), and ACPML	2016
[35]	The policy updating problem in the big data is solved by proposing the attribute-based encryption (ABE) scheme; then, DAC scheme for the big data is proposed which enables the owners to dynamic policy updating; different policy updating algorithms for different types of access control are proposed	This approach enables the checking of directly updating ciphertext by the cloud; there is no need for policy updating in data decryption	The outsource ABE, DAC scheme, Boolean formulas, and access tree	2016
[36]	The static and dynamic load balancing techniques are proposed; the pairs of keys are generated to secure the data; load balancing attached to a cloud effectively avails the resources provided to nodes	The model performs better in terms of response time, throughput, resource utilization, fault tolerance, and scalability	Load balancing technique, dynamic load balancing, and key pairs	2018
[37]	The data security monitoring method based on autonomic computing is proposed; different modules are proposed to gather the data stream and then evaluate the processed data to determine the abnormality; the data collection and analysis of storage are the core of the proposed model	It can accurately evaluate the degree of abnormal; the cost can be reduced by the architecture of integrating modules	The data monitoring process on autonomic computing and data mining algorithm based on the chaotic algorithm and abnormal behavior detection based on Poisson	2018
[10]	The sharing of data between a mobile system and cloud system is discussed in this work; the AES technique is used to encrypt the high definition (HD) video; different keys are used to make secure the system are public, private, and security keys (PUK, PRK, and SK)	The delay minimizes due to utilizing the computational power	AES, HD video, HEVC, PUK, PRK, and SK	2017

data transmission between the user and the service provider. A key generation process to encrypt and decrypt the data has been applied which is valid to perform this job. A pair of keys with a password or PIN is generated to completely perform the job of encryption and decryption of the data between each the user and the cloud. The load balancing is a technique used to balance all the load of the server on the nodes and gives all resources to nodes, which minimize the time response. The solution of overload, under load or dill server, is presented in the proposed model. Dynamic load balancing has four main steps. In the first, the transfer strategy is discussed to transfer a job from local to the remote node. In the second, the selection strategy is discussed to choose a perfect processor that performs well according to the input job. In the third, the selection of destination is introduced. In the fourth, all the information of nodes is collected in this stage. The cloud is arranged in a cluster form and the cluster is arranged in a hybrid form, i.e., hierarchical and distributed manner. The service array of each cloud provider is zero initially. When a user wants to avail the services of a cloud, then it requests to the cloud. Then, the service array checks by the cluster head if it meets the user array or not. If yes, then it can avail the service, otherwise it avails the services of service provider. This checking of service array repeated until the CSP is selected for the user. The authentication server has the password of the users which are particular to service providers. The client enters the authentication password which perfectly matched; hence, the cloud service is open for it and the data is made secured by the user. The proposed framework outperformed in throughput, overhead, fault tolerance, resource utilization, response time, and scalability.

The data stored in the cloud platform may be affected by the attacks; therefore, data monitoring is a necessary process. In [37], a cloud data monitoring system is proposed on the cloud platform. It monitors the cloud data whether it is abnormal or not and then analyzes the security of data according to the monitoring results. The proposed approach mainly focuses on the security of the cloud data; for this purpose, the approach follows different steps. The approach

proposes a model which efficiently and safely monitors the cloud data on time. The system adjusts the monitoring system in such a way that it automatically protects the data. The approach proposed a mining algorithm in which an improved-based chaotic algorithm, data mining method is proposed for the appearing of abnormal data in the cloud. To obtain the accurate test results, the approach also designs abnormal behavior detection based on the Poisson. All abnormal behavior monitoring data security implemented on autonomic computing. The model is mainly composed of five different modules called the NMM, the data analysis module, the response strategy module, the system implementation module, and the knowledge-based module. The NMM is used to gather the data of the system by collecting the data stream and generates the original data. The processed data is evaluated and extracts useful information by the data analysis module to determine whether the data extracted is abnormal; then, this data is fed back to the response strategy module to adjust the monitoring period. The core of the proposed approach is the data collection and analysis of storage. These two provide essential data monitoring information. In short, the idea of abnormal data monitoring and autonomic computing system is proposed. Then, the data security monitoring method based on autonomic computing is proposed. By doing this the changes of data in the cloud are monitored to ensure security. The simulation results show that the cost can be controlled and reduce with the architecture of integrated modules of collecting data, analysis, monitoring service, and volume-based monitoring cycle adjustment.

In [10], the sharing of data between the mobile and cloud in the secure lightweight, robust, and efficient schemes is presented. The media files such as video, audio, and images can be uploaded to or from wireless servers. The transferring of data between mobile and cloud should be authentic and free from risk. Previously, all presented schemes regarding the sharing of data between the cloud and mobile have the limitation of memory support, processing load, and data size. This framework considers the HEVC with IEVS for data hiding. The AES technique is used to encrypt the data in the proposed framework. The analyses of the work are to implement this model in real-time processing in such a way that the energy cannot be consumed more. A lot of abundant information is contained by high definition (HD) videos; the intradomain is used to encrypt the video as a result of which more compression occurs, and it also provides an extra slot for abundant to hide the secret data. The cryptographic approach is used to encode the HD videos; it creates the overload on the input video sequence in execution on the mobile devices. The encrypted data are then shared on Pb which is semitrusted for downloading the uploaded videos; the public key (PUK) and the private key (PRK) are provided, which will decrypt the video instead of encoding. The decryption is performed only by an authentic user who has SK. An extra feature has been introduced in the proposed framework in which the user mobility can be traced by authenticating it remotely. The PUK, PRK, and SK do not acquire any synchronization at any stage. An uploading user is responsible to encrypt and upload a video and must be announced to all the other authentic users that a new video has been uploaded. It is a choice of the user whether he wants to download and decrypt the video or not; the video can be decrypted by the SK. The proposed work performs his job in an efficient way in the Pr, i.e., the model saves the computational power in a cloud system while the decryption of the data performs. It is not necessary for the presented model that the receiver must have the same computational resources as the model has for the encryption process. The simulation results show that the processing time decreases up to 4.76 percent, correspondingly the data size approximately increases up to 0.72 percent, and the proposed framework applies to real-time cloud media.

2.6. Application-as-a-Service. Table 8 shortly summarizes the application-as-a-service. The abstraction layer system and method used to secure cloud computing is presented in [38]. The development and deployment of at least one software workload for a virtualization environment are presented. Through the software policy, for a workload associated with a metamodel virtualization environment, a security zone is defined by the developer. And then apply at least one type of security policy with respect to the security zone including all types of security zone policy in the metamodel. In such a way, the security zone policy can be associated with the development of the software workload. If the security zone is related to the software workload, then it automatically applies the security policy when the software workload is deployed within the security zone. The preference for an infrastructure environment is decided by the software development life cycle. This includes the definition of security, policy, and management. Unlike previously presented algorithms, in the proposed work, a user can plan a cloud, build a cloud, publish a cloud, service by consumption, or run the cloud computing service. The best fit values can be selected between the internal and external service providers. Several modules are used to perform different actions on a cloud system to ensure security. Consumption module 32 is used in collaboration and to access the cloud service published for consumption. Module 26 called a manager module is used to configure one or more clouds for services or computer workload to monitor the cloud.

The connected vehicular cloud computing (CVCC) a hybrid technology used for the security and privacy of the cloud model involves computing resources such as the cloud, roadside infrastructure, and vehicles presented in [39]. Unlike the traditional cloud, the proposed CVCC vehicular resources include applications and possible services when the cloud and roadside infrastructure are not available. On the security and privacy challenges of the CVCC, researchers pay attention to the advanced cryptographic primitive technique. In this technique, pairing computation consumes more time; therefore, it is good to use outsource pairing computation for the vehicles. The primitive is built on bilinear groups called G_1, G_2, and G_T. The first and last groups are additive cyclic and multiplicative cyclic, respectively. In CVCC, the time

Table 8: Application-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[38]	The visualization environment is used to update the development and deployment of a software workload; different zone security policies are adopted by the metamodel framework; the SDLC used to find perfect infrastructure environments	Highly improves the security of cloud computing	The SDLC and different cloud modules are used, i.e., builder module, consumption module, and manager module	2018
[39]	Three different entities involving cloud, roadside infrastructure, and vehicles with additional work of advanced primitive cryptographic primitive make the connected vehicular cloud computing more secure and confidential	The service of the CVCC provides more reliability to the cloud users due to the availability of the service even in the absence of cloud and roadside infrastructure	Advanced cryptographic primitives, outsource pairing computation, and rich applications	2018
[40]	Different algorithms and steps are proposed in cloud security; among them, one algorithm is proposed here; the security uses service level agreements; in this work, two techniques called the QPT and QHP are used on the cloud system	This proposed work ensures the security of the user's data at different levels	The SecSLA, QPT, and QHP are applied	2017

consuming in the bilinear groups is the pairing computation for the vehicles. Three different kinds of entities in CVCC are used called the cloud, the roadside infrastructure, and the vehicle. Among these three entities, the cloud has the most powerful computation capability. At second, roadside units (RSUs) have the powerful computation capability. And these two entities act as a server in pairing outsource services. The cloud is used as a helping model for the rest two entities such as the vehicles update their system through the cloud. The roadside infrastructure has many RSUs, and the roadside infrastructure provides all the services to the vehicles. The proposed framework has two kinds of pairing outsourcing models. Among these two, one is is outsourcing the pairing computation with only one server and the other is with two servers. In CVCC, every server vehicle acts like a malicious. Hence, the second outsourcing model not completely fits CVCC. Compared to the traditional cloud, computing the CVCC is superior in developing secure cloud data center model and Big Data concept analysis. According to the input properties, the pairing computations have seven types, but the proposed framework is related to only one amongst the seven.

The economic and technological benefits are not attractive in front of the security issues of the CSP. However, different steps or algorithms and software are developed to minimize security issues. One of them is SecSLA. In [40], the two-state of art security evaluation techniques are called the QPT and QHP. The QPT and QHP techniques are applied to cloud systems to provide quantitative- and analytical-based security. These techniques provide flexible security to ensure users using CSP. The extension is carried out in the proposed framework regarding the state of the art and standardization. The QPT and QHP techniques are empirically validated through a couple of case studies using real-world CSP data obtained from the CSA. The limitation and advantages of the QPT and QHP can be analyzed by experimental validation of these algorithms to further guide the adopters. The visual security can be judged by the users in CSP through the prototype of decision-making security. Overall, the system security is in the satisfaction category judge by the users. The techniques provide security insurance at different levels. However, the security evaluation is not in an end-to-end domain.

2.7. Platform-as-a-Service. Table 9 is provided for the ease of this classification component, which summarizes this component in a short way. The two heterogeneous tasks are presented in [41]. These algorithms are cost paying and the users should notice the cost while renting the virtual machines from the cloud data centers. The workflow is heterogeneous and needs different instantaneous series of computing, memory, and storage optimization. The proposed work is based on different optimization techniques called metaheuristic optimization technique, PSO, and the coding strategy. Through the PSO, various tasks are mapped corresponding to the virtual machine in terms of pricing. These optimized techniques minimize the cost of the total work-flow execution while providing the desired deadline and risk rate. The cloud features include dynamic provisioning and the unlimited heterogeneity of computing resources along with the hourly pricing model. To meet the desired requirements, the scheduling and resource provisioning methods are used into MDMCO problems. The coding strategy for PSO workflow scheduling has been applied to solve MDMCO problems. The practicality of the proposed model is using the three different real-worlds and workflow applications (LIGO, SIPHT, and Cyber-Shake)

Table 9: Platform-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[41]	The risk rate and cost are analyzed using three different workflow techniques using cloudSim simulator; the workflow needs a periodically based series of computing, storage, and memory optimization	This framework provides risk minimizing according to the cost for a user in the cloud	The MHA, the PSO, and the coding strategy	2016
[42]	The authors propose a new framework to provide managed security services via SDK; different techniques are involved in outperforming the model; the BT works to provide security to data for Pb and Pr in a multicloud environment	Through this model, a user can secure its data for different cloud services such as private, public, and hybrid	The horizontal service model, BT service store, and managed security service obtained through SDK	2016
[43]	For the data security purpose, encryption technique is applied in this presented work; the encryption approach is applied using AES	The algorithm provides strong security for data in SaaP	AES algorithm is applied to the cloud system	2018
[44]	Three best strategies are applied to achieve the best response in data security and openness services; then, between these two, a relation has been drawn to help the cloud provider to choose the best strategy for both security and service	This proposed work decreases the investment for the security purpose and puts the cloud user in optimal service openness and best security environment	Specific probability with macroanalysis, best strategies for investment in terms of security, and Nash equilibrium	2019

on the cloudSim simulation to demonstrate the effectiveness of the risk rate and cost.

Not only the security of the cloud is important but also financial discipline has the main role in SaaS and PaaS. In [42], the horizontal service model is designed to investigate how the managed capabilities migrated as security applications via software development kit (SDK). Through this approach and its associated SDK, the customers are allowed to implement and enforce different security policies to a Pr, Pb, and HCs. The reassurance of business continuity by protecting the application level of Iaas and Paas has been achieved by the cloud computing security model. This model provides the high security model with the structured methodology that the customers will like our cloud applications. The model provides CEPM along with the information technology (IT) cost optimization to the security budgets. The BT service store is a web portal that contains applications and where the customers can configure a Pr and Pb. BT is responsible to make security at the customer's end. BT has a set of security solutions via the horizontal model. In the proposed model, the security enforcement issues and compliance are demonstrated and show the elasticity, load balancing, and indeed security. The integration of the service store will be carried out through SDK. However, much software are assumed to be present in SDK because HS is accessible remotely, and there is a separation between management and enforcement configuration. The SDK involves the capabilities of deployment, enforcement, disablement, and removal. The proposed framework is flexible and can support any kind of security solution regarding any cloud platform through SDK.

In [43], the Heroku cloud is considered the service as a platform (SaaP). The dyno app is responsible to run the Heroku and this app is like the heart of Heroku. It supports different types of programming languages. The strongest is the cryptographic approach called the AES which is applied. The AES is a concern with the security, speed, and symmetric key algorithm. The Heroku contains different steps that make the data secure in SaaP. One step is the replacing of each byte with the byte of the substitution table. The second step is the cyclically shifting of row towards the left. The third step is a fixed polynomial which is multiplied by each column.

A complete security control and complete service openness are presented using Nash equilibrium in terms of investment in service, and security by cloud system is presented in [44]. Two assessment methods based on quantitative analysis have been accomplished for the investments. Amongst them, one is for security and the other one is for service openness, which are discussed. Then, a relationship has been drawn which helps out the cloud provider to make a decision: the best strategy in terms of both service and security. However, openness brings security problems in terms of illegal benefit. To make sure the security investment increases which helps in improving the security detection technology, there should be a balance between these two, i.e., investment and security. Therefore, this proposed model decreases investment and security investments. Three optimal strategies are applied to meet the desired response in terms of best service openness and strong security which are provided to cloud users. Firstly, the macroanalysis has been carried out with specific probability. Secondly, the best investment strategy is accomplished to

meet balance investment for security and services. Thirdly, the Nash equilibrium point is calculated to fulfill the optimal service openness and security conditions.

2.8. Integration-as-a-Service. In order to shortly analyze this classified component, Table 10 is provided. The CCAF security suitable for business clouds is present in [45]. Three major security technologies are developed and integrated with CCAF called firewall, IM, and encryption based. All technical issues in a cloud system are analyzed here, and the model is a business-based cloud. The framework provides its explanation with the help of three examples. In the first, the framework gives a solution for bioinformatics and cloud storage. All structured query languages (SQL) are blocked, which protects data in real time in CCAF multilayered. The bioinformatics service can simulate DNA, proteins, genes, tumors, and many other organs of the humans. In the second, CCAF is also used as a guideline for financial modeling; the price and practice are changed according to the risk. In the third, the model investigates the hacking methods as a part of prototype requirements. In CCAF 1.1 version, different techniques are proposed, namely, security policies and recommendation techniques, and the technologies are updated and emphasized. In this framework, mostly advanced computational techniques are discussed and used for the calculation of the risks and the volatility of the market. The proposed concepts are essential for big data in a cloud system. The backup of thousands of terabyte in size is delivered to storage services. The framework provides security assurance to all incoming and outgoing data to cloud systems which are based on thousands of virtual machines. By using this model, huge datasets can be processed in a cloud system. The simulation results show the viruses and trojans are blocked and detected up to 99.95 percent and in continuous 100 h attacks; the blockage ratio of the viruses is 85 percent. The value obtained from the detection and blocking of the virus and trojan is 0.012. The quantity, quickness, and variety for the big data are beneficial in the proposed CCAF. However, this model is limited to verification, encoding, and customer with license access.

In [46], to manage more than one virtual machine (VM) connectivity through a data center, a software stack called industrial technology research institute (ITRI) with security is proposed. In this work, trustable security inside the environment of the cloud is provided. Different modules and components are used to meet security requirements. The system implemented data volume isolation, role-based distributed firewall module, SLA-based traffic shaping, address resolution protocol (ARP) spoofing, and the DDoS to reduce the attacks. Furthermore, web applications' firewall protection, web application firewall (WAF) and lightweight directory access protocol (LDAP), is also used in the system. To provide security to a cloud system using ITRI cloud OS, virtual data center (VDC) isolation, role-based distributed, firewall distributed, WAF protection, and SLA-based distributed shaping are designed, and then implement them inside the cloud system. In VDC isolation first, the media access control layer isolation mechanism is performed which supports the isolation in the multitenant environment. In the second phase, an algorithm is designed, which isolates the volume of the data on VDC. In the role-based distributed firewall, different policies for the security issues are implemented to access control between the networks which in turn protect the VM from high Internet traffic. The proposed work efficiently provides security to the cloud network. The experimental results show, the bandwidth between two virtual machines dynamically shares and is approximately equal to bandwidth allocation.

In [47], the BC technology is used to tackle the security and performance issues in distributed systems. Cloud computing took advantage of the BPDPP services. The functions required for two services assured data derivation and distributed assets. The tamper-proof environment can be achieved by the BC (s) mechanism where the set of authentic minors is used on digital assets to verify the users. Moreover, by using strong cryptography, method block of the transaction is chained together to enable unbeatable records. To achieve vulnerability in BC and provenance in the cloud, the BWH attack is applied in a BC by considering the pool reward mechanism. To add successfully a BC, there is always a need for solving crypto-puzzle in miners which are hard tasks in the computational domain. Therefore, the crypto-puzzle is costly in terms of power and hardware etc.; due to these reasons, the honest miners applied in the pool. A well-known scheme called BWH, in which the malicious pool members joined for truthfully mining block, actually never published any mining blocks before. Hence, the revenue of the attackers in the pool decreases by withholding the valid blocks, but its own reward increases by submitting many shares to the pool manager. The vulnerability in BC cloud arises due to computational power required to obtain the PoW based on consensus; therefore, the BC is implemented on rely PoW to obtain the consensus. The BWH attack occurs in the BC cloud during the pool mining to identify the constraints on the attacker's hashing power to defeat the purpose of pool mining. The simulation results show the attacker's access used more computational power to disrupt the honest mining operation of BC. The strategy of the attackers realizes in two pools where reward schemes are different. The PPLN scheme is useful in keeping the impact lesser of the attackers than the proportional reward scheme.

The trustworthiness parameter of the cloud provider has been improved by using an enhanced QoS-based model in [48]. In this work, the accumulative value of the trust is obtained by updating dynamically the transaction after a specific period. The trust is concluded after analyzing the current status of the transaction. In order to find the user's feedback mechanism as well to find the data credibility, the covariance mathematical technique is used. The best cloud provider has been chosen by the user in this proposed model. For individual cloud service provider, the trustworthiness could be found by lative or computed trust value (ATV). Different SLAs deal with the availability, reliability, data integrity, and turn around efficiency. The SLA parameters include truthfulness, security, and honesty. The resource

Table 10: Integration-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[45]	The multilayered framework proposed is used to integrate and develop the three security technologies; the bioinformatics service can simulate DNA, genes, tumor, and organs of the human body	It ensures the security and safety of all incoming and outgoing data to the cloud; large datasets can be processed on a cloud system and are beneficial for the velocity and volume of the big data	CCAF version 1.1, firewall, identity management, and encryption structured query language (SQL) are used in this model	2016
[46]	This work proposed many modules and algorithms which are implemented at different levels to provide security to the cloud network; different security policies are designed and implemented according to situations	It provides the ability to share the bandwidth efficiently between the two virtual machines	VDC isolation, HTTP, VMs, role-based distributed firewall, and distributed WAF are implemented	2015
[47]	The BC technology used cryptographic enforced distributed ledger system for the security assurance in the cloud; it guarantees the data provenance and the vulnerabilities in the cloud; the BWH attack in BC like the distinct pool reward mechanism	The mechanism used for the security of data is highly strong, and for the attackers, it is difficult to attack because attackers must use extra computational power	The BC technology, BWH, crypto-puzzle, and PoW	2017
[48]	The trustworthiness of the cloud data is the key parameter along with the user credibility feedback; basic assessment model, Qos-based trust assessment, and covariance mathematical models are used to cloud data; the running time, trustworthiness, and user feedback are calculated	The proposed model is best in terms of users' data trustful, confidentiality, and consuming less time	Qos-based trust model, mathematical covariance technique, and basic assessment model are used with cloudSim simulation	2020

availability means users can access the cloud sources at any time. SLA is an agreement between the cloud provider and the user. In short, the basic assessment model, trust assessment model, QoS-based trust assessment model, and covariance mathematical model are used to find out the best cloud provider amongst many and evaluate the credibility of the user's feedback along with the security, reliability, and availability of the data. The cloud Sim platform is used to simulate the best response.

2.9. Security-as-a-Service. Table 11 tells about the references, overviews, techniques used, advantages, and the papers publication years for each research article included in this classification. To build trust in cloud computing is a difficult and complicated task due to the distributed, dynamic, and nontransparent environment. The proposed work trying to win the trust of the users in cloud computing by introducing new methods identifies the fake feedbacks [49]. The one is feedback evaluation and the second is the Bayesian game model. There is a direct relationship between the feedback and trust values, i.e., the attacked feedback has inaccurate value and vice versa. The feedback evaluation model is used to analyze fake feedback. The model indicates the fake feedbacks on the previous average feedbacks. The trust average value increases by rectifying fake feedbacks in comparison with the after fake feedback injection. The task of the feedback evaluation is to identify and rectify fake feedbacks. The malicious users are identified by using the game-theoretic approach. The feedback received by the malicious users is considered as the fake feedbacks and their feedbacks are prevented and not input for the further process to predict the trust. The component evaluates and updates the received feedback from the CSU after receiving service. It qualitatively identifies and rectifies the fake feedbacks. This prevents the circumvention, collusion, latency, and impersonation attacks. The second model called the Bayesian model is used to detect malicious users and prevent their fake feedbacks. A CSU requests the service from a CSP. The game will end if no suitable CSU is found, else the game continues. The trust values of the CSU are calculated using the received feedback. The Bayesian game model is deployed in a fuzzy logic approach. The Bayesian game is between the CSPs and CSUs. Each player has secret information, but it is not shared with any other player. There is no joint comparison between the payoff and cost while the Bayesian model is presented based on cost and payoff. The payoff in a Bayesian model is a qualitative measurement while the cost is a quantitative measurement. The simulation results show the proposed model correctly identifies and rectifies the fake feedback by the feedback evaluation method. The analytical results is matched with the Bayesian model to correctly recognize the malicious users and concluded that the feedbacks received by the malicious users are the fake feedbacks. To prevent the fake feedbacks in a Bayesian model which is due to the strong mathematical model, in this model, a variable delta parameter is used which

Table 11: Security-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[49]	Two new models are proposed to stop the fake feedbacks by the malicious users called the FEBM; the previous feedback average is considered to indicate fake feedback	The variable delta decreases the positive and negative false error with high accuracy	The delta variable factor, the Bayesian game model, and the feedback evaluation	2017
[50]	The attack-defense game-theoretic approach called Stackelberg game is proposed; the players in this game are called the defender and the attacker; the strategies of the defender are open and the attackers follow these strategies; the equilibrium point will be found by the active and passive structured	The Stackelberg model with active and passive structured is used	The attackers achieve the maximum gain of the defender	2018
[51]	In the proposed approach, an integrated solution to cloud security based using the clear framework and the BPMN is discussed; three-layered models are analyzed for the strong security and blocking of threats	The multilayered CCAF security model has 20 percent better performance than the single-layered security models which can block 7348 viruses and trojans; a quick locking system is achieved which can block and quarantine the 9919 trojans and viruses in quick response	The CCAF, BPMN, 10,000 trojans, and 10 PB data in the data center	2016
[52]	The security issues are analyzed on Security-as-a-Service (SecaaS) in this work; a new pattern called leveraging is applied to SaaS; it gives self-managing, automating, and scalable to SaaS	The simulation results show that it outperforms regarding security; the security of the SaaS improves by the cloud-native application	The CNA is used	2017
[53]	The proposed work in this article provided security to a Pr system; the whole work is divided into three steps to provide security to big data; two types of scanning obtained called vulnerability scanning and log scanning which then are correlated to each other to find the attacks on big data	The experimental result shows the attacks count, host computer used, and security tips count to guarantee the speed of analysis	Nikto scanning tools with Nagious and conical correlation are used	2016
[54]	The updated chain VM service proposes to handle the high traffic input to the chain; the halts and deadlock option provide security assurance; the repetition of the previous input data block in new updated chain VM	The technique achieves the increase in the percentage of the security and upgrading and optimizing the security; it configures and runs the desired security stack	Seamless flow, dispatcher VM, and SNAT docker container	2018
[55]	In this model, co-resident attacks encounter instead of looking to the solutions of the attacker after co-locating with their targets; the probability of the attackers co-locating with the targets mitigates in this approach	The cloud Sim and open stack simulators are used which shows the attackers first need to co-locate their VM according to target VM and the attackers achieve hardly up to 40 percent	The virtual machine allocation policy, the PSSP policy, workload balance, and low power consumption	2017
[56]	The approach presents different models to detect and track the already existing threats in the database and new incoming threats to a cloud system	The simulation results show the framework efficiently detects the anomaly security system up to 90%	The signature method, intellectual model, big data technology, and Weka application	2017
[57]	The covert channel analysis performs to secure data at multilevel which enables to secure the data in the presence of unauthorized personnel; different steps follow the data to achieve security	The work protects the data in the presence of an unauthorized person who achieves the trust of users	Covert channel and prototype approach is developed	2015
[58]	The assessment framework proposes to solve a security threat related to the specific client; in this model, different six threats are analyzed, find the concerned client for each threat, and give a secure environment	The model is not fixed to any specific network but can be fitted to any system; the model solves the problem of previously existing assessment framework problems and finds a threat for the concerned client	Spiral network and STRIDE categorizing model is used	2018

decreases the false positive and false negative errors. The feedback trust values distribution is compared in three states called before fake feedback injection, after feedback injection, and after rectification. In short, the Bayesian model is responsible to identify the fake feedbacks with higher accuracy.

The useful decision-maker technique in the attack-defense called the Stackelberg game is proposed in [50]. Two kinds of players are characterized in Stackelberg game called the defender and the attacker. The strategies of the defender are defined in advance and the attacker obeys them. The roles defined for the competitive players called the defender and the attacker is the natural roles. The cloud defender can be a cloud provider or the administrator of the cloud system. The attacker can be any kind of advanced security threats or hackers. The game-theoretic approach gives the tools needed to analyze the behavior and the strategies dependent on payoff functions. The defender wants to minimize the cost during its protection by enhancing the availability of the attacked CI. While on the other hand, the attacker wants to maximize the damages which minimize the cost of the attack itself. The game model formulation includes the player's actions and utility functions. The game structured in two stages: a passive and an active one. By using these two stages, the estimation of the model parameters and the game equilibrium point would be found easily. The equilibrium point can be found by finding the best defense strategy. According to the proposed game model, the defender can rationally choose the right strategy to incorporate the attacks in a proposed way. Besides, the potential cloud attack scenario is modeled between the attackers and the cloud providers as a nonzero-sum SSG. As an output, the attackers achieve a maximum reward which then enhances the defender gain. In short, the proposed game model defines the strategy which maximizes the reward and gain for the attackers and defenders. The empirical analyses prove that, by the security attack management, the attacks prevented the cloud service and data. In this model, a profitable strategy is chosen under a certain attack. The active and passive stages of the SSG provide information about the attacker's behavior.

