Wireless Communications and Mobile Computing

Cyber-Physical Mobile Computing, Communications, and Sensing for Industrial Internet of Things

Lead Guest Editor: Mohammad Khosravi
Guest Editors: Alireza Jolfaei and Pooya Tavallali

Cyber-Physical Mobile Computing, Communications, and Sensing for Industrial Internet of Things

Wireless Communications and Mobile Computing

Cyber-Physical Mobile Computing, Communications, and Sensing for Industrial Internet of Things

Lead Guest Editor: Mohammad Khosravi
Guest Editors: Alireza Jolfaei and Pooya Tavallali

Chief Editor

Associate Editors

Academic Editors

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 7154587, 17 pages
https://doi.org/10.1155/2021/7154587

Research Article

Intrusion Detection in Industrial Internet of Things Network-Based on Deep Learning Model with Rule-Based Feature Selection

Joseph Bamidele Awotunde,[1] **Chinmay Chakraborty**,[2] **and Abidemi Emmanuel Adeniyi**[3]

[1]*Department of Computer Science, University of Ilorin, Ilorin, Nigeria*
[2]*Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Jharkhand, India*
[3]*Department of Computer Science, Landmark University, Omu-Aran, Nigeria*

Correspondence should be addressed to Joseph Bamidele Awotunde; awotunde.jb@unilorin.edu.ng

Received 19 May 2021; Revised 21 July 2021; Accepted 4 August 2021; Published 3 September 2021

Academic Editor: Alireza Jolfaei

The Industrial Internet of Things (IIoT) is a recent research area that links digital equipment and services to physical systems. The IIoT has been used to generate large quantities of data from multiple sensors, and the device has encountered several issues. The IIoT has faced various forms of cyberattacks that jeopardize its capacity to supply organizations with seamless operations. Such risks result in financial and reputational damages for businesses, as well as the theft of sensitive information. Hence, several Network Intrusion Detection Systems (NIDSs) have been developed to fight and protect IIoT systems, but the collections of information that can be used in the development of an intelligent NIDS are a difficult task; thus, there are serious challenges in detecting existing and new attacks. Therefore, the study provides a deep learning-based intrusion detection paradigm for IIoT with hybrid rule-based feature selection to train and verify information captured from TCP/IP packets. The training process was implemented using a hybrid rule-based feature selection and deep feedforward neural network model. The proposed scheme was tested utilizing two well-known network datasets, NSL-KDD and UNSW-NB15. The suggested method beats other relevant methods in terms of accuracy, detection rate, and FPR by 99.0%, 99.0%, and 1.0%, respectively, for the NSL-KDD dataset, and 98.9%, 99.9%, and 1.1%, respectively, for the UNSW-NB15 dataset, according to the results of the performance comparison. Finally, simulation experiments using various evaluation metrics revealed that the suggested method is appropriate for IIOT intrusion network attack classification.

1. Introduction

A modern industrial revolution brings deep change and human growth, resulting in "Automation of Everything." It uses computer networks to link both digital devices, data mining, and real-world application management [1]. This revolution's opportunity helps everybody to access trillions of data and information that brings new opportunities. Significant increases in efficiency in the physical and digital industries may be felt by humans, resulting in a better quality of life and a more prosperous society. The creation of vast quantities of data from various sensors is popular in the Industrial Internet of Things (IIoT) world. These applications can be felt in various industries like healthcare, retail, automotive, and transport. In many industries, the IIoT can greatly increase efficiency, productivity, and operational efficiency. The IIoT will first develop existing processes and facilities, but the ultimate aim is to create completely new and vastly enhanced goods and services. Many companies recognize how and where IIoT innovations and solutions can lead to organizational changes, new and improved goods and services, and entirely new business models. On the IIoT, machine learning and deep learning algorithms can increase reliability, production, and customer satisfaction by combining technological innovations, sensors, programs, and applications.

Anything necessitates a wide range of technology that must be carefully integrated and orchestrated. These advancements in technology allow intelligent machines,

machinery, appliances, and integrated automation systems [2, 3] to automate routine operations and solve complex problems without human interference. Improvements in the smart workplace, smart data exploration, cognitive automation, and other aspects of business smartness should all be included. A digital twin is a virtual representation of physical assets, systems, and so on. It is commonly known as the Internet of Things (IoT), which is constantly developing all of these appliances and supplying us with an incredibly growing dataset that can be evaluated for efficiency, architecture, maintenance, and a host of other issues. A key feature of any digital twin is that it is constantly updated and "learns" any changes that arise in real-time. The IoT concept and its solutions have made a lot of changes in the physical world.

Since the cloud has altered how individuals and organizations communicate and perform business online, cyberspace plays an important role in today's societies and economies [4, 5]. As a result, the IIoT encompasses a variety of devices, software, and facilities that bridge the gap between the virtual and physical worlds [6]. Due to the connectivity of information technology (IT) and organizational technology (OT), industrial systems that depend on locked and exclusive communication systems are vulnerable to a wide range of interference activities [7, 8].

Machine-to-machine (M2M) and machine-to-person (M2P) connections to the network are used in IIoTs using the TCP/IP interface using various IIoT protocols [9, 10]. The number IIoTs have the number of flaws and bugs that can be abused using a range of advanced attack methods which has increased significantly. The attackers attempt to take advantage of these processes to steal sensitive information, commit financial funds, and corrupt device resources [11]. If the cybersecurity domain does not discover interesting mitigation strategies for stopping cyberthreats to the IIoT, it is estimated that they will cost up to $90 trillion by 2030 [12].

Protecting vital services and infrastructure is becoming a more critical problem in every organization as the volume of IIoT devices and implementations continues to grow [13]. Among the most frequent risks in IIoT networks is malware that abuses zero-day vulnerabilities. The perpetrators infect vulnerable computers to track and change their activities, using a variety of techniques like Progressive Determined Risk (PDR), Denial-of-Service (DoS), and Decentralized DoS. (DDoS). For instance, in 2010, the Stuxnet worm attacked Iran's nuclear program, in 2013, Iranian hackers hacked into the ICS of a dam in New York, and in 2015, the black-energy passive attack was explicitly equivalent to approximately 80.000 power outages in Ukraine [14, 15]. These nefarious practices showed that conventional cyberthreat methods, like security protocols, cryptography, access controls, and biometrics Interruption Discovery Systems (IDSs), are no longer sufficient for delivering successful vital infrastructure protection.

As a network security tonic, the network intrusion detection system (NIDS) is important in detecting and addressing all Internet attacks. The IIoT has become an essential portion of present machinery for data and knowledge transfer, necessitating the need for global network security [16]. To safeguard workstation schemes from multiple grid invasions, network intrusion detection systems (NIDS) are often used to recognize system traffic. In [17], intrusion is a framework that attempts to break information system's security services. Researchers have been inspired to create new IDSs in response to the threats posed by these invasive frameworks. Several intrusion detection systems (IDS) have previously been developed and upgraded, but they are still susceptible to a range of assaults. An increasing interest in anomaly detection research is due to IDS' ability to track and forecast malicious behavior unknown assaults. However, current machine learning-based irregularity discovery methods still have a high false alarm rate [18].

Recently, findings indicate that feature extraction is now at the core of a more accurate IDS [19, 20]. In most detection methods, the feature selection technique is used to pick the fitness values which input attributes for classification models, with the goal of aggregate discovery performance and reducing error rate in NIDS [8]. In particular, classifier feature vectors are massive, and not all of them apply to the groups to be categorized, requiring the use of a feature selection strategy. Conversely, feature selection approaches can be divided into three categories: filter approach, wrapper approach, and embedded approach [21]. The most popular feature selection strategy focused on selecting the best-fitted functionality which relies on dataset measurements lacking seeing classifier's performance. The wrapper method, on the other hand, is superior since the classifier feedback is used to evaluate the quality of the feature subclass, leading to higher prediction performance. The integrated process is analogous to wrapper approaches in that an intrinsic process modeling function in the classifier could be used to improve the learning algorithm's search efficiency.

Until now, several various categories for IDSs have been planned. Depending on the classification algorithm utilized, intrusion detection systems (IDSs) may be categorized as rule-based, misappropriation discovery, or diverse schemes. IDSs may either be classified as real-time if they use persistent system tracking or as sporadic or inactive if the tracking occurs only occasionally taking place at fixed times or even offline using data collected and processed over some time. Furthermore, new classifications have recently been introduced while discussing Industrial Control Systems (ICSs) with unique criteria and characteristics. The authors of [22] suggested a new classification system for IDSs called ICS, which are classified into three types: protocol review, traffic processing, and control process modeling.

Countermeasures are taken based on the information gathered from the detection systems about the identified attacks. The more accurately the type of attack is classified, the more effective the chosen countermeasures will be, and the less they will interfere with the device or network's proper operation. Furthermore, in some situations, countermeasures may have more severe effects than the attack itself if we do not detect the same form of attack. As a result, we aim to develop an intrusion prevention method that has proven expertise in each type of attack Moreover, for both

routine and irregular assaults, our system must have a low false alarm rate and a high detection accuracy, allowing limited processing to correctly classify. The latter function is important because intrusion detection systems are used in industrial control systems that operate critical infrastructures, where reliable and timely warning of cyberthreats is critical [23].

The feature extraction strategy is effective for the design and execution of legitimate security solutions, as well as for improving IDS performance [24–26]. In certain phenomenon detection methods, the need for greater accuracy and a lowered false alarm rate inspired the concept of data preprocessing and identification as the two mutual levels for IDS prototypes [27, 28]. The preprocessing phase removes the identification process which uses the reduction of attributes after removing redundant features from the dataset, retaining a decreased feature set that can be used to generate a high-performance version to predict attack classes using the base classifier.

Therefore, based on [8], this paper integrates the emerging infrastructure for applications of the Industrial Internet of Things. The authors reviewed the proposed scheme, offered the incorporation of work into a three-tier design for IIoT systems, and tested it against the NSL-KDD and UNSW-NB15 datasets. A rule-based model and a genetic search tool were used for the hybrid feature selection; thus, the evaluator subset was used to compute the connection between the class and each feature. The highest correlation from the attribute and class relationship is then chosen for selection. The merits of each attribute were then evaluated; function selection is known as the genetic search method, which produces attributes with the greatest value. If two attribute segments have the same performance score, the rule-based algorithm (rule assessment phase) produces the feature subset with the fewest volume of subset features. Finally, the features that have been selected are loaded into the ANN for template matching and assault selection. The ability of rule-based schemes combined with learning techniques to improve output precision has been demonstrated [29].

This was inspired by the assumption that integrating classifier optimization techniques into the feature representation and driving it with a rule-based algorithm would improve the performance of an IDS. The paper identifies intrusion in the IIoT network using the proposed model. The datasets used has huge features and parameters; an effective feature selection has to be employed to effectively reduce the high dimensionality of the datasets. This was done to reduce the burden this will have on the classifier. Furthermore, a feature extraction technique would make it easier for the classifier to select the most relevant qualities and exclude those that have a detrimental impact on classifier's performance. This motivated the creation of a new model using rule-based feature selection to effectively select the most relevant features from the datasets. The DFFNN classifier is used to train the features selected using this hybrid rule-based feature selection. The suggested model's performance is then assessed using current methodologies.

This paper's contributions are as follows:

(i) A system for intrusion prevention in the Industrial Internet of Things network is suggested

(ii) A hybrid approach focused on hybrid deep learning and rule-based feature selection for in-depth intrusion detection analysis

(iii) A relation of the current approach to prior methods for intrusion detection in the IIoT network is made. The proposed approach is stable, more efficient, and less resource-intensive, according to experimental results

2. Industrial Internet of Things Analytics Overview

Manufacturing, transportation, electricity, and healthcare are all affected by the Industry 4.0 revolution, which necessitates a change in industries that depend heavily on operational technology (OT). Previously, fog and edge computing [30] technologies were needed for Industrial IoT to ensure the required integration across Industry 4.0. However, this uprising introduces a new interrelated aspect that is critical for IIoT Analytics. The DL algorithms improve big data analytics capabilities, while IIoT Technologies enhance the utility of each of these categories. These algorithms can aid in the identification, categorization, and decision-making of each of these data types. The DL in combination with big data technologies generates practical and valuable data for policymaking. DL will be critical in IIoT and data analytics for effective and efficient selection, specifically in the field of streaming data and real-time insights in conjunction with edge computing systems [1].

Several business verticals, such as healthcare, grocery, automobiles, and transportation, are using IIoT applications. In many industries, the IIoT can greatly increase dependability, performance, and service quality. The IIoT will first develop current procedures and facilities, but the eventual aim is to create completely novel and vastly enhanced goods and services. Many companies recognize how and where IoT innovations and solutions can lead to organizational changes, new and improved goods and services, and entirely new business models. On the IIoT, machine learning and deep learning algorithms can increase reliability, production, and customer satisfaction by combining various machinery, procedures, apps, and applications. Anything necessitates a wide range of technology that must be carefully integrated and orchestrated.

These advancements in technology allow intelligent machines, tools, engines, and integrated control systems to execute repetitive duties to solve difficult problems without the need for human involvement [31]. Smart workplace advancements, intelligent data discovery, cognitive automation, and other aspects of business smartness should all be included. A digital twin is a virtual representation of physical assets, systems, and so on. This is generally alluded to as a result of the Internet of Things, which increasingly extends all of these appliances while providing us with a similarly increasing data collection that can be evaluated for

efficiency, architecture, and repair, among other things. Any digital twin's main advantage is that it is continually updated and "learns" any updates that occur in almost real time. The IIoT model and its applications are creating significant disruptions in the market globally.

2.1. The Four Key Components of Industrial IoT Architecture. Intelligent Edge Gateway: An intelligent edge gateway is a computer program closely aligned with sensor nodes that can capture, aggregate, and sanitize light data streaming. It allows one to upload tabulated and relevant data to the Internet of Things network. It acts as a connection between the hardware and the cloud IoT network in general.

IoT Cloud: The main IoT framework that uses data processing, machine knowledge, and artificial intelligence methods to handle massive quantities of data. The processing capabilities including device control, stream analytics, event management, a rules engine, alerts, and updates are all available. It offers components like big data analytics, as well as authorization, virtualization, end-to-end encryption, SDKs, and application APIs.

Business Incorporation and Platform: This is a backend framework that connects many IT schemes to certify that computer data is collected and processed in the full operational loop. ERP, QMS, planning and scheduling, and other systems are examples of such systems. Data analysis can be divided into three groups depending on the form of the result obtained. There are three types of analysis: descriptive, predictive, and prescriptive. Figure 1 displays IIoT architecture with four (4) layers including things, intelligent gateway, IoT cloud, and business application and integrations.

3. Related Work

As Anomaly Detection System (ADS) is an essential security management system that functions as a sniffer and deciding driver for routing traffic and spot suspicious activities [32], it functions as a packet capture and decoding engine for ensuring security and recognizing anomalous behavior. Since it can track both visible and invisible (zero-day) threats, the focus is on creating a pattern from standard data and treating any variance from it as an intrusion [33]. For example, the aim of [34, 35] centered on finding ADS using Particle Swarm Optimization (PSO) techniques for optimizing the performance of the One-Class Support Vector Machine (OCSVM) method by harvesting Modbus/TCP message network streams for testing and verifying the system. In [36], the authors built an IDS/ADS centered on this design, which was learned on offline data from a SCADA setting using network traces.

In [37], the authors constructed an IDS centered on the Modbus/TCP protocol setting using a K-NN classifier. While the aforementioned mechanisms performed admirably in certain cases, they were designed for particular configurations with a strong FPR. Similarly, in [38], the authors proposed an improved intrusion detection system (IDS) for matching the diverse structures of SCADA schemes using diverse OCSVM frameworks to select the right one for efficiently identifying multiple assaults. When operating, nevertheless, this computer used a large amount of computational power and had a high false warning rate for identification. Using SCADA mechanisms to obtain different aspects of contact events and using an SVM algorithm to identify attacks, authors in [39] suggested an ADS for detecting Modbus/TCP protocol-infiltrated assaults. The detection method, on the other hand, was ineffective in detecting irregular behaviors.

To prevent the effects of factors associated with the OCSVM's ability to track network attacks successfully, the authors in [40] merged the OCSVM method, and the recurrent K-means clustering algorithm was used. In another valuable effort, [41] proposed a critical infrastructure intrusion detection system centered on an artificial neural network (ANN) method that trained a multiperceptron ANN to identify anomalous network activity using fault back-propagation and Levenberg-Marquard features. Using a virtual network, in a relevant try, [42] used an ANN to detect DoS/DDoS attacks in IoTs, and in [43], the authors proposed a decentralized IDS based on artificial immunity for IoT devices. In [44], another set of researchers projected a Possibility Risk Identification-centered Intrusion Detection System (PRI-IDS) method for detecting replay attacks by inspecting Modbus TCP/IP protocol network traffic. However, these schemes had a high rate of false alarms and had trouble identifying certain new attacks.

In a related effort, the authors of [45] create a learning firewall that receives tagged samples and automatically configures itself by writing conservative preventive rules to avoid false alerts. We create a novel classifier family called classifiers that, unlike standard classifiers that just focus on accuracy, use zero false positive as the decision-making criterion. The authors first illustrate why naïve modifications of current classifiers, such as SVM, do not produce acceptable results and then present a generic iterative technique to achieve this goal. The proposed classifier, which is based on CART, is used to create a firewall for a Power Grid Monitoring System. We also put the technique to the test on the KDD CUP'99 dataset to see how well it works. The outcomes support the efficacy of our strategy.

IDSs have indeed been analyzed utilizing subsurface networks for identifying irregular findings from host and network-based systems by several researchers [46–48]. An ANN with a shallow network has one or two hidden layers, while a deep network has several hidden states of various architectures [49]. Deep learning is a form of a common machine-learning technique used by academic and industrial researchers because it can learn a detailed computational mechanism that mimics the normal behaviors of the human mind [50].

Several researcher has proved that the swiftness which received system signals is converted into massive datasets which pose a significant obstacle to IDS architectures' ability to analyze the subsequent large amounts of data for actual processing [51–53]. The authors in [54] suggested a new rule-based approach for detecting DoS assaults that relied on domain expert knowledge. For identifying DoS attacks, a rule-based classification algorithm was used, and the final classification was carried out by applying the rules from

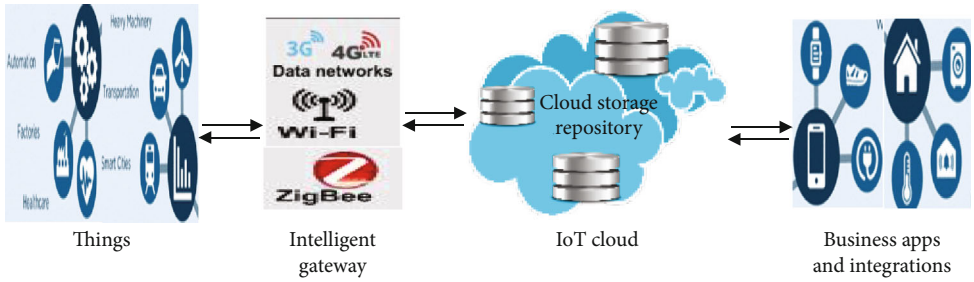


Figure 1: Industrial Internet of Things architecture.

the rule base and was confirmed using a domain expert. Feature selection techniques, also known as spatial removal, can aid in the conversion of databases from an elevated to a lesser spatial domain that better represents the problem space with the same efficacy [55, 56]. Unconnected variables may be eliminated without lowering data's importance to the detection model, which is the foundation of introduced feature collection [57, 58]. Most datasets have several attributes but few examples, according to [59, 60], possibly requiring and using feature selection techniques.

In [61], the authors provide an attack taxonomy based on the several layers of the IoT stack, such as device, infrastructure, communication, and service, as well as the specific characteristics of each layer that can be exploited by adversaries. Furthermore, we explain IoT-related vulnerabilities, exploitation techniques, attacks, impacts, and potential mitigation mechanisms and defense strategies using nine real-world cybersecurity incidents that attacked IoT devices deployed in the consumer, commercial, and industrial sectors. These with various additional examples emphasize the fundamental security vulnerabilities of IoT systems and indicate the possible attack implications of such interconnected ecosystems, while the suggested taxonomy provides a systematic approach for categorizing attacks based on the impacted layer and its impact.

A rule-based classifier-based data reduction strategy has been proposed in [62]. The suggested dimension reduction technique is an innovative data preparation technique that decreases both attributes and occurrences in testing specimens while keeping classifier precision. In [63], the authors suggested a fuzzy-based semisupervised learning method for IDS that constitutes a significant quantity of unlabelled data powered by labeled data to increase classification performance. The authors used the fuzzy measure to produce a trained independent hidden node feedforward neural network used to generate a fuzzy set vector of small, medium, and large specimen classification on unlabelled data. The training set is reused after using each vector of data classification independently in the initial training dataset. In a related work, the authors in [21] suggested a new method wrapper-based NIDS architecture based on Bayesian networks. The feature selection technique is used in this context to extract the appropriate features from the sample so that the Bayesian network classifier can reliably predict attack types.

For intrusion detection, [64] suggested a crossbreed method combining SVM and the ant colony. The aim of integrating the two machine learning techniques is to account for the shortcomings and capabilities of both methods to provide a more precise occurrence grouping. Similarly, in [65], the authors projected a wrapper approach for lightweight malware discovery based on decision trees. The suggested technique has four processes: preprocessing or removal of duplicate attack patterns, feature selection centered on a genetic algorithm (GA), postprocessing for standardized results, and traffic classification techniques centered on a neurotree technique. Similarly, in [66], a wrapping suitability purpose centered on a violation word for a wide amount of attributes with good classification precision and strict enforcement. The suggested wrapping fitness value is effective for feature extraction while maintaining prediction performance, according to the experiments. In [67], the paper suggested a decision tree classifier-based NIDS function collection depending on GA. The researchers used a GA to derive input data for decision trees as a classification algorithm to improve identification and reduce false alarms in cyberthreat detection.

In [53], the authors proposed a smart rule-based identification scheme for detecting Deprivation of Service (DoS) assaults in cloud servers scheme. The study used scoring and rating algorithms to simulate a cloud service, assault identities, and choose the best functionality. To discover assaults, a rule-based grouping procedure grounded on quality expertise was used to the selected features. The key benefit of their proposed model is a lower rate of false alarms and increased protection. But, due to the complex nature of attacks, the risk of confusion was not addressed. A modern feature selection strategy and a more effective KKN classifier were proposed in [68] for intrusion detection. The introduced feature set significantly reduces the existence of irrelevant features, thus improving KNN classifier's classification ability to distinguish kinds of invasion. Furthermore, the suggested feature selection algorithm reduces classifier's error warning rate dramatically. In a related work in [69], the authors suggested a new sophisticated artificial potential field technique for selecting features, as well as the implementation of a phased architecture as a base classifier for assault identification, when the suggested algorithm had a better classification exactness and a low wrong alarm degree as opposed to other approaches.

A machine learning classification algorithm to extract malware photographs with a mix of local and global characteristics was propped in [70, 71]. Their processes had a classification precision of 98.4% on a broad-scale study, using

9339 samples from 25 malware relatives in the Malimg dataset. Their methods achieved 99.21% classification accuracy in small-scale research, with 5288 samples from 8 malware relatives in the Malimg dataset. The authors in [72] created a CNN model that is used to separate threats from a corpus of binary executables. Moreover, this method had a classification accuracy of 98.52% when tested against a dataset of 9339 samples from 25 malware executables. Besides that, this template is used to randomly select 10% of samples in each loop to assess a malware family. In [73], the authors proposed a CNN-based malware classification model. From a dataset of 9339 samples, this model had a 98% accuracy. In each loop, a random method is used to pick 10% of samples to evaluate the malware family in question.

CNN used to create a malware classification model in [74]. The study used a corpus of 9339 samples from 25 diverse malware groups; this method had a 94.5% accuracy rate. In the same vein, in [75], the authors created a deep convolutional neural network that uses color image visualization to discover malware assaults on the Internet. Their findings showed that their classification efficiency for measuring cybersecurity threats had improved. The authors in [76] suggested a system built on Random Coefficient Selection and Mean Adjustment Method (RCSMMA). RCSMMA performs well against a variety of modern cyberattacks. Authors in [77] outlined the most important smart city applications and discussed the major issues of privacy and protection in smart city application architecture as a result of malware attacks. To avoid antagonists in the global sensor network, the authors in [78] proposed a stable steering and watching protocol using multivariant tuples.

To establish a powerful defense system against invaders, authors in [79] recommend developing strong intrusion detection systems that can detect intruders. In this paper, an ensemble classifier based on Crowd-Search is employed to categorize the UNSW-NB15 dataset, which is based on IoT. The most important characteristics from the dataset are first identified using the Crow-Search method and then provided to the ensemble classifier for training using the linear regression, Random Forest, and XGBoost algorithms. The proposed model's performance is then compared to that of state-of-the-art models to ensure that it is effective. The experimental results show that the suggested model outperforms the other models studied.

The widespread use of the internet in all aspects of human existence has raised the possibility of malicious attacks on the network. Intrusion detection systems have emerged as a result of the ease with which activities carried out via the network can spread. The patterns of attacks are also dynamic, necessitating effective cyberattack classification and prediction. To identify intrusion detection system (IDS) datasets, in [80], the authors proposed a hybrid principal component analysis (PCA)-firefly-based machine learning model. The dataset for this study was obtained from Kaggle. For the transformation of the IDS datasets, the model first uses One-Hot encoding. For dimensionality reduction, the hybrid PCA-firefly method is used. For classification, the XGBoost algorithm is used on the reduced dataset. To demonstrate the superiority of our suggested strategy, we undertake a detailed evaluation of the model using state-of-the-art machine learning approaches. The results of the experiments show that the suggested model outperforms the existing machine learning models.

From the existing related work, it can be seen that DL algorithms can considerably be used to increase the efficiency of IDS for IIoT by achieving the highest prediction performance while maintaining a low false alarm rate. Thus, motivating the use of the DL model with a hybrid rule-based technique for the automatic feature selection and sensing anomaly trends in data as suspect vectors using data transmission depth coverage. The proposed DFFNN based with hybrid rule-based feature selection model contains a rule-based model using a genetic search engine to select the relevant features and the DAE-DFFNN algorithm to classify IIoT network by classifying the constraint values of the DAE. It can find a good approximation for communication networks and transform high-dimensional data to low-dimensional data using DAE-DFFNN model's decreased layer, as explained in the following subsections.

4. The Proposed Intrusion Detection for Industrial Internet of Things Network

In this analysis, the deep feedforward neural network (DFFNN) is used to generate an effective ADS for IIoT locations. In the testing stage, a dual feature extraction employs a genetic search system as well as a rule-based algorithm. The subsection assesses or calculates the connection between individual features as well as the category. The class-attribute interaction with the highest similarity is used for filtering. This is referred to as function assessment. The genetic search procedure determines the qualities of each feature based on this function assessment and returns the attributes with the uppermost suitability value. If two attribute subsections have the same performance score, the rule-based algorithm (rule assessment phase) yields the feature vectors with the fewest quantity of subsection attributes. Finally, the chosen attributes are fed into the ANN, which is used to create models and classify attacks. These parameters are used to set up a standard DFFNN for discovering current and new attack instances. The DFFNN is used to detect mischievous vectors during the testing process. By translating the reduced hidden units, various hidden layers in the methodology will properly develop a detailed feature vector and grab the most important features. The subsections go into the specifics of the proposed system methodology.

4.1. Deep Feedforward Neural Network (DFFNN). The fundamental deep learning models are deep feedforward networks, often known as feedforward neural networks or multilayer perceptrons (MLPs). A feedforward network's purpose is to approximate a function f^x. For example, $y = f^x(\mathrm{x})$ transfers an input x to a category y in a classifier. A feedforward network learns the values of the parameters that result in the best function approximation by defining a mapping $y = f(\mathrm{x})$. Because information flows through the function being evaluated from x, the intermediate calculations

necessary to define f, and finally, to the output y, these models are referred to as feedforward models. There are no feedback links; therefore, model's outputs do not feedback into it. Recurrent neural networks are feedforward neural networks that have been extended to incorporate feedback connections. The DFFNN is usually described as an ANN method with input neurons, several hidden nodes, and an output neuron that are all directly connected without the use of a cycle [81].

The secret surface of each node reflects indistinct attributes dependent on the preceding stage's display, which are dynamically computed and processed in multiple layers to produce the outputs. This strategy is trained using a stochastic slope descent back-propagation methodology [82]. You can give a deeper feedforward neural network the ability to capture more complicated representations by creating a deeper feedforward neural network. If the complexity is justifiable, this could be justified. It has the advantage of being able to readily represent more complex functions.

The source data is fed into input nodes before being forwarded on to the hidden units, which generates a nonlinear manipulation of the information before being moved on to the output nodes in this deep-learning technique. To calculate the quality of the result, a feature role or back-propagation fault [83] is calculated, which is the discrepancy between the predicted and real presentation, and its value is transmitted backward across the unknown nodes to change the masses. The loss function is measured utilizing sole or minibatch specimens of the training examples rather than the whole set, with loads calibrated during each test to determine that the model is correctly suited.

This computation training data approach is based on the random chance of neural network variable activation, which results in the template being put in minima solutions with poor normalization [84]. To improve the convergence rate and the results of supervised learning, pretraining unsupervised strategies, specifically an AE, can be used to build the activation specifications [11].

4.2. Deep Autoencoder (DAE). A DAE is a feedforward neural network strategy for fast unsupervised computing execution [85]. It investigates the estimation of a unique task, where the result (x) is equal to the input $(\check{x})$ to construct a definition of a collection of data, that is, $(x \longrightarrow \check{x})$, (x). Its schematic representation consists of vectors $(x^{(\tilde{i})})$ in the input nodes and several concealed units of nonlinear initiation attributes. To learn compact features of the input data, the extracted features employ fewer neurons than the input nodes. As a result, it knows the most significant attributes and lowers spatial size and views the input data as an abstraction. At the end of the method, the output layer $(\check{x}^i)$ is shown as a close depiction of the input layer.

An AE's simplest framework comprises three layers: input, secret, and output. If the training data $(x^{(\tilde{i})})$ has n samples, each $(x^{(\tilde{i})})(i\epsilon(1, \cdots, n))$ has several proportions, as well as a spatial function vector $(d0)$; the Tanhinitiation function [85] is used and calculated using

$$T(t) = \frac{1 - e^{-2t}}{1 + e^{-2t}}. \tag{1}$$

The encoder and decoder are the two key components of the AE algorithm [86, 87]. A deterministic mapping called an encoder method $(f\theta)$ is used [86] to transform the input vector $(x^{(\tilde{i})})$ into a hidden layer representation $(z^{(\tilde{i})})$, and the dimensionality $x^{(\tilde{i})}$ is reduced to provide the right number of codes.

$$f\theta\left(x^{(i)}\right) = T(W_{x^{(i)}} + b), \tag{2}$$

where W is a d^0 x d^h, d^h weight matrix, d^h is the number of neurons in a concealed level $(d^0 < d^h)$, b is the bias vector, T is the Tanhinitiation utility, and θ, $[W, b]$ are the mapping parameters.

The product of the concealed layer's depiction is plotted, and the translator method is calculated by the deterministic plotting $(g\theta')$ as an approximation $(\check{x}^i)$ to restructure the input as an estimate $(\check{x}^i)$.

$$g\theta'\left(x^{(i)}\right) = T\left(W_{z^{(i)}}' + b'\right). \tag{3}$$

W′ is a d^0 x d^h weight matrix, b' is a bias vector, and θ' represents the mapping parameters $[W', b']$.

The information in that compressed representation is then used as inputs to reconstruct the original information after being transformed to fit the secret surface. The reform mistake (i.e., the alteration between the raw document and its low-dimensional reproduction) for a standard or minibatch training set(s) is calculated by the training process.

$$E\left(x, \check{x}\right) = \frac{1}{2}\sum_{i}^{s}\left\|x^{(i)} - \check{x}^{(i)}\right\|^2, \theta = \{W, b\} = \operatorname{argmin}_{\theta} E\left(x, \check{x}\right). \tag{4}$$

Feature selection phase:

Definition 1 (subset). A feature V_i is said to be relevant if there exists some v_i and c for which $p(V_i = v_i) > 0$ such that

$$p(C = c \mid V_i = v_i) \neq p(C - c). \tag{5}$$

Definition 2 (SubsetEval). If the connection between an individual component of when the association between a function and the outside parameter is understood, as well as the intercorrelation across each set of parameters, the connection between a standardized test made up of the combined modules and the outside parameter can be estimated in (6).

$$r_{zc} = \frac{k_{\overline{r_{zi}}}}{\sqrt{k + k(k-l)\overline{r_{ii}}}}, \tag{6}$$

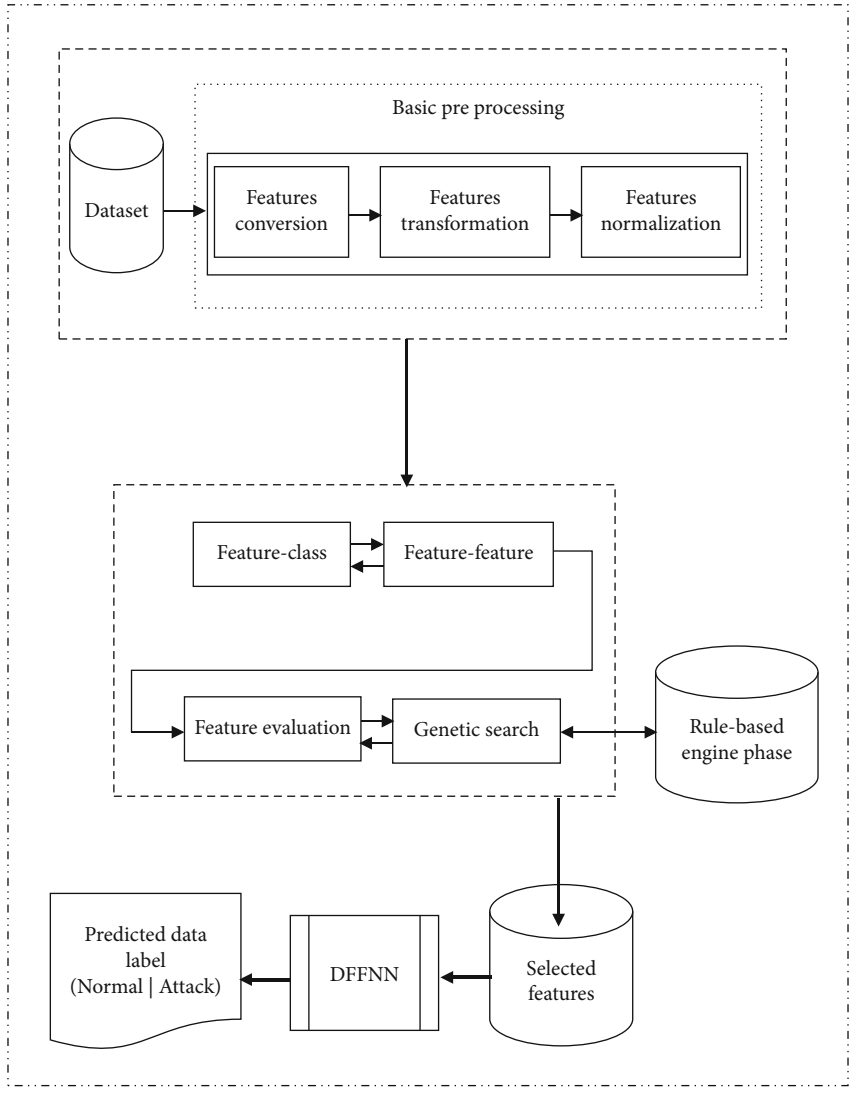


Figure 2: The proposed intrusion detection in industrial internet of things network.

where r_{zc} seems to be the association between the sum of the modules and the external parameter and k is the number of elements, r_{zi} is the average of the component-to-outside-variable correlations, and r_{ii} is the average component-to-outside-variable intercorrelation.

Definition 3 (Genetic search). The term "genetic search" refers to an exploration that is motivated by normal progression. A suitability task that is a lined grouping of an accuracy duration and an effortlessness duration is used in this genetic search.

$$\text{Fitness}(X) = \frac{3}{4}A + \frac{1}{4} = \left(1 - \frac{S+F}{2}\right), \tag{7}$$

where X represents a function subset, A represents DFFNN's average cross-validation precision, S represents the number of instances or training samples, and F represents the number of subset features.

Definition 4 (Rule engine). If there are several feature subsets $(F>)$ with identical fitness values, the rule-based instrument yields a feature subgroup (V_i) with fewer features (X_F), else, it yields the feature subgroup with the uppermost appropriateness value (F_{hi}) to the base classifier as in (8).

$$R = \begin{cases} V_i, & \text{if } V_i \epsilon F > \cap X_f, \\ V_i, & \text{if } F_{hi \cap \varnothing}. \end{cases} \tag{8}$$

This study suggests an effective intrusion discovery model for safeguarding the IIoT system against the mischievous activity. Figure 2 displayed the architecture of the projected model with the training and testing phases.

Figure 2 shows the proposed intrusion detection in IIoT network. In an IIoT setting, the proposed scheme investigates and chooses critical information from large-scale data. The first phase in the suggested method is data preprocessing, which includes function translation and regularization model.

4.2.1. Feature Transformation. Since the suggested framework only embraces mathematical properties, the apiece rhetorical attribute value is transformed into a mathematical formula; for instance, the NSL-KDD dataset contains multiple figurative attributes like procedure natures with reference values like ICMP, TCP, and UDP, which are plotted to 1, 2, and 3, respectively.

4.2.2. Feature Normalization. Since DL relies on different features based on masses. Data could be skewed into various spots due to levels, causing some values to update quicker than others [8, 11]. As a result, it is important to deal with this problem using statistical normalization, in which the Z − score function for each feature value $(v^{(i)})$ is calculated by

$$Z^{(i)} = \frac{v^{(i)} - \mu}{\sigma}, \tag{9}$$

where σ is the standard deviation and is the mean of the μ values for a given function $(v^{(i)}(i\epsilon\, 1, 2, 3, \cdots, n))$.

Since networks have high dimensionality, it is important to minimize it to increase computing resources and develop a compact and flexible ADS strategy [88]. As a result, the suggested DAE-DFFNN method is used to decrease high proportions to low proportions via a main reduced surface. More specifically, the model contains a nonlinear mechanism that encrypts a lot of features into the lesser feature set in the reduced hidden state, requiring dimensionality reduction to be realistic without the necessity for professional acquaintance. The purpose of the rule based with DAE-DFFNN dimension reduction is to identify excellently embodiments from the unclear framework in the probability model in terms of increased learning and also processed and decreased attributes.

4.3. Details of the Datasets Used. The testing, examining, and assessing of the behavior of the discovery scheme depends solely on the dataset, and this plays a vital role in getting a better result. A high-performance one not only yields effective outcomes for an offline device but can be successful in an actual setting. Most authors also used the well-known NSL-KDD datasets, which is a revised variant of the KDD CUP 99 database that solves the KDD CUP 99's main problems by deleting duplicate information and selecting documents concerning their proportions. It comprises 148,517 documents (77,054 standards and 71,460 assaults) after preprocessing, apiece of which includes 41 attributes and a class mark. Probing, DoS, user to root (U2R), remote to local (R2L), and normal are the five classes [89, 90]. However, despite being commonly used in IDSs, it is now obsolete [91]. As a result, a novel dataset called UNSW-NB15 is used to effectively test our proposed work. It includes contemporary synthesized attack activities and represents actual current normal behaviors [92]. It has a total of 257,673 records (93,000 regular and 164,673 attacks), each with 41 features and a classification mark. Fuzzers, examination, backdoors, DoS, vulnerabilities, standard, reconnaissance, shellcode, and worm are all among the ten separate class labels, one standard, and nine attacks.

4.4. Performance Analysis. To assess the performance and comparison of the proposed algorithm using DL and hybrid rule-based model with other existing models, the following performance metrics were used. The amount of correct and incorrect outcomes in a classification problem was summed and compared; the results were with the reference results. Accuracy, precision, recall, specificity, and $F1$-score are just a few of the most common matrices. True-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) statistical indices were calculated to solve the confusion matrix, as shown in Equations (10)–(16).

$$\text{Accuracy} : \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \tag{10}$$

$$\text{Precision} : \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{11}$$

$$\text{Sensitivity or Recall} : \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{12}$$

$$\text{Specificity} : \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{13}$$

$$\text{F1} - \text{score} : \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{14}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{15}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{16}$$

From Equations (10) to (16), accuracy denotes how often the prediction is correct, whereas precision denotes how often the class will be correct during prediction. However, recall indicates how much of the all-positive class was correctly predicted, whereas specificity assesses how well the negatives were identified. The $F1$-score is a combination of exactness and recall. The quantity of correct negative estimates distributed by the total quantity of negative forecasts is known as specificity. The true-positive rate (TPR) is defined as the proportion of properly recognized attacks over the total quantity of dataset classes, as seen in Equation (15). The TPR stands for discovery rate. The false alarm rate (FAR) is calculated by dividing the number of records wrongly denied by the total number of normal records. Equation (16) defines the FAR evaluation metric. As a result, in the IIoT system, the impetus for intrusion detection prediction is to achieve a higher accuracy and detection rate (DR) with a lower false alarm rate.

5. Results and Discussion

The R programming language platforms were used to implement the proposed model, and the evaluation was done using the explained performance metrics. Both datasets with the relevant DAE-DFFNN with dual rule-based design are used to seamlessly incorporate all characteristics. The NSL-KDD dataset contains 77,054 regular documents and 71,460 assault documents, as well as different samples from the UNSW-NB15 dataset, which contains

Table 1: Proposed method evaluation.

Dataset	TP rate	FP rate	Precision	Recall	*F*-measure	ROC	Class
UNSW-NB15	0.998	0.1	0.967	0.998	0.989	0.989	Attack
	0.999	0.001	0.998	0.996	0.967	0.998	Normal
NSL-KDD	0.996	0.1	0.984	0.999	0.997	0.997	Attack
	0.999	0.001	0.998	0.993	0.969	0.998	Normal

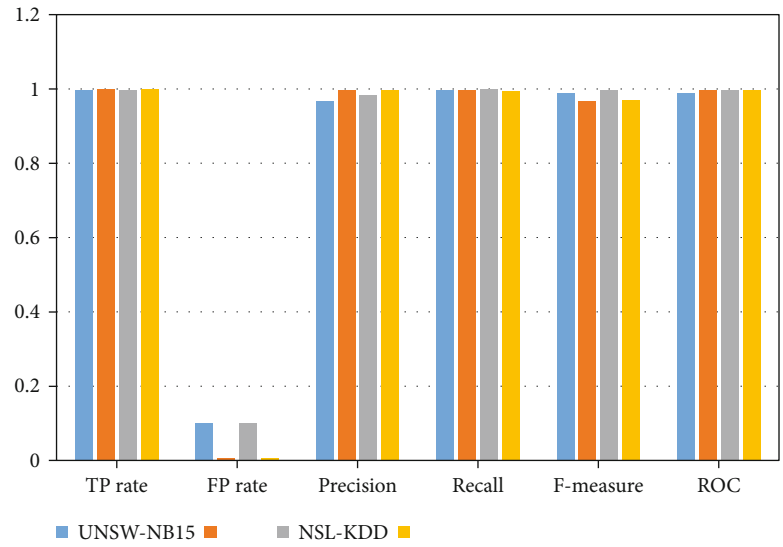


Figure 3: The performance evaluation of the proposed model on both datasets.

93,000 regular documents and 92,000 assault documents, with 20 percent of the normal records represents 40%, 20%, and 60% of the testers which were used for testing, respectively.

The network structures and parameters adopted based on the experiments yield the peak DR and lowermost FPR. The proposed model used the best network structures for both datasets after the best features are selected using the hybrid rule-based genetic search engine in addition to the DAE feature selection model are 41 nodes for one input layer, 10, 3, and 10 nodes for the three hidden layers and 41 nodes for the output layer for the DAE technique, and 2 nodes for the DFFNN model for the output layer with 2 nodes. For the NSL-KDD dataset, 0.0015 is the learning rate and 0.2 momenta start, and for UNSW-NB15 dataset, L1 and L2 regularizations of L1 = L2 = $1e-6$, momentum start of 0.2, momentum stable of 0.4, $1e7$ ramp momentum, annealing rate of $2e$-6, and 100 epochs; 0.002 learning rate for the Tanh activation function was used.

Table 1 and Figure 3 show the performance of the projected model using both NSL-KDD and UNSW-NB15 datasets using numerous metrics. The results obtained using various metrics show that the projected model is very important and relevant in intrusion detection of IIoT network for attack prediction and classification.

Table 2 displays the accuracy, detection rate, and FPR of the proposed model on the datasets. The findings reveal that the model outperforms the UNSW-NB15 dataset on NSL-KDD, with a precision of 99.0 percent, a detection rate of 99.0 percent, and an FPR of 1.0 percent.

Table 2: Evaluation of performances for two datasets.

Dataset	Accuracy	Detection rate	FPR
NSL-KDD	99.0%	99.0%	1.0%
UNSW-NB15	98.9%	99.9%	1.1%

Table 3 shows the discovery rates for the classes in the NSL-KDD and UNSW-NB15 datasets using the projected model. The outcomes for the UNSW-NB15 dataset are displayed in Table 3 and Figure 3 for the discovery rates of the record types classes: analysis (92.3%), backdoor (95.2%), DoS (97.3%), exploits (98.0%), fuzzer (67.1%), generic (99.8%), normal (99.6%), salicode (90.3%), worm (81.7%), reconnaissance (92%), and shellcode (90.2%), respectively. The results for the NSL-KDD dataset using the projected model are displayed in Table 3 and Figures 4 and 5 to determine the records types like DoS, normal, U2R, R2L, and probe with discovery rates of 99.2%, 99.7%, 75.5%, 94.3%, and 99.0%, respectively. The proposed model demonstrated overall better performance for intrusion detection in both used datasets even though some results like U2R, fuzzer, and worms are not too high in both datasets.

5.1. The Comparison of the Proposed Model with Existing Methods. To show how feature selection affects classification algorithm's detection efficiency, Table 4 compares the

Table 3: Detection rates for NSL-KDD and UNSW-NB15 dataset classes.

Dataset	DoS	Normal	Backdoor	Worm	Shellcode	Probe	Exploits	U2R	Salicode	Analysis	R^2	Fuzzer	Generic	Reconnaissance
NSL-KDD	99.2%	99.7%	—	—	—	99.0%	—	75.5%	—	—	94.8%	—	—	—
UNSW-NB15	97.3%	99.6%	95.2%	81.7%	90.2%	—	98.0%	—	90.3%	92.3%		67.1%	99.8%	92.0%

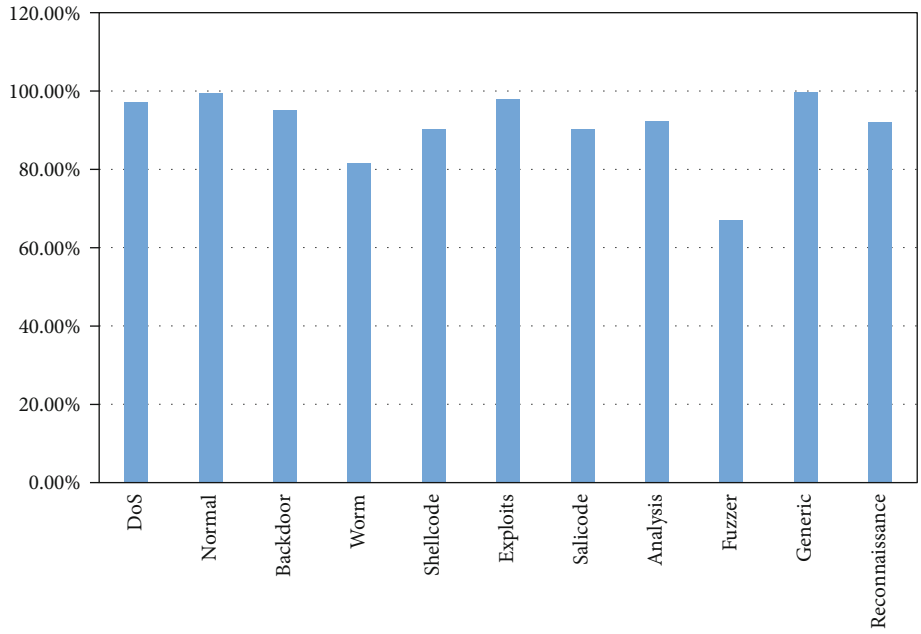


Figure 4: Detection rates for UNSW-NB15 dataset classes.

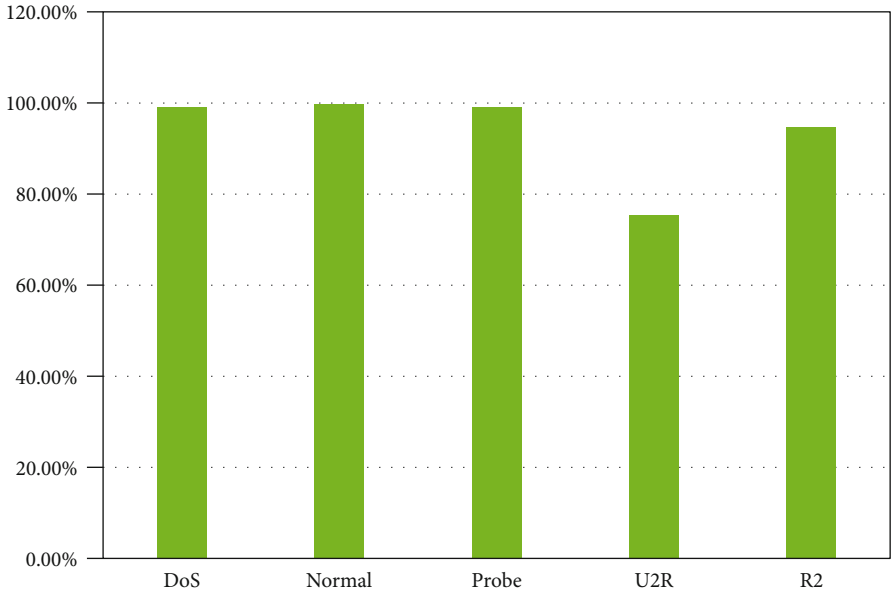


Figure 5: Detection rates for NSL-KDD dataset classes.

proposed approach to several known approaches. Table 4 displays the cumulative performance measures for the proposed system and other models using the decreased UNSW-NB15 dataset. The precision and FPR of the suggested approach are better than those of other approaches. The suggested network intrusion detection method, in general, has a 98.9 percent accuracy, which is 0.1 percent higher than the modified KNN with the second-highest accuracy. Similarly, as compared to other classifiers, the proposed method's FPR has a very low error percentage of 1.1 percent. When equated to other techniques using the reduced UNSW-NB15 dataset, the proposed approach performed better across all evaluation metrics. The proposed method's marginally higher accuracy is due to its robust feature selection and rule-based fitness assessment.

The suggested model's efficiency is contrasted to that of nine recently developed anomaly detection techniques, including the ADS system based on DL, the Filter-based Support Vector Machine (F-SVM) [95], the Computer Vision Method (CVT) [96], the Dirichlet Mixture Model (DMM) [91], the Triangular Area Nearest Neighbors (TANN) [97], DBN [98], RNN [52], DNN [81], and Ensemble-DNN [99]. Table 5 compares the identification rate and false-positive rate of our proposed system to other

Table 4: Summary of performance comparison for UNSW-NB15 dataset.

Model	Performance metrics					
	Accuracy (%)	FPR (%)	*F*-score (%)	Recall (%)	Precision (%)	ROC curve (%)
Wrapper + neurotree [67]	98.38	1.62	0.984	0.980	0.989	0.998
SVM+EML+K-means [58]	95.75	1.87	0.944	0.997	0.897	0.986
GA +SVM [93]	97.3	0.017	0.966	0.997	0.938	0.981
CNN+LSTM [94]	94.12	—	0.956	0.989	0.925	0.984
Modified KNN [70]	98.7	1.3	0.992	0.996	0.988	0.998
CfsSubsetEval + GA+RuleEval+ANN [8]	98.8	1.2	0.989	0.989	0.989	0.998
Proposed model	98.9	1.1	0.989	0.998	0.967	0.989

Table 5: For the NSL-KDD dataset, the results of the proposed model have been compared with nine other classifiers.

Technique	Detection rate	FPR
F-SVM [95]	92.2%	8.7%
CVT [96]	95.3%	5.6%
DMM [91]	97.2%	2.4%
TANN [97]	91.1%	9.4%
DBN [98]	95.1%	4.5%
RNN [52]	73%	3.6%
DNN [81]	76%	15%
Ensemble-DNN [99]	98%	14.7%
ADS-DL [11]	99%	1.8%
Proposed model	98.9	1.1%

models tested on the NSL-KDD dataset. Our developed scheme delivers the desired performance, with 99 percent DR and 1.8 percent FPR. The first four models demonstrated rational results in identifying destructive events after a feature selection process. F-SVM used shared information to solve linear and nonlinear data properties, which was then paired with the SVM for attack detection. Nevertheless, to improve IDS efficiency, this model's search strategy must be refined. CVT and TANN used the PCA technique to reduce the data measurements.

The F-SVM has a detection rate of 92.2% and FPR of 8.7%, CVT with a detection rate of 95.3% and FPR of 5.6%, DMM with a detection rate of 97.2% and FPR of 2.4%, TANN with a detection rate of 91.1% and FPR of 9.4%, DBN with a detection rate of 95.1% and FPR of 4.5%, RNN with a detection rate of 73.0% and FPR of 3.6%, DNN with a detection rate of 76.0% and FPR of 15%, ensemble-DNN with a detection rate of 98.0% and FPR of 14.7%, and ADS with detection rate of 99.0% and FPR of 1.8%. The proposed model differs from previous DL-based IDSs in that it uses a basic mathematical algorithm (DAE) and a hybrid rule-based function selection to estimate parameters that are appropriate DFFNN input to create its classification effectively and efficiently. Moreover, the model knows and examines high-level functionality, automatically decreases data dimensionality, and effectively portrays important features due to the reduced hidden layer. As a consequence, the proposed model is optimal for use in a real-world industrial environment with a vast amount of unlabeled and unstructured data, such as IIoT.

6. Conclusion

This paper proposes an ADS model for identifying destructive activities in IIoT networks utilizing data from TCP/IP packets. It employs unsupervised DL strategies that are hybrid rule-based with automated dimensionality reductions to provide a good description of standard network structures for unsupervised learning. The suggested DAE-DFFNN with hybrid rule-based design is successfully used to develop and remove essential features that improve its overall efficiency. As compared to other strategies developed in recent research, the proposed model achieves the maximum identification rate of 99.0 percent and the fewest false alarms of 1.0 percent when checked on different data samples from the NSL-KDD and NSW-NB15 datasets. Both NSL-KDD and NSW-NB15 were included in the proposed model since they are often used by researchers in intrusion detection and as a benchmark. The use of hybrid rule-based feature collection improves the consistency of the proposed model by using only appropriate features for class classification in the datasets. The future analysis would consider the use of real-world data gathered by the IIoT system to determine the effectiveness of its operation in these settings. In addition, in future work, the proposed model will be extended to accommodate different protocols.

Data Availability

No data available.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. Ambika, "Machine learning and deep learning algorithms on the Industrial Internet of Things (IIoT)," *Advances in Computers*, vol. 117, no. 1, pp. 321–338, 2020.

[2] R. Ashima, A. Haleem, S. Bahl, M. Javaid, S. K. Mahla, and S. Singh, "Automation and manufacturing of smart materials in Additive Manufacturing technologies using the Internet of Things towards the adoption of Industry 4.0," *Materials Today: Proceedings*, vol. 45, pp. 5081–5088, 2021.

[3] L. M. Gladence, V. M. Anu, R. Rathna, and E. Brumancia, "Recommender system for home automation using IoT and artificial intelligence," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–9, 2020.

[4] T. Sherasiya, H. Upadhyay, and H. B. Patel, "A survey: intrusion detection system for internet of things," *International Journal of Computer Science and Engineering (IJCSE)*, vol. 5, no. 2, pp. 91–98, 2016.

[5] J. B. Awotunde, R. G. Jimoh, S. O. Folorunso, E. A. Adeniyi, K. M. Abiodun, and O. O. Banjo, "Privacy and security concerns in IoT-based healthcare systems," *Internet of Things*, pp. 105–134, 2021.

[6] E. A. Adeniyi, R. O. Ogundokun, and J. B. Awotunde, "IoMT-based wearable body sensors network healthcare monitoring system," in *IoT in Healthcare and Ambient Assisted Living*, pp. 103–121, Springer, Singapore, 2021.

[7] K. Amit and C. Chinmay, "Artificial intelligence and Internet of Things based healthcare 4.0 monitoring system," *Wireless Personal Communications*, pp. 1–14, 2021.

[8] F. E. Ayo, S. O. Folorunso, A. A. Abayomi-Alli, A. O. Adekunle, and J. B. Awotunde, "Network intrusion detection based on deep learning model optimized with rule-based hybrid feature selection," *Information Security Journal: A Global Perspective*, vol. 29, no. 6, pp. 267–283, 2020.

[9] M. Abdulraheem, J. B. Awotunde, R. G. Jimoh, and I. D. Oladipo, "An efficient lightweight cryptographic algorithm for IoT security," in *Communications in Computer and Information Science*, pp. 444–456, Springer, 2021.

[10] A. Bakhtawar, R. J. Abdul, C. Chinmay, N. Jamel, R. Saira, and R. Muhammad, "Blockchain and ANFIS empowered IoMT application for privacy preserved contact tracing in COVID-19 pandemic," *Personal and Ubiquitous Computing*, 2021.

[11] A. H. Muna, N. Moustafa, and E. Sitnikova, "Identification of malicious activities in industrial internet of things based on deep learning models," *Journal of information security and applications*, vol. 41, pp. 1–11, 2018.

[12] E. Sitnikova, E. Foo, and R. B. Vaughn, "The power of hands-on exercises in SCADA cybersecurity education," in *Information Assurance and Security Education and Training*, pp. 83–94, Springer, Berlin, Heidelberg, 2013.

[13] S. Dash, C. Chakraborty, S. K. Giri, S. K. Pani, and J. Frnda, "BIFM: big-data driven intelligent forecasting model for COVID-19," *IEEE Access*, vol. 9, pp. 97505–97517, 2021.

[14] G. Tzokatziou, L. A. Maglaras, H. Janicke, and Y. He, "Exploiting SCADA vulnerabilities using a human interface device," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 7, pp. 234–241, 2015.

[15] D. Kushner, "The real story of stuxnet," *IEEE Spectrum*, vol. 50, no. 3, pp. 48–53, 2013.

[16] P. W. Khan and Y. Byun, "A blockchain-based secure image encryption scheme for the industrial Internet of Things," *Entropy*, vol. 22, no. 2, p. 175, 2020.

[17] Q. Yan and F. R. Yu, "Distributed denial of service attacks in software-defined networking with cloud computing," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 52–59, 2015.

[18] A. C. Enache and V. Sgârciu, "Anomaly intrusions detection based on support vector machines with an improved bat algorithm," in *2015 20th International Conference on Control Systems and Computer Science*, pp. 317–321, Bucharest, Romania, May 2015.

[19] O. Folorunso, F. E. Ayo, and Y. E. Babalola, "Ca-NIDS: a network intrusion detection system using combinatorial algorithm approach," *Journal of Information Privacy and Security*, vol. 12, no. 4, pp. 181–196, 2016.

[20] H. Zhang, D. D. Yao, N. Ramakrishnan, and Z. Zhang, "Causality reasoning about network events for detecting stealthy malware activities," *Computers & Security*, vol. 58, pp. 180–198, 2016.

[21] M. R. Kabir, A. R. Onik, and T. Samad, "A network intrusion detection framework based on Bayesian network using a wrapper approach," *International Journal of Computer Applications*, vol. 166, no. 4, pp. 13–17, 2017.

[22] Y. Hu, A. Yang, H. Li, Y. Sun, and L. Sun, "A survey of intrusion detection on industrial control systems," *International Journal of Distributed Sensor Networks*, vol. 14, no. 8, 2018.

[23] T. Cruz, L. Rosa, J. Proenca et al., "A cybersecurity detection framework for supervisory control and data acquisition systems," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 6, pp. 2236–2246, 2016.

[24] J. Camacho, A. Pérez-Villegas, P. García-Teodoro, and G. Maciá-Fernández, "PCA-based multivariate statistical network monitoring for anomaly detection," *Computers & Security*, vol. 59, pp. 118–137, 2016.

[25] M. Grill, T. Pevný, and M. Rehak, "Reducing false positives of network anomaly detection by local adaptive multivariate smoothing," *Journal of Computer and System Sciences*, vol. 83, no. 1, pp. 43–57, 2017.

[26] L. A. Maglaras, J. Jiang, and T. J. Cruz, "Combining ensemble methods and social network metrics for improving accuracy of OCSVM on intrusion detection in SCADA systems," *Journal of Information Security and Applications*, vol. 30, pp. 15–26, 2016.

[27] R. O. Ogundokun, J. B. Awotunde, E. A. Adeniyi, and F. E. Ayo, "Crypto-Stegno based model for securing medical information on IOMT platform," *Multimedia tools and applications*, pp. 1–23, 2021.

[28] J. Soto and M. Nogueira, "A framework for resilient and secure spectrum sensing on cognitive radio networks," *Computer Networks*, vol. 115, pp. 130–138, 2017.