The CCAF multilayered security model is proposed in [51]. It is believed that the cloud security is only achieved by such an approach that is systematic, adaptable, and well structured. The components, rationale, and overview are explained in the proposed framework to protect the data from different threats. The huge data exist in the data center up to 10 petabytes (PB) which is difficult to protect in real time. The BPMN is used to simulate the data. The CCAF multilayered security is applicable to protect the real time data and consists of three different security layers called the firewall and access control, identity management, intrusion prevention, and, the last but not the least, convergent encryption. A total of 10,000 different trojans and viruses are analyzed in the experiment with two sets of ethical hacking. The proposed CCAF can block the 9,919 viruses and trojans in seconds while the one remaining is isolated or quarantined. The continuous injection of trojans and viruses decrease the blocking percentage of the CCAF, and, in such case, 97.43 percent is quarantined. To make the CCAF more efficient, the BPMN model is combined with the CCAF for the strong security and efficient penetrating testing results. The penetration testing and more other related experiments provided the robustness and precision measurement to the proposed model to justify it from the other approaches. The CCAF multilayer provides multiple protections and a suitable method with the help of which security of the data improves and handles the 10 PB in the data center. The BPMN model is used to understand how the data can be used in rest and motion state within 2 seconds. The time taken by the BPMN to protect the 2 PB is 50 hours; it means an integrated approach is required than the FGSM algorithm and is used for the injection of 10,000 trojans and viruses in the data center. Two different experiments are performed at this stage: one experiment shows the firewall, identity management, and encryption which could block the 54,233,742 and 842 viruses and Trojans, and the remaining 81 could be quarantined. While the other experiment shows the continuous injection of 10,000 trojans and viruses makes the blocking rate decrease from 99.19 to 76.00 percent in 125 hours.

In [52], the authors design a new pattern to improve cloud security. The new design is Security-as-a-Service (SecaaS); it is available just after the SaaS. Leveraging the cloud gives self-managed, automated, and scalable facilities. The developing and deployment of the native design patterns recently appears as an outstanding approach for the application of the clouds. The cloud-native application (CNA) improves the SecaaS. The CNA provides scalability to the cloud system. The experimental results show the better performance in terms of security. However, the design pattern of CNA is complex and has no defined steps for the requirement of detail planning.

In this [53], the authors try to find out an efficient mechanism to provide security to big data on cloud computing. For this purpose, the proposed work divided the design scheme into three main steps. This division includes a vulnerability scan, system log collection, and correlation analysis. This division detects the attacks, illegal approaches, potential threats, and many other security events of the attackers in time based on big data. For vulnerability scan, different tools such as regular using of Nikto and many more are used for the cloud system to carry out the detailed vulnerability scan report for regular network security. The system log collection uses Spelunk, Nagious, and many other tools to collect the detailed system log report. Correlation analysis, also called canonical correlation analysis, is used on the vulnerability and log scanning reports to find out the attacker attacks, corresponding to which the system will issue the warning in time. The cloud computing system has several aspects of security named as to prevent advanced persistent threats, user access control, integrating tools and processes, and real-time data analysis. The experimental results show the system took a total of 136 virtual machines. A Pr is formed by the host machines. The overall cloud system first scans with the help of which a detailed vulnerability scanning report is obtained. This report shows 136 host computers are used, 149 security vulnerability, 178

security warnings, and 497 security tips. After obtaining the vulnerability report in the second step, the system collects the log files for every minute. Then, through conjunction, the vulnerability and log reports are correlated through correlation analysis. The last minute is used to analyze the speed. The log system is used if the suspect is found; then, the log system decides whether this suspicion is an attack or not. The model cannot discover 100 percent attacks of the attacker.

The smooth update service for cloud-based security services is presented in [54]. The virtual machine (VM) executing in a cloud data center receives the network access request by the client on a remote trusted location to a nontrusted remote site and then route it on a chain of security services. The updated facility has by the VM in this approach, in which the VM transfers the network traffic from the initial chain to the updated chain seamlessly and updates the cloud service seamlessly. The halt and the deadlock action is performed on the previous version by the dispatcher VM after confirming the correctness of the updated version. A small test signal flows on the updated chain service to ensure the correctness of the updated chain service. High traffic is input to the updated chain service to analyze the traffic handling operation. The previous stage input data is the block in the updated chain service and checks to ensure that all the input data to the updated chain service are correct and transfer without any problem. The dispatcher VM waits for a specific period after the previous version stops to receive more traffic. The output of a chain VM is input to the next VM in a chain form. The presented work efficiently facilitates in providing the upgrading and optimizing cloud-based security service. The approach mainly performs to configure and can run the desired security stack using a platform called cloud-based security service. The customized and redundant security can be achieved by one or more cloud computing services. The approach efficiently increases the security but with the limitation of scalability issues, lack of customization, and depends on the single point failure. The approach is not limited to one or all the capabilities of firewalls, antivirus, and antimalware. In the experiment, different databases, clients, users, devices, appliances, and cloud-based storage are used and the security results with different ranges are analyzed.

The focus of [55] is on the threat co-resident attack, in which side channels extract the information from virtual machines (VMs) co-located on the same server which are focused. The approach makes many difficulties for the attackers to co-locate with their target by improving the VM allocation policies. The proposed work especially targets the model access attacks by defining security matrices. Then, the matrices of the model notice the difficulties of achieving co-residence under commonly used three policies, after which new policy is designed; this not only minimizes the threat of attacks but also fulfills the requirement of workload balance and power consumption. Then, it practically analyzes these steps by implementing them in cloud Sim and Open Stack simulators. A prototype security policy called PSSF was used earlier. However, the limitation of this prototype is workload balance and power consumption. This work targets these points (workload and power consumption) along with security to achieve high security of the PSSF model, which becomes more applicable in clouds. The MMO techniques are used to improve the PSSF. Firstly, in the proposed work, secure matrices are defined which measures the safety of the allocate policy of the VM which defends against the co-resident attackers. Secondly, these matrices are the model under three basic commonly used VM allocation policies. Thirdly, a new security policy that decreases the probability of the co-locating attackers along with the workload and power consumption is implemented. To avoid obtaining attacker's co-residence, a game-theoretic approach is used which compares three different commonly used VM allocation policies in terms of security, workload, and power consumption. There are two types of VM placements called initial placement and live migration, while the applied model is based on the initial placement. In short, the framework shown before the attackers can extract any private information of the user, and the attacker first needs to co-locate their VM with the target VM. The results show that the simulation was performed on cloudSim and Open Stack; the attackers can achieve the co-resident efficiency hardly up to 40 percent.

An anomaly detection system to secure the cloud environment is discussed in [56]. The wide range of memory is used to store the data in the cloud system which creates many security threats that are uneasy to solve. The cloud computing services have many advantages over the traditional systems, but the cloud systems lack in the concern of security. In the cloud, all the data can be lost due to different security threats and hackers. Therefore, the cloud model needs a data center network that facilitates the model to access a large number of datasets and detects different security threats. Generally, the signature method follows to detect the threats in the cloud model which compare the incoming traffic with the database threats. However, by using this method if a new threat arrives and do not have the database, then it cannot be indicated as a threat. In this scenario, it is more necessary and efficient to use the intellectual method to detect the threats. In the presented model, both the detection systems are added, i.e., intellectual can handle new incoming threats while the signature method used to handle the threats already has by the database. To process the data in a dynamic mode and with speed in big data, different methods and tools are used, i.e., DFS and parallel computing on many servers. They achieve a secure environment, and the proposed model performs different tasks on a cloud system which includes developing secure cloud data center model, developing an anomaly detection system, and big data analysis. The data center model can eliminate the deficiencies of security. The accomplishment becomes possible due to the use of technology architecture, high-speed communications, and unified computing structures. The addition function is added to the data center model for the detection of anomaly secure environments. The Weka application is used which makes the anomaly detection model more secure and accurate. The anomaly methods are mostly used in the area of cloud computing environment, fraud detection in banking, mobile

areas, monitoring of information systems hardware, network's intrusion detection system, processing CCTV images, and suspicious websites. The simulation result shows the developing system provides the high percentage of up to 90 percent of the anomaly detection in the secure cloud environment.

In [57], the covert channel enables to communicate directly the cloud and users in the presence of unauthorized persons while not affecting the data. The protection of data is achieved by multilevel security using the covert channel. The security, feasibility, and performance of the cloud data are achieved by the prototype approach. Different steps involve securing the data while the attackers are present. In this approach, a predefined agreement table is signed between the users and the cloud. An extra fake 8 bits value is generated and transmitted over the wireless channel. The users receive the 8 bits extra value and original data, and the users matched and found out the accurate data according to the agreement able. In this way, the data are secured in the presence of a third party. The simulation result shows the system has acceptance time in cloud data. However, the approach generates extra 8 bits combinations which affect the cloud storage space.

The researchers have worked previously on the assessment approaches for conventional security risk, but they were unable to explicit the concerned risk with respect to a specific attack. In [58], the authors propose such a security model that evaluates the security with respect to a specific threat and having an assessment framework. In this model, the threats which are the concern to the specific client are solved. The risk assessments are mainly divided into three different stages and have a relationship with each other. The three stages of the risk assessment are elicitation, analysis, and control; these all are arranged on a spiral model. The elicitation phase concerns with the assets, vulnerabilities and security requirements, and threat and attack scenarios. The analysis phase is related to the value of assets, the impact, and the relevance of the threats by vulnerabilities. And the last stage but not the least is the controlled phase which describes the management of the vulnerabilities, mitigating risk, and recovery attacks planning. The process of solving threats and concerns with the specific client gives the quantification of security risk. More beneficially, the proposed model can be implemented on any network, and it is not fixed to any one network. The proposed model is called the spiral model and all the stages in this model have a dependency on each other. In this mechanism, the threats for the specific client can be found out and only the concerned client and security are taken into account. The threats are computed in their specific functionalities. All the security risks are taken as a common framework. To categorize the different threats, a model called STRIDE is considered, which has six different threats. The threats are spoofing (S), tampering (T), repudiation (R), information disclosure (I), DoS, and elevation of privilege (E). In short, this model is used for a specific threat which relates to a client that will evaluate to secure the cloud system. The model used has six types of different threats.

2.10. Management-as-a-Service. In order to view the summary of the Management-as-a-service, Table 12 is provided.

In [2], the authors discussed the security issues; they not only concern with inside the network but also relate with the outside of a Pr. As the security of the information is a concern with the third party. Different software scanners are used to ensure the security concern and manage it in a good manner. Three different scanners are used to look after the security of information inside and outside the network. The scanners used for inside the network security management are Metasploit, Nessus, and open VAS. While outside the network security insurance task is performed by using web App tenable IO. In all the network ports, usage information is noticed and a list made for all the ports, that is, which port is used inside or outside the network. Then, according to port (either inside or outside the network), different rules are applied to each port. An Internet protocol identity is provided to a virtual machine through any software tool (Metasploit, Nessus, open VAS, and web app tenable IO). According to experimental results, the Metasploit scanner outperforms than all the other scanners.

In [59], the authors propose the security information and event management (SIEM) architecture to protect the cloud from different intelligent threats. To analyze and recognize the intelligent cyber threats based on virtualization technologies, the SIEM architecture deployed to the Security-as-a-Service platform. The SIEM is an important component of a business platform and network infrastructure which can collect data, aggregate it, normalize it, store it, and correlate the data with the traditional security systems. The traditional systems deployed on the host and network domains are firewall, IDPS, and antimalware systems. The cloud-based service can be protected in several of security events by SIEM architecture. To manage and analyze different log events generated by the cloud security sensors in the security-on-air (SOA) projects, there is a need for SIEM regarding manage log, security events, and correlation analytic by recognizing cyber threats. The data enrichment can be achieved by adding Open SOE and are complemented by adding to the SIEM architecture. The main goal of the SIEM architecture is to provide valuable security and large data correlation to detect cyber threats. The presented model mainly exists by the SIEM engine to collect different data from the users; SIEM storage to store the correlated data, to ensure the security service SIEM user layer is used. The SIEM engine can support the threat analytic and execute it by the virtual machine. The data identifier manager (DIM) is used to recognize the data from various security sensors. The attacker name and traffic information are stored on big data carried out through SIEM baseline and log data frequency. The statistical database uses big data analytics, data mining methodologies by IP address, and the port for data enrichment from collected data and then provides the data in the correlated form to analyze and recognize the cyber threat; these all are achieved by SIEM architecture. The SIEM architecture correlates the collected security features with the aggregate threat datasets. Highly intelligent security performed by the SIEM architecture model on the cloud data.

Table 12: Management-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[2]	The inside and outside security of the Pr network is analyzed here; three different scanners are used to make sure the security inside and outside the cloud; for inside security, Metasploit, Nessus, and open vas tools are used; the outside security of the cloud is performed by web app tenable IO	It secure the data for inside and outside the Pr from hijackers, etc.	The scanners Nessus, Metasploit, open VAS, and web app tenable IO tools are used	2018
[59]	The presented model calculates the baseline values based on day, time, and log data frequency to show out the attacker's name and traffic information; the correlation performs in the model which makes the threat identification easy	The proposed model analyzes and recognizes the cyber threat by obtained correlated information; the SIEM architecture provides the engine to store data and storing the ability to store data	The SOA, OpenSOE, and complemented DIM	2017
[60]	The model presents the security of a cloud system; the quantitative evaluation provided security; the problems of evaluation and comparing the security level offered the presence of complex service supply chain	The presented model provides a well-defined set of security related to the acquired service	Reference evaluation model (REM), SecSLAs algorithms, and complex service supply chain techniques are presented to achieve security goals	2016
[8]	The proposed work ensures the users in the security domain, while they deal in a cloud network; the work enables the users to check and control the data at any stage and level; different models are applied to analyze the security and accountability	It makes the cloud system according to the demand of the users who can check and control their data at any stage	The PM in the cloud, RSASS system, data security model, and CIA models are used	2015
[61]	Various old policies of the cloud network security are compared with network security; besides, some additional new knowledge-based policies are added to cloud security, and then experimentally analyze and find the correctness and effectiveness of the proposed work	The experimental results proved the design policies stood well to old policies for cloud security	A multidimensional integer space conflict detection algorithm is proposed	2016
[62]	According to the job of the user and computer security provided to the cloud system in this framework, through GA-based job security, two new concepts are achieved in this work called security on demand and improved trust level	By using the algorithms and techniques in this work, the simulation results show the high-level security with the minimum cost of time	The GA-based job security technique is used	2015
[63]	Two security threats called timing attacks and DoS are analyzed and secure the cloud system from these threats; the security-aware and NDTMSI with the agent-based monitoring system is proposed	The proposed approach provides magnificence security without substantial modification and provides integrity and confidential and authentication monitoring to the cloud	The safe random number generator (BBS), secure generic scheduler, nondeterministic model, a secure hashing algorithm (SHA) scoring workers, security bias model, and leveraging genetic heuristic approach	2017
[64]	The hybrid cryptographic system (HCS) is applied to the cloud system to secure the data; in this work, symmetric and asymmetric approaches are analyzed to encrypt and decrypt the data	In this work, the system from start to end is secure which achieves the trust of the users until the data stores on the cloud	In the hybrid cryptographic system, asymmetry data encryption RSA and symmetry data encryption AES are applied	2017

The per-service SecSLA discusses in [60]. The authors show the provider and consumer reach an agreement point on the features of the security of each service instead of leased. In this framework, each customer fixed a different security level agreement (SLA) for each leased service. A well-defined set of security rules guarantees the acquire service with respect to providers. The reference evaluation model (REM) provides a quantitative level of security. In the presence of a complex service supply chain, the security offered by cloud providers involving resource acquisition from different CSPs. In this model to build up per-service security, SLA starts from currently available public repositories called STAR repository. The security control by the providers is used to fill a declarative section of respective security SLA; it identifies the security features provided by the CSPs. Subsequently, the combined security SLA is built by generating an enhanced consensus assessment initiative questionnaire (CAIQ) in the availability of a complex supply chain. The SLA can monitor the additional security capabilities introduced in the supply chain, and the security SLAs are compared to the REM technology. In short, the combined per-service security SLA is currently obtained by simply putting all together with the controls declared by the CSPs in the supply chain. The model outperforms in terms of security.

In [8], the authors describe the importance of security and accountability in cloud computing using different models. The researchers focus on security availability and performance. The top issue is security, and it is still important in the cloud, as the users have no control in the security section. All the security actions, planes, and policies on data are implemented by the CSP and the users have no interference. The modification and deletion of the data are performed by the users; in this case, the service provider must maintain the sequence and order of the data as data have previously. The role of the distributed protocol is important as the data is stored at different places due to which latency is generated. Accountability is another parameter in the cloud which ensures the users about the security of the data. Accountability enables the user to check security, transparency, and control. It brings the confidence; the user can control, check, and verify data at any stage according to its expectation. The log recorder uses to check and analyze the actions taken by the users are suitable or not. Accountability builds the trust of users on the CSP. Different models are used to ensure security and accountability in the cloud: PM in cloud, RSASS system, data security model, and cloud information accountability framework. After implementing all these models on data, the results show for the security purpose the best models are data security model, PM, and CIA framework. For accountability, the outperformed model is cloud information accountability (CIA) and the RSASS system. Overall, the CIA model is best for both (security and accountability). The main drawback of the CIA model is its larger size.

In [61], different old policies are discussed and then the new policies are presented for cloud security. In this work, the policy and knowledge-based comparison are made between cloud security and network security. A multidimensional space and sum up of a conflict detection method are proposed for cloud security policy, in which mapping of the security field to an integer multidimensional space set is according to the condition of field mapping principle. The experimental results are carried out through different central processing units (CPUs), which find the correctness and effectiveness of security policy collision detection algorithm on multidimensional integer space. The result shows the time computational of the proposed policies when the experiments are repeatedly carried out. However, due to the time dynamic of the proposed work this model is not stable in the time domain.

The job security problem of the cloud security is proposed in [62]. The proposed work divided the whole cloud system into four architecture-based layers called SOA architecture, management middle-ware, resource virtualization, and physical resource. The chromosomes and genes are used to functionalize the algorithm and are called GA-based job security. In this algorithm, each chromosome is utilized to show the schedule of a set of jobs on many computers. The gene is used to show the relation between jobs and computers. Many users and computers are considered in which each user has a task which is input to the computer; the proposed algorithm then analyzes the security level according to the job of a computer. In summary, two new concepts are analyzed in this framework called security demand, to improve the trust level of the users. The simulation result proves that the proposed framework provides a high level of security to a cloud system by consuming minimum time.

The DoS and timing attack prevention are the computational cloud challenges and presented in [63]. A novel architectural model is based on a multiagent scheme and security-aware nondeterministic metascheduler driven by genetic heuristic to enforce the cloud security. The proposed approaches prevent the cloud system from the DoS and timing attacks in an Open Stack platform. The basic security objectives in modern clouds are the preservation of confidentiality, integrity, and availability. The proposed work incorporates the first two issues in this article. The information security methods are used for the cryptography service designed and implemented for the theoretical model. To solve the above proposed security issues, the cloud architecture divided into multiple components to secure the whole cloud system. Amongst the total eighteen security control points, the proposed work encounters the access control, protection, audit and accountability planning, security assessment and authorization, risk assessment, identification and authentication, system and communication protection, incident response and system, and the last information integrity. The proposed model is like the resource management system; it supports the inside secureness of the CI by task distribution according to required security demands coming from the cloud consumers. To meet this security option, the batch scheduler leveraging a genetic heuristic approach is used. To monitor the task flow, characterizing the scheduled processing supports the multiagent system in cloud computing. By doing such security operation, the model is protected by DoS and timing attacks. The scheduling, monitoring, and reporting activities are

carried out in nondeterministic time intervals. Two additional models to secure the model more are proposed. One is called scoring worker model and the other one is called security bias. The scoring worker model states the virtual machines are used for the scheduling tasks only and are characterized by the security level at least equal to or higher than the demands of customers. The averaged time associated with security operation processing is added to the fee for the total runtime for each customer. Second, the security bias model considered the cryptographic operations associated with each task. The enlarging of time required for the task processing in a scheduling process is modeled in security bias. In short, the simulation results show the proposed model effectively increase safety without substantial modification. The model is dynamic and used for different schedule types. This property of the model ensures proper security. The elements which are incorporated in scheduling are a denial of service (DoS) and task injection prevention and integrity monitoring and authenticity checking with a standard.

The CSPs increase their demands day by day, but the security of the data in the cloud is still a question mark. In [64], the authors propose an ecosystem in which different techniques are applied to secure the data. Many researchers propose different algorithms and techniques; most of them concern with specific threats. However, the proposed framework designs such techniques and algorithms which secure the whole system from the start to end. The hybrid cryptographic system (HCS) is analyzed in the proposed task. The focus of the work is in the encryption and decryption of the data efficiently. The cloud environment is secured by analyzing HCS for the symmetric and asymmetric encryption. The rehabilitation services administration (RSA) and AES algorithms are used to perform symmetric and asymmetric operations to secure the whole system until the end. The RSA algorithm generates the keys privately and publicly which are later used for the encryption. Different parameters should be noticed while analyzing the user data which are data protection, traffic hijacking, isolation of resources, and malicious insider. The hashing and salting techniques are also applied to the ecosystem to strengthen the encryption process at different stages. The proposed framework outperforms to secure the whole system. This achieves the trust of the users, and the data are secure until it stores on the cloud.

2.11. Testing-as-a-Service. To minimize the time consumption for the researchers the whole summary of this classification component is presented in Table 13.

In [10], the authors analyze and test the security issues of an online education system. The security issues are highly important to secure the data of a system. The online education system is mostly popular nowadays; therefore, hackers and attackers may want to expose and leak the data. The data store in the cloud is not in the control of the user. The App Scan tool is used to analyze and test the security of the education system. The main working principle of App scan: first, it finds the web with the help of which the directories and site parameters can be analyzed regarding security. Second, a modified HTTP tool is used to check the attempt attacks on the data on the basis of which rules and arrangements are implemented by the security responsible authority. At last, it tests the overall education network by the load-runner tool. According to experimental results, high risk issues are analyzed and resolved through app scanner after resolving no higher-level and middle-level risks exist.

In [65], the MapReduce framework for cloud computing is provided. The MapReduce currently is the most popular and dominant programming model. Therefore, it is essential to protect the probity of the data processing of the MapReduce model. The malicious workers are categorized into collusive and noncollusive; the workers try to degrade the results to find the easiest way to attack cloud computing. The important task is to find out malicious workers in an efficient manner. In the proposed framework, security protection is designed to find the malicious workers. The proposed security system consists of two mechanisms called credit-based replicate tasks and verification which can be combined to behave efficiently. The integrity in the map phase of the MapReduce is obtained in this work. MapReduce is a programming model and generates large datasets. The jobs of the users are submitted to a scheduling machine. The input datasets are broken into small independent pieces from 16 to 64 MB in size by the MapReduce model. Then, the chunks are managed by the map task in a parallel way. The outputs are sort by the proposed map which then inputs to reduce the tasks. Both input and output jobs are stored in DFS. The proposed MapReduce is divided into two phases: one phase is the map phase and the second is the reduce phase. The workers in the map and reduce phases are called the mappers and reducers, respectively. In the map phase, the input datasets are divided into small chunks. And in the reduce phase, the intermediate results are distributed to the reducers to reduce the task associated with reducers. When the reducers receive a reduce task, then they wait for the notification about the map task completion and then receive intermediate results as their inputs. After this, the reducers input their data to FDS and then the reducers notify the master that they return the results to the user job. To detect the malicious workers the framework consists of the replicate task and verification approach. When the task queue picked by the master, it will be forwarded to any two workers. The two workers execute this task which is called the replicate task. By obtaining the two results from the two workers, the master compares these results if they are different; then, it means one of the workers is a noncollusive worker. To detect which one worker is noncollusive, the improved replicate task which is a credit-based replicate task is proposed. The simulation results enable the users to detect all the collusive workers with different malicious ratio. When the quiz threshold is set to 6 the overhead of the proposed model was no longer than 2.5, even the malicious ratio equals 0.5. The response time is almost 3 percent longer than the traditional models. In short, high security with malicious workers is obtained for cloud computing which is a sign of stable and efficient.

Table 13: Testing-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[10]	The online education system is made secure in this work; however, different attackers and hijackers want to leak and expose the data; the App Scan tool is used to make secure and analyze the data; however, a modified HTTP tool is used to find the attempt attacks	In this work, the security of the cloud is arranged according to attempt attacks	App Scan and modified HTTP are used	2018
[65]	The novel approach called MapReduce is the programming model which is presented; the big datasets are divided into small pieces and then input to the scheduling machine; the whole framework is divided into two phases called map and reduce phases	The overhead mechanism is almost eliminated with high accuracy and in an efficient manner	The MapReduce programming model, malicious workers (collusive and noncollusive), protection framework, credit-based replicate tasks, and verification	2016
[66]	The optimal decryption method has been proposed to encrypt cloud data; the RRA-CPA-based PKE has been improved to avoid random attacks on the cloud data	This model presents the optimal result in terms of avoiding the random attacks on cloud data along with the strong decryption	IND-CCA secure PKE scheme and RRA-CPA-based PKA scheme have been presented	2020

In [66], the different random attacks on the data encryption techniques have been discussed. The encryption techniques include RRA-CPA and have optimal decryption algorithms. In the PKE scheme to avoid random attacks, an improved version has been discussed of the PKE scheme which encounters the different random attacks. For the purpose of encryption, the information, firstly, the hardcore function has been applied to gain the output; then efficient decryption and cipher text size have been applied to this encryption. This proposed model efficiently enhances the decryption.

2.12. Infrastructure-as-a-Service. Table 14 explains the short summary of the Infrastructure-as-a-Service.

In [67], the software-defined network (SDN) and information-centric networking (ICN) are analyzed in a cloud model. The purpose is to put both SDN and ICN in a cloud and to ensure the security of the data. With the help of these flexible infrastructures and stable networks, performance is achieved in the cloud environment as per service level agreement. The ICN is considered the building block of the proposed framework because of the following reasons: the architecture related to ICN is more efficient than the IP-based networking and the host-oriented communication model by default replaces to a content centric model. The information is accessed by the identifier rather than the host in the ICN, whereas the SDN decouples the control plane in an efficient manner which is responsible to make decisions to route a packet and processing of the packet to the desired destination. This property of the SDN administrates the network to provide services of routing, policy enforcement, and the last security. The proposed framework provides autoconfiguration to topology and policy of the network so that a user faster avails the cloud service. Through the naming manager, the data is given to each user in the proposed framework, which ensures the integrity of the security.

The security system of the enterprise is proposed in [68]. Infrastructure construction is needed to combine the information security establishment. The Pr network security is methodical, and there is a chance of damaging the network. The big data security and basic network security condition consider the Pr network security as an arrival point. Then, they analyze the relevant evaluation indexes and establish the evaluation system model along with the key technologies in cloud enterprises. The artificial intelligence provides some new possibilities to solve the problems of network security such as complex security data, lack of real time, feasible information, and professional protection technology. The construction system and information security management system are the supplement to each other. The core of enterprise information security is the security management strategy which was implemented perfectly and strictly. Only network security that has improved the speed of information construction people can enjoy the ease of network safety. The key technology for the Pr is divided into five different layers called device security, network security, system security, and application security and data security layers. By analyzing the threats on each layer and then performing a positive action according to threat gives us a secure, confidentially, integrity, and controllability network.