[29] M. S. Abadeh, J. Habibi, and C. Lucas, "Intrusion detection using a fuzzy genetics-based learning algorithm," *Journal of Network and Computer Applications*, vol. 30, no. 1, pp. 414–428, 2007.

[30] M. Aazam and E. N. Huh, "Fog computing microdata center-based dynamic resource estimation and pricing model for IoT," in *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*, pp. 687–694, Gwangju, Korea, March 2015.

[31] C. Cecchinel, M. Jimenez, S. Mosser, and M. Riveill, "An architecture to support the collection of big data in the internet of

things," in *2014 IEEE World Congress on Services*, pp. 442–449, Anchorage, AK, USA, June 2014.

[32] N. Moustafa, J. Hu, and J. Slay, "A holistic review of network anomaly detection systems: a comprehensive survey," *Journal of Network and Computer Applications*, vol. 128, pp. 33–55, 2019.

[33] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Machine learning models for secure data analytics: a taxonomy and threat model," *Computer Communications*, vol. 153, pp. 406–440, 2020.

[34] N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," *Information Security Journal: A Global Perspective*, vol. 25, no. 1-3, pp. 18–31, 2016.

[35] W. Shang, P. Zeng, M. Wan, L. Li, and P. An, "Intrusion detection algorithm based on OCSVM in industrial control system," *Security and Communication Networks*, vol. 9, no. 10, p. 1049, 2016.

[36] L. A. Maglaras and J. Jiang, "Intrusion detection in SCADA systems using machine learning techniques," in *2014 Science and Information Conference*, pp. 626–631, London, UK, August 2014.

[37] P. Silva and M. Schukat, "On the use of k-nn in intrusion detection for industrial control systems," in *Proceedings of The IT&T 13th International Conference on Information Technology and Telecommunication*, pp. 103–106, Dublin, Ireland, August 2014.

[38] B. Stewart, L. Rosa, L. A. Maglaras et al., "A novel intrusion detection mechanism for scada systems which automatically adapts to network topology changes," *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 4, no. 10, 2017.

[39] W. Shang, J. Cui, M. Wan, P. An, and P. Zeng, "Modbus communication behavior modeling and SVM intrusion detection method," in *Proceedings of the 6th International Conference on Communication and Network Security*pp. 80–85, Singapore, November 2016.

[40] L. A. Maglaras and J. Jiang, "Ocsvm model combined with k-means recursive clustering for intrusion detection in scada systems," in *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, pp. 133-134, Rhodes, Greece, August 2014.

[41] O. Linda, T. Vollmer, and M. Manic, "Neural network-based intrusion detection system for critical infrastructures," in *2009 International Joint Conference on Neural Networks*, pp. 1827–1834, Atlanta, GA, USA, June 2009.

[42] E. Hodo, X. Bellekens, A. Hamilton et al., "Threat analysis of IoT networks using artificial neural network intrusion detection system," in *2016 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–6, Yasmine Hammamet, Tunisia, May 2016.

[43] R. Chen, C. M. Liu, and C. Chen, "An artificial immune-based distributed intrusion detection model for the internet of things," in *Advanced materials research*, pp. 165–168, Trans Tech Publications Ltd, 2012.

[44] T. Marsden, N. Moustafa, E. Sitnikova, and G. Creech, "Probability risk identification based intrusion detection system for SCADA systems," in *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pp. 353–363, Springer, Cham, Switzerland, 2017.

[45] M. S. Haghighi, F. Farivar, and A. Jolfaei, "A machine learning-based approach to build zero false-positive IPSs for industrial IoT and CPS with a case study on power grids security," *IEEE Transactions on Industry Applications*, pp. 1–9, 2020.

[46] N. Gao, L. Gao, Q. Gao, and H. Wang, "An intrusion detection model based on deep belief networks," in *2014 Second International Conference on Advanced Cloud and Big Data*, pp. 247–252, Huangshan, China, November 2014.

[47] B. Abolhasanzadeh, "Nonlinear dimensionality reduction for intrusion detection using auto-encoder bottleneck features," in *2015 7th Conference on Information and Knowledge Technology (IKT)*, pp. 1–5, Urmia, Iran, May 2015.

[48] Y. Li, R. Ma, and R. Jiao, "A hybrid malicious code detection method based on deep learning," *International Journal of Security and Its Applications*, vol. 9, no. 5, pp. 205–216, 2015.

[49] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 5, pp. 21954–21961, 2017.

[50] J. C. van Dijk and P. Williams, "The history of artificial intelligence," in *Expert Systems in Auditing*, pp. 21–26, Palgrave Macmillan, London, 1990.

[51] A. Karkouch, H. Mousannif, H. Al Moatassime, and T. Noel, "Data quality in internet of things: a state-of-the-art survey," *Journal of Network and Computer Applications*, vol. 73, pp. 57–81, 2016.

[52] Y. Qin, Q. Z. Sheng, N. J. Falkner, S. Dustdar, H. Wang, and A. V. Vasilakos, "When things matter: a survey on data-centric Internet of Things," *Journal of Network and Computer Applications*, vol. 64, pp. 137–153, 2016.

[53] F. Zafar, A. Khan, S. U. R. Malik et al., "A survey of cloud computing data integrity schemes: design challenges, taxonomy and future trends," *Computers & Security*, vol. 65, pp. 29–49, 2017.

[54] R. Rajendran, S. V. N. Santhosh Kumar, Y. Palanichamy, and K. Arputharaj, "Detection of DoS attacks in cloud networks using intelligent rule based classification system," *Cluster Computing*, vol. 22, no. S1, pp. 423–434, 2019.

[55] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," *Expert Systems with Applications*, vol. 67, pp. 296–303, 2017.

[56] L. C. Leonard, "Web-based behavioral modeling for continuous user authentication (CUA)," in *Advances in Computers*, pp. 1–44, Elsevier, 2017.

[57] C. Sammut and G. I. Webb, "Feature selection," in *In Encyclopedia of Machine Learning Edited by Claude Sammut, Geoffrey I. Webb*, pp. 429–433, Springer, 2010.

[58] S. Goswami, A. K. Das, A. Chakrabarti, and B. Chakraborty, "A feature cluster taxonomy based feature selection technique," *Expert Systems with Applications*, vol. 79, pp. 76–89, 2017.

[59] D. Acarali, M. Rajarajan, N. Komninos, and I. Herwono, "Survey of approaches and features for the identification of HTTP-based botnet traffic," *Journal of network and computer applications*, vol. 76, pp. 1–15, 2016.

[60] V. Snášel, J. Nowaková, F. Xhafa, and L. Barolli, "Geometrical and topological approaches to Big Data," *Future Generation Computer Systems*, vol. 67, pp. 286–296, 2017.

[61] C. Xenofontos, I. Zografopoulos, C. Konstantinou, A. Jolfaei, M. K. Khan, and K. K. R. Choo, "Consumer, commercial and

industrial IoT (in) security: attack taxonomy and case studies," *IEEE Internet of Things Journal*, 2021.

[62] V. Herrera-Semenets, O. Andrés Pérez-García, R. Hernández-León, J. van den Berg, and C. Doerr, "A data reduction strategy and its application on scan and backscatter detection using rule-based classifiers," *Expert Systems with Applications*, vol. 95, pp. 272–279, 2018.

[63] R. A. R. Ashfaq, X. Z. Wang, J. Z. Huang, H. Abbas, and Y. L. He, "Fuzziness based semi-supervised learning approach for intrusion detection system," *Information Sciences*, vol. 378, pp. 484–497, 2017.

[64] W. K. Kirnapure and A. R. B. Patil, "Classification, detection and prevention of network attacks using rule based approach," *International Research Journal of Engineering and Technology*, vol. 4, no. 4, pp. 1453–1459, 2017.

[65] S. S. Sivatha Sindhu, S. Geetha, and A. Kannan, "Decision tree based light weight intrusion detection using a wrapper approach," *Expert Systems with Applications*, vol. 39, no. 1, pp. 129–141, 2012.

[66] B. Chakraborty and A. Kawamura, "A new penalty-based wrapper fitness function for feature subset selection with evolutionary algorithms," *Journal of Information and Telecommunication*, vol. 2, no. 2, pp. 1–18, 2018.

[67] G. Stein, B. Chen, A. S. Wu, and K. A. Hua, "Decision tree classifier for network intrusion detection with GA-based feature selection," in *Proceedings of the 43rd annual southeast regional conference on - ACM-SE 43*, pp. 136–141, Kennesaw, Georgia, 2005.

[68] B. Senthilnayaki, K. Venkatalakshmi, and A. Kannan, "Intrusion detection system using fuzzy rough set feature selection and modified KNN classifier," *International Arab Journal of Information Technology*, vol. 16, no. 4, pp. 746–753, 2019.

[69] S. Ganapathy and A. Kannan, "Energy-aware cluster and neuro-fuzzy based routing algorithm for wireless sensor networks in IoT," *Computer Networks*, vol. 151, pp. 211–223, 2019.

[70] H. Naeem, B. Guo, M. R. Naeem, F. Ullah, H. Aldabbas, and M. S. Javed, "Identification of malicious code variants based on image visualization," *Computers & Electrical Engineering*, vol. 76, pp. 225–237, 2019.

[71] H. Naeem, B. Guo, F. Ullah, and M. R. Naeem, "A cross-platform malware variant classification based on image representation," *KSII Transactions on Internet & Information Systems*, vol. 13, no. 7, 2019.

[72] M. Kalash, M. Rochan, N. Mohammed, N. D. Bruce, Y. Wang, and F. Iqbal, "Malware classification with deep convolutional neural networks," in *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pp. 1–5, Paris, France, February 2018.

[73] R. Kumar, Z. Xiaosong, R. U. Khan, I. Ahad, and J. Kumar, "Malicious code detection based on image processing using deep learning," in *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence - ICCAI 2018*, pp. 81–85, Chengdu, China, March 2018.

[74] Z. Cui, F. Xue, X. Cai, Y. Cao, G. G. Wang, and J. Chen, "Detection of malicious code variants based on deep learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3187–3196, 2018.

[75] F. Ullah, H. Naeem, S. Jabbar et al., "Cyber security threats detection in internet of things using deep learning approach," *IEEE Access*, vol. 7, pp. 124379–124389, 2019.

[76] N. N. Hurrah, S. A. Parah, J. A. Sheikh, F. Al-Turjman, and K. Muhammad, "Secure data transmission framework for confidentiality in IoTs," *Ad Hoc Networks*, vol. 95, p. 101989, 2019.

[77] F. Al-Turjman, H. Zahmatkesh, and R. Shahroze, "An overview of security and privacy in smart cities' IoT communications," *Transactions on Emerging Telecommunications Technologies*, pp. 1–19, article e3677, 2019.

[78] B. D. Deebak and F. Al-Turjman, "A hybrid secure routing and monitoring mechanism in IoT-based wireless sensor networks," *Ad Hoc Networks*, vol. 97, article 102022, 2020.

[79] G. Srivastava, G. Thippa Reddy, N. Deepa, B. Prabadevi, and M. Praveen Kumar Reddy, "An ensemble model for intrusion detection on the Internet of Softwarized Things," in *Adjunct Proceedings of the 2021 International Conference on Distributed Computing and Networking*, pp. 25–30, Nara, Japan, January 2021.

[80] S. Bhattacharya, S. R. K. S, P. K. R. Maddikunta et al., "A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU," *Electronics*, vol. 9, no. 2, p. 219, 2020.

[81] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho, "Deep learning approach for network intrusion detection in software-defined networking," in *2016 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pp. 258–263, Fez, Morocco, 2016, October.

[82] M. Li, T. Zhang, Y. Chen, and A. J. Smola, "Efficient mini-batch training for stochastic optimization," *In proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, 2014, pp. 661–670, New York, USA, 2014.

[83] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and Intelligent Laboratory Systems*, vol. 39, no. 1, pp. 43–62, 1997.

[84] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?," in *In proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 201–208, Sardinia, Italy, March 2010.

[85] X. Tao, D. Kong, Y. Wei, and Y. Wang, "A big network traffic data fusion approach based on fisher and deep auto-encoder," *Information*, vol. 7, no. 2, p. 20, 2016.

[86] W. Hardy, L. Chen, S. Hou, Y. Ye, and X. Li, "DL4MD: a deep learning framework for intelligent malware detection," in *Proceedings of the International Conference on Data Science (ICDATA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).*, p. 61, Las Vegas, USA, 2016.

[87] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2014.

[88] M. Yousefi-Azar, V. Varadharajan, L. Hamey, and U. Tupakula, "Autoencoder-based feature learning for cyber-security applications," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3854–3861, Anchorage, AK, USA, May 2017.

[89] S. Rathore, A. Saxena, and M. Manoria, "Intrusion detection system on KDDCup99 dataset: a survey," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 4, 2015.

[90] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6, Ottawa, ON, Canada, July 2009.

[91] N. Moustafa, G. Creech, and J. Slay, "Big data analytics for intrusion detection system: statistical decision-making using finite Dirichlet mixture models," in *Data Analytics and Decision Support for Cybersecurity*, pp. 127–156, Springer, Cham, Switzerland, 2017.

[92] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, pp. 1–6, Canberra, ACT, Australia, November 2015.

[93] B. M. Aslahi-Shahri, R. Rahmani, M. Chizari et al., "A hybrid method consisting of GA and SVM for intrusion detection system," *Neural Computing and Applications*, vol. 27, no. 6, pp. 1669–1676, 2016.

[94] C. M. Hsu, H. Y. Hsieh, S. W. Prakosa, M. Z. Azhari, and J. S. Leu, "Using long-short-term memory-based convolutional neural networks for network intrusion detection," in *International Wireless Internet Conference*, pp. 86–94, Taipei, Taiwan, 2018.

[95] M. A. Ambusaidi, X. He, P. Nanda, and Z. Tan, "Building an intrusion detection system using a filter-based feature selection algorithm," *IEEE Transactions on Computers*, vol. 65, no. 10, pp. 2986–2998, 2016.

[96] Z. Tan, A. Jamdagni, X. He, P. Nanda, R. P. Liu, and J. Hu, "Detection of denial-of-service attacks based on computer vision techniques," *IEEE Transactions on Computers*, vol. 64, no. 9, pp. 2519–2533, 2014.

[97] C. F. Tsai and C. Y. Lin, "A triangle area based nearest neighbors approach to intrusion detection," *Pattern Recognition*, vol. 43, no. 1, pp. 222–229, 2010.

[98] M. Z. Alom, V. Bontupalli, and T. M. Taha, "Intrusion detection using deep belief networks," in *2015 National Aerospace and Electronics Conference (NAECON)*, pp. 339–344, Dayton, OH, USA, 2015.

[99] S. A. Ludwig, "Intrusion detection of multiple attack classes using a deep neural net ensemble," in *2015 National Aerospace and Electronics Conference (NAECON)*, pp. 1–7, Dayton, OH, USA, November 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9970972, 12 pages
https://doi.org/10.1155/2021/9970972

Research Article

A Time-Overlapping Multiplex VLC System for End-Edge Data Transmission

Tingting Fu,[1] **Huanghong Zhu**,[1] **Han Hai**,[2] **and Haksrun Lao**[3]

[1]*School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China*
[2]*College of Information Science and Technology, Donghua University, Shanghai, China*
[3]*Center of Engineering and Design, Chong Cheng Chinese School, Phnom Penh, Cambodia*

Correspondence should be addressed to Haksrun Lao; haksrunlao@hotmail.com

Received 5 March 2021; Revised 11 June 2021; Accepted 5 July 2021; Published 10 August 2021

Academic Editor: Mohammad R. Khosravi

Communication is one of the most important foundations in the Internet of Things. Although some cutting-edge technologies, such as 5G, have greatly empowered edge computing, electromagnetic interference and pollution make them impracticable in many environments. The visible light communication (VLC) is a new type of wireless communication technology with appealing benefits not presented in radio communications. VLC allows a lamp or other light source to not only serve as illumination but also simultaneously transmit data. Although traditional VLC multiplexing technologies have been able to achieve a high-speed data transmission rate, they require all receivers to use the same modulation means. In many scenarios, various-type receivers coexist; it is costly to incorporate multiple senders to implement adaptive content distribution in on-demand services. In this paper, we propose a new type of VLC multiplexing system, which realizes end-edge data transmission through pulse position modulation (PPM), pulse width modulation (PWM), and pulse amplitude modulation (PAM) simultaneously. Therefore, one edge server can serve multiple types of end-users without interference. In order to evaluate the performance of the system, we conduct experiments with different settings of communication distance, communication angle, and different environmental light conditions. For three modulations, the proposed system can achieve a transmission speed three times as that for a single modulation, and reach the accuracy rate of up to 99% within the proper communication range.

1. Introduction

Along with the incredible growth of mobile devices in the Internet of Things (IoT) and the explosion of demands for energy and resource-hungry applications, such as online shopping [1], video streaming [2], data processing [3], data sharing [4], and space-air-ground communications [5], the multiaccess edge computing demonstrates the possibility to provide infrastructure, platform, and software as a service for end-users from edge servers with a fixed or wireless network connection [6]. In many cases, an edge server may need to serve multiple end-users simultaneously, where congestion and latency could be very high due to the spectrum crunch problem of wireless communications. Only extending spectrum resources cannot solve the end-edge data transmission problem effectively [7]. Therefore, to reduce the wireless transmission latency between end-users and edge servers, future wireless communications with different radio access technologies, transmission backhauls, and network slices are evaluated in the emerging edge computing paradigm [8].

Different from the emerging RF radio communication system which needs complex signal processing and spectrum sensing means [9], visible light communication (VLC) uses low-cost, energy-saving, and efficient light-emitting diode (LED) to encode the data into high-frequency changes of light intensity that cannot be sensed by human eyes [10]. Various optical sensors (photodiode) can demodulate data by monitoring the change of light intensity. In many cases, the off-the-shelf equipment can be used to realize data communication on the basis of lighting [11]. VLC works in the unregulated spectrum range, and the bandwidth is 10^4 times of the RF [12]. Furthermore, it is free of electromagnetic interference and pollution; therefore, it can be applied in the electromagnetism-sensitive environment, such as inside

airplanes, hospitals, and scientific machineries. Another important advantage of VLC is being able to provide better security and privacy in some scenarios since light cannot penetrate walls and requires line-of-sight contact with the receiver [13]. For these reasons, VLC can bring communication capabilities between end-users and edge servers, which is characterized by low latency, low cost, good scalability, and privacy protection.

One-to-many communication instances are very common in the edge computing, such as user-specific task services and personalized information deliveries [14]. In such a scenario, multiplexing, which refers to the use of a transmission medium to achieve multichannel signal communication such as MIMO [15] (multi-input and multioutput), can improve the utilization of the link and achieve Burst-Mode communications [16]. In VLC systems, common multiplexing technologies include WDM (Wavelength Division Multiplexing), SDM (Space Division Multiplexing), TDM (Time Division Multiplexing), and FDM (Frequency Division Multiplexing) [17]. WDM and SDM need to use multiple LEDs as transmitters to form a MIMO system. TDM will bring down the transmission rate under the same hardware conditions. FDM is much complex to achieve, and the hardware equipment is more expensive, which is not suitable for simple text data transmission. The transmission rate of these highly complicated VLC devices can reach Gbps, but so far, none of them has been put into use. Therefore, our goal is to use single LED and existing low-cost hardware to design a simple VLC multiplex data communication system and improve the system transmission capacity, where multiple end-users share a communication link.

In this paper, we propose a multiplex data transmission system based on pulse position modulation (PPM) [18], pulse width modulation (PWM) [19], and pulse amplitude modulation (PAM) [20]. The key idea is to use these three modulation methods to encode three groups of data, respectively, then combine three signals into one signal, and send the data stream through a single LED to three categories of end-users simultaneously. Each end-user uses a different demodulation rule to get its dedicated data information, with time-overlapping. This is very useful in many scenarios, especially in a dynamic environment, where end-users come and go frequently. They may expect for different message contents in terms of the group they belong to. By combining multiple modulation methods into one signal, the sender can broadcast different contents to different end-users at one time, without interference.

For instance, the proposed VLC system can be deployed in a vehicular environment where mobility is the main challenge in vehicular edge computing [21, 22]. The free space in the frequency spectrum for all normal communication specifications is very full. Therefore, using radio wireless communication to transfer information may have interference from other wireless signals, which could lead to a traffic accident for vehicles. As can be seen from Figure 1, three vehicles have overlapped communication range (shown in different colors), which is potential to cause conflicts. If personalized information has to be transmitted, the time needs to be divided into small slots. Therefore, it is important to use new technologies to avoid this happen. Also, the VLC multiplexing system has better and low-cost privacy performance since each group of users uses a different coding scheme. Regarding wireless multiplexing, applying different cipher keys is a common solution. The proposed VLC system is suitable for the problem. Also, it is energy efficient since it is based on LED, and very healthy for humans since the change of lights is undetectable by eyes.

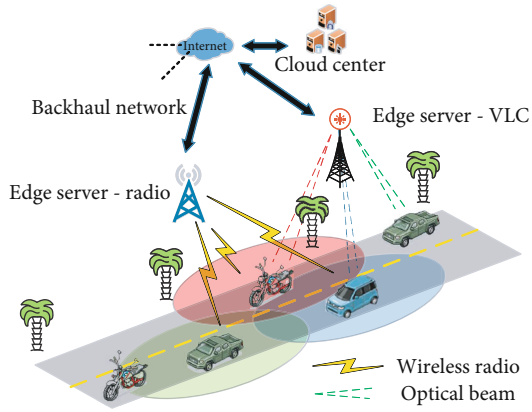


Figure 1: Demonstration of multiplexing in edge computing.

The rest of the paper is organized as follows. We first briefly introduce the latest literatures on data transmission in edge computing and VLC-related research. Then, we draw the big picture of the overall model of the system. After that, the system design, including modulation, demodulation, hardware implementation, and software coding, is described in detail, followed by the experiment results and performance analysis. Finally, we conclude the paper and suggest some open issues for future work.

2. Related Work

2.1. Communication in Edge and Industrial Computing. Communication is the key foundation in Edge and Industrial Computing, through which task offloading and data transmission must go with. To enhance communication capability, MIMO is often used in IIoT [23], where multiple antennas are installed. In [24], QoE (Quality of Experience) of the edge-cloud architecture is enhanced by adopting application-level coding (e.g., transcoding, rate control) to match the estimated capacity at the radio downlink. In the Edge Computing-Internet of Vehicles system, two communication models between edge servers and users are considered, i.e., the single-hop communication and the multihop communication [25]. Tasks from user vehicles may be offloaded to 5G base stations for processing. 5G can provide fast and reliable data transmission between base stations and cloud servers; however, interference management of wireless communications within a base station deserves further investigations, not to mention security communication problem for drones [26]. In another scenario, where service robots are deployed for healthcare, data is collected from people and sent to the cloud. Robots also act as edge servers to alleviate the burden of the cloud as well as reduce latency [27]. The result of edge computing is exchanged through rapid

machine-to-machine communication. Low-cost large-scale communication is also very important for mobile edge computing in the maritime Internet of Things, where the channel allocation and power allocation problems are two major means to optimize the offloading policy [28]. With the rapid growth of federated learning in the end-edge-cloud orchestra, in most cases, it may suffer from a large number of rounds to convergence, which leads to high communication costs. Some work [29] proposes novel compression techniques called FedAvg to produce communication-efficient.

2.2. VLC System and Multiplexing. The early research of VLC was initiated by a team led by Nakagawa of Japan. They envisioned the combination of VLC and PLC to provide indoor network communication [30] and firstly studied the environmental characteristics of VLC indoor propagation [10]. Since then, increasing communication rate has become a major research direction of VLC, including filtering yellow light [31], designing complex modulation schemes (for example, CSK [32]), and using multiple inputs and multiple outputs to realize parallel data stream transmission [33]. At present, the fastest VLC communication system can reach a speed of 10 Gbps.

Professor Chi Nan of the Fudan University introduced the use of several traditional multiplexing technologies in VLC and verified the feasibility of multiplexing technology in improving the transmission capacity of the VLC system [12]. The multiplexing technologies that can be used include wavelength division multiplexing, space division multiplexing, polarization multiplexing, and frequency division multiplexing [12]. Wavelength division multiplexing (WDM) refers to using visible light of different wavelengths as carrier waves to modulate signals. For RGB-LED, red, green, and blue visible light of different wavelengths can be used to modulate different signals, respectively. Space division multiplexing uses multiple transmitters to send data and multiple receivers to receive data at the same time, to realize parallel transmission of space multiplexing. Polarization multiplexing is to modulate the signal to the linearly polarized light in different directions by using the visible polarizer and to carry out multichannel parallel transmission. Frequency division multiplexing (FDM) is to realize multiclient parallel transmission in frequency by using subcarrier modulation signals with different center frequencies of LED. The transmission rate of these traditional multiplexing technologies can be up to gigabits per second.

At present, there are a bunch of researches on VLC modulation methods. We can divide it into two categories: (1) single carrier modulation mechanism [34], such as pulse width modulation (PWM), pulse amplitude modulation (PAM), and pulse position modulation (PPM); (2) multicarrier modulation mechanism [35], which requires more complex hardware support, such as orthogonal frequency division multiplexing (OFDM). Different from existing research, we developed a new VLC multiplexing system for low data rate and cost-sensitive applications.

3. System Model

The overall diagram of the core time-overlapping multiplex VLC system is shown in Figure 2. For different applications, LEDs can vary in power, and the driver circuit should be designed correspondingly. The proposed communication system can support up to three categories of end-users in the same direction. Each category can have multiple end-users. In the figure, each category is represented by one end-user. The edge server is used to transmit the data stream to the agent microcontroller. The microcontroller modulates the data going to be sent, uses the modulated signal to control the MOSFET, and loads it on the LED light source. Under the control of the MOSFET, the LED lights up quickly, and finally, the visible light is carried in the communication channel with data dissemination. At the receiving end, the rapid intensity change of LED will be captured by photodiodes. Each photodiode converts the optical signal containing information into an electrical signal and then uses the microcontroller connected with the photodiode to sample the electrical signal generated by the photodiode. Finally, each end-user will process the sampling data according to its dedicated demodulation rule, restore the original information as it been sent.

4. System Design

The VLC multiplexing data transmission system encodes the data of different categories and then sends the data to each end-user simultaneously through the modulation mode of PPM, PWM, and PAM. Each end-user uses different demodulation methods to obtain their data. Each end-user does not interfere with another end-user, nor affects each other, and is not able to extract data belonging to other end-users.

4.1. Channel Model. Let us assume that the proposed visible light communication system is composed of LED transmitters and end-users. Then, the received signal can be considered as in Equation (1) [36]:

$$y(t) = \sqrt{\rho} H(t) \otimes s(t) + n(t), \tag{1}$$

where $\rho = (r^2 \bar{P}_K^2)/\sigma^2$ denotes the average electrical SNR at each receive unit, $\otimes$ is time convolution, $\bar{P}_K^2 = (1/K)\Sigma_{i=1}^{K} P_K^{(i)}$ represents the average received optical power, and r denotes the photodiode responsivity. The vector $n(t)$ indicates K -dimensional noise. The noise includes the receiver thermal noise and shot noise due to ambient light. Thus, $n(t)$ can be modeled as independent and identically distributed additive white Gaussian noise with power spectral density $\sigma^2 = \sigma_{\text{shot}}^2 + \sigma_{\text{thermal}}^2$, where σ_{shot}^2 is the shot noise variance and $\sigma_{\text{thermal}}^2$ is the thermal noise variance [37].

For VLC systems, the direct line-of-sight (LOS) and the nondirect line-of-sight (NLOS) are the two general models. In this paper, we only consider the LOS propagation path since LOS accounts for the most total received optical power at the receiver. Let $\mathbf{H}_k \in \mathbb{R}^{1 \times N_T}$ denotes the channel matrix between transmitter and receiver (as shown in Equation (2)):

$$\mathbf{H}_k = \left[h_{k1}, h_{k2}, \cdots, h_{kN_T}\right], \tag{2}$$

where h_{ki} represents the direct current gain between the k-th

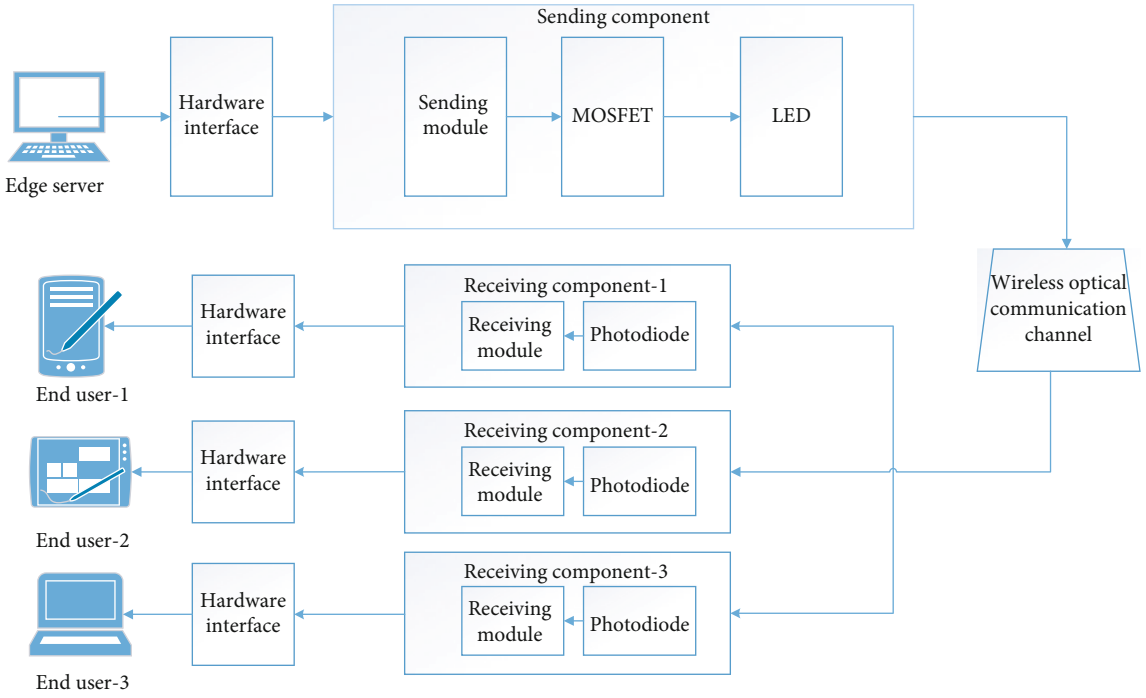


Figure 2: Core model of the proposed time-overlapping multiplex VLC system.

end-user and the i-th transmitter. In terms of the LOS case, h_{ki} is given by [10] as shown in Equation (3):

$$h_{ki} = \begin{cases} \frac{A_r}{d_{ki}^2} L(\phi) T_s(\psi_{ki}) g(\psi_{ki}) \cos(\psi_{ki}), & 0 \le \psi_{ki} \le \Psi_c, \\ 0, & \psi_{ki} > \Psi_c. \end{cases} \tag{3}$$

where A_r is the active area of the photodiode and d_{ki} is the distance from the LED to the photodiode. $T_s(\psi_{ki})$ is the gain of the optical filter and ψ_{ki} denotes the angle of incidence. The optical field of the view of the photodiode can be denoted by Ψ_c. Finally, $g(\psi_{ki})$ denotes the optical concentrator gain.

4.2. Modulation. The PPM is based on the position of the pulse. PPM is to divide the time of a cycle into several time slots of equal intervals, and the transmission of the pulse is done in any one of the slots. According to the corresponding relationship between data and pulse position, the sender chooses to send a pulse signal in a certain time slot to realize data transmission. The design principle of PPM is shown in Figure 3(a). A cycle is divided into two time slots. The first slot has a pulse signal indicating data bit "1," and the second slot has a pulse signal indicating bit "0."

The PWM uses pulse duration to control the LED drive current; thus, it can adjust the brightness. The advantage of the PWM is that it does not suffer from the wavelength shift due to the current variation in the intensity or amplitude modulation-based scheme. In PWM, brightness level with a wide range (0-100%) can be achieved by directly adjusting the modulation index. Moreover, the human eyes cannot sense the current switching since the dimming signal frequency is usually above 100 Hz. PWM uses different widths of the pulse signal to realize data modulation. The design principle of PWM is shown in Figure 3(b). Different width of the pulse signal in a cycle time represents bit "1" and bit "0."

PAM refers to a modulation mode in which the pulse height changes with the encoding. It is a bandwidth efficient scheme since it can improve spectral efficiency. Data is modulated into different amplitudes of the signal pulse. PAM may suffer nonlinearity in LEDs' luminous efficacy due to the modulation schemes employing different intensity levels. Since the light emitted by an LED depends on the input current and temperature, it changes at multiple symbol levels of the PAM along with changes in the drive current. The design principle of PAM is shown in Figure 3(c). The different heights of the pulse signal in a cycle time are used to represent bit "1" and bit "0."

The visible light communication system adopts intensity modulation and direct detection mechanism. The power of LED can be expressed as Equation (4):

$$P_{\text{led}} = V_{\text{led}} * I_{\text{led}}. \tag{4}$$

where V_{led} is the voltage of LED light source and I_{led} is the current of LED light source. The peak intensity of LED light source is mainly affected by the LED power. Therefore, by using different input voltage, it can generate pulse signals of different heights and, finally, act on the LED to change the peak intensity of the LED light source.

In this paper, we propose to use PPM, PWM, and PAM modulation technologies to implement the VLC multiplexing data transmission system for end-edge data transmission. By modulating the pulse position, pulse width, and pulse amplitude simultaneously, we can send data to three categories of

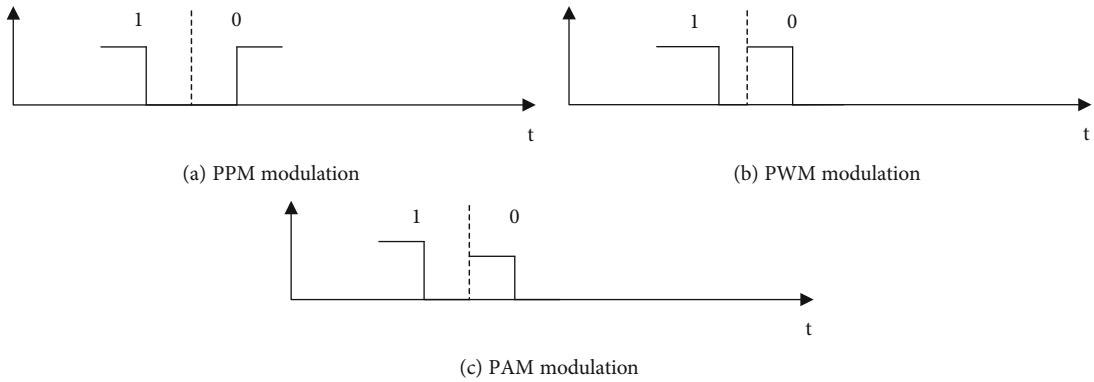


Figure 3: PPM, PWM, and PAM modulation.

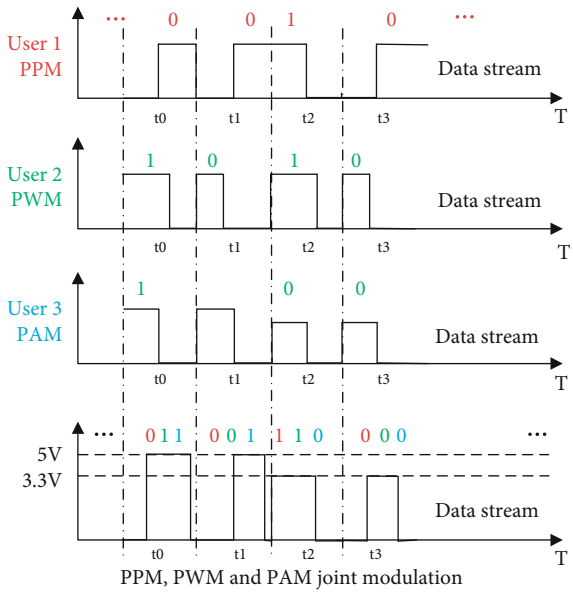


Figure 4: PPM, PWM, and PAM joint modulation signal diagram.

end-users at the same time, for instance, end-user 1 with PPM, end-user 2 with PWM, and end-user 3 with PAM. Suppose that data stream "0010" is going to be sent to the end-user 1, data stream "1010" is going to be sent to the end-user 2, and data stream "1100" is going to be sent to the end-user 3. The modulation design is shown in Figure 4; the server first encodes data into light pulse modulations for each end-user and then integrates them into one signal with changing position, width, and height. The first bits of three users are, respectively "0," "1," and "1." Therefore, the corresponding joint modulation signal starts with a low pulse and ends with a high pulse in the time slot, and the high pulse has a larger width and a higher amplitude. Each end-user can extract the exclusive data from the same signal waveform.

However, in some hybrid system where both 5 V and 3.3 V power are mixed, if a previous data starts with a low pulse, ends with a high pulse of 5 V power supply, and the following data starts with a high pulse of 3.3 V power supply, ends with a low pulse, there will be a direct conversion from 5 V to 3.3 V between two data. The problem is before the 5 V power supply has been completely turned off, the 3.3 V power supply has been turned on, so chaotic data will be generated, increasing the difficulty of decoding. Our solution is to check whether a signal starts with a low pulse and ends with a high pulse, immediately add a low pulse delay after the high pulse to ensure that the power supply is completely shut down.

4.3. Demodulation. The receiving end-user continuously senses the intensity of the incident light through the photodiode, processes the collected intensity value, and restores the original data. For each end-user, first of all, we need to locate the rising edge and the falling edge of each periodic optical pulse [13]. We can find the local maximum value to locate the edge of the optical pulse by calculating the first derivative $I'(x)$ of the intensity value as shown in Equation (5):

$$I'(x) = -\frac{1}{2} \cdot I(x-1) + 0 \cdot I(x) + \frac{1}{2} \cdot I(x+1), \quad (5)$$

where $I(x)$ is the intensity perceived by the photodiode.

Each end-user applies different demodulation methods. For end-user 1, the data is decoded by the PPM principle to find the relative positions of high pulse and low pulse in the cycle, and the original data information is restored according to different positions of the pulse. For end-user 2, data is decoded by the PWM principle, the high pulse width is calculated by a rising edge and a falling edge of the pulse, and original data information is restored according to different width values. For end-user 3, we decode the data by the PAM principle, find the maximum light intensity detected in each cycle, and compare it with the peak intensity of LED when connecting to 5 V power supply and 3.3 V power supply, respectively, and finally, restore the original data information. However, the peak intensity of LED decays with the square of the distance between the LED and the photodiode, and the different distance between the LED and the photodiode will produce a difference in the peak intensity detected by the photodiode. Therefore, our solution is to add two bits of light intensity calibration before sending each data packet, where the first bit uses a 5 V power supply, and the second one uses a 3.3 V power supply (as shown in Figure 5). Finally, when restoring the data information, the peak intensity of the data bits only needs to be compared with that of the first two calibration bits.

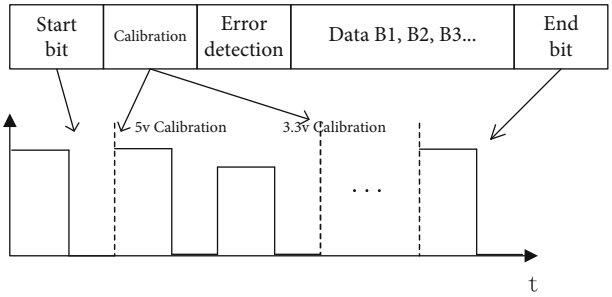


Figure 5: Design of calibration in PAM model.

In many practical data transmission processes, the length of data stream to be sent to each end-user is different. A binary bit with no meaning could be recognized as data for end-users improperly. Therefore, we add a stop bit at the end of each data stream, according to which the end-user judge whether to continue to accept the data. Bit "1" means continuing to receive and bit "0" means stopping receiving.

4.4. Error Detection and Correction. In this paper, we build an Arduino UNO-based experiment platform as the validation system, due to the limited sampling rate; the pulse width corresponding to a bit may be inaccurate during decoding, resulting in decoding error of that bit in the process of data transmission. Moreover, consider that many VLC systems only need a one-way transmission, where it is not possible to get the receiver's feedback information; hence, we use FEC (Forward Error Correction) to detect and correct errors to achieve reliable transmission. Hamming code is a type of linear block code, which has been widely used in the telecommunication system. It can correct random errors and burst errors. In our design, Hamming code is adopted to add redundant bits in data to realize error detection and correction. It organizes data bits into groups, determines whether there are any errors in the group through parity check, in terms of the principle of odd or even matching. Concretely, the total length of the data frame is 14 bits, which includes 2 calibration bits, 8 data bits, and 4 parity check bits. Hamming code can only recover 1 bit error. If more than 1 data bits flip during the transmission, Hamming code will not work. However, during the experiment, we find that very few bit errors occur and the vast majority of them is 1 bit error, providing that the communication range is within the proper distance. The reason we chose Hamming code is that it is very simple and easy to implement. It also puts little cost over transmission.

4.5. Design of the Validation System. The evaluation system includes a sending module and a receiving module. The sending module mainly includes one microcontroller, one MOSFET Driver Module, and one LED light source. The receiving end mainly includes one microcontroller, one resistor, and one photodiode. The specifications of the LED light source and the photodiode used in the experiment are shown in Table 1.

On the transmitter side, the microcontroller is Arduino UNO [38], with a clock frequency of 16 MHz, which can generate microsecond level pulses. Two pins (pin 9 and pin 10) of the microcontroller are connected with two p-type MOSFET driver modules. The two p-type MOSFET modules are, respectively, connected with 3.3 V and 5 V voltage, and one LED light source. Through controlling the two MOSFET drivers, the LED can be quickly switched on and off. The control circuit of the transmitter is shown in Figure 6.

Table 1: Component parameters.

Component	Parameter
Parameters of LED	Voltage range: 3.3-5 V
	Power: 5 W
	Angle of view: 90°
	Intensity: 200-300 lx
	Wavelength: 300-760 nm
Parameters of Photodiode	Photocurrent: 10 μA
	Rise/fall time: 12 ns
	Peak wavelength: 850 nm

The receiver also uses Arduino UNO as the microcontroller to connect the photodiode at pin A0 and sample the data of the photodiode with a sampling frequency of 50 kHz. It can map the voltage collected from the photodiode to a value from 0 until 1023. A 1 mΩ resistor is connected at the receiving end to improve the gain of the photodiode. The control circuit of the receiving end is shown in Figure 7.

5. Results and Analysis

We test the proposed time-overlapping multiplex VLC system from two aspects: (1) system function test: whether the transmitter (edge server) and the receiver (end-user) can reliably conduct one to three communications; (2) system performance test: test the speed and accuracy of the system under different link distance, different perspective, and different ambient light conditions. Due to a limited budget, we only build a minimal system that can fulfill the evaluation purpose. The prototype of the experiment system is shown in Figure 8. One LED controlled by the Arduino board acts as the edge server (Figure 8(a)) while three Arduino boards with photodiodes are end-users (each of which uses PPM, PWM, and PAM, respectively, Figure 8(b)). By default, the experiment is carried out indoor and under the condition of turning on the fluorescent lamp, the photodiode is directly below the LED, with a height of 13 cm. The period of transmitting signal is set as 220 μs, and the pulse width modulated high pulse width is 140 μs and 100 μs, respectively.

5.1. System Function Experiment. The system function test experiment is carried out by sending text data to receivers. The text data needed to be sent to end-users is input from the console. In the test, we arbitrarily select the characters "a," "6," and "$*$," corresponding to end-user 1, end-user 2, and end-user 3, respectively. The received data printed on the end-user console validates that the letter, the number, and the symbol can be correctly transmitted.

In order to ensure that the system can work normally, we draw the pulse shape diagram of the transmitter and the receiver (as shown in Figure 9). The pulse shape diagram

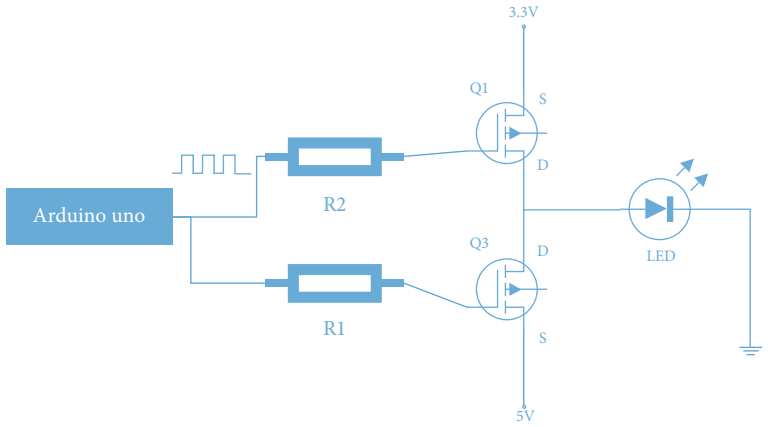


Figure 6: Hardware design of the transmitter.

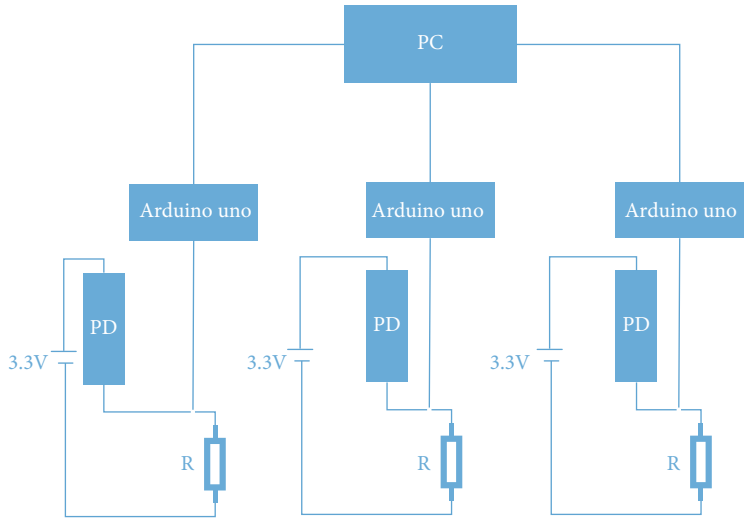
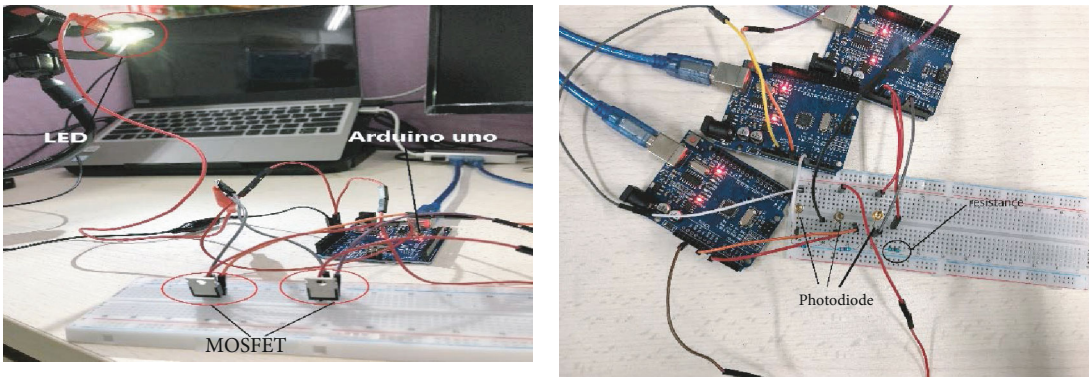


Figure 7: Hardware design of receivers.



(a) Transmitter subsystem (b) Receiver subsystem

Figure 8: Experiment system implementation.

shows that the receiver can correctly receive the signal from the transmitter.

5.2. System Performance Experiment. By testing the throughput and accuracy of the VLC multiplexing system, we investigate the influence of communication channel parameters and environmental factors on the system performance. Throughput is the number of bits received correctly per second, and accuracy is the ratio of the number of bits received correctly to all bits sent. In each experiment, the sender sends

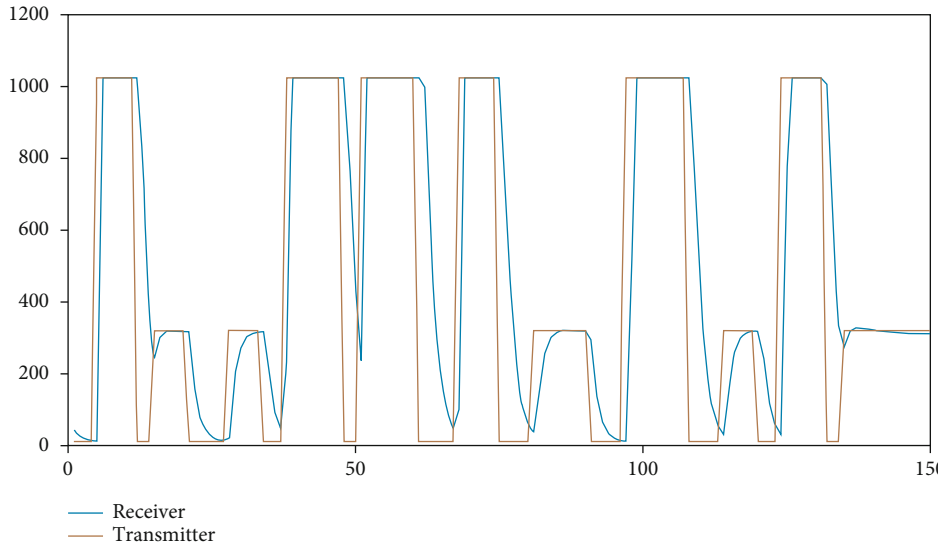
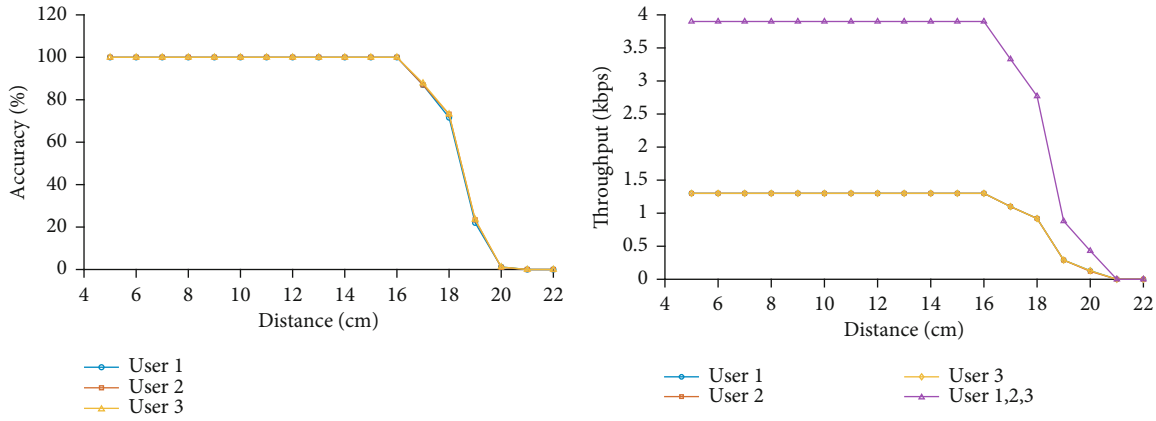


Figure 9: Pulse signal diagram of the transmitter and receiver.



(a) Influence of communication distance on system accuracy

(b) Influence of communication distance on throughput

Figure 10: Influence of communication distance on the system.

512 random bits of data to each end-user at the same time. We repeat the experiment ten times to calculate the average throughput and accuracy.

5.2.1. The Influence of Link Distance on the System. The link distance between LED and photodiode is one of the important parameters that affect the accuracy of data transmission. We adjust the distance from the LED to the photodiode from 5 cm to 22 cm and keep the LED directly below the photodiode. The experimental results are shown in Figures 10(a) and 10(b). We can see that the system can support a 16 cm communication distance, and the throughput of each end-user is about 1.29 Kbps; the total throughput is 3.87 kbps. As long as the distance between the sender and the receiver is within the communication range, the data can be transmitted reliably. When the communication distance is more than 20 cm, its data transmission accuracy becomes 0%. We can see that no matter what kind of modulation means are used, the curves are highly coincident. This is because once the communication range is over length, the current generated by the photodiode induced light intensity changes very little, which makes Arduino unable to sample it.

Due to the low power of the LED (5 W) in the validation system, the valid distance is around 16 cm. The power used in related work is around 20 W to 30 W. Therefore, the effective distance can be extended by applying a LED with larger power. Under the same condition, it is expected that our multiplexing system can achieve the same distance performance as in the related work (up to 100 meters). Furthermore, the distance can be extended by adding an optical lens at the receiving end. Finally, increasing the frequency of the microcontroller (in our system, Arduino UNO is a low-end microcontroller) can directly improve the data rate to a much higher level (15 Gbps in a report).

5.2.2. The Influence of Different Ambient Light Conditions on the System. To test the robustness of the system, we carry out experiments at 2 : 00 pm and 7 : 00 pm, respectively, under different ambient light conditions and turn on the fluorescent lamp in both cases. The experimental results of end-user 1

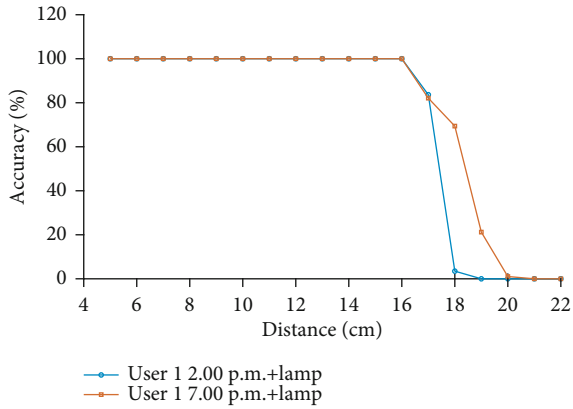


Figure 11: Influence of ambient light conditions on the system.

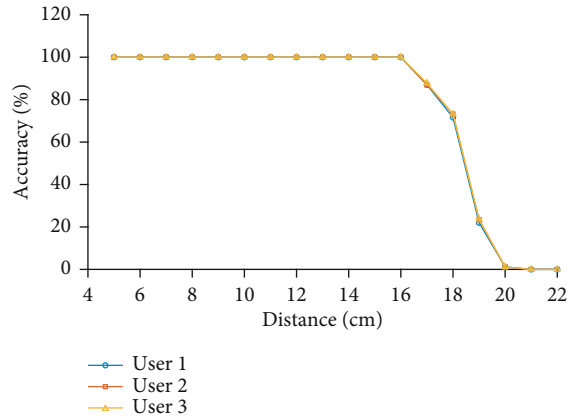


Figure 12: Influence of different angle of view on the system.

under different ambient light conditions are shown in Figure 11. We can see that at two o'clock in the afternoon, when the ambient illumination is high, the maximum supported communication distance of the system is 16 cm, which is the same as the experiment at 7 : 00 pm. When the communication distance is more than 18 cm, the effect of the ambient light starts to appear; however, the result got at 2 : 00 pm has little improvement. Therefore, we can say that the system has a strong tolerance to ambient light.

5.2.3. The Influence of Angle of View on the System. The angle of view is also another important parameter to measure the system. It refers to the angle between the LED and photodiode line and the LED normal. We adjust the viewing angle from 0° to and 35°. The experiment is carried out at 7 : 00 pm and the distance between LED and photodiode is 13 cm. The experimental results are shown in Figure 12. We can see that the angle of view supported by the system is about 40° (±20°).

5.2.4. The Performance Comparison with PWM Multiplexing. As most of recent work is moved to use multiple-input multiple-output (MIMO) hybrid multiplexing in VLC due to achievable very high data rate, there are few researches insisting on using single LED. However, not all applications need such high data rate. They may concern more on deploying cost, where multiplexing with a single LED is very useful. To our best knowledge, there is no such work combining PPM, PWM, and PAM as we do, therefore, we compare our method with a traditional multiplexing method only using pulse width modulation (PWM).

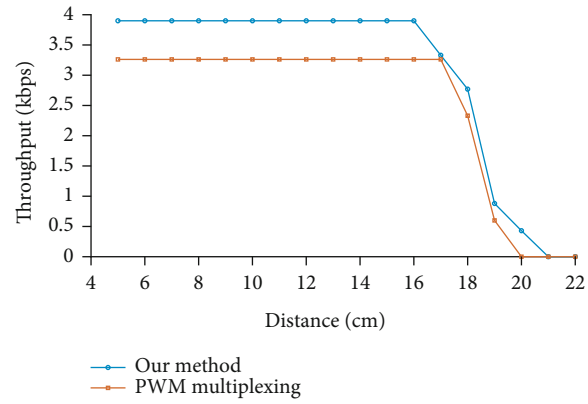


Figure 13: The performance comparison with PWM multiplexing.

As shown in Figure 13, the total throughput of PWM multiplexing with three users is around 3.27 Kbps, which is only 84% of the throughput of our method. Furthermore, with the increase of the number of users, the number of pulse width will also increase in an exponential manner, which will greatly lower per user throughput.

5.2.5. The Test on Vast Transmission. Since large data files, such as images and videos, are very common in data transmission, to test the robustness of the proposed system, we have carried out image transmission experiments. We choose the most famous image "Lena" as the test object. First, we convert the standard ".tif" file into ".eps" format, which size is around 1560 KB. Second, we send the image using PPM, PWM, and PAM modulation simultaneously. In this step, we read the eps file byte by byte and apply Hamming code before transmission. The actual size of each image transmitted is 2730 KB. Third, we analyze the difference of the original image and the received image. Since the data rate is about 1.29 Kbps, it costs around 2100 seconds (35 minutes) to finish one transmission. We repeat the experiment for 100 times and find that, in 93% of the cases, we get the equal images. In the remaining 7 cases, the differences between two images are less than 2 bytes in 5 results. Most of bit errors have already been corrected by Hamming code. As shown in Figure 14, (a) is the original image and (b) is the received image. The differences between them are 41 bytes (possessing 0.00256% of total bits), which is the worst case in the 100 time experiments. However, the human eyes still cannot tell the difference between them.

Therefore, the proposed system can handle most of practical application scenarios with reasonable cost. However, if the hard robustness data transmission needs to be guaranteed, either a two-way communication means or much complex Forward Error Correction Encoding must be adopted, which violates the design purpose of this system.

(a) Lena-original (b) Lena-received

Figure 14: Image transmission result.

5.3. Potential Applications. Besides end-edge data transmission, the proposed system can be applied in many IoT applications, such as underwater communication, medical treatment, and indoor positioning.

Because of the high attenuation rate of electromagnetic wave, it is difficult to use radio waves for underwater wireless communication. Compared with traditional underwater sonar communication, visible light optical communication has higher directivity and confidentiality. Taiyo Yuden and Toyo Electric Co jointly developed a high-speed underwater wireless communication device based on visible light. Using the general blue LED with low attenuation rate in water, the maximum communication speed of 50 Mbps is achieved.

For hospital and medical field, visible light communication is considered to be the most suitable wireless communication mode since it does not cause electromagnetic interference to electronic equipment. Whether it is to realize the dynamic monitoring of patients or the communication between medical devices, VLC provides a safe and reliable way of information transmission. The hospital is likely to be the first large-scale popularization of VLC system.

The proposed VLC system uses LED as transmitter, which are currently being installed in most buildings, for instance, large shopping malls, office, and classroom. Since the global positioning system (GPS) signals are difficult to pass through building walls, the VLC-based indoor positioning system can be a good substitute. It estimates location by the geometric properties of triangles. Specifically, this technique uses difference reference points to get target position, where the reference points are LEDs, and target is the optical receiver. Comparing to radio frequency-based indoor positioning system, VLC is electromagnetic interference free and can achieve a good accuracy.

6. Conclusion

In this paper, we proposed a new VLC system to use PPM, PWM, and PAM modulation technology to realize the VLC multiplexing communication system for end-edge data transmission. We designed the modulation methods and solved the problem of the nonfixed length of individual information. Also, a delay was added to fit the mixed voltage system. We used a low-cost hardware solution, in which an off-the-shelf LED and photodiode are adopted to build the preliminary validation prototype. The proposed VLC system successfully completed a one-to-three communication. The communication rate of a single client can reach 1.29 Kbps, the communication distance can reach 16 cm, and the accuracy is near 99%.

Future research issues include further improving communication range and data rate and efficient two-ways VLC communications. With the more powerful LEDs and microcontrollers, the range and the rate can be easily extended to a level that has the potential to be used in many scenarios, such as underwater communication, medical treatment, and indoor positioning to avoid radio communication interference and also improve safety.

Data Availability

Data is available on request; please contact the corresponding author Haksrun Lao (haksrunlao@hotmail.com).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

This is a joint work with Mr. Huanghong Zhu. Prof. Tingting Fu was the supervisor of Mr. Zhu and Mr. Lao when Haksrun Lao was an undergraduate student at Hangzhou Dianzi University. Dr. Hai contributes in improving and finalizing the system model.

Acknowledgments

This work was in part supported by the Chong Cheng Chinese School, Phnom Penh, Cambodia.

References

[1] C. Huang, W. Jiang, J. Wu, and G. Wang, "Personalized review recommendation based on users' aspect sentiment," *ACM Transactions on Internet Technology*, vol. 20, no. 4, pp. 1–26, 2020.

[2] C. Chen, B. Liu, S. Wan, P. Qiao, and Q. Pei, "An edge traffic flow detection scheme based on deep learning in an intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1840–1852, 2020.

[3] T. Yang, H. Feng, C. Yang, Y. Wang, J. Dong, and M. Xia, "Multivessel computation offloading in maritime mobile edge computing network," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4063–4073, 2019.