In [69], the authors presented different security issues and their suitable solutions. The proposed work categorizes cloud computing into two sections called the deployment model and service delivery model. The deployment model includes public, private, hybrid, and community clouds, while the service delivery model includes SaaS, PaaS, and IaaS. Providing the security and reliability to the cloud environment is the responsibility

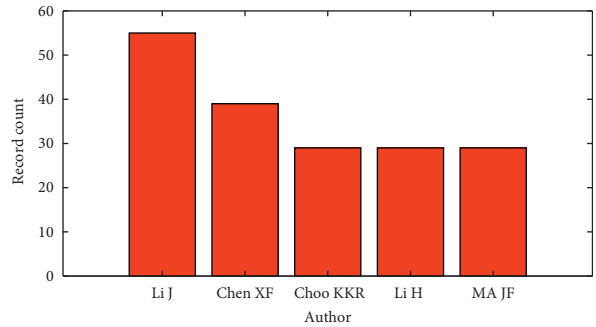


Figure 4: Author-based Survey.

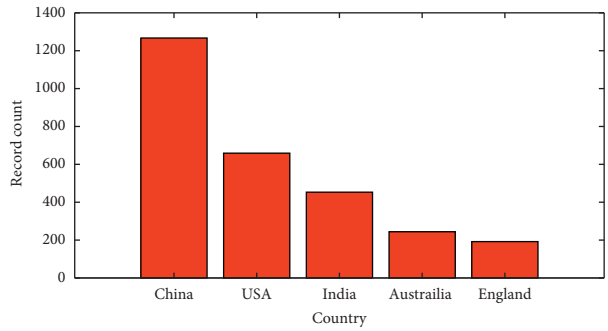


Figure 5: Country-based Survey.

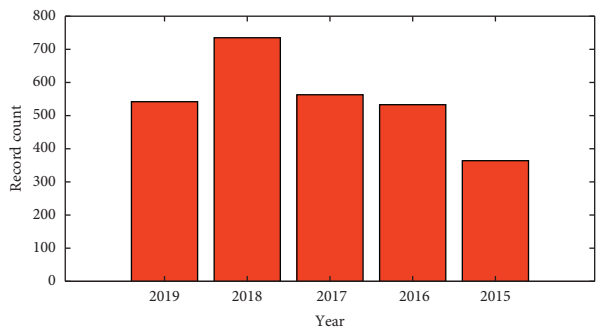


Figure 6: Year-based Survey.

of both user and cloud providers. The security issues and their corresponding solutions are presented in this article. The multitenant architecture problems are overcome through isolation and segmentation techniques. The perishing attacks and other breach threats are overcome through web browsers. Through service level agreement, the problems of the maintenance of the QoS, data loss, and other threats are solved. Using strong two-factor authentication techniques, the problems of account hijacking and related threats are solved. The DoS attack and the abuse of cloud service are prevented by using a honeypot system. In this article, different attacks are mentioned through CSA which are the topmost nine (09) attacks on the cloud system until 2013: data breaches, data loss, account hijacking, insecure APIs, DoS, malicious insiders, abuse of cloud services, insufficient due diligence, and shared technology issues.

The control blockchain-based framework called the Auth-Privacy chain is adopted to minimize the leakage and temper the data in the cloud [70]. First, the identity is provided to a node and this address is stored in blockchain; then, according to this address encryption is made. The process is called access control authorization and authorization revocation in the Auth-privacy chain. Along with these enterprises, the operating system is also implemented. After applying all these frameworks, the outcomes show that the proposed work not only minimizes the hackers and administration from illegal accessing but also provides protection to authorized privacy. Besides this, the Auth-privacy chain is responsible to provide integrity, confidentiality, and accountability along with the protection from many inside and outside unauthentic attacks. Blockchain is an open plain temper proof such as a distributed database. In DO command, a user can upload the data in the cloud, while

Table 14: Infrastructure-as-a-Service.

References	Overviews	Advantages	Techniques	Years
[67]	Two networks called software-defined network (SDN) and information-centric networking (ICN) are analyzed on the cloud model to secure the data; the ICN is considered the building block of the framework, and SDN decouples the control plane	This flexible infrastructure highly provides data security	Software defines network, and information-centric networking are applied	2017
[68]	The private network analyzes the big data to secure the enterprises; different threats on different layers are analyzed and give a solution to each threat; by doing this a secure enterprise Pr network is achieved	The framework gives different solutions for the different threats on five layers, and the Pr becomes more secure and confidential	The big data, key technology, five layers, and antivirus technology	2018
[69]	The cloud providers and the users both are responsible to deliver reliable and secure data over the cloud; different security threats and their solutions are analyzed here; the CSP protects the sharing of technology issues; the security provided is architecture and layer base	Highly secure techniques are discussed which fully protects the data on any deployment model	Isolation and segmentation, web browser, service level agreement, two-factor authentication, and many more techniques are used	2016
[70]	The leakage and temper of the cloud data is minimized; along with these, the privacy and unauthentic scenarios are being analyzed using a framework called the Auth-privacy chain, a blockchain framework, and enterprise operating system	This proposed framework eliminates the data leakage and temper of data in the cloud system and provides privacy	Blockchain-based access control, the Auth-privacy chain, and enterprise operating system	2020

DU command is responsible for accessing the data if the cloud allows. Cloud is a semitrusted platform, and in this framework, blockchain is assumed trustful. DO command uploads the data or resources to the cloud, and DU command sends a request to the cloud; then, the cloud checks whether blockchain has premium, and hence, finally replies according to the response.

3. Conclusion

In this survey, different challenges are highlighted for cloud computing security based on cloud components. These challenges are related to the outside as well as inside the cloud systems. Several techniques, approaches, models, algorithms etc., are applied at different levels to cloud components to make sure the security of the cloud models. In this paper, the classifications are made in which different issues related to security are solved with different techniques. The classifications involve the overviews of the proposed schemes in which specific issues related to cloud security are analyzed; their advantages and the techniques used to solve the concern issues are discussed. On the basis of cloud component classification, the researcher can pick out the desired technique related to the concern security issue while dealing with cloud security. Bibliometric survey based on authors, countries, and years has been conducted in the field of cloud computing security as can be seen in Figures 4–6, respectively. The bar graph of this survey shows the authors, countries, and years with their record in the field of cloud computing security. Moreover, the future directions are given which are concluded after analyzing all the problems in cloud security components which are to be improved yet. Table 1 shows the acronyms used in this manuscript.

4. Future Directions

The future directions for cloud computing security are

(i) The signature method is used to capture a specific threat and not allow the threats to interfere with the cloud data. This method accurately focused on the specific threat with a lower false rate. However, the signature method recognizes only threats that are known to its directory; this is unable to recognize the unknown threats. In the future, this method can be improved to catch all kinds of threats.

(ii) The cloud-native application pattern is used for cloud data security purpose. However, designing the cloud-native application (CNA) pattern is a complex procedure that required detailed planning and has no defined steps. Therefore, in future work, the researchers can optimize this technique.

(iii) The advanced encryption standard technique is mostly used to encrypt the data which provides strong security against the attackers. However, the AES does not withstand the brute force attack and

linear cryptographic analysis. This is a common kind of threat which is mostly occurring and not handled by the AES. In future work, AES needs to modify for the brute force threat.

(iv) The security service level agreement algorithms have been proposed to make secure the cloud data. Different techniques in this approach are used to evaluate security issues such as reference evaluation model and complex service supply chain techniques. However, in these techniques, the security evaluation for the cloud is not concerned with the end-to-end domain. Furthermore, the techniques considered in this approach generate the extra 8 bits attachment which affects the cloud storage space. Therefore, in the future, researchers can improve these techniques according to need.

(v) The virtual machines (VMs) efficiently enhance the cloud security, but it fails in scalability issues and depends on the single point failure and lack of customization. Therefore, in future work, these issues can be fixed and improve.

(vi) Most of the traditional algorithms are applied to secure the cloud systems at different stages and levels such as data encryption standard, advanced encryption standard, and Rivest–Shamir–Adleman cryptosystem, which are not automated. The future challenge is to make them in an automatic form which will enhance the accuracy and reduces the time consumption.

(vii) Different software, tools, and cryptographic approaches are analyzed to secure the cloud storage and layers, but these techniques are complex and more time consuming. In future work, these techniques can be converted to a noncomplex process, which consumes minimum time.

(viii) The App scan tool is used to secure the cloud data, but it is unable to stop all the threats. It can miss those threats and allow entering the cloud data which are unknown to its directory. In future, the researchers can expand and increase the domain of the App scan tool in terms of the directory which can capture and remove all the threats.

(ix) For the Big data, technology-based tools (MapReduce etc) are used to secure the large structure or unstructured data. Such traditional software cannot handle the massive data which results in the poor security of the big data. To overcome such an issue the researchers should design the optimal software tools that are best in their performances.

(x) A load balancing technique is used to ensure the security resources and performances used on it. While, it is also used to avoid data overloading and underloading in virtual machines which itself is a big challenge. The researchers should minimize the overloading and under loading of the data in the future and should maintain an intermediate data bunch to overcome this problem.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by King Saud University, Riyadh, Saudi Arabia, through Researchers Supporting Project number RSP-2020/184 and partially supported by the Faculty of Computer Science and Information Technology, University of Malaya under Postgraduate Research Grant (PG035-2016A).

References

[1] A. Tahir, F. Chen, H. U. Khan et al., "A systematic review on cloud storage mechanisms concerning e-healthcare systems," *Sensors*, vol. 20, no. 18, p. 5392, 2020.

[2] S. Narula and A. Jain, "Cloud computing security: Amazon web service," in *Proceedings of the 2015 Fifth International Conference on Advanced Computing & Communication Technologies*, pp. 501–505, Rohtak, Haryana, India, Feburary 2015.

[3] B. Liao, Y. Ali, S. Nazir, L. He, and H. U. Khan, "Security analysis of IoT devices by using mobile computing: a systematic literature review," *IEEE Access*, vol. 8, pp. 120331–120350, 2020.

[4] I. Gordin, A. Graur, A. Potorac, and D. Balan, "Security assessment of OpenStack cloud using outside and inside software tools," in *Proceedings of the 2018 International Conference on Development and Application Systems (DAS)*, pp. 170–174, Suceava, Romania, May 2018.

[5] R. Aluvalu and L. Muddana, "A survey on access control models in cloud computing," *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI)*, , 2015.

[6] C. Feng, A. Muhammad, A. Ahmad, A. Ullah, and H. U. Khan, "Towards energy-efficient framework for IoT big data healthcare solutions," *Scientific Programming*, vol. 2020, Article ID 7063681, , 2020.

[7] H. Xu, W. Yu, D. Griffith, and N. Golmie, "A survey on industrial Internet of Things: a cyber-physical systems perspective," *IEEE Access*, vol. 6, pp. 78238–78259, 2018.

[8] I. Khalil, A. Khreishah, and M. Azeem, "Cloud computing security: a survey," *Computers*, vol. 3, no. 1, pp. 1–35, 2014 Mar.

[9] M. Iglesias-Urkia, A. Orive, A. Urbieta, and D. Casado-Mansilla, "Analysis of CoAP implementations for industrial internet of Things: a survey," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 7, pp. 2505–2518, 2019.

[10] K. El Makkaoui, A. Ezzati, A. Beni-Hssane, and C. Motamed, "Cloud security and privacy model for providing secure cloud services," in *Proceedings of the 2016 2nd International*

Conference on Cloud Computing Technologies and Applications (Cloud Tech), pp. 81–86, Marrakech, Morocco, May 2016.

[11] G. Beier, S. Niehoff, and B. Xue, "More sustainability in Industry through industrial Internet of Things?" *Applied Sciences*, vol. 8, no. 2, p. 219, 2018.

[12] A. Singh and K. Chatterjee, "Cloud security issues and challenges: a survey," *Journal of Network and Computer Applications*, vol. 79, pp. 88–115, 2017.

[13] G. L. Reddy and B. M. Krishna, "Survey of cloud computing and its application," *Journal Impact Factor*, vol. 3, p. 24, 2018.

[14] X. Sun, "Critical security issues in cloud computing: a survey," in *Proceedings of the 2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing,(HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 216–221, Omaha, NE, USA, May 2018.

[15] R. Kaur and J. Kaur, "Cloud computing security issues and its solution: a review," in *Proceedings of the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, New Delhi, India, pp. 1198–1200, March 2015.

[16] N. C. Paxton, "Cloud security: a review of current issues and proposed solutions," in *Proceedings of the 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, pp. 452–455, Pittsburgh, PA, USA, November 2016.

[17] W. Dashti, A. Qureshi, A. Jahangeer, and A. Zafar, "Security challenges over cloud environment from service provider prospective," *Cloud Computing and Data Science*, vol. 1, no. 1, pp. 12–20, 2020.

[18] D. M. Vistro, A. U. Rehman, M. S. Farooq, A. Abid, and M. Idrees, "A survey ON the role OF security and integrity issues IN cloud," *Journal of Critical Reviews*, vol. 7, pp. 1456–1469, 2020.

[19] J. B. Hong, A. Nhlabatsi, D. S. Kim, N. A. Hussein, and K. M. Khan, "Systematic identification of threats in the cloud: a survey," *Computer Networks*, vol. 150, pp. 46–69, 2019.

[20] M. Anisetti, C. A. Ardagna, E. Damiani, and F. Gaudenzi, "A security benchmark for openstack," in *Proceedings of the 2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*, pp. 294–301, Honolulu, HI, USA, June 2017.

[21] D. K. Tosh, S. Shetty, X. Liang, C. A. Kamhoua, K. A. Kwiat, and L. Njilla, "Security implications of blockchain cloud with analysis of block withholding attack," in *Proceedings of the 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pp. 458–467, Madrid, Spain, May 2017.

[22] O. V. Lindqvist, G. Fitzgerald, Z. Wu, Y. Wu, F. Shen, and X. Luo, "Osmotic interrelationship between blood and gut fluid in the isopod Porcellio scaber Latr. (Crustacea)," in *Proceedings of the 2018 5th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2018 4th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pp. 57–59, Shanghai, China, June 2018.

[23] U. Ghosh, P. Chatterjee, D. Tosh, S. Shetty, K. Xiong, and C. Kamhoua, "An SDN based framework for guaranteeing security and performance in information-centric cloud networks," in *Proceedings of the 2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*, pp. 749–752, Honolulu, HI, USA, June 2017.

[24] S. Siadat, A. M. Rahmani, and H. Navid, "Identifying fake feedback in cloud trust management systems using feedback evaluation component and Bayesian game model," *The Journal of Supercomputing*, vol. 73, no. 6, pp. 2682–2704, 2017.

[25] A. Jakóbik, D. Grzonka, and F. Palmieri, "Non-deterministic security driven meta scheduler for distributed cloud organizations," *Simulation Modelling Practice and Theory*, vol. 76, pp. 67–81, 2017.

[26] Y. Wu, Y. Lyu, and Y. Shi, "Cloud storage security assessment through equilibrium analysis," *Tsinghua Science and Technology*, vol. 24, no. 6, pp. 738–749, 2019.

[27] J. Luna, A. Taha, R. Trapero, and N. Suri, "Quantitative reasoning about cloud security using service level agreements," *IEEE Transactions on Cloud Computing*, vol. 5, no. 3, pp. 457–471, 2015.

[28] F. Kong, Y. Zhou, B. Xia, L. Pan, and L. Zhu, "A security reputation model for IoT health data using S-AlexNet and dynamic game theory in cloud computing environment," *IEEE Access*, vol. 7, pp. 161822–161830, 2019.

[29] R. Y. Chou, G. W. Bannister, and inventors, "Nubeva Inc, assignee. Seamless service updates for cloud-based security services," United States Patent US 10,530,815, 2020.

[30] F. R. Martinez and E. Pulier, "Csc Agility Platform Inc, assignee. System and method for a cloud computing abstraction layer with security zone facilities," United States Patent Application US 16/058,688, 2019.

[31] S. Kang, B. Veeravalli, and K. M. Aung, "A security-aware data placement mechanism for big data cloud storage systems," in *Proceedings of the 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 327–332, New York, NY, USA, April 2016.

[32] J.-H. Lee, S. K. Young, H. K. Jong, and K. K. Ik, "Toward the SIEM ar-chitecture for cloud-based security services," in *Proceedings of the 2017 IEEE Conference on Communications and Network Security (CNS)*, pp. 398-399, Las Vegas, NV, USA, October 2017.

[33] Y. Han, J. Chan, T. Alpcan, and C. Leckie, "Using virtual machine allocation policies to defend against co-resident attacks in cloud computing," *IEEE Transactions on Dependable and Secure Computing*, vol. 14, no. 1, pp. 95–108, 2015.

[34] A. De Benedictis, V. Casola, M. Rak, and U. Villano, "Cloud security: from per-provider to per-service security slas," in *Proceedings of the 2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS)*, pp. 469–474, Ostrava, Czech Republic, September 2016.

[35] V. Chang and M. Ramachandran, "Towards achieving data security with the cloud computing adoption framework," *IEEE Transactions on Services Computing*, vol. 9, no. 1, pp. 138–151, 2015.

[36] V. Chang, Y.-H. Kuo, and M. Ramachandran, "Cloud computing adoption framework: a security framework for business clouds," *Future Generation Computer Systems*, vol. 57, pp. 24–41, 2016.

[37] H. Cheng, C. Rong, K. Hwang, W. Wang, and Y. Li, "Secure big data storage and sharing scheme for cloud tenants," *China Communications*, vol. 12, no. 6, pp. 106–115, 2015.

[38] Z. Li, J. Ge, H. Yang et al., "A security and cost aware scheduling algorithm for heterogeneous tasks of scientific workflow in clouds," *Future Generation Computer Systems*, vol. 65, pp. 140–152, 2016.

[39] Z. Hu, S. Gnatyuk, O. Koval, V. Gnatyuk, and S. Bondarovets, "Anomaly detection system in secure cloud computing environment," *International Journal of Computer Network and Information Security*, vol. 9, no. 4, p. 10, 2017.

[40] J. Agarkhed and R. Ashalatha, "An efficient auditing scheme for data storage security in cloud," in *Proceedings of the 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pp. 1–5, Delhi, India, April 2017.

[41] A. Nhlabatsi, J. B. Hong, D. S. Kim, R. Fernandez, N. Fetais, and K. M. Khan, "SpiralŜRA: a threat-specific security risk assessment framework for the cloud," in *Proceedings of the 2018 IEEE International Conference on Software Quality, Reliability and Security (QRS)*, pp. 367–374, Lisbon, Portugal, July 2018.

[42] P. Y. Wang and M. Q. Hong, "A secure management scheme designed in cloud. In2016 IEEE 2nd International conference on big data security on cloud (BigDataSecurity)," in *Proceedings of the IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 158–162, New York, NY, USA, April 2016.

[43] S. A. Hande and S. B. Mane, "An analysis on data Accountability and Security in cloud," in *Proceedings of the 2015 International Conference on Industrial Instrumentation and Control (ICIC)*, pp. 713–717, Pune, India, May 2015.

[44] J. Xu, C. Liang, H. K. Jain, and D. Gu, "Openness and security in cloud computing services: assessment methods and investment strategies analysis," *IEEE Access*, vol. 7, pp. 29038–29050, 2019.

[45] J. H. Lee, Y. S. Kim, J. H. Kim, and I. K. Kim, "Toward the SIEM architecture for cloud-based security services," in *Proceedings of the 2017 IEEE Conference on Communications and Network Security (CNS)*, pp. 398-399, Las Vegas, NV, USA, October 2017.

[46] T. S. Fatayer and K. A. Timraz, "MLSCPC: multi-level security using covert channel to achieve privacy through cloud computing," in *Proceedings of the 2015 World Symposium on Computer Networks and Information Security (WSCNIS)*, pp. 1–6, Hammamet, Tunisia, September 2015.

[47] J. Li, Y. Zhang, X. Chen, and Y. Xiang, "Secure attribute-based data sharing for resource-limited users in cloud computing," *Computers & Security*, vol. 72, pp. 1–12, 2018.

[48] H. Hassan, A. I. El-Desouky, A. Ibrahim, E.-S. M. El-Kenawy, and R. Arnous, "Enhanced QoS-based model for trust assessment in cloud computing environment," *IEEE Access*, vol. 8, pp. 43752–43763, 2020.

[49] D. Gonzales, J. M. Kaplan, E. Saltzman, Z. Winkelman, and D. Woods, "Cloud-trust—a security assessment model for infrastructure as a service (IaaS) clouds," *IEEE Transactions on Cloud Computing*, vol. 5, no. 3, pp. 523–536, 2015.

[50] N. S. Darwazeh, R. S. Al-Qassas, and F. AlDosari, "A secure cloud computing model based on data classification," *Procedia Computer Science*, vol. 52, pp. 1153–1158, 2015.

[51] H. Cui, Y. Li, X. Liu, N. Ansari, and Y. Liu, "Cloud service reliability modelling and optimal task scheduling," *Iet Communications*, vol. 11, no. 2, pp. 161–167, 2017.

[52] S. Feng, Z. Xiong, D. Niyato, P. Wang, and S. S. Wang, "Joint pricing and security investment for cloud-insurance: a security interdependency perspective," in *Proceedings of the 2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, Marrakech, Morocco, April 2018.

[53] D. H. Sharma, C. A. Dhote, and M. M. Potey, "Implementing intrusion management as security-as-a-service from cloud," in *Proceedings of the 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 363–366, Bengaluru, India, October 2016.

[54] M. Usman, M. Ahmad Jan, and X. He, "Cryptography-based secure data storage and sharing using HEVC and public clouds," *Information Sciences*, vol. 387, pp. 90–102, 2017.

[55] M. Ramachandran and V. Chang, "Towards performance evaluation of cloud service providers for cloud data security," *International Journal of Information Management*, vol. 36, no. 4, pp. 618–625, 2016.

[56] K. K. Gola, R. Rathore, and S. Rastogi, "Secure: dynamic distributed load balancing technique in cloud computing," *International Journal of Advanced Research in Computer Science*, vol. 9, no. 1, 2018.

[57] A. Arora, A. Khanna, A. Rastogi, and A. Agarwal, "Cloud security ecosystem for data security and privacy," in *Proceedings of the 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*, pp. 288–292, Noida, India, January 2017.

[58] D. Zhe, W. Qinghong, S. Naizheng, and Z. Yuhan, "Study on data security policy based on cloud storage," in *Proceedings of the 2017 IEEE 3rd International Conference on Big Data Security on Cloud (Bigdatasecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 145–149, Beijing, China, May 2017.

[59] C. Prakash and S. Dasgupta, "Cloud computing security analysis: challenges and possible solutions," in *Proceedings of the 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp. 54–57, Chennai, India, March 2016.

[60] G. Ducatel, J. Daniel, T. Dimitrakos, F. El-Moussa, R. Rowlingson, and A. Sajjad, "Managed security service distribution model," in *Proceedings of the 2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 404–408, Beijing, China, August 2016.

[61] M. Derfouf, A. Mimouni, and M. Eleuldj, "Vulnerabilities and storage security in cloud computing," in *Proceedings of the 2015 International Conference on Cloud Technologies and Applications (Cloud Tech)*, pp. 1–5, Marrakesh, Morocco, June 2015.

[62] T. C. Chiueh, E. J. Chang, R. Huang, H. Lee, V. Sung, and M. H. Chiang, "Security considerations in ITRI cloud OS," in *Proceedings of the 2015 International Carnahan Conference on Security Technology (ICCST)*, pp. 107–112, Taipei, Taiwan, September 2015.

[63] W. Zhu and C. Lee, "A security protection framework for cloud computing," *JIPS*, vol. 12, no. 3, pp. 538–547, 2016.

[64] B. H. Lee, E. K. Dewi, and M. F. Wajdi, "Data security in cloud computing using AES under HEROKU cloud," in *Proceedings of the 2018 27th Wireless and Optical Communication Conference (WOCC)*, pp. 1–5, Hualien, Taiwan, April 2018.

[65] L. Qing, Z. Boyu, W. Jinhua, and L. Qinqian, "Research on key technology of network security situation awareness of private cloud in enterprises," in *Proceedings of the 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 462–466, Chengdu, China, April 2018.

[66] P. Liu, "Public-key encryption secure against related randomness attacks for improved end-to-end security of cloud/edge computing," *IEEE Access*, vol. 8, pp. 16750–16759, 2020.

[67] K. A. Torkura, M. I. Sukmana, F. Cheng, and C. Meinel, "Leveraging cloud native design patterns for security-as-a-service applications," in *Proceedings of the 2017 IEEE*

International Conference on Smart Cloud (Smart Cloud), vol. 3, pp. 90–97, New York, NY, USA, November 2017.

[68] H. Li, R. Lu, J. Misic, and M. Mahmoud, "Security and privacy of connected vehicular cloud computing," *IEEE Network*, vol. 32, no. 3, pp. 4–6, 2018.

[69] H. Zhang, "Research on job security scheduling strategy in cloud computing model," in *Proceedings of the 2015 International Conference on Intelligent Transportation, Big Data and Smart City*, pp. 649–652, Halong Bay, Vietnam, December 2015.

[70] C. Yang, L. Tan, N. Shi, B. Xu, Y. Cao, and K. Yu, "AuthPrivacyChain: a blockchain-based access control framework with privacy protection in cloud," *IEEE Access*, vol. 8, pp. 70604–70615, 2020.

Hindawi
Scientific Programming
Volume 2020, Article ID 8810634, 10 pages
https://doi.org/10.1155/2020/8810634

Review Article

Use of Big Data Tools and Industrial Internet of Things: An Overview

Yingzi Wang,[1] **Muhammad Nazir Jan**,[2] **Sisi Chu**,[3] **and Yue Zhu**[3]

[1]*College of Intelligence and Computing, TianJin University, TianJin 300300, China*
[2]*Department of Computer Science, University of Swabi, Swabi, Pakistan*
[3]*Automotive Data of China Co.,Ltd., TianJin 300300, China*

Correspondence should be addressed to Yingzi Wang; wangyingzi@catarc.ac.cn

Received 8 September 2020; Revised 26 September 2020; Accepted 3 October 2020; Published 21 October 2020

Academic Editor: Habib Ullah Khan

Big data is ever playing an important role in the industry as well as many other organizations. With the passage of time, the volume of data is increasing. This increase will create huge bulk of data which needs proper tools and techniques to handle its management and organization. Different techniques and tools are being used to properly handle the management of data. A detailed report of these techniques and tools is needed which will help researchers to easily identify a tool for their data and take help to easily manage the data, organize the data, and extract meaningful information from it. The proposed study is an endeavour toward summarizing and identifying the tools and techniques for big data used in Industrial Internet of Things. This report will certainly help researchers and practitioners to easily use the tools and techniques for their need in an effective way.

1. Introduction

With the passage of time, the volume of data is increasing. In today's digital world, the information surges with the extensive use of the Internet and global communication systems. This increase will create huge bulk of data which needs proper tools and techniques to handle its management and organization. Big data is ever playing an important role in the industry as well as many other organizations. Huge bulk of data is produced from the healthcare information systems, electronic records, wearables, smart devices, handheld devices, and so on. The recent increase in medical big data and the development of computational techniques in the field of information technology enable researchers and practitioners to extract and visualize big data in a new spectrum of use.

The industry is leading toward the spreading out and developments of IIoT with the incorporation of emerging technologies and applications of IoT. The aim of the IIoT is to achieve high efficiency of operations for management of industrial assets and to increase the productivity of industries. More attention is given to the applications of IoT with its integration to industries. The applications of IoT are obvious in every field of life from industry to education, healthcare, and to other places. A number of studies are available related to the applications, uses, and different approaches to handle big data [1–8]. Different techniques and tools are being used for extracting important information from big data. The data are mostly unstructured which need proper structure, shape, and management through which the data can easily be accessed and processed. The role of visualization is to capture the important information from the data and to visualize it for the easiness of practitioners. Some of the programming tools which deal with big data are Informatica PowerCenter, Apache Hadoop, and Tableau, which analyze data extremely efficiently and enable the visualization of meaningful insights extracted from big data.