[4] K.-P. Yu, L. Tan, M. Aloqaily, H. Yang, and Y. Jararweh, "Blockchain-enhanced data sharing with traceable and direct revocation in iiot," *IEEE Transactions on Industrial Informatics*, p. 1, 2021.

[5] S. Wan, J. Hu, C. Chen, A. Jolfaei, S. Mumtaz, and Q. Pei, "Fair-hierarchical scheduling for diversified services in space, air and ground for 6g-dense internet of things," *IEEE Transactions on Network Science and Engineering*, 2020.

[6] S. Shahzadi, M. Iqbal, T. Dagiuklas, and Z. Ul-Qayyum, "Multi-access edge computing: open issues, challenges and future perspectives," *Journal of Cloud Computing*, vol. 6, no. 1, p. 30, 2017.

[7] A. Zhou, S. Wang, S. Wan, and L. Qi, "Lmm: latency-aware micro-service mashup in mobile edge computing environment," *Neural Computing and Applications*, vol. 32, no. 19, pp. 15411–15425, 2020.

[8] H. Wu, X. Li, and Y. Deng, "Deep learning-driven wireless communication for edge-cloud computing: opportunities and challenges," *Journal of Cloud Computing*, vol. 9, no. 1, p. 21, 2020.

[9] M. Raja, "Application of cognitive radio and interference cancellation in the l-band based on future air-to-ground communication systems," *Digital Communications and Networks*, vol. 5, no. 2, pp. 111–120, 2019.

[10] T. Komine and M. Nakagawa, "Fundamental analysis for visible-light communication system using led lights," *IEEE transactions on Consumer Electronics*, vol. 50, no. 1, pp. 100–107, 2004.

[11] L. U. Khan, "Visible light communication: applications, architecture, standardization and research challenges," *Digital Communications and Networks*, vol. 3, no. 2, pp. 78–88, 2017.

[12] C. Nan, W. Yiguang, and W. Yuanyuan, "Research on multidimensional multiplexing technology of visible light communication," *Optics and Optoelecronics Technology*, vol. 12, no. 4, pp. 1–6, 2014.

[13] T. Zhao, K. Wright, and X. Zhou, "Lighting up the internet of things with darkvlc," in *Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications, HotMobile 2016*, pp. 33–38, St. Augustine, FL, USA, 2016.

[14] S. Safavat, N. N. Sapavath, and D. B. Rawat, "Recent advances in mobile edge computing and content caching," *Digital Communications and Networks*, vol. 6, no. 2, pp. 189–194, 2020.

[15] C. Qing, B. Cai, Q. Yang, J. Wang, and C. Huang, "Elm-based superimposed csi feedback for fdd massive mimo system," *IEEE Access*, vol. 8, pp. 53408–53418, 2020.

[16] C. Qing, W. Yu, B. Cai, J. Wang, and C. Huang, "Elm-based frame synchronization in burst-mode communication systems with nonlinear distortion," *IEEE Wireless Communications Letters*, vol. 9, no. 6, pp. 915–919, 2020.

[17] A. Jovicic, J. Li, and T. Richardson, "Visible light communication: opportunities, challenges and the path to market," *IEEE Communications Magazine*, vol. 51, no. 12, pp. 26–32, 2013.

[18] C. Nan, *Led Visible Light Communication Technology*, Tsinghua University Press, Beijing, 2014.

[19] A. Pradana, N. Ahmadi, and T. Adiono, "Design and implementation of visible light communication system using pulse width modulation," in *Proceedings of the International Conference on Electrical Engineering and Informatics*, pp. 25–30, Denpasar, Indonesia, 2015.

[20] C. Yang, W. Liu, X. Li, Q. Yang, and Z. He, "Nyquist-pam-4 transmission using linear dpd and mlse for indoor visible light communications," in *Proceedings of the IEEE/CIC International Conference on Communications in China*, pp. 22–24, Qingdao, China, 2017.

[21] S. Raza, W. Liu, M. Ahmed et al., "An efficient task offloading scheme in vehicular edge computing," *Journal of Cloud Computing*, vol. 9, no. 1, p. 28, 2020.

[22] C. Chen, Y. Ding, S. Guo, and Y. Wang, "Davt: an error-bounded vehicle trajectory data representation and compression framework," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 10606–10618, 2020.

[23] Y. Gong, L. Zhang, R. P. Liu, K. Yu, and G. Srivastava, "Nonlinear mimo for industrial internet of things in cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5533–5541, 2020.

[24] S. Chen, Y. Wang, and M. Pedram, "A semi-markovian decision process based control method for offloading tasks from mobile devices to the cloud," *GLOBECOM*, pp. 2885–2890, 2013.

[25] S. Wan, R. Gu, T. Umer, K. Salah, and X. Xu, "Toward offloading internet of vehicles applications in 5g networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 1–9, 2020.

[26] C. Feng, K. Yu, A. K. Bashir et al., "Efficient and secure data sharing for 5g flying drones: a blockchain-enabled approach," *IEEE Network*, vol. 35, no. 1, pp. 130–137, 2021.

[27] S. Wan, Z. Gu, and Q. Ni, "Cognitive computing and wireless communications on the edge for healthcare service robots," *Computer Communications*, vol. 149, pp. 99–106, 2020.

[28] T. Yang, H. Feng, S. Gao et al., "Two-stage offloading optimization for energy-latency tradeoff with mobile edge computing in maritime Internet of Things," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 5954–5963, 2020.

[29] J. Mills, J. Hu, and G. Min, "Communication-efficient federated learning for wireless edge intelligence in IoT," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 5986–5994, 2020.

[30] T. Komine and M. Nakagawa, "Integrated system of white led visible-light communication and power-line communication," *IEEE transactions on Consumer Electronics*, vol. 49, no. 1, pp. 71–79, 2003.

[31] H. Le Minh, D. O'Brien, G. Faulkner et al., "100-Mb/s NRZ visible light communications using a postequalized white led," *Photonics Technology Letters*, vol. 15, no. 1, pp. 1063–1065, 2009.

[32] S. Rajagopal, R. D. Roberts, and S.-K. Lim, "Ieee 802.15.7 visible light communication: modulation schemes and dimming support," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 72–82, 2012.

[33] L. Zeng, D. C. O'Brien, H. L. Minh et al., "High data rate multiple input multiple output (mimo) optical wireless communications using white led lighting," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 9, pp. 1654–1662, 2009.

[34] Q. Wang, D. Giustiniano, and D. Puccinelli, "Openvlc: software-defined visible light embedded networks," in *Proceedings of the 1st ACM MobiCom Workshop on Visible Light Communication Systems, VLCS@MobiCom 2014*, pp. 15–20, Maui, Hawaii, USA, 2014.

[35] S. Dimitrov and H. Haas, *Principles of LED Light Communications: Towards Networked Li-Fi*, Cambridge University Press, Cambridge, 2015.

[36] T. V. Pham, H. L. Minh, and A. T. Pham, "Multi-user visible light communication broadcast channels with zero-forcing precoding," *IEEE Transactions on Communications*, vol. 65, no. 6, pp. 2509–2521, 2017.

[37] J. M. Kahn and J. R. Barry, "Wireless infrared communications," *Proceedings of the IEEE*, vol. 85, no. 2, pp. 265–298, 1997.

[38] S. Das, A. Chakraborty, D. Chakraborty, and S. Moshat, "PC to PC data transmission using visible light communication," in *Proceedings of the International Conference on Computer Communication and Informatics Coimbatore*, pp. 22–27, Coimbatore, India, 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3279059, 9 pages
https://doi.org/10.1155/2021/3279059

Research Article

Image-Based Indoor Localization Using Smartphone Camera

Shuang Li,[1,2] Baoguo Yu,[1] Yi Jin,[3] Lu Huang,[1,2] Heng Zhang,[1,2] and Xiaohu Liang[1,2]

[1]*State Key Laboratory of Satellite Navigation System and Equipment Technology, China*
[2]*Southeast University, China*
[3]*Beijing Jiaotong University, China*

Correspondence should be addressed to Yi Jin; yjin@bjtu.edu.cn

Received 17 April 2021; Revised 30 May 2021; Accepted 20 June 2021; Published 5 July 2021

Academic Editor: Mohammad R. Khosravi

With the increasing demand for location-based services such as railway stations, airports, and shopping malls, indoor positioning technology has become one of the most attractive research areas. Due to the effects of multipath propagation, wireless-based indoor localization methods such as WiFi, bluetooth, and pseudolite have difficulty achieving high precision position. In this work, we present an image-based localization approach which can get the position just by taking a picture of the surrounding environment. This paper proposes a novel approach which classifies different scenes based on deep belief networks and solves the camera position with several spatial reference points extracted from depth images by the perspective-*n*-point algorithm. To evaluate the performance, experiments are conducted on public data and real scenes; the result demonstrates that our approach can achieve submeter positioning accuracy. Compared with other methods, image-based indoor localization methods do not require infrastructure and have a wide range of applications that include self-driving, robot navigation, and augmented reality.

1. Introduction

According to statistics, more than 80 percent of people's living time is in an indoor environment such as shopping malls, airports, libraries, campuses, and hospitals. The purpose of the indoor localization system is to provide accurate positions in large buildings. It is vital to applications such as evacuation of trapped people at fire scenes, tracking of valuable assets, and indoor service robot. For these applications to be widely accepted, indoor localization requires an accurate and reliable position estimation scheme [1].

In order to provide a stable indoor location service, a large number of technologies are researched including pseudolite, bluetooth, ultrasonic, WiFi, ultra wideband, and LED [2, 3]. It is almost impossible to obtain very accurate results for a radio-based approach in view of the multipath interference through arrival time and arrival angle methods. The time-varying indoor environment and the movement of pedestrians also have adverse effects on the stability of fingerprint information [4–6]. In addition, the high cost of hardware equipment, construction, and installation as well as maintenance and update is also an important factor limiting the development of indoor positioning technology. Besides, these kinds of methods can only output the position (X, Y, and Z coordinates) but not the view angle (pitch, yaw, and roll angles).

The vision-based positioning method is a kind of passive positioning technology which can achieve high positioning accuracy and does not need extra infrastructure. Moreover, it can not only output the position but also the view angle at the same time. Therefore, it has gradually become a hotspot of indoor positioning technology [7, 8]. Such methods typically involve four steps: first, establishing an indoor image dataset collected by depth cameras with exact positional information; second, comparing the images collected by a camera to the images in the database which established the last step; third, retrieving some of the most similar pictures, then extracting the feature and matching the points; at last, solving the perspective-*n*-point problem [9–12]. However, the application of scene recognition to mobile location implies several challenges [13–15]. The complex three-dimensional shape of the environment results in occlusions, overlaps, shadows, and reflections which require a robust description of the scene [16]. To address these issues,

we propose a particularly efficient approach based on a deep belief network with local binary pattern feature descriptors. It enables us to find out the most similar pictures quickly. In addition, we restrict the search space according to adaptive visibility constraints which allows us to cope with extensive maps.

2. Related Work

Before presenting the proposed approach, we review previous work on image-based localization methods and divide these methods into three categories roughly.

Manual mark-based localization methods completely rely on the natural features of the image which lacks robustness, especially under conditions of varying illumination. In order to improve the robustness and accuracy of the reference point, special coding marks are used to meet the higher positioning requirements of the system. There are three benefits: simplify the automatic detection of corresponding points, introduce system dimensions, and distinguish and identify targets by using a unique code for each mark. Common types of marks include concentric rings, QR codes, or patterns composed of colored dots. The advantage is raising the recognition rate and effectively reducing the complexity of positioning methods. The disadvantage is that the installation and maintenance costs are high, some targets are easily obstructed, and the scope of application is limited [17, 18].

Natural mark-based localization methods usually detect objects on the image and match them with an existing building database. The database contains the location information of the natural marks in the building. The advantage of this method is that it does not require additional local infrastructure. In other words, the reference object is actually a series of digital reference points (control points in photogrammetry) in the database. Therefore, this type of system is suitable for large-scale coverage without increasing too much cost. The disadvantage is that the recognition algorithm is complex and easy to be affected by the environment, the characteristics are easy to change, and the dataset needs to be updated [19–22].

Learning-based localization methods have emerged in the past few years. It is an end-to-end method that directly obtains 6dof pose, which has been proposed to solve loop-closure detection and pose estimation [23]. This method does not require feature extraction, feature matching, and complex geometric calculations and is intuitive and concise. It is robust in weak textures, repeated textures, motion blur, and lighting changes. In the training phase, the calculative scale is very large, and GPU servers are usually required, which cannot run smoothly on mobile platforms [20]. In many scenarios, learning-based features are not as effective as traditional features such as SIFT, and the interpretability is poor [24–27].

3. Framework and Method

In this section, first, we introduce the overview of the framework. Then, the key modules are explained in more detail in the subsequent sections.

3.1. Framework Overview. The whole pipeline of the visual localization system is shown in Figure 1. In the following, we briefly provide an overview of our system.

In the offline stage, the RGB-D cameras are held to collect enough RGB images and depth images around the indoor environment. At the same time, the pose of the camera and the 3D point cloud are constructed. The RGB image is used as a learning dataset to train the network model, and then, the network model parameters are saved until the loss function value does not decrease. In the online stage, after the previous step is completed, anyone enters the room, downloads the trained network model parameters to the mobile phone, and takes a picture with the mobile phone, and the most similar image is identified according to the deep learning network. The unmatched points are eliminated, and the pixel coordinates of the matched points and the depth of the corresponding points are extracted. According to the pinhole imaging model, the n-point perspective projection problem-solving method can be used to calculate the pose of the mobile phone in the world coordinate system. Finally, the posture is converted into a real position and displayed on the map.

3.2. Camera Calibration and Image Correction. Due to the processing error and installation error of camera lens, the image has radial distortion and tangential distortion. Therefore, we must calibrate the camera and correct the images in the preprocessing stage. The checkerboard contains some calibration reference points, and the coordinates of each point are disturbed by the same noise. Establishing the function γ:

$$\gamma = \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| p_{ij} - p\wedge(A, R_i, t_i, P_i) \right\|^2, \tag{1}$$

where p_{ij} is the coordinate of the projection points on image i for reference point j in the three-dimensional space. R_i and t_i are the rotation and translation vectors of image i. P_i is the three-dimensional coordinate of reference point i in the world coordinate system. $\widehat{p}(A, R_i, t_i, P_i)$ is the two-dimensional coordinate in the image coordinate system.

3.3. Scene Recognition. In this section, we use the deep belief network (DBN) to categorize the different indoor scenes. The framework includes image preprocessing, LBP feature extracting, DBN training, and scene classification.

3.3.1. Local Binary Pattern. The improved LBP feature is insensitive to rotation and illumination changes. The LBP operator can be specifically described as the following: the gray values in the window center pixel are defined as the threshold, and the gray values of the surrounding 8 pixels are, respectively, compared with the threshold in a clockwise direction, and if the gray value is bigger than the threshold, then mark the pixel as 1; otherwise, mark 0, and then get an 8-bit binary number through the comparison. After the decimal conversion, get the LBP value of the center pixel in this window. The value reflects the texture information of the point at this position. The calculation process is shown in Figure 2.

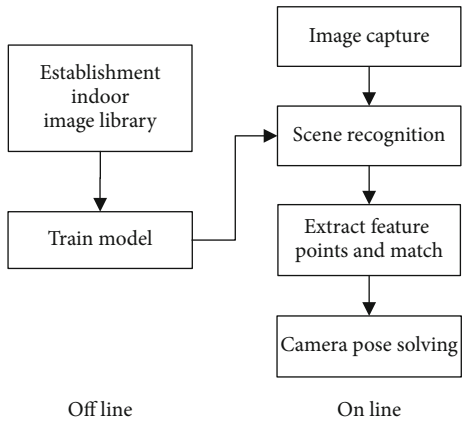


Figure 1: The framework of the visual localization system.

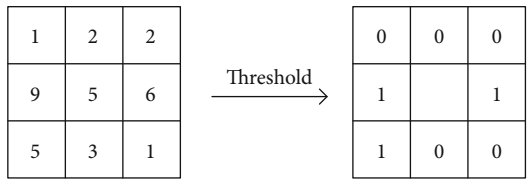


Figure 2: Local binary pattern calculation process.

The formula of local binary pattern:

$$\mathrm{LBP}(x_c, y_c) = \sum_{n=0}^{N-1} 2^n s(i_n - i_c),$$
$$s(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{else}, \end{cases} \tag{2}$$

where (x_c, y_c) is the horizontal and vertical coordinate of the center pixel; N is number 8; i_c, i_n are the gray values of the center pixel and the neighborhood pixel, respectively; and $s(\cdot)$ is the two-valued symbol function.

The earliest proposed LBP operator can only cover a small range of images, so the optimization and improvement methods for the LBP operator are constantly proposed by researchers. We adopt the method which improves the insufficiency of the window size of the original LBP operator by replacing the traditional square neighborhood with a circular neighborhood and expanding the window size as shown in Figure 3.

In order to make the LBP operator have rotation invariance, the circular neighborhood is rotated clockwise to obtain a series of binary strings, and the minimum binary value is obtained, and then, the value is converted into decimal, which is the LBP value of the point. The process of obtaining the rotation-invariant LBP operator is shown in Figure 4.

3.3.2. Deep Belief Network. The deep belief network consists of a multirestricted Boltzmann machine (RBM) and a backpropagation (BP) neural network. The Boltzmann machine is a neural network based on learning rules. It consists of a visible layer and a hidden layer. The neurons in the same layer and the neurons in different layers are connected to each other. There are two types of neuron output states: active and inactive, represented by numbers 1 and 0. The advantage of the Boltzmann machine is its powerful unsupervised learning ability, which can learn complex rules from a large amount of data; the disadvantages are the huge amount of calculation and the long training time. The restricted Boltzmann machine canceled the connection between neurons in the same layer; each hidden unit and visible layer unit are independent of each other. Roux and Bengio theoretically prove that as long as the number of neurons in the hidden layer and the training samples are sufficient, the arbitrary discrete distribution can be fitted. The structure of BM and RBM is shown in Figure 5.

The joint configuration energy of its visible and hidden layers is defined as

$$E(v, h|\theta) = -\sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} c_j h_j - \sum_{i=1}^{m}\sum_{j=1}^{n} v_i w_{ij} h_j, \tag{3}$$

where $\theta = \{W_{ij}, b_i, c_j\}$ are parameters in RBM, b_i is bias of visible layer i, c_j is bias of visible layer j, and w_{ij} is the weight.

The output of the hidden layer unit is

$$h_j = \sum_{j=1}^{n} v_i w_{ij} + b_j. \tag{4}$$

When the parameters are known, based on the above energy function, the joint probability distribution of (v, h)

$$P(v, h|\theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)},$$
$$Z(\theta) = \sum_{v,h} e^{-E(v,h|\theta)}, \tag{5}$$

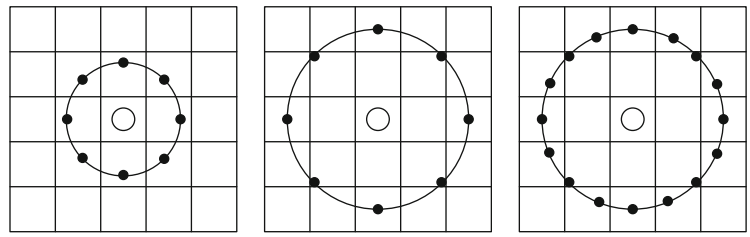
where $Z(\theta)$ is the normalization factor. Distribution of v is $P(v|\theta)$, joint probability distribution $P(v, h|\theta)$:

$$P(v|\theta) = \sum_{h} p(v, h|\theta) = \frac{1}{Z(\theta)} \sum_{h} e^{-E(v,h|\theta)}. \tag{6}$$

Since the activation state of each hidden unit and visible unit is conditionally independent, therefore, when the state of the visible and hidden units is given, the activation probability of the first implicit unit and visible elements is

$$P(h_j = 1|v, \theta) = \sigma\left(b_j + \sum_{i=1}^{m} v_i w_{ij}\right),$$
$$P(v_i = 1|h, \theta) = \sigma\left(c_i + \sum_{j=1}^{n} h_j w_{ij}\right), \tag{7}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid activation function.



(a) LBP_5^1 (b) LBP_5^2 (c) LBP_{16}^2

Figure 3: Three types of LBP.

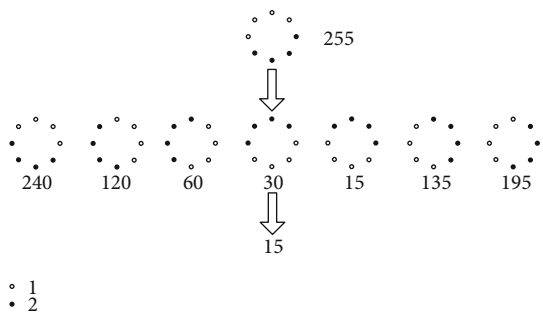


Figure 4: Rotation-invariant LBP schematic.

3.4. Feature Point Detection and Matching. In this paper, we propose a multifeature point fusion algorithm. The combination of the edge detection algorithm and the ORB detection algorithm enables the detection algorithm to extract the edge information, thereby increasing the number of matching points with fewer textures. The feature points of the edge are obtained by the Canny algorithm to ensure that the object with less texture has feature points. ORB have scale and rotation invariance, and the speed is faster than SIFT. The BRIEF description algorithm is used to construct the feature point descriptor [28–31].

The Brute force algorithm is adopted as the feature matching strategy. It calculates the Hamming distance between each point of the template image and each feature point of the sample image. Then compare the minimum Hamming distance value with the threshold value; if the distance is less than the threshold value, regard these two points as the matching points; otherwise, they are not matching points. The framework of feature extraction and matching is shown in Figure 6.

3.5. Pose Estimation. The core idea is to select four noncoplanar virtual control points; then, all the spatial reference points are represented by the four virtual control points, and then, the coordinates of the virtual control points are solved by the correspondence between the spatial reference points and the projection points, thereby obtaining the coordinates of all the spatial reference points. Finally, the rotation matrix and the translation vector are solved. The specific algorithm is described as follows.

Given n reference points, the world coordinate is $\tilde{P}_i^w = (x_i, y_i, z_i)^T$, $i = 1, 2, \cdots, n$. The coordinates of the corresponding projection point in the image coordinate system are $\tilde{u}_i = (u_i, v_i)^T$, and the corresponding homogeneous coordinates are $P_i^w = (x_i, y_i, z_i, 1)^T$ and $u_i = (u_i, v_i, 1)^T$. The correspondence between the reference point P_i^w and the projection point u_i:

$$\lambda_i u_i = K[R\,t]P_i^w, \tag{8}$$

where λ_i is the depth of the reference point and K is the internal parameter matrix of the camera:

$$K = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{9}$$

where $f = f_u = f_v$ is the focal length of the camera and $(u_0, v_0) = (0, 0)$ is the optical center coordinate.

First, select four noncoplanar virtual control points in the world coordinate system. The relationship between the virtual control points and their projection points is shown in Figure 7.

In Figure 7, $\mathbf{C}_1^w = [0, 0, 0, 1]^T$, $\mathbf{C}_2^w = [1, 0, 0, 1]^T$, $\mathbf{C}_3^w = [0, 1, 0, 1]^T$, and $\mathbf{C}_4^w = [0, 0, 1, 1]^T$. $\{\mathbf{C}_j^c, j = 1, 2, 3, 4\}$ are homogeneous coordinates of the virtual control point in the camera coordinate system, $\{\tilde{\mathbf{C}}_j^c, j = 1, 2, 3, 4\}$ is the corresponding nonhomogeneous coordinate, $\{\mathbf{C}_j, j = 1, 2, 3, 4\}$ is the homogeneous coordinate of the projection point corresponding in the image coordinate system, and $\{\tilde{\mathbf{C}}_j, j = 1, 2, 3, 4\}$ is the corresponding nonhomogeneous coordinate. $\{\mathbf{P}_i^c, i = 1, 2, \cdots, n\}$ is the homogeneous coordinate of the reference point in the camera coordinate system; $\{\tilde{\mathbf{P}}_i^c, i = 1, 2, \cdots, n\}$ is the corresponding nonhomogeneous coordinate. The relationship between the spatial reference points and the control points in the world coordinate is as follows:

$$\mathbf{P}_i^w = \sum_{j=1}^{4} \alpha_{ij} \mathbf{C}_j^w,\ i = 1, 2, \cdots, n, \tag{10}$$

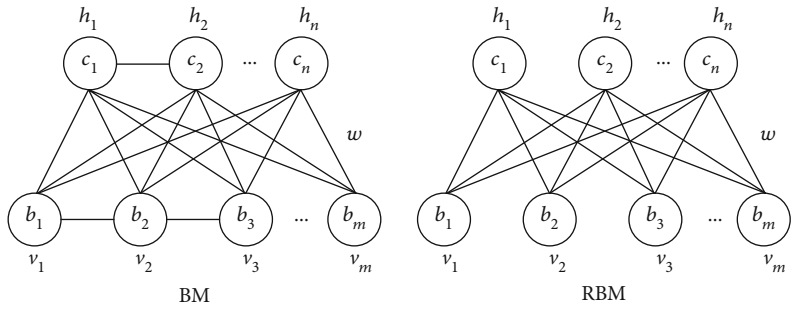


Figure 5: Boltzmann machine and restricted Boltzmann machine. v is the visible layer, m indicates the number of input data, h is the hidden layer, and w is the connection weight between two layers,$\forall i, j, v_i \in \{0, 1\}, h_j \in \{0, 1\}$.

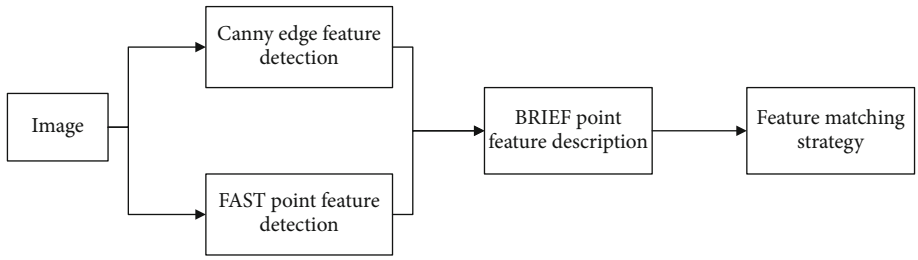


Figure 6: The process of multifeature fusion extraction and matching.

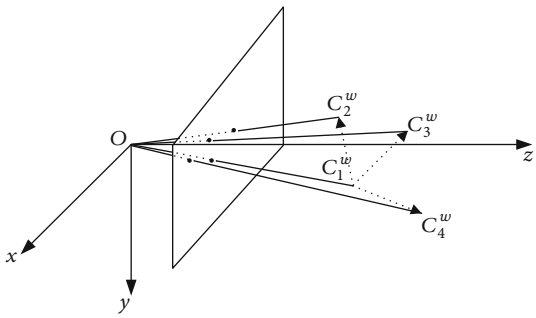


Figure 7: Virtual control point and its projection point correspondence.

where vector $[\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \alpha_{i4}]^T$ is the coordinate of the Euclidean space based on the control point $\mathbf{C}_i^c$. From the invariance of the linear relationship under the Euclidean transformation,

$$\begin{aligned} \mathbf{P}_i^c &= \sum_{j=1}^{4} \alpha_{ij} C_j^c, \quad i = 1, 2, \cdots, n, \\ \lambda_i u_i &= \mathbf{K}\tilde{\mathbf{P}}_i^c = \mathbf{K} \sum_{j=1}^{4} \alpha_{ij} \tilde{\mathbf{C}}_j^c, \quad i = 1, 2, \cdots, n. \end{aligned} \tag{11}$$

Assume $\tilde{\mathbf{C}}_j^c = [x_j^c, y_j^c, z_j^c]^T$, then

$$\lambda_i = \sum_{j=1}^{4} \alpha_{ij} z_j^c. \tag{12}$$

Then, obtain the equation:

$$\begin{aligned} \sum_{j=1}^{4} \alpha_{ij} f x_j^c - \alpha_{ij} u_i z_j^c = 0, \\ \sum_{j=1}^{4} \alpha_{ij} f y_j^c - \alpha_{ij} v_i z_j^c = 0. \end{aligned} \tag{13}$$

Assume $Z = [Z_1^{cT}, Z_2^{cT}, Z_3^{cT}, Z_4^{cT}]^T$, $Z_j^c = [f x_j^c, f y_j^c, z_j^c]^T$, $j = 1, 2, 3, 4$, then the equations are obtained from the correspondence between spatial points and image points as follows:

$$MZ = 0. \tag{14}$$

The solution Z is the kernel space of the matrix M:

$$Z = \sum_{i=1}^{N} \beta_i W_i, \tag{15}$$

where W_i is the eigenvector of $M^T M$, N is the dimension of the kernel, and β_i is the undetermined coefficient. For a perspective projection model, the value of N is 1, resulting in

$$Z = \beta W, \tag{16}$$

where $W = [w_1^T, w_2^T, w_3^T, w_4^T]^T$, $w_j = [w_{j1}, w_{j2}, w_{j3}]^T$; then, the image coordinates of the four virtual control points are

$$\mathbf{c}_j = \left\{ \frac{w_{j1}}{w_{j3}}, \frac{w_{j2}}{w_{j3}}, 1 \right\}, \quad j = 1, 2, 3, 4. \tag{17}$$

The image coordinates of the four virtual control points obtained by the solution and the camera focal length obtained during the calibration process are taken into the absolute positioning algorithm to obtain the rotation matrix and the translation vector.

4. Experiments

We conducted two experiments to evaluate the proposed system. In the first experiment, we compare the proposed algorithm with other state-of-the-art algorithms on public datasets and then perform numerical analysis to show the accuracy of our system. The second experiment evaluated the performance of accuracy in the real world.

4.1. Experiment Setup. The experimental devices include an Android mobile phone (Lenovo Phab 2 Pro) and a depth camera (Intel RealSense D435) as shown in Figure 8. The user interface of the proposed visual positioning system on a smart mobile phone running in an indoor environment is shown in Figure 9.

4.2. Experiment on Public Dataset. In this experiment, we adopted the ICL-NUIM dataset which consists of RGB-D images from camera trajectories from two indoor scenes. The ICL-NUIM dataset is aimed at benchmarking RGB-D, Visual Odometry, and SLAM algorithms [32–34]. Two different scenes (the living room and the office room scene) are provided with ground truth. The living room has 3D surface ground truth together with the depth maps as well as camera poses and as a result perfectly suits not only for benchmarking camera trajectory but also for reconstruction. The office room scene comes with only trajectory data and does not have any explicit 3D model with it. The images were captured at 640∗480 resolutions.

Table 1 shows localization results for our approach compared with state-of-the-art methods. The proposed localization method is implemented on Intel Core i5-4460 CPU@3.20 GHz. The total procedure from scene recognition to pose estimation takes about 0.17 s to output a location for a single image.

4.3. Experiment on Real Scenes. The images are acquired by a handheld depth camera at a series of locations. The image size is 640 × 480 pixels, and the focal length of the camera is known. Several images of the laboratory are shown in Figure 10.

Using the RTAB-Map algorithm, we get the 3D point cloud of the laboratory. It is shown in Figure 11. The blue points are the position of the camera, and the blue line is the trajectory.

The 2D map of our laboratory is shown in Figure 12. The length and width of the laboratory are 9.7 m and 7.8 m, respectively. First, select a point in the laboratory as the origin of the coordinate system and establish a world coordinate system. Then, hold the mobile phone, walk along different routes, and take photos, respectively, as indicated by the arrows.

In the offline stage, we get a total of 144 images. Due to some images captured at different scenes being similar, we divide them into 18 categories. In the online stage, we captured 45 images at different locations on route 1 and 27 images on route 2. The classification accuracy formula is

$$P = \frac{N_i}{N}, \tag{18}$$

where N_i is the correct classified number of scene images and N is the total number of scene images. The classification accuracy of our method is 0.925.

Most mismatched scenes concentrate in the corner, mainly due to the lack of significant features or mismatches. Several mismatched scenes are shown in Figure 13.

After removing the wrong matched results, the error cumulative distribution function graph is shown in Figure 14.

The trajectory of the camera is compared with the predefined route. After calculating the Euclidean distance between the results through our method and the true position, we get the error cumulative distribution function

Figure 8: Intel RealSense D435 and Lenovo mobile phone.

Figure 9: The user interface of the proposed visual positioning system on a smart mobile phone running in an indoor environment.

Table 1: Comparison of mean error in ICL-NUIM dataset.

Method	Living room	Office room
PoseNet	0.60 m, 3.64°	0.46 m, 2.97°
4D PoseNet	0.58 m, 3.40°	0.44 m, 2.81°
CNN+LSTM	0.54 m, 3.21°	0.41 m, 2.66°
Ours	0.48 m, 3.07°	0.33 m, 2.40°

Figure 10: Images captured from different scenes.

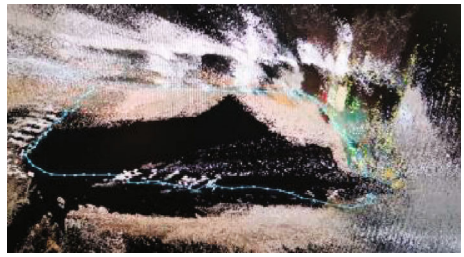


Figure 11: 3D point cloud of laboratory.

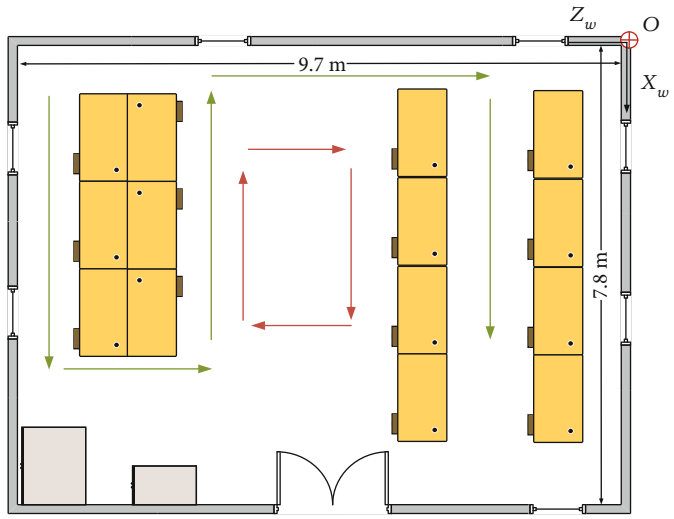


Figure 12: Environmental map and walking route.

Figure 13: Mismatched scene.

graph (Figure 14). It can be seen that the average positioning error is 0.61 m. Approximately 58% point positioning error is less than 0.5 m, about 77% point error is less than 1 m, about 95% point error is less than 2 m, and the maximum error is 2.55 m.

Since the original depth images in our experiment are based on RTAB-Map, its accuracy is not accurate. For example, in an indoor environment, intense illumination and strong shadows may lead to inconspicuous local features. It is also difficult to construct a good point cloud model. In

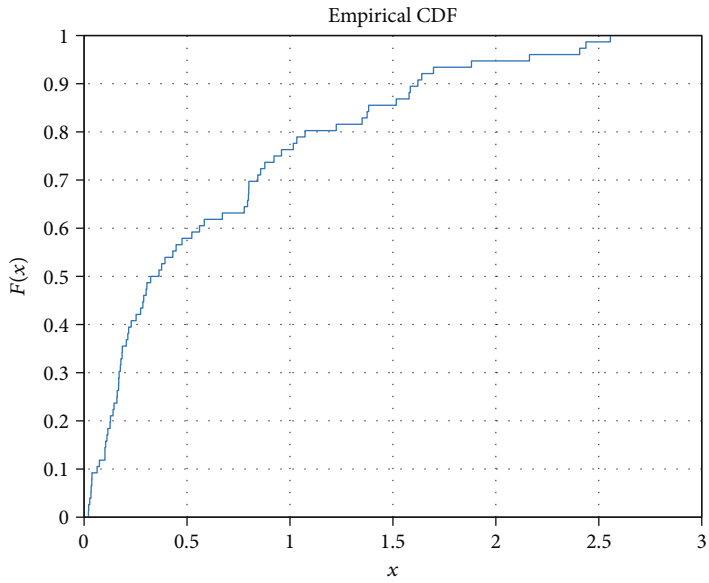


Figure 14: Error cumulative distribution function graph.

the future, we plan to use laser equipment to construct a point cloud.

5. Conclusions and Future Work

In this article, we have presented an indoor positioning system based only on cameras. The main work is to use deep learning to identify the category of the scene and use 2D-3D matching feature points to calculate the location. We implemented the proposed approach on a mobile phone and achieved a positioning accuracy of decimeter level. The preliminary indoor positioning experiment result is given in this paper. But the experimental site is a small-scale place. The following work needs to be done in the future: with the rapid development of deep learning, it can generate high-level semantics and effectively solve the limitations caused by artificial design features, use a more robust lightweight image retrieval algorithm, and carry out tests under different lighting and dynamic environments, system tests under large-scale scenarios, and long-term performance tests.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was partially supported by the Key Research Development Program of Hebei (Project No. 19210906D).

References

[1] J. Wu, S. Guo, H. Huang, W. Liu, and Y. Xiang, "Information and communications technologies for sustainable development goals: state-of-the-art, needs and perspectives," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2389–2406, 2018.

[2] P. Lazik, N. Rajagopal, O. Shih, B. Sinopoli, and A. Rowe, "Alps: s bluetooth and ultrasound platform for mapping and localization," in *Proceedings of the 13th ACM Conferenceon Embedded Networked Sensor Systems, ACM*, pp. 73–84, New York, NY, USA, 2015.

[3] S. He and S. Chan, "Wi-Fi fingerprint-based indoor positioning: recent advances and comparisons," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 466–490, 2017.

[4] C. L. Wu, L. C. Fu, and F. L. Lian, "WLAN location determination in e-home via support vector classification," in *Networking Sensing and Control, IEEE International Conference, 2004*, pp. 1026–1031, Taipei, Taiwan, 2004.

[5] G. Ding, Z. Tan, J. Wu, and J. Zhang, "Efficient indoor fingerprinting localization technique using regional propagation model," *IEICE Transactions on Communications*, vol. 8, pp. 1728–1741, 2014.

[6] G. Ding, Z. Tan, J. Wu, J. Zeng, and L. Zhang, "Indoor fingerprinting localization and tracking system using particle swarm optimization and Kalman filter," *IEICE Transactions on Communications*, vol. 3, pp. 502–514, 2015.

[7] C. Toft, W. Maddern, A. Torii et al., "Long-term visual localization revisited," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2020.

[8] A. Xiao, R. Chen, D. Li, Y. Chen, and D. Wu, "An indoor positioning system based on static objects in large indoor scenes by using smartphone cameras," *Sensors*, vol. 18, no. 7, pp. 2229–2246, 2018.

[9] E. Deretey, M. T. Ahmed, J. A. Marshall, and M. Greenspan, "Visual indoor positioning with a single camerausing PnP," in *In Proceedings of the 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 1–9, Banff, AB, Canada, October 2015.

[10] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2969–2976, Colorado Springs, CO, USA, 2011.

[11] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2d-to-3dmatching," in *2011 IEEE International Conference on Computer Vision, IEEE*, pp. 667–674, Barcelona, Spain, 2011.

[12] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," in *European Conference on Computer Vision (ECCV)*, Berlin, Heidelberg, 2012.

[13] M. Larsson, E. Stenborg, C. Toft, L. Hammarstrand, T. Sattler, and F. Kahl, "Fine-grained segmentation networks: self-supervised segmentation for improved long-term visual localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 31–41, Seoul, Korea, 2019.

[14] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool, "Night-to-day image translation for retrieval-based localization," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 5958–5964, Montreal, QC, Canada, 2019.

[15] J. X. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: large-scale scene recognition from abbey to zoo," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, San Francisco, CA, USA, 2010.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarseto fine: robust hierarchical localization at large scale," in *Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12716–12725, California, 2019.

[18] Q. Niu, M. Li, S. He, C. Gao, S.-H. Gary Chan, and X. Luo, "Resource efficient and automated image-based indoor localization," *ACM Transactions on Sensor Networks*, vol. 15, no. 2, pp. 1–31, 2019.

[19] Y. Chen, R. Chen, M. Liu, A. Xiao, D. Wu, and S. Zhao, "Indoor visual positioning aided by CNN-based image retrieval: training-free, 3D modeling-free," *Sensors*, vol. 18, no. 8, pp. 2692–2698, 2018.

[20] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *IEEE International Conference on Robotics & Automation*, pp. 4762–4769, Stockholm, Sweden, 2016.

[21] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 39, no. 9, pp. 1744–1756, 2016.

[22] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 39, no. 7, pp. 1455–1461, 2016.

[23] B. Zeisl, T. Sattler, and M. Pollefeys, "Camera pose voting for large-scale image-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 2704–2712, Santiago, Chile, 2015.

[24] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: a convolutional network for real-time 6-dof camera relocalization," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946, Santiago, Chile, 2015.

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, Salt Lake City, Utah, 2018.

[26] Z. Chen, A. Jacobson, N. Sunderhauf et al., "Deep learning features at scale for visual place recognition," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, 2017.

[27] S. Lynen, B. Zeisl, D. Aiger et al., "Large-scale, real-time visual–inertial localization revisited," *The International Journal of Robotics Research*, vol. 39, no. 9, pp. 1–24, 2020.

[28] M. Dusmanu, I. Rocco, T. Pajdla et al., "D2-net: a trainable cnn for joint description and detection oflocal features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8092–8101, California, 2019.

[29] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *IEEE International Conference on Robotics and Automation*, pp. 1848–1853, Kobe, Japan, 2009.

[30] A. Xu and G. Namit, "SURF: speeded-up robust features," *Computer Vision & Image Understanding*, vol. 110, no. 3, pp. 404–417, 2008.

[31] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision*, pp. 2564–2571, Barcelona, Spain, 2012.

[32] A. Handa, T. Whelan, J. Mcdonald, and A. J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *2014 IEEE international conference on Robotics and automation (ICRA)*, pp. 1524–1531, Hong Kong, China, 2014.

[33] M. Labbe and F. Michaud, "RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, 2019.

[34] Z. Gao, Y. Li, and S. Wan, "Exploring deep learning for view-based 3D model retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1, pp. 1–21, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 2939334, 13 pages
https://doi.org/10.1155/2021/2939334

Research Article

Explainable Artificial Intelligence for Sarcasm Detection in Dialogues

Akshi Kumar[1,2] **Shubham Dikshit**[3] **and Victor Hugo C. Albuquerque**[1,4]

[1]*Graduate Program on Telecommunication Engineering, Federal Institute of Education, Science and Technology of Ceará, Fortaleza, CE, Brazil*
[2]*Department of Computer Science and Engineering, Delhi Technological University, Delhi, India*
[3]*Department of Computer Science and Engineering, IMS Engineering College, Ghaziabad, India*
[4]*Graduate Program on Teleinformatics Engineering, Federal University of Ceará, Fortaleza, CE, Brazil*

Correspondence should be addressed to Akshi Kumar; akshikumar@dce.ac.in

Received 17 May 2021; Revised 11 June 2021; Accepted 21 June 2021; Published 3 July 2021

Academic Editor: Mohammad R. Khosravi

Sarcasm detection in dialogues has been gaining popularity among natural language processing (NLP) researchers with the increased use of conversational threads on social media. Capturing the knowledge of the domain of discourse, context propagation during the course of dialogue, and situational context and tone of the speaker are some important features to train the machine learning models for detecting sarcasm in real time. As situational comedies vibrantly represent human mannerism and behaviour in everyday real-life situations, this research demonstrates the use of an ensemble supervised learning algorithm to detect sarcasm in the benchmark dialogue dataset, MUStARD. The punch-line utterance and its associated context are taken as features to train the eXtreme Gradient Boosting (XGBoost) method. The primary goal is to predict sarcasm in each utterance of the speaker using the chronological nature of a scene. Further, it is vital to prevent model bias and help decision makers understand how to use the models in the right way. Therefore, as a twin goal of this research, we make the learning model used for conversational sarcasm detection interpretable. This is done using two post hoc interpretability approaches, Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP), to generate explanations for the output of a trained classifier. The classification results clearly depict the importance of capturing the intersentence context to detect sarcasm in conversational threads. The interpretability methods show the words (features) that influence the decision of the model the most and help the user understand how the model is making the decision for detecting sarcasm in dialogues.

1. Introduction

Natural language is a vital information source of human sentiments. Automated sarcasm detection is often described as a natural language processing (NLP) problem as it primarily requires understanding the human expressions, language, and/or emotions articulated via textual or nontextual content. Sarcasm detection has attracted growing interest over the past decade as it facilitates accurate analytics in online comments and reviews [1, 2]. As a figurative literary device, sarcasm makes use of words in a way that deviates from the conventional order and meaning thereby misleading polarity classification results. For example, in a statement "*Staying up till 2:30am was a brilliant idea to miss my office meeting*," the positive word "*brilliant*" along with the adverse situation "*miss my office meeting*" conveys the sarcasm, because sarcasm has an implied sentiment (negative) that is different from surface sentiment (positive due to presence of "brilliant"). Various rule-based, statistical, machine learning, and deep learning-based approaches have been reported in pertinent literature on automatic sarcasm detection in single sentences that often rely on the content of utterances in isolation. These include a range of techniques such as sense disambiguation [3] to polarity flip detection in text [4] and multimodal (text +image) content [5, 6]. Furthermore, its use on social media platforms like Twitter and Reddit is primarily to convey user's frivolous intent, and therefore, the dialect is more casual and includes the use of microtext like

wordplay, neologism, emojis, and slangs. Few recent works have taken into account the additional contextual information along with the utterance to deal with these challenges. Different researchers have considered varied operational cues to typify context. In 2019, Kumar and Garg [7] defined five broad categories of context, namely, social-graph, temporal, content, modality, and user-profile based which can be used for improving the classification accuracy. Evidently, it is essential to capture the operational concern, that is, the pragmatic meaning defined by "context" as sarcasm. But the use of sarcasm in dialogues and conversational threads has further added to the challenges making it vital to capture the knowledge of the domain of discourse, context propagation during the course of dialogue, and situational context and tone of the speaker. For example, recently, several Indian airlines took to Twitter to engage users in a long thread meant to elicit laughs and sarcastic comebacks amid the coronavirus lockdown that has kept passengers and airlines firmly on the ground. IndiGo playfully teased its rivals by engaging in a Twitter banter resulting in comic wordplays on airlines' advertising slogans. IndiGo began by asking Air Vistara "*not flying higher?*" in reply to which the airlines tagged peer GoAir, punning on its tagline "fly smart" and what followed was other key airlines like AirAsia and SpiceJet joining the thread exchange equipped with witty responses using each other's trademark business taglines (https://www.deccanherald.com/business/coronavirus-indigo-vistara-spicejet-engage-in-banter-keep-twitterati-in-splits-amid-lockdown-blues-823677.html).

As seen in Figure 1, it is not only important to capture the intrasentence context but the intersentence context too to detect sarcasm in conversational threads. Moreover, the sarcastic intent of the thread is difficult to comprehend without the situational context as in this case is the unprecedented travel restrictions, including the grounding of all domestic and international passenger flights, to break the chain of the coronavirus disease (COVID-19) transmission.

But as sarcasm is a convoluted form of expression which can cheat and mislead analytic systems, it is equally important to achieve high prediction accuracy with decision understanding and traceability of actions taken. As models cannot account for all the factors that will affect the decision, explainability can account for context and help understand the included factors that will affect decision making so that one can adjust prediction on additional factors. Explainable artificial intelligence (XAI) [8, 9] is the new buzzword in the domain of machine learning which intends to justify the actions and understand the model behaviour. It enables building robust models with better decision-making capabilities.

Thus, in this paper, we firstly demonstrate the role of context in conversational threads to detect sarcasm in the MUStARD dataset [5], which is a multimodal video corpus for research in automated sarcasm discovery compiled using dialogues from famous sitcoms, namely, "Friends" by Bright, Kauffman, Crane Productions, and Warner Bros. Entertainment Inc., "The Big Bang Theory" by Chuck Lorre, Bill Prady, CBS, "The Golden Girls" by Susan Harris, NBC, and "Sarcasmaholics Anonymous." The data is labelled with true

Figure 1: Online sarcastic conversational thread.

and false for the sarcastic and nonsarcastic dialogues using the sequential nature of scenes in the episodes, and we use eXtreme Gradient Boosting (XGBoost) method [10] to primarily investigate how conversational context can facilitate automatic prediction of sarcasm. As a twin goal of this research, we aim to make the supervised learning models used for conversational sarcasm detection interpretable with the help of XAI. The goal is to show the words (features) that influence the decision of the model the most.

Using dialogue dataset from sitcoms can invariably relate to any real-life utterance making this work relevant for various sentiment analysis-based market and business intelligence applications for assessing insights from conversational threads on social media. Most situational comedies or sitcoms are led by the comedy of manners, vaudeville, and our tacit perceptions of everyday life. These are the story of our psychodynamics and sociodynamics on situations that could arise in everyday life and unfold the unexpected and ironic comedy of human behaviour in real-life situations. For example, in Friends, season 10, episode 3, Ross walks in with a clearly overdone tan to the point that his skin color is very dark and looks truly ridiculous. He tells Chandler that he went to the tanning place his wife (Monica) suggested. And Chandler came up with a sarcastic statement "*Was that place the sun?*" as it looked like the only tanning place that could make someone's skin look like that would be sitting directly beneath the scorching sun! The sarcasm in Chandler's dialogue could only be understood considering the entire conversation and not taking his dialogue in isolation (Figure 2).

XAI in a typical NLP task setting can offer twofold advantages, namely, transferability, as machine learning models are trained in a controlled setting, deployment in real time should also ensure that the model has truly learned to detect underlying phenomenon, and secondly, it can help determining the contextual factors that affect the decision. The terms interpretability and explainability are often used

Figure 2: Friends: season 10, episode 3.

interchangeably as both play a complementary role in understanding predictive models [11]. The term interpretability tells us what is going on in the algorithm, i.e., it enables us to predict what will happen if there are some changes in the parameters or input, and explainability tells the extent to which the internal working of any machine learning or deep learning model can be explained in human terms. Characteristically, interpretable machine learning systems provide explanations for their outputs. According to Miller [12], interpretability is defined as the capability to understand the decision and means that the cause and effect can be determined. Interpretable machine learning (ML) describes the process of revealing causes of predictions and explaining a derived decision in a way that is understandable to humans. The ability to understand the causes that lead to a certain prediction enables data scientists to ensure that the model is consistent with the domain knowledge of an expert. Furthermore, interpretability is critical to obtain trust in a model and to be able to tackle problems like unfair biases or discrimination. One way to apply interpretable ML is by using models that are intrinsically interpretable and known to be easy for humans to understand such as linear/logistic regression, decision trees, and K-nearest neighbors [13]. Alternatively, we can train a black-box model and apply post hoc interpretability techniques [14] (Figure 3) to provide explanations.

In this paper, we use two post hoc model agnostic explainability techniques called Local Interpretable Model-agnostic Explanations (LIME) [15, 16] and Shapley Additive exPlanations (SHAP) [17, 18] to analyze the models on the dataset by checking the evaluation metrics and select the model where explanation can be separated from the models. The intent is to evaluate the black-box model much easily on how each word plays an important role in the prediction of the sarcastic dialogues by the speaker using the sequential nature of a scene in the TV series. Thus, the key contributions of this research are as follows:

	Global	Local
Model-specific	Model internals; Intrinsic feature importance	Rule sets (Tree structure)
Model-agnostic	Partial dependence plots; Feature importance; Global surrogate models	Individual conditional expectation; Local surrogate models

Figure 3: Post hoc interpretability techniques.

(i) Using sequence of utterances to detect sarcasm in real-time dialogues

(ii) Using post hoc model-agnostic local surrogate machine learning interpretability methods to comprehend which words within a dialogue are the most important for predicting sarcasm

The scope of the research can be extended to real-time AI-driven sentiment analysis for improving customer experience where these explanations would help the service desk to detect sarcasm and word importance while predicting sentiment. The organization of the paper is as follows: the next section briefs about the taxonomy of machine learning interpretability methods followed by related work within the domain of sarcasm detection specifically in conversational data in Section 3. Section 4 discusses the key techniques used in this research followed by the results and conclusion in Section 5 and Section 6, respectively.

2. Taxonomy of Machine Interpretability Methods

Artificial intelligence (AI) is gradually participating in day-to-day experiences. Its entrusted adoption and encouraging acceptance in various real-time domains are highly contingent upon the transparency, interpretability, and explainability of models built. Particularly in customer-centric environments, trust and fairness can help customers achieve better outcomes. Introduced in the early 1980s, XAI is a framework and tool which helps humans to understand the model behaviour and enables building robust models with better decision-making capabilities. It is used for understanding the logic behind the predictions made by the model and justifies its results to the user.

A trade-off between the model interpretability and predictive power is commonly observed as shown in Figure 4. As the model gets more advanced, it becomes harder to explain how it works. High interpretability models include traditional regression algorithms (linear models, for example), decision trees, and rule-based learning. Basically, these are approximate monotonic linear functions. On the other hand, low interpretability models include ensemble methods and deep learning where the black-box feature extraction offers poor explainability.

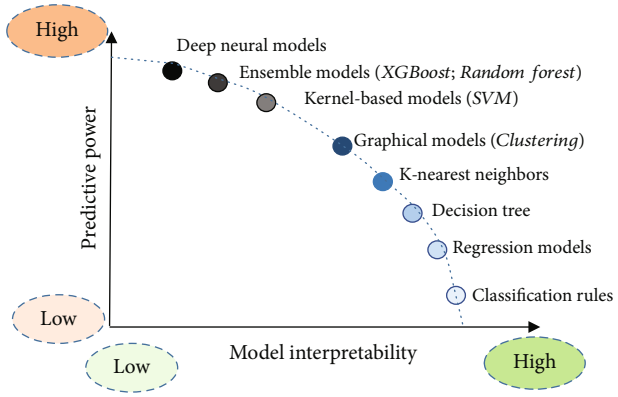


Figure 4: Predictive power vs. interpretability trade-off.

Machine interpretability methods are often categorized along three main criteria [19, 20]. The first discriminates based on the coverage of explanation as local or global for explanation for at instance-level (individual predictions) or model-level (entire model), respectively. Global interpretability methods explain the entire ML model at once from input to prediction, for example, decision trees and linear regression. Local interpretability methods explain how predictions change for when input changes and are applicable for a single prediction or a group of predictions. The second criteria differentiate between the explanations based on the interpretable design capabilities as intrinsically interpretable models and post hoc models (Figure 5). Intrinsically interpretable models are models that are interpretable by design, and no postprocessing steps are needed to achieve interpretability. These are self-explaining, and explainability is often achieved as a by-product of model training. On the other hand, in post hoc methods, explainability is often achieved after the model is trained and it requires postprocessing using external methods to achieve interpretability.

The third criterion to categorize interpretability methods is the applicability limitation to specific models or any ML model. Based on these criteria, the methods are divided into model-specific and model-agnostic methods. Model-specific techniques can be used for a specific architecture and require training the model using a dataset. Intrinsic methods are by definition model-specific. On the contrary, model-agnostic methods can be used across many black-box models without considering their inner processing or internal representations and do not require training the model. Post hoc methods are usually model-agnostic.

Post hoc interpretability methods consider interpretability of predictions made by black-box models after they have been built. These can further be categorized into four categories as surrogate models, feature contribution, visualisations, and case-based methods [19, 21]. Figure 6 shows the key model-agnostic methods available in literature [14].

In this work, we use two popular Python libraries, SHAP and LIME, to interpret the output and leverage model explanations.

3. Related Work

There is notable literary evidence apropos the versatile use of machine learning and deep learning algorithms for automated sarcasm detection. In the past, rule-based algorithms were employed initially to detect sarcasm [22]. Later, many researchers [23–29] used ML algorithms to detect sarcasm in textual content. Naive Bayes and fuzzy clustering models were employed by Mukherjee et al. [30] for sarcasm detection in microblogs. The researchers concluded that Naive Bayes models are more effective and relevant than the fuzzy clustering models. Prasad et al. [31] analyzed and compared various ML and DL algorithms to conclude that gradient boost outperforms the other models in terms of accuracy. In 2018, Ren et al. [32] employed contextual information for sarcasm detection on Twitter dataset by utilizing two different context-augmented neural models. They demonstrated that the proposed model performs better than the other SOTA models. In 2019, Kumar and Garg [33] compared various ML techniques like SVM, DT, LR, RF, KNN, and NN for sarcasm detection on Twitter and Reddit datasets. A hybrid deep learning model of soft attention-based bi-LSTM and convolution neural network with GloVe for word embeddings was proposed by Kumar et al. [34]. The results demonstrated that the proposed hybrid outperforms CNN, LSTM, and bi-LSTM. Kumar and Garg [4] reported a study on context-based sarcasm detection on Twitter and Reddit datasets using a variety of ML techniques trained using tf-idf and DL techniques using GloVe embedding.

Recent studies have also been reported on multimodal sarcasm detection. In 2019, Cai et al. [35] used bi-LSTM for detection of sarcasm in multimodal Twitter data. In the same year, Kumar and Garg [6] employed various supervised ML techniques to study context in sarcasm detection in typographic memes and demonstrated that multilayer perceptron is best among all the models. In 2020, a study by Kumar et al. [36] built a feature-rich support vector machine and proposed a multihead attention-based bi-LSTM model for sarcasm detection in Reddit comments. Few studies on sarcasm detection in online multilingual content have also been reported. In 2020, Jain et al. [2] had put forward a hybrid of bi-LSTM with softmax attention and CNN for sarcasm detection in multilingual tweets. In 2021, Farha et al. [37] compared many transformer-based language models like BERT and GPA on Arabic data for sarcasm detection. Faraj et al. [38] proposed a model based on ensemble techniques with an AraBERT pretrained model for sarcasm detection in Arabic text with an accuracy of 78%.

Sarcasm detection in conversations and dialogues has created a great interest with NLP researchers. Ghosh et al. [39] used conditional LSTM and LSTM with sentence-level attention to understand the role of context in social media discussions. Hazarika et al. [40] proposed a CASCADE (a ContextuAl SarCasm DEtector) model which extracted contextual information from online social media discussions on Reddit to detect sarcasm by taking into consideration

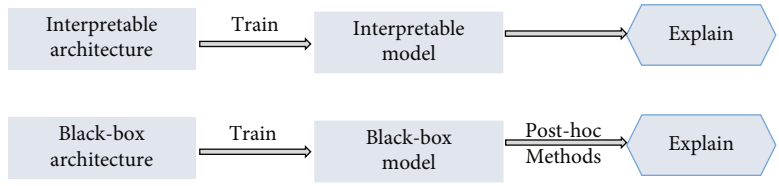


Figure 5: Intrinsic vs. post hoc interpretability.

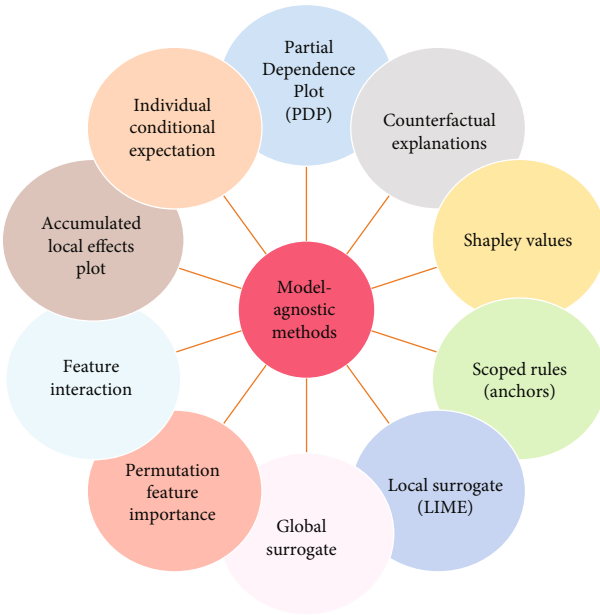


Figure 6: Model-agnostic methods.

stylometric and personality features of the users and trained a CNN for content-based feature extraction. Castro et al. [5] proposed the MUStARD dataset which contains audio-visual data from popular sitcoms and showed how multimodal cues enhance the primary sarcasm classification task. In 2020, Baruah et al. [41] implemented BERT, bi-LSTM, and SVM classifiers for sarcasm detection utilizing the context of conversations. Jena et al. [42] performed the task of sarcasm detection in conversations using a C-Net model which comprised BERT models. Recently, Zhang et al. [43] proposed a model based on quantum theory and fuzzy logic to detect sarcasm in conversations in MUStARD and Reddit datasets.

The use of explainable AI for interpretability of the underlying ML techniques for sarcasm detection has been studied by few researchers. In 2018, researchers Tay et al. [44] improved the interpretability of the algorithms by employing multidimensional intra-attention mechanisms in their proposed attention-based neural model. The proposed model was validated on various benchmark datasets of Twitter and Reddit and compared with other baseline models. Akula et al. [45] focused on detecting sarcasm in texts from online discussion forums of Twitter, dialogues, and Reddit datasets by employing BERT for multihead self-attention and gated recurrent units, to develop an interpretable DL model as self-attention is inherently interpretable.

4. XAI for Sarcasm Detection in Dialogue Dataset

Black-box ML models have observable input-output relationships but lack transparency around inner workings. This is typical of deep-learning and boosted/random forest models which model very complex problems with high nonlinearity and interactions between inputs. It is important to decompose the model into interpretable components and simplify the model's decision making for humans. In this research, we use XAI to provide insights into the decision points and feature importance used to make a prediction about sarcastic disposition of conversations. The architectural flow of the research undertaken in this paper is shown in Figure 7.