To facilitate the management of data for easy access and to operate, there should be a detailed report on the existing tools and techniques which can easily access, manage, operate, and execute useful information from the data for different purposes. Therefore, to facilitate this process, a detailed report of the existing literature is presented in this study. This detailed report will help researchers and scholars

to devise new algorithms, techniques, and tools for the analysis and management of big data.

The organization of the paper is as follows. Section 2 shows the related work to big data tools and support of the industry. Section 3 presents the existing approaches to support big data in IIoT. Section 4 shows the support of IIoT regarding big data tools and techniques. The paper is concluded in Section 5.

2. Big Data Tools and Support of the Industry

With the advancements in Industrial Internet of Things (IIoT) sensing, communication, technology characterizations, and high throughput instrumentation, the level of data generation is expected to grow exponentially [9]. Lin et al. [10] presented an approach of integrating sensing data from diverse sources and equipment to apply on IIoT. The industrial Micro Control Unit is connected to interface with actuator, data sources, and equipment. The experimental results show that IIoT can reduce the problem of heterogeneous protocol and database manufacturing data transmission. This article demonstrates the complexity and unique nature of multimedia big data (MMBD) computing for Internet of Things (IoT) applications as well as builds up an inclusive taxonomy used for MMBD abstracted into a new process model reflecting MMBD over IoT. Many research challenges linked with MMBD, for example, quality of service requirements, heterogeneity, reliability, accessibility, and scalability, are addressed by the process model. The process model is discussed through a case study [11]. In this work, architecture for flood forecasting and monitoring is proposed by means of convergence between HPC and big data. This architecture can analyze, store, and collect big data as well as help in the flood prediction result generation [12]. Mobile computing services can be used in IoT by using services of mobile phones, apps, or through M-Health care system [13]. Alexopoulos et al. [14] presented the IIoT architecture and its development details to support the industrial product service system life cycle.

In this article, a novel model is developed in the perspective of manufacturing progression that reviews the key big data analytics (BDA) capabilities. The findings are beneficial for the companies in order to understand big data potential implications as well as their analytics capabilities for their manufacturing processes and efficient BDA-enabler infrastructure design [15]. Boyes et al. [16] presented the concept of IIoT and the association to the ideas such as cyber physical systems and Industry 4.0. IoT-related taxonomies were analyzed and an analysis framework was developed for IIoT that can be used to list and characterize the devices of IIoT when analyzing security vulnerability and threats. For the big data sentiment analysis (BDSA) and for best or optimal decision selection, a framework was proposed and also applied as a mathematical algorithm [17]. In this study, for big data and Cognitive Internet of Things (CIoT), a new architecture is proposed. The planned architecture helps the computing systems through combining data lake (DL) and warehouse (DWH), and for the collection of heterogeneous data, a tool is defined [18]. Urquhart and Mcauley [19] presented an approach for the risks of IIoT drawn both on the regulatory and technical perspectives. In this study, functional and structural properties of cloud manufacturing (CMfg) were analyzed, and a business intelligence architecture was proposed that plans to empower distributing pertinent KPIs identified with intrigued process data, with the helpful layer of dependability [20].

An overview of big data in smart manufacturing was directed, and an applied framework was proposed from the viewpoint of item life cycle. This framework permits examining key advantages and potential applications, and the debate of future research directions and current challenges gives essential insights for the industry and scholarly world [21]. This paper examines the current big data analytics (BDA) technologies, strategies, and algorithms that can prompt the improvement of insightful Industrial Internet of Things (IIoT) frameworks. We devise a scientific classification by characterizing and classifying the literature based on essential factors (for example, analytics types, industrial analytics applications, requirements, analytics techniques, analytics tools, and data sources). The case studies and frameworks of different endeavours were presented which have been profited by BDA [22]. This paper investigates how firms can capture an incentive from big data to improve green commitment by giving an applied model through an exhaustive and all-encompassing writing that relates big data sources to the reception of various green systems. The principle finding of the examination is that organizations that need to execute clean innovation strategy frequently allude to outside accomplice to build up the essential architecture expected to abuse enormous information potentialities [23]. Apart from these approaches, the big data and IoT have several other applications in diverse issues of the real world [24–28].

3. Existing Approaches to Support Big Data in IIoT

Humayun et al. [29] presented a comprehensive report of the evolution, prevention, and mitigation of ransomware in the context of IoT. For smart factories, construction path and reference architecture were proposed by examining IIoT technology as well as their application in assembling workshops. Joined with the examination of business as usual and requirements of the discrete assembling undertaking workshops, this paper structures the overall theoretical model architecture of the framework [30]. In this examination, a blockchain-dependent data sharing scheme was proposed that entirely considers efficiency as well as security of data sharing. In this plan, a Hyperledger Fabric and identity authentication-dependent secure data sharing structure was designed for the data sharing security. Additionally, a network recognition algorithm was proposed to partition the customers into various data sharing networks as per the comparability of mark data. The exploratory outcomes demonstrate that the proposed colloboration is successful for efficient and secure data sharing among various customers [31].

This paper discusses about the IoT data management concepts and current and survey solutions, talks about the most encouraging solutions, and recognizes important open exploration issues on the theme giving rules to assist further contributions [32]. In this article, for a scalable pipeline to distribute as well as process data as of blend of shop-floor sources, an architecture was proposed. The architecture was implemented in order to explore the feasibility of this methodology and bring together ad hoc power data and MTConnect-compliant machine to help analytics applications [33]. This work presents a procedure data examination stage which worked around the idea of Industry 4.0. The platform uses the big data software tools, ML algorithms, and state-of-the-art IIoT platforms. The results indicated that in situations where process information about the procedure within reach is restricted, information-driven delicate sensors are helpful instruments for predictive data investigation [34]. For industrial data processing, an Industrial Internet of Things cloud-fog hybrid network (ITCFN) framework was proposed. The results have shown that the proposed framework effectively reduces the processing delay of industrial data [35].

In this study, a systematic strategy was used to review the weaknesses as well as strengths of open-source technologies for stream processing and big data to set up its usage for Industry 4.0 use cases [36]. A framework was developed for the additive manufacturing enterprises by combining sustainable smart manufacturing technologies, additive manufacturing, and big data analytics. The proposed framework is beneficial for additive manufacturing industry leaders to take the right decision at the beginning stage of the product life cycle [37]. The big data characteristic of the testbed was studied by using an inhouse-developed IoT-enabled manufacturing testbed [38]. A distributed service-oriented architecture was provided for the solution of problem of product tracing [39]. The distributions of droplet size with high-velocity airblast atomization were examined [40]. In this article, an interactive data investigation framework was proposed, which poses a service-oriented perspective on the smart factory [23]. This article investigates the potential of artificial intelligence (AI) as well as machine learning (ML) to lever big data and Internet of Things (IoT) in smart cities in personalised service development. IoT smart city applications are suggested so as to benefit from this work [41]. Gierej [42] presented the idea of a business model for the companies implementing IIoT technologies. The approach is developed to help traditional companies in the transition of the digital market.

The proof procurement challenge is examined. A contextual investigation of a smart city venture with IoT administrations gathering big data which are put away in the cloud processing condition is presented. The strategies can be summed up to other big data in the cloud environment [43]. A fault prediction technique dependent on industrial big data is presented, which legitimately exhumes the connection between the data, for example, the status as well as sound data, and the equipment faults by machine learning techniques [44]. Distributed growing self-organizing map (DGSOM) and a novel distributed self-adaptive neural network algorithm were presented to tackle unsupervised machine learning need of big data [45]. Younan et al. [46] presented a study with a comprehensive review of the existing challenges in the literature and recommended technologies for enabling the analysis of data and search in the future IoT search engines. Two case studies are presented to show promising growth on smartness and intelligence of applications of IoT based on the integration of information and communication technologies. The applications of smart phones enable the patients to know about their diseases after the analysis in the field of gynaecology and paediatrics [47]. In this article, an architecture based on Internet of Things is proposed for big data that is used for diverse smart cities. The results demonstrated that this kind of method has the potential of the applicability to give beneficial services of smart cities, for example, detection of travel profiles in smart transport, comfort in smart buildings, and management of the energy consumption [48]. Jiang [49] presented an approach which studies the IoT developments and technologies related to cloud computing and smart cities and then focussed on the IoT technologies and cloud computing. Dachyar et al. [50] conducted an in-depth analysis of the 26420 papers published in the area of IoT. This article aims to adapt and detect concept drift dependent on cognitive learning principles. The approach executes to detect concept drift, determines concept drift type as well as in automated time windows [51]. Table 1 shows the existing approaches, methods, and tools to support big data.

4. Support of IIoT regarding Big Data Tools and Techniques

Several studies exist related to the applications of big data in IIoT. The study presented an enhanced platform of industrial big data for the reduction of time and data storage space of data processing [54]. The aim of the paper is to assess the impact of different serialization and compression methods on the platform of big data and then attempt to select the most suitable method for the platform of industry. The aim of the study is to propose a fabric which is a technique of blockchain-based data transmission for IIoT [56]. The approach uses secret sharing mechanism based on blockchain. The paper presented an approach of city geospatial dashboard for the collection, sharing, and visualization of the data collected from different sources like satellite data, IoT devices, and other big data [58]. The contribution of the paper is to present the concept of constructing community-based platform of cross IIoT service through utilizing the existing mobile and fixed facilities as wireless IoT gateway in a city which facilitates the easy implementation of IoT gateway at local service for bringing economical and social values [59]. The study focussed on the spatiotemporal modeling to organize the data in temporal, attributive, and spatial dimensions [60]. To manage the multisource manufacturing data, ontology-based big data integration mechanism is presented. The authors proposed an ADTT—advanced distributed tensor-

Table 1: Existing approaches, methods, and tools to support big data.

S.No	Reference	Title
1	[9]	Big data analytics tool based on statistical process monitoring for smart manufacturing
2	[11]	Multimedia big data computation and applications of IoT
3	[12]	IoT, big data, and HPC-based smart flood management framework
4	[15]	Big data analytics for manufacturing processes
5	[17]	An algorithmic implementation of entropic ternary reduct soft sentiment set using soft computing technique on big data sentiment analysis for optimal selection of a decision based on real-time update in online reviews
6	[18]	Architecture for Cognitive IoT and big data
7	[20]	Challenges and opportunities for publishing IIoT data in manufacturing
8	[21]	A comprehensive review of big data analytics throughout product life cycle to support sustainable smart manufacturing
9	[22]	Role of big data analytics in IIoT
10	[23]	Big data and natural environment
11	[30]	Intelligent manufacturing production line data monitoring system for IIoT
12	[31]	A secure and efficient data sharing scheme based on blockchain in IIoT
13	[32]	Data management techniques for IoT
14	[33]	Scalable data pipeline architecture to support the IIoT
15	[34]	Industry 4.0-based process data analytics platform
16	[35]	Optimization of IIoT data processing latency
17	[36]	Big data and stream processing platforms for Industry 4.0 requirements mapping for a predictive maintenance use case
18	[37]	Framework of big data for sustainable and smart additive manufacturing
19	[38]	Feature engineering in big data analytics for IoT-enabled smart manufacturing
20	[39]	An architecture for aggregating information from distributed data nodes for IIoT
21	[40]	Application of big data analysis technique on high-velocity airblast atomization
22	[23]	Interactive data exploration as a service for the smart factory
23	[41]	Smart city services using machine learning, IoT, and big data
24	[43]	Digital forensics challenges to big data in the cloud
25	[44]	On fault prediction based on industrial big data
26	[45]	Apache spark-based distributed self-organizing map algorithm for sensor data analysis
27	[48]	Techniques of big data to smart city deployments
28	[51]	A cognitive data stream mining technique for context-aware IoT systems
29	[52]	Implementation of the FSO2
30	[53]	An intelligent outlier detection method with one class support tucker machine and genetic algorithm toward big sensor data in IoT
31	[54]	Big data-based improved data acquisition and storage system for designing industrial data platform
32	[55]	Cybersecurity in an IIoT environment
33	[56]	A secure fabric blockchain-based data transmission technique for IIoT
34	[57]	Concept drift detection and adaption in big imbalance IIoT data using an ensemble learning method of offline classifiers
35	[58]	City geospatial dashboard
36	[59]	A community-based IoT service platform to locally disseminate socially valuable data
37	[60]	The spatiotemporal modeling and integration of manufacturing big data in job shop
38	[61]	A big data-enabled consolidated framework for energy efficient software defined data centers in IoT setups
39	[62]	A parallel military dog-based algorithm for clustering big data in cognitive IIoT
40	[63]	Big data cleaning based on mobile edge computing in industrial sensor cloud
41	[64]	A highly efficient distributed tensor-train decomposition method for IIoT big data
42	[65]	Big data-driven edge-cloud collaboration architecture for cloud manufacturing

train—decomposition approach along with a computational method for the IIoT big data processing [64]. The existing literature was searched in order to identify the associated materials related to the proposed study. For this purpose, the popular libraries such as ACM, IEEE, ScienceDirect, and Springer were considered to show the related materials. The reason behind these libraries was that these libraries publish quality materials which are peer reviewed. Figure 1 shows the number of papers published in the given years in the library of ScienceDirect. The last five years were considered as the latest research published in these recent years.

Figure 2 shows the article type along with the number of publications in the given library.

Figure 3 shows publication titles and percentage of publications.

Figure 4 shows the articles types and number of publications in the library IEEE.

Figure 5 shows the publication topics and percentages of number of publications.

Figure 6 shows the media format and number of publications in the ACM library.

Figure 7 shows the publication types and number of papers published in the given library.

Figure 8 shows the number of publications in the given years.

Figure 9 shows the article types and percentages of publication in the Springer library.

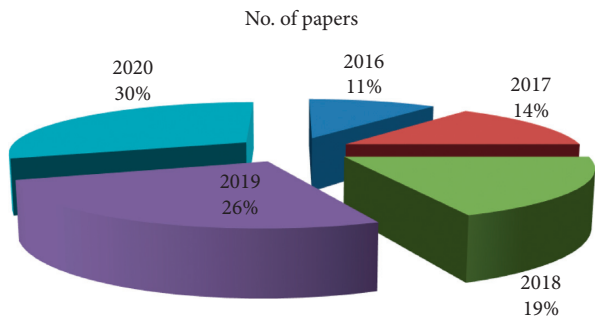


Figure 1: Number of papers in the given year for ScienceDirect.

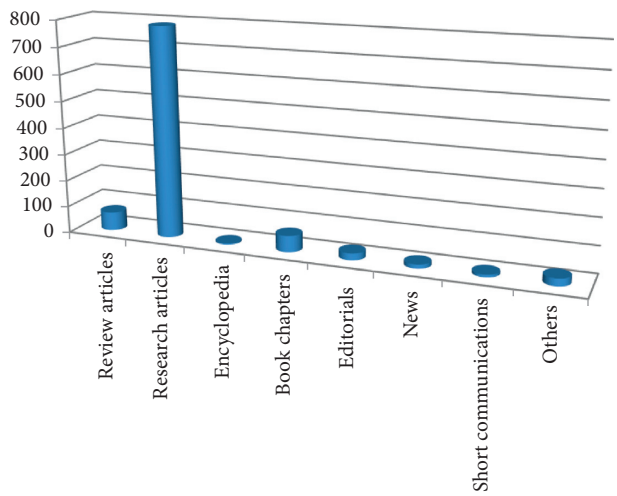


Figure 2: Article type and number of publications.

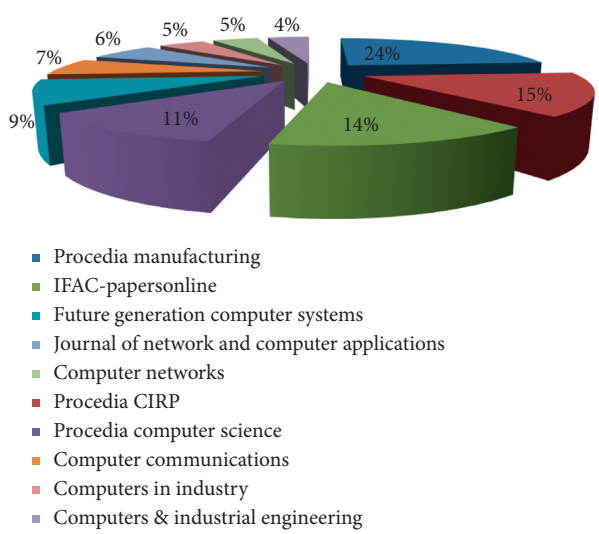


Figure 3: Publication titles and number of publications.

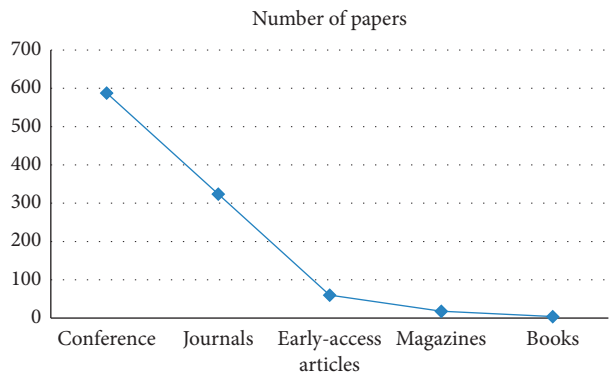


Figure 4: Articles type and number of publications.

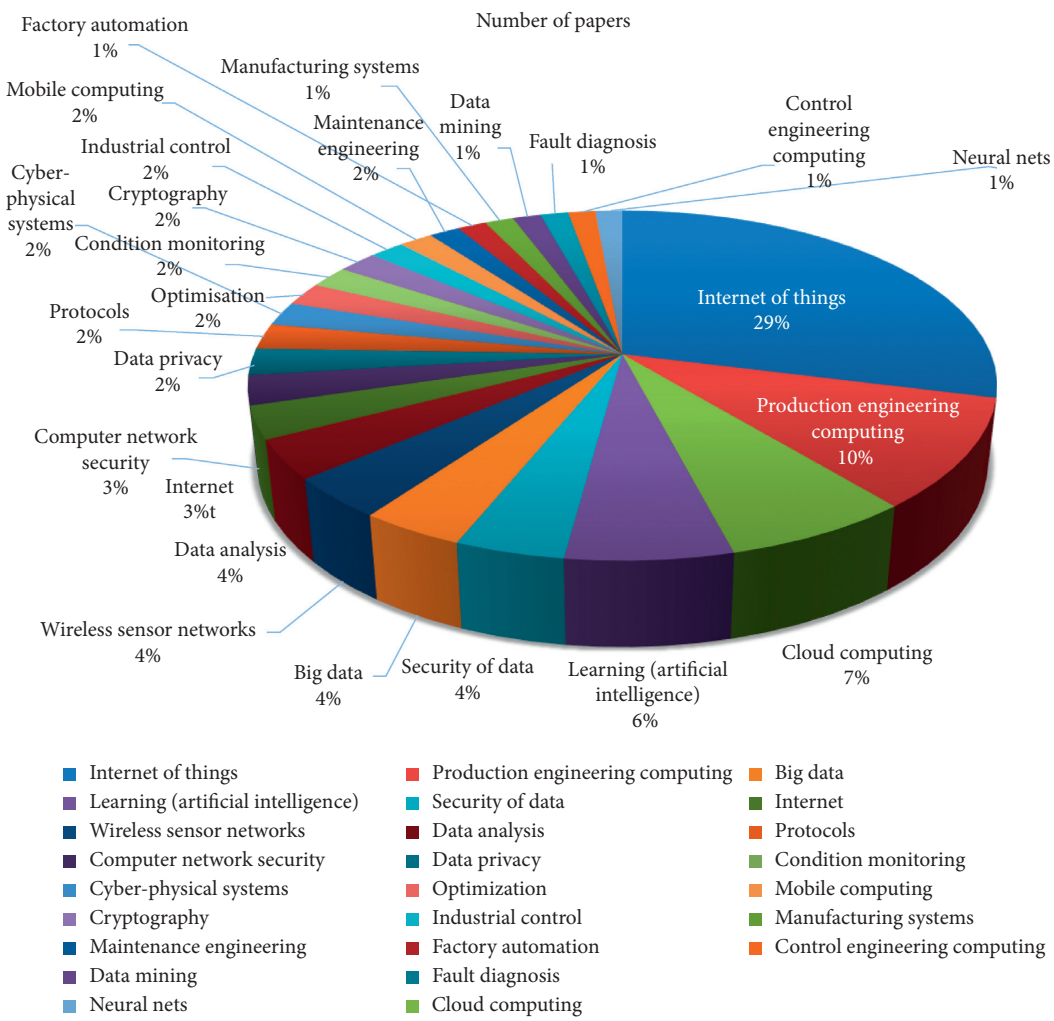


Figure 5: Publication topics and percentage of publications.

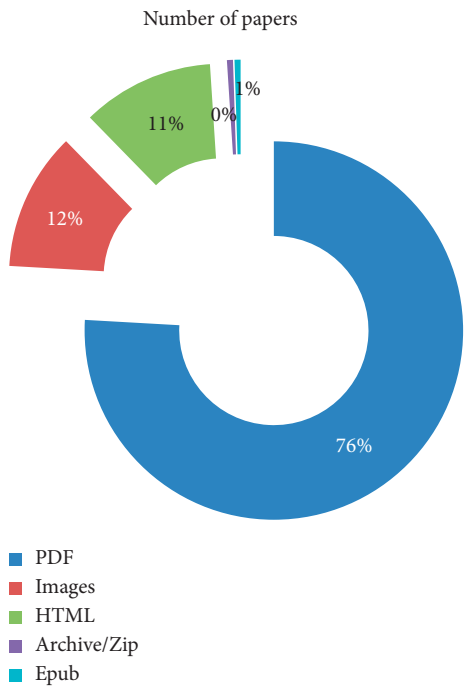


Figure 6: Media format and number of publications.

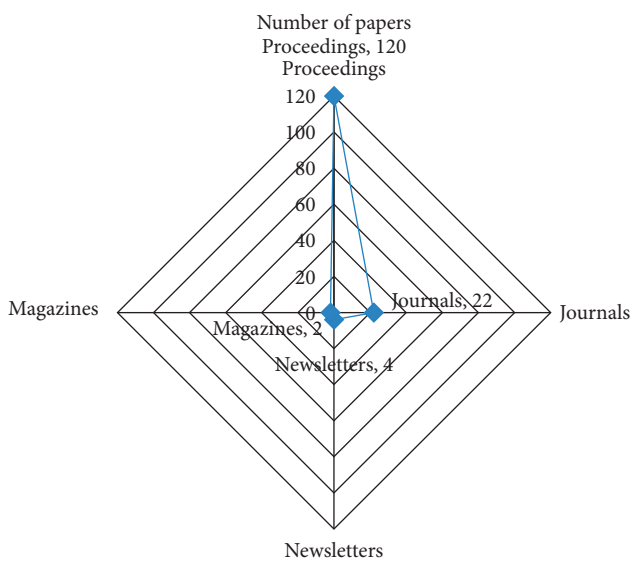


Figure 7: Publication types and number of papers.

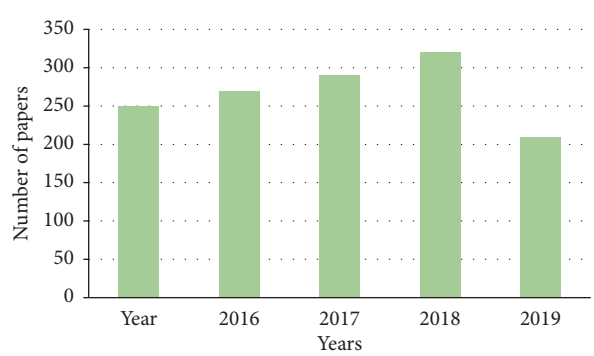


Figure 8: Number of papers in the given years.

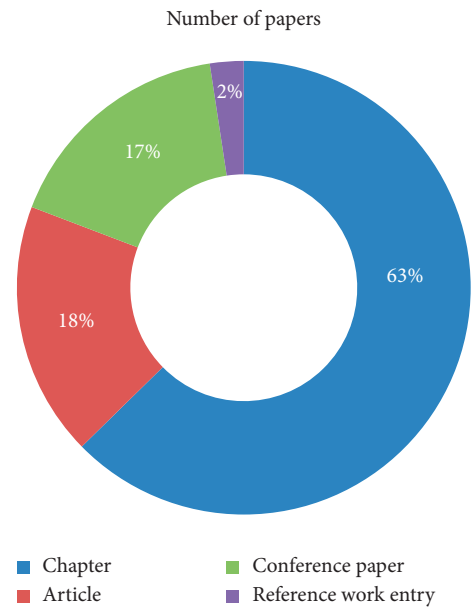


Figure 9: Content types and percentage of publications.

5. Conclusion

With the passage of time, the volume of data is increasing. This increase will create huge bulk of data which needs proper tools and techniques to handle its management and organization. Big data is ever playing an important role in the industry as well as many other organizations. Huge bulk of data is produced from the healthcare information systems, electronic records, wearables, smart devices, handheld devices, and so on. The recent increase in medical big data and the development of computational techniques in the field of information technology enable researchers and practitioners to extract and visualize big data in a new spectrum of use. Different techniques and tools are being used to properly handle the management of data. A detailed report of these techniques and tools is needed which will help researchers to easily identify a tool for their data and take help to easily manage the data, organize the data, and extract meaningful information from it. The proposed study is an endeavour toward summarizing and identifing the tools and techniques for big data used in IIoT. This report will help researchers and practitioners to easily use the tools and techniques for their need in an effective way and will devise new solutions for the industry of big data.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this paper.

Acknowledgments

This study was sponsored in part by the Intelligent Manufacturing Project of Tianjin (20193155).

References

[1] E. Yadegaridehkordi, M. Nilashi, L. Shuib et al., "The impact of big data on firm performance in hotel industry," *Electronic Commerce Research and Applications*, vol. 40, p. 100921, 2020.

[2] S. Asadi, R. H. Abdullah, M. Safaei, and S. Nazir, "An integrated sem-neural network approach for predicting determinants of wearable healthcare devices adoption," *Mobile Information Systems*, vol. 2019, Article ID 8026042, 9 pages, 2019.

[3] S. Nazir, S. Khan, H. U. Khan et al., "A comprehensive analysis of healthcare big data management, analytics and scientific programming," *IEEE Access*, vol. 8, pp. 95714–95733, 2020.

[4] S. Nazir, M. Nawaz, A. Adnan, S. Shahzad, and S. Asadi, "Big data features, applications, and analytics in cardiology-A systematic literature review," *IEEE Access*, vol. 7, no. 1, pp. 143742–143771, 2019.

[5] S. Nazir, M. Nawaz Khan, S. Anwar et al., "Big data visualization in cardiology-A systematic review and future directions," *IEEE Access*, vol. 7, no. 1, pp. 115945–115958, 2019.

[6] A. U. Haq, "Intelligent machine learning approach for effective recognition of diabetes in the E-healthcare using clinical data," *Sensors*, vol. 20, 2020.

[7] S. Nazir, S. Ali, M. Yang, and Q. Xu, "Deep learning algorithms and multi-criteria decision making used in big data- a systematic literature review," *Security and Communication Networks*, vol. 2020, Article ID 2836064, 19 pages, 2020.

[8] S. Nazir, *A Comprehensive Analysis of Healthcare Big Data Management, Analytics and Scientific Programming*, IEEE Access, Piscataway, NJ, USA, 2020.

[9] Q. P. He and J. Wang, "Statistical process monitoring as a big data analytics tool for smart manufacturing," *Journal of Process Control*, vol. 67, pp. 35–43, 2018.

[10] Y. J. Lin, C.-B. Lan, and C.-Y. Huang, "A realization of cyber-physical manufacturing Control system through industrial internet of things," *Procedia Manufacturing*, vol. 39, pp. 287–293, 2019.

[11] A. Kumari, S. Tanwar, S. Tyagi, N. Kumar, M. Maasberg, and K.-K. R. Choo, "Multimedia big data computing and internet of things applications: a taxonomy and process model," *Journal of Network and Computer Applications*, vol. 124, pp. 169–195, 2018.

[12] S. K. Sood, R. Sandhu, K. Singla, and V. Chang, "IoT, big data and HPC based smart flood management framework," *Sustainable Computing: Informatics and Systems*, vol. 20, pp. 102–117, 2018.

[13] S. H. Almotiri, M. A. Khan, and M. A. Alghamdi, "Mobile health (m-health) system in the context of IoT," in *Proceedings of the 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops*, IEEE, Vienna, Austria, pp. 39–42, August 2016.

[14] K. Alexopoulos, S. Koukas, N. Boli, and D. Mourtzis, "Architecture and development of an industrial internet of things framework for realizing services in industrial product service systems," *Procedia CIRP*, vol. 72, pp. 880–885, 2018.

[15] A. Belhadi, K. Zkik, A. Cherrafi, S. R. M. Yusof, and S. El fezazi, "Understanding big data analytics for manufacturing processes: insights from literature review and multiple case studies," *Computers & Industrial Engineering*, vol. 137, Article ID 106099, 2019.

[16] H. Boyes, B. Hallaq, J. Cunningham, and T. Watson, "The industrial internet of things (IIoT): an analysis framework," *Computers in Industry*, vol. 101, pp. 1–12, 2018.

[17] A. Dwivedi and R. P. Pant, "An algorithmic implementation of entropic ternary reduct soft sentiment set (ETRSSS) using soft computing technique on big data sentiment analysis (BDSA) for optimal selection of a decision based on real-time update in online reviews," *Journal of King Saud University-Computer and Information Sciences*, 2019.

[18] M. S. Hadj Sassi, F. G. Jedidi, and L. C. Fourati, "A new architecture for cognitive internet of things and big data," *Procedia Computer Science*, vol. 159, pp. 534–543, 2019.

[19] L. Urquhart and D. McAuley, "Avoiding the internet of insecure industrial things," *Computer Law & Security Review*, vol. 34, no. 3, pp. 450–466, 2018.

[20] J. Ordieres-Meré, J. Villalba-Díez, and X. Zheng, "Challenges and opportunities for publishing IIoT data in manufacturing as a service business," *Procedia Manufacturing*, vol. 39, pp. 185–193, 2019.

[21] S. Ren, Y. Zhang, Y. Liu, T. Sakao, D. Huisingh, and C. M. V. B. Almeida, "A comprehensive review of big data analytics throughout product life cycle to support sustainable smart manufacturing: a framework, challenges and future research directions," *Journal of Cleaner Production*, vol. 210, pp. 1343–1365, 2019.

[22] M. H. ur Rehman, I. Yaqoob, K. Salah, M. Imran, P. P. Jayaraman, and C. Perera, "The role of big data analytics in industrial internet of things," *Future Generation Computer Systems*, vol. 99, pp. 247–259, 2019.

[23] F. Calza, A. Parmentola, and I. Tutore, "Big data and natural environment. How does different data support different green strategies?" *Sustainable Futures*, vol. 2, Article ID 100029, 2020.

[24] V. F. Brock and H. U. Khan, "Are enterprises ready for big data analytics? A survey-based approach," *International Journal of Business Information Systems*, vol. 25, no. 2, pp. 256–277, 2017.

[25] V. Brock and H. U. Khan, "Big data analytics: does organizational factor matters impact technology acceptance?" vol. 4, no. 1, p. 21, 2017.

[26] B. Liao, Y. Ali, S. Nazir, L. He, and H. U. Khan, "Security analysis of IoT devices by using mobile computing: a systematic literature review," *IEEE Access*, vol. 8, pp. 120331–120350, 2020.

[27] M. Madhuri, A. Q. Gill, and H. U. Khan, "IoT-enabled smart child safety digital system architecture," in *Proceedings of the 2020 IEEE 14th International Conference on Semantic Computing*, IEEE, San Diego, CA, USA, pp. 166–169, 2020.

[28] A. Q. Gill, G. Beydoun, M. Niazi, and H. U. Khan, "Adaptive architecture and principles for securing the IoT systems," in *Proceedings of the International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pp. 173–182, Springer, Lodz, Poland, 2020.

[29] M. Humayun, N. Z. Jhanjhi, A. Alsayat, and V. Ponnusamy, "Internet of things and ransomware: evolution, mitigation and prevention," *Egyptian Informatics Journal*, 2020.

[30] W. Chen, "Intelligent manufacturing production line data monitoring system for industrial internet of things," *Computer Communications*, vol. 151, pp. 31–41, 2020.

[31] J. Chi, Y. Li, J. Huang et al., "A secure and efficient data sharing scheme based on blockchain in industrial internet of things," *Journal of Network and Computer Applications*, vol. 167, Article ID 102710, 2020.

[32] B. DiÃ¨ne, J. J. P. C. Rodrigues, O. Diallo, E. L. H. M. Ndoye, and V. V. Korotaev, "Data management techniques for internet of things," *Mechanical Systems and Signal Processing*, vol. 138, Article ID 106564, 2020.

[33] M. Helu, T. Sprock, D. Hartenstine, R. Venketesh, and W. Sobel, "Scalable data pipeline architecture to support the industrial internet of things," *CIRP Annals*, vol. 69, no. 1, pp. 385–388, 2020.

[34] J. C. Kabugo, S.-L. Jämsä-Jounela, R. Schiemann, and C. Binder, "Industry 4.0 based process data analytics platform: a waste-to-energy plant case study," *International Journal of Electrical Power & Energy Systems*, vol. 115, Article ID 105508, 2020.

[35] W. Liu, G. Huang, A. Zheng, and J. Liu, "Research on the optimization of IIoT data processing latency," *Computer Communications*, vol. 151, pp. 290–298, 2020.

[36] R. Sahal, J. G. Breslin, and M. I. Ali, "Big data and stream processing platforms for Industry 4.0 requirements mapping for a predictive maintenance use case," *Journal of Manufacturing Systems*, vol. 54, pp. 138–151, 2020.

[37] A. Majeed, Y. Zhang, S. Ren et al., "A big data-driven framework for sustainable and smart additive manufacturing," *Robotics and Computer-Integrated Manufacturing*, vol. 67, Article ID 102026, 2021.

[38] D. Shah, J. Wang, and Q. P. He, "Feature engineering in big data analytics for IoT-enabled smart manufacturing - comparison between deep learning and statistical learning," *Computers & Chemical Engineering*, vol. 141, Article ID 106970, 2020.

[39] T. Zhu, S. Dhelim, Z. Zhou, S. Yang, and H. Ning, "An architecture for aggregating information from distributed data nodes for industrial internet of things," *Computers & Electrical Engineering*, vol. 58, pp. 337–349, 2017.

[40] A. Urbán, A. Groniewsky, M. Malý, V. Józsa, and J. Jedelský, "Application of big data analysis technique on high-velocity airblast atomization: searching for optimum probability density function," *Fuel*, vol. 273, Article ID 117792, 2020.

[41] J. Chin, V. Callaghan, and I. Lam, "Understanding and personalising smart city services using machine learning, the Internet-of-Things and Big Data," in *Proceedings of the 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*, pp. 2050–2055, Edinburgh, Scotland, June 2017.

[42] S. Gierej, "The framework of business model in the context of industrial internet of things," *Procedia Engineering*, vol. 182, pp. 206–212, 2017.

[43] X. Feng and Y. Zhao, "Digital forensics challenges to big data in the cloud," in *Proceedings of the 2017 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. 858–862, Exeter, UK, June 2017.

[44] Q. Han, H. Li, W. Dong, Y. Luo, and Y. Xia, "On fault prediction based on industrial big data," in *Proceedings of the 2017 36th Chinese Control Conference (CCC)*, pp. 10127–10131, Dalian, China, July 2017.

[45] M. Jayaratne, D. Alahakoon, D. D. Silva, and X. Yu, "Apache spark based distributed self-organizing map algorithm for sensor data analysis," in *Proceedings of the IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*, pp. 8343–8349, Beijing, China, November 2017.

[46] M. Younan, E. H. Houssein, M. Elhoseny, and A. A. Ali, "Challenges and recommended technologies for the industrial internet of things: a comprehensive review," *Measurement*, vol. 151, p. 107198, 2020.

[47] Y. Karaca, M. Moonis, Y.-D. Zhang, and C. Gezgez, "Mobile cloud computing based stroke healthcare system," *International Journal of Information Management*, vol. 45, pp. 250–261, 2019.

[48] M. V. Moreno, F. Terroso-Saenz, A. Gonzalez-Vidal et al., "Applicability of big data techniques to smart cities deployments," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 800–809, 2017.

[49] D. Jiang, "The construction of smart city information system based on the Internet of Things and cloud computing," *Computer Communications*, vol. 150, pp. 158–166, 2020.

[50] M. Dachyar, T. Y. M. Zagloel, and L. R. Saragih, "Knowledge growth and development: internet of things (IoT) research, 2006â€"2018," *Heliyon*, vol. 5, no. 8, Article ID e02264, 2019.

[51] D. Nallaperuma, D. D. Silva, D. Alahakoon, and X. Yu, "A cognitive data stream mining technique for context-aware IoT systems," in *Proceedings of the IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*, pp. 4777–4782, Beijing, China, November 2017.

[52] A. Bamrungwong, "Implementation of the FSO2 life extension program by using big data and IIoT," in *Proceedings of the 2019 Petroleum and Chemical Industry Conference Europe (PCIC EUROPE)*, pp. 1–8, Paris, France, May 2019.

[53] X. Deng, P. Jiang, X. Peng, and C. Mi, "An intelligent outlier detection method with one class support tucker machine and genetic algorithm toward big sensor data in internet of things," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 6, pp. 4672–4683, 2019.

[54] D. Geng, C. Zhang, C. Xia, X. Xia, Q. Liu, and X. Fu, "Big data-based improved data acquisition and storage system for designing industrial data platform," *IEEE Access*, vol. 7, pp. 44574–44582, 2019.

[55] F. A. B. JuÃ¡rez, "Cybersecurity in an industrial internet of things environment (IIoT) challenges for standards systems and evaluation models," in *Proceedings of the 2019 8th International Conference On Software Process Improvement*, vol. 23, pp. 1–6, Leon, Mexico, October 2019.

[56] W. Liang, M. Tang, J. Long, X. Peng, J. Xu, and K. C. Li, "A secure FaBric blockchain-based data transmission technique for industrial internet-of-things," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3582–3592, 2019.

[57] C.-C. Lin, D.-J. Deng, C.-H. Kuo, and L. Chen, "Concept drift detection and adaption in big imbalance industrial IoT data using an ensemble learning method of offline classifiers," *IEEE Access*, vol. 7, pp. 56198–56207, 2019.

[58] K. K. Lwin, Y. Sekimoto, W. Takeuchi, and K. Zettsu, "City geospatial dashboard: IoT and big data analytics for geospatial solutions provider in disaster management," in *Proceedings of the 2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pp. 1–4, December 2019.

[59] Y. Shoji, K. Nakauchi, W. Liu, Y. Watanabe, K. Maruyama, and K. Okamoto, "A community-based IoT service platform to locally disseminate socially-valuable data :best effort local data sharing network with no conscious effort?" in *Proceedings of the 2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, pp. 724–728, April 2019.

[60] W. Fang, Y. Guo, W. Liao, S. Huang, C. Yang, and K. Cui, "The spatio-temporal modeling and integration of manufacturing big data in job shop: an ontology-based approach," in *IEEE 7th International Conference Oon Industrial Engineering Aand Applications (ICIEA), 16-21 April 2020 2020*, pp. 394–398, 2020.

[61] K. Kaur, S. Garg, G. Kaddoum, E. Bou-Harb, and K.-K. R. Choo, "A bBig dData-eEnabled cConsolidated fFramework for eEnergy eEfficient sSoftware dDefined dData cCenters in IoT sSetups," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2687–2697, 2020.

[62] A. K. Tripathi, K. Sharma, M. Bala, A. Kumar, V. G. Menon, and A. K. Bashir, "A parallel military dog based algorithm for clustering big data in cognitive industrial internet of things," *IEEE Transactions on Industrial Informatics*, p. 1, 2020.

[63] T. Wang, H. Ke, X. Zheng, K. Wang, A. K. Sangaiah, and A. Liu, "Big data cleaning based on mobile edge computing in industrial sensor-cloud," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1321–1329, 2020.

[64] X. Wang, L. T. Yang, L. Song, H. Wang, L. Ren, and J. Deen, "A tensor-based multi-attributes visual feature recognition method for industrial intelligence," *IEEE Transactions on Industrial Informatics*, p. 1, 2020.

[65] C. Yang, S. Lan, L. Wang, W. Shen, and G. G. Q. Huang, "Big data driven edge-cloud collaboration architecture for cloud manufacturing: a software defined perspective," *IEEE Access*, vol. 8, pp. 45938–45950, 2020.

Hindawi
Scientific Programming
Volume 2020, Article ID 8859454, 7 pages
https://doi.org/10.1155/2020/8859454

Research Article

Evaluating the Role of Big Data in IIOT-Industrial Internet of Things for Executing Ranks Using the Analytic Network Process Approach

Xiaoqun Liao,[1] Mohammad Faisal,[2] Qing QingChang,[3,4] and Amjad Ali[5]

[1]*Information and Network Center, Xi'an University of Science and DSTechnology, Xi'an 710054, China*
[2]*Department of Computer Science & IT, University of Malakand, Chakdara, Pakistan*
[3]*School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China*
[4]*School of Information Management, Shanghai Linxin University of Accounting and Finance, 995 Shangchuan Road, Pudong New District, Shanghai 201209, China*
[5]*Department of Computer Science, University of Swat, Mingora, Pakistan*

Correspondence should be addressed to Qing QingChang; changqingqing54@sina.com

Received 29 July 2020; Revised 2 September 2020; Accepted 14 September 2020; Published 20 October 2020

Academic Editor: Habib Ullah Khan

Due to the enhancements of Internet of Things (IoT) and sensors deployments, the production of big data in Industrial Internet of Things (IIoT) is increased. The accessing and processing of big data become a challenging issue due to the limited storage space, computational time, networking, and IoT devices end. IoT and big data are well thought-out to be the key concepts when describing new information architecture projects. The techniques, tools, and methods that help to provide better solutions for IoT and big data can have an important role to play in the architecture of business. Different approaches are being practiced in the literature for evaluating the role of big data in IIoT. These techniques are not handling the situations when complexity of dependency arises among parameters of the alternatives. The proposed research uses the approach of Analytic Network Process (ANP) for evaluating the role of big data in IIoT. The results show that the proposed research works well for evaluating the role of big data in IIoT.

1. Introduction

The developments in the field of data processing, Internet, and electronic communications have enabled the world to easy access to different physical devices. The whole world is covered with different devices embedded with the actuators and sensors. Huge bulk of data is produced from the communication of these heterogeneous devices which need to be researched for mining useful insights. These useful information and insights will play an important role in decision-making and optimum management or services and resources.

The integration of emerging trends, technologies, and applications of IoT in the industrial environment is leading toward the expansion and developments of IIoT. The IIoT is collection of smart devices and objects which sense, collect, process, and communicate real-time occurrences in the industrial setup. IIoT aims to achieve high efficiency of operation to manage industrial assets and increase the productivity of industry with the support of product customization. In the last few years, special attentions were given to the applications of IoT with emerging into the industry. Several IoT devices are brought to facilitate the human life. The applications of IoT technologies in industry of automation have been expended to the IIoT which supports Cyber Physic System in which human and machine interact [1]. Internet of Things has several applications in the daily life and has made life very easy. From industry to education, healthcare, and to other place, the IoT is mostly used. Internet of Medical Things is the advanced version of

IoT which has a key role in healthcare. In healthcare, the devices are sometimes connected through heterogeneous environment with the support of different IoT devices.

IIoT takes help from the communication of IoT in the applications of business which mostly focus on the interoperability among machines. As the daily life objects and things connected to the Internet are increasing with the passage of time, which made the IoT be dynamic network of networks, more challenges such as dynamicity, heterogeneity, volume, and velocity of data makes the services of IoT inconsistent, incomplete, inaccurate, and incorrect results which resultant can make complexity for the applications of IIoT such as healthcare, wearable, smart transportation, and industry [2].

The contribution of the proposed research is to use the approach of Analytic Network Process for evaluating the role of big data in IIoT. The process of pairwise comparison was carried out for the criteria and available alternatives. The results show that the proposed research works well for evaluating the role of big data in IIoT.

The organization of the paper is as follows. Section 2 presents the related work to the role of big data in IIoT. Section 3 shows the research method of the proposed research. The results and discussions are shown in Section 4. The paper is concluded in Section 5.

2. Related Work

With the enhancement in IoT and sensors deployments, the production of big data in IIoT is increased. The accessing and processing of big data become a challenging issue due to the limited storage space, computational time, networking, and IoT device end. Researchers try to use different techniques and tools for evaluating and extracting meaningful insights from such data. Ur Rehman et al. [3] identified technologies of big data analytics, techniques, and algorithms used in the development of intellectual industrial Internet of Things systems. Classification and categorization taxonomy were devised for the literature based on significant parameters such as analytics tools, data sources, requirements, analytics techniques, and industrial applications. Different frameworks and case studies are presented to show the applications of big data analytics in various enterprises. Hadj Sassi et al. [4] proposed an architecture of big data and cognitive Internet of Things. The approach integrates the Data Lake and Data Warehouse and then defines the tool for collection of heterogeneous data. Sunhare et al. [5] presented a comprehensive and systematic review of the different data mining techniques employed in small- and large-scale IoT for formulating an intelligent environment. The system of cloud-based IoT big data mining is also discussed. Yao et al. [6] presented the applications of deep learning model for diagnosing gallbladder stone from Internet of Health Things (IoHT) big data. The research categorises the characteristics of gallstone for improving the presented model to determine the chemical compositions of gallstone. The model obtain smart healthcare data from IoT for supporting diagnose and recommendation of treatment of gallbladder stone with the aim to build smart IoHT.

Khan et al. [7] presented the concept of IIoT in a new way to help reader to understand the IIoT. The existing efforts made for research in the IIoT are described. The study highlights the enabling technologies for IIoT and the challenges to the IIoT. Gulati and Kaur [8] analysed the key opportunities integrated from the concept of SIoT into the industry with proposing reference architecture. An ontological model is designed to present the model from a semantic perspective. For the relationship management, an approach among manufacturing resources is introduced. The authors [1] presented the ideas of data management in IoT, the literature of data management in IoT, the most relevant solutions, and traced open research challenges. Aceto et al. [9] provide the detail description of the key technologies and approaches used in association to Healthcare 4.0, the key applications scenario, benefits, multidisciplinary challenges, and the derivations. Younan et al. [2] presented a study with comprehensive review of the existing challenges in the literature and recommending technologies for enabling the analysis of data and search in the future IoT search engines. Two case studies are presented to show promising growth on smartness and intelligence of applications of IoT based on the integration of information and communication technologies. The applications of smart phone enable the patients to know about their diseases after the analysis in the field of gynaecology and paediatrics [10]. Mobile computing services can be used in IoT by using services of mobile phones, apps, or through M-Health care system [11].

Feldner and Herber [12] presented a qualitative approach for evaluation of IPv6 for IIoT. Interviews were conducted with five experts of IPv6 users. The experimental results indicate the key challenges are (a) tool help and available libraries are immature, (b) user can configure the IP communication manually, and (c) complicated one-fits-all protocol is prevalent. Boyes et al. [13] presented the concept of IIoT and the association to the ideas such as cyber physical systems and Industry 4.0. IoT-related taxonomies were analysed and developed an analysis framework for IIoT that can be used to list and characterise the devices of IIoT when analysing security vulnerability and threats. Alexopoulos et al. [14] presented the IIoT architecture and its development details to support the industrial product service system life cycle. Ge et al. [15] presented a survey on the technologies of big data in different domains of the IoT for facilitating and stimulating sharing of knowledge across the domains of IoT. The similarities and differences among the technologies of big data in different domains with the reusability of technologies in the domain of IoT are discussed. Lin et al. [16] presented an approach of integrating sensing data from diverse sources and equipments to apply on IIoT. The industrial Micro Control Unit is connected to interface with actuator, data sources, and equipments. The experimental results show that IIoT can reduce the problem of heterogeneous protocol and database manufacturing data transmission.

Humayun et al. [17] presented a comprehensive report of the evolution, prevention, and mitigation of Ransomware in the context of IoT. Jiang [18] has presented an approach

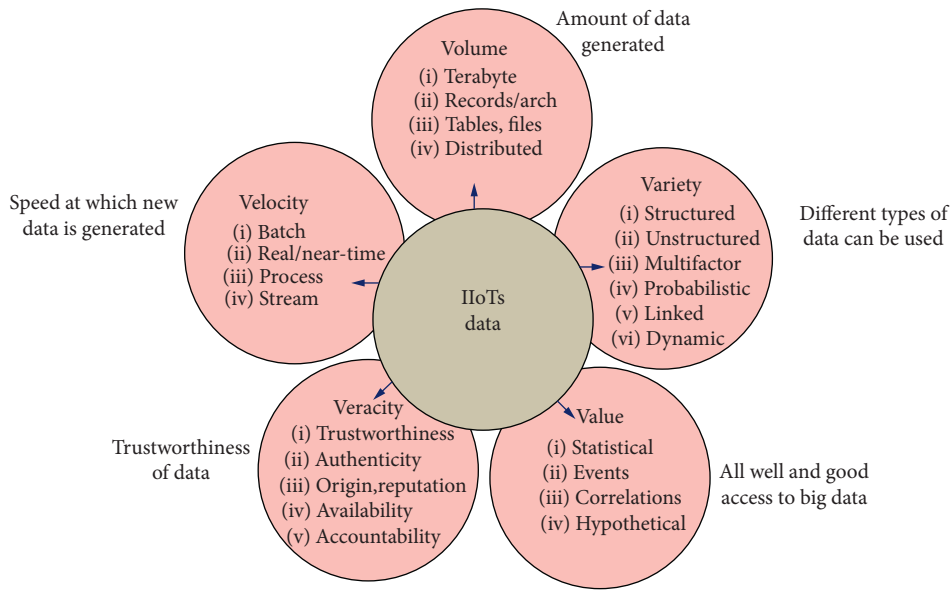


Figure 1: 5 Vs of big data.

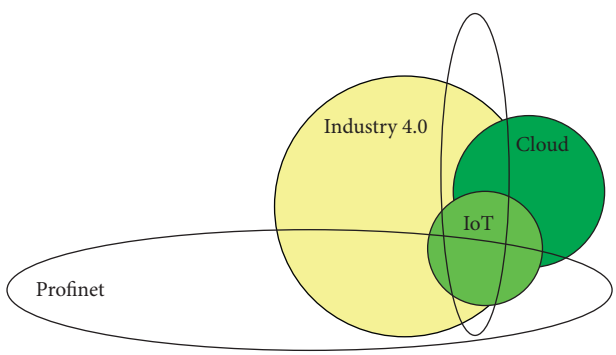


Figure 2: Generic IIoT.

which firstly studies the IoT developments and technologies related to cloud computing and smart cities and then focussed on the IoT technologies and cloud computing. Urquhart and Mcauley [19] presented an approach for the risks which lie for IioT, drawn both on the regulatory and technical perspectives. Gierej [20] presented the idea of business model for the companies implementing IIoT technologies. The approach is developed to help traditional companies in the transition of the digital market. Dachyar et al. [21] conducted in-depth analysis of the 26420 papers published in the area of IoT.

3. Research Method

Various taxonomies such as analytics tools, data sources, analytics techniques, industrial analytics applications, and requirements are used in the literature for analysing the data. Figure 1 shows the 5 Vs of big data.

Figure 2 shows the generic IIoT.

The proposed method has used the Analytic Network Process for the evaluation of role of big data in IIoT. The ANP has several applications in different areas [22–25]. The ANP approach was selected for evaluating the role of IIoT due to the reasons as this method works very well in situation, where complexity of dependencies exists among criteria and alternatives. The ANP method consists of three parts: (a) the goal, (b) criteria, and (c) alternatives. Five criteria along with subcriteria and three conditions were available as alternatives. The details regarding the ANP can be found in saaty (1996), while the following are the main steps of the ANP:

(a) Division of problem into subcriteria

(b) Quantitative scale of measure between 1 and 9

(c) Pairwise comparison for criteria and alternatives to achieve goal

(d) Eigenvalue and the related Eigenvector of the comparison matrix

(e) The consistency of matrix is measured

The "Consistency Index (CI)" and "Consistency Random (CR)" of pairwise comparison matrix are computed by equations (1) and (2):

Table 1: Random consistency index.

	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

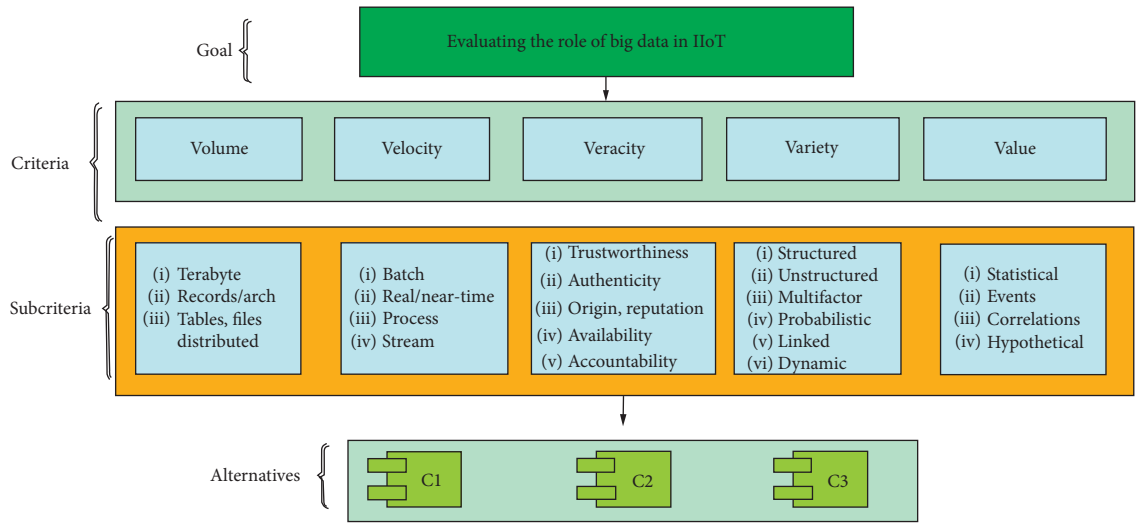


Figure 3: Goal, criteria, and alternatives.

Table 2: Comparison with respect to condition 1.

	Volume	Variety	Value	Veracity	Velocity	E.V
Volume	1	3	2	5	4	0.417
Variety	1/3	1	2	3	2	0.219
Value	1/2	1/2	1	3	3	0.196
Veracity	1/5	1/3	1/3	1	2	0.091
Velocity	1/4	1/2	1/3	1/2	1	0.077

CR = 0.065.

Table 3: Comparison with respect to condition 2.

	Volume	Variety	Value	Veracity	Velocity	E.V
Volume	1	2	3	2	5	0.383
Variety	1/2	1	2	3	2	0.246
Value	1/3	1/2	1	2	3	0.172
Veracity	1/2	1/3	1/2	1	2	0.124
Velocity	1/5	1/2	1/3	1/2	1	0.076

CR = 0.054.

$$C_i = \frac{\lambda_{\max} - n}{n - 1}, \tag{1}$$

$$\mathrm{CR} = \frac{\mathrm{CI}}{\mathrm{RI}}. \tag{2}$$

Random consistency (RI) table is given by satty and is shown in Table 1 [26].