The MUStARD dataset used for this research consists of 690 dialogues by the speakers from four famous television shows. It is publicly available and manually annotated for sarcasm. The dataset consists of details about the speaker, utterance, context, context speakers, and sarcasm. For example, the dataset entry for a conversational scene as given in Figure 8 from Friends, season 2, episode 20, is shown in Table 1.

It is noted that most of the dialogues in this dataset are from two most popular shows, namely, the Big Bang Theory and Friends. The data is balanced with an equal number of sarcastic and nonsarcastic dialogues. Figure 9 shows the dataset distribution for the respective TV shows.

The following subsections discuss the details.

4.1. Supervised Machine Learning for Sarcasm Detection in Dialogues. The data was cleaned as the dialogues obtained had some errors in spelling, emoticons, and unnecessary brackets and names of the subtitle providers; any column which had any missing values or wrong data was removed from the dataset. The evaluation of an utterance relies strongly on its context. The contextual interaction between associated chronological dialogues is based on conversational common ground and thereby raising it to prominence in the current context as shown in Figure 10.

Therefore, we use the punch-line utterance, its accompanied context, and the sarcastic/nonsarcastic label to train our model. tf-idf vectorization [46] is done to transform the textual features into representation of numbers. The data is trained using an ensemble learning approach, eXtreme Gradient Boosting (XGBoost). As a popular implementation of gradient tree boosting, XGBoost provides superior classification performance in many ML challenges. In gradient

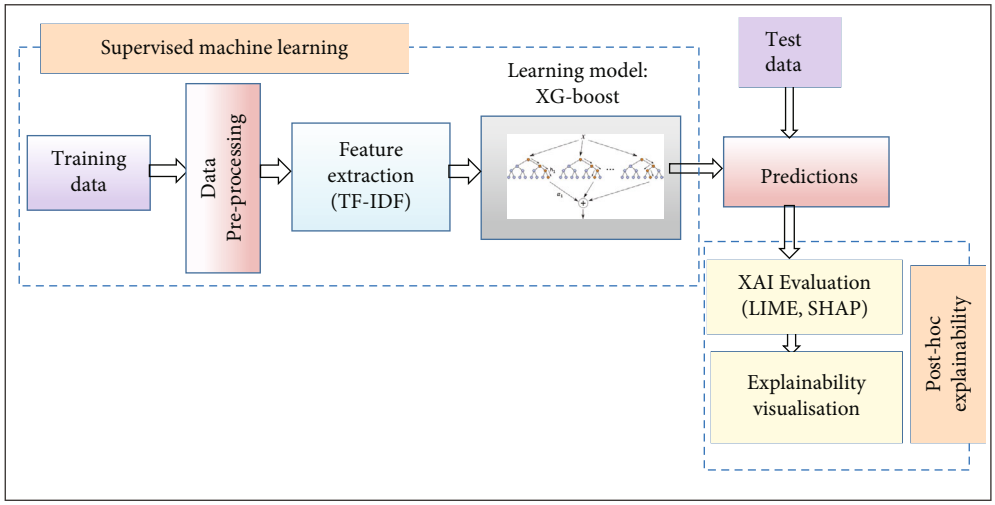


Figure 7: The architectural flow of the research undertaken.

Figure 8: Friends, season 2, episode 20.

boosting, a shallow and weak tree is first trained and then the next tree is trained based on the errors of the first tree. The process continues with a new tree being sequentially added to the ensemble, and the new successive tree improves on the errors of the ensemble of preceding trees. The key advantages of using XGBoost are that it is highly flexible, leverages the power of parallel processing, supports regularization, handles missing values, allows tree pruning, and has built-in cross-validation and high computational speed. On the flip side, explaining the XGBoost predictions seems hard and powerful tools are required for confidently interpreting tree models such as XGBoost. Subsequently, we discuss the two model-agnostic methods selected for seeking explanations that justify and rationalize the black-box model of XGBoost for sarcasm detection in dialogues.

4.2. Post Hoc Explainability Models for Sarcasm Detection in Dialogues. Post hoc interpretability approaches propose to generate explanations for the output of a trained classifier in a step distinct from the prediction step. These approximate the behaviour of a black box by extracting relationships between feature values and the predictions. Two widely accepted categories of post hoc approaches are surrogate models and counterfactual explanations [14]. Surrogate

Table 1: Dataset entry for the given conversational scene from Friends.

Utterance	An utterance is a unit of speech bound by breaths or pauses. The dialogues are spoken by the speaker with respect to the context in the scene.	But younger than some buildings!
Speaker	The character of the series who is giving the dialogue delivery.	Chandler
Context speakers	The side characters to whom the dialogue is being uttered by the main character of that scene.	Chandler
Context	The reason or the scene on the series which led to the dialogue utterance by the speaker.	I know Richard's really nice and everything, it's just that we do not know him really well you know, plus he is old *(Monica glares)* -er than some people.
Show	Name of the show	Friends
Sarcasm	This is the feature in the data to show whether the dialogue utterance by the speaker is sarcastic or nonsarcastic utterance is given as true and nonsarcastic comment is given as false in the dataset.	True

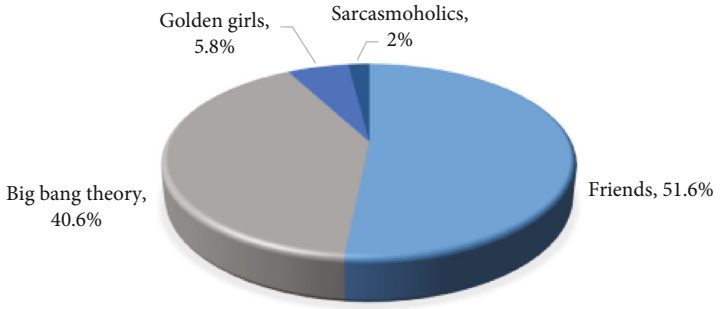


Figure 9: Dataset distribution for TV shows.

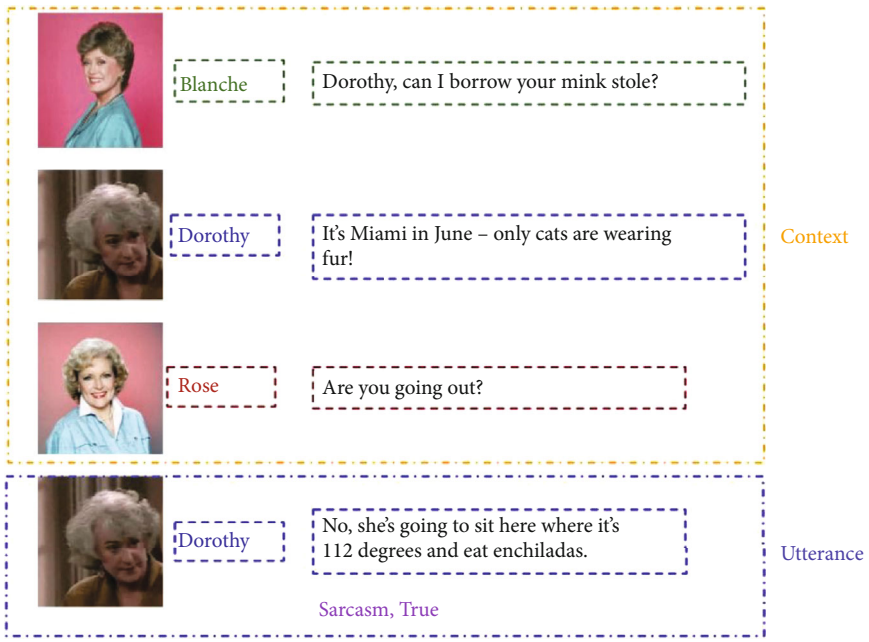


Figure 10: Utterance and associated context from a scene in The Golden Girls.

model approaches are aimed at fitting a surrogate model to imitate the behaviour of the classifier while facilitating the extraction of explanations. Often, the surrogate model is a simpler version of the original classifier. Global surrogates are aimed at replicating the behaviour of the classifier in its entirety. On the other hand, local surrogate models are

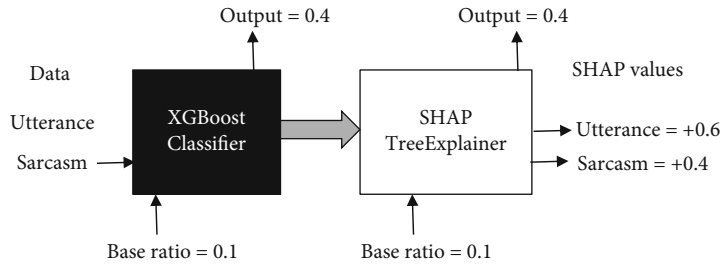


Figure 11: SHAP model architecture.

Table 2: Performance results using utterance + context.

Learning models	Accuracy	Precision	Recall	F-1 score
XGBoost	0. 931	0.965	0.887	0.924
Random forest	0.586	0.402	0.637	0.492
SVM [5]	—	0.579	0.545	0.541

Table 3: Performance results using only utterance.

Learning models	Accuracy	Precision	Recall	F-1 score
XGBoost	0.879	0.852	0.918	0.883
Random forest	0.547	0.369	0.579	0.405
SVM [5]	—	0.609	0.596	0.598

trained to focus on a specific part of the rationale of the trained classifier. In this research, we use two different post hoc local surrogate explainability methods, namely, Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP). The methods fundamentally differ in terms of the interpretability technique used to explain the working of the black-box model. Typically, LIME creates a new dataset or text from the original text by randomly removing words from the original test and gives the probability to each word to eventually predict based on the calculated probability. SHAP, on the other hand, does not create a separate dataset but uses Shapley values to explain the prediction of any input by computing the contribution of each feature for prediction.

4.2.1. LIME. LIME is available as an open-source Python package. It is a local surrogate approach that specifies the importance of each feature to an individual prediction. LIME does not work on the training data; in fact, it gives the prediction by testing it with variations of the data. It trains a linear model to approximate the local decision boundary for that instance, which then generates a new dataset consisting of all the permutation samples along with their corresponding predictions. New data is created by randomly removing words from the original data. The dataset is represented with binary features for each word. A feature is set to 1 if the corresponding word is included and 0 if it is not included. The new dataset of the LIME then trains the interpretable model, i.e., the RF model which is then weighted by the proximity of the sampled instances to the instance of interest. The learned model should be able to give the general idea of the machine learning model prediction locally, but it may not be a good global approximation. The generic steps of LIME include sampling of instances followed by training the surrogate model using these instances to finally generate the final explanation given to the user through a visual interface provided with the package. Mathematically, LIME explanations are determined using

$$\text{explanation}(x) = \text{argarg min } g \in G\, L(f, g, \pi_x) + \Omega_g. \quad (1)$$

According to the mathematical formula, the explanation model for instance x is the ML model (random forest, in our case) which then minimises the loss L, such as mean square error (MSE). This L measures the closeness of the explanation to the prediction of the original model f, while keeping the model complexity $\Omega(g)$ low. G is the pool of possible explanation, and π_x is the proximity measure of how large the neighborhood is around the instance x. LIME optimizes only the loss part of the data.

The idea for training the LIME model is simple:

(i) Select the instance which the user wants to have explanation of the black-box prediction

(ii) Add a small noisy shift to the dataset and get the black-box prediction of these new points

(iii) Weight the new point samples according to the proximity of the instance x

(iv) Weighted, interpretable models are trained on the dataset with the variations

(v) With the interpretable local model, the prediction is explained

4.2.2. SHAP. SHAP is aimed at explaining individual explanations based on the cooperative game theory Shapley values. Shapley values are used for the prediction to be explained by the assumption of each feature value of the instance as a "player." These values tell the user how fairly the distribution is among the "players" in game. The Shapley value is the average marginal contribution of a feature value across all possible coalitions. The reason to choose SHAP as our second explainable model was because SHAP computes the contribution of each feature of the prediction. These features act as "players" which will then be used to see if the payoff of the distribution is fair or not. It needs to satisfy the local accuracy, missingness, and consistency properties making predictions [17].

Utterance +context

	Predicted 0	Predicted 1
Actual 0	TN = 309	FP = 36
Actual 1	FN = 11	TP = 334

Only utterance

	Predicted 0	Predicted 1
Actual 0	TN = 290	FP = 28
Actual 1	FN = 55	TP = 317

Figure 12: Confusion matrix of XGBoost on MUStARD dataset.

SHAP explains the output of the black-box model by showing the working of the model to explain the prediction of an instance computing each feature's contribution to the prediction. As given in (2), mathematically, SHAP specifies explanation of each prediction as it gives out the local accuracy of the represented features

$$g\left(z'\right) = \varnothing_0 + \sum_{j=1}^{M} \phi_j z'_j, \tag{2}$$

where g is the explanation model and $z' \epsilon \{0,1\}^M$ is the coalition vector in the dataset. M denotes the maximum size of the coalition in SHAP where entry 1 represents that the feature is present and 0 represents that the feature is absent.

SHAP basically follows three properties for the result, and those properties are as follows:

(i) *Local Accuracy*. Local accuracy means that the explanation model should match the original model as given in

$$f(x) = g\left(x'\right) = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j \tag{3}$$

(ii) *Missingness*. Missing feature gets the attribution score of 0 where 0 represents the absence of the feature. It means that the simplified input feature and the original input feature should be the same so that it does not have any impact. It is given as shown in

$$x'_j = 0 \Longrightarrow \phi_j = 0 \tag{4}$$

(iii) *Consistency*. Consistency means that the values increase or remain the same according to the marginal contribution of the feature values of the model. It is given by

$$f'_x\left(z'\right) - f'_x\left(z'_j\right) \geq f_x\left(z'\right) - f_x\left(z'_j\right) \tag{5}$$

In the paper, the features which are used for the target prediction and the SHAP value for the contribution of that feature are the difference between the actual prediction and the mean prediction. SHAP provides both local and global interpretability by calculating SHAP values on the local level for feature importance and then providing a global feature importance by summing the absolute SHAP values for each of the individual predictions. The SHAP model architecture is shown in Figure 11.

We use KernelSHAP (https://docs.seldon.io/projects/alibi/en/stable/methods/KernelSHAP.html) in this work for the estimation of the instance x of each feature contribution. KernelSHAP uses weighted local linear regression to estimate the Shapley values for any model.

5. Results and Discussion

We implemented the model using scikit-learn, a framework in Python. The classification performance of XGBoost was evaluated using accuracy, F1 score, precision, and recall as metrics. The training : test split was 70 : 30. The model is trained with default parameters using the Python XGBoost package. The performance of XGBoost was compared with another ensemble learning method—random forest and superior results were observed using XGBoost. Also, the primary goal of this research was to investigate the role and importance of context we trained and tested the model with and without context. A comparison with the existing work [5] that uses support vector machines (SVM) as the primary baseline for sarcasm classification in speaker-independent textual modality is also done. The results obtained using the punch-line utterance and its associated context are shown in Table 2 whereas the results obtained using only the punch-line utterance that is without using context as a feature are shown in Table 3.

It is evident from the results that sarcastic intent of the thread is more efficiently captured using context, improving the accuracy by nearly 5%. The confusion matrix for the XGBoost classifier with and without context is shown in Figure 12. To compute the confusion matrices, we take a count of four values as follows:

(i) *True Positives* (TP): number of sarcastic utterance correctly identified

(ii) *False Positives* (FP): number of nonsarcastic utterance that was incorrectly identified as sarcastic utterance

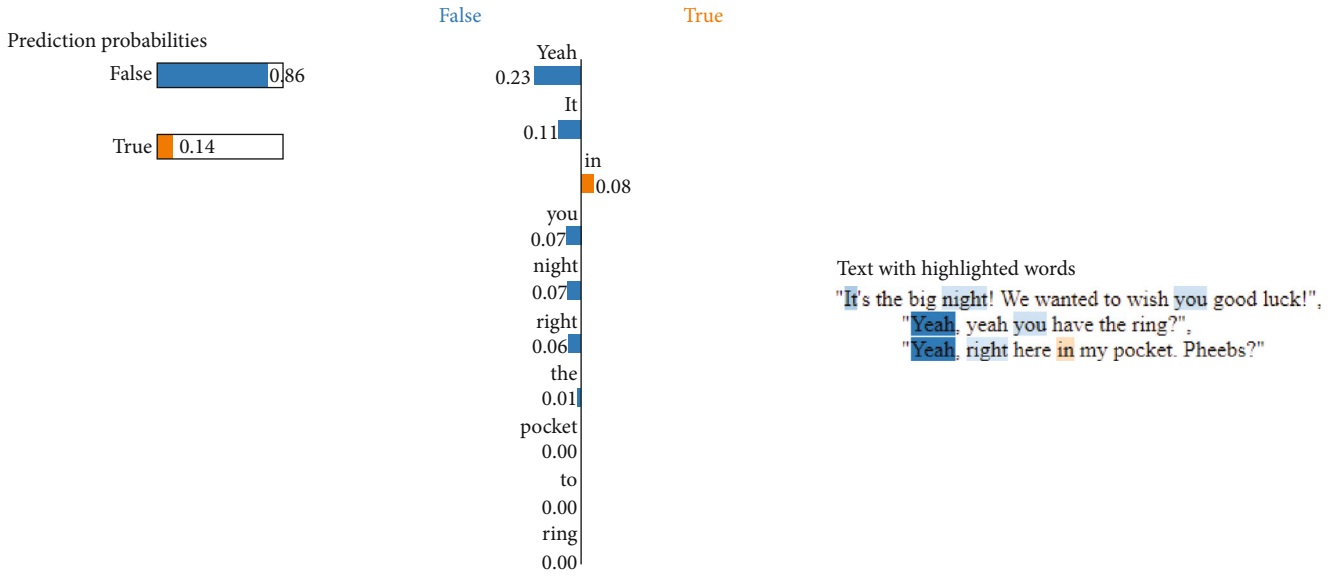


Figure 13: LIME visualisation.

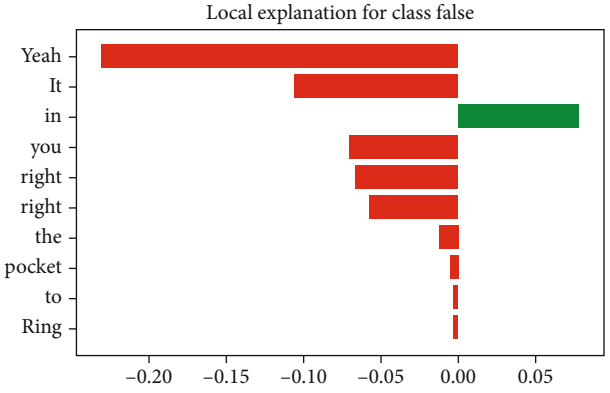


Figure 14: Local explanation for false class.

(iii) *False Negatives* (FN): number of sarcastic utterance that was incorrectly identified as nonsarcastic utterance

(iv) *True Negatives* (TN): number of nonsarcastic utterance correctly identified

The objective was not only to produce higher results but also to produce a better analysis. Therefore, after the evaluation of the learning algorithm, explainable models of LIME and SHAP were used for prediction interpretability. LIME text classifier and LIME text explainer were used to obtain the explanation model for LIME. The class names were set to true and false according to the label, for the LIME text explainer with random state of 42. For SHAP, it was trained and tested on the training and testing vectors generated by tf-idf vectors with 200 background samples to generate the force plot and summary plot of the XGBoost using utterance and context as features.

The explanation model for LIME and SHAP shows which words in the dialogues of the characters influence the model

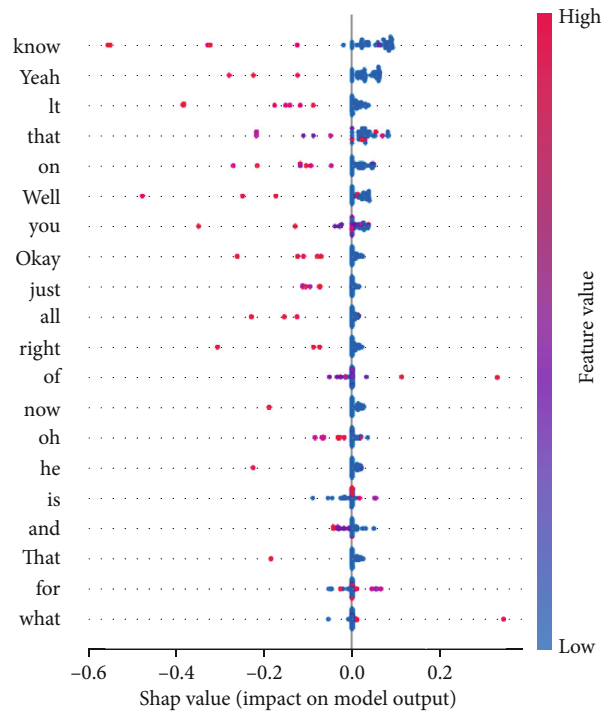


Figure 15: SHAP summary plot.

to label the utterance as sarcastic or not. The explainability scores from each of the methods are generated for every feature in the dataset. Evidently, for an utterance with sarcasm, certain words receive more importance than others. Figure 13 shows the LIME visualisation, where it can be observed that only some parts of the dialogue (taken arbitrarily) are being used to determine the probability of the sarcasm of the utterance by the speaker. As we randomly select an utterance in the test set, it happens to be an utterance that is labelled as nonsarcastic, and our model predicts it as

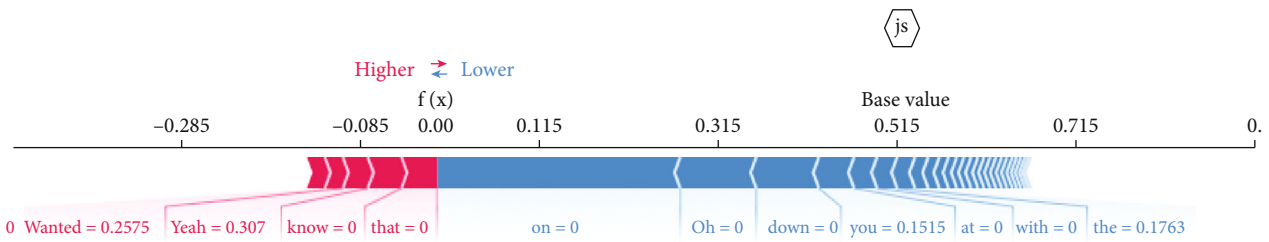


Figure 16: SHAP force plot.

nonsarcastic as well. Using this utterance, we generate explanations.

Noticeably, for this conversation, word "*Yeah*" has the highest negative score for class sarcasm and our model predicts this conversation should be labelled as nonsarcastic with the probability of 86%.

Figure 14 shows how the weights are trained, and the weights of each word given in the utterance are used to determine the sarcasm of the utterance by the speaker.

The same goes for the visualisation of the SHAP model as given in Figure 15, which helps the user understand how the model is making the decision for detecting sarcasm in dialogues. It is using each and every word as a "player" and giving the coalition of whether the model can equally pay off or not. This is a very helpful view that shows at a global level in which direction each feature contributes as compared to the average model prediction. The y-axis in the right side indicates the respective feature value being low vs. high. Each dot represents 1 instance in the data, and the cluster of dots indicates there are many instances in the data with that particular SHAP value.

Thus, the SHAP summary plot combines the feature effect with the importance of the feature. In the SHAP summary plot, each point is the Shapley value for a feature and an instance. The y-axis and x-axis in the summary plot show the feature and the Shapley values, respectively. The colors in the summary plot indicate the impact of the feature from high to low, and the overlapping points in the plot show the distribution of the Shapley values per feature.

Another way to understand the explainability of the utterance using SHAP can be done using the force plot of the data. A force plot helps visualising Shapley values for the features. Feature values in pink cause to increase the prediction. The size of the bar shows the magnitude of the feature's effect. Feature values in blue cause to decrease the prediction. Sum of all feature SHAP values explains why model prediction was different from the baseline. Figure 16 gives the multiprediction force plot used in the given instance with utterance and context for the analysis of the prediction path. Again, the word "*Yeah*" has higher feature importance.

The results support the hypothesis that how each word in the utterance with respect to the context of the dialogues is important for sarcasm detection.

6. Conclusion

With the accelerated use of sentiment technologies in online data streams, companies have integrated it as an enterprise solution for social listening. Sarcasm is one of the key NLP challenges to sentiment analysis accuracy. Context incongruity can be used to detect sarcasm in conversational threads and dialogues where the chronological statements formulate the context of the target utterance. We used an ensemble learning method to detect sarcasm in benchmark sitcom dialogue dataset. Results clearly establish the influence of using context with the punch-line utterance as features to train XGBoost. Further, the predictions given by the black-box XGBoost are explained using LIME and SHAP for local interpretations. These post hoc interpretability methods demonstrate that how few words unambiguously contribute to the decision and word importance is the key to accurate prediction of the sarcastic dialogues. As a future work, we would like to evaluate other XAI methods such as PDP for the detection of sarcasm. Also, temporal context and span analysis for context incongruity are another promising line of work. Gauging other rhetorical literary devices in online data streams is also an open domain of research. Auditory cues such as tone of the speaker and other acoustic markers such as voice pitch, frequency, empathetic stress and pauses, and visual cues for facial expressions that can assist sarcasm detection in audio-visual modalities need further investigation.

Data Availability

Publicly accessible data has been used by the authors.

Additional Points

Code Availability. Can be made available on request.

Ethical Approval

The work conducted is not plagiarized. No one has been harmed in this work.

Conflicts of Interest

The authors certify that there is no conflict of interest in the subject matter discussed in this manuscript.

Authors' Contributions

All the authors have equally contributed in the manuscript preparation. All the authors have given consent to submit the manuscript. The authors provide their consent for the publication.

References

[1] N. Majumder, S. Poria, H. Peng et al., "Sentiment and sarcasm classification with multitask learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, 2019.

[2] D. Jain, A. Kumar, and G. Garg, "Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN," *Applied Soft Computing*, vol. 91, article 106198, 2020.

[3] D. Ghosh, W. Guo, and S. Muresan, "Sarcastic or not: word embeddings to predict the literal or sarcastic meaning of words," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1003–1012, Lisbon, Portugal, 2015.

[4] A. Kumar and G. Garg, "Empirical study of shallow and deep learning models for sarcasm detection using context in benchmark datasets," *Journal of Ambient Intelligence and Humanized Computing*, 2019.

[5] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629, Florence, Italy, 2019.

[6] A. Kumar and G. Garg, "Sarc-m: sarcasm detection in typographic memes," in *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019*, Uttaranchal University, Dehradun, India, 2019.

[7] A. Kumar and G. Garg, "The multifaceted concept of context in sentiment analysis," in *Cognitive Informatics and Soft Computing*, P. Mallick, V. Balas, A. Bhoi, and G. S. Chae, Eds., vol. 1040 of Advances in Intelligent Systems and Computing, pp. 413–421, Springer, Singapore, 2020.

[8] D. Gunning, *Explainable artificial intelligence (XAI)*, Defense Advanced Research Projects Agency (DARPA), 2017, https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf.

[9] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model," *Complexity*, vol. 2021, Article ID 6634811, 11 pages, 2021.

[10] C. Li, X. Zheng, Z. Yang, and L. Kuang, "Predicting short-term electricity demand by combining the advantages of ARMA and XGboost in fog computing environment," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 5018053, 18 pages, 2018.

[11] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: a survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 210–215, Opatija, Croatia, 2018.

[12] T. Miller, "Explanation in artificial intelligence: insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[13] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 5, article e1379, 2020.

[14] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: a survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": explaining the predictions of any classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 1135–1144, San Diego, California, 2016.

[16] C. Meske and E. Bunde, "Transparency and trust in human-AI-interaction: the role of model-agnostic explanations in computer vision-based decision support," in *International Conference on Human-Computer Interaction, HCII 2020*, H. Degen and L. Reinerman-Jones, Eds., vol. 12217 of Lecture Notes in Computer Science, pp. 54–69, Springer, Cham, 2020.

[17] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, https://arxiv.org/abs/1705.07874.

[18] A. Messalas, Y. Kanellopoulos, and C. Makris, "Model-agnostic interpretability with shapley values," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–7, Patras, Greece, 2019.

[19] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE access*, vol. 6, pp. 52138–52160, 2018.

[20] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: an overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89, Turin, Italy, 2018.

[21] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2019.

[22] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 704–714, Seattle, Washington, USA, 2013.

[23] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 757–762, Beijing, China, 2015.

[24] J. Tepperman, D. Traum, and S. Narayanan, ""Yeah right": sarcasm recognition for spoken dialogue systems," in *Ninth international conference on spoken language processing*, Pittsburgh, Pennsylvania, 2006https://www.isca-speech.org/archive/archive_papers/interspeech_2006/i06_1821.pdf.

[25] R. Kreuz and G. Caucci, "Lexical influences on the perception of sarcasm," in *Proceedings of the Workshop on Computational Approaches to Figurative Language - FigLanguages '07*, pp. 1–4, Rochester, New York, 2007.

[26] O. Tsur, D. Davidov, and A. Rappoport, "ICWSM—a great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, 2010.

[27] D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcasm in Twitter and Amazon," in *Proceedings of the fourteenth conference on computational natural language learning*, pp. 107–116, Uppsala, Sweden, 2010.

[28] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in Twitter," *Language Resources and Evaluation*, vol. 47, no. 1, pp. 239–268, 2013.

[29] D. Bamman and N. Smith, "Contextualized sarcasm detection on Twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 9, Oxford, UK, 2015.

[30] S. Mukherjee and P. K. Bala, "Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering," *Technology in Society*, vol. 48, pp. 19–27, 2017.

[31] A. G. Prasad, S. Sanjana, S. M. Bhat, and B. S. Harish, "Sentiment analysis for sarcasm detection on streaming short text data," in *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, pp. 1–5, London, UK, 2017.

[32] Y. Ren, D. Ji, and H. Ren, "Context-augmented convolutional neural networks for Twitter sarcasm detection," *Neurocomputing*, vol. 308, pp. 1–7, 2018.

[33] A. Kumar and G. Garg, "Sarcasm detection using feature-variant learning models," *Proceedings of ICETIT 2019*, P. Singh, B. Panigrahi, N. Suryadevara, S. Sharma, and A. Singh, Eds., , pp. 683–693, Springer, Champ, 2020.

[34] A. Kumar, S. R. Sangwan, A. Arora, A. Nayyar, and M. Abdel-Basset, "Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network," *IEEE Access*, vol. 7, pp. 23319–23328, 2019.

[35] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in Twitter with hierarchical fusion model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2506–2515, Florence, Italy, 2019.

[36] A. Kumar, V. T. Narapareddy, V. Aditya Srikanth, A. Malapati, and L. B. M. Neti, "Sarcasm detection using multi-head attention based bidirectional LSTM," *IEEE Access*, vol. 8, pp. 6388–6397, 2020.

[37] I. A. Farha and W. Magdy, "Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pp. 21–31, Kyiv, Ukraine (Virtual), 2021.

[38] D. Faraj and M. Abdullah, "Sarcasmdet at sarcasm detection task 2021 in arabic using arabert pretrained model," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pp. 345–350, Kyiv, Ukraine (Virtual), 2021.

[39] D. Ghosh, A. R. Fabbri, and S. Muresan, "The role of conversation context for sarcasm detection in online interactions," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 186–196, Saarbrücken, Germany, 2017.

[40] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea, "Cascade: contextual sarcasm detection in online discussion forums," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1837–1848, Santa Fe, New Mexico, USA, 2018.

[41] A. Baruah, K. Das, F. Barbhuiya, and K. Dey, "Context-aware sarcasm detection using BERT," in *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 83–87, Online, 2020.

[42] A. K. Jena, A. Sinha, and R. Agarwal, "C-net: contextual network for sarcasm detection," in *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 61–66, Online, 2020.

[43] Y. Zhang, Y. Liu, Q. Li et al., "CFN: a complex-valued fuzzy network for sarcasm detection in conversations," *IEEE Transactions on Fuzzy Systems*, p. 1, 2021.

[44] Y. Tay, L. A. Tuan, S. C. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1010–1020, Melbourne, Australia, July 2018.

[45] R. Akula and I. Garibay, "Explainable detection of sarcasm in social media," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 34–39, Washington D.C., 2021.

[46] J. Ramos, "Using tf-idf to determine word relevance in document queries," *Proceedings of the First Instructional Conference on Machine Learning*, vol. 242, no. 1, pp. 29–48, 2003.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6697467, 11 pages
https://doi.org/10.1155/2021/6697467

Research Article

A Secure IoT-Based Cloud Platform Selection Using Entropy Distance Approach and Fuzzy Set Theory

Alakananda Chakraborty,[1] **Muskan Jindal**,[1] **Mohammad R. Khosravi**,[2] **Prabhishek Singh**,[1] **Achyut Shankar**,[1] **and Manoj Diwakar**[3]

[1]*Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India*
[2]*Department of Computer Engineering, Persian Gulf University, Iran*
[3]*Graphic Era Deemed to be University Dehradun, India*

Correspondence should be addressed to Prabhishek Singh; prabhisheksingh88@gmail.com

Received 22 December 2020; Revised 9 April 2021; Accepted 4 May 2021; Published 18 May 2021

Academic Editor: Federico Tramarin

With the growing emergence of the Internet connectivity in this era of Gen Z, several IoT solutions have come into existence for exchanging large scale of data securely, backed up by their own unique cloud service providers (CSPs). It has, therefore, generated the need for customers to decide the IoT cloud platform to suit their vivid and volatile demands in terms of attributes like security and privacy of data, performance efficiency, cost optimization, and other individualistic properties as per unique user. In spite of the existence of many software solutions for this decision-making problem, they have been proved to be inadequate considering the distinct attributes unique to individual user. This paper proposes a framework to represent the selection of IoT cloud platform as a MCDM problem, thereby providing a solution of optimal efficacy with a particular focus in user-specific priorities to create a unique solution for volatile user demands and agile market trends and needs using optimized distance-based approach (DBA) aided by Fuzzy Set Theory.

1. Introduction

One of the greatest inventions of Gen Z, Internet has rapidly emerged over the last two decades, connecting people and organizations together into one giant family. This connectivity has generated the urgency of Internet of Things (IoT) [1], which involves sensors, software devices, and other technologies, for the purpose of maintaining the security and privacy of humongous data transmission among other devices and systems [2]. For this sole purpose, several distinct IoT platforms have come into existence with their own unique cloud service providers (CSP) at the backend. But, like every coin has two sides, i.e., it has also led to a problematic situation when it comes to the selection of ideal CSP for a selected set of attributes that purview a finite set of requirements, to assist the process of decision-making where one has to deliberate multiple attributes, possible scenarios, market trends, and user biases [3].

According to the best of research and knowledge and observation, no compatible and comprehensive study and solutions been done for this integrated set of requirements in the field of cloud service provider selection (CSPS). However, there exist a lot of work that includes some of our set of factors quality which elaborately and accurately formulates algorithm for some quality factors and some for the technical aspects [4, 5]. Divergent from the preexisting schemes, thus, providing a flexible, realistic, and compatible methodology towards cloud service selection (CSS) considering all the possible factors under the sun required for an ideal cloud service.

1.1. Significance of the Research. In spite of the existence of many software solutions for this decision-making problem, they have been proved to be inadequate considering the distinct attributes unique to individual user.

1.1.1. Research Gaps in the IoT-Based Cloud Service Providers. The requirement of decision-making among the various cloud service providers amalgamated with IoT applications has led to the emergence of several software solutions in

recent years. However, multiple demerits can be observed in the performance and efficiency of these platforms [6]. The current short comings present in the IoT-based cloud service provider solutions involve numerous dimensions including the inefficiency of the platforms to extend for supporting the heterogeneous sensing technologies. Other demerits include the proprietorship of data, providing insinuations of privacy and security [7]. The processing and sharing of information can also be counted as another gap especially in scenarios where it is essential to support novel services. The absence of assistance provided by application developers is another shortcoming faced by several IoT cloud platforms [4]. Furthermore, most of these IoT platforms do not possess the property of expansion for the addition of new components to withstand the emergence of new technologies and provide economies of scale. Lastly, the delivery of the purchased software to the respective connected devices is also not supported by a majority of the marketplaces dedicated for IoT applications.

Multicriteria Decision-Making (MCDM) techniques provide a scientific and easy solution. MCDM deals with organizing variegated attributes which come under the purview of decision-making. It specializes in handling issues where the proximity of attributes is close, and human cognitive abilities are not able to take the logical decisions. It does so by performing bargains or trade-offs by replacing one criterion by equivalent another. This paper presents an integrated set of factors that contribute the solution to the problematic issue of the selection of an optimal IoT cloud platform.

In a nutshell, the qualities mentioned below make the proposed methodology novel when compared to state-of-the-art techniques:

(1) Identification and categorization of selection attributes (SA): after thorough and detailed studying of more than fifty research papers, few factors, i.e., selection attributes (SA) were filtered out. About 90 factors were carefully studied, and explicitly observed and relevant factors were mined out by removing redundant elements and those which were similar to each other. Finally, these factors were then categorized into broad three categories after extensive reasoning and filtration, namely,

(i) Quality factors

(ii) Technical factors

(iii) Economic factors

(2) Prioritization of identified selection factors (SA): the identified selection attributes (SA) are then prioritized according to Fuzzy Set Theory. Here, prioritization is done on this basis of calculated priority weights of each selection attribute individually

(3) Development of a hybrid decision-making framework: once the selection attributes are successfully selected and filtered out and then weighed, respectively, then the hybrid decision-making framework is developed using mainly two methodologies:

(i) Fuzzy Set Theory

(ii) Matrix Multicriteria Decision-Making Method

2. Literature Review

This section of the research is concerned with the existing studies in the field of selecting optimal cloud computing service provider for IoT-based applications, where the problem of service selection has been represented as an MCDM problem. To search the relevant data, the keywords like Cloud Computing for IoT, Cloud Platforms for IoT Services, IoT based Cloud Service Selection, IoT Service Selection Attributes, and Cloud Service Selection for IoT were used. As a result of this research, a total of 104 research papers from various highly reputed journals and conferences were analyzed in detail. Now, these papers were screened by examining their primary focus, whether it is related to the cloud service selection or not. Then, in the second screening, the approaches used and the case studies mentioned along with the selection attributes (SA) which were mentioned in these research papers were deliberate to make a comparative study of the same. The comprehensive tabular literature survey is shown in Table 1.

This paper presents and develops a hybrid decision-making framework using two methodologies, namely, Fuzzy Set Theory and Matrix Multicriteria Decision-Making (MMCDM) where identification and categorization of 14 selection attribute are prepared into three categories, namely, quality factors, technical factors, and economic factors. After removing redundant features and filtering unnecessary information, thereby making this framework relatively less vulnerable to prerequisites and limitations as compared to available frameworks and techniques in cloud computing service selection.

3. Methodology

3.1. Security and Privacy Challenges in Cloud-Based IoT Platforms. While IoT and its applications are well explored and secure, the cloud-based IoT platforms are still comparatively less explored and nascent in nature [18]. Categorized in two purviews, static and mobile-based platforms both have variegated challenges on grounds of security and privacy. There are multiple security challenges including identity privacy that deals with protection of details of user of the cloud devices like his/her personal real-world information. Other threats include disclosure of the real-time location of user termed as location privacy [19]. Node compromising attack is also one of the most enduring threats to user's privacy as it includes planned attacked to gain access to user's private information [20]. Removal or addition of transmission multiple layers is a very mundane breach performed by various IoT users; it involves manipulating the concept of reward

Table 1

Citation/name	Methodology	Advantages	Disadvantages
[8]	This study proposes a multistep approach to evaluate, categorize, and rate cloud-based IoT platforms via implementing Multicriteria Decision-Making (MCDM), probabilistic linguistic term sets (PLTSs), and finally, a probabilistic linguistic best-worst (PLBW) is used to score all platforms	Though the proposed method seems complex but a real-time implementation via case study provides cogent proof of its efficiency. It also outperforms individual scoring, classification, and evaluating methods.	The data used in the case study is limited which explains the flow of the method but falls back to prove its cogency. Moreover, inclusion of latest hybrid techniques in the domain for comparative analysis could further edify the study's significance.
[9]	Cloud service provider selection approach is proposed via application of Multicriteria Decision-Making (MCDM), analytical hierarchical process (AHP), technique for order of preference by similarity to ideal solution (TOPSIS), and the best-worst method (BWM). Case study is presented to support the same.	The study successfully identifies and provides solutions the drawbacks of classical multicriteria decision-making (MCDM) approaches in terms of accuracy, time required, and complexity of computation. AHP is outperformed by proposed approach.	The use case scenario presented used stimulated scenarios and data that raises question against the cogency of the proposed study.
[10]	Additive manufacturing based cloud-based service providing framework is proposed to include both hard and soft services for the ease of customer use. These include data-based testing, design, 3D printing, remote control of printers, and face recognition using AI.	This study understands and provides solution to the real-time consumer or customer problems. Its feature providing framework proves to be easy, feasible, and effective.	The study only provides a framework along with it merit without any details of implementing or developing the framework for real world application.
[11]	The study is aimed at identifying various determinants that cause deprecation of various ministry of micro, small, and medium enterprises in India, contributing a huge impact on Indian economy. Data is collated from 500 Indian MSMEs. Multiple criteria include social influence, Internet of Things, perceived ease of use, trust, and perceived IT security risk among others.	This study evaluates real-time data from 500 MSMEs that proves its cogency. Moreover, it provides insight that can be directly deliberated by policy makers to create maximum impact.	A comparative analysis with other policy-making insight provider algorithms along with impact of implementing the recommended changes would create more clarity and value for the research.
[12]	A comparative analysis is performed to obtain the best cloud-based IoT platform for any business or organization by deliberating multiple criteria, functional and nonfunctional requirements among five giants, namely, Azure, AWS, SaS, ThingWorx, and Kaa IoT by application various techniques like analytical hierarchical process (AHP), K-means clustering, and statistical tests.	The hierarchal method of requirement classification provides edge to the method and various statistical tests implemented on the results obtained creates increased sense of cogency or significance to the study.	The cloud-based IoT platforms are limited creating false sense of performance in terms of evaluating more than 5 platforms. Moreover, requirement classification into hierarchy is very time and effort intensive.
[13]	IoT applications built via cloud-based platforms are assayed for any kind of security challenge or data inconsistency issues that arise due to third party auditors, phishing attacks. It also provides strategies to prevent the same.	The objective of the study is very relevant to the need of the hour, providing valid and much needed information. It also provides recommendations handle the same.	The scope of study is limited to theoretical analysis without any real data implementation or case study to prove the cogency of the points mentioned in the paper.

Table 1: Continued.

Citation/name	Methodology	Advantages	Disadvantages
[14]	This study poses to create a need for authorization in cloud-enabled IoT systems by assaying various security threats that such a set up encounters via two case studies in order. Proposing control-based authorization system	The aim of the study very cogent and current, deliberating recent developments of cloud-based IoT applications. The case studies presented aid to the cause of study while contributing to the significance of the proposed framework.	The framework proposed for control-based authorization lacks any sense of implementation or efforts towards prototype development.
[15]	An attack distribution detector is proposed to prevent malfunctioning of trust bounders in IoT-based applications, leading to severe data theft. A downsampler-encoder-based cooperative data generator is proposed to discriminate noisy data that may lead of data theft that malfunction trust boundaries.	The continuous updating and verification of the model provides it optimal results and performance to detection probable data thefts. The model outperforms primordial machine learning and deep learning techniques.	Inclusion of latest hybrid techniques in the domain for comparative analysis could further edify the study's significance.
[16]	Various cogent issues with IoT middleware are brought to attention while proposing a state of art IoT middleware that can integrate with MQTT, CoAP, and HTTP as application-layer protocols.	The problem addressed by In.IoT framework is cogent, and its relevance has been shown very accurately in the study.	A comparative analysis with classical middleware and latest hybrid techniques could further edify the significance of the study.
[17]	An intrusion detection technique for cloud-based IoT application is proposed by implementing machine learning, to obtain state of art accuracy and in-depth analysis of source or type of intrusion.	The survey of 95 developments in intelligence-based intrusion detection techniques provides the study significant relevance and ground for comparative analysis with proposed technique.	Though the study shows optimal accuracy, false-positive results still hamper its cogency.

distribution when transmitting. Malicious IoT users add or remove the distribution layers to hamper the cycle or amount of reward distributed [25].

3.2. Distance-Based Approach. Distance-based approach (DBA) is an effective and efficient MCDM method. Identifying and defining the optimized state of the multiple attributes that are part of the process is the initial gradation in the proposed method. The optimal state represented by the vector OP is the set of best values of criteria over a range of alternatives. The best values can be maximum or minimum, defining the type of criteria.

Reference to Figure 1, as indicated, vector "OP" is the optimal point in a multidimensional space. It acts as a reference point to which the other values of all the alternatives are analyzed to one another quantitatively. In other words, an arithmetical difference of the current values of alternatives from their corresponding optimal values is taken, which represents the ability of the considered alternatives to achieve the optimal state. The decision-making issue which needs to be dealt with is searching for a viable solution on basis of its proximity to the optimal state.

In Figure 1, "H" represents the feasible region and "Alt" as the alternative. The distance-based technique is aimed at determining a point in the "H" region and is in closest proximity to the optimal point.

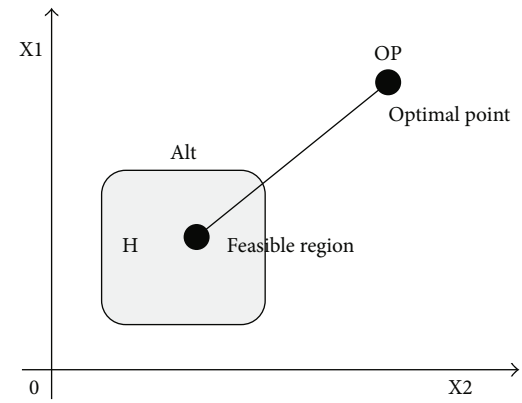


Figure 1: Graphical Representation of DBA Methodology.

To implement the above approach, let $i = 1, 2, 3, 4\ldots$ $n =$ alternatives, and $j = 1, 2, 3, 4\ldots$ $m =$ selection attributes. A matrix is created to represent the entire set of alternatives along with their respective criteria, which is shown in (1).

$$[d] = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix}. \quad (1)$$

This matrix is known as the decision matrix $[d]$. Now, we take the priority weights of these attributes according to the opinions of various experts and calculate their averages. We take the sum of these averages and divide the sum by each of these averages. The result is the creation of another matrix with only one row and columns equal to the number of attributes. This matrix is known as priority weights matrix [PW] as shown in (2).

$$[\mathrm{PW}] = [\mathrm{PW}_{11}\mathrm{PW}_{12} \cdots \mathrm{PW}_{1m}]. \tag{2}$$

Using the following Equations (3), (4) and (5), the decision matrix is standardized to minimize the impact of different units of measurement and to simplify the process.

$$\overline{d_j} = \frac{1}{n} \sum_{i=1}^{n} d_{ij}, \tag{3}$$

where $\overline{d_j}$ is the average of each attribute for all alternatives.

$$S_j = \left[\frac{1}{n}\sum_{i=1}^{n}\left(d_{ij} - \overline{d_j}\right)^2\right]^{1/2}, \tag{4}$$

where S_j is the standard deviation of each attribute for all alternatives.

$$d'_{ij} = \frac{d_{ij} - \overline{d_j}}{S_j}, \tag{5}$$

where d_{ij} is the value of each attribute for an alternative and d'_{ij} is the standardized value of each attribute for an alternative.

The final matrix is known as the standardized matrix $[d']$ and is represented as in (6).

$$\left[d'\right] = \begin{bmatrix} d'_{11} & d'_{12} & \cdots & d'_{1m} \\ d'_{21} & d'_{22} & \cdots & d'_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d'_{n1} & d'_{n2} & \cdots & d'_{nm} \end{bmatrix}. \tag{6}$$

The best value of each attribute is selected over the set of alternatives. The best values can be maximum or minimum values, depending on the type of attribute specified. The matrix formed using this set of values is known as the optimal matrix $[O]$ as shown in (7).

$$[O] = [O_{11}O_{12} \cdots O_{1m}]. \tag{7}$$

The distance of each of the alternatives from its optimal state is calculated as the numerical difference between the values of each of the attributes and their corresponding optimum counterparts. The resulting values form a matrix called the distance matrix $[O']$ as represented in (8).

$$\left[O'\right] = \begin{bmatrix} O_{11} - d'_{11} & O_{12} - d'_{12} & \cdots & O_{1m} - d'_{1m} \\ O_{11} - d'_{21} & O_{12} - d'_{22} & \cdots & O_{1m} - d'_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ O_{11} - d'_{n1} & O_{12} - d'_{n2} & \cdots & O_{1m} - d'_{nm} \end{bmatrix}. \tag{8}$$

Each value of this matrix is then squared and multiplied by their corresponding priority weights, as explained by Equation (9).

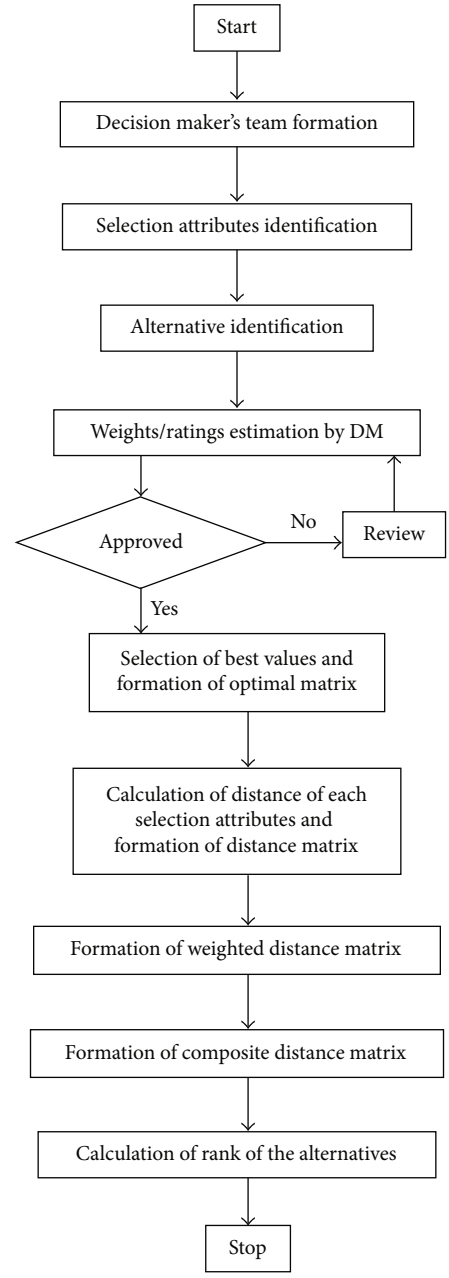


Figure 2: Model development of DBA methodology.

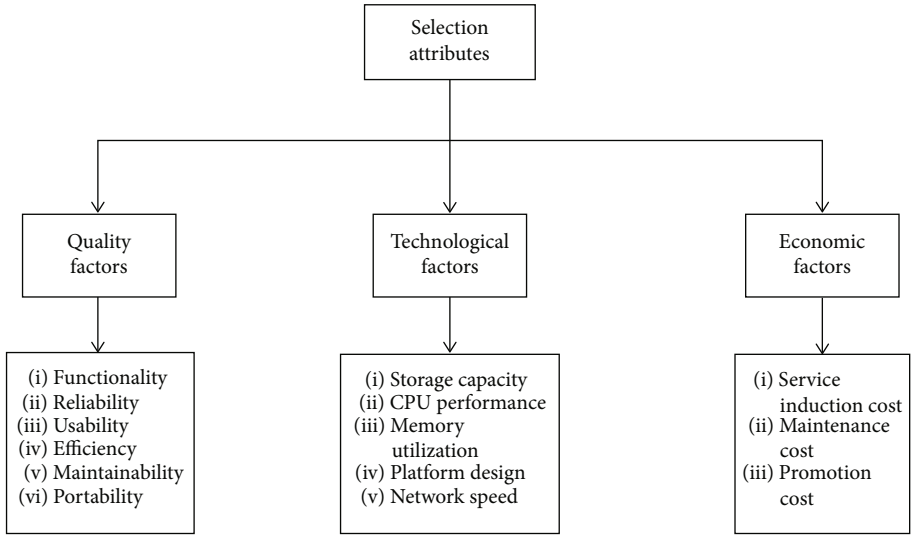


Figure 3: Classification of selection attributes.

$$W_{ij} = O'^{2}_{ij} \times PW_j. \tag{9}$$

The resulting matrix is called the weighted distance matrix $[W]$ as shown in (10).

$$[W] = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1m} \\ W_{21} & W_{22} & \cdots & W_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nm} \end{bmatrix}. \tag{10}$$

Equation (11) is used to calculate the composite distance, "CD" between each alternative to the optimal state.

$$\mathrm{CD}_i = \left[\sum_{j=1}^{m} W_{ij}\right]^{1/2}. \tag{11}$$

The one-column matrix formed as a result of this equation is called the composite distance matrix [CD] as shown in (12).

$$[\mathrm{CD}] = \begin{bmatrix} \mathrm{CD}_{11} \\ \mathrm{CD}_{21} \\ \vdots \\ \mathrm{CD}_{\mathrm{n1}} \end{bmatrix}. \tag{12}$$

The last step of this method involves calculating the rank of each alternative by using their composite distance values. The smallest value gets the 1st rank, the second smallest value gets the 2nd rank, and so on. This is how the DBA MCDM approach is used for cloud service provider selection. Below, Figure 2 represents the model development of the methodology.

Table 2: TFN scale for priority weights.

Symbols	TFN values
Extremely most important (EMI)	[1, 1, 1]
Very most important (VMI)	[0.8, 0.9, 1]
Most important (MI)	[0.6, 0.7, 0.8]
Important (I)	[0.4, 0.5, 0.6]
Less important (LI)	[0.2, 0.3, 0.4]
Very less important (VLI)	[0.0, 0.1, 0.2]
Extremely less important (ELI)	[0.0, 0.0. 0.0]

Table of TFN scale for importance of the 14 attributes during cloud service provider selection.

Table 3: TFN scale of performance ratings.

Symbols	TFN values
Very high (VH)	[1, 1, 1]
High (H)	[0.8, 0.9, 1]
Above average (AA)	[0.6, 0.7, 0.8]
Average (A)	[0.4, 0.5, 0.6]
Below average (BA)	[0.2, 0.3, 0.4]
Low (L)	[0.0, 0.1, 0.2]
Very low (VL)	[0.0, 0.0. 0.0]

Table of TFN scale for performance ratings of the 9 CSPs over the selection attributes.

3.3. Estimation-of-Distribution Algorithms. These algorithms are general metaheuristics applied in optimization to represent a recent alternative to classical approaches [21]. EDAs build probabilistic models of promising solutions by repeatedly sampling and selecting points from the underlying search space. EDAs typically work with a population of candidate solutions to the problem, starting with the population generated according to the uniform distribution over all admissible solutions [22]. Many distinct approaches have been proposed for the estimation of probability distribution,

$$D = \begin{bmatrix} 0.6 & 0.6 & 0.74 & 0.58 & 0.38 & 0.54 & 0.1 & 0.12 & 0.58 & 0.42 & 0.58 & 0.78 & 0.62 & 0.58 \\ 0.52 & 0.7 & 0.68 & 0.73 & 0.46 & 0.3 & 0.28 & 0.26 & 0.44 & 0.54 & 0.72 & 0.56 & 0.52 & 0.34 \\ 0.48 & 0.64 & 0.26 & 0.74 & 0.52 & 0.68 & 0.7 & 0.58 & 0.56 & 0.32 & 0.42 & 0.44 & 0.64 & 0.6 \\ 0.38 & 0.32 & 0.58 & 0.38 & 0.48 & 0.54 & 0.44 & 0.48 & 0.54 & 0.66 & 0.4 & 0.38 & 0.36 & 0.34 \\ 0.3 & 0.36 & 0.4 & 0.51 & 0.78 & 0.48 & 0.82 & 0.62 & 0.44 & 0.78 & 0.52 & 0.36 & 0.56 & 0.12 \\ 0.44 & 0.62 & 0.3 & 0.28 & 0.52 & 0.56 & 0.38 & 0.42 & 0.7 & 0.52 & 0.44 & 0.3 & 0.62 & 0.26 \\ 0.46 & 0.72 & 0.28 & 0.6 & 0.52 & 0.72 & 0.72 & 0.66 & 0.66 & 0.5 & 0.32 & 0.28 & 0.78 & 0.52 \\ 0.4 & 0.52 & 0.7 & 0.64 & 0.68 & 0.3 & 0.38 & 0.48 & 0.48 & 0.48 & 0.68 & 0.6 & 0.42 & 0.18 \\ 0.42 & 0.62 & 0.6 & 0.65 & 0.38 & 0.72 & 0.28 & 0.12 & 0.66 & 0.32 & 0.46 & 0.66 & 0.56 & 0.6 \end{bmatrix}$$

(a)

$$PW = [0.078 \quad 0.078 \quad 0.074 \quad 0.068 \quad 0.075 \quad 0.071 \quad 0.066 \quad 0.071 \quad 0.075 \quad 0.073 \quad 0.074 \quad 0.062 \quad 0.072 \quad 0.064]$$

(b)

Figure 4: (a) Decision matrix is represented by D. (b) Performance rating matrix represented by P.

$$D' = \begin{bmatrix} 1.92 & 0.25 & 1.28 & 0.08 & -1.16 & 0.01 & -1.56 & -1.53 & 0.19 & -0.60 & 0.60 & 1.80 & 0.47 & 1.05 \\ 0.93 & 1.00 & 0.95 & 1.11 & -0.52 & -1.58 & -0.77 & -0.80 & -1.33 & 0.25 & 1.71 & 0.46 & -0.38 & -0.30 \\ 0.43 & 0.55 & -1.33 & 1.18 & -0.03 & 0.94 & 1.07 & 0.85 & -0.02 & -1.31 & -0.67 & -0.27 & 0.64 & 1.17 \\ -0.79 & -1.85 & 0.41 & -1.29 & -0.35 & 0.01 & -0.06 & 0.33 & -0.24 & 1.10 & -0.83 & -0.63 & -1.74 & -0.30 \\ -1.78 & -1.55 & -0.56 & -0.39 & 2.06 & -0.38 & 1.60 & 1.05 & -1.33 & 1.96 & 0.12 & -0.75 & -0.03 & -1.54 \\ -0.05 & 0.40 & -1.11 & -1.97 & -0.03 & 0.14 & -0.33 & 0.02 & 1.50 & 0.11 & -0.51 & -1.12 & 0.47 & -0.75 \\ 0.19 & 1.15 & -1.22 & 0.22 & -0.03 & 1.21 & 1.16 & 1.26 & 1.06 & -0.03 & -1.47 & -1.24 & 1.84 & 0.71 \\ -0.54 & -0.35 & 1.06 & 0.49 & 1.25 & -1.58 & -0.33 & 0.33 & -0.89 & -0.17 & 1.40 & 0.70 & -1.23 & -1.20 \\ -0.30 & 0.40 & 0.52 & 0.56 & -1.16 & 1.21 & -0.77 & -1.53 & 1.06 & -1.31 & -0.35 & 1.07 & -0.03 & 1.17 \end{bmatrix}$$

Figure 5: Standardized matrix represented by D'.

$$O = [1.92 \quad 1.15 \quad 1.28 \quad 1.18 \quad 2.06 \quad 1.21 \quad 1.60 \quad 1.26 \quad 1.50 \quad 1.96 \quad 1.71 \quad -1.24 \quad -1.74 \quad -1.54]$$

(a)

$$O' = \begin{bmatrix} 0 & 0.90 & 0 & 1.09 & 3.24 & 1.19 & 3.17 & 2.79 & 1.30 & 2.56 & 1.11 & -3.05 & -2.22 & -2.60 \\ 0.98 & 0.15 & 0.32 & 0.06 & 2.58 & 2.79 & 2.38 & 2.07 & 2.83 & 1.70 & 0 & -1.07 & -1.36 & -1.24 \\ 1.48 & 0.60 & 2.62 & 0 & 2.10 & 0.26 & 0.52 & 0.41 & 1.52 & 3.27 & 2.39 & -0.97 & -2.39 & -2.72 \\ 2.71 & 3.00 & 0.87 & 2.47 & 2.43 & 1.19 & 1.67 & 0.93 & 1.74 & 0.85 & 2.55 & -0.61 & 0 & -1.24 \\ 3.70 & 2.70 & 1.85 & 1.58 & 0 & 1.59 & 0 & 0.20 & 2.83 & 0 & 1.59 & -0.48 & -1.71 & 0 \\ 1.97 & 0.75 & 2.40 & 3.16 & 2.10 & 1.06 & 1.94 & 1.24 & 0 & 1.85 & 2.23 & -0.12 & -2.22 & -0.79 \\ 1.72 & 0 & 2.51 & 0.96 & 2.10 & 0 & 0.44 & 0 & 0.43 & 1.99 & 3.19 & 0 & -3.59 & -2.26 \\ 2.46 & 1.50 & 0.21 & 0.68 & 0.81 & 2.79 & 1.94 & 0.93 & 2.40 & 2.13 & 0.31 & -1.95 & -0.51 & -0.34 \\ 2.22 & 0.75 & 0.76 & 0.61 & 3.24 & 0 & 2.38 & 2.79 & 0.43 & 3.27 & 2.07 & -2.32 & -1.71 & -2.72 \end{bmatrix}$$

(b)

Figure 6: (a) Optimal matrix represented by O. (b) Distance matrix represented by O'.

such as Independent Variables, Bivariate Dependencies, and Multiple Dependencies.