(f) Construction of weighted matrix

(g) Conversion of weighted super matrix into limit matrix

Table 4: Comparison with respect to condition 3.

	Volume	Variety	Value	Veracity	Velocity	E.V
Volume	1	5	2	3	7	0.453
Variety	1/5	1	2	2	3	0.196
Value	1/2	1/2	1	3	2	0.184
Veracity	1/3	1/2	1/3	1	2	0.103
Velocity	1/7	1/3	1/2	1/2	1	0.063

CR = 0.095.

(h) Decision of the alternatives

Figure 3 shows the goal, criteria, and alternatives of the proposed research.

4. Results and Discussion

The process of pairwise comparisons of the proposed study was carried out in order to decide about available alternatives. The experts from the industry involved to check the contents against the available alternatives and relevant score should be given. Table 2 shows the pairwise comparison with respect to condition 1 for the available criteria.

Table 3 shows the pairwise comparison with respect to condition 2 for the available criteria.

Table 4 shows the pairwise comparison with respect to condition 3 for the available criteria.

Table 5 shows the pairwise comparison with respect to criteria (volume) for the available alternatives (conditions).

Table 6 shows the pairwise comparison with respect to criteria (variety) for the available alternatives (conditions).

Table 7 shows the pairwise comparison with respect to criteria (value) for the available alternatives (conditions).

Table 5: Comparison with respect to volume.

	Condition 1	Condition 2	Condition 3	E.V
Condition 1	1	3	4	0.623
Condition 2	1/3	1	2	0.239
Condition 3	1/4	1/2	1	0.137

CR = 0.22.

Table 6: Comparison with respect to variety.

	Condition 1	Condition 2	Condition 3	E.V
Condition 1	1	2	2	0.490
Condition 2	1/2	1	2	0.312
Condition 3	1/2	1/2	1	0.198

CR = 0.052.

Table 7: Comparison with respect to value.

	Condition 1	Condition 2	Condition 3	E.V
Condition 1	1	4	2	0.579
Condition 2	1/4	1	1	0.187
Condition 3	1/2	1	1	0.234

CR = 0.069.

Table 8: Comparison with respect to veracity.

	Condition 1	Condition 2	Condition 3	E.V
Condition 1	1	3	4	0.623
Condition 2	1/3	1	2	0.239
Condition 3	1/4	1/2	1	0.137

CR = 0.022.

Table 9: Comparison with respect to velocity.

	Condition 1	Condition 2	Condition 3	E.V
Condition 1	1	2	3	0.525
Condition 2	1/2	1	3	0.334
Condition 3	1/3	1/3	1	0.142

CR = 0.056.

Table 10: Weighted super matrix.

Node lable		IIoTs big data					Alternatives		
		Volume	Variety	Value	Veracity	Velocity	Condition 1	Condition 2	Condition 3
IIoTs big data	Volume	0.000	0.000	0.000	0.000	0.000	0.417	0.383	0.453
	Variety	0.000	0.000	0.000	0.000	0.000	0.219	0.246	0.196
	Value	0.000	0.000	0.000	0.000	0.000	0.196	0.172	0.184
	Veracity	0.000	0.000	0.000	0.000	0.000	0.091	0.124	0.103
	Velocity	0.000	0.000	0.000	0.000	0.000	0.077	0.076	0.063
Alternatives	Condition 1	0.623	0.490	0.579	0.623	0.525	0.000	0.000	0.000
	Condition 2	0.239	0.312	0.187	0.239	0.334	0.000	0.000	0.000
	Condition 3	0.137	0.198	0.234	0.137	0.142	0.000	0.000	0.000

Table 11: Limit matrix.

Node lable		IIoTs big data					Alternatives		
		Volume	Variety	Value	Veracity	Velocity	Condition 1	Condition 2	Condition 3
IIoTs big data	Volume	0.41	0.41	0.41	0.41	0.41	0.00	0.00	0.00
	Variety	0.22	0.22	0.22	0.22	0.22	0.00	0.00	0.00
	Value	0.19	0.19	0.19	0.19	0.19	0.00	0.00	0.00
	Veracity	0.10	0.10	0.10	0.10	0.10	0.00	0.00	0.00
	Velocity	0.07	0.07	0.07	0.07	0.07	0.00	0.00	0.00
Alternatives	Condition 1	0.00	0.00	0.00	0.00	0.00	0.58	0.58	0.58
	Condition 2	0.00	0.00	0.00	0.00	0.00	0.25	0.25	0.25
	Condition 3	0.00	0.00	0.00	0.00	0.00	0.17	0.17	0.17

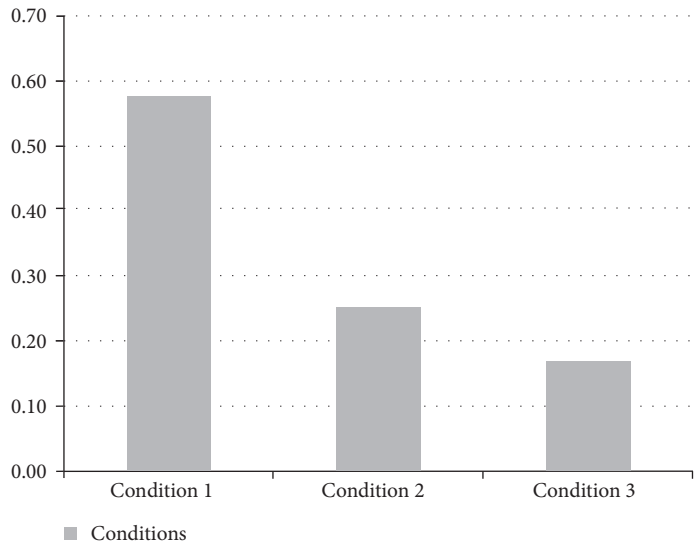


Figure 4: Ranking of available alternatives.

Table 8 shows the pairwise comparison with respect to criteria (veracity) for the available alternatives (conditions).

Table 9 shows the pairwise comparison with respect to criteria (velocity) for the available alternatives (conditions).

After pairwise comparisons, the E.V values of all the tables were merged into the weighted matrix. Table 10 shows the weighted matrix.

The weighted matrix (Table 10) was converted into the limit matrix in order to get the most appropriate condition in the available alternatives. Table 11 shows the limit matrix of the proposed study.

From limit matrix (Table 11), it is concluded that condition 1 is the best option among the available alternatives followed by condition 2, and then condition 3. Figure 4 shows the decision of the available alternatives with their rank.

5. Conclusion

In recent years, the accessing and processing of big data become a challenging issue due to the limited storage space, computational time, networking, and IoT device end. IoT and big data are well thought-out to be the key concepts when describing new information architecture projects. With the enhancements and development in the Internet of Things and sensor deployments, the production of big data in Industrial Internet of Things is increased. Researchers are using different approaches and techniques for evaluating the role of big data in IIoT. These techniques are not handling the situations when complexity of dependency arises among parameters of the alternatives. The contribution of the proposed research is to use the approach of Analytic Network Process for evaluating the role of big data in IIoT. The process of pairwise comparison was carried out for the criteria and available alternatives. The results show that the proposed research works well for evaluating the role of big data in IIoT.

Data Availability

No data are available

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Science and Technology Project of State Grid Xizang Electric Power Co., Ltd. (SGXZJY00JHJS2000007), Influence of Energy Storage Technology Application on Power Grid, and Science and Technology Project of State Grid Zizang Electric Power Co., Ltd. (SGXZJY00JHJS2000008), Research Technology Service of Multi Energy Complementary Demonstration Application.

References

[1] B. Diã¨Ne, J. J. P. C. Rodrigues, O. Diallo, E. L. H. M. Ndoye, and V. V. Korotaev, "Data management techniques for Internet of Things," *Mechanical Systems and Signal Processing*, vol. 138, Article ID 106564, 2020.

[2] M. Younan, E. H. Houssein, M. Elhoseny, and A. A. Ali, "Challenges and recommended technologies for the industrial Internet of Things: a comprehensive review," *Measurement*, vol. 151, Article ID 107198, 2020.

[3] M. H. Ur Rehman, I. Yaqoob, K. Salah, M. Imran, P. P. Jayaraman, and C. Perera, "The role of big data analytics in industrial Internet of Things," *Future Generation Computer Systems*, vol. 99, pp. 247–259, 2019.

[4] M. S. Hadj Sassi, F. G. Jedidi, and L. C. Fourati, "A new architecture for cognitive Internet of Things and big data," *Procedia Computer Science*, vol. 159, pp. 534–543, 2019.

[5] P. Sunhare, R. R. Chowdhary, and M. K. Chattopadhyay, "Internet of Things and data mining: an application oriented survey," *Journal of King Saud University-Computer and Information Sciences*, 2020.

[6] C. Yao, S. Wu, Z. Liu, and P. Li, "A deep learning model for predicting chemical composition of gallstones with big data in medical Internet of Things," *Future Generation Computer Systems*, vol. 94, pp. 140–147, 2019.

[7] W. Z. Khan, M. H. Rehman, H. M. Zangoti, M. K. Afzal, N. Armi, and K. Salah, "Industrial Internet of Things: recent advances, enabling technologies and open challenges," *Computers & Electrical Engineering*, vol. 81, Article ID 106522, 2020.

[8] N. Gulati and P. D. Kaur, "Towards socially enabled internet of industrial things: architecture, semantic model and relationship management," *Ad Hoc Networks*, vol. 91, Article ID 101869, 2019.

[9] G. Aceto, V. Persico, and A. Pescapé, "Industry 4.0 and health: Internet of Things, big data, and cloud computing for healthcare 4.0," *Journal of Industrial Information Integration*, vol. 18, Article ID 100129, 2020.

[10] Y. Karaca, M. Moonis, Y.-D. Zhang, and C. Gezgez, "Mobile cloud computing based stroke healthcare system," *International Journal of Information Management*, vol. 45, pp. 250–261, 2019.

[11] S. H. Almotiri, M. A. Khan, and M. A. Alghamdi, "Mobile health (*m*-health) system in the context of IoT," in *Proceedings of the 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pp. 39–42, Vienna, Austria, August 2016.

[12] B. Feldner and P. Herber, "A qualitative evaluation of IPv6 for the industrial Internet of Things," *Procedia Computer Science*, vol. 134, pp. 377–384, 2018.

[13] H. Boyes, B. Hallaq, J. Cunningham, and T. Watson, "The industrial internet of things (IIoT): an analysis framework," *Computers in Industry*, vol. 101, pp. 1–12, 2018.

[14] K. Alexopoulos, S. Koukas, N. Boli, and D. Mourtzis, "Architecture and development of an industrial internet of things framework for realizing services in industrial product service systems," *Procedia CIRP*, vol. 72, pp. 880–885, 2018.

[15] M. Ge, H. Bangui, and B. Buhnova, "Big data for Internet of Things: a survey," *Future Generation Computer Systems*, vol. 87, pp. 601–614, 2018.

[16] Y. J. Lin, C.-B. Lan, and C.-Y. Huang, "A realization of cyber-physical manufacturing control system through Industrial Internet of Things," *Procedia Manufacturing*, vol. 39, pp. 287–293, 2019.

[17] M. Humayun, N. Z. Jhanjhi, A. Alsayat, and V. Ponnusamy, "Internet of Things and ransomware: evolution, mitigation and prevention," *Egyptian Informatics Journal*, 2020.

[18] D. Jiang, "The construction of smart city information system based on the Internet of Things and cloud computing," *Computer Communications*, vol. 150, pp. 158–166, 2020.

[19] L. Urquhart and D. Mcauley, "Avoiding the internet of insecure industrial things," *Computer Law & Security Review*, vol. 34, no. 3, pp. 450–466, 2018.

[20] S. Gierej, "The framework of business model in the context of Industrial Internet of Things," *Procedia Engineering*, vol. 182, pp. 206–212, 2017.

[21] M. Dachyar, T. Y. M. Zagloel, and L. R. Saragih, "Knowledge growth and development: internet of things (IoT) research, 2006–2018," *Heliyon*, vol. 5, no. 8, Article ID e02264, 2019.

[22] S. Nazir, S. Shahzad, I. Zada, and H. Khan, "Evaluation of software birthmarks using fuzzy analytic hierarchy process," in *Proceedings of the Fourth International Multi-Topic Conference*, pp. 171–175, Jamshoro, Pakistan, Decembe 2015.

[23] S. Nazir, S. Shahzad, Z. Hussain, M. Iqbal, and A. Keerio, "Evaluating student grades using analytic network process," *Sindh University Research Journal (Science Series)*, vol. 47, pp. 1–5, 2015.

[24] S. Nazir, S. Anwar, S. A. Khan et al., "Software component selection based on quality criteria using the analytic network process," *Abstract and Applied Analysis*, vol. 2014, Article ID 535970, 12 pages, 2014.

[25] S. Nazir, S. Shahzad, M. Nazir, and H. U. Rehman, "Evaluating security of software components using analytic network process," in *Proceedings of the 11th International Conference on Frontiers of Information Technology (FIT)*, IEEE, Islamabad, Pakistan, pp. 183–188, 2013.

[26] T. L. Saaty, "Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process," *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, vol. 102, no. 2, pp. 251–318, 2008.

Hindawi
Scientific Programming
Volume 2020, Article ID 8850096, 7 pages
https://doi.org/10.1155/2020/8850096

Research Article

Big Data, Scientific Programming, and Its Role in Internet of Industrial Things: A Decision Support System

Ju Li,[1] **Muhammad Nazir Jan**,[2] **and Mohammad Faisal**[3]

[1]*Chongqing Technology and Business Institute, Chongqing 401520, China*
[2]*Department of Computer Science, University of Swabi, Swabi, Pakistan*
[3]*Department of Computer Science and IT, University of Malakand, Chakdara, KP, Pakistan*

Correspondence should be addressed to Ju Li; liju87031@gmail.com

Received 17 August 2020; Revised 11 September 2020; Accepted 15 September 2020; Published 6 October 2020

Academic Editor: Habib Ullah Khan

Big data is a challenging issue as its volume, shape, and size need to be modified in order to extract important information for a specific purpose. The amount of data is rising with the passage of time. This increase in volume can be a challenging issue to analyze the data for smooth industry and the Internet of things. Several tools, techniques, and mechanisms are available to support the handling and management process of such data. Decision support systems can be one of the important techniques which can support big data in order to make decisions on time. The proposed study presents a decision support system to deal with big data and scientific programming for the Industrial Internet of Things. The study has used the tool of SuperDecisions to plot the hierarchy of situations of big data and scientific programming and to select the best alternative among the available.

1. Introduction

Big data is termed to be a hot area of research which needs to be shaped in order to derive and extract meaningful information for the specific purpose of research. The data exist in different forms including structured, unstructured, and semistructured. Deriving and extracting meaningful insights from data is quite difficult. Several tools and techniques exist to overcome these deficiencies. Still, researchers try to come with a solution to extract meaningful information from the data of Industrial Internet of Things (IIoTs) in an effective and efficient way.

Decision-making based on multicriteria is one of the most efficient problems solving mean to select appropriate decision among the number of choices. Due to its effectiveness and potential, it is exploited in various domains such as computer science and IT, agriculture, and business sector. The research finds novel means to make the decision support system for the problems of various application domains by using multiple criteria in integration with machine learning and artificial intelligence. Decision support system plays an important role in real life [1–7]. It is one of the major and difficult tasks which ultimately results in success or failure. For example, in the context of a business or organization, the decision support system serves the mid- and higher-level management and assists people in making decision about certain problems that rapidly change and also not easy to be specified. Decision-making becomes more difficult in situations where it is based on multiple criteria. The current approaches for solving issue of decision support system is capable with power of multicriteria decision support system in order to support the decision makers in taking the right decision in situation of complexity. Researchers try to use multicriteria-based decision support system to integrate the effectiveness of power of machine learning algorithms to provide an intelligent decision-making alternative [8–10]. Decision support systems can be exploited in almost any domain to solving decision problems. Various domains exploit the theories and methods used for decision-making based on simple to more advanced [11, 12].

The contribution of the proposed study is to offer a decision support system for selecting the most appropriate alternative (vendor) from the available choices. The tool of

SuperDecisions was considered for experimental process of the proposed research. The goal of the research was based on the defined criteria for selecting the appropriate choice. The results show that the method is effective for decision-making of selecting vendor.

The organization of the paper is as follows: Section 2 represents the related work to the proposed research. Section 3 shows the research methodology with experimental setup and use of tool for decision support system. The paper is concluded in Section 4.

2. Related Work

Several studies exist related to the big data and its scientific programming for Industrial Internet of Things. In this examination, the authors proposed a blockchain-dependent data sharing scheme that entirely considers efficiency as well as security of data sharing. In this plan, a hyperledger fabric and identity authentication-dependent secure data sharing structure was designed for the data sharing security. Additionally, a network recognition algorithm was proposed to partition the customers into various data sharing networks as per the comparability of mark data. The exploratory outcomes demonstrate that the proposed conspire is successful for efficient and secure data sharing among various customers [13]. This work presents a procedure data examination stage worked around the idea of industry 4.0 [14]. The platform uses the big data software tools, ML algorithms, and state-of-the-art IIoT platforms. The displaying results indicated that, in situations where process information about the procedure wonders within reach is restricted, information-driven delicate sensors are helpful instruments for predictive data investigation. In this article, a novel model is developed in the perspective of manufacturing progression that reviews the key big data analytics (BDA) capabilities. The findings are beneficial for the companies in order to understand big data potential implications as well as their analytics capabilities for their manufacturing processes and efficient BDA-enabler infrastructure design [15]. In this study, for big data and cognitive Internet of things, a new architecture is proposed. The planned architecture helps the computing systems through combining Data Lake and WareHouse, as well as for the collection of heterogeneous data, a tool is defined [16].

In this article, for industrial data processing, an Industrial Internet of Things cloud-fog hybrid network framework was proposed. The results have shown that the proposed framework reduces effectively the delay processing of industrial data [17]. In this study, a systematic strategy was used to review the weaknesses as well as strengths of open-source technologies for stream processing and big data to set up its usage for industry 4.0 use cases [18]. In this study, functional and structural properties of cloud manufacturing were analyzed, and a business intelligence architecture was proposed that plans to empower distributing pertinent KPIs identified with intrigued process data, with the helpful layer of dependability [19]. This paper examines the current big data analytics technologies, strategies, as well as algorithms that can prompt the improvement of insightful Industrial Internet of Things frameworks. We devise a scientific classification by characterizing and classifying the literature based on essential factors (for example, analytics types, industrial analytics applications, requirements, analytics techniques, analytics tools, and data sources). The case studies and frameworks of the different endeavours were presented that have profited by BDA [20].

Table 1: Different areas of decision support system.

Application domain	Reference
Business sector	[9, 21–23]
Computer field	[24–27]
General problem solving	[28–34]
Industrial domain	[35–40]
Medical field	[8, 41, 42]
People safety	[43, 44]
Supply chain	[10, 45–47]
Sustainable computing	[48]
Waste management	[49–51]
Energy sector	[52–54]
Disaster management	[55, 56]
Information about DSS	[57, 58]
Environmental side	[59]

3. Research Method

With the passage of time, the size and volume of data are increasing. These data will reach a situation where their management and analysis will be a challenging issue. For the analysis and management of big data, there is need of tools and techniques to properly analyze, organize, and extract meaningful information for a specific purpose. The role of decision support system is obvious in different areas of research. Decision support system is a system where the decision can be made based on some criteria to evaluate particular circumstances. The proposed research identified some of the areas where decision support system is helpful in making the decision. Table 1 shows some of the research areas of decision support system.

Figure 1 represents some of the research area domains of the decision support system along with the number of publications.

The current study proposes decision support system in order to select the most appropriate vendor from the available alternatives. The tool of SuperDecisions was used for the experimental process of the proposed study. The goal of the study is to select the most suitable alternative based on some defined criteria. Figure 2 shows the hierarchy followed for plotting the goal, criteria, and alternatives of the proposed study. The attributes of criteria are selected generally which cover most of the criteria of the vendor for the purpose of selection.

The process of giving suitable weights to each criterion against alternatives, and vice versa, was given through experts in the field. Figure 3 shows criteria with their scores for the alternative (Vendor 1). The consistency ratio was checked for Vendor 1 which was calculated as 0.094.

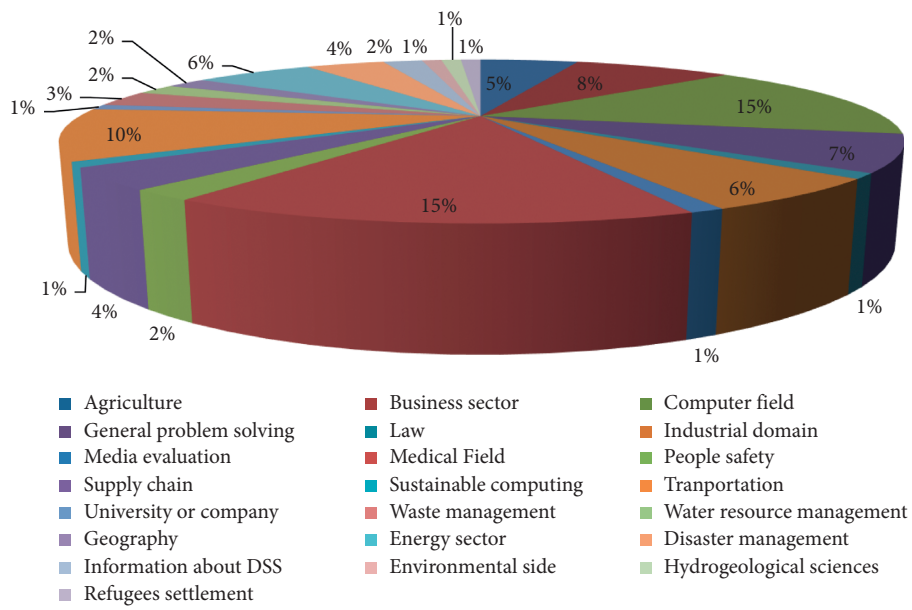


Figure 1: Applications of DSS in different domains along with the number of publications.

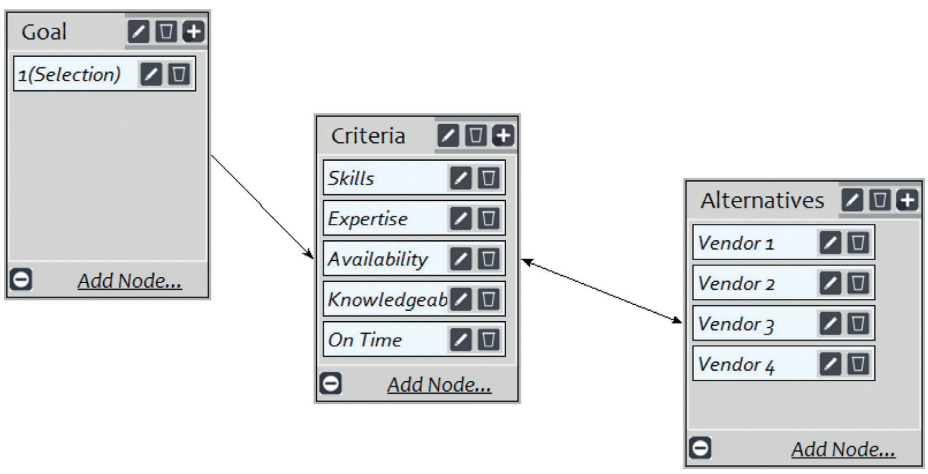


Figure 2: Hierarchy of the proposed study using SuperDecision.

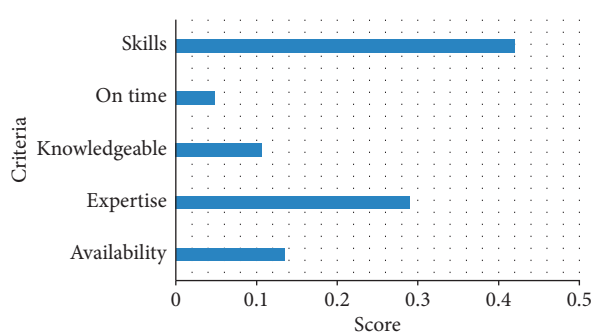


Figure 3: Criteria with their scores for the alternative (Vendor 1).

Figure 4 shows criteria with their scores for the alternative (Vendor 2). The consistency ratio was checked for Vendor 2 which was calculated as 0.043.

Figure 5 shows criteria with their scores for the alternative (Vendor 3). The consistency ratio was checked for vendor 3 which was calculated as 0.045.

Figure 6 shows criteria with their scores for the alternative (Vendor 4). The consistency ratio was checked for Vendor 4 which was calculated as 0.082.

After calculating all the pairwise comparisons for the goal, criteria, and alternatives, a summarized matrix of unweighted and weighted matrix were obtained. These were then converted to limit matrix which is shown in Table 2.

The overall score for the goal, criteria, and alternatives are shown in Figure 7. This figure shows the normalization by cluster and the limitation of the calculations.

Figure 8 shows the final score of the vendors for the ideal, normal, and raw cases. This was obtained from the overall process of calculations. From this, it is concluded that

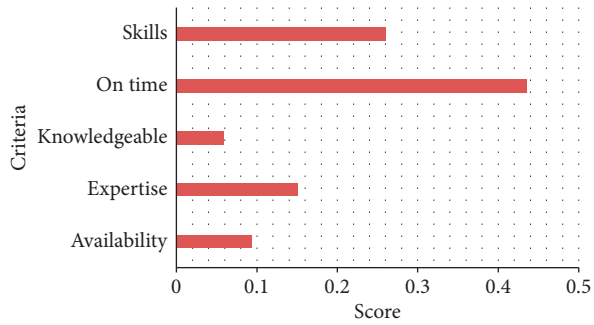


Figure 4: Criteria with their scores for the alternative (Vendor 2).

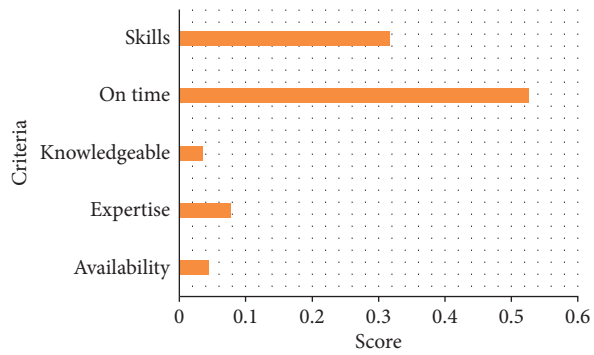


Figure 5: Criteria with their scores for the alternative (Vendor 3).

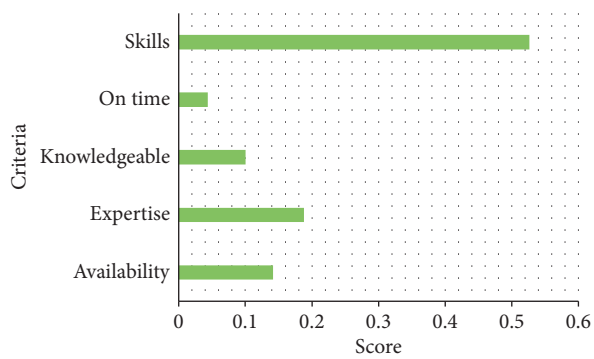


Figure 6: Criteria with their scores for the alternative (Vendor 3).

Table 2: Limit matrix.