4. Implementation, Results, and Discussions

Evaluating various cloud service providers using DBA (distance-based approach) methodology with Fuzzy Set Theory to calculate ranks based upon selection attributes is described by the following steps:

(1) Identification and selection of cloud service providers (CSP): with the aid of an experienced group of dedicated decision-makers, specialized in the field of Computer Science and Technology, an analysis of several cloud service providers was conducted, and 9 most widely used CSPs, namely, (1) Amazon Web Services(AWS), (2) Digital Ocean, (3) Google Cloud Platform, (4) Microsoft Azure, (5) RackSpace, (6) SoftLayer, (7) IBM Smart Cloud Enterprise, (8) GoGrid, and (9) Google Compute Engine, were filtered out by evaluating them on various conceptual and practical criteria based on both quality and usability, thereby selecting them after immense brainstorming sessions and collective use of elimination principle

(2) Identification of selection attributes: three major factors were identified which were, namely, quality factors, technical factors, and economic factors. They were further classified as: quality factors (functionality, reliability, usability, efficiency, maintainability, and portability), technical factors (storage capacity, CPU performance, memory utilization, platform design, and network speed), and economic factors (service induction cost, maintenance cost, and promotion cost) after the detailed analysis and intensive study of the cloud service providing industry and its various prerequisites along with understanding the market where this industry thrives (Figure 3).

(3) Application of Fuzzy Set Theory: fuzzy logic forms its basis upon human acumen of decision-making regarding vague and nebulous information. It grossly

$$W = \begin{bmatrix} 0 & 0.06 & 0 & 0.08 & 0.78 & 0.10 & 0.66 & 0.55 & 0.12 & 0.47 & 0.09 & 0.57 & 0.35 & 0.43 \\ 0.07 & 0 & 0.01 & 0 & 0.50 & 0.55 & 0.37 & 0.30 & 0.60 & 0.21 & 0 & 0.18 & 0.13 & 0.09 \\ 0.17 & 0.02 & 0.51 & 0 & 0.33 & 0.01 & 0.01 & 0.01 & 0.17 & 0.78 & 0.42 & 0.05 & 0.41 & 0.47 \\ 0.57 & 0.70 & 0.05 & 0.41 & 0.44 & 0.10 & 0.18 & 0.06 & 0.22 & 0.05 & 0.48 & 0.02 & 0 & 0.09 \\ 1.07 & 0.57 & 0.25 & 0.16 & 0 & 0.18 & 0 & 0 & 0.60 & 0 & 0.18 & 0.01 & 0.21 & 0 \\ 0.30 & 0.44 & 0.42 & 0.67 & 0.33 & 0.08 & 0.24 & 0.10 & 0 & 0.25 & 0.36 & 0 & 0.35 & 0.04 \\ 0.23 & 0 & 0.46 & 0.06 & 0.33 & 0 & 0.01 & 0 & 0.01 & 0.29 & 0.75 & 0 & 0.92 & 0.32 \\ 0.47 & 0.17 & 0 & 0.03 & 0.04 & 0.55 & 0.24 & 0.06 & 0.43 & 0.33 & 0.01 & 0.23 & 0.01 & 0.01 \\ 0.38 & 0.04 & 0.04 & 0.02 & 0.78 & 0 & 0.37 & 0.55 & 0.01 & 0.78 & 0.31 & 0.33 & 0.21 & 0.47 \end{bmatrix}$$

(a)

$$CD = \begin{bmatrix} 2.079 \\ 1.747 \\ 1.844 \\ 1.852 \\ 1.807 \\ 1.800 \\ 1.850 \\ 1.624 \\ 2.085 \end{bmatrix}$$

(b)

Figure 7: (a) Weighted distance matrix represented by W. (b) Composite distance matrix represented by CD.

Table 4: Rank of cloud service providers.

Cloud service providers	Composite distances	Rank
Amazon Web Services	2.079	8
Digital Ocean	1.747	2
Google Cloud Platform	1.844	5
Microsoft Azure	1.852	7
RackSpace	1.807	4
SoftLayer	1.800	3
IBM Smart Cloud Enterprise	1.850	6
GoGrid	1.624	1
Google Compute Engine	2.085	9

Table for ranking of various cloud service providers.

distinguishes real-world problems based upon human comprehensive skills rather than absolute Boolean logic. In other words, the fuzzy system implements scales rather than 0/1 for coherent human understanding where 0 represents absolute fallacy, 1 represents absolute truth, and the middle values represent the fuzziness or the fuzzy values. In this study, we have implemented a triple fuzzy number scale which uses a triplet set of the form [a, b, c] with a sensory scale (Tables 2 and 3) [28]. A survey was conducted among a group of 40 selected experts associated with the technical field. The (Table 2) questionnaire consisted of 14 pristine questions, based upon which a priority weights matrix was created consisting of the weights or values of the assorted attributes. While in the second questionnaire (Table 3), the nine already selected CSPs were appraised on the grounds of the 14 categorized selection attributes by an adept team of 5 experts. The extracted data from the questionnaires mentioned above were converted from a literal scale to TFN (Triple Fuzzy Number) scale, thereby averaged to a fuzzy number

(4) Determination of weights and performance ratings: the expert-assigned linguistic terms were first converted into corresponding TFNs using the fuzzy scale and then defuzzied to get crisp score values. The data was extracted from the questionnaires and then evaluated using a combination of the mathematical formulas and concepts of aggregation and average

(5) Creating performance rating matrices: a decision matrix of the performance ratings (Figure 4(a)) and a single-row matrix of the priority weights (Figure 4(b)) were created under the supervision of expert guidance using the fuzzy scale and MCDM

(6) Calculating standardized matrix: root mean square of each selection attributes is carefully evaluated; furthermore, the mean previously determined is subtracted from each value and simultaneously divided by the corresponding root mean square of that particular selection attribute to get the standardized matrix (Figure 5).

(7) Creating optimal and distance matrix: the optimal matrix is estimated by targeting the best values of each selected attribute of the standardized matrix (Figure 6(a)), i.e., the maximum values for quality and technical factors and minimum values for economic factors. Additionally, the distance matrix is calculated by finding the distance between each value of a particular selection attribute with its corresponding best value (Figure 6(b)).

(8) Calculating weighted and composite distance matrix: by squaring the respective values of the distance matrix and multiplying them to the corresponding priority weights, the weighted distance matrix was obtained (Figure 7(a)). The matrix formed was then used to evaluate the composite distance matrix by calculating the square root of the total sum of each alternative (Figure 7(b)).

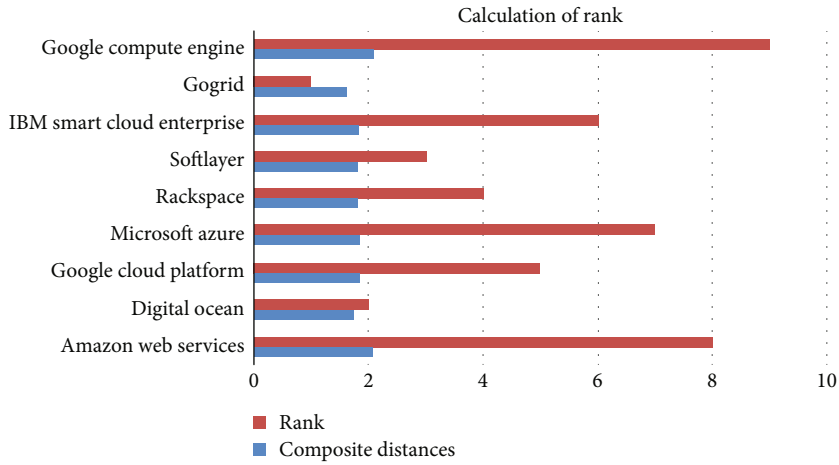


Figure 8: Calculation of rank.

(9) Ranking of cloud service providers: finally, the alternatives are ranked in decreasing order of their corresponding values in the composite distance matrix. Therefore, the least rank or rank 1 is most preferable while the maximum rank, i.e., rank 9, is least preferable considering the given set of alternatives (Table 4).

The selection of cloud service providers is a problematic task as many decision-making parameters are taken into consideration like security, cost optimization, availability, reliability, and fault tolerance, to name a few. Most of the mentioned factors are not constant and individualistic wherein every consumer who requires a cloud service provider has an almost unique set of demands and requisites, each having selected set of attributes has a different weight than other, i.e., prioritized attributes are not rare. Considering the presented scenario, Multicriteria Decision-Making technique has shown significant efficacy and is implemented in the field widely as it provides both individualistic and concurrent results.

Table 4 shows the ranking of nine cloud service providers based on fourteen carefully discerned attributes categorized in three categories, namely, quality factors (functionality, reliability, usability, efficiency, maintainability, and portability), technical factors (storage capacity, CPU performance, memory utilization, platform design, and network speed), and economic factors (service induction cost, maintenance cost, and promotion cost). The step-by-step gradation procedure described above where priority weights and decision matrix are extracted from surveys data using a fuzzy scale, which is then optimized-standardized and refined by priority weights as extracted from user survey data proves to be a simple, effective, and reliable method of action for selection of optimal cloud service provider. Figure 8 shows the graphical representation of the same.

5. Conclusion and Future Work

Taking into consideration the current extensive use in the cloud-based IoT services for building computation, storage, infrastructure, and other needs has led to a greater demand for an efficient methodology to drive which given cloud service provider meets one's unique and individualistic demands for fulfilling ever-changing solutions in the field of IoT. Given the scenario of current cloud-based IoT applications with multiple service providers and vivid requirements, a lot of decision-making criteria and methodologies already exist, some of which include TOPSIS that is a useful and straightforward technique for ranking several possible alternatives according to closeness to the ideal solution, or AHP, or VIKOR which is based on the aggregating fuzzy merit that represents the closeness of an alternative to the ideal solution by compromising between two or more options to get a unified opinion between multiple criteria, and PROMETHEE compares various measures available by the technique of outranking.

Despite the preexisting techniques to classify, evaluate, and rate various IoT-based cloud service providers due to continuously changing set of attributes, challenges user's privacy and security, user authentication, location privacy, disparate demands of customers, and colossus pool of available attributes ranging from performance, cost optimization to quality, it becomes very peculiar to get virtually concurrent results for one's ever-changing characteristics and agility even after discerning the given set of methodologies in the available literature. Therefore, this research meets all the scenarios mentioned above, demands and unique characteristics by using an optimized matrix methodology aided by distance-based approach (DBA), some of its salient features are as follows:

(a) Considering a broad set of categories that are further graded into subattributes, i.e., performance, technical and economic factors are individually optimized to get ultimate efficacy

(b) Simple, straightforward, reader-friendly, and easily captured and understood by anyone concept and procedure

(c) It is obtained by taking into consideration priorities among attributes as extracted from a user by surveyed data

Conclusively, in the presented research methodology, a distance-based approach with an optimized approach to

consider priorities as set by data extracted from user survey in a simple, lucid yet compelling procedure to select an ideal cloud service provider for IoT applications. This study finds nine alternatives or popular cloud service providers and 14 attributes or deciding criteria.

Privacy and security are the two most emerging challenges in IoT applications as provided by cloud service providers due to the nascent nature of the field. Though IoT-based applications have already been explored from the aspects of privacy and security, implementing IoT applications via cloud-based platforms leads to a new set of possible threats. In future work, this study intends to evaluate variegated cloud-based IoT platforms from the aspects of security and privacy by analysing the same under the purview of three criteria. Firstly, the future work will deal with user's individualistic threats of privacy and security like location privacy, breach of personal information, protection of user's hardware and software devices, and user profile authentication. Other criteria include privacy and security challenges for a multilevel organization, namely, secure route establishment, isolation of malicious nodes, self-stabilization of the security protocol, and preservation of location privacy. Lastly, this study would assay multiple case studies of leading cloud-based IoT platforms' breach of security and privacy to perform a comparative analysis of the same.

Data Availability

The data will be provided based on a request by the evaluation team.

Consent

All the authors of this paper have shown their participation voluntarily.

Conflicts of Interest

The authors of this research article declare that there is no conflict of interest in preparing this research article.

Authors' Contributions

Alakananda Chakraborty is responsible for the methodology; Muskan Jindal for the software and data curation; Mohammad R. Khosravi for the data curation and writing—original draft preparation; Achyut Shankar for the visualization and investigation; Prabhishek Singh for the supervision; Manoj Diwakar for the software and validation; and Achyut Shankar for the writing—reviewing and editing.

References

[1] R. Sikarwar, P. Yadav, and A. Dubey, "A survey on IOT enabled cloud platforms," in *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 120–124, Gwalior, India, 2020, April.

[2] N. Almolhis, A. M. Alashjaee, S. Duraibi, F. Alqahtani, and A. N. Moussa, "The security issues in IoT-cloud: a review," in *2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, pp. 191–196, Langkawi, Malaysia, 2020, February.

[3] M. Ullah, P. H. Nardelli, A. Wolff, and K. Smolander, "Twenty-one key factors to choose an IoT platform: theoretical framework and its applications," 2020, https://arxiv.org/pdf/2004.04924.

[4] H. M. Alabool and A. K. Mahmood, "Trust-based service selection in public cloud computing using fuzzy modified VIKOR method," *Australian Journal of Basic and Applied Sciences*, vol. 7, no. 9, pp. 211–220, 2013.

[5] Z. Rehman, O. K. Hussain, and F. K. Hussain, "Parallel cloud service selection and ranking based on QoS history," *International Journal of Parallel Programming*, vol. 42, no. 5, article 276, pp. 820–852, 2014.

[6] P. Ganguly, "Selecting the right IoT cloud platform," in *2016 International Conference on Internet of Things and Applications (IOTA)*, pp. 316–320, Pune, India, 2016, January.

[7] P. Varga, J. Peto, A. Franko et al., "5G support for industrial IoT applications–challenges, solutions, and research gaps," *Sensors*, vol. 20, no. 3, p. 828, 2020.

[8] M. Lin, C. Huang, Z. Xu, and R. Chen, "Evaluating IoT platforms using integrated probabilistic linguistic MCDM method," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11195–11208, 2020.

[9] A. E. Youssef, "An integrated MCDM approach for cloud service selection based on TOPSIS and BWM," *IEEE Access*, vol. 8, pp. 71851–71865, 2020.

[10] Y. Wang, Y. Lin, R. Y. Zhong, and X. Xu, "IoT-enabled cloud-based additive manufacturing platform to support rapid product development," *International Journal of Production Research*, vol. 57, no. 12, pp. 3975–3991, 2018.

[11] V. S. Narwane, B. E. Narkhede, R. D. Raut, B. B. Gardas, P. Priyadarshinee, and M. S. Kavre, "To identify the determinants of the CloudIoT technologies adoption in the Indian MSMEs: structural equation modelling approach," *International Journal of Business Information Systems*, vol. 31, no. 3, pp. 322–353, 2019.

[12] A. Mijuskovic, I. Ullah, R. Bemthuis, N. Meratnia, and P. Havinga, "Comparing apples and oranges in IoT context: a deep dive into methods for comparing IoT platforms," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1797–1816, 2021.

[13] L. Surya, "Security challenges and strategies for the IoT in cloud computing," *International Journal of Innovations in Engineering Research and Technology ISSN*, vol. 3, pp. 2394–3696, 2016.

[14] S. Bhatt, A. T. Lo'ai, P. Chhetri, and P. Bhatt, "Authorizations in cloud-based internet of things: current trends and use cases," in *2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC)*, pp. 241–246, Rome, Italy, 2019, June.

[15] M. M. Hassan, M. R. Hassan, S. Huda, and V. H. C. de Albuquerque, "A robust deep learning enabled trust-boundary protection for adversarial industrial IoT environment," *IEEE Internet of Things Journal*, vol. 8, 2020.

[16] M. A. da Cruz, J. J. Rodrigues, P. Lorenz, V. Korotaev, and V. H. C. de Albuquerque, "In. IoT—a new middleware for internet of things," *IEEE Internet of Things Journal*, vol. 8, 2020.

[17] K. A. da Costa, J. P. Papa, C. O. Lisboa, R. Munoz, and V. H. C. de Albuquerque, "Internet of Things: a survey on machine

learning-based intrusion detection approaches," *Computer Networks*, vol. 151, pp. 147–157, 2019.

[18] S. Satyadevan, B. S. Kalarickal, and M. K. Jinesh, "Security, trust and implementation limitations of prominent IoT platforms," in *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*, pp. 85–95, Bhubaneswar, Odisha, India, 2015.

[19] X. Li, Q. Wang, X. Lan, X. Chen, N. Zhang, and D. Chen, "Enhancing cloud-based IoT security through trustworthy cloud service: an integration of security and reputation approach," *IEEE Access*, vol. 7, pp. 9368–9383, 2019.

[20] C. Stergiou, K. E. Psannis, B. G. Kim, and B. Gupta, "Secure integration of IoT and cloud computing," *Future Generation Computer Systems*, vol. 78, pp. 964–975, 2018.

[21] J. Zhou, Z. Cao, X. Dong, and A. V. Vasilakos, "Security and privacy for cloud-based IoT: challenges," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 26–33, 2017.

[22] M. Hauschild and M. Pelikan, "An introduction and survey of estimation of distribution algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 3, pp. 111–128, 2011.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5513582, 9 pages
https://doi.org/10.1155/2021/5513582

Research Article

A New Semantic and Statistical Distance-Based Anomaly Detection in Crowd Video Surveillance

Fariba Rezaei and Mehran Yazdi

School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran

Correspondence should be addressed to Mehran Yazdi; yazdi@shirazu.ac.ir

Received 5 February 2021; Revised 9 April 2021; Accepted 7 May 2021; Published 17 May 2021

Academic Editor: Alireza Jolfaei

Recently, attention toward autonomous surveillance has been intensified and anomaly detection in crowded scenes is one of those significant surveillance tasks. Traditional approaches include the extraction of handcrafted features that need the subsequent task of model learning. They are mostly used to extract low-level spatiotemporal features of videos, neglecting the effect of semantic information. Recently, deep learning (DL) methods have been emerged in various domains, especially CNN for visual problems, with the ability to extract high-level information at higher layers of their architectures. On the other side, topic modeling-based approaches like NMF can extract more semantic representations. Here, we investigate a new hybrid visual embedding method based on deep features and a topic model for anomaly detection. Features per frame are computed hierarchically through a pretrained deep model, and in parallel, topic distributions are learned through multilayer nonnegative matrix factorization entangling information from extracted deep features. Training is accomplished through normal samples. Thereafter, K-means is applied to find typical normal clusters. At test time, after achieving feature representation through deep model and topic distribution for test frames, a statistical earth mover distance (EMD) metric is evaluated to measure the difference between normal cluster centroids and test topic distributions. High difference versus a threshold is detected as an anomaly. Experimental results on the benchmark Ped1 and Ped2 UCSD datasets demonstrate the effectiveness of our proposed method in anomaly detection.

1. Introduction

Automatic video surveillance has recently attracted the attention of researchers since a large number of cameras, installed in surrounding places, may not let human based-surveillance be error free. Thus, computer vision and machine learning come to help analyze the output videos for various tasks of automatic recognition and anomaly detection. Originally, raw signals are used to extract information through machine learning techniques [1]. However, the high dimensionality of video signals captured by high-resolution video cameras makes traditional methods computationally complex. Thereby, to combat the issue of curse of dimensionality, dimensionality reduction techniques have received more attention. Linear and nonlinear dimensionality reduction approaches can be applied as task-dependent techniques. PCA, MDS, LLE, and autoencoder are some to name a few. Generally speaking, all computer vision-based feature extraction methods like handcrafted features (SIFT, HOG, etc...) can also be considered a kind of dimensionality reduction.

New emerging embedding methods, basically introduced in natural language modeling/processing (NLP), map the original high-dimensional signals to embed spaces and consecutively capture high-level information, which besides the compression, the semantic relations of signals are also preserved [2, 3]. Embedding techniques in NLP are based on representing each word as a vector in a vector space model. Preliminary one hot encoding suffers from lack of preservation of semantic relations, since orthogonally between words neglects the probable coherence between them. Topic-based representations such as LSA, probabilistic LSA, LDA, and NMF try to capture semantics [3].

Embedding can also be applied to vision tasks to bridge the semantic gap in image or video analysis. Recently, deep

learning architectures (CNN, RNN, AE, RBM, etc.) have been well studied for anomaly detection [4]. Diving into the high-level features, they have shown considerable results in comparison to handcrafted features. Supervised CNNs consist of both convolution and fully connected (FC) layer for feature extraction and classification/recognition, respectively. Ultraparameters in CNN are caused by those terminative FC layers, which may cause overfitting in limited dataset regimes when training from scratch. Therefore, attention is trended toward using only pretrained convolutional layers for feature extraction and powerful image representations, putting aside FC layers.

In most researches, anomaly detection is investigated based on defining a model(s) on normal samples and detecting anomalies as deviation from this normality. This deviation can be measured either by likelihood or similarity. In [5], an anomaly was defined based on interaction forces between pedestrians using the social force model (SFM), and LDA was used to compute likelihood for test set to evaluate deviation from a normal model in a probabilistic framework, whereas in [6, 7], normal training samples were used to create a dictionary model and deviation was calculated as high sparse reconstruction cost between an original test sample and its reconstruction through a linear combination of normal bases in the Euclidean space.

In this paper, we investigate a combination of the deep model, topic model, and statistical distance for anomaly detection. In contrast to previous methods which were based on either handcrafted or deep features, neglecting semantic and interpretable information, we analyze the combination of a deep model with a topic model hierarchically to produce semantic representation. We apply a pretrained deep model for hierarchical feature extraction from different layer levels, for each training image. Thereafter, we take the advantages of nonnegative matrix factorization (NMF) as a topic modeling approach in capturing semantic features. Specially, we applied a multilayer NMF, for hierarchical topic representation injecting information extracted from hierarchical layers of a deep model in hierarchical decompositions. After learning topic distribution per frame in the training stage, we apply K-means clustering to compute cluster centroids as typical normal topic-based representations. At test time, in a similar pipeline for feature extraction at the train stage, semantic representation for test frames is calculated and compared to typical normal topic distributions through a statistical distance metric. Here, the earth mover distance (EMD) metric is chosen as a distance metric since it has shown efficient performance in comparing distributions.

Our main contributions are as follows:

(1) We take the advantages of both the deep model (pretrained VGG-Net) and the topic model (multilayer NMF), hierarchically and in combination to reach high-level and semantic frame representation

(2) Since topic distributions are extracted at the final level as the frame representations, after K-means clustering, some normal representative topic distributions for normality are achieved, and then, EMD statistical distance metric is applied in clustering-based anomaly detection framework

The organization of the rest of this paper is as follows: literature review in three domains of anomaly detection, topic modeling, and statistical learning methods are provided in Section 2. Section 3 introduces our proposed pipeline for crowd anomaly detection. Experimental results are reported in Section 4. Finally, Section 5 concludes this paper.

2. Literature Review

In this section, we review researches in anomaly detection, topic modeling, and statistical distance separately.

2.1. Anomaly Detection. Video surveillance studies for anomaly detection was started by using traditional handcrafted feature extraction and model learning and improved over the years by applying end-to-end deep architectures. Formerly, low-level features like color, texture, and its variants, like mixture of dynamic texture (MDT), SIFT, SURF, optical flow, and trajectories, were extracted either from appearance, motion, or both, depending on the anomaly definition. At model learning stages, binary classifiers like SVM, decision tree, and NN have been applied for supervised scenarios [1]. However, in semisupervised and unsupervised scenarios, given only normal videos at the training stage, a model for normal behavior is created and an anomaly is detected as a deviation from this model. This has been done for instance by one-class SVM (OCSVM) or fitting a Gaussian model on normal samples. Some researchers took the idea of the inherent sparsity of vision. A dictionary was learned from normal samples, and at the test time, a large reconstruction error was interpreted as an anomaly. Reconstruction was done as a linear combination of dictionary bases which are representative of all normal samples. Dictionary can be learned offline through codebook generation or online through updating along with observing new normal samples [8].

Recently, deep learning methods have commenced entering to the practical realm like vision, lexical, and speech. The intermediate image representations learned through CNN, especially when trained on large-scale datasets like ImageNet, have been proven to be powerful image descriptors.

In [9], anomalous behaviors were captured through a novel concept of aggregation of ensembles (AOE), based on fine-tuning different pretrained ConvNets and a pool of classifiers. They assumed that different CNN architectures learn different levels of representation from crowd videos, and thus, an ensemble of CNNs will enable enriched feature sets to be extracted. Autoencoder-based architectures were also studied where a large reconstruction error was considered a sign of anomaly score. The autoencoder can reduce dimensionality and is vastly used in unsupervised learning problems or as the preliminary stage of supervised task [10]. In particular, after training an AE or sparse AE on normal samples, the bottleneck layer can be considered feature extraction layers for any test samples. Some researchers tried to incorporate both handcrafted and deep features in a unified

configuration. In [11], a trajectory-pooled deep convolutional descriptor was introduced combining dense trajectories and convolutional feature maps which results in high discriminative features. Convolutional networks outperform both traditional low-level features and their compositional forms like BoW, Fisher Kernel, and VLAD, [12] although sometimes are used cooperatively. In [12], features extracted from within layers of a convolutional network were used in VLAD to compress the data and subsequently feed to SVM for classification. Wimmer et al. [13] applied Fisher vector encoding to the output feature maps of CNN to find fixed-length representation for image classification.

Sabokrou et al. investigated video anomaly detection through different deep architectures [14–21]. Autoencoder-based anomaly detection and localization using sparsity was introduced in [14, 15]. An architecture based on deep 3D autoencoder, deeper 3D convolutional neural network (CNN), and cascade of two cascaded classifiers was proposed in [16] for anomaly detection. High speed and accurate detection and localization of anomalies were achieved in [18] using fully convolutional neural networks (FCNs) and cascaded outlier detection. Some researches applied generative adversarial networks and its variants for image anomaly detection [17, 19, 22]. Semisupervised anomaly detection was analyzed in [23] based on information theory. A novel self-supervised representation learning based on integration of a neighbourhood-relational encoding (NRE) among the training data and an encoder-decoder structure was proposed in [20]. In [21], they propose an adversarial training approach to detect out-of-distribution samples in an end-to-end model through jointly training two deep neural networks which collaborate at test time to detect novelties.

2.2. Topic Modeling. Topic modeling is an unsupervised method, originally introduced for text analysis, but has been also noticed in vision. It is based on the idea that documents containing similar contents will likely use a similar set of words that are indicated by topics. Topic modeling discovers patterns as low-dimensional latent representation given unlabeled collection of documents constituted of words. pLSA, LDA, and NMF are among the most common probabilistic topic modeling approaches [24–26]. Topic models take as input a set of documents J, a set of words V, and in a cooccurrence matrix of words and documents $F = \|n_{wj}\|_{w\epsilon V.j\epsilon J}$ (or BoVW representation, and produce a set of topic T, or more especially $P(w \mid k)$ and $p(k \mid j)$, for $w \in V.j \in J.k \in T$, as word distribution per topic and topic distribution per document, respectively. Consider n_{wj} as the number of times the word w appears in document j, then documents can be represented as mixtures of topics.

F can be decomposed into two matrices $F = \Phi\Theta$, where $\Phi = \{\phi_{wk}\}_{w\epsilon V.k\epsilon K}$ is a word-topic matrix with $\phi_{wk} = p(w \mid k)$ and $\phi_k = \{\phi_{wk}\}_{w\epsilon V}$, and $\Theta = \{\theta_{kj}\}_{k\epsilon K.j\epsilon J}$ is a topic-document matrix with $\theta_{kj} = p(k \mid j)$ and $\theta_j = \{\theta_{kj}\}_{k\epsilon K}$. The decomposition can be solved through the various topic model algorithms with a different assumption. For instance, LDA uses a predefined number of topics, whereas hierarchical Dirichlet process (HDP) [27] estimates the best number of topics based on the training dataset.

In [28], Niebles et al. studied the application of latent topic models, namely, pLSA and LDA, for action categorization. Especially, they extract spatiotemporal interest points along the input volumes followed by codebook generation. In an unsupervised fashion, they succeeded in detecting and localizing actions, which were considered latent topics. New learning algorithms based on EM and variational Bayes inference were proposed in [29] for activity analysis in videos where the description of activities and behaviors was made by the dynamic topic model. The activities and behaviors were described by a dynamic topic model. They also evaluated anomaly localization procedures in the topic modeling framework. In [30], scene classification was made by discovering objects per image in an unsupervised fashion using pLSA. They subsequently used object distribution in each image for scene classification using supervised kNN. Topic modeling-based abnormal behavior recognition has been previously investigated in [5, 31]. In almost all cases, low likelihood corresponds to abnormal test samples. An unsupervised topic model (pLSA) anomaly detection and localization were studied in [32] based on extra information of location and size beside quantized spatiotemporal gradient descriptors to create a more informative vocabulary over visual clips. Each document (frame) is fully described by a corresponding distribution over topics.

2.3. Statistical Distance. Statistical distances try to find the distance between two statistical objects, and when accompanied with a symmetric property, they are known as a metric. In the anomaly detection area, distance measures such as Jensen Shannon divergence or Z score value were applied for comparing query observation to those extracted patterns from normal samples [33]. According to the evaluation of this distance concerning the threshold, the anomaly can be detected. As a powerful statistical distance, earth mover distance (EMD), also known as the Wasserstein metric, was applied in the image domain [34, 35] to compare two probability distributions, mainly based on low-level features like color or texture. It is based on computing statistical distance between two signatures. The typical signature consists of a list of pairs:

$$S = \{(x_1.m_1).(x_2.m_2) \cdots (x_n.m_n)\}, \tag{1}$$

where each x_i is a certain feature, and m_n is its mass (how many times that feature occurs in the record). Considering two signatures P and Q which contain m and n clusters, respectively,

$$P = \{(p_1.w_{p1}).(p_2.w_{p2}) \cdots (p_m.w_{pm})\}, \tag{2}$$

$$Q = \{(q_1.w_{q1}).(q_2.w_{q2}) \cdots (q_n.w_{qn})\}, \tag{3}$$

and $p_i(q_i)$ is the cluster representative and $w_{pi}(w_{qi})$ is the weight of cluster i. Also, consider $D = [d_{i.j}]$ as the ground distance between clusters p_i and q_j. It can be chosen or learned

according to the problem at hand. The aim is to find flow matrix $F = [f_{i.j}]$, where $f_{i.j}$ is the flow between p_i and q_j, such that the below overall cost is minimized with its related constraints.

$$\begin{aligned} &\min \sum_{i=1}^{m}\sum_{j=1}^{n} f_{i.j}d_{i.j}, \\ &f_{i.j} \geq 0\, 1 \leq i \leq m.1 \leq j \leq n, \\ &\sum_{j} f_{i.j} \leq w_{pi} 1 \leq i \leq m, \\ &\sum_{i} f_{i.j} \leq w_{qi}.1 \leq j \leq n, \\ &\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i.j} = \min\left\{\sum_{j=1}^{m} W_{pi}.\sum_{j=1}^{n} W_{qj}\right\}. \end{aligned} \tag{4}$$

This optimization can be solved via linear programming. It is based on solving a kind of transportation problem. Once the flow F is calculated, then the EMD is defined as the work normalized by the total flow:

$$\mathrm{EMD}(P.Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i.j}d_{i.j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i.j}}. \tag{5}$$

EMD suffers from high computational complexity $O(N^3 \log N)$. Wavelet EMD was proposed in [36] to reach a linear time algorithm for approximating the EMD for low-dimensional histograms using the sum of absolute values of the weighted wavelet coefficients of the difference histogram.

Rare studies have gained from EMD in anomaly detection. To the best of our knowledge, only in [7], wavelet EMD was applied in conjunction with sparse representation for anomaly detection instead of the Euclidean distance, for its robustness. In this paper, we investigate wavelet EMD on our proposed clustering-based anomaly detection.

3. Proposed Method

In this paper, we analyze anomaly detection at frame level in crowded scenes. Our proposed architecture is shown in Figure 1. The pipeline consists of two stages: (1) feature extraction and (2) anomaly detection. The feature extraction stage itself consists of two parts entangled with each other: (1) hierarchical feature extraction through pretrained VGG-Net [37] and (2) hierarchical latent representation from multilayer NMF. Both architectures start from low-level features and increase in depth to high-level information resulting in ultimate representation.

In the second stage, we applied clustering-based anomaly detection. Precisely, K-means is applied to all processed training samples' ultimate representations, to create typical normal clusters. Since the training dataset consists of only normal samples, thus, cluster centroids are normal frame representatives. At test time, test frames are processed to be represented in learned topic space from the training stage and compared to each cluster centroids. A large statistical distance from all centroids is detected as an anomaly. In the following, we explain each part in more detail.

3.1. Preprocessing and Feature Extraction. The dataset is separated into two subsets as train and test set. Let $X_{\text{train}} = [x_1.x_2 \cdots x_{n_{\text{Train}}}]^T \in R^{n_{\text{Train}} \times B_0}$, where n_{Train} is the number of frames in the train dataset, $B_0 = m \times n \times c$ and m, n, and c are the width, height, and number of channel, respectively, for the original captured image.

3.1.1. Deep Representation. Pretrained model is applied for feature extraction in problems encountering scarcity of training datasets, since training from scratch may result in overfitting. As higher layer feature maps are task specific, we extract more general features from lower layers. We resized each frame to be in a compatible size as the input for VGG-Net model ($m_0 \times n_0 \times c_0$) and extract features hierarchically from different depths of the architecture. Let $a^0 = x(\in R^{m_0 \times n_0 \times c_0})$ be a typical train image in compatible size with VGG input layer. Then,

$$a^l = f\left(w^{l-1}a^{l-1} + b^{l-1}\right) \in R^{m_l \times n_l \times c_l}, \tag{6}$$

is the output feature map from layer l. w^{l-1} and b^{l-1} are VGG weights and biases pretrained, respectively, for layer l. $m_l \times n_l$ is the spatial size of the feature map, and c_l is the feature map's depth at layer l. We extract feature maps from L different depths $(l = 1 \cdots .L)$; then, feature maps at each layer $l\,(l = 1.2 \cdots .L)$ are separately feed to the global average pooling (GAP) layer to get representations in vector format. GAP layers take input volumes of size $m_l \times n_l \times c_l$ and create $1 \times c_l$ dimensional vector by spatial averaging. Therefore, for each frame x, now, we have L vector representations, $f_{Dl} \in R^{c_l}(l = 1.2 \cdots .L)$. Considering all training samples, now we have L different size matrices, $M_l \in R^{n_{\text{Train}} \times f_{Dl}}$.

3.1.2. Topic-Based Representation. In parallel, we try to capture semantic information based on the topic model. Specially, we applied multilayer NMF since multilayer has been shown to improve performance by capturing more semantic features [38]. We adopt a similar approach to [39] by considering a frame as a document and trying to extract topic distribution per document. However, we apply multilayer NMF for hierarchical topic modeling. Single-layer NMF decomposes a nonnegative matrix V into two low-rank nonnegative basis and coefficient matrices W and H.

$$\mathrm{V} = \mathrm{WH}'.\mathrm{V} \in R^{m \times n}.\mathrm{W} \in R^{m \times k}.\mathrm{H} \in R^{n \times k}, \tag{7}$$

where H is the new low-dimensional representation for V. The decomposition is solved as an optimization problem through a multiplicative update approach. In multilayer NMF, computed latent representation in preceding layers is decomposed hierarchically in subsequent layers. Consider $X_{\text{train-pca}} = \mathrm{PCA}(X_{\text{train-vec}})$ and $X_{\text{train-pca}} = [x_1.x_2 \cdots x_{n_{\text{Train}}}]^T \in R^{n_{\text{Train}} \times D_0}$, where PCA applied to each vectorized frame to decrease dimensionality from $m_0 \times n_0$ to $D_0 < m_0 \times n_0$ per

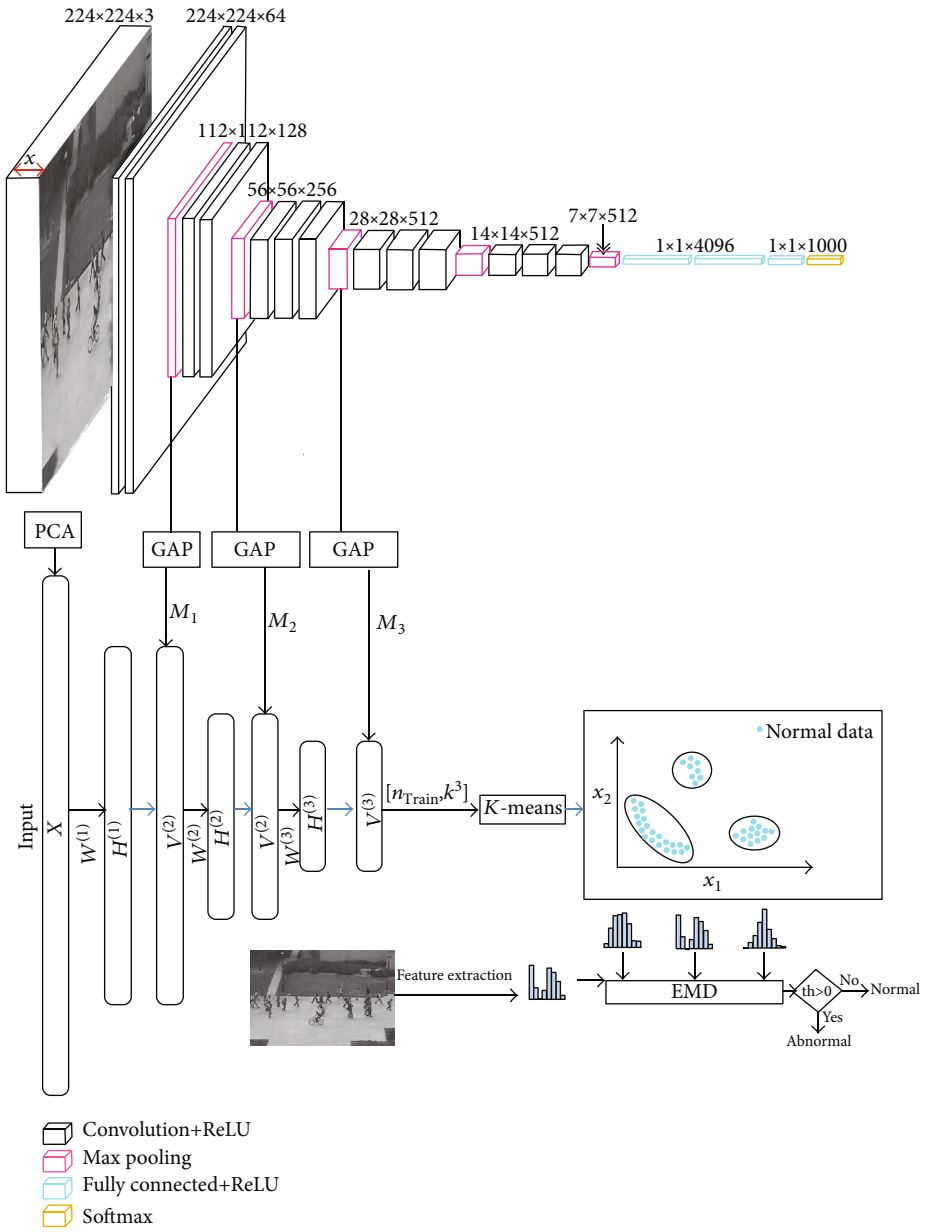


Figure 1: Our proposed architecture for anomaly detection. It consists of two stages of hierarchical feature representation and cluster-based anomaly detection.

frame and standardized to stay in range [0-1] . Let $H_0 = X_{\text{train-pca}}$ as input to the first stage of multilayer NMF. Then, it can be decomposed as $H_0 = W_1 H_1$. Instead of directly applying the second NMF to H_1, as the new low-dimensional representation, H_1 is processed to V_1 before being introduced to the next layer. V_l is computed as $V_l = f(H_l.M_l).l = 1 \cdots L$ where $f(.)$ is the nonlinear function, like softmax, and M_l is feature representation from pretrained VGG-Net at layer l .

$$V_l = f(H_l.M_l) = W_{l+1}H'_{l+1}.W_{l+1} \in R^{D_{l-1}\times D_l}.H_{l+1} \in R^{n_{\text{Train}}\times D_l}. \quad (8)$$

Table 1: UCSD dataset in detail.

Dataset	Resolution	Number of training sequences	Number of test sequences
Ped1	158 × 238	34~200 images	36~200 images
Ped2	240 × 360	16120~200 images	12120~200 images

Here, we use softmax as a nonlinear function to have a distribution-like representation. Since the ReLu activation function has been applied in deep architecture, nonnegativity is preserved. Bringing in M_ls in multilayer NMF decomposition results in both high-level and semantic information, which can improve the performance of the subsequent tasks.

Figure 2: Typical normal and abnormal samples of the UCSD dataset. Left to right: normal frame and abnormal frame for Ped1 and normal frame and abnormal frame for Ped2.

By decomposing V_l in the next layer, we force the architecture to learn how to combine information from the previous layer; therefore, $D_l < D_{l-1}$. Training separately each NMF layer, to learn W_l and H_l, ultimate data representation V_L is acquired. Finally, V_L integrates features throughout the deep model and topic model.

3.2. Anomaly Detection. Upon training completion, $V_L \in R^{n_{\text{Train}} \times D_L}$ is acquired from normal frames in the training set. We apply K-means algorithm to V_L to find K cluster centroids as normality representatives. Therefore, now, we have K cluster centroids $s_i.i = 1 \cdots .K$ which are used in cluster-based anomaly detection. Each test frame x_{test} is fed to our learned feature extraction block from the training phase, and ultimate representation $V_{L.test}$ is acquired. $V_{L.test}$ can be considered as the final topic distribution for x_{test}. $V_{L.\text{test}}$ is compared to each s_i and exceedance of statistical wavelet EMD distance from threshold th is detected as an anomaly.

$$\min_{i=1:K} \left(d_{\text{EMD}.i}(V_{L.\text{test}}.s_i)\right) > \text{th} \rightarrow V_{L.\text{test}}, \tag{9}$$

is an abnomal frame.

4. Results and Discussion

We conducted experimental analysis on UCSD dataset as one of the benchmark datasets in crowd anomaly detection introduced in [40], recorded with a static camera at 10 fps. This dataset contains two scenes as Ped1 and Ped2, each of which is split into train and test sequences. The nonpedestrian objects, like bikers, skaters, and small carts, are considered anomalies. More details about this dataset are provided in Table 1. Typical normal and abnormal sample frames for Ped1 and Ped2 datasets are also shown in Figure 2.

When originally introduced, VGG [37] was trained on the ImageNet dataset which only consists of object classes; however, recently, pretrained VGG on both the ImageNet and Places dataset is provided which consider scene classes, as well. 1000 classes from the ImageNet and the 365 classes from the Places365Standard [41] were merged to train a VGG16-based model (Hybrid1365-VGG [42]). We use VGG model pretrained both on the ImageNet and Places datasets to improve the capability of our deep feature extraction block in capturing both objects and scenes features. For this paper, our algorithms have been implemented in Python and run on a PC with 2.9 GHz Core i5 GPU, with GTX1080 GPU, and 16G RAM. Original frames are resized to be compatible with VGG, as VGG accepts input of size $224 \times 224 \times 3$. Feature maps from different depths, namely, block2 – pool, block3 – pool, and block4 – pool of VGG architecture, were extracted and resulted in $(56 \times 56 \times 128)$, $(28 \times 28 \times 256)$, and $(14 \times 14 \times 512)$ feature maps, respectively. Then, we applied global average pooling to each feature map separately which results in $f_{D1} : 128D$, $f_{D2} : 256D$, and $f_{D3} : 512D$ representation vectors in hierarchical order. On the other hand, we applied multilayer NMF with $L = 3$ on our train set with reduced dimensionality by PCA (2000D vector each frame). W_0, W_1, and W_2 are learned separately with a multiplicative updates. D_1, D_2, and D_3 are chosen as 512, 256, and 128, respectively. K-means clustering with $K = 50$ is applied to the final representation $V_L \in R^{n_{\text{Train}} \times D_L}$ to generate typical representative centroids. In the UCSD dataset, there are $n_{\text{Train}} = 6800$ for Ped1 and $n_{\text{Train}} = 2550$ for Ped2 datasets.

Table 2: Fixed parameter used in the proposed algorithm.

Dataset/parameters	Ped1	Ped2
Number of training samples	2550	6800
L (number of levels for feature hierarchies)	3	
K (K-means clustering)	50	
Threshold (for WEMD distance comparison)	0.33	0.24

In our experiment, there are some parameters that we investigate their values and fixed after evaluation. These parameters are shown in Table 2.

VGG16 consists of several layers (C11-C12-P1-C21-C22-P2-C31-C32-C33-P3-C41-C42-C43-P4-C51-C52-C53-P5-FC1-FC2-FC3). Convolutions and fully connected layers have trainable parameters. Three last fully connected layers provide task specific features. So, we focus on first 5 convolution layers. We chose $L = 3$ to achieve a trade-off between accuracy and complexity. The number of clusters in K-means clustering was also evaluated for $K = 30,40,50,60$ and chosen as $K = 50$ based on accuracy evaluation. We decided on the value of threshold for WEMD comparison based on average distance from training samples representations, since the training dataset consists only of normal samples.

For Ped 1, we compare our proposed approach both to traditional methods (SRC [6], MPPCA [43], and MDT [40]) and high-level deep learning-based methods (AVID [19], Sabokrou [8], and deep cascade [16]). As introduced and calculated in [26], evaluation metrics such as equal error rate (EER) and area under curve (AUC) are computed at frame level and compared to the state-of-the-art methods. EER indicates the point where false positive rate equals to false negative rate. The lower the EER is, the higher accuracy can be achieved. A comparison of EER of our proposed

Table 3: Comparison of AUC performance for the UCSD Ped1 dataset at frame level.

Method	SRC [6]	MPPCA [43]	MDT [40]	AVID [19]	Sabokrou [8]	Deep cascade [16]	Proposed approach
EER	19	40	25	12.3	8.4	9.1	8.1
AUC	86	59	81.8	—	93.2	—	93.9

Table 4: Comparison of EER performance for the UCSD Ped2 dataset at the frame level.

Method	SF [5]	MPPCA [43]	MDT [40]	Conv-AE [44]	AVID [19]	Deep anomaly [18]	Deep cascade [16]	ALOCC [17]	ST-AE [45]	Proposed approach
EER	42	36.0	24.0	21.7	14.	13.5	9.	13	12.0	6.1
AUC	63	71	85	90	—	—	—	—	87.4	97.3

Table 5: Accuracy criteria for the Ped 1 and Ped 2 datasets.

Dataset/criteria	Ped1	Ped2
Accuracy	90.3	95.4

approach to the previous method is shown in Table 3 for Ped1. Results show the comparable performance for our proposed method. Besides, AUC as the area under ROC curve is computed and compared to the state-of-the-art. Results show the outperformance of our proposed approach in AUC, as well.

For Ped 2, The Ped1 dataset suffers from the perspective problem. For this reason, most researches have been conducted on Ped2. We compare our proposed approach both to traditional methods (SF [5], MPPCA [43], and MDT [40]) and high-level deep learning-based methods (Conv-AE [44], AVID [19], deep anomaly [18], deep cascade [16], ALOCC [17], and ST-AE [45]). A comparison of EER of our proposed approach to the previous method is shown in Table 4 for Ped2. Results show the comparable performance for our proposed method. Besides, AUC is computed and compared to the state-of-the-art. Results show the outperformance of our proposed approach in AUC.

Moreover, we evaluated accuracy as

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}. \tag{10}$$

The results, shown in Table 5 for the Ped1 and Ped2 datasets, indicate the high performance of our proposed method.

5. Conclusions

In this paper, we discussed a new semantic and statistical distance-based crowd anomaly detection at the frame level. In particular, inspired by the earth mover distance metric applied previously on low-level vision features, we applied this statistical distance to hierarchically learned features, through pretrained deep convolutional neural network and topic model, for anomaly detection. Features from VGG-Net, pretrained on hybrid dataset (Places dataset and ImageNet dataset) and multilayered NMF as semantic interpretable features, were computed in combination as hierarchical representation and used in clustering-based anomaly detection using wavelet EMD statistical distance. Experimental results show the outperformance of our proposed approach. In the future, we will investigate anomaly localization by patch analysis through the kernel convolutional network (CKN) [46] and EMD in a similar framework to localize anomalies.

Data Availability

The readers can access the UCSD Ped1 and Ped2 datasets in http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

Conflicts of Interest

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

References

[1] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, 2015.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems the MIT press:neurIPS proceedings*, pp. 3111–3119, 2013.

[3] P. Wiriyathammabhum, D. Summers-Stay, C. Fermuller, and Y. Aloimonos, "Computer vision and natural language processing: recent approaches in multimedia and robotics," *ACM Computing Surveys (CSUR)*, vol. 49, no. 4, pp. 1–44, 2016.

[4] R. Wang, K. Nie, T. Wang, Y. Yang, and B. Long, "Deep learning for anomaly detection," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 894–896, Houston, TX, USA, 2020.

[5] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 935–942, Miami, FL, USA, 2009.

[6] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR 2011*, pp. 3449–3456, Colorado Springs, CO, USA, 2011.

[7] X. Zhu, J. Liu, J. Wang, C. Li, and H. Lu, "Sparse representation for robust abnormality detection in crowded scenes," *Pattern Recognition*, vol. 47, no. 5, pp. 1791–1799, 2014.

[8] M. Sabokrou, M. Fathy, Z. Moayed, and R. Klette, "Fast and accurate detection and localization of abnormal behavior in crowded scenes," *Machine Vision and Applications*, vol. 28, no. 8, pp. 965–985, 2017.

[9] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned ConvNets," *Neurocomputing*, vol. 371, pp. 188–198, 2020.

[10] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semisupervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

[11] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deepconvolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4305–4314, Boston, MA, USA, 2015.

[12] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1798–1807, Boston, MA, USA, 2015.

[13] G. Wimmer, A. Vécsei, M. Häfner, and A. Uhl, "Fisher encoding of convolutional neural network features for endoscopic image classification," *Journal of Medical Imaging*, vol. 5, no. 3, article 034504, 2018.

[14] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowded scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 56–62, Boston, MA, USA, 2015.

[15] M. Sabokrou, M. Fathy, and M. Hoseini, "Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder," *Electronics Letters*, vol. 52, no. 13, pp. 1122–1124, 2016.

[16] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992–2004, 2017.

[17] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3379–3388, Salt Lake City, UT, USA, 2018.

[18] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, vol. 172, pp. 88–97, 2018.

[19] M. Sabokrou, M. Pourreza, M. Fayyaz et al., "Avid: adversarial visual irregularity detection," in *Asian Conference on Computer Vision*, pp. 488–505, Springer, Cham, 2018.

[20] M. Sabokrou, M. Khalooei, and E. Adeli, "Self-supervised representation learning via neighborhoodrelational encoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8010–8019, Seoul, Korea (South), 2019.

[21] M. Sabokrou, M. Fathy, G. Zhao, and E. Adeli, "Deep end-to-end one-class classifier," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 2, pp. 675–684, 2021.

[22] L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft, "Image anomaly detection with generative adversarial networks," in *Joint European conference on machine learning and knowledge discovery in databases*, pp. 3–17, Springer, Cham, 2018.

[23] L. Ruff, R. A. Vandermeulen, N. Görnitz et al., "Deep semi-supervised anomaly detection," 2019, arXiv preprint arXiv:1906.02694.

[24] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 147–153, 2015.

[25] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception by hierarchical Bayesian models," in *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8, Minneapolis, MN, USA, 2007.

[26] X. Wang and E. Grimson, "Spatial latent Dirichlet allocation," in *Advances in neural information processing systems. MIT Press: Cambridge*, pp. 1577–1584, MA, USA, London, UK, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[28] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[29] O. Isupova, D. Kuzin, and L. Mihaylova, "Learning methods for dynamic topic modeling in automated behavior analysis," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 9, pp. 3980–3993, 2018.

[30] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *European Conference on Computer Vision*, pp. 517–530, Springer, Berlin, Heidelberg, 2006.

[31] O. P. Popoola and Kejun Wang, "Video-based abnormal human behavior recognition| a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 865–878, 2012.

[32] D. Pathak, A. Sharang, and A. Mukerjee, "Anomaly localization in topic based analysis of surveillance videos," in *2015 IEEE Winter Conference on Applications of Computer Vision*, pp. 389–395, Waikoloa, HI, USA, 2015.

[33] Bo du and Liangpei Zhang, "A discriminative metric learning based anomaly detection method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 6844–6857, 2014.

[34] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[35] M. A. Ruzon and C. Tomasi, "Edge, junction, and corner detection using color distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1281–1295, 2001.

[36] S. Shirdhonkar and D. W. Jacobs, "Approximate earth mover's distance in linear time," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Anchorage, AK, USA, 2008.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv preprint arXiv:1409.1556.

[38] H. A. Song, B. K. Kim, T. L. Xuan, and S. Y. Lee, "Hierarchical feature extraction by multi-layer non-negative matrix factorization network for classification task," *Neurocomputing*, vol. 165, pp. 63–74, 2015.

[39] X. Wan, "A novel document similarity measure based on earth mover's distance," *Information Sciences*, vol. 177, no. 18, pp. 3718–3730, 2007.

[40] Weixin Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, 2014.

[41] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: a 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[42] L. Wang, Z. Wang, W. Du, and Y. Qiao, "Objectscene convolutional neural networks for event recognition in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 30–35, Boston, MA, USA, 2015.

[43] J. Kim and K. Grauman, "Observe locally, infer globally: a spacetime MRF for detecting abnormal activities with incremental updates," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 2921–2928, Miami, Fla, USA, 2009.

[44] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742, Las Vegas, NV, USA, 2016.

[45] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X. S. Hua, "Spatio-temporal autoencoder for video anomaly detection," in *Proceedings of the 25th ACM international conference on multimedia, Mountain View*, pp. 1933–1941, California, USA, 2017.

[46] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid, "Convolutional kernel networks," *Advances in Neural Information Processing Systems the MIT press:neurIPS proceedings*, pp. 2627–2635, 2014.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5555362, 10 pages
https://doi.org/10.1155/2021/5555362

Review Article

Searchable Encryption with Access Control in Industrial Internet of Things (IIoT)

Jawhara Bader[1,2] **and Anna Lito Michala**[1]

[1]*School of Computing Science, University of Glasgow, Glasgow G12 8RZ, UK*
[2]*Department of Computer Science, Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Saudi Arabia*

Correspondence should be addressed to Jawhara Bader; j.alamri.1@research.gla.ac.uk

Received 4 February 2021; Revised 31 March 2021; Accepted 29 April 2021; Published 17 May 2021

Academic Editor: Mohammad R. Khosravi

The technological advancements in the Internet of Things (IoT) and related technologies lead to revolutionary advancements in many sectors. One of these sectors, is the industrial sector red that leverages IoT technologies forming the Industrial Internet of Things (IIoT). IIoT has the potential to enhance the manufacturing process by improving the quality, trace-ability, and integrity of the industrial processes. The enhancement of the manufacturing process is achieved by deploying IoT devices (sensors) across the manufacturing facilities; therefore, monitoring systems are required to collect (from multiple locations) and analyse the data, most likely in the cloud. As a result, IIoT monitoring systems should be secure, preserve the privacy, and provide real-time responses for critical decision-making. In this review, we identified a gap in the state-of-the-art of secure IIoT and propose a set of criteria for secure and privacy preserving IIoT systems to enhance efficiency and deliver better IIoT applications.

1. Introduction

The Internet of Things (IoT) has gained enormous popularity in the last decade, which consists of interconnected devices such as mobile phones, computers, sensors, and many more. These devices helped to develop and improve many sectors, such as Smart Cities, Smart Homes, and Healthcare [1]. The significant improvement added to these sectors encouraged the industrial sector to introduce IoT into the manufacturing paradigm. As a result, this led to a new industrial revolution: Industry 4.0. A new term Industrial Internet of Things (IIoT) has been used to collectively refer to proposed IoT solutions in this space [2].

The applications of IIoT can be classified into four categories (Figure 1). The first class includes production flow, quality control, and energy consumption. This class is aimed at improving production processes. The second class is operation-oriented management, which includes supply chain and enterprise decision management. The third class focuses on the allocation and collaboration of resources. This class includes collaborative manufacturing and customization technology. Finally, the last class mainly focuses on product life cycle management. Additionally, it focuses on service optimization, such as remote maintenance and product traceability [2].

The growing population has led to an increasing demand for products, which has saturated the manufacturing industry and even more so in the recent COVID crisis. As a result, the manufacturing segment is expected to have the highest and fastest-growing market segment by end-user at a compound annual growth rate (CAGR) of 27.94%. To meet the growing demand, an efficient manufacturing system has become mandatory. This demand can only be achieved by the integration of the latest technologies, such as IoT within the manufacturing process [3]. However, the stringent regulatory requirements (COMAH, IEC, and SIL) for safety must be satisfied for this shift to become usable in real-world applications.

To demonstrate the relevance of IIoT and its ability to meet regulatory requirements, we present two examples of manufacturers currently using IIoT schemes [4]. The first one is Airbus, the European aircraft manufacturer. Airbus currently integrates IoT technologies into its products and its workers' tools in the manufacturing process. Also, Airbus

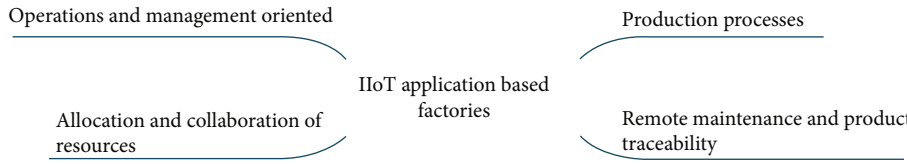


Figure 1: IIoT application-based factories.

is harnessing IoT technologies to clear a backlog of orders and boost revenues. It is clear that IoT is transforming the aviation industry by enabling a more seamless passenger journey, increasing operational efficiency, and driving a new age of "connected aviation." The second example is the global tech firm Client Global Insights (CGI). CGI has teamed up with Microsoft to deliver a predictive maintenance solution for elevators by leveraging IoT. CGI claims that it has developed a solution which securely connects thousands of sensors and systems within elevators and monitors everything from motor temperature to shaft alignment. The data are collected and processed on the cloud using Microsoft's cloud-based Azure Intelligent Systems Service. These elevators enable technicians to use real-time IIoT data to spot defects and repair them before a breakdown occurs.

Thus IIoT, such as monitoring systems, help industries improve their resources and meet their clients' needs while ensuring high-quality production. IIoT achieves this by providing ubiquitous connectivity, efficient data analytics tools, better decision support systems, and applications [5, 6]. As IIoT applications deal with complicated processes, these applications have a critical impact on several parties. For example, a failure in an IIoT application may put the employees' lives in the factory at severe risk. Similarly, the business resources may be at risk which has cost implications [7]. Therefore, the accuracy, precision, and risk impact of application-failure metrics of IIoT applications should be higher than in IoT applications. Moreover, IIoT applications must fulfil the stringent requirements of real-time processing and feedback, time synchronisation, and regular communication [8].

The three essential components of security are confidentiality, integrity, and availability, which are known as the CIA triangle. Confidentiality ensures that only authorised users can read the data of a system. Integrity ensures that no changes are made to the data, and availability means that all services and data are available [9]. Availability and integrity of data are considered more essential than confidentiality for industrial environments. This, however, does not diminish the need for confidentiality. With the internet-connected systems of IIoT, all three aspects should be brought up to an acceptable level. Thus, in the development of new IIoT and Industry 4.0 systems that leverage the existing network and cloud infrastructure, confidentiality and integrity should be weighed equally to availability [10].

IIoT applications, such as monitoring systems in smart factories, work by collecting data from multiple locations and analysing the data on the cloud. However, collecting and processing data on the cloud compromise the data privacy and security, leading to sensitive information leakage [11, 12]. Data-at-rest must be securely stored and processed on the cloud without compromising its security and privacy [11, 13]. This goal is challenging as it requires processing the encrypted data (not the plaintext) on the cloud. Moreover, some IIoT applications require access control (AC) policies to allow specific users, such as a manager or a third-party contractor to access and query the data. This AC requirements adds a more significant challenge [14, 15]. To highlight the need for data confidentiality and integrity in IIoT applications, the authors in [16] demonstrated how collecting air-quality-related data on unsecure servers can misinform the public or mislead policymakers. The authors showed that any modification to the sensors' data can lead to false-negative emergency alerts or wrong decisions. An example of wrong decisions is triggering the evacuation alarm or stopping a production line. Therefore, it is crucial for IIoT applications to secure the collected data. This can be achieved by encrypting the collected data during transmission and at-rest [17].

The degree of severity in relation to violating privacy on the IIoT differs from that of the IoT. In IoT, unauthorised access may lead to privacy problems such as data theft. On the other hand, violating privacy on IIoT may lead to a disastrous decision that can cause the entire system to fail [7]. IoT and IIoT may share similar security threats. Yet, there is a substantial difference between the degree of severity in the event of a security breach in both IoT and IIoT [18]. In other words, authenticating an illegitimate device may cause a normal IoT system to experience some problems, such as privacy invasion. On the other hand, a similar scenario in an IIoT system could cause serious consequences. For example, disrupting the network or forcing the network to take hazardous actions. Thus, IIoT requires a higher level of security. To do so, there are several factors to consider, such as the applications' requirements, the type of IIoT devices, and a recovery technique in the case of cybersecurity attack [19].

In addition to the previous security and privacy requirements, factories need to share the data with other parties, such as insurance or/and consulting companies, customers, and/or employees. To control and manage data access, IIoT systems must deploy AC mechanisms on the encrypted data on the cloud [20].

Several recent studies have addressed the security and privacy issues for IIoT, from different perspectives. For example, the authors of [17] categorised the security challenges for both IoT and IIoT. The authors also specified whether these challenges are applicable to IoT or IIoT or both. On the other hand, the study demonstrates the security challenges in the IIoT and stresses the need to design practical solutions. Also, the study shows that various IIoT scenarios require application-specific designs. However, solutions to the challenge of appropriate designs are not suggested by the authors.

In [21], the authors analysed the security challenges of IIoT and provided a comparative analysis of the available solutions. This study set out to identify some open research problems related to system integration, communication, energy factor, preventive and detective measures, authorisation, and architecture of IIoT. However, the study does not suggest feasible and practical solutions. Similarity, Tange et al. [10] provide a systematic literature review of IIoT security requirements. The authors demonstrate how fog computing can address these requirements. Additionally, the authors identified some research opportunity to use secure fog computing for IIoT.

uilding on existing findings from [10, 17, 21], in this article, we examined the practical considerations of embedding security and privacy solutions to IIoT system architectures moving away from the cloud paradigm to minimise exposure to threats. Thus, we focus on combining searchable encryption and access control methods in a cloud-Edge architecture to assess their suitability and efficiency from the privacy, security, and response time perspectives.

2. Objectives

In the context of IIoT, privacy refers to protecting the confidentiality of the IIoT device and its collected readings (data). For example, exposing the sensor's location is considered a security and safety threat. The lack of privacy preservation causes security threats, such as in the case of utility monitoring, which will affect the network's process [19]. To address the aforementioned requirements for security and privacy in IIoT monitoring systems, we must consider the following challenges:

(1) *The Limited Resources in IIoT Devices, such as Low Computational Power, Low Power Consumption, and Low Storage*. Therefore, deploying and running encryption algorithms on these devices may add significant performance overhead.

(2) *The Adaptation of Searchable Encryption (SE)*. Searching the encrypted data on the cloud requires adapting and enhancing searchable encryption (SE) algorithms to work on both the cloud and IIoT devices.

(3) *The Critical Real-Time Requirement for IIoT Systems*. IIoT monitoring systems should fulfill the critical real-time requirement, which significantly affects the decision-making process. Thus, each component in the system should be optimised to reduce the overall execution time.

The security aspects and requirements in IIoT can be identified as follows [22]:

(i) *Impact of Attack*. A successful attack on an IIoT system has a high impact due to the critical nature of this industry.

(ii) *Secure Communication*. It is important to maintain a secure connection between IIoT parties. Unsecure communication channels can expose sensitive information.

(iii) *Authentication and Authorisation*. IIoT requires authentication and authorisation for all connected devices. This includes but not limited to, sensors, internal users, and external users.

(iv) *Accountability*. It is important to keep track of all actions and incidents in an IIoT system to identify and recover from any possible incidents. According to [23], the main security concerns are authentication and access control. The reason is that users with improper access rights can severely affect these systems.

In this paper, we aimed to examine the state-of-the-art in security and privacy of IIoT application systems and focus on the combination between searchable encryption and access control methods applied on Edge computing to assess their suitability and efficiently from a privacy, security, and response time perspective. Hence, the present article is aimed at fostering scientific discussion regarding IIoT from the privacy and security perspective under constraints by revealing the current state of research as well as identifying areas to be addressed by future research efforts. By doing so, the following research questions are pursued:

(i) *RQ1*. Which research areas and methods have addressed the security, privacy, and efficiently performance of IIoT and to what extend so far?

(ii) *RQ2*. Which research areas or methods can be proposed to further security and privacy improvement in the future of smart factories?

To answer the abovementioned questions, a systematic review of relevant literature is applied as presented in Section 3. The results of the literature search are presented in Section 4 where we taxonomise relevant research. A discussion of the findings along with open challenges and suggested areas of further research is presented in Section 5 with conclusions presented in Section 6.

3. Methods

We selected a list of publications related to IIoT systems to be included in the comparative analysis. The databases and sources used in this systematic review include (1) IEEE, (2) Springer, (3) websites of smart factories found through a generic Google search, and (4) Google Scholar (including ResearchGate).

We focused on several topics, including security in IIoT, enhanced searchable encryption algorithms, and a combination between searchable encryption, and access control methods. The search keywords alongside the number of results are presented in Table 1.

As the above combination of data sources and keywords returned a vast amount of results, we selected the following inclusion criteria to identify the most relevant sources: (1) language: English, (2) date range: within the past five years (2017-2020), (3) the article presents a review or a survey

Table 1: Results of the literature search.

Topic	Online library	Number of results
Industrial Internet of Things applications	IEEE	2,845
	Springer	44,937
	Google search	53,000
	Google Scholar	100,000
Searchable encryption for IIoT	IEEE	9
	Springer	8
	Google search	124
	Google Scholar	120
Access control for IIoT	IEEE	101
	Springer	898
	Google search	5,600
	Google Scholar	5,000
Privacy preserving for IIoT applications	IEEE	19
	Springer	114
	Google search	1,000
	Google Scholar	1,110
Edge computing in IIoT application	IEEE	89
	Springer	483
	Google search	3,150
	Google Scholar	3,000
Requirements of IIoT	IEEE	238
	Springer	1,057
	Google search	5,000
	Google Scholar	4,580
Searchable encryption with access control	IEEE	105
	Springer	1,168
	Google search	2,300
	Google Scholar	10,000
Monitoring system using IIoT	IEEE	125
	Springer	967
	Google search	5,200
	Google Scholar	15,100
Security of IIoT applications	IEEE	197
	Springer	853
	Google search	4,600
	Google Scholar	5,030

related to IIoT, and (4) relevance: searchable encryption with access control for Edge-based IIoT application is necessary. The exclusion criteria are as follows: (1) nonrelated to the relevance inclusion criteria, (2) implicitly related the relevance inclusion criteria, (3) duplicate articles that appear multiple times in one or more databases, and (4) nonresearch article.

Those four exclusion criteria and four inclusion criteria as illustrated above improved the objectivity of this review paper. A filtering process was carried out to exclude those articles that fulfill the exclusion criteria. The remaining articles were classified according to the four inclusion criteria, and data of interest was collected.

4. Results

The literature search returned a total of 268,507 results. Overall, we read 235 sources, as we excluded the majority by reading the abstracts. A total of 54 sources remained for analysis and were taxonomised as seen in Figure 2.

4.1. State-of-the-Art in IIoT with Embedded Security Mechanisms. IIoT systems can benefit from the massive amount of collected data to generate a useful approach. This approach can improve the performance of the system and minimise unplanned downtime [24]. IIoT systems utilise cloud servers to store and process the generated massive data [24]. However, the data need time to be transferred to centralized data centres, which degrades the IIoT system efficiency. This implies that processing data on an Edge server could help the IIoT system meet real-time requirements and reduce the decision-making latency [25]. The survey presented in [17] identified two constraints when protecting data confidentiality in IIoT systems through data encryption. One of these constraints is related to the limited resources of IIoT devices.

Gebremichael et al. [19] describe the privacy challenges in IIoT based on the levels of the architecture as follows: device, platform, and application layers. The solutions provide access control methods, authentication mechanisms, data encryption, and secure channels to ensure the privacy at the device layer, for example, protecting Edge nodes against a fake node insertion attack. They also describe several points that developers need to consider when designing privacy solutions for the IIoT. These points can be described as follows:

(i) Cryptographic mechanisms are generally employed to enforce privacy policies. The challenge is to design a lightweight privacy-enhancing cryptosystem suitable for IIoT devices. These IIoT devices have limited resources. Thus, it is crucial to prevent heavy computations to meet the IIoT real-time requirement

(ii) Further research is needed to provide lightweight cryptosystem solutions with anonymised data methods. Also, advanced data analytics tools to process the collected data

(iii) Reducing the amount of data collected by Edge devices to the minimum data points that are required for system operations while continuing to provide anonymisation techniques on user data

(iv) Illustrating data access policies and implementing appropriate access control methods that are capable of identifying authorised users that have access rights to Edge node data

Several solutions can protect IIoT systems' privacy, such as encryption, access control, processing data on the Edge, and anonymisation. Privacy in IIoT systems is challenging as these systems usually store and process data in third-party cloud services.

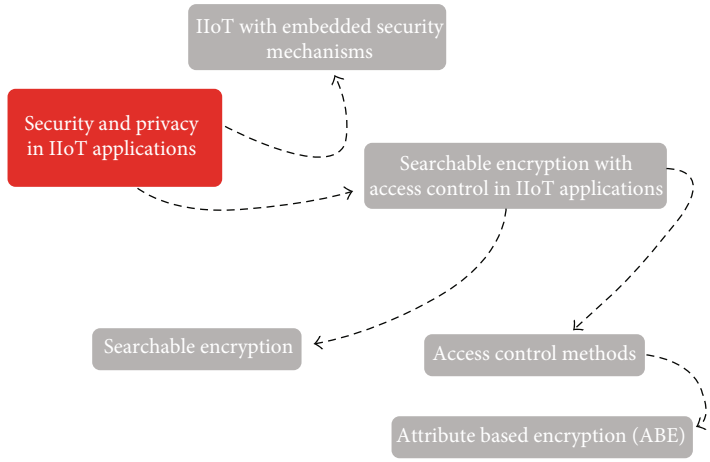


Figure 2: Literature article categories.

Yu et al. state that the data generated by IoT devices has increased dramatically. At the same time, Cisco predicted that the volume of data generated by IoT devices would reach 500 ZB by 2019 [26]. However, massive data need time to be transferred directly to the cloud for processing, which adds computation overhead. This computation overhead increases the latency, bandwidth, and may even lead to the unavailability of IIoT applications [2]. To address this issue, the concept of Edge/fog computing has been defined, and data can now be processed much closer to the source. This is because some cloud services are brought to the Edge of the network. In this context, fog computing differs from Edge computing in that it uses the interconnection between endpoints. Edge computing, on the other hand, focuses on isolated endpoints [26]. This implies that processing data on the Edge server helps the IIoT system to meet the real-time requirement and reduce decision-making latency, especially for delay-sensitive applications [2]. Edge computing is applied to manufacturing based on IoT to meet these requirements [27].

Many researchers introduced improvements to searchable encryption algorithms that would make them lightweight for IIoT, such as [28]. Yet, this method is not tested for its applicability in industrial plants. Wazid et al. [29] review the access control in IIoT such as a monitoring system of an industrial plant. They state that authentication is the most important security requirement in cloud-based IIoT while this requirement is still needed to improve the proposed solution.

The following subsections will discuss data analtyics, searchable encryption, and access control state-of-the-art methods that have the potential to address the challenges identified in this subsection.

4.2. Edge Data Analytics for IIoT Applications. Data analytics is the most important step in the monitoring system's life cycle. IoT data analytics improves fault detection, disaster forecasting, service, and smart decision-making [30]. Moreover, they help the smart factory extract the knowledge from raw data with the support of IIoT applications, for example, to better understand technological enabler behaviour or to relate issues derived from combined and statistical data processing [31]. The usage of feature extraction methods provides more accurate data analysis results. Besides, meeting the real-time requirement for IIoT manufacturing applications, for instance, a robust incremental feature extraction method based on PCA (Principal Component Analysis) is proposed to meet the real-time requirement [30]. Extracting data features from the data by applying such techniques allows Edge servers to take smart decisions for delay-sensitive applications [27]. Applying Edge analytics directly reduces the volume of data to be transmitted to the cloud. This, in turn, reduces the information that must be encrypted, which makes the encryption overhead minor. However, this reduction introduces other challenges in terms of accuracy and traceability, especially in regard to the route cause fault finding capabilities. Thus, appropriate Edge data analytics methods must be identified to optimise the trade-off between benefit and side effects.

4.3. Searchable Encryption. Searchable encryption (SE) is a cryptographic technique that allows secure searching over encrypted data [32]. SE allows a user (or an automated program) to perform a secure query for a specific event without compromising the data confidentiality. For example, using SE to encrypt data on the cloud prevents the cloud provider or any unauthorised person (including the system administrator) from accessing or querying the encrypted data. There are two SE schemes [33]; one of these schemes is Symmetric Searchable Encryption (SSE). SSE requires a private key to be distributed between users, which is not suitable for multiple user scenarios [34]. The other scheme is Public Key Encryption with Keyword Search (PEKS) [35]. PEKS is a public-key cryptosystem that allows search over encrypted data using a public key instead of private keys, allowing multiple parties to query the data without compromising the data owner's private key.

4.4. Access Control Methods. There are several known AC mechanisms, including but not limited to attribute-based, key-policy-based, role-based, and trust-based [36]. However, the most commonly used AC mechanisms with PEKS are

role- and attribute-based access control. Table 2 summarises the difference between these two AC mechanisms, based on two recent publications [37, 38]. The following subsections will further discuss those approaches and their capacity to be combined with SE and critically compare them in the context of IIoT.

4.5. Role-Based Access Control (RBAC) with PEKS. RBAC is a security mechanism that allows users to access data based on their roles within an organisation [39]. The authors in [40] introduced RBAC to PEKS using free bilinear, as bilinears have high computational cost. The authors used the RBAC mechanism to simplify the frequent user's permission assignment within a large organisation. However, using RBAC with PEKS makes it hard to manage third parties' access policies (users outsides the organisation), which is an essential requirement for a monitoring system in the IIoT. Besides, using RBAC with PEKS is inflexible as it must be painstakingly managed.

4.6. Attribute-Based Encryption (ABE). Attribute-based access control (ABAC) is a security mechanism that allows organisations to grant access to users based on some attributes, such as their division or title [41]. On the other hand, Attribute-Based Encryption (ABE) combines searchable encryption with the ABAC approach [42]. In ABE, a message is encrypted for a specific receiver using a set of attributes. Thus, only the person who holds a key for the matching attributes can decrypt the message [39]. ABE has two paradigms: Key-Policy ABE (KP-ABE) and Ciphertext-Policy ABE (CP-ABE). In KP-ABE, the user's private key is associated with a specified access policy, and the ciphertext is encrypted under a set of attributes. The user can decrypt the ciphertext if the attributes in ciphertext satisfy the access policy in the user's key. Thus, KP-ABE mechanism answers the following question, "what type of data should the user access?". Differently, the CP-ABE answers the question, "What attributes must a user have to access the encrypted data?". Typically, CP-ABE is considered an adjustable scheme because it guarantees more control to the user over the encrypted data [43].

Rasori [43] improved ABE and reduced the communication overhead by 35 per cent compared with existing ABE for medical applications. This novel CP-ABE is more efficient and could be a suitable solution for low-power communication protocols in IIoT. Sathya and Kumar [44] proposed a medical system that collects patient's data during emergencies and shares the data with the doctors. The authors' proposed system combines blowfish encryption and an ABE scheme. The authors evaluated their proposed system using several symmetric encryption algorithms, encryption time, decryption time, and total computation time. Their evaluation shows that the blowfish algorithm has better performance to encrypt data when used with CP-ABE to grant the authorised users' access to medical data. The main advantage of this work is the fast transmission of medical data, while the main disadvantage of using the blowfish algorithm is the linear relationship between the size of ciphertext and the number of attributes. When the number of attributes increases, so does the size of ciphertext.

Miao et al. [45] proposed a higher security level PEKS with CP-ABE approach that supports access control with multiple permissions as well as hidden access policies. Also, the authors employed traceability techniques to prevent dishonest data users from leaking their private key to others. Their evaluations show that the computation costs for encryption and decryption increase linearly as the number of user attributes does.

Yang et al. [46] proposed a system to monitor the patient's status with two access control modes. The first mode is for normal situations where the doctors, nurses, and technical staff have access under an access policy. The second mode is for emergencies where the first-aider needs access to the patient's historical data. To achieve these controlled access modes, the authors applied ABE for normal access and break-glass algorithm for emergency access. However, their approach provides data security but does not provide a revocation mechanism to the emergency access policy, once the situation is resolved.

4.7. Attribute-Based Keyword Search (ABKS). In the Attribute-Based Keyword Search (ABKS) scheme, the keywords are encrypted by an AC policy and the data with attributes. The user can generate a trapdoor that can be used to search over encrypted data [47]. The ABSE (attribute-based searchable encryption) scheme has exactly the contrary where the owner transmits the valid search query to the user and allows them to decrypt the data when its attributes satisfy the access policy [48]. However, ABKS schemes provide efficient search operations which allow retrieving encrypted data for multiple authorised users with flexible access policy [49].

Guo et al. [50] proposed a new ABKS to support encryption for both keyword and messages where most existing ABKS encrypts the keyword. In their proposed ABKS, there is no need for a secure channel to transmit the search tokens to the cloud. Also, it is a robust scheme against resisting offline keyword guessing attacks by inside attackers (i.e., the honest-but-curious servers). This scheme is evaluated and applied to a telemedicine system that is used to support healthcare services at multiple locations. However, the communication time in this scheme is high and is not suitable for time-sensitive applications.

4.8. Combining Searchable Encryption with Access Control. To achieve strong confidentiality, SE must be combined with access control [51, 52]: if a ciphertext appears as a search result, we learn something about the underlying document, even if the access control does not allow us to access the document. This illustrates the need for a linked search and access control, so that search results present to users only data to be accessed by the users [53]. Thus, the SE protects data confidentiality, and AC schemes protect user access privileges [54].

It is essential to protect data that travels through the IIoT network. Thus, SE covers cryptographic protection across all networks by (1) protecting the Edge and cloud networking and (2) protecting endpoint connectivity [9]. Encryption techniques protect the privacy of big data in the data storage phase. Confidentiality, the first consideration when the encrypted data is stored in cloud servers can be secured by

Table 2: Comparing role and attribute-based access control.

Feature	Access control mechanism	
	Role-based access control (RBAC)	Attribute-based access control (ABAC)
Access control granularity	Coarse-grain access control	Fine-grain access controls
User addition mechanism	Creating access control groups defined as roles with presetup privileges. Users can be added into the group for their desired access privileges.	Users are assigned attributes to describe their properties. The access control system needs to focus on the required access control policies that are described by a set of attributes to check the user's privileges to decide if the access should be granted or not.
Structure of access policy	Policies are assigned (operation/object pairs) to groups before the access request is made.	Using Boolean rule structure to express the policies.
The input of authorisation decisions	Users are assigned to roles and inherit the permissions assigned to the roles they have. Roles are often organised in a role hierarchy, which defines the inheritance of permissions between roles.	They are used as input for authorisation decisions with many criteria, such as department, job code, time of day, IP address, and user location.
Decision level	Only related to functionality	Relate to access in both the data level and the field level, but also to functionality.
Access level	Do not allow access for nonemployees to organisation assets.	Allow limited access for third parties to organisational assets.
Model status	One of the main problems is that it is not an automatic model, needs to be painstakingly managed, and often involves significant manual intervention. The role-based mechanism, by itself, is inadequate to address the dynamic requirements of cloud-based IoT.	The ABAC model is a dynamic model. The system dynamically deploys access control by using attributes, i.e., a flexible access control approach.

efficient encryption techniques. However, when the data user sends the request to retrieve the data from the cloud, the cloud server cannot reply to the user's request, because it cannot decrypt the encrypted data or search over encrypted data. Searchable encryption schemes could address these challenges.

While the Attribute-Based encryption (ABE) methods might secure information transmission and the fine-grained sharing of encrypted IIoT data, they additionally need to overcome new application deterrents in IIoT-cloud frameworks: (1) restricted resource IoT devices; (2) difficulty in encrypted data recovery at cloud servers: the encrypted records limit the adaptability and accuracy of information recovery, leading to unessential or incorrect outcomes; (3) lack of successful key administration: once CA is compromised, all previously encrypted files can be leaked because of the keys generated by a central authority (CA). To address the above difficulties, a novel lightweight searchable encryption method is needed for IIoT-cloud frameworks [55].

4.9. Searchable Encryption with Access Control in IIoT Applications. The literature survey of Zhou et al. [33], which spanned 2014 to 2019, identified schemes that combined PEKS with Attribute-Based Encryption (PEKS-ABE) for cloud-based applications. Moreover, this survey demonstrated that the PEKS-ABE provides efficient data sharing and searching ability, but it needs to improve the privacy of user keys. However, they do not also apply it to IIoT wherein to improve the privacy of the user keys, an Edge processing and storage approach could be utilised.

The following two works focus on improving either SE or AC for IIoT environments, but they do not combine them. Chen et al. [28] proposed lightweight searchable encryption for cloud-based IIoT applications with security improvements. In [56], published in 2020, they improve CP-ABE in many aspects:

(1) *Using a Hybrid Cloud Infrastructure*. Public cloud to store encrypted IoT data and the private cloud to execute CP-ABE tasks over the data

(2) *Guaranteeing Data Privacy at the User Level against the Private Cloud*. The author achieved this by proposing two encryption techniques. These techniques work by protecting IoT data privacy at the item level and preventing the user-key leakage problem.

(3) Enabling the private cloud to execute CP-ABE encryption/decryption tasks in batches and executing the CP-ABE reencryption tasks regardless of the size of IoT data, thus improving the performance of IIoT applications

Chen et al. [28] proposed lightweight searchable encryption for cloud-based IIoT applications with security improvements. To achieve more precise data retrieval, Miao et al. [57] proposed an improved ABE scheme with multikeyword search to support simultaneous numeric attribute comparison, thereby greatly enhancing the flexibility of ABE encryption in a dynamic IoT environment. Furthermore, attribute-based

multikeyword search schemes were also investigated in [58]. Nevertheless, this CP-ABE scheme inevitably concentrates on the single authority environment in which a CA essentially controls all attributes' authorisation. The single authorisation cannot effectively generate and manage the public/secret keys in the IIoT.

However, these studies did not improve the bandwidth of data that is outsourced to the cloud, which is important to minimise the computational cost. Zhang et al. [55] proposed a lightweight SE-AC scheme by providing lower computational complexity. Moreover, their framework enhanced privacy by preventing leakage during data outsourcing to a cloud server. In summary, they provide fine-grained AC, multikeyword search, lightweight decryption, and a multi-authority environment. They provide low latency as well as improved security against the chosen-keyword attack and the chosen-plaintext attack. Their LSABE and LSABE-MA schemes can support single keyword and multikeyword searching while maintaining the lightweight decryption on many practical testing platforms (PC, mobile phone, and Raspberry Pi models). Moreover, their schemes meet the low-latency requirement of IIoT applications. Therefore, their schemes are suitable for practical IIoT environments. However, their work did not consider the accuracy and data bandwidth, which is regarded as requirements of IIoT applications. In addition, the encryption time for their schemes is 24 seconds. Simultaneously, latency is an important metric in the encryption phase for the real-world IIoT environment. Thus, encrypted privacy-sensitive data must upload to the cloud immediately. Hence, we identify a gap in extracting the useful information from the raw data before encrypting them to minimise the encryption time and the bandwidth and to improve the overall performance to meet IIoT requirements.

5. Discussion

Several studies have combined SE with AC to query encrypted data with different AC policies. However, studies that combined PEKS and AC mechanisms, such as CP-ABE, still suffer from low privacy for user keys, high volumes of data transmission, or a high ratio of error for returned data (reduced accuracy). Some studies combined these algorithms in the medical domain to improve the privacy of medical data and the security level against external and internal attacks. Furthermore, some systems still have a high computational cost, which is not practical for a computationally restricted environment such as IIoT. This high computational cost prevents studies from meeting the real-time requirement for the time-sensitive IIoT applications. Therefore, IIoT applications must minimise the computational cost and improve performance to meet the near real-time requirements. Gebremichael et al. [19] discussed the further research that needs to be considered in the IIoT applications. The authors argue that using SE or homomorphic encryption (HE) can maintain security and privacy for systems that rely on cloud providers. Besides, SE provides fast and secure data delivery from the cloud for time-critical applications. Leading from the above discussion, we identify four research questions and open challenges as follows:

(i) *RQ1*. How do we adopt and deploy a lightweight version of the Public Key Encryption with Keyword Search (PEKS) algorithm on both the IIoT devices and the cloud to achieve a near real-time performance that is suitable for time-sensitive IIoT systems?

(ii) *RQ2*. How can we introduce, investigate, and evaluate the combination of PEKS and CP-ABE mechanisms in the cloud versus Edge architecture while achieving the best performance for time-sensitive IIoT systems?

(iii) *RQ3*. How do we investigate the performance overhead for deployment on the Edge vs. the cloud server on various IIoT applications and identify the proper architecture for each application type?

(iv) *RQ4*. How to design and develop a framework with an efficient CP-ABE mechanism and PEKS algorithm tailored to a suitable cloud and Edge deployment for IIoT systems to provide a secure and privacy-preserving solution for IIoT systems with AC support?

6. Conclusions

This study provided an unambiguous literature review that specifically focused on SE with AC for IIoT in a systematic manner. We demonstrated that the existing approaches and articles do not meet all the requirements of IIoT to support smart factory needs. Our review highlights the efficient combination between AC or CP-ABE and SE or PEKS to preserve privacy and minimise the execution time. These improvements can assist in taking smart decisions, specifically if deployed in an cloud-Edge architecture. However, the remaining open challenges need to be addressed to evaluate if these solutions can provide an efficient and reliable framework for IIoT applications.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by a scholarship through University of Tabuk in Saudi Arabia (TBU331). We would like to thank Dr. Jeremy Singer, Dr. Tim Storer, and Dr. Christos Anagnostopoulos for their feedback.

References

[1] M. Hermann, T. Pentek, and B. Otto, "Design principles for Industrie 4.0 scenarios: a literature review," in *2016 49th Hawaii international conference on system sciences (HICSS)*, pp. 3928–3937, Koloa, HI, USA, 2016.

[2] Y. Yu, R. Chen, H. Li, Y. Li, and A. Tian, "Toward data security in edge intelligent IIoT," *IEEE Network*, vol. 33, no. 5, pp. 20–26, 2019.

[3] P. R. Newswire, *Global Industrial IoT Market: Research Report 2015-2019*, Lon-Reportbuyer, 2015.

[4] F. Roberts, *9 examples of manufacturers making IIoT work for them*, Internet of Business, 2016.

[5] C. Liu, F. Chen, J. Zhu, Z. Zhang, C. Zhang, and C. Zhao, *Industrial IoT Technologies and Applications, Vol. 202*, Springer International Pu, 2017.

[6] P. Mathur, *IoT Machine Learning Applications in Telecom, Energy, and Agriculture*, Springer Nature Switzerland AG, 2020.

[7] G. Drosatos, K. Rantos, D. Karampatzakis, T. Lagkas, and P. Sarigiannidis, "Privacy-preserving solutions in the Industrial Internet of Things," in *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pp. 219–226, Marina del Rey, CA, USA, May 2020.

[8] STHE Web and FORA-r Devices, *IoT FOR BUSINESS Take Manufacturing's Shift Your Manufacturing Shift to Lightspeed to Lightspeed*, 2020.

[9] J. Wan, J. Li, M. Imran, D. Li, and Fazal-e-Amin, "A blockchain-based solution for enhancing security and privacy in smart factory," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3652–3660, 2019.

[10] K. Tange, M. De Donno, X. Fafoutis, and N. Dragoni, "A systematic survey of Industrial Internet of Things security: requirements and fog computing opportunities," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 4, pp. 2489–2520, 2020.

[11] M. S. Hossain and G. Muhammad, "Cloud-assisted Industrial Internet of Things (IIoT) - enabled framework for health monitoring," *Computer Networks*, vol. 101, pp. 192–202, 2016.

[12] A.-R. Sadeghi, C. Wachsmann, and M. Waidner, "Security and privacy challenges in Industrial Internet of Things," in *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, San Francisco, CA, USA, June 2015.

[13] P. P. Jayaraman, X. Yang, A. Yavari, D. Georgakopoulos, and X. Yi, "Privacy preserving Internet of Things: from privacy techniques to a blueprint architecture and efficient implementation," *Future Generation Computer Systems*, vol. 76, pp. 540–549, 2017.

[14] H. F. Atlam, A. Alenezi, R. K. Hussein, and G. B. Wills, "Validation of an adaptive risk-based access control model for the Internet of Things," *International Journal of Computer Network and Information Security*, vol. 10, no. 1, pp. 26–35, 2018.

[15] 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.

[16] L. Luo, Y. Zhang, B. Pearson, Z. Ling, H. Yu, and X. Fu, "On the security and data integrity of low-cost sensor networks for air quality monitoring," *Sensors*, vol. 18, no. 12, p. 4451, 2018.

[17] X. Yu and H. Guo, "A survey on IIoT security," in *2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS)*, pp. 1–5, Singapore, 2019.

[18] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of Things: architecture, enabling technologies, security and privacy, and applications," *Ieee Internet of Things Journal*, vol. 4, no. 5, pp. 1125–1142, 2017.

[19] T. Gebremichael, L. P. Ledwaba, M. H. Eldefrawy et al., "Security and privacy in the Industrial Internet of Things: current standards and future challenges," *IEEE Access*, vol. 8, pp. 152351–152366, 2020.

[20] K. Pothong, I. Brass, M. Carr et al., *Editors of the Cybersecurity of the Internet of Things: PETRAS Stream Report 03 Privacy and Trust 05 Adoption and Acceptability*, PETRAS Stream Report, 2019.

[21] P. Jayalaxmi, R. Saha, G. Kumar, N. Kumar, and T.-h. Kim, "A taxonomy of security issues in Industrial Internet-of-Things: scoping review for existing solutions, future implications, and research challenges," *IEEE Access*, vol. 9, pp. 1–1, 2021.

[22] A. Sciences, *Security and Privacy Trends in the Industrial Internet of Things*, Springer, Berlin, Germany, 2019.

[23] J. Sengupta, S. Ruj, and S. Das Bit, "A comprehensive survey on attacks, security issues and blockchain solutions for IoT and IIoT," *Journal of Network and Computer Applications*, vol. 149, article 102481, 2020.

[24] M. S. Virat, S. M. Bindu, B. Aishwarya, B. N. Dhanush, and M. R. Kounte, "Security and privacy challenges in Internet of Things," in *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics*, pp. 454–460, Tirunelveli, India, 2018.

[25] T. Qiu, J. Chi, X. Zhou, Z. Ning, M. Atiquzzaman, and D. O. Wu, "Edge computing in Industrial Internet of Things: architecture, advances and challenges," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2462–2488, 2020.

[26] I. Ungurean and N. C. Gaitan, "A software architecture for the Industrial Internet of Things—a conceptual model," *Sensors*, vol. 20, p. 5603, 2020.

[27] B. Chen, J. Wan, A. Celesti, D. Li, H. Abbas, and Q. Zhang, "Edge computing in IoT-based manufacturing," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 103–109, 2018.

[28] B. Chen, L. Wu, N. Kumar, K.-K. R. Choo, and D. He, "Lightweight searchable public-key encryption with forward privacy over IIoT outsourced data," *IEEE Transactions on Emerging Topics in Computing*, 2019.

[29] M. Wazid, A. K. Das, R. Hussain, G. Succi, and J. J. P. C. Rodrigues, "Authentication in cloud-driven IoT-based big data environment: Survey and outlook," *Journal of Systems Architecture*, vol. 97, pp. 185–196, 2019.

[30] X. Kong, J. Chang, M. Niu, X. Huang, J. Wang, and S. I. Chang, "Research on real time feature extraction method for complex manufacturing big data," *International Journal of Advanced Manufacturing Technology*, vol. 99, no. 5-8, pp. 1101–1108, 2018.

[31] T. P. Raptis, A. Passarella, and M. Conti, "Data management in industry 4.0: state of the art and open challenges," *IEEE Access*, vol. 7, pp. 97052–97093, 2019.

[32] K. Chamili, M. J. Nordin, W. Ismail, and A. Radman, "Searchable encryption: a review," *International Journal of Security and Its Applications*, vol. 11, no. 12, pp. 79–88, 2017.

[33] Y. Zhou, N. Li, Y. Tian, D. An, and L. Wang, "Public key encryption with keyword search in cloud: a survey," *Entropy*, vol. 22, no. 4, pp. 1–24, 2020.

[34] S. T. Hsu, C. C. Yang, and M. S. Hwang, "A study of public key encryption with keyword search," *International Journal of Network Security*, vol. 15, no. 2, pp. 71–79, 2013.

[35] S. Yu, C. Wang, K. Ren, and W. Lou, "Achieving secure, scalable, and fine-grained data access control in cloud computing," in *2010 Proceedings IEEE INFOCOM*, San Diego, CA, USA, March 2010.

[36] R. Charanya and M. Aramudhan, "Survey on access control issues in cloud computing," in *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, Pudukkottai, India, 2016.

[37] M. U. Aftab, Z. Qin, N. W. Hundera et al., "Permission-based separation of duty in dynamic role-based access control model," *Symmetry*, vol. 11, no. 5, p. 669, 2019.

[38] S. Bhatt, L. A. Tawalbeh, P. Chhetri, and P. Bhatt, "Authorizations in cloud-based Internet of Things: current trends and use cases," in *2019 4th International Conference on Fog and Mobile Edge Computing, FMEC 2019*, vol. 1, pp. 241–246, Rome, Italy, 2019.

[39] S. Shekhar and H. Xiong, "Geo-Role-Based Access Control," in *Encyclopedia of GIS*, pp. 368–368, Springer, 2008.

[40] K. Rajesh Rao, I. G. Ray, W. Asif, A. Nayak, and M. Rajarajan, "R-PEKS: RBAC enabled PEKS for secure access of cloud data," *IEEE Access*, vol. 7, pp. 133274–133289, 2019.

[41] P. J. Sun, "Privacy protection and data security in cloud computing: a survey, challenges, and solutions," *IEEE Access*, vol. 7, pp. 147420–147452, 2019.

[42] Z. Fu, X. Wu, C. Guan, X. Sun, and K. Ren, "Toward efficient multi-keyword fuzzy search over encrypted outsourced data with accuracy improvement," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 12, pp. 2706–2716, 2016.

[43] M. Rasori, "*fABElous: an attribute-based scheme for Industrial Internet of Things*," in *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, Washington, DC, USA, June 2019.

[44] D. Sathya and P. G. Kumar, "Secured remote health monitoring system," *Healthcare Technology Letters*, vol. 4, no. 6, pp. 1–5, 2017.

[45] Y. Miao, X. Liu, K. K. R. Choo et al., "Privacy-preserving attribute-based keyword search in shared multi-owner setting," *IEEE Transactions on Dependable and Secure Computing*, vol. 4, pp. 1–15, 2019.

[46] Y. Yang, X. Zheng, W. Guo, X. Liu, and V. Chang, "Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system," *Information Sciences*, vol. 479, pp. 567–592, 2019.

[47] Q. Zheng, S. Xu, and G. Ateniese, "VABKS: verifiable attribute-based keyword search over outsourced encrypted data," in *IEEE INFOCOM 2014-IEEE conference on computer communications*, pp. 522–530, Toronto, ON, Canada, May 2014.

[48] H. Yin, J. Zhang, Y. Xiong et al., "CP-ABSE: a ciphertext-policy attribute-based searchable encryption scheme," *IEEE Access*, vol. 7, pp. 5682–5694, 2019.

[49] Q. Li, Y. Yue, and Z. Wang, "Deep Robust Cramer Shoup Delay Optimized Fully Homomorphic for IIOT secured transmission in cloud computing," *Computer Communications*, vol. 161, pp. 10–18, 2020.

[50] L. Guo, Z. Li, W. C. Yau, and S. Y. Tan, "A decryptable attribute-based keyword search scheme on eHealth cloud in Internet of Things platforms," *IEEE Access*, vol. 8, pp. 26107–26118, 2020.

[51] Y. W. Hwang, I. Y. Lee, and K. Yim, "A study on access control scheme based on ABE using searchable encryption in cloud environment," in *Advances in Internet, Data and Web Technologies*, vol. 47 of Lecture Notes on Data Engineering and Communications Technologies, pp. 215–221, 2020.

[52] D. Ziegler, A. Marsalek, B. Prünster, and J. Sabongui, "Efficient access-control in the IIoT through attribute-based encryption with outsourced decryption," in *Proceedings of the 17th International Joint Conference on e-Business and Telecommunications: SECRYPT*, Portugal, 2020.

[53] N. Löken, "Searchable encryption with access control," in *Proceedings of the 12th International Conference on Availability, Reliability and Security*, Reggio Calabria, Italy, 2017.

[54] P. Chaudhari and M. L. Das, "Privacy preserving searchable encryption with fine-grained access control," *IEEE Transactions on Cloud Computing*, vol. 7161, no. c, pp. 1–1, 2019.

[55] K. Zhang, J. Long, X. Wang, H.-N. Dai, K. Liang, and M. Imran, "Lightweight searchable encryption protocol for Industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4248–4259, 2021.

[56] S. Qi, Y. Lu, W. Wei, and X. Chen, "Efficient data access control with fine-grained data protection in cloud-assisted IIoT," *IEEE Internet of Things Journal*, vol. 4662, no. c, pp. 1–1, 2020.

[57] Y. Miao, J. Ma, X. Liu, X. Li, Z. Liu, and H. Li, "Practical attribute-based multi-keyword search scheme in mobile crowdsourcing," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 3008–3018, 2018.

[58] Y. Miao, X. Liu, R. H. Deng et al., "Hybrid keyword-field search with efficient key management for industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3206–3217, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6668771, 8 pages
https://doi.org/10.1155/2021/6668771

Research Article

Generative Adversarial Network for Image Raindrop Removal of Transmission Line Based on Unmanned Aerial Vehicle Inspection

Changbao Xu,[1] Jipu Gao,[1] Qi Wen,[1] and Bo Wang[2]

[1]*Electric Power Research Institute of Guizhou Power Grid Co., Ltd., Guiyang 550000, China*
[2]*School of Electrical and Automation Engineering, Wuhan University, China*

Correspondence should be addressed to Bo Wang; whwdwb@whu.edu.cn

Received 9 December 2020; Revised 9 February 2021; Accepted 9 March 2021; Published 23 March 2021

Academic Editor: Mohammad R. Khosravi

In the process of UAV line inspection, there may be raindrops on the camera lens. Raindrops have a serious impact on the details of the image, reducing the identification of the target transmission equipment in the image, reducing the accuracy of the target detection algorithm, and hindering the practicability of UAV line inspection technology in cyber-physical energy systems. In this paper, the principle of raindrop image formation is studied, and a method of raindrop removal based on generation countermeasure network is proposed. In this method, the attention recurrent network is used to generate the raindrop attention map, and the context code decoder is used to generate the raindrop image. The experimental results show that the proposed method can remove the raindrops in the image and repair the background image of raindrop coverage area and can generate a higher quality raindrop removal image than the traditional method.

1. Introduction

UAV inspection image is the most important information carrier in Industrial Internet of Things (IIoT). The purpose of intelligent inspection can be achieved through the target detection and fault location of the machine inspection image. Sometimes there are raindrops on the camera in the process of UAV line patrol, which will cover the information of the target object in the background image and reduce the image quality. Raindrops make the transmission line equipment absorb a wider range of environmental light when imaging, and the superposition of these refracted light and the reflected light of the target object causes the image degradation. In addition, the camera should focus on the transmission line equipment when the UAV takes photos during the line patrol, and the presence of raindrops will affect the camera's focus, making the image background virtual, and the image detail information loss is serious, so the follow-up operation of the machine patrol image with raindrops will be extremely difficult. Therefore, the existence of raindrops will lead to the uneven quality of the machine patrol image, which will affect the extraction and utilization of image information and reduce the accuracy and reliability of target detection.

In the field of image processing, single image raindrop removal is an extremely complex technology. There are not many existing methods to carry out relevant technical research for a long time. These methods can be roughly divided into traditional raindrop removal methods and CNN-based raindrop removal methods. The traditional raindrop removal methods are divided into filtering and the learned dictionary plus sparse coding methods. Filtering includes guided filter [1], improved guided filter [2], multi-guided filter [3], LO smoothing filter [4], and nonlocal means filter [5]. The image of raindrop removal generated by filtering is fuzzy, and some raindrops cannot be removed. Fu et al. [6] use the filter to filter the image containing raindrops to get the high-frequency and the low-frequency image, use learned dictionary plus sparse coding to remove raindrops from the high-frequency image, and then combine the high-frequency image and the low-frequency image to get the raindrop image. On this basis, Kang et al. [7] introduced the raindrop HOG feature and used the K-means clustering method to cluster the high-frequency images to obtain the

rain dictionary and the rain-free dictionary and then sparse coding, respectively, to obtain the high-frequency rain-free image and the high-frequency raindrop-free image and the low-frequency image fusion to obtain the raindrop-free image. The image background obtained by this method is clearer than that obtained by Fu's method. Lou et al. [8] proposed a discriminative sparse coding method to remove image raindrops. This coding method has certain discrimination ability, which can reduce the error rate of raindrop discrimination and improve the effect of raindrop removal. In 2013, David et al. [9] first used convolutional neural network for image raindrop removal. Firstly, a sample database containing raindrop-free image pairs was constructed, and the corresponding image was segmented by a sliding window with step length of 1. Then, the network was trained by the mean square error between corresponding image blocks, and finally, the convolutional neural network model capable of raindrop removal was obtained. After that, Fu [10, 11] and others fused the convolution neural network and image decomposition, using the convolution neural network to extract the raindrop feature in the image, as the raindrop feature in the high-frequency component to achieve the raindrop removal in the high-frequency component, and eventually improve the quality of the raindrop removal effect image.

Through the research on the existing methods, we found that most of the traditional methods of raindrop removal are based on the model. The traditional model is used to describe raindrops, rain lines, and background images, respectively, and with the corresponding algorithm, using step by step iterative optimization to remove the raindrop. The traditional method is not ideal for the image processing with dense raindrops; the background image covered by raindrops cannot be repaired precisely. The method based on convolution neural network can fully extract the feature information of the image, and the effect of using this method to remove the raindrop is better.

However, with the increase of network depth, the network is prone to overfitting, and the effect of raindrop removal is difficult to be further improved. In view of the shortcomings of the above algorithm, this paper analyzes the principle of raindrop image generation and then discusses the basic structure of GAN. On this basis, the raindrop image generation model is integrated into the GAN, and a raindrop removal method based on the GAN is proposed. The raindrop image obtained by this method is closer to the real image.

2. Single Image Raindrop Removal Model

2.1. Image Generation Model with Raindrops. In the process of image raindrop removal, the raindrop image is usually modelled as a linear combination of background image and raindrop layer, and the mathematical expression is shown as equal

$$I(x) = (1 - M(x)) \odot B(x) + R(x). \tag{1}$$

I represents the raindrop image taken by the UAV during the line patrol, x is the pixel position in the image, and B is the background image, that is, the UAV takes clear transmission line equipment. R is the impact of raindrops on the image, and M is the binary mask, which is used to represent the impact of raindrops on the background image.

2.2. Generative Adversarial Networks. In recent years, with the continuous development of deep learning, scholars put forward the generative adversarial network (GAN), which has good performance in dealing with complex data distribution and is one of the most promising methods in the field of unsupervised learning. The model contains generating module and discriminating module. In the aspect of image restoration, by the game between the two modules, high-quality images can be output.

The core idea of GAN is game. The generation model is used to generate a realistic sample, and the discrimination model is used to judge the authenticity of the generated image. The discrimination network needs to be able to distinguish whether the input image is a real picture or a picture generated from the generated network. If it is a real picture, output 1; otherwise, output 0. The network generates new pictures according to the pattern of the real pictures. By playing games with the discrimination network, the quality of the generated pictures is as close to the real pictures as possible so that the discriminator cannot recognize the image from the generator. In order to achieve this function, the generation network and the GAN need to be trained alternately and iteratively.

Learning complex data distribution quickly is the strong point of GAN. Also, the network does not need complex constraint functions, and the whole learning process does not need human intervention. Another feature of GAN is that it can update the loss function of the network by itself depending on the distribution of sample data. In the process of training the generative adversarial network, the discriminative network can be used as the loss function of the generative network, which plays a role of supervision and guidance for the optimization of the generated network. The process of judging network parameter updating is also the process of optimizing the network loss function.

2.3. Raindrop Removal Model Based on Generative Adversarial Network. Same as the basic structure of GAN, the raindrop model based on generative adversarial network mainly includes generative network and discriminative network. Under the guidance of attention map, clear and real raindrop removal images are generated as far as possible. The overall architecture of raindrop removal network is shown in Figure 1. The improved generative network and discrimination network will be described in detail below.

The whole loss function of raindrop model based on GAN is shown in

$$\min_G \max_D \left\{ E_{R\sim P\text{clean}}[\log(D(R))] + E_{I\sim P\text{raindrop}}[\log(1 - D(G(I)))] \right\}, \tag{2}$$

where G stands for generating network and D stands for

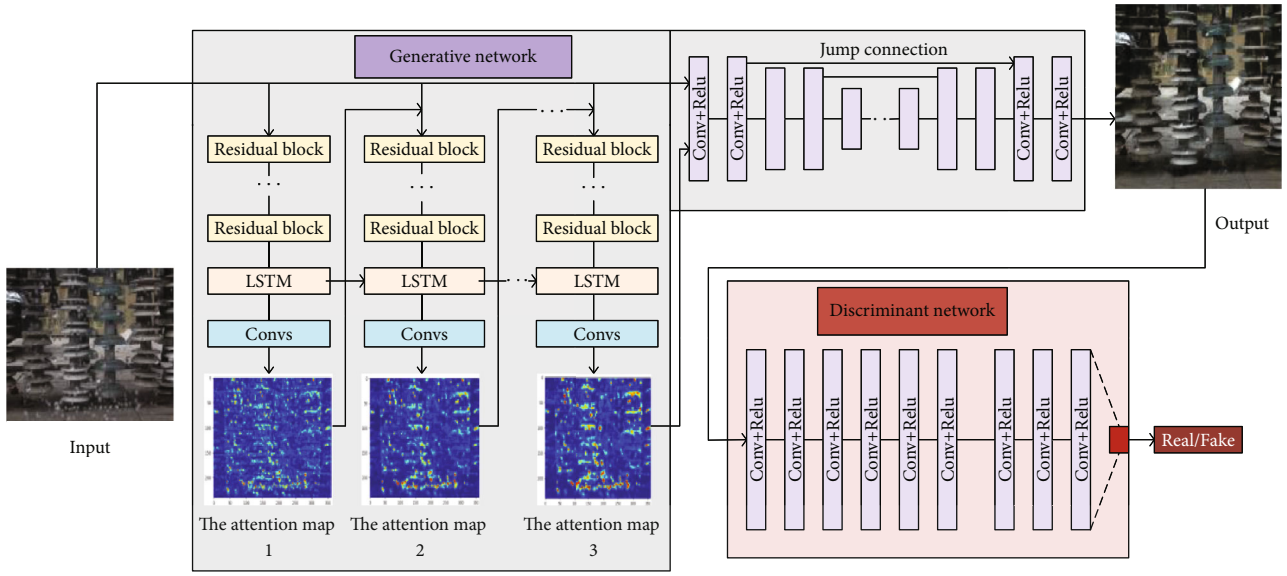


Figure 1: Diagram of the improved generative network consists of two subnetworks: attention recurrent network and context autoencoding decoding network.

discriminating network. I is the image with raindrop, $G(I)$ is the image after raindrop removal, and R is the real sample without raindrop.

2.3.1. Improved Generative Network. As shown in Figure 1, the improved generative network consists of two subnetworks: attention recurrent network and context autoencoding decoding network. LSTM network is included in the attention recurrent network [12], which generates attention map by cyclic iteration. Attention map contains the location and shape information of raindrops in raindrop image, which guides the context codec to focus on raindrops and their surrounding areas.

(1) Attention Recurrent Neural Network. The attention recurrent network is used to locate the target area in the visual attention model to improve the accuracy of target recognition [13–16]. Inspired by this, this paper applies this structure to the raindrop removal network and uses the visual attention guidance generative network and distinguish network to find the location of raindrops in the image. As shown in the generator part of Figure 1, the attention recurrent network consists of four circulation modules, each of which contains a packet residual network [17, 18], an LSTM unit, and a convolution layer, wherein the residual module is used to extract the raindrop feature information from the input image and the attention map generated by the previous recurrent module, and the LSTM unit [19, 20] and the convolution layer are used to generate a 2D attention map.

Binary mask plays a key role in the generation of attention map. There are only two numbers 0 and 1 in the mask. 0 means there is no raindrop in this pixel, and 1 means there is raindrop in this pixel. The mask image and the raindrop image are input into the first recurrent module of the attention cycle network for the generation of the initial attention map. The mask image is obtained by subtracting the clear image from the image with raindrops and then setting a certain threshold value to filter. Although the obtained mask image is relatively rough, it has a great effect on the generation of fine attention map. The biggest distinction between attention graph and mask graph is that the mask graph only contains 0 and 1, and the value of attention graph is [0, 1]. The larger the median value of the attention graph indicates that the more attention should be paid to the pixel, that is, the more likely there are raindrops at the pixel. Even in the same raindrop area, the value of attention map will be different, which is related to the shape and thickness of raindrops and also reflects the influence of raindrops on different pixels of background image.

The attention recurrent network contains a LSTM (Long Short-Term Memory). The LSTM unit includes an input gate i_t, a forgetting gate f_t, an output gate o_t, and a unit status C_t. The interaction between state and gate in time dimension is defined in

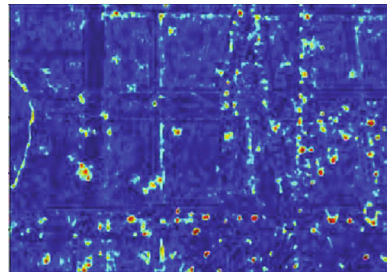
$$\begin{aligned} i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \odot C_{t-1} + b_i), \\ f_t &= \sigma\left(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \odot C_{t-1} + b_f\right), \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tanh\left(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c\right), \\ H_t &= o_t \odot \tanh\left(C_t\right), \end{aligned} \tag{3}$$

where X_t is the image feature generated by the residual network, C_t represents the state feature to be transferred to the next LSTM unit, H_t is the output feature of LSTM unit, $\odot$ is matrix multiplication, and $*$ is convolution operation.

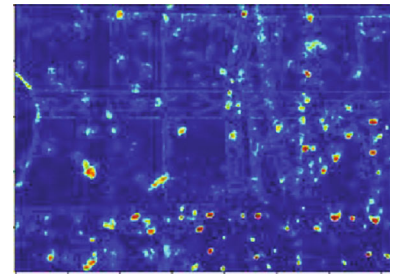
The input of the generated network is an image pair with the same background scene, one with raindrops and one without raindrops. The loss function of each recurrent

(a) The image contains raindrops

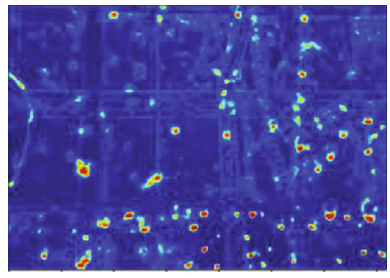
(b) Raindrop removal image



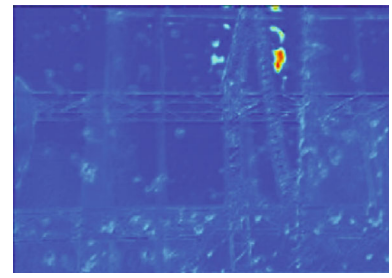
(c) The attention map of raindrops



(d) The attention map of raindrops 2



(e) The attention map of raindrops 3



(f) The attention map of raindrops 4

Figure 2: Raindrop removal image and attention map.

module is defined as the mean square error (MSE) between the output attention map and the binary mask M. For recurrent cycle network, the front-module loss function is given a smaller weight, and the back-module loss function is given a larger weight. The loss function is shown in

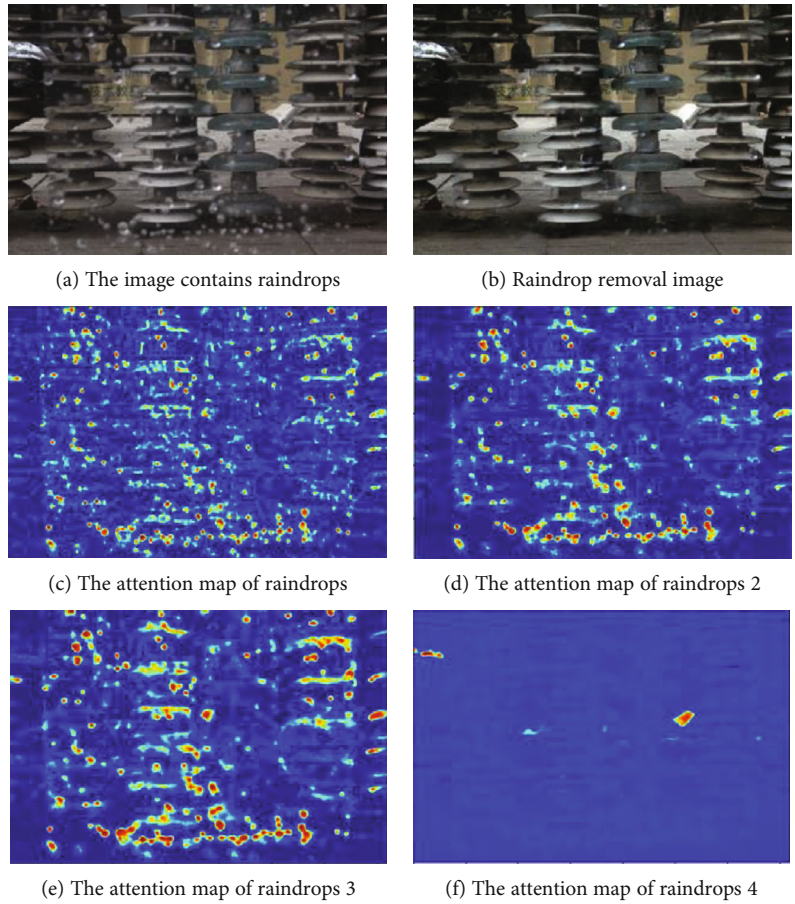
$$L_{\mathrm{ATT}}(\{A\}, M) = \sum_{t=1}^{N} \theta^{N-t} L_{\mathrm{MSE}}(A_t, M), \tag{4}$$

where A_t is the attention graph generated by the cyclic network in time step t. $A_t = \mathrm{ATT}_t(F_{t-1}, H_{t-1}, C_{t-1})$, F_{t-1} represents the fusion of the image with raindrops and the output attention map of the previous recurrent unit. In the whole recurrent network, the larger N is, the finer attention map is generated. But the larger N is, the more memory is needed to store the intermediate parameters. It is found that the network efficiency is the highest when $N = 4, \theta = 0.8$.

(2) Context Automatic Encoder-Decoder. The input of the context auto codec is the attention map generated by the raindrop image and the attention recurrent network. The raindrop removal and background restoration are achieved under the guidance of the attention map. There are 16 conv-relu modules in the context autoencoder-decoder. The structure of coding and decoding is symmetrical. Skip connection is added between corresponding modules to prevent the image from being blurred. There are two loss functions used in the context autoencoder-decoder, multiscale loss and perceptual loss. Multiscale loss function extracts image feature information from different layers of decoder and makes full use of multilevel image information to optimize the model to obtain clear image of raindrop removal. The multiscale loss function is defined as

$$L_M(\{S\}, \{A\}) = \sum_{i=1}^{M} \lambda_i L_{\mathrm{MSE}}\left(S_i, A_{N_i}\right), \tag{5}$$

where S_i represents the image features extracted from the i-th layer of the encoder, A_{N_i} represents the real image which has the same scale with S_i, and $\{\lambda_i\}_{i=1}^{M}$ is the weight of different scales. The design of loss function pays more attention to feature extraction on large-scale image, and the smaller size image contains less information which has little influence on model optimization. The output image sizes of the last layer, the last third layer, and the last fifth layer of the decoder are 1/4, 1/2, and 1 of the original sizes, respectively, and the corresponding weights λ are set to 0.6, 0.8, and 1.0.



(a) The image contains raindrops (b) Raindrop removal image

(c) The attention map of raindrops (d) The attention map of raindrops 2

(e) The attention map of raindrops 3 (f) The attention map of raindrops 4

Figure 3: Raindrop removal image and attention map.

Table 1: PSNR and SSIM of deraindrop image.

Method	PSNR	SSIM
Yang raindrop removal method	19.1538	0.7128
Fu raindrop removal method	19.8693	0.8176
Raindrop removal based on GAN	31.5710	0.9023

In addition to the pixel-based scale loss, this paper also increases the perceptual loss [21] to obtain the global difference between the output of the automatic context encoder-decoder and the corresponding clear picture. Perceptual loss measures the difference between the raindrop removed image and the real image from the global perspective, which will make the raindrop image closer to the real sample. The image global information can be extracted by vgg16, and the network pretraining needs to be completed on the ImageNet data set in advance. The perceptual loss function is defined as

$$L_P(O, T) = L_{\mathrm{MSE}}(\mathrm{VGG}(O), \mathrm{VGG}(T)). \tag{6}$$

VGG is a pretrained CNN, which can complete the feature extraction of a given input image. O is the output image of the automatic encoder, $O = G(I)$, and T is a real image sample without raindrops. To sum up, the loss function of the generated network is defined as

$$L_G = 10^{-2} L_{\mathrm{GAN}}(O) + L_{\mathrm{ATT}}(\{A\}, M) + L_M(\{S\}, \{A\}) + L_P(O, T), \tag{7}$$

where $L_{\mathrm{GAN}}(O) = \log(1 - D(O))$.

2.3.2. Improved Discrimination Network. The function of discriminating network is to distinguish true and false samples. The discriminator in GAN usually uses global discriminator [22–24]. Determine the difference between the image output by the generator and the real sample. Only using global information to judge whether the image is true or false is not conducive to the restoration of local image information by generating network. For image raindrop removal, this method hopes to restore the details of the image as much as possible, so as to carry out the subsequent target detection work. The existing discrimination network cannot be used directly. Therefore, this paper combines the global discriminator and the local discriminator to determine the true and false output samples of the generated network together.

The use of the local discriminator is based on knowing the location information of raindrops in the image. The attention map is generated in the attention cycle network of the image restoration stage, which solves the problem of

Table 2: Target detection results.

Method	Tower failure (AP)	Small size fittings (AP)	Ground conductor failure (AP)	Insulator failure (AP)	mAP
Yang raindrop removal method	0.5164	0.4436	0.5019	0.5482	0.5025
Fu raindrop removal method	0.5548	0.4931	0.5326	0.5931	0.5343
Raindrop removal based on GAN	0.7684	0.5689	0.6143	0.6849	0.6591

location of raindrops in the image. Therefore, attention map can be introduced into the discriminator network to guide the local discriminator to automatically find the raindrop area in the image. CNN is used to extract features from the inner layer of the discriminator. At the same time, it also extracts features from the raindrop image generated by the generator. Then, the loss function of the local discriminator is formed by combining the obtained feature image and attention image. The existence of attention map will guide the discrimination network to pay more attention to the raindrop area in the image. In the last layer of the discrimination network, the full connection layer is used to judge the authenticity of the input image. The overall structure of the discrimination network is shown in the lower right part of Figure 1. The whole loss function of the discrimination network can be expressed as

$$L_D(O, R, A_N) = -\log(D(R)) - \log(1 - D(O)) + \gamma L_{map}(O, R, A_N), \tag{8}$$

where γ is 0.05, the first two terms of the formula are the loss function of the global discriminator, L_{map} represents loss function of local discriminator, and the loss function of local discriminator is shown in

$$L_{map}(O, R, A_N) = L_{\text{MSE}}\left(D_{\text{map}}(O), A_N\right) + L_{\text{MSE}}\left(D_{\text{map}}(R), 0\right). \tag{9}$$

D_{map} represents the two-dimensional attention mask function generated by the discrimination network, and R represents the sample image extracted from the real and clear image database. 0 represents the attention map with only 0 value, which represents there is no raindrop in the real image, so attention map is not required to guide the network to extract features.

The discriminant network in this paper consists of seven convolution layers, the core of which is (3, 3), the full connection layer is 1024, and the single neuron uses the Sigmoid activation function.

3. Model Training

3.1. Data Set Formation. For the training of raindrop removal network proposed in this paper, a set of transmission line equipment image pairs is needed. Each pair of images contains exactly the same background scene, one of which contains raindrops and the other has no raindrops.

Error reporting in order to make the method proposed in this paper suitable for the image raindrop removal in the scene of UAV line patrol, this paper simulates the real scene of transmission line as much as possible when making the data set. UAV carries two cameras with two identical glasses when making the data set, one to spray water and the other to keep clean. Spray water on the glass plate to simulate raindrops on the camera in rainy days. The thickness of the glass plate is 3 mm. Set the distance between the glass and the camera to 2 to 5 cm to produce different raindrop images and minimize the reflection effect of the glass. During the shooting process, keep the relative position of the camera and the glass lens unchanged, and ensure that the background images taken by the two cameras are the same. Also, ensure that the atmospheric conditions (such as sunlight and cloud) and the background objects should be static during the image acquisition process. Finally, 2000 pairs of images including transmission line equipment scenes were taken.

3.2. Raindrop Removal Online Training Details. The 2000 pairs of pictures in the data set are allocated according to 8 : 2, among which 1600 pairs are used as model training sets and 400 pairs are used as model test sets. The super parameters of the model are set, in which the initial learning rate is set to 0.001, the batch size is set to 16, and the number of iterations is set to 40000. Using Adam optimization algorithm, it is found that the rate of gradient descent is relatively low in the process of training. Therefore, it is changed to momentum optimization algorithm, and it is found that the convergence speed of the model is significantly faster. After 40000 times of iterative training, the model is verified by test set, and it is found that the raindrop model based on the network of resistance generation has good portability.