	Vendor 1	Vendor 2	Vendor 3	Vendor 4	Availability	Expertise	Knowledgeable	On time	Skills	Selection
Vendor 1	0.000	0.000	0.000	0.000	0.463	0.402	0.476	0.485	0.651	0.000
Vendor 2	0.000	0.000	0.000	0.000	0.275	0.325	0.254	0.276	0.167	0.000
Vendor 3	0.000	0.000	0.000	0.000	0.177	0.157	0.177	0.153	0.119	0.000
Vendor 4	0.000	0.000	0.000	0.000	0.085	0.115	0.093	0.087	0.063	0.000
Availability	0.135	0.093	0.044	0.142	0.000	0.000	0.000	0.000	0.000	0.136
Expertise	0.290	0.151	0.078	0.188	0.000	0.000	0.000	0.000	0.000	0.302
Knowledgeable	0.107	0.059	0.036	0.101	0.000	0.000	0.000	0.000	0.000	0.068
On time	0.048	0.436	0.526	0.044	0.000	0.000	0.000	0.000	0.000	0.035
Skills	0.420	0.261	0.317	0.526	0.000	0.000	0.000	0.000	0.000	0.459
1 (selection)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

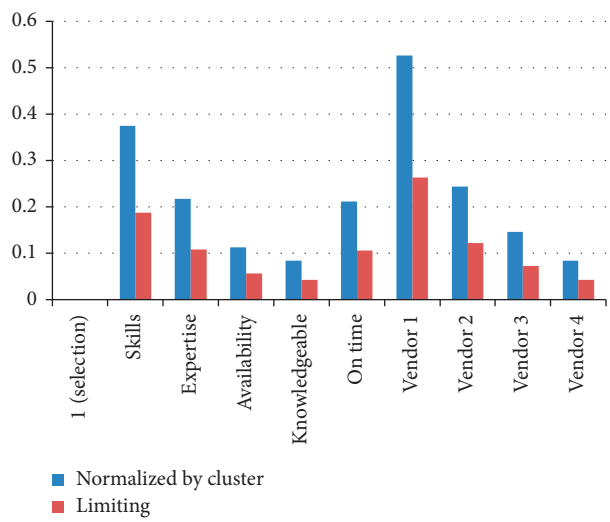


Figure 7: Overall score of goal, criteria, and alternative.

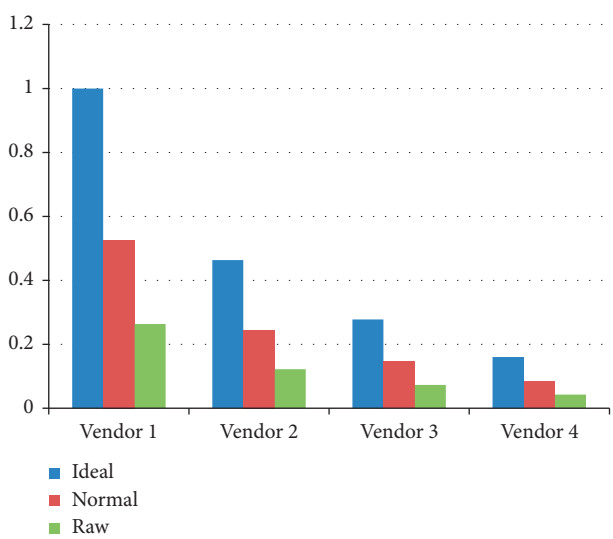


Figure 8: Calculated score of vendors for three cases (ideal, normal, and raw).

Vendor 1 is the best choice among the available alternatives followed by Vendor 2, and so on.

4. Conclusions

With the advancements in technology and communication of different devices, the size and volume of data are increasing with the passage of time. For the analysis and management of big data, there is a need of tools and techniques to properly analyze, organize, and extract meaningful information for a specific purpose. Big data is a challenging issue as its volume, shape, and size need to be modified in order to extract important information for a specific purpose. The amount of data is rising with the passage of time. Several tools, techniques, and mechanisms are available to support the handling and management process of such data. Decision support systems can be one of the important techniques which can support big data in order to make decisions on time. The proposed study presents a decision support system to deal with big data and scientific programming for the Industrial Internet of Things by using the tool of SuperDecisions to plot the hierarchy of situations of big data and scientific programming and to select the best alternative among the available. Results of the experiments show that the proposed decision support system is effective in terms of selecting the most appropriate alternative which is vendor in situation of multicriteria.

Data Availability

No data are available.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors acknowledge the Chongqing Municipal Education Commission Science and Technology Research Program and Computer System Robot Gas Detection Research Funding Project (KJQN201804006).

References

[1] T. Ahmad, Y. Ma, M. Yahya, B. Ahmad, S. Nazir, and A. U. Haq, "Object detection through modified YOLO neural network, an intelligent decision support system," *Scientific Programming*, vol. 2020, Article ID 8403262, 10 pages, 2020.

[2] H.-U. Rahman, H. K. Bamma, S. Nazir, S. Shahzad, and T. Hodosi, "A sourcing decision model for application maintenance services," in *Proceedings of the 3rd International Conference on Science in Information Technology (ICSITech)*, Bandung, Indonesia, October 2017.

[3] A. Khan, J. Li, A. U. Haq et al., "Partial observer decision process model for crane-robot action," *Scientific Programming*, vol. 2020, Article ID 6349342, 14 pages, 2020.

[4] S. Nazir, S. Ali, M. Yang, and Q. Xu, "Deep learning algorithms and multi-criteria decision making used in big data—a systematic literature review," *Security and Communication Networks*, vol. 2020, Article ID 2836064, 18 pages, 2020.

[5] S. Nazir, S. Shahzad, S. Mahfooz, and M. N. Jan, "Fuzzy logic based decision support system for component security evaluation," *International Arab Journal of Information and Technology*, vol. 15, pp. 1–9, 2018.

[6] S. Nazir, S. Shahzad, A. Ullah, and A. Hussain, "Identification and analysis of project attributes affecting the decision of requirement elicitation technique," in *Proceedings of the 2017 National Graduate Conference*, Islamabad, Pakistan, April 2017.

[7] J. Zhang, S. Nazir, A. Huang, and A. Alharbi, "Multicriteria decision and machine learning algorithms for component security evaluation: library-based overview," *Security and Communication Networks*, vol. 2020, Article ID 8886877, 14 pages, 2020.

[8] S. Safdar, S. Zafar, N. Zafar, and N. F. Khan, "Machine learning based decision support systems (DSS) for heart disease diagnosis: a review," *Artificial Intelligence Review*, vol. 50, no. 4, pp. 597–623, 2018.

[9] I. Aouadni and A. Rebai, "Decision support system based on genetic algorithm and multi-criteria satisfaction analysis (MUSA) method for measuring job satisfaction," *Annals of Operations Research*, vol. 256, pp. 3–20, 2017.

[10] M. Jemmali, M. Alharbi, and L. K. B. Melhim, "Intelligent decision-making algorithm for supplier evaluation based on multi-criteria preferences," in *Proceedings of the 2018 1st International Conference on Computer Applications & Information Security (ICCAIS)*, pp. 1–5, Riyadh, Saudi Arabia, April 2018.

[11] L. S. R. Supriadi and L. Sui Pheng, "Knowledge based decision support system (KBDSS)," in *Business Continuity Management in Construction Singapore*, pp. 155–174, Springer, Singapore, 2018.

[12] A. Alaeddini and K. G. Murty, "DSS (decision support system) for allocating appointment times to calling patients at a medical facility,""DSS (decision support system) for allocating appointment times to calling patients at a medical facility," in *Case Studies in Operations Research: Applications of Optimal Decision Making*, K. G. Murty, Ed., pp. 83–109, Springer, New York, NY, USA, 2015.

[13] J. Chi, Y. Li, J. Huang et al., "A secure and efficient data sharing scheme based on blockchain in industrial internet of things," *Journal of Network and Computer Applications*, vol. 167, Article ID 102710, 2020.

[14] J. C. Kabugo, S.-L. Jämsä-Jounela, R. Schiemann, and C. Binder, "Industry 4.0 based process data analytics platform: a waste-to-energy plant case study," *International Journal of Electrical Power & Energy Systems*, vol. 115, Article ID 105508, 2020.

[15] A. Belhadi, K. Zkik, A. Cherrafi, S. R. M. Yusof, and S. El Fezazi, "Understanding big data analytics for manufacturing processes: insights from literature review and multiple case studies," *Computers & Industrial Engineering*, vol. 137, Article ID 106099, 2019.

[16] M. S. Hadj Sassi, F. G. Jedidi, and L. C. Fourati, "A new architecture for cognitive internet of things and big data," *Procedia Computer Science*, vol. 159, pp. 534–543, 2019.

[17] W. Liu, G. Huang, A. Zheng, and J. Liu, "Research on the optimization of IIoT data processing latency," *Computer Communications*, vol. 151, pp. 290–298, 2020.

[18] R. Sahal, J. G. Breslin, and M. I. Ali, "Big data and stream processing platforms for industry 4.0 requirements mapping for a predictive maintenance use case," *Journal of Manufacturing Systems*, vol. 54, pp. 138–151, 2020.

[19] J. Ordieres-Meré, J. Villalba-Díez, and X. Zheng, "Challenges and opportunities for publishing IIoT data in manufacturing as a service business," *Procedia Manufacturing*, vol. 39, pp. 185–193, 2019.

[20] M. H. Ur Rehman, I. Yaqoob, K. Salah, M. Imran, P. P. Jayaraman, and C. Perera, "The role of big data analytics in industrial internet of things," *Future Generation Computer Systems*, vol. 99, pp. 247–259, 2019.

[21] X. Fu, X.-J. Zeng, X. Luo, D. Wang, D. Xu, and Q.-L. Fan, "Designing an intelligent decision support system for effective negotiation pricing: a systematic and learning approach," *Decision Support Systems*, vol. 96, pp. 49–66, 2017.

[22] J. Korczak, H. Dudycz, B. Nita, P. Oleksyk, and A. Kaźmierczak, "Extension of intelligence of decision support systems: manager perspective," in *Information Technology for Management: New Ideas and Real Solutions*, pp. 35–48, Springer, Cham, Switzerland, 2017.

[23] G. Silahtaroğlu, "Implementing adaptive strategies of decision support systems during crises,""Implementing adaptive strategies of decision support systems during crises," in *Global Business Strategies in Crisis: Strategic Thinking and Development*, Ü. Hacioğlu, H. Dinçer, and N. Alayoğlu, Eds., Springer International Publishing, Cham, Switzerland, pp. 287–302, 2017.

[24] A. Kaklauskas, "Intelligent decision support systems," in *Biometric and Intelligent Decision Making Support*, pp. 31–85, Springer International Publishing, Cham, Switzerland, 2015.

[25] A. Kaklauskas and R. Gudauskas, "Intelligent decision-support systems and the internet of things for the smart built environment," in *Start-Up Creation*, pp. 413–449, Woodhead Publishing, Cambridge, UK, 2016.

[26] J. C. Leyva López, P. A. Álvarez Carrillo, D. A. Gastélum Chavira, and J. J. Solano Noriega, "A web-based group decision support system for multicriteria ranking problems," *Operational Research*, vol. 17, no. 2, pp. 499–534, 2017.

[27] A. Martin, P. Zarate, and G. Camillieri, "A multi-criteria recommender system based on users' profile management,""A multi-criteria recommender system based on users' profile management," in *Multiple Criteria Decision Making: Applications in Management and Engineering*, C. Zopounidis and M. Doumpos, Eds., Springer International Publishing, Cham, Switzerland, pp. 83–98, 2017.

[28] T. Bakshi, A. Sinharay, B. Sarkar, and S. K. Sanyal, "A new DST-belief theoretic project selection model for multi-criteria decision support system," *Journal of the Institution of Engineers (India): Series C*, vol. 96, no. 4, pp. 337–349, 2015.

[29] O. E. Bukharov and D. P. Bogolyubov, "Development of a decision support system based on neural networks and a genetic algorithm," *Expert Systems with Applications*, vol. 42, no. 15-16, pp. 6177–6183, 2015.

[30] M. Kalinina, "Multi criteria decision support system: preference information and robustness," in *New Contributions in Information Systems and Technologies*, pp. 641–651, Springer International Publishing, Cham, Switzerland, 2015.

[31] C.-S. Wang, H.-L. Yang, and S.-L. Lin, "To make good decision: a group DSS for multiple criteria alternative rank and selection," *Mathematical Problems in Engineering*, vol. 2015, Article ID 186970, 15 pages, 2015.

[32] V. Ferretti and G. Montibeller, "Key challenges and meta-choices in designing and applying multi-criteria spatial decision support systems," *Decision Support Systems*, vol. 84, pp. 41–52, 2016.

[33] A. Deaconescu, "Decision support system based on robust design methods," in *Proceedings of the 2017 4th International Conference on Biomedical and Bioinformatics Engineering—ICBBE 2017*, November 2017.

[34] S. Bouzayane and I. Saad, "Intelligent multicriteria decision support system for a periodic prediction," in *Decision Support Systems IX: Main Developments and Future Trends*, pp. 97–110, Springer International Publishing, Cham, Switzerland, 2019.

[35] I. Karlsson, A. H. C. Ng, A. Syberfeldt, and S. Bandaru, "An interactive decision support system using simulation-based optimization and data mining," in *Proceedings of the 2015 Winter Simulation Conference (WSC)*, December 2015.

[36] R. Sadeghian and M. R. Sadeghian, "A decision support system based on artificial neural network and fuzzy analytic network process for selection of machine tools in a flexible manufacturing system," *The International Journal of Advanced Manufacturing Technology*, vol. 82, no. 9–12, pp. 1795–1803, 2016.

[37] M. Gomes, F. Silva, F. Ferraz, A. Silva, C. Analide, and P. Novais, "Developing an ambient intelligent-based decision

support system for production and control planning," in *Advances in Intelligent Systems and Computing*, pp. 984–994, Springer International Publishing, Cham, Switzerland, 2017.

[38] S.-J. Shin, D. B. Kim, G. Shao, A. Brodsky, and D. Lechevalier, "Developing a decision support system for improving sustainability performance of manufacturing processes," *Journal of Intelligent Manufacturing*, vol. 28, no. 6, pp. 1421–1440, 2017.

[39] J. Gąbka and G. Filcek, "Multiple criteria decision support system for making the best manufacturing technologies choice and assigning contractors," in *Advances in Intelligent Systems and Computing*, pp. 314–323, Springer International Publishing, Cham, Switzerland, 2018.

[40] M. M. Mabkhot, A. M. Al-Samhan, and L. Hidri, "An ontology-enabled case-based reasoning decision support system for manufacturing process selection," *Advances in Materials Science and Engineering*, vol. 2019, Article ID 2505183, 18 pages, 2019.

[41] Y. Jiang, B. Qiu, C. Xu, and C. Li, "The research of clinical decision support system based on three-layer knowledge base model," *Journal of Healthcare Engineering*, vol. 2017, Article ID 6535286, 8 pages, 2017.

[42] Y.-F. Chen, C.-S. Lin, K.-A. Wang et al., "Design of a clinical decision support system for fracture prediction using imbalanced dataset," *Journal of Healthcare Engineering*, vol. 2018, Article ID 9621640, 13 pages, 2018.

[43] M. Camacho-Collados and F. Liberatore, "A decision support system for predictive police patrolling," *Decision Support Systems*, vol. 75, pp. 25–37, 2015.

[44] M. Marzouk and B. Mohamed, "Integrated agent-based simulation and multi-criteria decision making approach for buildings evacuation evaluation," *Safety Science*, vol. 112, pp. 57–65, 2019.

[45] J. Lee, H. Cho, and Y. S. Kim, "Agile supply chain decision support system,""Agile supply chain decision support system," in *Reshaping Society through Analytics, Collaboration, and Decision Support: Role of Business Intelligence and Social Media*, L. S. Iyer and D. J. Power, Eds., Springer International Publishing, Cham, Switzerland, pp. 29–50, 2015.

[46] G. Dellino, T. Laudadio, R. Mari, N. Mastronardi, and C. Meloni, "A reliable decision support system for fresh food supply chain management," *International Journal of Production Research*, vol. 56, no. 4, pp. 1458–1485, 2018.

[47] J. Mar-Ortiz, M. D. Gracia, and N. Castillo-García, "Challenges in the design of decision support systems for port and maritime supply chains," in *Exploring Intelligent Decision Support Systems: Current State and New Trends*, pp. 49–71, Springer International Publishing, Cham, Switzerland, 2018.

[48] A. K. Sangaiah, A. Abraham, P. Siarry, and M. Sheng, "Intelligent decision support systems for sustainable computing," in *Intelligent Decision Support Systems for Sustainable Computing: Paradigms and Applications*, pp. 1–6, Springer International Publishing, Cham, Switzerland, 2017.

[49] A. Schwenk-Ferrero and A. Andrianov, "Nuclear waste management decision-making support with MCDA," *Science and Technology of Nuclear Installations*, vol. 2017, Article ID 9029406, 20 pages, 2017.

[50] S. Kapilan and K. Elangovan, "Potential landfill site selection for solid waste disposal using GIS and multi-criteria decision analysis (MCDA)," *Journal of Central South University*, vol. 25, no. 3, pp. 570–585, 2018.

[51] O. Rybnytska, F. Burstein, A. V. Rybin, and A. Zaslavsky, "Decision support for optimizing waste management," *Journal of Decision Systems*, vol. 27, no. sup1, pp. 68–78, 2018.

[52] R. Mukhamediev, R. Mustakayev, K. Yakunin, S. Kiseleva, and V. Gopejenko, "Multi-criteria decision support system for RES evaluation," in *Proceedings of the 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–6, Almaty, Kazakhstan, October 2018.

[53] R. I. Mukhamediev, R. Mustakayev, K. Yakunin, S. Kiseleva, and V. Gopejenko, "Multi-criteria spatial decision making supportsystem for renewable energy development in Kazakhstan," *IEEE Access*, vol. 7, pp. 122275–122288, 2019.

[54] S. Torabi Moghadam and P. Lombardi, "An interactive multi-criteria spatial decision support system for energy retrofitting of building stocks using CommuntiyVIZ to support urban energy planning," *Building and Environment*, vol. 163, Article ID 106233, 2019.

[55] M. Esmaelian, M. Tavana, F. J. Santos Arteaga, and S. Mohammadi, "A multicriteria spatial decision support system for solving emergency service station location problems," *International Journal of Geographical Information Science*, vol. 29, no. 7, pp. 1187–1213, 2015.

[56] M. S. E. Mohamed and A. A. Binsultan, "Developing an intelligent decision support system approach for crisis preparedness," in *Advances in Intelligent Systems and Computing*pp. 690–699, Cham, Switzerland, 2017.

[57] G. Zhang, J. Lu, and Y. Gao, "Decision making and decision support systems," in *Multi-Level Decision Making: Models, Methods and Applications*, pp. 3–24, Springer, Berlin, Germany, 2015.

[58] A. D. Haidar, "Techniques for intelligent decision support systems," in *Construction Program Management–Decision Making and Optimization Techniques*, pp. 159–183, Springer International Publishing, Cham, Switzerland, 2016.

[59] P. Hirsch, M. Grochowski, and K. Duzinkiewicz, "Decision support system for design of long distance heat transportation system," *Energy and Buildings*, vol. 173, pp. 378–388, 2018.

Hindawi
Scientific Programming
Volume 2020, Article ID 8875069, 10 pages
https://doi.org/10.1155/2020/8875069

Research Article

A New Scalable and Expandable Access Control Model for Distributed Database Systems in Data Security

Mehmet Guclu, Cigdem Bakir, and Veli Hakkoymaz

Department of Computer Engineering, Yildiz Technical University, Istanbul 34220, Turkey

Correspondence should be addressed to Mehmet Guclu; mehmetguclu007@gmail.com

Received 5 August 2020; Revised 21 August 2020; Accepted 29 August 2020; Published 10 September 2020

Academic Editor: Habib Ullah Khan

Access control models are an important tool developed for securing today's data systems. Institutions use the access control models specifically to define who their employees are, what they can do, which resources they can reach, and which processes they can perform and use them to manage the whole process. This is a very hard and costly process for institutions with distributed database systems. However, access control models cannot be implemented in a qualified way due to the fact that the conditions for defining users' demands to reach resources distributed on different servers, one of which is consequentially bound to the other, the verification and authorization of those user demands, and being able to monitor the actions of the users cannot be configured in an efficient way all the time. With our model suggested in this study, the aim is to automatically calculate the permissions and access levels of all users defined in the distributed database systems for the objects, and, in this way, we will reach a more efficient decision as to which objects the users can access while preventing their access to the information they do not need. Our proposed model in this study has been applied to real life data clusters from organizations providing health and education services and a public service. With the proposed model, all models have been run on servers sharing resources in a private network. The performance of the proposed model has been compared to that of traditional access models. It was confirmed that the proposed model presented an access control model providing more accurate access level results as well as being scalable to many distributed database systems.

1. Introduction

Today, there are new threats damaging the information systems and resources: armored viruses, ransomware, and cryptoLocker malware [1]. Despite the most enterprising steps taken to protect the systems from these harmful threats, the attackers can sometimes be successful. Every phenomenon causing a violation of any one of the principles of confidentiality, integrity, and accessibility—the three main elements of information security—is a violation of security [2]. While some violations deliberately make the systems inaccessible and interrupt services, some of them occur due to accidental software or hardware failures. Either by accident or malice, security violations seriously affect the activity and reliability of an institution.

Denial of service attacks, distributed denial of service attacks, inappropriate surfing behaviors on the Web, wiretapping, access to resources using a backdoor means of access, and accidental or deliberate data interchanges are leading factors causing security violations [3]. Deliberately or accidentally interchanged data affects the integrity principle of security in computing systems and in particular plays a significant role in the occurrence of deliberate or accidental data interchange phenomena [4]. There is a need for a good access model designed according to the scale of the organization and to the confidential access rights necessary for the users to cope with these kinds of problems. In this study, the aim was to automatically calculate the permission and access levels of all users defined in the distributed database systems based on the objects. In this way, a more efficient decision can be reached as to which objects users can access and to prevent their access to information they do not need, in real time.

Access control-based models are one of the most important principal measures used to prevent unauthorized access and minimize the impact of security violation [5]. Today, there are access control models specifically designed. However, it is seen that these models cannot completely

meet the needs of the rapidly increasing number of systems that are becoming more complex day by day, place a serious financial burden on the system, cannot completely ensure data flow control, and, to a great degree, cause a loss of flexibility in the application [6–9]. For this reason, it is also seen that it is not only sufficient for the access control models to be configured to protect the information systems from unauthorized accesses, malicious users, and erroneous use; it is also important that they be easily manageable and scalable in accordance with the organizational structure and that the access control functionality be designed consistently.

There are many real-world applications where static access control models such as judicial network information systems, defense systems, and hospital management systems are not effective. The main reasons for this are as follows: we can list the confidentiality, security restrictions, and level of access difference according to the organizational structure, the initially decided security policies cannot be dynamically changed in accordance with the changing corporate or commercial conditions and business requirements, and the access controls are not easily managed. In our study, referring to such problems that we experience in real system applications allows dynamically changing the permission and access level of the user on the object, based on the current status of the user and object within the organizational structure and/or their status/level of change/updated over time (authority level, some privileges, exceptions, degree of privacy, etc.) that can be adapted to different systems. An access control technique is presented.

The access control model developed in this study addressed the problems frequently faced in applications and provided a model that is more functional, more easily manageable, and more scalable and can deliver more consistent results. The main contribution and aim of this study was to automatically calculate the permission and access levels of all users having an active role in distributed database systems, avoid overauthorization, and deliver more efficient decisions on which objects users can access and prevent their access to information they do not need.

The remaining parts of this study have been organized as follows: there is an introduction in the first section, section two looks at related studies, and materials and methods are covered in section three, while the experimental study is detailed in section four, and section five covers the conclusions.

2. Related Works

Cloud computing is one of the advanced areas in the Information Technology (IT) sector today. Because there are many computer pirates and malicious users on the Internet, it is very important to ensure the confidentiality of the data in the medium of the cloud. For this purpose, it is seen that the number of cloud computing-based advanced access control models has been recently and rapidly increasing [10,11]. Behera and Khilar [12] developed a new access control method. The suggested method authorizes the user according to the user's value before entering the cloud environment. For this, the value of both the user and the cloud resources is calculated. If the value of both the users and the cloud resources is higher than the threshold values, it is deemed as reliable. In another study explaining the validity of current access control models for cloud computing and their services, an access control model increasing the security and preventing unauthorized users from accessing the cloud resources was presented [13].

In the current Distribution Version Control System (DVCS) where the access control principles are distributed across many heterogeneous systems, it is hard to respect the principle of least privilege. In some studies, the main hardships experienced in advance towards a more thorough and manageable access control model in distributed systems have been mentioned [7,9,14]. In one study, an access control architecture that can be adapted by the Industrial Control System (ICS) community has been presented for controlling any access via policies in accordance with the least privilege principle [14]. The aim was to protect central policy management and every bound field device in the suggested architecture. Bertolissi and Fernandez [15] defined a model for their access control design by considering the confidentiality requirements of the distributed media. In this study, a framework was suggested for the implementation of access control policies by taking into consideration the local policies determined by every member on a distributed system consisting of various sites so that each one of them will protect their own resources.

Due to their widespread use, IoT devices are highly likely to contain various security vulnerabilities and threats. Therefore, dealing with IoT-related attacks, vulnerabilities, security, and privacy challenges requires a strong security mechanism. Liao et al. demonstrated that a strong security mechanism can be achieved better with mobile computing, which provides both hardware- and software-based security solutions [16].

There are also some of methodologies put forward to evaluate the security of software components. Fuzzy Logic (FL) approach is modeled to evaluate the safety of components in [17]. The research has shown that the proposed methodology based on ISO/IEC 18028-2 security attributes is useful in situations of uncertainty, thus helping to select the most secure software component. In another study, a method that evaluated the security of software components to enable the software development process is presented [18]. The security of the software component was evaluated using the ANP model based on specific security attributes provided of ISO/IEC 27002. Another study, which presented a system-based differential mathematical model for software birthmark-based comparisons and evaluation of security in end-to-end communication systems, evaluated the smoothing of software piracy and theft detection process and the security of end-to-end communication systems [19]. In another study, it has been shown how data security is ensured with machine learning algorithms [20].

According to the research findings of Rehman et al., it was stated that some people provide fake information to the websites of social media and nonprofit organizations because they think that users collected too much information for the sake of security and individuals feel insecure about

the personal information provided to them [21]. In the study, it was emphasized that it is important to read the minds of the users and get feedback from them in order to minimize this gap between users and service providers.

Data access can be statically inspected using role-based or policy-based access models. However, it was seen that there was still a large gap in the issue of ensuring data access security in the great data age where many studies are conducted on storing today's huge amounts of unconfigured data [22, 23]. There are many real-world applications where static access control systems are not efficient, such as airport search/observation, defense, and hospital management systems [24]. There is a need for a system that learns and adapts according to the user reality. The current role-based access control system easily attracts uninvited guests. Again, in policy-based access control, a deficiency in adaptation occurs because a policy decided at the beginning cannot be dynamically changed. Risk-adaptive access control—suggested by Srivastav and Shekokar —presents a framework that understands the user reality, calculating the risk and acting on this basis afterwards [24]. This framework considers many real-world qualifiers, such as access period, access place, previous history of the request (how many times the same request has been repeated), and the precision of the requested data. Though that study shows similarity to this study in terms of purpose and scope, the previous actions and access requests of the user are not considered in this current study. The model suggested in this study assigns a value to every user in different dimensions appropriate to the organizational structure and relates the access permission to the object with the dimension values and access levels of the user. In other words, it calculates the access permission and level of a user related to an object according to the abilities or values owned by that user.

3. Traditional Access Control Models

In the Mandatory access control model, access of users to the resources is controlled in accordance with certain rules predetermined by a central authority. This type of access control is widely observed in military confidentiality classification. In the Discretionary access control model, users can give access authorizations to other users within the limits assigned to them or they can determine the limitations. This type of access control is commonly seen in folder and file authorizations of operating systems. RBAC provides access rights based on the roles and privileges of the users. RBAC requires users to be assigned to different roles to get the associated permissions. However, the problems of role explosion limit its use to enterprise systems only. Here, a user may have multiple roles or capacities within a given organization. Thus, when the subject is seeking access to an object, the user must first indicate the role within which the request is being made [25].