4. Experiment Results

4.1. Comparison of Effect Pictures of Raindrop Removal. Randomly select a picture from the image data set containing raindrops for raindrop removal, and the results are shown in Figures 2 and 3.

The background image in Figures 2 and 3 is the tower and insulator string; Figures 2(a)–2(f) and Figures 3(a)–3(f) are the original image, the raindrop removal image, and the attention map generated by four recurrent networks, respectively. The original image contains dense raindrops. The raindrop removal method proposed in this paper can remove most of the raindrops in the image and repair the background image of the raindrop covered part. It can be seen from the attention map that the location and size of

raindrops in the original image can be clearly determined. From the comparison of Figures 2(a) and 2(b) and Figures 3(a) and 3(b), it can be seen that the contrast, brightness, and target edge information of the raindrop removal image and the original image are basically the same.

4.2. Comparison of Raindrop Removal Image Indexes. Randomly select a picture from the data set containing raindrops, use Yang raindrop removal method [25–27] and the method proposed in this paper to remove raindrops, and calculate the PSNR value and SSIM value of the two methods to obtain the image; the results are shown in Table 1.

It can be seen from Table 1 that the PSNR and SSIM of the image obtained by the method proposed in this paper are higher than those of Yang and Fu, which indicates that the similarity between the raindrop image obtained by the method proposed in this paper and the original clear background image is higher, which proves that the effect of the raindrop method based on the generated antinetwork is better than that of Yang and Fu.

4.3. Target Detection Result Comparison. Randomly select 50 inspection images of transmission line with raindrops including tower fault, small size hardware fault, ground wire fault, and insulator fault from the test set. Yang's raindrop removal method and the raindrop removal method proposed in this paper are, respectively, used for image raindrop removal. The Faster Rcnn target detection algorithm is used to detect the device defect target of raindrop image, Yang raindrop image, and raindrop image of the method proposed in this paper. Then, calculate the AP value of four kinds of faults and the mAP value of each group of images, respectively. The results are shown in Table 2.

From the AP value and the mAP value in Table 2, it can be seen that the target detection accuracy of the image after raindrop removal is higher than that without image enhancement. At the same time, the proposed method is better than the previous methods in the aspects of raindrop removal and image restoration.

5. Conclusion

The discrimination network uses a combination of global and local discriminators to distinguish the generated raindrop images. Using the test set in this paper to test the model, the experiment shows that the method proposed in this paper can completely remove the raindrop in the image and repair the background image, and the raindrop image is closer to the real image. Using the method in this paper to process the image raindrop can restore the image details and improve the accuracy of the target detection algorithm.

Data Availability

The data used to support the findings of this study are included within the article. The project was supported by Science Support Project of Guizhou Province ([2020]2Y039).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Science Support Project of Guizhou Province ([2020]2Y039).

References

[1] J. Xu, W. Zhao, P. Liu, and X. Tang, "Removing rain and snow in a single image using guided filter," in *IEEE International Conference on Computer Science & Automation Engineering*, pp. 304–307, Zhangjiajie, China, 2012.

[2] J. Xu, W. Zhao, P. Liu, and X. Tang, "An improved guidance image based method to remove rain and snow in a single image," *Computer and Information Science*, vol. 5, no. 3, pp. 1–11, 2012.

[3] X. Zheng, Y. Liao, W. Guo, X. Fu, and X. Ding, "Single-image-based rain and snow removal using multi-guided filter," in *International Conference on Neural Information Processing*, pp. 258–265, Berlin, Heidelberg, 2013.

[4] X. Ding, L. Chen, X. Zheng, Y. Huang, and D. Zeng, "Single image rain and snow removal via guided L0 smoothing filter," *Multimedia Tools & Applications*, vol. 75, no. 5, pp. 2697–2712, 2016.

[5] J. H. Kim, C. Lee, J. Y. Sim, and C. S. Kim, "Single-image deraining using an adaptive nonlocal means filter," in *IEEE International Conference on Image Processing*, pp. 914–917, Melbourne, VIC, Australia, 2014.

[6] Y. H. Fu, L. W. Kang, C. W. Lin, and C. T. Hsu, "Single-frame-based rain removal via image decomposition," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 782no. 1, pp. 1453–1456, Prague, Czech Republic, 2011.

[7] L. W. Kang, C. W. Lin, and Y. H. Fu, "Automatic single-image-based rain streaks removal via image decomposition," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1742–1755, 2012.

[8] Y. Luo, Y. Xu, and H. Ji, "Removing rain from a single image via discriminative sparse coding," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3397–3405, Santiago, Chile, 2015.

[9] D. Eigen, D. Krishnan, and R. Fergus, "Restoring an image taken through a window covered with dirt or rain," in *IEEE International Conference on Computer Vision*, pp. 633–640, Columbus, Ohio USA, 2014.

[10] Z. Gao, Y. Li, and S. Wan, "Exploring deep learning for view-based 3D model retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 1, pp. 1–21, 2020.

[11] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: a deep network architecture for single-image rain removal," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2944–2956, 2017.

[12] S. Ding, S. Qu, Y. Xi, and S. Wan, "A long video caption generation algorithm for big video data retrieval," *Future Generation Computer Systems*, vol. 93, pp. 583–595, 2019.

[13] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network,"

in *IEEE Conference on Computer Vision & Pattern Recognition*, pp. 1715–1723, Honolulu, HI, USA, 2017.

[14] Y. Zhao, H. Li, S. Wan et al., "Knowledge-aided convolutional neural network for small organ segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1363–1373, 2019.

[15] S. Ding, S. Qu, Y. Xi, and S. Wan, "Stimulus-driven and concept-driven analysis for image caption generation," *Neurocomputing*, vol. 398, pp. 520–530, 2020.

[16] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017.

[17] Y. Xi, Y. Zhang, S. Ding, and S. Wan, "Visual question answering model based on visual relationship detection," *Signal Processing: Image Communication*, vol. 80, p. 115648, 2020.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[19] L. Wang, H. Zhen, X. Fang, S. Wan, W. Ding, and Y. Guo, "A unified two-parallel-branch deep neural network for joint gland contour and segmentation learning," *Future Generation Computer Systems*, vol. 100, pp. 316–324, 2019.

[20] Y. Tanaka, A. Yamashita, T. Kaneko, and K. T. Miura, "Removal of adherent waterdrops from images acquired with a stereo camera system," *IEICE Transactions on Information and Systems*, vol. 89, no. 7, pp. 2021–2027, 2006.

[21] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, pp. 694–711, Cham, 2016.

[22] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–14, 2017.

[23] Z. Gao, H. Xue, and S. Wan, "Multiple discrimination and pairwise CNN for view-based 3D object retrieval," *Neural Networks*, vol. 125, pp. 290–302, 2020.

[24] Y. Li, S. Liu, J. Yang, and M. H. Yang, "Generative face completion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3911–3919, Honolulu, HI, USA, 2017.

[25] M. Khosravi and S. Samadi, "Reliable data aggregation in internet of ViSAR vehicles using chained dual-phase adaptive interpolation and data embedding," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2603–2610, 2020.

[26] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1685–1694, Honolulu, HI, 2017.

[27] L. Li, T. T. Goh, and D. Jin, "How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis," in *Neural Computing and Applications*, vol. 32, pp. 4387–4415, Springer, London, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5579857, 14 pages
https://doi.org/10.1155/2021/5579857

Research Article

NAAM-MOEA/D-Based Multitarget Firepower Resource Allocation Optimization in Edge Computing

Liyuan Deng,[1] Ping Yang,[1] Weidong Liu,[1] Lina Wang,[2] Sifeng Wang,[2] and Xiumei Zhang[3]

[1]*Xi'an Research Institute of High-Technology, China*
[2]*School of Computer Science, Qufu Normal University, China*
[3]*School of Computer Science and Software Engineering, University of Science and Technology Liaoning, China*

Correspondence should be addressed to Xiumei Zhang; aszxm2002@126.com

Received 11 January 2021; Revised 9 February 2021; Accepted 27 February 2021; Published 22 March 2021

Academic Editor: Mohammad R. Khosravi

In the edge environment, the multiobjective evolutionary algorithm based on decomposition (MOEA/D) has been widely used in the research of multitarget firepower resource allocation. However, as the MOEA/D algorithm uses a fixed neighborhood update mechanism, it is impossible to rationally allocate computing resources based on the difficulty of each subproblem optimization, which results in some problems such as reduced population evolution efficiency and poor evolution quality during the calculation process. In order to solve these problems, a decision mechanism for subproblems and population evolution stages is designed, and on this basis, a MOEA/D algorithm based on the neighborhood adaptive adjustment mechanism is proposed to adapt to the edge environment. The optimization model of multiobjective firepower resource allocation based on the maximization of damage effect and the minimization of strike cost is constructed and solved. Using the ZDT series of test functions for comparative experiments, the simulation results show that the proposed algorithm can balance the distribution and convergence of population evolution and obtain satisfactory optimization results.

1. Introduction

In the edge environment, due to the limited computing resources of edge clients, the allocation of firepower resources based on factors such as battlefield situation, weapon performance, and combat objectives reasonably deploying and allocating various types and quantities of weapons and equipment to obtain the best combat effect is an important part of combat planning [1]. The firepower resource allocation optimization problem in edge environment usually constructs a single-objective firepower resource allocation optimization model based on the damage probability objective function, using heuristic genetic algorithm [2], simulated annealing genetic algorithm [3], particle swarm algorithm [4], and ant colony algorithm to solve the model. In practical problems, the objective function that only considers the single factor of damage probability is obviously not realistic. Literature [5] establishes interception benefit maximization and loss minimization models and used multiobjective quantum behavior particle swarm algorithm with a single/dual potential trap to solve the model. Literature [6] uses a genetic algorithm based on reference point nondominated sorting to solve the optimization problem of multispace-based ground strike weapon multitarget firepower resource allocation. Literature [7] uses the multitarget discrete particle swarm-gravity search algorithm (MODPSO-GSA) to achieve the solution of the multitarget allocation model of coordinated air combat weapons. The decomposition-based multiobjective evolutionary algorithm decomposes the high-dimensional and

complex multiobjective optimization problem into multiple single-objective subproblems by referring to the decomposition strategy in mathematical programming and optimizes the subproblems separately. It has the advantages of high algorithm efficiency and simple operation [8, 9].

The MOEA/D algorithm has been used in the study of multitarget firepower resource allocation in edge environment. Literature [10] comprehensively considers the influence of factors such as weapon type, target number, and damage probability and uses the MOEA/D algorithm as the framework to construct the WMOM/D algorithm for solving the multitarget fire distribution model. Simulation experiments prove that the WMOM/D algorithm has the advantage of solving the problem of small-scale fire distribution. Literature [11] applies the MOEA/D algorithm to the multiobjective fire optimization problem of aircraft carrier formation antisubmarine warfare and proposes the GD-MOEA/D algorithm combining differential evolution and Gaussian mutation operation, which greatly improves the speed of solving the problem and the quality of the solution. Literature [12] integrates the MOEA/D algorithm with the multilevel coevolutionary algorithm and uses the multilevel cooperative MOEA/D algorithm to solve the multiobjective optimization model of the joint fire strike target assignment problem. The simulation experiment proves that the algorithm has good convergence and uniformity.

However, because the MOEA/D algorithm uses a fixed neighborhood update mechanism, the ability to reasonably allocate computing resources is low especially in the edge environment with limited computing resources. So the problems such as reduced population evolution efficiency and poor evolution quality will occur in the calculation process.

To this end, this paper considers the impact of subproblems and the degree of population evolution on the performance of the algorithm, designs the decision mechanism for subproblems and population evolution stages, and proposes a MOEA/D algorithm based on the neighborhood adaptive adjustment mechanism. Compared with the traditional MOEA/D algorithm, the NAAM-MOEA/D algorithm can better balance the convergence and distribution and improve the quality of the solution.

In the simulation experiment, the NAAM-MOEA/D algorithm was compared with the MOEA/D algorithm, the MOEA/D-DE algorithm, and the NSGA-III algorithm. The algorithm running time was reduced by 82.1%, 108.1%, and 153.6%, respectively; the GD value was reduced by 84%, 59%, and 35%, respectively; and the IGD value of the algorithm was reduced by 75%, 56%, and 40%, respectively.

The main innovations of this article are summarized as follows:

(1) Aiming at the defects of the traditional MOEA/D algorithm's fixed neighborhood update mechanism in solving the multiobjective fire resource allocation problem, a MOEA/D algorithm based on the neighborhood adaptive adjustment mechanism is proposed, which greatly improves the efficiency and quality in edge

(2) A method for judging the evolution stage of the population based on the attribution of the weight vector and the degree of evolution of the subproblems is proposed, which provides a reliable basis for judging the evolution state of the population

(3) Based on the population evolution stage judgment method, a neighborhood adaptive adjustment mechanism is constructed and used in the MOEA/D algorithm to improve the convergence and distribution of the algorithm

The organizational structure of the paper is as follows:

Firstly, the related work is discussed in Section 2. Then, the optimization model of firepower resource allocation in edge environment is established in Section 3.1, the construction and decomposition of subproblems are discussed in Section 3.2.1, the shortcomings of traditional MOEA/D algorithm are analyzed in Section 3.2.2, and a mechanism for judging population evolution state is proposed in Section 3.2.3. The neighborhood adaptive adjustment mechanism is proposed in Section 3.2.4 and the steps of the NAAM-MOEA/D algorithm are summarized in Section 3.2.5. Finally, the simulation experiment is carried out in Section 4, and the performance of the algorithm is tested.

2. Related Work

In order to improve the performance of traditional MOEA/D algorithms in edge environment, researchers have proposed a variety of improved algorithms. The MOEA/D-DE algorithm proposed in literature [13] uses a difference operator instead of an evolution operator to enrich the diversity of the population, but the difference operator used by the algorithm is only applicable to a population of a specific size. The MOEA/D-DRA algorithm proposed in literature [14] allocates corresponding computing resources according to the complexity of specific problems and improves the performance of the algorithm by dynamically adjusting resource allocation; however, the proposed resource allocation criteria also have certain limitations. The MOEA/D-GL algorithm proposed in literature [15] embeds the grouping and statistical learning mechanism in the traditional MOEA/D algorithm, which prevents the population from falling into local optimization and improves the diversity of the population, but the overall performance improvement of the algorithm is not significant. The CD-MOEA/D-DE algorithm proposed in literature [16] controls the operation process of the algorithm by formulating control parameters ∂ and balances the performance of a multiobjective optimization problem solving and adaptive ability; however, the algorithm has a certain randomness in the value of the control parameter ∂ and does not have universal applicability.

In addition, the researchers have proposed many specific improvement measures for the shortcomings of the fixed neighborhood update mechanism of the MOEA/D algorithm in solving multiobjective optimization problems especially in the edge conditions with limited computing resources; however, the article does not elaborate on the

mechanism of how the neighborhood size affects the performance of the MOEA/D algorithm. Literature [17] points out that the size of the neighborhood will have an important impact on the performance of the MOEA/D algorithm, which provides important research directions for subsequent researchers. Literature [18] believes that different multiobjective optimization problems require different neighborhood sizes, and that the same multiobjective optimization problem also requires different neighborhood sizes at different stages of the algorithm, and proposes the ENS-MOEA/D algorithm with neighborhood adaptive adjustment capability; however, the ENS-MOEA/D algorithm may fall into local optimization in the later stage of operation. The ADEMO/D-ENS algorithm proposed in literature [19] combines the adaptive differential evolution algorithm with the variable neighborhood decomposition method to achieve the optimization of the algorithm. The MOEA/D-AGR algorithm proposed in literature [20] introduces an adaptive global replacement strategy in the neighborhood update method, which makes up for the shortcomings of the traditional MOEA/D algorithm in terms of global search capabilities. The MOEA/D-NMO algorithm proposed in literature [21] combines mutation strategies with different characteristics and neighborhoods of different sizes to select the best evolutionary combination to ensure the convergence of the algorithm while maintaining the diversity of the algorithm. The algorithms proposed in literature [19], literature [20], and literature [21] have all made improvements to the fixed field, but they all have certain limitations in application.

Although the current improved methods for fixed neighborhoods have improved the performance of traditional MOEA/D algorithms, the neighborhood adaptive strategies used by these algorithms do not consider the impact of population evolution on neighborhoods. Literature [22] proposes a neighborhood adaptive adjustment mechanism based on population evolution stage and individual fitness value, so that every individual has a corresponding neighborhood value at different evolution stages, but its neighborhood adjustment method does not consider the evolution status of the subproblems. Although the MOEA/D-ANS algorithm proposed in literature [23] adopts the ANS mechanism that adaptively adjusts the size of the neighborhood according to the evolution state of the population and subproblems, it can balance the convergence and distribution of population evolution, but it does not give a clear method on the statistical evolution of the number of better subquestions.

3. Method

3.1. Optimization of Fire Resource Allocation Model. The multitarget firepower resource allocation optimization problem in the edge environment can be described as follows: on the basis of satisfying the maximum damage effect and the minimum combat cost, determine the number of various weapons and equipment used to strike specific targets to obtain a feasible combat plan.

Set the target set of the enemy's combat system as $D = \{D_1, D_2, \cdots, D_M\}$; D_i represents the i-th target. There are a total of N types of weapons available for use.

$B = \{B_1, B_2, \cdots, B_N\}$, and B_j represents the j-th types of weapons. If there is a total of M_j class to choose from the j-th type of weapons, then $B_j = \{B_j^1, B_j^2, B_j^3, \cdots, B_j^{M_j}\}$. Select the j-th weapon in the weapon set B to strike the i-th target in the target set D; the probability of the target being destroyed is p_{ij} and the cost of each use of the j-th weapon is C_j. Suppose that the damage ability of the j-th type of weapons to target D_i is

$$P_i = 1 - \prod_{n=1}^{M_j} \left(1 - m_{ij}^n p_{ij}\right). \tag{1}$$

Among them, only when the j-th type of weapons of class n weapon is used to strike target D_i, there is $m_{ij}^n = 1$; otherwise, $m_{ij}^n = 0$. The purpose of firepower resource allocation is to maximize the damage effect under limited conditions. It is necessary to consider the priority of attacking the targets with high importance. Therefore, the calculation model of damage capability can be defined as

$$\max f(x) = \sum_{i=1}^{M} \sum_{j=1}^{N} \omega_i \left[1 - \prod_{n=1}^{M_j} \left(1 - m_{ij}^n p_{ij}\right)\right]. \tag{2}$$

Among them, ω_i is the importance of the i-th target.

In addition, the minimum operational cost calculation model is defined as follows:

$$\min C(x) = \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{n=1}^{M_j} m_{ij}^n C_j. \tag{3}$$

The constraints of the model are as follows:

(1) *Damage lower bound constraint*: if the target is to be destroyed to a certain extent so that it will lose certain combat capability, it is necessary to reach its damage lower bound. If the damage lower bound of target i is defined as β_i, then

$$P_i = 1 - \prod_{n=1}^{M_j} \left(1 - m_{ij}^n p_{ij}\right) \geq \beta_j. \tag{4}$$

(2) *Constraints on the number and types of weapons used*: it is stipulated that one weapon can only attack one target at most:

$$\sum_{i=1}^{M} m_{ij}^n \leq 1, \quad j = 1, 2, 3, \cdots, N, n = 1, 2, 3, \cdots, M_j. \tag{5}$$

It is stipulated that one type of weapon can only attack one type of target:

$$\sum_{j=1}^{N} m_{ij}^{n} \le 1, \quad i = 1, 2, 3, \cdots, M, n = 1, 2, 3, \cdots, M_j. \tag{6}$$

In summary, the multiobjective firepower resource allocation optimization model can be defined as

$$\begin{cases} \max f(x) = \sum_{i=1}^{M}\sum_{j=1}^{N} \omega_i \left[1 - \prod_{j=1}^{M_j}\left(1 - m_{ij}^{n} p_{ij}\right)\right], \\ \min C(x) = \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{n=1}^{M_j} m_{ij}^{n} C_j, \\ \text{s.t.} \\ \sum_{i=1}^{M} m_{ij}^{n} \le 1, \quad j = 1, 2, 3, \cdots, N, n = 1, 2, 3, \cdots, M_j, \\ \sum_{j=1}^{N} m_{ij}^{n} \le 1, \quad i = 1, 2, 3, \cdots, M, n = 1, 2, 3, \cdots, M_j, \\ P_i = 1 - \prod_{n=1}^{M_j}\left(1 - m_{ij}^{n} p_{ij}\right) \ge \beta_j. \end{cases} \tag{7}$$

3.2. Detailed Introduction of NAAM-MOEA/D Algorithm

3.2.1. Construction and Decomposition of Subproblem. The core of constructing the subproblem of the MOEA/D algorithm is to construct the weight vector of an objective function subproblem. Suppose the weight vector of the subproblem of the objective function is

$$\varphi^{r} = \left(\frac{r-1}{N-1}, \frac{N-r}{N-1}\right). \tag{8}$$

In the formula, N is the number of subproblems after decomposition, $r = 1, 2, \cdots, N$.

The core of the MOEA/D algorithm is the decomposition operation, usually using aggregate functions to decompose the multiobjective constraint problem into single-objective subproblems. Commonly used decomposition methods are weighted sum method, Chebyshev method, and boundary crossing method based on penalty. This paper adopts the Chebyshev method, and its decomposition principle is

$$g^{te}(x|\varphi^{r}, Z^{*}) = \max_{1 \le i \le m}\{\varphi_i^{r} | f_i(x) - Z_i^{*}\}. \tag{9}$$

Among them, $\varphi^{*} = \{\varphi_1^{*}, \varphi_2^{*}, \cdots, \varphi_m^{*}\}$ is the weight vector corresponding to the subproblem r. $Z^{*} = \{Z_1^{*}, Z_2^{*}, \cdots, Z_m^{*}\}$ is the ideal point. $f_i(x)$ is the i-th objective function, and φ_i^{r} is the i-th component of the weight vector φ^{r}. Z_i^{*} is the i-th component of the ideal point Z^{*}.

The single-objective optimization function of the i-th subproblem of objective function constructed by the Chebyshev method can be expressed as follows:

$$g^{te}(\mu|\gamma^{r}, Z) = \max_{1 \le i \le m}\left\{\gamma_i^{r}\left|\frac{f_i(x) - Z_i}{c_i}\right|\right\}. \tag{10}$$

In the formula, $f_i(x)$ is the i-th objective function, Z is the reference vector, Z_i is the i-th component of the reference vector Z, γ^{r} is the weight vector, and γ_i^{r} is the i-th component of the weight vector γ^{r}.

3.2.2. Defects of Traditional MOEA/D Algorithm. The MOEA/D algorithm maintains the power of population evolution from the update strategy of the neighborhood. The parent gene of an individual comes from the neighborhood, and it adopts a coevolution model based on neighborhood update. The evolution of an individual is carried out on the basis of the neighborhood. While evolving by itself, it drives the evolution of other neighborhoods by optimizing other individuals in the neighborhood. The MOEA/D algorithm uses a fixed neighborhood strategy. For different subproblems, the MOEA/D algorithm divides it into a neighborhood of the same size. In fact, the computational complexity of each subproblem in the objective function is different. The subproblems have different requirements for the size of the neighborhood at different stages. The size of the neighborhood has a very important impact on the evolution of the subproblems. When the size of the neighborhood is large, the probability of other individuals in the neighborhood being replaced by offspring individuals increases, and the population convergence speeds up, but the distribution of the population will become worse as the neighborhood size increases, making it easy for the algorithm to fall into local find the best. When the size of the neighborhood is small, the probability of other individuals in the neighborhood being replaced by offspring individuals decreases, the population convergence speed slows, the algorithm convergence decreases, and the overall evolution speed of the population decreases accordingly.

3.2.3. Judging Mechanism of Population Evolution State. From the previous analysis, we can see that in the MOEA/D algorithm, subproblems and populations have different requirements for neighborhood size at different evolution stages. Then, how to judge the evolution state of the population and whether it can find a mechanism that can effectively evaluate the evolution stage of the population is the core problem that the new algorithm needs to solve.

Some scholars propose to use the individual density of subproblems to assess the degree of population evolution. The individual density of the subproblem is equivalent to the number of individuals in the subinterval. If the individual density of the subproblem is smaller, the surrounding individuals are denser, the better the degree of evolution of the individual is, and the greater the probability of the problem being solved. If the individual density of the subproblem is smaller, the surrounding individuals are sparser, then the degree of evolution of the individual is smaller, and the problem is less likely to be solved.

Input: the threshold d_m, ω_m, ε_m; JM^i_{wei} is the attribution judging mechanism of weight vector; JM^i_{sub} is the subproblem evolution degree judgment mechanism; JM^i_{pop} is the population evolution degree judgment mechanism; d_{wi} is the distance between the weight vector and the individual; ω is the number of individuals owned by the weight vector; ε represents the number of subproblems with better evolution.
Output: the population evolution state
1 Determine the attribution of the weight vector;
2 **for**$d_{wi} \le d_m$, **do**
$JM^i_{wei} = 1$; determine that the individual belongs to the weight vector
else do$JM^i_{wei} = 0$
3 Calculate the number of individuals owned by the weight vector: $\omega = \sum_{i=1}^{K} JM^i_{wei}$;
4 Determine the degree of evolution of subproblems;
5 **for**$\omega \ge \omega_m$, **do**
$JM^i_{sub} = 1$ and determine the degree of evolution of subproblems is better
else do$JM^i_{sub} = 0$;
6 Calculate the number of subproblems with better evolution: $\varepsilon = \sum_{i=1}^{K} JM^i_{sub}$
7 Determine the evolution stage of the population;
8 **for**$\varepsilon > \varepsilon_m$, **do**
$JM^i_{pop} = 2$ and determine that the current population evolution degree is too fast, belonging to an overevolution state;
else for$\varepsilon < \varepsilon_m$, **do**
$JM^i_{pop} = 0$ and determine that the current population is slowly evolving and belongs to a state of lagging evolution
else for $\varepsilon = \varepsilon_m$, **do**
$JM^i_{pop} = 1$ and determine that the current population has a good evolution speed and belongs to a normal evolutionary state
9 **end**

Algorithm 1: Judgment mechanism of population evolution status.

Input: the initial population $X = \{x_1, x_2, \cdots, x_n\}$; the size of the initial neighborhood corresponding to the subproblem within the population $T_s = \{T_{s1}, T_{s2}, \cdots, T_{sn}\}$; the population initial neighborhood T_i
Output: the size of the neighborhood corresponding to the subproblem within the population $T_s = \{T_{s1}, T_{s2}, \cdots, T_{sn}\}$, the current population size T^*.
1 Initialize: $T^* = T_i = T_s$.
2 evolution
3 **for**$x_i \in X$**do**
Take x_i corresponding to the individuals in neighborhood $B(i) = \{i_1, i_2, \cdots, i_T\}$ to perform crossover and mutation operations to obtain offspring individuals;
4 Determine the degree of evolution of the subproblems and the evolution status of the population according to the mechanism provided in Section 3.2.3;
5 Neighborhood adaptive adjustment
6 **for**$JM^i_{sub} = 1$, **do**
The evolution of the previous generation is fast and uses formula (11) to appropriately reduce the current neighborhood size T;
else for$JM^i_{sub} = 0$, **do**
The evolution of the previous generation is slow and uses formula (11) to appropriately increase the current neighborhood size T;
7 **for**$JM^i_{pop} = 1$, **do**
The evolution rate of the previous generation population is moderate;
else for$JM^i_{pop} = 2$, **do**
The evolution rate of the previous generation population is fast and uses formula (12) to appropriately reduce the current population size T^*;
else for$JM^i_{pop} = 0$, **do**
The evolution rate of the previous generation population is slow and uses formula (12) to appropriately increase the current population size T^*;
8 Output $T_s = \{T_{s1}, T_{s2}, \cdots, T_{sn}\}$ and T^*.

Algorithm 2: Neighborhood adaptive adjustment mechanism.

Input: optimal model of multiobjective fire resource allocation; termination criteria; population size N; population crossover probability p_c; probability of population variation p_m;
Output: optimal plan for firepower resource allocation
1 Initialize
2 EP = ∅
3 Initialize population individuals $x_1, x_2, \cdots, x_N$;
4 Generate weight vector $\varphi_1, \varphi_2, \ldots, \varphi_N$;
5 Calculate the Euclidean distance between any two weight vectors. For each weight vector, find the T nearest weight vectors to form its neighborhood. $i = 1, 2, \cdots, N$ and $B(i) = \{i_1, i_1, \cdots, i_T\}$. Among them, $\varphi_{i_1}, \varphi_{i_2}, \ldots, \varphi_{i_T}$ are the T weight vectors closest to φ_i;
6 Initialize the ideal point $Z^* = (Z_1, Z_2, .., Z_m)^T$;
7 Evolution
8 **for** $i = 1, 2, \cdots, N$, **do**
9 Crossover and mutation: randomly select two individuals in $B(i)$ to perform crossover and mutation operations to obtain offspring individual y;
10 Update ideal point Z^*
11 **for each** $j = 1, 2, \cdots, m$.
if $Z_j < f_j(y)$, **then** $Z_j = f_j(y)$
end
12 Set adaptive neighborhood
13 Judge the subproblems and the evolution status of the population through the criteria provided in Section 2.1;
14 Use the method provided in Section 2.2 to obtain the population neighborhood T^* and the neighborhood T corresponding to the subproblem;
15 Update neighborhood $B(i)$
for each $j \in B(i)$
if $g^{\omega s}(y' | \lambda_j, z) \le g^{\omega s}(x^j | \lambda_j, z)$ then
$x^j = y'$, $FV^j = F(y')$
end
end
16 Remove all individuals dominated by $F(y')$ in EP and add individuals not dominated to EP at the same time;
17 **end**
18 **Stop operation**
19 After the algorithm evolves to the maximum algebra G_max, it stops and outputs the optimal solution. If the stopping condition is not met, it returns to Step 7.

Algorithm 3: The framework of NAAM-MOEA/D algorithm.

Table 1: Optimization model parameters of firepower allocation resources.

Project	W1	W2	W3	W4	Importance
T1	0.82	—	—	—	0.22
T2	—	0.95	—	—	0.31
T3	—	—	0.87	—	0.28
T4	—	—	—	0.85	0.19
Unit cost	5	10	8	4	—
Total number	10	10	10	10	—

Some scholars propose that if the distance between the subproblem and a certain solution in space is used as the evaluation criterion, if the distance between them is relatively close, it can be judged that the solution belongs to the subproblem. In the spatial coordinate system, the solution corresponds to the individual in the coordinate system, and the subproblem corresponds to the weight vector. Therefore, the problem of determining the attribution of the solution can be transformed into the problem of finding the distance between the weight vector and the individual.

This paper proposes a mechanism for evaluating the evolutionary stage of a population (see Algorithm 1):

3.2.4. Neighborhood Adaptive Adjustment Mechanism. In order to meet the needs of balancing the convergence and distribution of the MOEA/D algorithm, according to the population evolution state judgment mechanism in Section 3.2.3, this paper proposes a neighborhood strategy that adaptively adjusts the population size based on the different evolution stages of the population, which can also be called a neighborhood adaptive adjustment mechanism (NAAM) as in Algorithm 2.

The setting adjustment formula is as follows:

$$T_s = \begin{cases} T^* \left[1 - \frac{1}{T^*} \omega \left(\frac{mT^*}{N} \right)^{\theta} \right], & \mathrm{JM}_{\mathrm{sub}}^i = 1, \\ T^* \left[1 + \frac{1}{T^*} \omega \left(\frac{mT^*}{N} \right)^{\theta} \right], & \mathrm{JM}_{\mathrm{sub}}^i = 0, \end{cases} \tag{11}$$

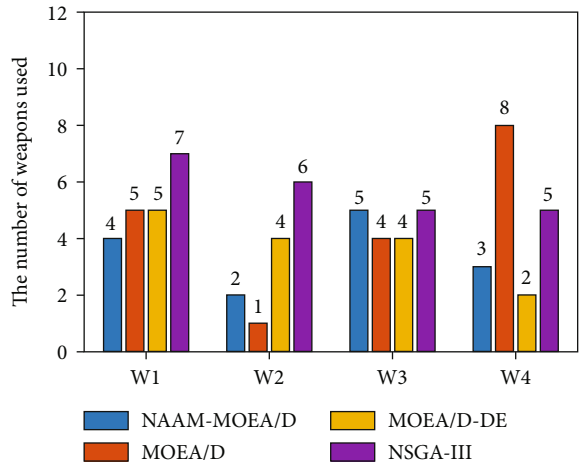


Figure 1: Number of weapons of each type used by the 4 algorithms.

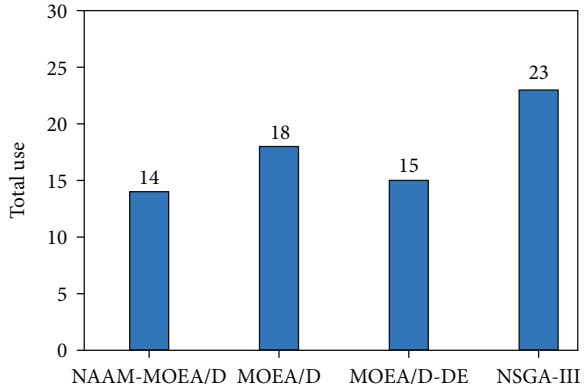


Figure 2: Total number of weapons used by the 4 algorithms.

$$T^* = \begin{cases} T_i\left[1 - \frac{1}{T_i}\omega\left(\frac{mT_i}{N}\right)^{\theta}\right], & \mathrm{JM}_{\mathrm{pop}}^{i} = 2, \\ T_i, & \mathrm{JM}_{\mathrm{pop}}^{i} = 1, \\ T_i\left[1 + \frac{1}{T_i}\omega\left(\frac{mT_i}{N}\right)^{\theta}\right], & \mathrm{JM}_{\mathrm{pop}}^{i} = 0. \end{cases} \quad (12)$$

3.2.5. The Framework of NAAM-MOEA/D Algorithm. The framework of the NAAM-MOEA/D algorithm can be described as in Algorithm 3:

4. Experiment and Simulation

4.1. Example Analysis of Algorithm. There are 4 types of weapons to strike at 4 targets in the enemy's combat system. Combining the content of Section 3.1, we assume that the model satisfies various constraints, and the model parameters are given in Table 1.

MATLAB 2020 is selected to write the algorithm program. The running environment is a Windows 7 R64-bit operating system, 4 GB memory, Intel Pentium processor. The NAAM-MOEA/D algorithm, MOEA/D algorithm, MOEAD/D-DE algorithm, and NSGA-III algorithm are selected for the simulation operation.

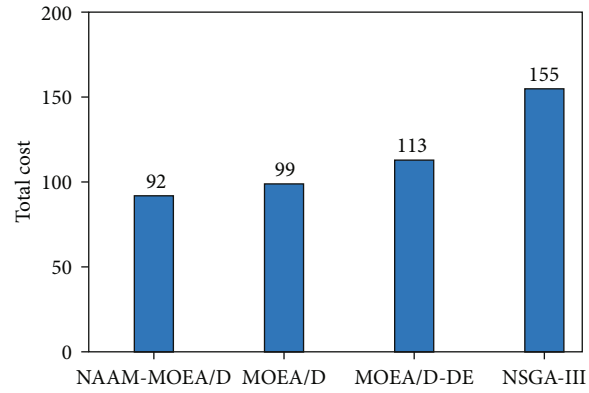


Figure 3: Total computational cost of the 4 algorithms.

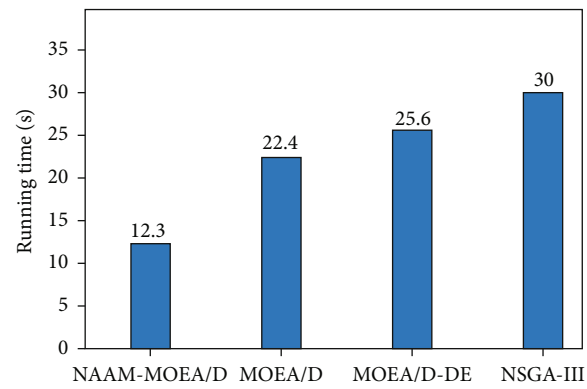
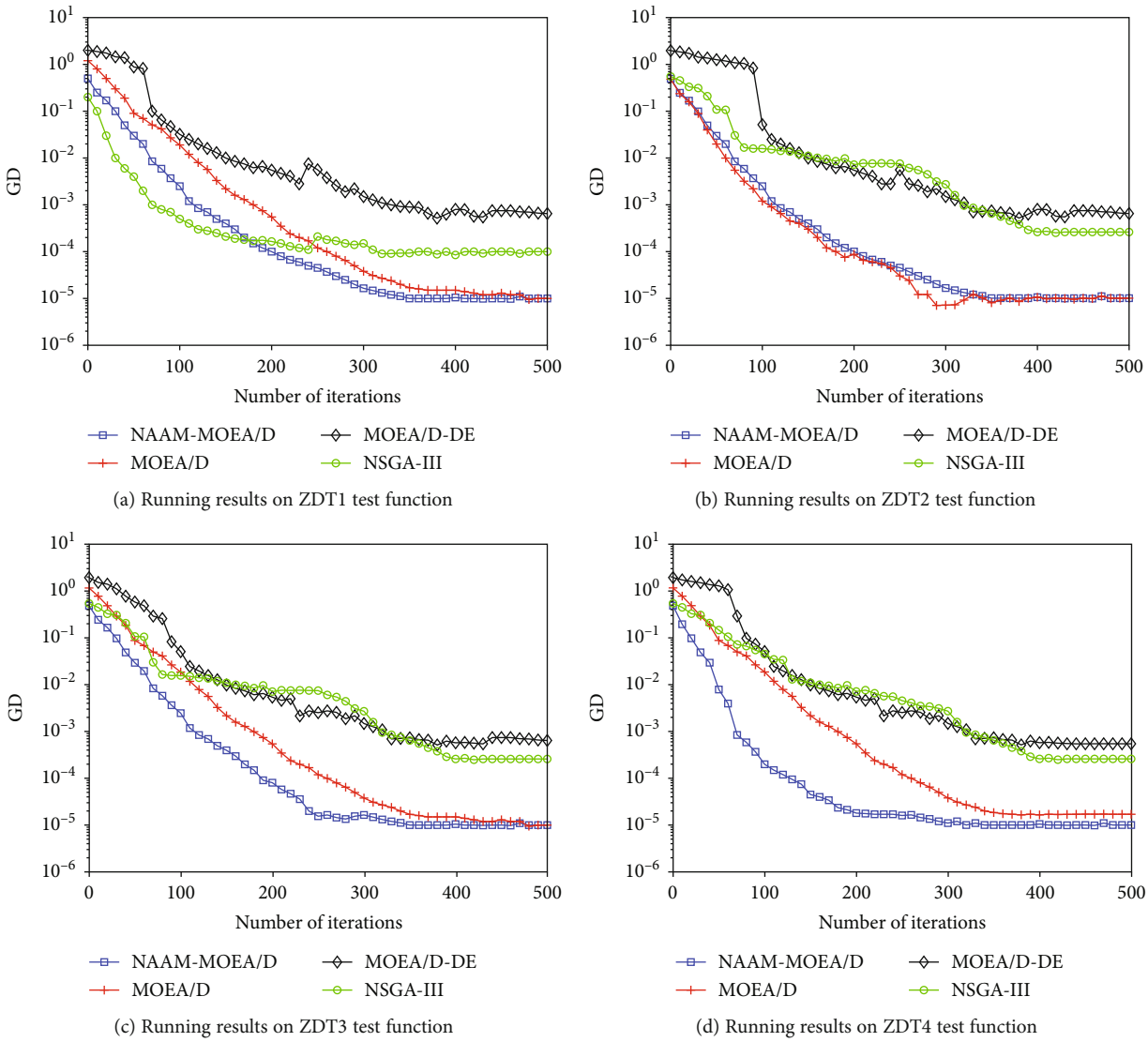


Figure 4: Calculation time spent of the 4 algorithms.

Table 2: Statistics of fire resource distribution.

Project	W1 T1	W2 T2	W3 T3	W4 T4	Total use	Total cost	Running time
NAAM-MOEA/D	4	2	5	3	14	92	12.3 s
MOEA/D	5	1	4	8	18	99	22.4 s
MOEAD/D-DE	5	4	4	2	15	113	25.6 s
NSGA-III	7	6	5	5	23	155	30 s

Figure 1 counts the number of various weapons used by the four algorithms. It can be seen from Figure 1 that the NSGA-III algorithm uses 7 W1 weapons, which is more than the MOEA/D algorithm and the MOEA/D-DE algorithm; both algorithms use 5 W1 weapons, and the NAAM-MOEA/D algorithm uses 4 W1 weapons. The NSGA-III algorithm uses 6 W2 weapons; the MOEA/D-DE algorithm and the NAAM-MOEA/D algorithm use 4 and 2 W2 weapons, respectively; while the MOEA/D algorithm uses the least number of W2 weapons and only one is used. The NAAM-MOEA/D algorithm and the NSGA-III algorithm both use 5 W3 weapons, which is more than the MOEA/D algorithm and the MOEA/D-DE algorithm. Both algorithms use 4 W3 weapons. The MOEA/D algorithm uses 8 W4



(a) Running results on ZDT1 test function

(b) Running results on ZDT2 test function

(c) Running results on ZDT3 test function

(d) Running results on ZDT4 test function

Figure 5: Variation curve of GD with algorithm iteration number.

weapons, which is more than the NSGA-III algorithm. The NAAM-MOEA/D algorithm and the MOEA/D algorithm use 3 and 2 W4 weapons, respectively.

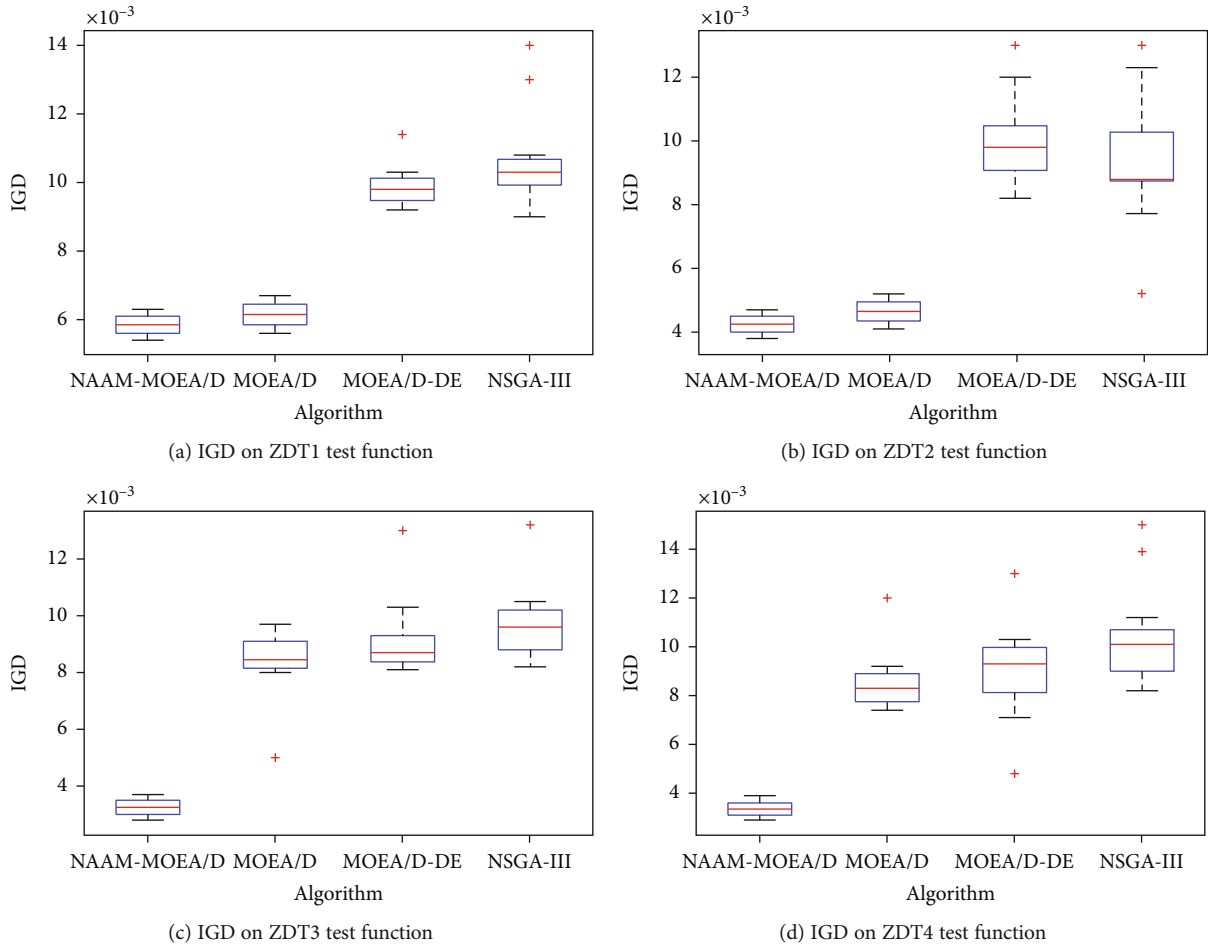
Figure 2 counts the total number of weapons used by the four algorithms. It can be seen from Figure 2 that the NSGA-III algorithm uses the largest number of weapons, using 23 weapons in total. The MOEA/D algorithm and the MOEA/D-DE algorithm use 18 and 15 weapons, respectively, and the NAAM-MOEA/D algorithm uses the least amount of weapons—only 14 weapons are used.

Figure 3 compares the total cost of weapon use of the four algorithms. It can be seen from Figure 3 that the NSGA-III algorithm costs the most weapons, with a total cost of 155, followed by the MOEA/D-DE algorithm, with a total cost of 113, while the MOEA/D algorithm and NAAM-MOEA/D algorithm had the least weapon use cost, costing 99 and 92, respectively.

Figure 4 compares the computing time of the four algorithms. It can be seen from Figure 4 that the NAAM-MOEA/D algorithm has the least computing time, which takes only 12.3 s, and the NSGA-III algorithm has the most computing time, which takes 30 s. The computing times of the MOEA/D algorithm and the MOEA/D-DE algorithm are, respectively, 22.4 s and 25.6 s.

The statistics of firepower resource allocation obtained through simulation calculation are shown in Table 2.

It can be seen from Table 2 that the number of weapons used and the total cost obtained by the NAAM-MOEA/D algorithm are better than those of the other three algorithms. The number of weapons used by the MOEA/D-DE algorithm is close to the number of weapons used by the NAAM-MOEA/D algorithm, but the total cost is about 23% higher. The total cost calculated by the MOEA/D-DE algorithm is close to the total cost calculated by the NAAM-MOEA/D algorithm, but 4 more weapons are used. The number of weapons used and the total cost obtained by the NSGA-III algorithm are significantly more than those of the other three algorithms, indicating that the algorithm has the worst performance. In addition, the running time of the NAAM-MOEA/D algorithm is 12.3 s, which is reduced by 82.1%, 108.1%, and 153.6% compared with the MOEA/D algorithm,



(a) IGD on ZDT1 test function

(b) IGD on ZDT2 test function

(c) IGD on ZDT3 test function

(d) IGD on ZDT4 test function

Figure 6: IGD box plot of the algorithm under different test functions.

Table 3: Comparison of IGD indicators of algorithms on ZDT series functions.

Test function	NAAM-MOEA/D Mean (std)	MOEA/D Mean (std)	MOEA/D-DE Mean (std)	NSGA-III Mean (std)
ZDT1	5.83*E*-03 (4.73*E*-04)	5.96*E*-03 (5.13*E*-04)	9.57*E*-03 (1.31*E*-02)	1.01*E*-02 (1.65*E*-02)
ZDT2	4.17*E*-03 (4.85*E*-04)	4.51*E*-03 (8.13*E*-04)	9.33*E*-03 (1.07*E*-04)	8.75*E*-03 (2.27*E*-03)
ZDT3	3.17*E*-03 (2.05*E*-04)	8.48*E*-03 (9.30*E*-04)	8.73*E*-03 (5.95*E*-04)	9.55*E*-03 (2.68*E*-04)
ZDT4	3.25*E*-03 (2.72*E*-04)	8.21*E*-03 (4.03*E*-04)	9.12*E*-03 (7.15*E*-04)	1.04*E*-02 (4.52*E*-03)

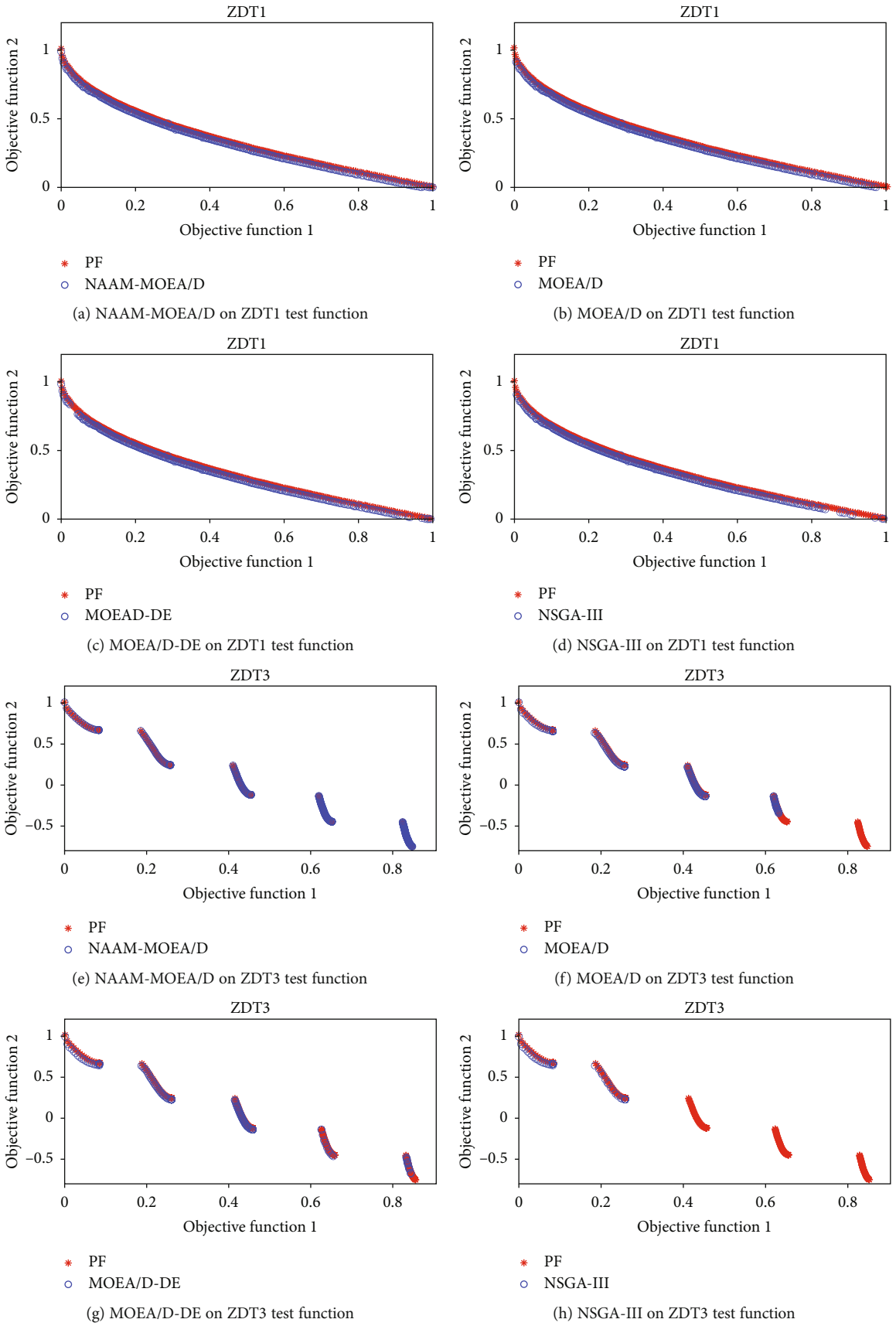
the MOEA/D-DE algorithm, and the NSGA-III algorithm, respectively, indicating that the NAAM-MOEA/D algorithm has obvious advantages in computing speed.

4.2. Performance Test of the Algorithm. In order to verify the performance of the NAAM-MOEA/D algorithm, ZDT series of test functions are selected to test the performance of the NAAM-MOEA/D algorithm with the MOEA/D algorithm, MOEA/D-DE algorithm, and NSGA-III algorithm.

In order to ensure the fairness and rationality of the algorithm evaluation, the population size and initial neighborhood size of the four algorithms are set to the same (population size $N = 100$, initial neighborhood $T = 100$). All algorithms adopt simulated binary crossover (crossover probability $p_c = 0.9$) and polynomial mutation (mutation probability $p_m = 1/n$, n is the dimension of decision variables). Each algorithm runs 20 times independently, and the evaluation times are set to 10000. Inverse generation distance (IGD) and generation distance (GD) were used as evaluation indexes. Each test function is run 20 times independently and averaged every 10 generations. The variation curve of GD with the number of iterations (0-500 generations) of the algorithm is shown in Figure 1.

As shown in Figure 5(a), the NAAM-MOEA/D algorithm tends to be stable on the test function ZDT1, and the convergence speed is slower than the NSGA-III algorithm and faster than the MOEA/D algorithm and the MOEA/D-DE algorithm.

As shown in Figure 5(b), on the test function ZDT2, the convergence speed of the NAAM-MOEA/D algorithm is



(a) NAAM-MOEA/D on ZDT1 test function

(b) MOEA/D on ZDT1 test function

(c) MOEA/D-DE on ZDT1 test function

(d) NSGA-III on ZDT1 test function

(e) NAAM-MOEA/D on ZDT3 test function

(f) MOEA/D on ZDT3 test function

(g) MOEA/D-DE on ZDT3 test function

(h) NSGA-III on ZDT3 test function

Figure 7: Continued.

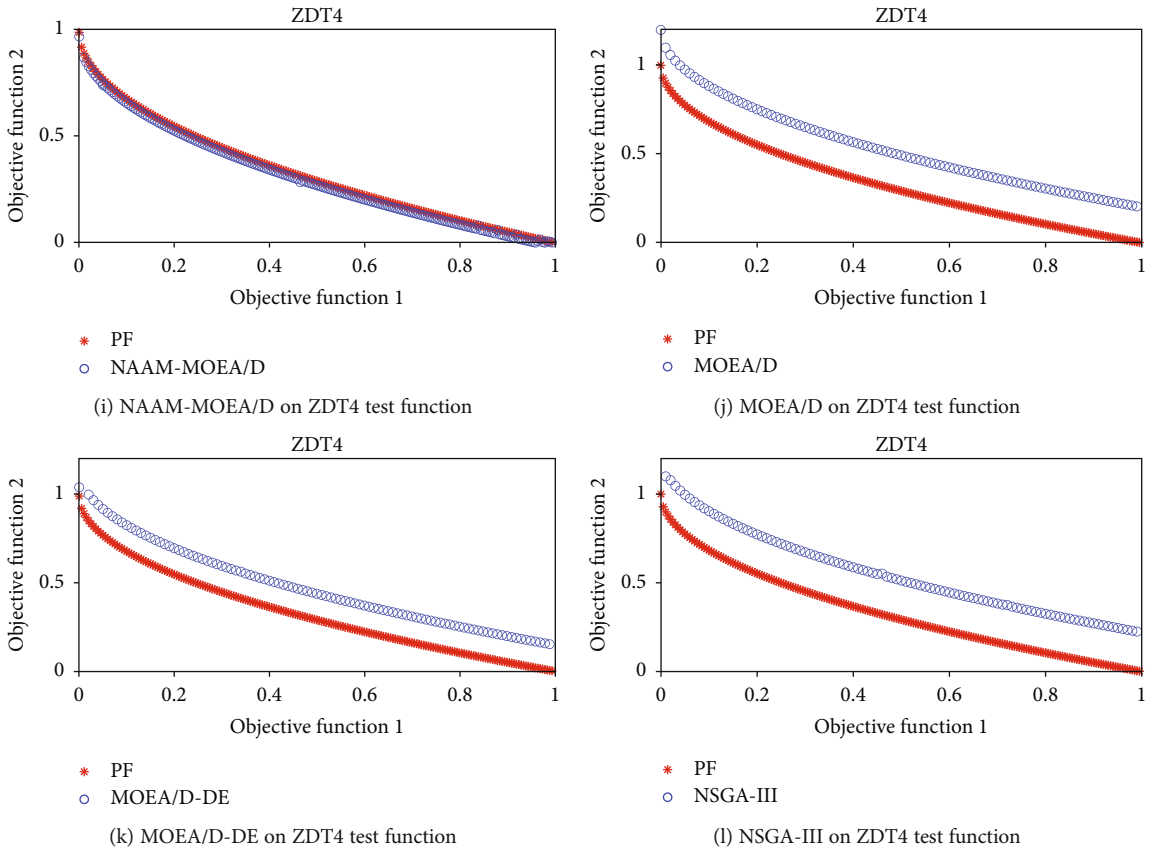


Figure 7: Comparison of Pareto front and ideal Pareto front on ZDT test function.

faster than that of the MOEA/D-DE algorithm and the NSGA-III algorithm. Although it is slightly slower than the MOEA/D algorithm, the population degradation degree of the MOEA/D algorithm is higher than that of the NAAM-MOEA/D algorithm.

As shown in Figure 5(c), on the test function ZDT3, the NAAM-MOEA/D algorithm converges faster than the other algorithms.

As shown in Figure 5(d), on the test function ZDT4, the NAAM-MOEA/D algorithm has a faster population convergence speed due to the advantages of the adaptive neighborhood adjustment mechanism adopted, and the algorithm convergence performance is significantly better than the MOEA/D algorithm, MOEAD/D-DE algorithm, and NSGA.

Therefore, the NAAM-MOEA/D algorithm not only ensures that the algorithm has a faster convergence rate but also solves the population degradation problem that occurs during the algorithm operation and ensures the stability of the algorithm operation, so that the algorithm can have more resources to improve the diversity of the population.

As shown in Figure 6, comparing the IGD box plots of various algorithms on the ZDT series test functions in the comparison Table 3, we can see that the NAAM-MOEA/D algorithm's mean, minimum, median (at the position of the red line in the figure), and interquartile range (key indicators such as box length) are lower than those of the MOEA/D algorithm, MOEA/D-DE algorithm, and NSGA-III algorithm. The probability and size of the abnormal value of the NAAM-MOEA/D algorithm are also lower than those of the other three algorithms, which show that the stability and quality of the NAAM-MOEA/D algorithm is higher.

On the test functions ZDT1 and ZDT2, the comprehensive performance of the NAAM-MOEA/D algorithm is slightly better than that of the MOEA/D algorithm and significantly better than that of the MOEA/D-DE algorithm and the NSGA-III algorithm. On the test functions ZDT3 and ZDT4, the comprehensive performance of the NAAM-MOEA/D algorithm is significantly better than that of the MOEA/D algorithm, the MOEA/D-DE algorithm, and the NSGA-III algorithm. This is because there are many discontinuous regions in the target space of test function ZDT3. These regions adopt the fixed neighborhood setting method, but do not use the adaptive neighborhood allocation strategy to reasonably allocate the algorithm, which leads to the waste of algorithm resources and the slowdown of population evolution speed.

Figure 7 shows the comparison of the Pareto front and the ideal Pareto front obtained by the four algorithms on the ZDT test function. Among them, the red meter character represents the ideal PF, and the blue circle represents the optimal solution of the Pareto frontier obtained by the various algorithms.

On the test function ZDT3, the improved MOEA/D solution set is more evenly distributed on the ideal Pareto front. In the other three algorithms, some leading edges are not completely found, and the solution set is missing to a certain extent. Among them, the MOEA/D algorithm and the MOEA/D-DE algorithm have a little poor distribution of

solution set, while the NSGA-III algorithm has the least distribution. This is because the other algorithms spend limited computing resources in the discrete region of test function ZDT3 and produce too many nondominated solutions, which hinders the evolution of the population.

On the test function ZDT4, the NAAM-MOEA/D algorithm has converged to the ideal, while the other algorithms have fallen into the local optimization state to varying degrees. It can be seen that the NAAM-MOEA/D algorithm has more advantages in reasonable allocation of computing resources and can better ensure the convergence of the algorithm.

Through the comparison, we can see that the Pareto frontier solution set obtained by the NAAM-MOEA/D algorithm almost uniformly converges to the PF of the ideal Pareto. However, the other three algorithms have different degrees of missing or uneven distribution of solution sets in various test functions. The NAAM-MOEA/D algorithm shows some performance advantages when dealing with simple test problems such as ZDT1, but the advantages are not obvious. However, the NAAM-MOEA/D can allocate computing resources reasonably and take into account the convergence and distribution of the algorithm due to its flexible neighborhood update strategy when dealing with relatively complex test problems such as ZDT3 and ZDT4.

5. Discussion

In this section, we establish a firepower resource allocation optimization model for edge environment based on given specific data, conduct simulation experiments, and test and evaluate the performance of the algorithm combined with the ZDT series of functions. However, several additional points should be pointed out and further analyzed in detail, which are specified as below.