4. The Proposed Model

The flow diagram of our proposed model is shown in Figure 1.

Data are expressed as objects in our proposed model. Users are classified according to the security dimensions. A security dimension explains the characteristics of a user, and every dimension contains some values that can be assigned to the users. For instance, some example values that can be owned by users in different security dimensions are called Unit, Security Classification, Business Title, and Operation, as shown in Table 1. Unit dimensions within the security dimensions that can be assigned to the user include Unit A, Unit B, Unit C, Unit D, and Unit E values. The Security Classification dimension consists of Top Secret, Secret, and Unclassified values; the Business Title dimension can be Head Doctor, Doctor, IT Personnel, Nurse, and Purchasing Personnel values; and the Operation dimension can consist of Process A, Process B, Process C, Process D, Process E, and Process F values.

A security dimension may have the following characteristics.

4.1. Ordered Dimension. If a dimension is ordered, the dimension values are ordered and the order is compositional; it also covers the values below the value assigned to a user. For example, the values of Top Secret, Secret, and Unclassified values occur in the dimension called Security Classification. A user assigned with the Secret value is also automatically conferred with the Confidential and Unclassified values.

4.2. Unordered Dimension. If a dimension is unordered, the dimension values are not ordered and more than one value can be assigned to a user. For instance, the values of Process A, Process B, Process C, Process D, Process E, and Process F are within the dimension called Operation. A user can take part in both Process C and Process E operations.

A user whose Business Title is Head Doctor assigns the dimension values to other users. At least one value should be assigned to every user from every dimension. However, other values can be assigned in other dimensions. For example, each one of the five users in 2 takes part in different units. While User 1, who is in Unit E and can perform processes C and D, is a nurse in the Confidential Security class, User 4, who is in Unit A and can perform processes A, B, and F, is a doctor from the Confidential Security class.

Each user uses dimension values for the access model. Each user can or cannot access an object according to these dimension values. Namely, a user may have read and write access to an object according to the values taken from all the security dimensions.

Access permission lists should contain a value from each access dimension. In addition, it can also contain some values from the same dimension (such as Process A, Process B, and Process F). The access level of users to objects is determined according to the dimension values taken from each security dimension. For instance, read and write access level can be used for the Purchasing Personnel taking place in the Business Title security dimension and the read only access level can be allocated to the Nurse.

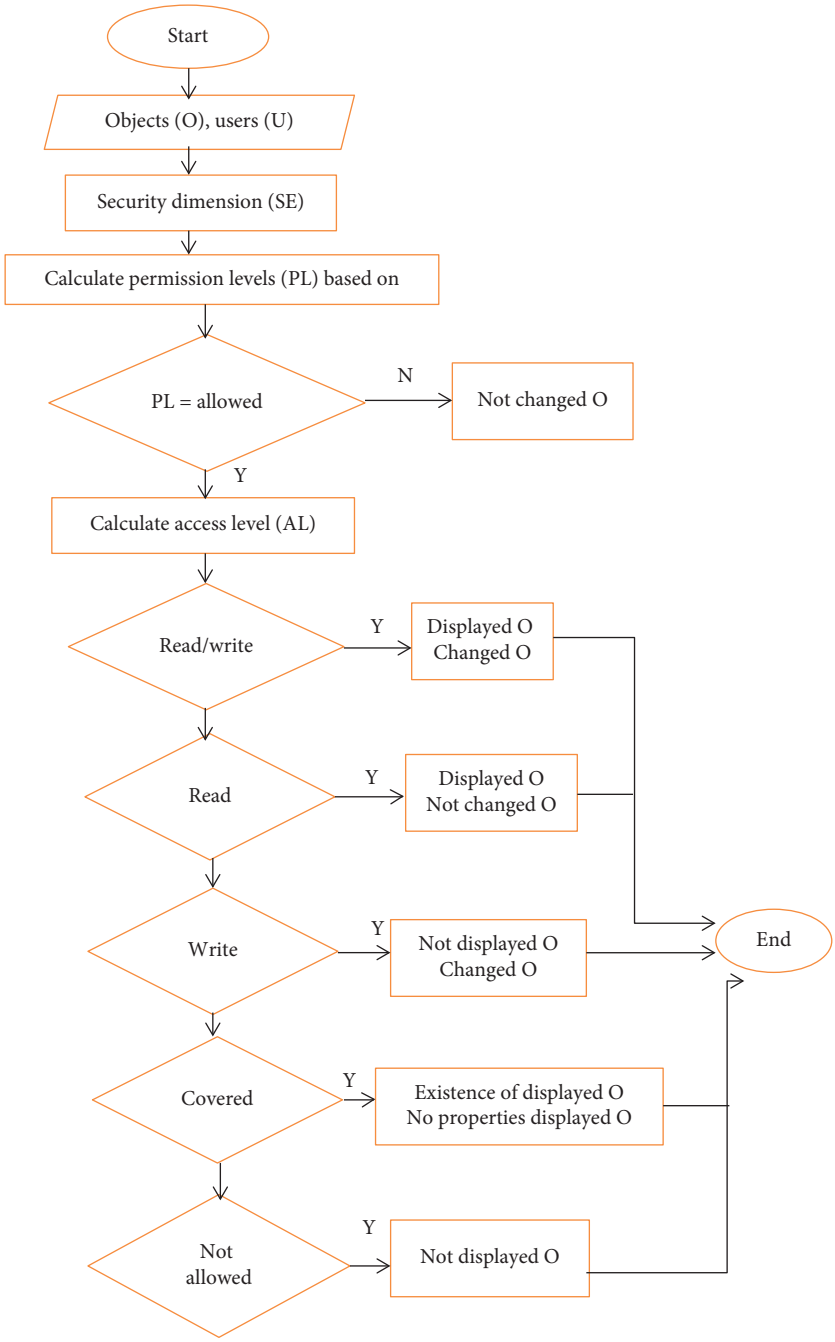


Figure 1: The flow diagram of the proposed model.

Table 1: Security dimensions.

	Security dimension		
Unit	Security Classification	Business Title	Process
Unit A	Top Secret	Head Doctor	Process A
Unit B	Secret	Doctor	Process B
Unit C	Confidential	IT Personnel	Process C
Unit D	Unclassified	Nurse	Process D
Unit E		Purchasing Personnel	Process E
			Process F

Table 2: The dimension values of five different users.

Security dimension	User 1	User 2	User 3	User 4	User 5
Unit	Unit E	Unit B	Unit D	Unit A	Unit C
Security Classification	Confidential	Secret	Confidential	Secret	Top Secret
Business Title	Nurse	IT Personnel	Nurse	Doctor	IT Personnel
Operation	Processes C and D	Processes B, E, and F	Processes C, D, and E	Processes A, B, and F	Processes A, B, C, D, E, and F

4.3. Permission Levels. Permission levels are the different ability levels that allow for changes to the security settings of an object. The permission level for an object is collectively determined from within the permissions assigned to that object:

If the permission level is *Allowed*, the security settings of the object can be changed.

If the permission level is *Not Allowed*, the security settings of the object cannot be changed.

So, if the permission level of an object is *Allowed*, the object can be queried and its security settings can be changed, but if the access level is *Not Allowed*, the object cannot be displayed.

4.4. Access Levels. Access levels are the different ability levels to see or change objects. The access level to an object is collectively determined within the access permissions of the object:

(i) If the access level of a user is *Read/Write*, the object can be displayed and changed

(ii) If the access level of a user is *Read Only*, the object can be displayed but cannot be changed

(iii) If the access level of a user is *Write Only*, the object cannot be displayed but can be changed

(iv) If the access level of a user is *Covered*, the existence of the object can be displayed, but its properties cannot be displayed

(v) If the access level of a user is *Not Allowed*, the object cannot be displayed, and the object does not show up in the query results

4.5. Access and Permission Levels. Access permission to an object is related to the dimension values (the value taken from each dimension by a user, such as the values defined for the 5 different users in Table 2) and access level (Read/Write, Read Only, Covered, or Not Allowed). In other words, the access permission and level are revealed for an object by a user according to the abilities or dimension values owned by that user. If the access level of the user is covered or above (Covered, Read Only, Write Only, and Read/Write), the access of that user to the object is allowed.

4.6. Access Level or Permission Level in One Dimension. Access permission to an object can be related to the values in a dimension for many access levels (e.g.,, while a user could only get the Read Only access level for Process B in the Operation dimension, they could get the access levels of Read and Write for Process C). Permission levels could also be valid for a similar situation. In these situations, the least restrictive access and permission levels are used.

If expressed in an example, the following dimension values could be assigned to a user (Table 3).

The user could display the object with access permissions given in Table 4.

The object could have the permissions given in Table 5.

Object access permissions specify that Process D user membership in the Operation dimension has been set up with *Read/Write* access. Because there is no access permission defined for Process A in the Operation dimension, the user membership for Process A in the Operation dimension is set up with the *Not Allowed* access level. The least restrictive of these access levels is the Read and Write level; for this reason, this access level is used for the Operation dimension.

Object access permissions specify that the Confidential Security Classification of the user has been set up with *Read Only* access—being the least restrictive in this case. Because the object does not have any permission to relate the Nurse Title to a permission level, the resulting permission is the *Not Allowed* level.

4.7. General Access or Permission Level of the Object. The calculation of the least restrictive access or permission level in each dimension may have different results for each dimension. In this situation, the least restrictive access or permission level is used each time.

The general calculation is shown in Figure 2. According to Figure 1, the *Read Only* level is used for the Unit dimension, the *Read/Write* access level is used for the Operation dimension, and the *Read Only* access level is used for the Security Classification dimension. Because the most restrictive one of these levels is *Read Only*, the general access level taken by the user for the object becomes *Read Only*.

5. Experimental Study

Three different real datasets taken from the institutions delivering health, education, and public services have been used in the study, and the success of the suggested access control model and other methods has been assessed according to the results attained from each dataset.

5.1. Datasets. The three datasets used in the study taken from different sectors underwent a preliminary process so

Table 3: Dimension values defined for User 1.

Security dimension	User 1
Unit	Unit E
Security Classification	Confidential
Business Title	Nurse
Operation	Processes C and D

Table 4: Access levels for an object.

Security dimension	Dimension value	Access level
Unit	Unit E	Read Only
	Secret	Write Only
Security Classification	Confidential	Read Only
	Confidential	Covered
Operation	Process A	Read
Operation	Process D	Read/Write

Table 5: Access permissions for the object.

Security dimension	Dimension value	Permission level
Business Title	Doctor	Allowed

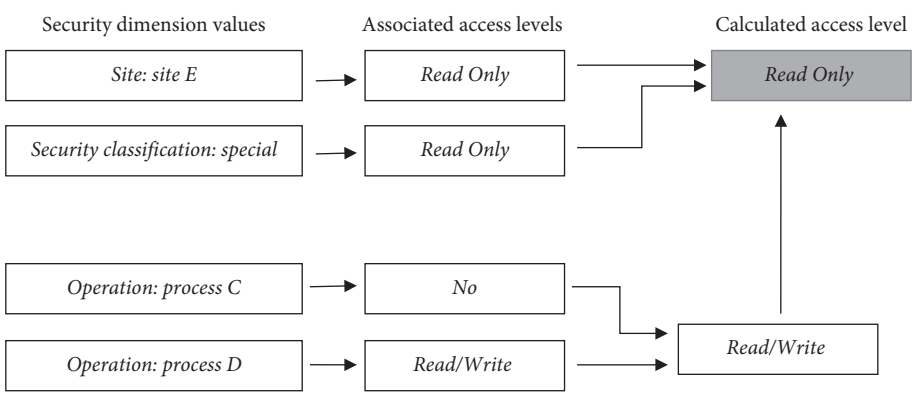


Figure 2: The access level.

that each user and object mentioned in the dataset was classified according to the security dimensions. Real classification scales for these institutions have been taken as the basis for the classification process. The dataset taken from the health sector consisted of 107 users, 36,251 objects, and 8 security dimensions; the dataset taken from the education sector consisted of 292 users, 72,988 objects, and 6 security dimensions, and the dataset taken from the public sector consisted of 1355 users, 752,220 objects, and 11 security dimensions. Datasets have been labeled as the health dataset, education dataset, and public dataset.

5.2. Experimental Analysis. Together with our suggested model, other access control models have been used on a platform operating a real distributed system, and all models have been separately applied to the three datasets. The permission and access level results attained for all models applied to each dataset were compared to the permission and access level results in use by the application belonging to the sector from which the dataset was taken, and the performance values of the methods were compared.

Table 6 shows the calculated access permission and level results for an object with ID number "1" in the health dataset. Table 7 shows the valid access permission and level results for an object with ID number "1" in the health dataset. When both tables were compared, the calculated access permissions appear to be 100% similar to the valid access permissions. While the valid access level for User 4 is Read/Write, the calculated access level is found as Only Write. In this case, while the accuracy rate of the calculated access permission for the object with ID number "1" is 100%, the accuracy rate of the calculated access level will be 75%.

The percentages of correct permission and access level detection for each method were taken as the basis for the performance assessment of the methods applied to the datasets.

Table 6: Calculated access permission and level results.

Dataset: health	Object ID: 1	Users	Permission level	Access level
		User 1	Allowed	Only Read
		User 2	Allowed	Read/ Write
		User 3	Not Allowed	Not Allowed
		User 4	Allowed	Only Write

Table 7: Access permission and level results in industry.

Dataset: health	Object ID: 1	Users	Permission level	Access level
		User 1	Allowed	Only Read
		User 2	Allowed	Read/ Write
		User 3	Not Allowed	Not Allowed
		User 4	Allowed	**Read/ Write**

5.3. Performance Results of the Proposed Model. Test results for the suggested model applied to the health, education, and public datasets are shown in Table 8. The testing showed that the suggested model achieved a correct permission level of 98.20% for the health dataset and access levels were correctly detected in 94.70% of cases where the object permission level had been correctly detected. For the education dataset, permission levels were correctly detected in 95.03% of cases, and access levels were correctly detected in 90.95% of cases where the object permission level had been correctly detected. For the public dataset permission levels were correctly detected in 97.91% of cases; access levels were correctly detected in 95.12% where the object permission level had been correctly detected.

When the results produced by the suggested model were assessed, it could be said that the suggested model achieved correct access permission and access level at 90% and above in the datasets belonging to the three different sectors. In addition, it was observed that as the security dimension (number of properties) increased, the success ratio for detection at the access level also increased (Figure 3). In addition, when the security dimension number was higher than other datasets, the success ratio was higher, as in the public dataset where the number of users and objects is higher compared to the others.

5.4. Performance Results for the RBAC. Test results for the role-based access control model on the health, education, and public datasets are shown in Table 9. The testing showed that this model achieved a correct permission level of 92.17% for the health dataset and access levels were correctly detected in 90.63% of cases where the object permission level had been correctly detected. For the education dataset, permission levels were correctly detected in 89.09% of cases, and access levels were correctly detected in 85.98% of cases where the object permission level had been correctly detected. For the public dataset, permission levels were correctly detected in 89.42% of cases; access levels were correctly detected in 82.77% where the object permission level had been correctly detected.

When the results rendered by the RBAC model were assessed, it was shown that the model detected correct access permission and access levels at 90% and above in the health dataset consisting of less users and objects, but a decrease was observed in the accuracy percentage of the access level, especially as the number of users and objects increased.

5.5. Performance Results for the MAC/DAC. Test results for the MAC/DAC on health, education, and public datasets are shown in 10. In the test results for this model, the performance percentage of the model for MAC and DAC with the higher access permission and access level accuracy has been taken as the basis for the assessment. The testing showed that this model achieved a correct permission level of 87.60% for the health dataset and access levels were correctly detected in 86.02% of cases where the object permission level had been correctly detected. For the education dataset, permission

Table 8: Permission and access level performance of the proposed model.

	Access permission	Access level
Health dataset	98.20%	94.70%
Education dataset	95.03%	90.95%
Public dataset	97.91%	95.12%

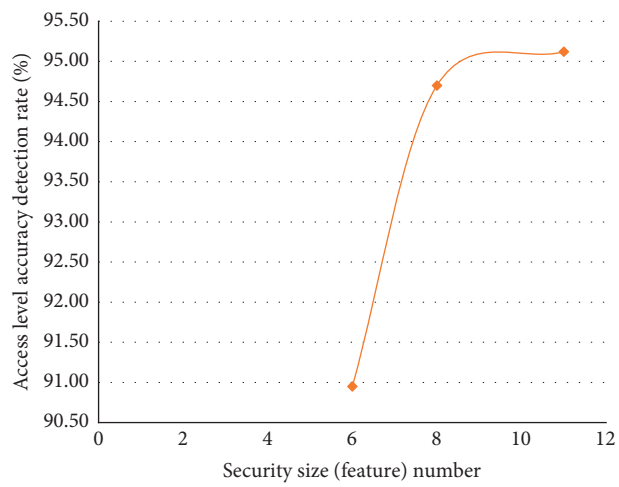


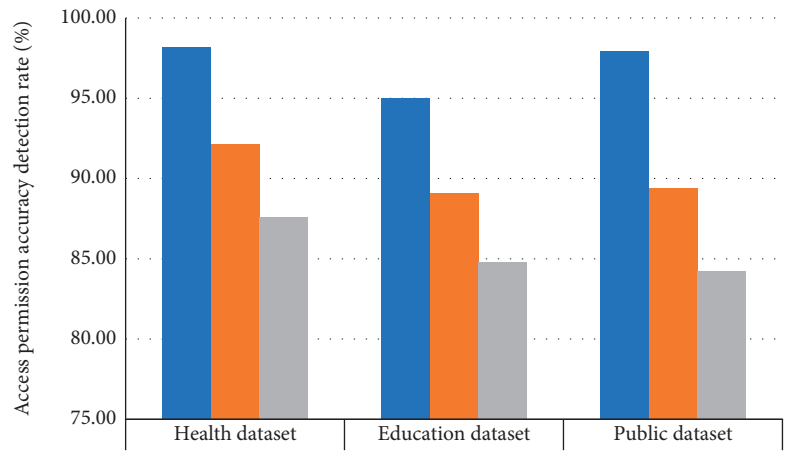
Figure 3: Access level success rate according to the number of security dimensions.

Table 9: Permission and access level performance for RBAC.

	Access permission	Access level
Health dataset	92.17%	90.63%
Education dataset	89.09%	85.98%
Public dataset	89.42%	82.77%

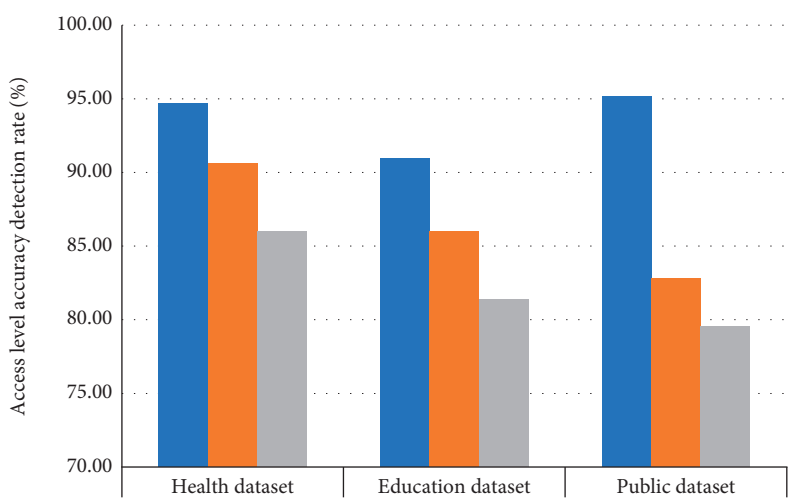
Table 10: Permission and access level performance of MAC/DAC.

	Access permission (%)	Access level (%)
Health dataset	87.60	86.02
Education dataset	84.79	81.39
Public dataset	84.21	79.54



	Health dataset	Education dataset	Public dataset
The suggested model	98.20%	95.03%	97.91%
RBAC	92.17%	89.09%	89.42%
MAC/DAC	87.60%	84.79%	84.21%

Figure 4: Correct detection rate for access permission in the three models.



	Health dataset	Education dataset	Public dataset
The suggested model	94.70%	90.95%	95.12%
RBAC	90.63%	85.98%	82.77%
MAC/DAC	86.02%	81.39%	79.54%

Figure 5: Correct detection rate for access level in the three models.

levels were correctly detected in 84.79% of cases, and access levels were correctly detected in 81.39% of cases where the object permission level had been correctly detected. For the public dataset, permission levels were correctly detected in 84.21% of cases; access levels were correctly detected in 79.54% where the object permission level had been correctly detected.

When the results produced by the MAC/DAC models were assessed, as in the RBAC model, it was shown that this model also delivered higher percentages of correct access permissions and access level in the health dataset, but a decrease was observed in the accuracy of access permissions and access levels, especially as the number of users and objects increased.

5.6. Performance Evaluation. Given that the suggested model delivered more successful results for the access permission and access level detection rate when compared to other techniques—as seen in Figures 4 and 5—it can be said that it achieved correct detection rate of 90% and above in all three datasets. The other techniques were less successful in datasets with high numbers of users and objects. So, this result showed a more expandable technique for different sector applications compared to other techniques and a more scalable technique for the same sector applications.

6. Conclusions

The proposed new access control model investigated in this study was applied to a real distributed system, and, in this way, calculations were made as to which users could access the data stored in different physical media with access permission and level.

With the access control that we proposed in the study, access permissions of users to an object in a distributed environment are associated with the dimension values and object access levels owned by the user. Compared to other access control methods based on performance evaluation, the proposed model dynamically calculates the user's access and level on an object based on the specific permissions and powers that a user has, the size values assigned to him, and the access permissions and levels of the object.

When the experimental results delivered by the suggested model were assessed, the suggested model was applied to the datasets belonging to three different sectors taken from real life and the performance of the suggested model was compared to the Traditional Access Control frequently encountered in real system applications. It has been shown that the suggested model delivered correct access permission and access level 90% and above cases in all three datasets and also delivered successful results in a way that was scalable for all three sectors when compared to other models. As a result, the particular problems frequently faced in distributed system applications were dealt with and the suggested model is expandable and scalable for distributed systems while delivering more consistent authorization results.

As a continuation of this study, the suggested model will be developed and a new framework, taking the access period, access place, and user behaviors as the basis for the design, will be presented.

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 they have no conflicts of interest.

References

[1] M. E. Whitman and H. J. Mattord, *Principles of Information Security, Course Technology*, Cengage Learning, Boston, MA, USA, 2012.

[2] M. G. Solomon and D. Kim, *Fundamentals of Information Systems Security*, Jones & Bartlett Learning, Burlington, MA, USA, 3rd edition, 2016.

[3] M. Guclu, C. Bakir, V. Hakkoymaz, and B. Diri, "Comparisons on intrusion detection and prevention systems in distributed databases," *Balkan Journal of Electrical and Computer Engineering*, vol. 7, pp. 446–455, 2019.

[4] J. Andress, *The Basics of Information Security Understanding the Fundamentals of InfoSec in Theory and Practice*, pp. 17–49, Elsevier, Amsterdam, Netherlands, 2011.

[5] A. S. M. Kayes, W. Rahayu, P. Watters, M. Alazab, T. Dillon, and E. Chang, *Achieving security scalability and flexibility using Fog-Based Context-Aware Access Control*, vol. 107, pp. 307–323, Elsevier, Amsterdam, Netherlands, 2020.

[6] M. Kotari and N. N. Chiplunkar, *Investigation of Security Issues in Distributed System Monitoring, Information Sciences*, pp. 609–634, Springer, Berlin, Germany, 2020.

[7] M. Kotari and D. N. N. Chiplunkar, "Framework of security mechanisms for monitoring adaptive distributed systems," *IOSR Journal of Computer Engineering*, vol. 18, no. 04, pp. 25–36, 2016.

[8] M. S. Shin, H. S. Jeon, Y. W. Ju, B. J. Lee, and S. P. Jeong, "Constructing RBAC based security model in u-healthcare service platform," *The Scientific World Journal*, vol. 2015, Article ID 937914, 13 pages, 2015.

[9] J. Reid, I. Cheong, M. Henrickson, and J. Smith, "A novel use of RBAC to protect privacy in distributed health care information systems," in *proceedings of the 8th Australasian Conference on Information Security and Privacy (ACISP 2003)*, Wollongong, Australia, July 2003.

[10] J. Li, Z. Liao, C. Zhang, and Y. Shi, "A 4D-role based access control model for multitenancy cloud platform," *Mathematical Problems in Engineering*, vol. 2016, Article ID 2935638, 16 pages, 2016.

[11] R. Lu, Y. Rahulamathavan, H. Zhu, C. Xu, and M. Wang, "Security and privacy challenges in vehicular cloud computing," *Mobile Information Systems*, vol. 2016, Article ID 6812816, 2 pages, 2016.

[12] P. K. Behera and P. M. Khilar, "A novel trust based access control model for cloud environment," in *Proceedings of the International Conference on Signal, Networks, Computing, and Systems*, Springer, Rourkela, India, pp. 285–295, 2016.

[13] S. Pandey, A. Dwivedi, J. Pant, and M. Lohani, "Security enforcement using TRBAC in cloud computing," in *Proceedings of the International Conference on Computing, Communication and Automation (ICCCA)*, pp. 1232–1238, IEEE, Noida, India, 2016.

[14] J. H. Huh, R. B. Bobba, T. Markham et al., "Next-generation access control for distributed control systems"" *IEEE Internet Computing*, vol. 20, no. 5, pp. 28–37, 2016.

[15] C. Bertolissi and M. Fernández, "A metamodel of access control for distributed environments: applications and properties," *Information and Computation*, vol. 238, pp. 187–207, 2014.

[16] B. Liao, Y. Ali, S. Nazir, L. He, and H. U. Khan, "Security analysis of IoT devices by using mobile computing: a systematic literature review," *IEEE Access*, vol. 8, pp. 120331–120350, 2020.

[17] S. Nazir, S. Shahzad, S. Mahfooz, and M. . Nazir, "Fuzzy logic based decision support system for component security evaluation," *The International Arab Journal of Information Technology*, vol. 15, pp. 1–9, 2015.

[18] S. Nazir, S. Shahzad, M. Nazir, and H. U. Rehman, "Evaluating security of software components using analytic network

process," in *Proceedings of the 11th International Conference on Frontiers of Information Technology (FIT)*, pp. 183–188, IEEE, Islamabad, Pakistan, 2013.

[19] M. Li, S. Nazir, H. U. Khan, S. Shahzad, and R. Amin, "Modelling features-based birthmarks for security of end-to-end communication system," *Security and Communication Networks*, vol. 2020, Article ID 8852124, 9 pages, 2020.

[20] J. Zhang, S. Nazir, A. Huang, and A. Alharbi, "Multicriteria decision and machine learning algorithms for component security evaluation: library-based overview," *Future of Information and Communication Conference*, vol. 2, pp. 964–974, 2019.

[21] H. U. Rehman, A. U. Rehman, S. Nazir, I. U. Rehman, and N. Uddin, "Privacy and security—limits of personal information to minimize loss of privacy," *Future of Information and Communication Conference*, vol. 2, pp. 964–974, 2019.

[22] K. Szczypiorski, L. Wang, X. Luo, and D. Ye, "Big data analytics for information security," *Security and Communication Networks*, vol. 2018, Article ID 7657891, 2 pages, 2018.

[23] P. Angin, B. Bhargava, and R. Ranchal, "Big data analytics for cyber security," *Security and Communication Networks*, vol. 2019, Article ID 4109836, 2 pages, 2019.

[24] K. Srivastava and N. Shekokar, *Machine Learning Based Risk-Adaptive Access Control System to Identify Genuineness of the Requester*, pp. 129–143, Springer, Berlin, Germany, 2020.

[25] D. Kim and M. G. ve Solomon, *Fundamentals of Information Systems Security*, Jones and Bartlett Publishers Inc., Burlington, MA, USA, 2016.