(1) The types of weapons and the number of samples given in Section 4.1 are not large enough (both are 4). Therefore, in the future simulation experiments, we should focus on large sample data sets to verify the performance of the method under the condition of large sample data

(2) In Section 4.2, the ZDT series functions are selected to test the performance of the algorithm. The simulation results show that the performance of the NAAM-MOEA/D algorithm is better than that of the other three algorithms. However, only one kind of test function verification is not convincing enough, so DLTZ, WFG, and other test functions should be selected to evaluate the algorithm, so as to provide more sufficient reference for the improvement of algorithm performance

6. Conclusion

This paper constructs a multiobjective firepower resource allocation optimization model for edge environment with limited computing resources, based on maximizing damage effect and minimizing combat cost. Aiming at the defects of the traditional MOEA/D algorithm fixed neighborhood update mechanism, a MOEA/D algorithm based on neighborhood adaptive adjustment mechanism is proposed and the model is solved. It can be seen from the simulation experiment that the MOEA/D algorithm based on the neighborhood adaptive adjustment mechanism has significantly improved its stability, convergence, and distribution.

In the next step, current work will continue to be improved by considering security and privacy issues [24–33]. In addition, more complex multiobjective solutions with more context factors [34–41] will be considered.

Abbreviations

MOEA/D: Multiobjective evolutionary algorithm based on decomposition
NAAM-MOEA/D: Neighborhood adaptive adjustment mechanism-multiobjective evolutionary algorithm based on decomposition
MODPSO-GSA: Multiobjective discrete particle swarm optimization-gravitational search algorithm
WMOM/D: Weapon-target assignment multiobjective model based on decomposition
GD-MOEA/D: Gauss mutation and differential evolution based on a multiobjective evolutionary algorithm based on decomposition
MOEA/D-DE: Multiobjective evolutionary algorithm based on decomposition-differential evolution
MOEA/D-DRA: Multiobjective evolutionary algorithm based on decomposition-dynamical resource allocation
ENS-MOEA/D: Ensemble neighborhood size-multiobjective evolutionary algorithm based on decomposition
ADEMO/D-ENS: Adaptive differential evolution for multiobjective problems-ensemble neighborhood size
MOEA/D-AGR: Multiobjective evolutionary algorithm based on decomposition-adaptive global replacement
MOEA/D-NMO: Multiobjective evolutionary algorithm based on decomposition-neighborhood mutation operator
MOEA/D-ANS: Multiobjective evolutionary algorithm based on decomposition-adaptive neighborhood strategy
NSGA-III: Nondominated sorted genetic algorithm-III.

Data Availability

The experiment dataset is generated randomly through simulation.

Conflicts of Interest

We declare that there is no conflict of interest regarding this submission.

Authors' Contributions

Liyuan Deng finished the English writing, review, and editing of the paper. Liyuan Deng, Ping Yang, and Weidong Liu finished the experiments. Lina Wang, Sifeng Wang, and Xiumei Zhang finished the algorithm design.

Acknowledgments

This work was supported by Xi'an Research Institute of High-Technology.

References

[1] L. I. Ping and L. I. Changwen, "Modeling and algorithm of weapon target cooperative fire assignment," *Command Control & Simulation*, vol. 37, no. 2, pp. 36–40, 2015.

[2] J. Zhang, Z. X. Wang, L. Chen, Z. B. Wu, and J. F. Lu, "Modeling and optimization on antiaircraft weapon-target assignment at multiple interception opportunity," *Acta Armamentarii*, vol. 35, no. 10, pp. 1644–1650, 2014.

[3] D. Chao-yang, L. Yao, and W. Qing, "Improved genetic algorithm for solve firepower distribution," *Acta Armamentarh*, vol. 37, no. 1, pp. 97–102, 2016.

[4] C. L. Fan, Q. H. Xing, and M. F. Zheng, "Weapon-target allocation optimization algorithm based on IDPSO," *Systems Engineering and Electronics*, vol. 37, no. 2, pp. 336–342, 2015.

[5] X. Hao, X. Qinghua, and W. Wei, "WTA for air and missile defense based on fuzzy multi-objective programming," *Systems Engineering and Electronics*, vol. 40, no. 3, pp. 563–570, 2018.

[6] L. Qingguo, L. Xinxue, W. Jian, L. Yaxiong, and C. Hao, "Optimization of fire distribution for multiple SGSW based on improved NSGA-III," *Systems Engineering and Electronics*, vol. 42, no. 9, pp. 1995–2002, 2020.

[7] J. J. Gu, J. J. Zhao, J. Yan, and X. Chen, "Cooperative weapon-target assignment based on multi-objective discrete particle swarm optimization-gravitational search algorithm in air combat," *Journal of Beijing University of Aeronautics and Astronautics*, vol. 41, no. 2, pp. 252–258, 2015.

[8] Q. Zhang and H. Li, "MOEA/D: a multiobjective evolutionary algorithm based on decomposition," *IEEE Transaction on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.

[9] S. Zhao, P. Suganthan, and Q. Zhang, "Decomposition-based multiobjective evolutionary algorithm with an ensemble of neighborhood sizes," *IEEE Transactions on Evolutionary Computation*, vol. 16, no. 3, pp. 442–446, 2012.

[10] Y. Zhang, R. N. Yang, J. L. Zuo, and X. Jing, "Weapon-target assignment based on decomposition-based evolutionary multi-objective optimization algorithms," *Systems Engineering and Electronics*, vol. 36, no. 12, pp. 2435–2441, 2014.

[11] L. Chen and Y. Ma, "Anti-submarine firepower optimization of aircraft carrier formation based on GD-MOEA/D algorithm," *Computer Simulation*, vol. 35, no. 10, pp. 33–38, 2018.

[12] C. Hui and Y. Ma, "Model of target assignment in joint fire strike operations," *Journal of Systems Simulation*, vol. 30, no. 8, pp. 2942–2949, 2018.

[13] H. Li and Q. Zhang, "Multiobjective optimization problems with complicated Pareto sets, MOEA/D and NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 2, pp. 284–302, 2009.

[14] Z. H. A. N. G. Qingfu, L. I. U. Wudong, and L. I. Hui, "The performance of a new version of MOEA/D on CEC09 unconstrained MOP test instances," in *2009 IEEE Congress on Evolutionary Computation*, pp. 203–208, Washington D.C., USA, 2009.

[15] L. Li, D. Liu, and X. Wang, *Multi-objective permutation flow shop scheduling problem based on improved MOEA/D algorithm*, Computer Integrated Manufacturing Systems, 2020.

[16] X. Zhou, W. Xuewu, and X. Gu, "MOEA/D based on constrained approach and differential evolution," in *Proceedings of the 38th Chinese Control Conference*, pp. 2034–2039, Guangzhou Baiyun International Convention Center, China, 2019.

[17] H. Ishibuchi, Y. Hitotsuyanagi, N. Tsukamoto, and Y. Nojima, "Use of biased neighborhood structures in multiobjective memetic algorithms," *Soft Computing*, vol. 13, no. 8-9, pp. 795–810, 2009.

[18] S.-Z. Zhao, P. N. Suganthan, and Q. Zhang, "Decomposition-based multiobjective evolutionary algorithm with an ensemble of neighborhood sizes," *IEEE Transactons on Evolusonary Computation*, vol. 16, no. 3, pp. 442–446, 2012.

[19] H. Xia, J. Zhuang, and D. Yu, "Combining crowding estimation in objective and decision space with multiple selection and search strategies for multi-objective evolutionary optimization," *IEEE Transactions on Cybernetics*, vol. 44, no. 3, pp. 378–393, 2013.

[20] Z. Wang, Q. Zhang, A. Zhou, M. Gong, and L. Jiao, "Adaptive replacement strategies for MOEA/D," *IEEE Transcations on Cybernetics*, vol. 46, no. 2, pp. 474–486, 2016.

[21] L. Liu and L. Zheng, "MOEA/D algorithm based on combinational optimization of neighborhood and mutation operator," *Computer Engineering*, vol. 43, no. 3, pp. 232–240, 2017.

[22] E. Li and R. Chen, "Improved MOEA/D algorithm based on adaptive mutation operator and neighborhood size," *Computer Engneering and Applications*, vol. 55, no. 9, pp. 49–55, 2019.

[23] H. Geng, W. Han, Y. Ding, and S. Zhou, "Improved MOEA/D algorithm based on adaptive neighborhood strategy," *Computer Engineering*, vol. 45, no. 5, pp. 161–168, 2019.

[24] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2016.

[25] Z. Sun, Y. Wang, Z. Cai, T. Liu, X. Tong, and N. Jiang, "A two-stage privacy protection mechanism based on blockchain in mobile crowdsourcing," *International Journal of Intelligent Systems*, 2021.

[26] Y. Xu, J. Ren, Y. Zhang, C. Zhang, B. Shen, and Y. Zhang, "Blockchain empowered arbitrable data auditing scheme for network storage as a service," *IEEE Transactions on Services Computing*, vol. 13, no. 2, pp. 289–300, 2020.

[27] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2020.

[28] L. Qi, C. Hu, X. Zhang et al., "Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4159–4167, 2020.

[29] T. Liu, Y. Wang, Y. Li, X. Tong, L. Qi, and N. Jiang, "Privacy protection based on stream cipher for spatiotemporal data in IoT," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 7928–7940, 2020.

[30] Z. Cai and Z. He, "Trading private range counting over big IoT data," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, Dallas, TX, USA, 2019.

[31] Y. Xu, C. Zhang, G. Wang, Z. Qin, and Q. Zeng, "A blockchain-enabled deduplicatable data auditing mechanism for network storage services," *IEEE Transactions on Emerging Topics in Computing*, p. 1, 2020.

[32] W. Zhong, X. Yin, X. Zhang et al., "Multi-dimensional quality-driven service recommendation with privacy-preservation in mobile edge environment," *Computer Communications*, vol. 157, pp. 116–123, 2020.

[33] Q. Liu, Y. Tian, J. Wu, T. Peng, and G. Wang, "Enabling verifiable and dynamic ranked search over outsourced data," *IEEE Transactions on Services Computing*, p. 1, 2019.

[34] L. Wang, X. Zhang, R. Wang, C. Yan, H. Kou, and L. Qi, "Diversified service recommendation with high accuracy and efficiency," *Knowledge-Based Systems*, vol. 204, article 106196, 2020.

[35] J. Li, T. Cai, K. Deng, X. Wang, T. Sellis, and F. Xia, "Community-diversified influence maximization in social networks," *Information Systems*, vol. 92, article 101522, 2020.

[36] H. Liu, H. Kou, C. Yan, and L. Qi, "Keywords-driven and popularity-aware paper recommendation based on undirected paper citation graph," *Complexity*, vol. 2020, Article ID 2085638, 15 pages, 2020.

[37] S. Zhang, Q. Liu, and Y. Lin, "Anonymizing popularity in online social networks with full utility," *Future Generation Computer Systems*, vol. 72, no. 7, pp. 227–238, 2017.

[38] Z. Chunjie, L. Ali, H. Aihua, Z. Zhiwang, Z. Zhenxing, and W. Fusheng, "Modeling methodology for early warning of chronic heart failure based on real medical big data," *Expert Systems with Applications*, vol. 151, article 113361, 2020.

[39] T. Cai, J. Li, A. S. Mian, R. Li, T. Sellis, and J. X. Yu, "Target-aware holistic influence maximization in spatial social networks," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2020.

[40] Q. Liu, P. Hou, G. Wang, T. Peng, and S. Zhang, "Intelligent route planning on large road networks with efficiency and privacy," *Journal of Parallel and Distributed Computing*, vol. 133, pp. 93–106, 2019.

[41] Y. Wang, G. Yang, Y. Li, and X. Tong, "A worker-selection incentive mechanism for optimizing platform-centric mobile crowdsourcing systems," *Computer Networks*, vol. 107, article 107144, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6610654, 12 pages
https://doi.org/10.1155/2021/6610654

Research Article

An Optimization Method for Mobile Edge Service Migration in Cyberphysical Power System

Qian Cao,[1] **Qilin Wu**,[1] **Bo Liu**,[1] **Shaowei Zhang**,[2] **and Yiwen Zhang**[3]

[1]*School of Information Engineering, Chaohu University, Chaohu 238000, China*
[2]*School of Computer Science and Technology, Anhui Wenda University of Information Engineering, Hefei 230000, China*
[3]*School of Computer Science and Technology, Anhui University, Hefei 230000, China*

Correspondence should be addressed to Qilin Wu; lingqiw@126.com

Received 1 December 2020; Revised 14 January 2021; Accepted 31 January 2021; Published 15 February 2021

Academic Editor: Mohammad R. Khosravi

To relieve the pressure of processing computation-intensive applications on mobile devices and avoid high latency during data transmission, edge computing is proposed to solve this problem. Mobile edge computing (MEC) allows the deployment of MEC servers at the edge of the network to interact with users on the premise of low transmission delay, thereby improving the quality of service (QoS) for users. However, due to the high mobility of users, with the continuous change of geographical location, when users exceed the signal range of the MEC server, the services they request on the MEC server will also be migrated to other MEC servers. The handoff process may involve high response delays caused by service forwarding, thereby greatly degrading QoS. For the above problems, in this paper, a service migration optimization method based on transmission power control is proposed. First, according to the transmission power of the MEC server, the user's activity range is divided into multiple subregions based on a Voronoi diagram. Therefore, there is one MEC server in each subregion, and the size of each subregion is adjusted by controlling the transmission power of the MEC server to minimize the number of wireless handoffs and the energy consumption of the MEC server. Then, the particle swarm optimization (PSO) is adopted to solve the above multiobjective optimization problem. Finally, the effectiveness of the proposed method is verified through extensive experiments.

1. Introduction

Nowadays, with the rapid development of mobile devices, mobile applications are becoming more and more complex, and mobile devices with limited resources usually cannot meet the needs of most applications. Therefore, the industry began to consider offloading such computation-intensive applications to the cloud [1]. However, the remote offloading in traditional cloud computing may involve high latency and cannot meet the low-latency requirements [2] of some latency-sensitive applications, including augmented reality (AR) and remote game control [3]. Meanwhile, the exponential growth of information caused by a large number of devices and applications has brought tremendous pressure to remote information transmission. To solve the above problems, mobile edge computing (MEC) has been proposed, and a large number of servers are placed at the network edge [4, 5]. MEC is regarded as a supplement to mobile devices with relatively limited computational and storage capacity, which can enable computation offloading and provide services to users. In MEC, a new computing device called an MEC server, which is deployed on the base station to provide services and computing resources for users, is deployed at the network edge to act as a small cloud data center, giving the network edge the ability to process data [6]. Clearly, MEC servers can provide users with cloud services closer to the end-users so that users can request services with low latency.

Service providers can deploy related services on the MEC server to improve user experience, expand the user market, and earn more benefits. The reason is that the use of the MEC server helps users' mobile devices meet the performance requirements of some applications, greatly reduces information transmission delays, improves users' QoS, and reduces the traffic between users and the core network, thereby reducing operating cost. However, in the mobile edge

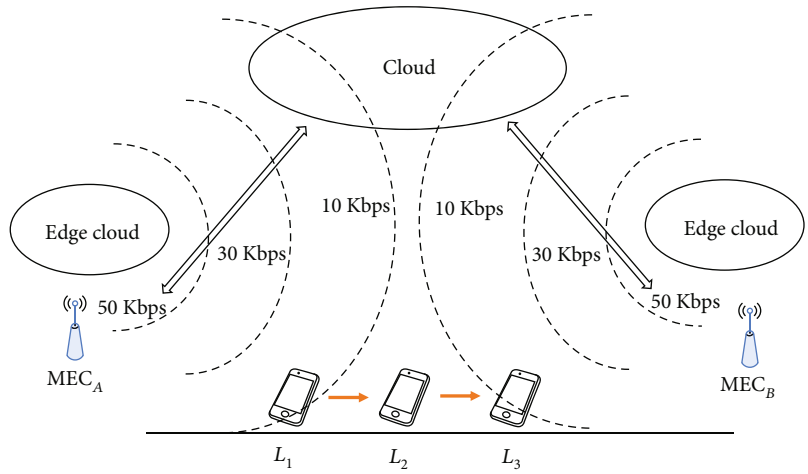


Figure 1: Example of edge service migration in mobile network.

computing environment, mobile devices often have high mobility, and service migration operations may be involved due to the time-varying location of their users. As shown in Figure 1, the edge computing system structure includes two MEC servers, MEC$_A$ and MEC$_B$, and the signal ranges corresponding to different powers are shown in the dotted arc in the figure. While the user moves from L_1 via L_2 to L_3, when the user walks out of the signal range of theMEC$_A$ server and enters the signal range of the MEC$_B$ server, the service requested by the user will be migrated from MEC$_A$ to MEC$_B$. Before the end of the handoff, the user can only use the services of the MEC$_A$. When the service request is forwarded from MEC$_A$ to MEC$_B$ through a limited capacity backhaul, the response time of the service will be significantly increased [7], resulting in a significant decrease in the quality of user experience.

Due to the limited coverage of the MEC server and the high time-varying location of users, users may switch servers frequently in the process of using mobile devices. So, the forwarding time and downtime of the request involved in this process will degrade QoS. At present, some studies propose to reduce the number and probability of service migration by deploying service copies in advance. However, at the same time point, users can only request services from one MEC server. The MEC servers that are not accessed by users but have deployed service copies will be occupied with unnecessary storage resources, resulting in a waste of resources.

In this paper, we focus on the problem of how to improve user QoS and effectively reduce server energy consumption when edge users have high mobility. Since the number of wireless handoffs and the energy consumption of the MEC are related to the coverage of the MEC server, it is necessary to reasonably control the coverage of the MEC server. However, the coverage of the MEC server is closely related to the transmission power of the MEC server. Therefore, by using transmission power control for service migration optimization, we can minimize the number of wireless handoffs and the energy consumption of an MEC server through service migration optimization. That is our motivation.

Compared with the existing methods, our main contributions can be summarized as follows:

(1) A service migration optimization method is proposed based on transmission power control. This method adjusts the size of each subarea according to the transmission power of the MEC server, so as to achieve the goal of minimizing the number of wireless handoffs and energy consumption of the MEC server

(2) The experimental scene is modeled by using the Voronoi diagram, and the multiobjective optimization problem is transformed into a single-objective optimization problem by using the weight coefficient transformation method. Furthermore, the PSO algorithm is used to solve the optimization problem, so as to achieve the goal of minimizing the user wireless handoff times and minimizing the energy consumption of the MEC server

(3) A large number of simulation experiments were carried out using a real base station data set, the Telecom Dataset [8–10], under the assumption that the user's mobile path is known, which verified the effectiveness and efficiency of the algorithm in this study

The remainder of this paper is organized as follows. Section 2 discusses and summarizes related work. In Section 3, we introduce the system model. After that, we introduce the PSO optimization method for minimizing the number of wireless handoffs of user equipment and minimizing the energy consumption of the MEC server in Section 4. Then, we give the experimental results and analysis in Section 5. Finally, the paper is concluded in Section 6.

2. Related Work

With the rapid development of the mobile Internet and the Internet of Things, a large number of delay-sensitive and computation-intensive applications have emerged. To meet

the low-latency and high-performance requirements, edge computing was proposed to provide solutions. Since the coverage of the MEC server is limited, as the user moves, the edge nodes that the user can connect to also change. If the user's service is always located on the MEC server where the user initially connected, the user's service request should be forwarded from the MEC server to the original MEC server through the backhaul link, which will increase the service delivery delay. Therefore, in order to maintain the user QoS, the edge service should be dynamically migrated between multiple MEC servers along with the user's movement.

At present, many studies have contributed to reducing service migration time. Taleb and Ksentini [11] proposed a Follow-Me Cloud (FMC) analysis model. In this work, the Markovian mobility model was proposed to analyze the performance of MEC when users move, but they did not consider whether or not services are migrated and where to migrate. In the scenario of a one-dimensional mobile model, Ksentini et al. [12] used the Markov decision process (MDP) to decide whether or not to carry out service migration by weighing the cost of service migration and the improved experience quality of users. However, their solution can be very time-consuming when MDP has a large number of states. Wang et al. [13] modeled the service migration problem as MDP and proved that when users follow a one-dimensional asymmetric random walk model, the best strategy for solving service migration is the threshold strategy, and they proposed an algorithm to find the best threshold. The algorithm proposed by Wang et al. is more effective than the standard solution of MDP. In addition, in [14], Chen et al. studied the service migration problem when users follow the two-dimensional mobile model. Afterwards, Wang et al. [15] proposed a layered migration architecture (base layer, application layer, and instance layer), which can effectively reduce transmission time. Machen et al. [16] proposed a service migration method based on container handoff, which uses a hierarchical storage system to reduce synchronization overhead of the file system, thereby reducing time cost for service migration. Furthermore, Ud Din et al. [17] studied the performance optimization of edge service under the constraint of long-term service migration overhead. To solve the problem of unpredictable user movement behavior, Ouyang et al. adopted the Lyapunov optimization to decompose the long-term optimization problem into a series of real-time optimization problems.

There are also some studies dedicated to reducing the probability and frequency of service migration to achieve the objective of optimizing user QoS. To reduce the delay caused by service migration, Ma et al. [18] proposed a CloudSpider architecture combining placement of a virtual machine (VM) replica and VM scheduling to reduce high migration delay caused by VM image transmission through low-bandwidth wide-area network (WAN) links. They used deduplication technology to compensate for the additional storage requirements caused by the placement of replicas, and studied the VM replica placement algorithm. Besides, Ouyang et al. [19] deployed service replicas on MEC servers near users in advance, aiming to minimize the probability of service migration and the number of service replicas. Yatao et al. [20] proposed an analysis model to compare the costs of service migration and service replica deployment. The model analyzes the impact of user movement mode and service duration on migration and replication costs, respectively. The above several studies deploy service replicas on MEC servers around users in advance to reduce the probability of service migration. However, the user can only request the resources of one MEC server at the same time, and the MEC server that is not accessed by the user but has service replicas will be occupied with storage resources. Since the resources on the MEC server are limited, the backup of useless resources will cause the resource waste of the MEC server.

In terms of reducing transmission delay, controlling transmission power is an effective method because transmission power is closely related to signal quality, interference, and channel capacity. Therefore, Bose et al. [21] proposed a cloud-aware power control method to maximize the result delivery rate under the condition of satisfying the delay requirement. Since the transmission power is also related to signal coverage, for the scenario of two MEC servers in [22], the coverage of MEC servers is controlled by transmission function, and VM migration is used to achieve load balancing, so the average service delay of MEC servers is reduced. Afterwards, Zhang et al. [23] extended the previous work in [24]. In the case of multiple MEC servers, they considered the mobility of users and studied how to maximize the cost-effectiveness, that is, to minimize the number of activated MEC servers, under the condition of meeting the service delay.

In summary, based on the transmission power control technology, this paper is aimed at solving the problem of user QoS degradation caused by user mobility in a mobile edge computing environment, and uses the weight coefficient transformation method and PSO algorithm to solve the multiobjective optimization problem, which can reduce the number of wireless handoffs and service migrations of user equipment and minimize the server energy consumption.

3. Problems and Models

3.1. Problem Description. Here, we consider n MEC servers and m users, which form two sets $E = \{e_1, e_2, \cdots, e_n\}$ and $U = \{u_1, u_{2,\cdots,} u_m\}$, respectively. The MEC server creates an isolated virtual machine environment for users. In order to meet the service delay requirements, we assume that the virtual machine is always placed on the MEC server connected by the user wirelessly. With the movement of users, radio network resources are also changing dynamically. So, service migration is always accompanied by wireless handoff. As a result, frequent wireless handoff and service migration will have a great impact on user experience.

Suppose there are three users u_1, u_2, and u_3, and their moving paths are shown in Figure 2. The coverage radius of the MEC server e_1 is configured as r_1 and r_3, and the coverage radius of e_2, it is configured as r_2 or r_4. Then, according to Figure 1, there will be four cases as follows:

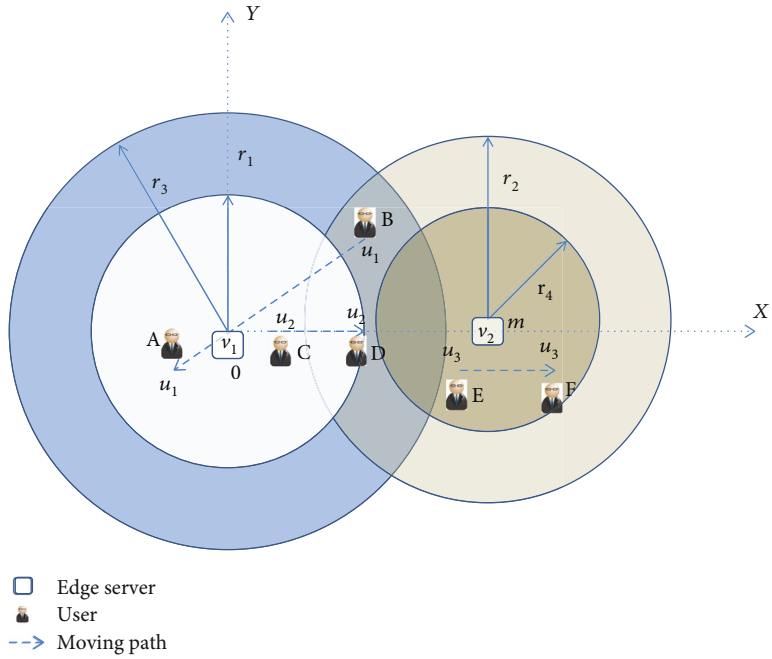


Figure 2: The impact of transmission power on service migration.

Case 1. If $r_{e_1} = r_1$, $r_{e_2} = r_2$, user u_1 leaves the coverage area of e_1 and wireless handoff and service migration need to be performed.

Case 2. If $r_{e_1} = r_1$, $r_{e_2} = r_4$, user u_1 cannot connect to any MEC server at point B.

Case 3. If $r_{e_1} = r_3$, $r_{e_2} = r_2$, u_1, u_2, and u_3 do not require wireless handoff and service migration.

Case 4. If $r_{e_1} = r_3$, $r_{e_2} = r_4$, as in case 3, all users do not need wireless handoff.

In summary, when the coverage radius is small, users will make frequent wireless handoffs, and even the connection will be interrupted (for case 2). When the coverage area is large, it will cause unnecessary high energy consumption and waste of resources (for cases 3 and 4, handoff can be avoided, and a smaller coverage area can meet the demand). The coverage of the MEC server is related to its transmission power [25]. Therefore, this section will study how to set the transmission power of the MEC server, so as to minimize the number of user wireless handoffs and the energy consumption of the MEC server.

3.2. System Model. The definitions of related concepts are given as follows:

Definition 1 (MEC server). The MEC server can be defined as a two-tuple $e = (p, \text{tp})$, where

(1) p is the location of the MEC server

(2) tp is the transmission power of the MEC server

Definition 2 (movement path). The user's movement path can be modeled as a triple mp = (time, location, M), where

(1) *Time* is the length of time that the user moves, which is composed of a series of discrete moments

(2) *Location* is the location of the user at the above moment

(3) M is a mapping relationship from time to location: M : time $\longrightarrow$ location

In this section, it is assumed that the user is always connected to the MEC server that provides the maximum received signal strength (RSS). Therefore, in this section, the user activity area is divided into an RSS Voronoi diagram. The RSS Voronoi diagram is defined as follows:

Definition 3 (RSS Voronoi diagram). Assuming that there is a group of MEC servers $E = \{e_1, e_2, \cdots, e_n\}$ in area A, the RSS Voronoi diagram divides area A into multiple V polygons, each V polygon has an MEC server, and the points in the V polygon are defined as follows:

$$V(e_i) = \left\{ u_k : \text{RSS}_{e_i}^{u_k} > \text{RSS}_{e_j}^{u_k} \forall j \neq i \right\}, \tag{1}$$

where $\text{RSS}_{e_i}^{u_k}$ is the received signal strength, user u_k is the receiver, and MEC server e_i is the sender.

The basic features of the Voronoi diagram are as follows:

(1) There is a generator in each subregion

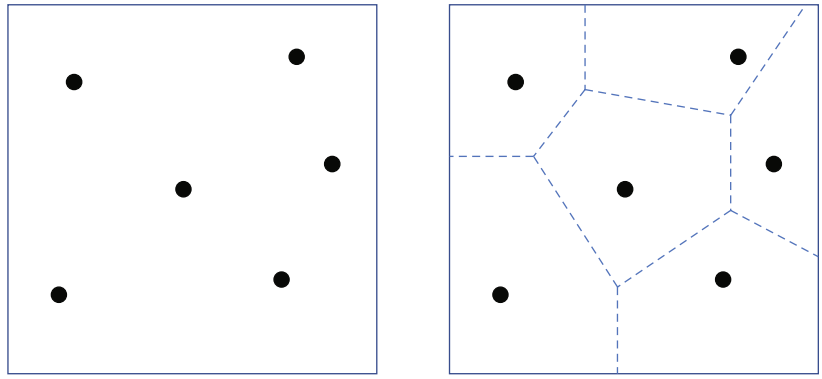


Figure 3: Examples of the Voronoi diagram.

(2) The distance from the point in the subregion to the generator is less than the distances to other generators

Different from the traditional Voronoi diagram, the standard of dividing the subregions is the received signal strength rather than Euclidean distance. Figure 3 is an example of the RSS Voronoi diagram, in which

(1) There is one MEC server in each subregion

(2) The signal strength received by the points in the subregion from the MEC server in the subregion is greater than that received from other MEC servers

(3) The points on the boundary of the subarea receive the same signal strength from the MEC server that generated the boundary

When the user's moving path spans multiple subregions, the user needs to perform multiple wireless handoffs. The division of subregions is related to the user's received signal strength, and the user's received signal strength is related to the transmission power of the MEC server. Therefore, the number of user wireless handoffs and service migrations can be reduced by controlling the transmission power of the MEC server. The relationship between the received signal strength of the user and the transmission power of the MEC server [26, 27] is as follows:

$$\mathrm{RSS}_{e_j}^{u_i} = \mathrm{tp}_{e_j} + G_{e_j} + G_{u_i} + H - L_{e_j}^{u_i}, \tag{2}$$

where tp_{e_j} is the transmission power of the MEC server e_j in decibels (dbm), G_{e_j} and G_{u_i} are the antenna gains of the sender and receiver, respectively, H is the Rayleigh power fading coefficient, and $L_{e_j}^{u_i}$ is the path loss between the sender and the receiver, which is calculated as follows [28–30]:

$$L_{e_j}^{u_i} = L_1 + 10n_1 \log_{10}\left(d\left(e_j, u_i\right)\right) + 10(n_2 - n_1)\left(1 + \frac{d\left(e_j, u_i\right)}{d_b}\right), \tag{3}$$

Table 1: The symbols commonly used in this section.

Symbols	Meaning
E	MEC server set
e_i	The ith MEC server in E
RSS	Received signal strength
Tp	Transmission power
G	Antenna gain
H	Rayleigh power fading coefficient
L	Path loss
n_1	Short-distance path loss index
n_2	Long-distance path loss index
d_b	Boundary value of long distance and short distance
EN_j	Energy consumption of MEC server j
Ht_i	Number of handoffs of user u_i

where L_1 is the path loss when the distance between the receiver and the sender is 1 meter; n_1 and n_2 are the long-distance and short-distance path loss indexes, respectively; $d(e_j, u_i)$ is the Euclidean distance between the MEC server e_j and the user u_i, d_b is the boundary value dividing long distance and short distance. The symbols commonly used in this section are shown in Table 1.

Through the mapping M : time $\longrightarrow$ location, the location of the user at any point in time can be derived, and the MEC server that the user is connected to can be obtained. According to the user's movement trajectory, the final number of user's handoffs can be obtained. Here, the user's mobile information can be obtained from many channels; for example, the user may apply map services (such as navigation). In addition, since the user's daily itinerary will not change much, the user's movement path can also be inferred through the user's past behavior. There are many studies on predicting user mobile behavior [31–33], which is beyond the scope of this section. In this section, it is assumed that the user's moving path is known.

As users move, MEC servers that can provide services to users are constantly changing. When the transmission power of the MEC server is low, the coverage area of the MEC server is small, causing frequent user wireless handoffs, and even

signal interruption may occur. If the transmission power is too large, the energy consumption of the MEC server will also increase. Moreover, when the MEC server has a large overlapping coverage area, the interference received by the MEC server will be too large, which in turn increases the service delay. Based on such fact, in this section, the service migration optimization problem is modeled as a multiobjective optimization problem, which is aimed at minimizing the number of user wireless handoffs and minimize the energy consumption of the MEC server.

The energy consumption of the MEC server can be expressed as follows:

$$\mathrm{EN}_j = \alpha * \mathrm{tp}_j. \tag{4}$$

In formula (4), EN_j is the energy consumption of MEC server e_j, which is proportional to the transmission power tp_j of e_j, and α is an adjustable parameter.

Let $C_i(t_k)$ represent the MEC server that user u_i is connected to at time t_k, and the user is always connected to the MEC server that provides the maximum received signal strength, so $C_i(t_k)$ can be expressed as follows:

$$C_i(t_k) = \left\{ j \middle| \mathrm{RSS}_{e_j}^{u_i} > \mathrm{RSS}_{e_l}^{u_i},\ \forall e_l \in E \right\}. \tag{5}$$

Next, the following variables are defined:

$$\begin{aligned} I_{ij}(t_k) &= \begin{cases} 1, & C_i(t_k) = j, \\ 0, & \text{otherwise}, \end{cases} \\ I_i(t_k) &= \begin{cases} 1, & C_i(t_k) \neq C_i(t_{k+1}), \\ 0, & \text{otherwise}. \end{cases} \end{aligned} \tag{6}$$

Since users can only connect to the same MEC server at a time, we can have

$$\sum_{e_j \in E} I_{ij}(t_k) = 1. \tag{7}$$

Let Ht_i represent the number of device handoffs during the user's movement, then Ht_i can be expressed as follows:

$$\mathrm{Ht}_i = \sum_k I_i(t_k). \tag{8}$$

Therefore, in this paper, the optimization objective can be expressed as follows:

$$\begin{aligned} \mathbf{P}: \quad & \min \quad \sum_{u_i \in U} \mathrm{Ht}_i \\ & \min \quad \sum_{e_j \in E} \mathrm{EN}_j. \end{aligned} \tag{9}$$

4. The Optimization Method

In this section, two optimization objectives are considered: minimizing the number of wireless handoffs of user equipment and minimizing the energy consumption of the MEC server. This is a multiobjective optimization problem. There are many ways to solve the multiobjective optimization problem. As in reference [34], this paper uses the weight coefficient transformation method to solve the problems raised in this section. Therefore, problem P can be transformed into problem $P1$:

$$P1: \quad \min \quad w_1 \sum_{u_i \in U} \mathrm{Ht}_i + w_2 \sum_{e_j \in E} \mathrm{EN}_j, \tag{10}$$

where $w_1 + w_2 = 1$. In this study, the PSO algorithm [34] was used to solve problem $P1$. Because of its simplicity and ease of implementation and the small number of parameters, the particle swarm algorithm is widely used in function optimization, neural network training, etc. In the PSO algorithm, the particle is a bird in the search space, the position of the particle is the solution to the optimization problem, and the speed of the particle determines the direction and distance of its flight. During each iteration, each particle updates its position according to its velocity. After multiple iterations, the optimal solution is finally obtained. Using PSO algorithm to solve the problem $P1$ mainly includes the following steps:

Step 1. Randomly initialize the particle swarm position and velocity matrix: xMatrix, vMatrix. The position and velocity of each particle are n-dimensional vectors, and the position is composed of the transmission power of n MEC servers. Assuming that the transmission power range of the MEC server is $[\mathrm{tp}_{\min}, \mathrm{tp}_{\max}]$, the position of particle i is initialized to n random numbers in $[\mathrm{tp}_{\min}, \mathrm{tp}_{\max}]$, that is, $x_i = (\mathrm{tp}_{i1}, \mathrm{tp}_{i2}, \cdots, \mathrm{tp}_{\mathrm{in}})$, and the velocity is initialized to n random numbers in $(0, 1)$, that is, $v_i = (v_{i1}, v_{i2}, \cdots, v_{\mathrm{in}})$.

Step 2. Calculate the fitness of the particles according to formulas (4)–(8), and get the fitness matrix fitMatrix

Step 3. Update the historical optimal position of each particle. In the first iteration, the historical optimal position of each particle is its random initial position $p_i = x_i$

Step 4. Update the global optimal position of the group. Initially, the global optimal position of the swarm is the particle position with the smallest fitness in the particle swarm position matrix xMatrix $p_g = \min_{\mathrm{fit}}(x\mathrm{Matrix})$

Step 5. Update the speed and position of each particle:

$$v_i^{t+1} = w \cdot v_i^t + c_1 \cdot r_1 \cdot \left(p_i^t - x_i^t\right) + c_2 \cdot r_2 \cdot \left(p_g^t - x_i^t\right), \tag{11}$$

where w is the inertia weight, c_1 is the local learning factor, c_2 is the global learning factor, r_1, r_2 are random numbers in $[0, 1]$.

Input: base station location, user moving path
Output: number of wireless handoffs, energy consumption
Begin
1: //initialization
2: **for each** particle i **do**
3: Randomly select n numbers from the interval $[tp_{min}, tp_{max}]$ as the position x_i of i
4: Randomly select n numbers from the interval $(0, 1)$ as the speed v_i of i
5: Apply formulas (4)–(8) to evaluate particle i
6: The historical optimal position of particle $ip_i = x_i$
7: **end for**
8: Global optimal position $p_g = \min_{fit}\{p_i\}$
9: **while** *iterations* < *IteratorNum* **do**:
10: **for** $i = 1$ **to** N **do**
11: Apply formula (11) to update velocity of particle i
12: Apply formula (12) to update position of particle i
13: // Update p_i and p_g
14: **if** $\text{fit}(x_i) < \text{fit}(p_i)$ **do**
15: $p_i = x_i$
16: **if** $\text{fit}(p_i) < \text{fit}(p_g)$ **do**
17: $p_g = p_i$
18: **end for**
19: **end while**
End

Algorithm 1: PSO algorithm to solve problem P1.

The position of the particle is updated to:

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{12}$$

Step 6. Repeat (Steps 2–5) until the end condition is met.

The algorithm is described as follows, where *iterations* represents the current number of iterations, *IteratorNum* represents the number of iterations, and N is the size of the particle swarm:

5. Experimental Results and Analysis

The experimental environment is PyCharm Community Edition, the programming language is Python 3.5, and the configuration of the experimental machine is 16 G memory, core i7-4790 3.60 GHz processor, Windows 7, 64-bit operating system. The Telecom Dataset is provided by Shanghai Telecom [8–10]. The data set has six parameters including month, day, start time, end time, base station location, and user ID. The data set contains a total of 7.2 million records, which are records of 9,481 mobile phones accessing the Internet through 3,233 base stations. In this experiment, 10 stations of three subways in Shanghai were selected as the user's moving routes, the latitude and longitude of these 30 sites were obtained on Baidu Maps, and 8 base stations were selected near the 30 sites to provide services to users. The experiment parameters are shown in Table 2:

Table 2: Experimental parameters.

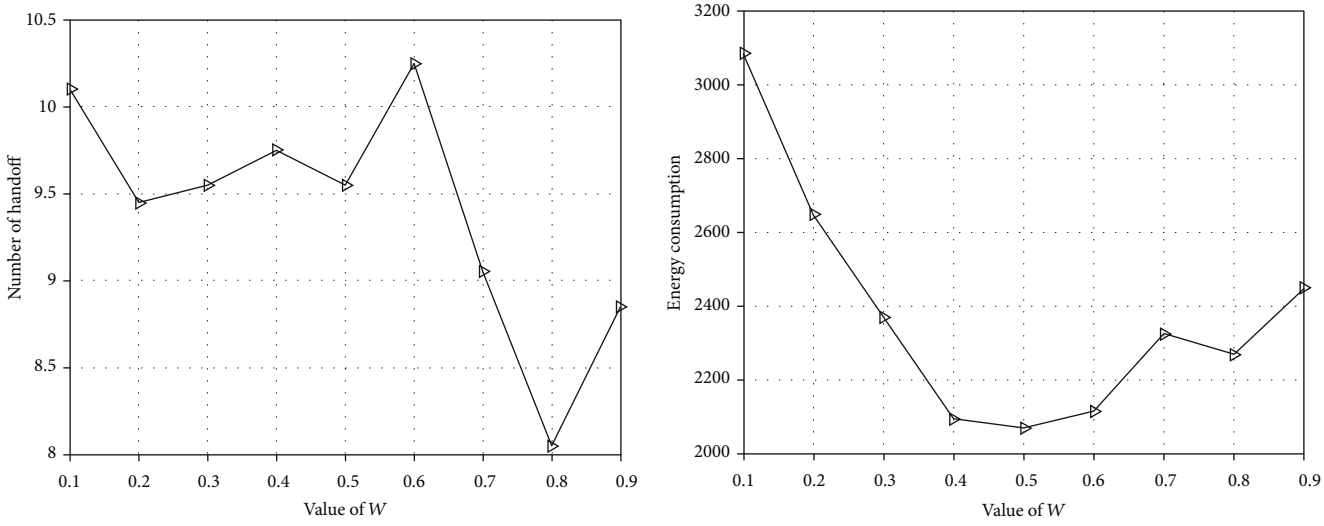
Parameters	Values
L_1	20 dBm
G_{ui}	8.35 dBi
G_{ej}	24.5 dBi
n_1	2
n_2	4
d_b	100 m
PSO number of particles	10
PSO number of iterations	100

Table 3: PSO parameters.

Parameter	w	c_1	c_2
Experiment 1	0.1 to 1	2	2
Experiment 2	0.5	0.5 to 4	2
Experiment 3	0.5	2	0.5 to 4

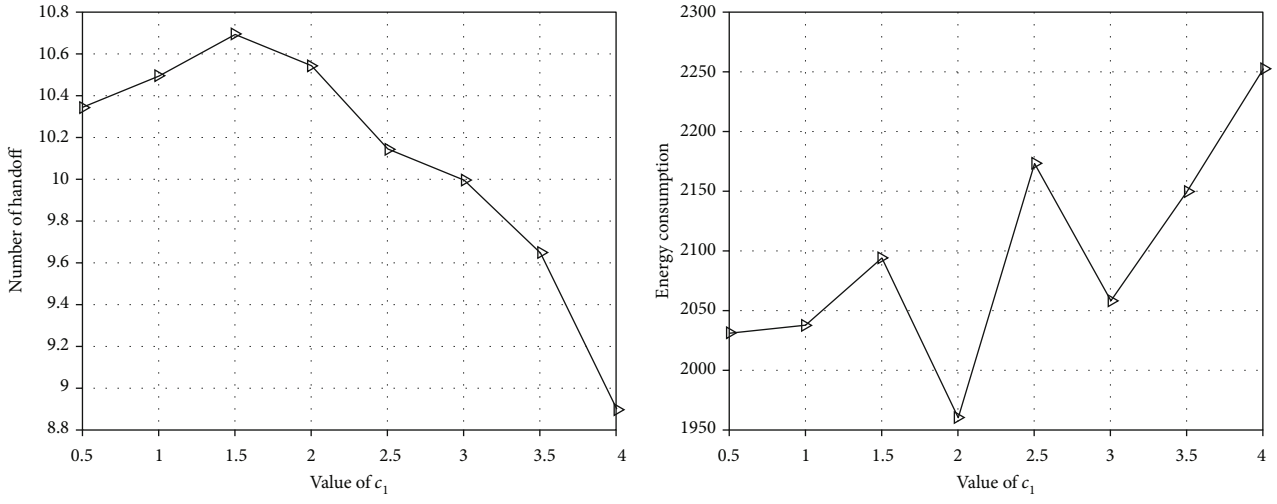
5.1. Experiment for PSO Parameter Selection. First, the parameters in the PSO algorithm were verified, including the impact of inertia weight w, local learning factor c_1, global learning factor c_2 on the number of handoffs, and energy consumption. The particle swarm size was set to 10, the number of iterations was set to 100, the particle swarm algorithm was executed 20 times, and the average value was taken as the experimental result. The experimental parameter settings are shown in Table 3.

The essence of PSO is to apply formula (11) to update the velocity of particles in each iteration. Formula (11) is



(a) Handoff vs. w (b) Energy consumption vs. w

Figure 4: The influence of parameter w on experimental results.



(a) Handoff vs. c_1 (b) Energy consumption vs. c_1

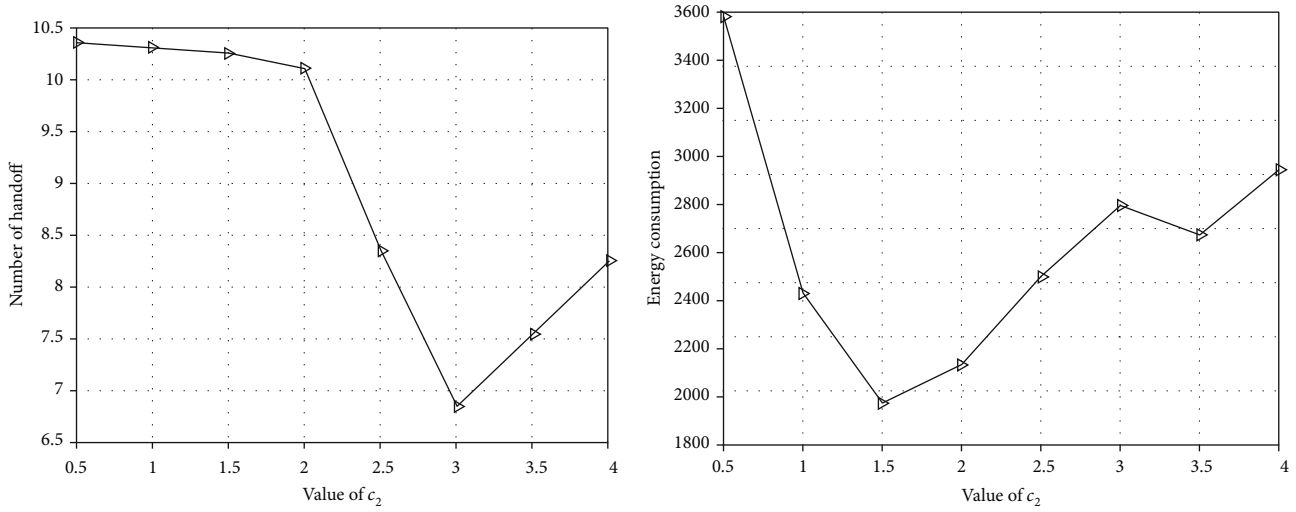
Figure 5: The influence of parameter c_1 on experimental results.

composed of three items: memory item, self-cognition item, and group recognition item. The inertia weight w determines the influence of the speed in the previous iteration on this iteration. To study the influence of the inertia weight w on the experiment, the values of other parameters are fixed: $c_1 = c_2 = 2$, w changes from 0.1 to 0.9, and the step size is 0.1. The experimental results are shown in Figure 4.

The figure above shows that, as the value of w increases, the trend of the number of handoffs is slowly increasing and then decreasing sharply. The minimum value is obtained when $w = 0.8$, and the minimum value is close to 8. The energy consumption drops first and then rises. When the energy consumption is small, the power of the MEC server is small and the coverage area is small, so the number of handoffs is relatively high. Therefore, when $w = 0.5$, the energy consumption achieved the minimum value, and the number of handoffs is 9.55, which is at a relatively high level. In this study, 20 repeated experiments were performed and the average value was taken as the result, so the number of handoffs may be a decimal.

The local learning factor c_1 and the global learning factor c_2 are the weights of self-cognition items and group recognition items. The purpose of the self-cognition item is that the speed of particles is affected by one's own experience. The group cognition item reflects the influence of knowledge sharing between particles on finding the optimal solution. To study the influence of these two parameters on the experiment, $w = 0.5$ was set and c_1, c_2 were set to range from 0.5 to 4, and the step length was 0.5. The experimental results are shown in Figures 5 and 6.

In Figure 5, the number of user handoffs decreases as the value of c_1 increases, while the energy consumption shows a fluctuating upward trend. The influence of the global learning factor on the number of user handoffs and the energy



(a) Handoff vs. c_2

(b) Energy consumption vs. c_2

FIGURE 6: The influence of parameter c_2 on the experimental results.

consumption of the MEC server is shown in Figure 5. The number of user handoffs and the MEC server energy consumption have a similar trend, and both decrease first and then increase. When $c_2 = 3$, the number of user handoffs achieves the extreme value, and when $c_2 = 1.5$, the energy consumption achieves the optimal value.

In Figure 6, the experimental results show that the two evaluation indicators of number of handoffs and energy consumption cannot reach the optimal value at the same time. In order to balance these two evaluation indicators, the parameters of PSO were set as $w = 0.8$, $c_1 = 3$, and $c_2 = 1$.

5.2. Comparative Analysis of Experiment Results. This experiment was compared with the following two algorithms:

(1) Genetic algorithm (GA) [24]: the genetic algorithm searches for the optimal solution of the problem by simulating the natural evolution process. The main steps are as follows: (1) initialize the population, (2) assess the individual fitness value, (3) select, (4) crossover, (5) mutate, and (6) repeat (2)–(5) until the end conditions are met

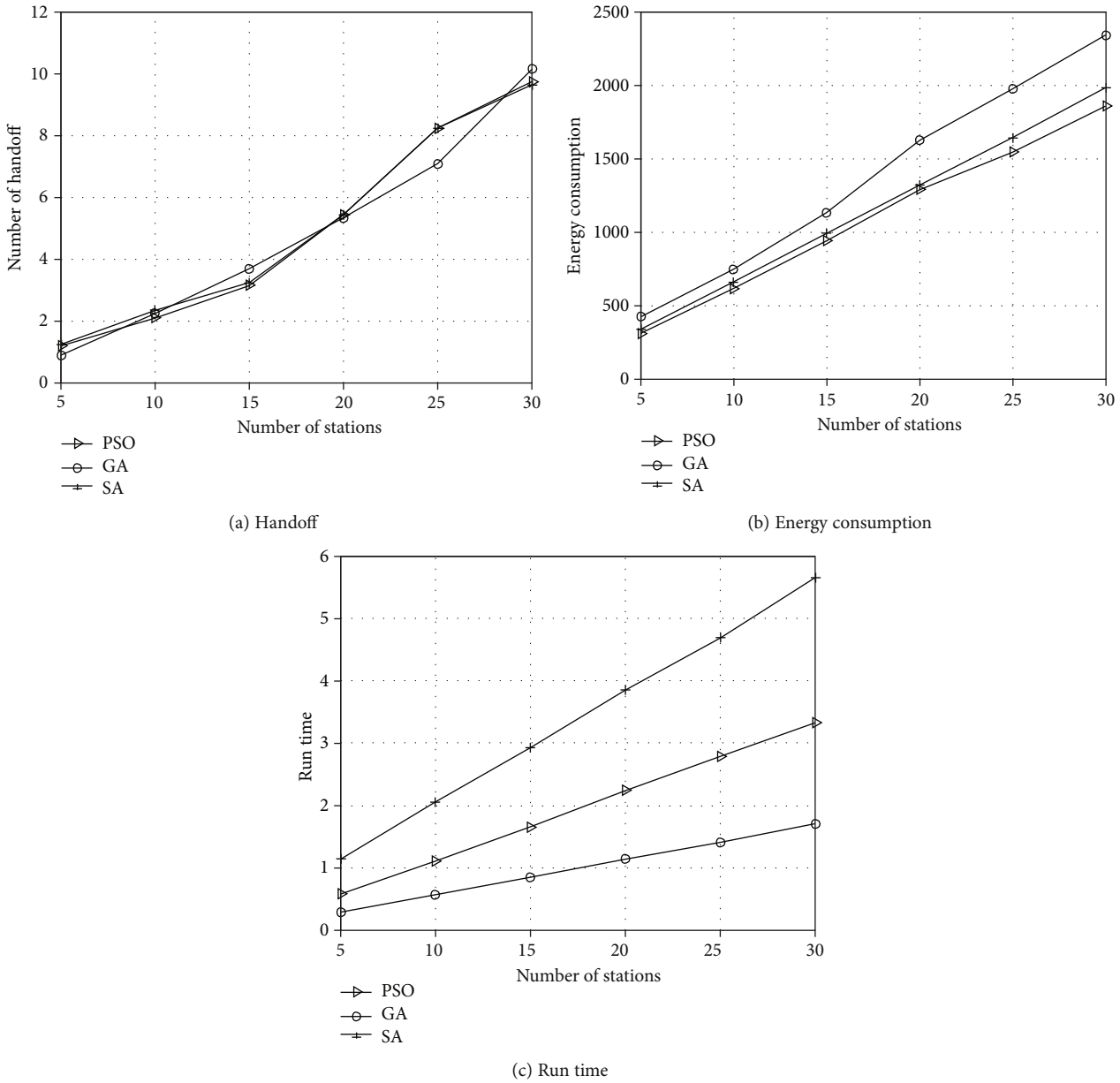
(2) Simulated annealing algorithm (SA) [25]: the principle of solid annealing is the theoretical basis of the simulated annealing algorithm. A solid is heated to a sufficiently high temperature, and then slowly cooled. During cooling, the particles are gradually ordered, and finally the internal energy is the smallest at room temperature. The basic steps of the simulated annealing algorithm are as follows: (1) initialize the solution T_{old}; (2) generate a new solution T_{new}; (3) apply the evaluation function to evaluate T_{old} and T_{new}; (4) if T_{new} is better than T_{old}, replace T_{old} with T_{new}, otherwise, accept T_{new} with a certain probability; (5) repeat (2)–(4) until the end condition is met. The parameters of the simulated annealing algorithm were set as number of iterations = 100, initial temperature = 100, attenuation factor = 0.85, and probability of accepting the difference = $\exp(-\Delta t/t)$, where Δt is the difference between T_{new} and T_{old}, and t is the current temperature

Three subway routes in Shanghai were selected as the user's movement trajectory. For each subway line, 10 stations were selected, and 8 base stations were selected around these 30 stations to provide services to users. Then, the influence of the number of sites on the number of handoffs and energy consumption was studied. Figure 7 is the result of the comparison of the three algorithms.

Figure 6(a) shows that, for the three algorithms, the difference in the number of handoffs is small and the results and trends of PSO and SA are very similar. When the number of stations is 20 and 25, the number of handoffs calculated by the two are equal. When the number of stations is less than 20, PSO is slightly better than SA. GA has the worst performance in the number of handoffs, and only when the number of sites is 25 is it better than SA and PSO. The performance of the three algorithms in terms of energy consumption is shown in Figure 6(b). The energy consumption increases as the number of sites increases. Obviously, the results calculated by PSO are optimal in terms of energy consumption. Finally, these three algorithms were also measured from the running time. Although GA performs poorly in the number of handoffs and energy consumption, its execution time cost is the lowest. Although PSO has a slight advantage over SA in the number of handoffs and energy consumption, its execution time cost is much lower than SA. Therefore, PSO can effectively solve this multiobjective optimization problem.

6. Conclusion

Deploying services on edge nodes can bring computing and storage resources closer to users, thereby reducing service delays and improving user experience. However, the dynamic feature of user equipment in mobile edge computing has led to



(a) Handoff

(b) Energy consumption

(c) Run time

Figure 7: The influence of the number of stations on the experimental results.

the need to migrate services between different MEC servers. Since the edge nodes communicate through the backhaul link, the service forwarding between the edge nodes will cause a high delay, which leads to a significant reduction in the quality of user experience. Therefore, reducing service migration is a very important task.

The study considers the use of transmission power control technology to reduce the number of service migrations during user movement. Under the assumption that the user's moving route is known, the edge node of the user's wireless connection is controlled through the transmission power of the MEC server, so the number of wireless handoffs during the user's movement is reduced. Meanwhile, in order to avoid energy waste caused by excessive transmission power, minimizing the energy consumption of the MEC server is also regarded as an optimization objective. The multiobjective optimization problem is transformed into a single-objective problem through the weight coefficient conversion method, and then the PSO algorithm is used to solve the problem. The experimental results show the effectiveness of the PSO algorithm in this multiobjective optimization problem.

Data Availability

All of the data used in this study are already available on the Internet and is easily accessible.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Talent Project of Anhui Province (Grant No. gxgnfx2019034), the Anhui Province Key Research and Development Program Project (Grant No. 201904a05020091) and its supporting projects (Grant No. PT04a05020094), the Excellent Talents Support Program of Colleges and Universities (Grant No. gxyq2020083), the key project of the Natural Science Research of Higher Education Institutions in Anhui Province (Grant No. KJ2020A0680), the 2018 Higher Education Research Project of Anhui Province (Grant No. 2018JYXM0334), the 2020 Quality Improvement Project of Chaohu College on Discipline Construction (Grant No. kj20xqyx03), and a scientific research project commissioned by the Hefei Saile Education Technology Co., Ltd. (Grant No. hxkt20210002).

References

[1] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "Thinkair: dynamic resource allocation and parallel execution in the cloud for mobile code offloading," in *2012 Proceedings IEEE INFOCOM*, pp. 945–953, Orlando, FL, USA, 2012.

[2] S. Wang, M. Zafer, and K. K. Leung, "Online placement of multi-component applications in edge computing environments," *IEEE Access*, vol. 5, pp. 2514–2533, 2017.

[3] B. G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "Clonecloud: elastic execution between mobile device and cloud," in *Proceedings of the sixth conference on Computer systems*, pp. 301–314, New York, 2011.

[4] S. Wang, A. Zhou, R. Bao, W. Chou, and S. S. Yau, "Towards green service composition approach in the cloud," *IEEE Transactions on Services Computing*, pp. 1–1, 2018.

[5] 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.

[6] M. R. Bonyadi and Z. Michalewicz, "Particle swarm optimization for single objective continuous space problems: a review," *Evolutionary Computation*, vol. 25, no. 1, pp. 1–54, 2017.

[7] Z. Becvar, J. Plachy, and P. Mach, "Path selection using handover in mobile networks with cloud-enabled small cells," in *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, pp. 1480–1485, Washington, DC, 2014.

[8] S. Wang, Y. Zhao, L. Huang, J. Xu, and C. H. Hsu, "QoS prediction for service recommendations in mobile edge computing," *Journal of Parallel and Distributed Computing*, vol. 127, pp. 134–144, 2019.

[9] Y. Guo, S. Wang, A. Zhou, J. Xu, J. Yuan, and C.-. H. Hsu, "User allocation-aware edge cloud placement in mobile edge computing," *Software: Practice and Experience*, vol. 50, no. 5, pp. 489–502, 2020.

[10] J. Xu, S. Wang, B. K. Bhargava, and F. Yang, "A blockchain-enabled trustless crowd-intelligence ecosystem on mobile edge computing," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3538–3547, 2019.

[11] T. Taleb and A. Ksentini, "An analytical model for follow me cloud," in *2013 IEEE Global Communications Conference (GLOBECOM)*, pp. 1291–1296, Atlanta, GA, USA, 2013.

[12] A. Ksentini, T. Taleb, and M. A. Chen, "Markov decision process-based service migration procedure for follow me cloud," in *2014 IEEE International Conference on Communications (ICC)*, pp. 1350–1354, Sydney, NSW, Australia, 2014.

[13] S. Wang, R. Urgaonkar, T. He, M. Zafer, K. Chan, and K. K. Leung, "Mobility-induced service migration in mobile micro-clouds," in *2014 IEEE Military Communications Conference*, pp. 835–840, Baltimore, MD, USA, 2014.

[14] C. Chen, Z. Liu, S. Wan, J. Luan, and Q. Pei, "Traffic flow prediction based on deep learning in internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2020.

[15] S. Wang, R. Urgaonkar, M. Zafer, T. He, K. Chan, and K. K. Leung, "Dynamic service migration in mobile edge-clouds," in *2015 IFIP Networking Conference (IFIP Networking)*, pp. 1–9, Toulouse, France, 2015.

[16] A. Machen, S. Wang, K. K. Leung, B. J. Ko, and T. Salonidis, "Live service migration in mobile edge clouds," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 140–147, 2017.

[17] F. Ud Din, A. Ahmad, H. Ullah, A. Khan, T. Umer, and S. Wan, "Efficient sizing and placement of distributed generators in cyber-physical power systems," *Journal of Systems Architecture*, vol. 97, pp. 197–207, 2019.

[18] L. Ma, S. Yi, and Q. Li, "Efficient service migration across edge servers via docker container migration," in *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, pp. 1–13, New York, NY, USA, 2017.

[19] T. Ouyang, Z. Zhou, and X. Chen, "Follow me at the edge: mobility-aware dynamic service placement for mobile edge computing," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 10, pp. 2333–2345, 2018.

[20] Y. Yang, Z. Zheng, X. Niu, M. Tang, Y. Lu, and X. Liao, "A location-based factorization machine model for web service QoS prediction," *IEEE Transactions on Services Computing*, pp. 1–1, 2019.

[21] S. K. Bose, S. Brock, R. Skeoch, and S. Rao, "CloudSpider: combining replication with scheduling for optimizing live migration of virtual machines across wide area networks," in *2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pp. 13–22, Newport Beach, CA, USA, 2011.

[22] I. Farris, T. Taleb, M. Bagaa, and H. Flick, "Optimizing service replication for mobile delay-sensitive applications in 5G edge network," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, Paris, France, 2017.

[23] Y. Zhang, G. Cui, S. Deng, F. Chen, Y. Wang, and Q. He, "Efficient query of quality correlation for service composition," *IEEE Transactions on Services Computing*, 2018.

[24] P. A. Frangoudis and A. Ksentini, "Service migration versus service replication in multi-access edge computing," in *2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pp. 124–129, Limassol, Cyprus, 2018.

[25] P. Mach and Z. Becvar, "Cloud-aware power control for real-time application offloading in mobile edge computing," *Transactions on Emerging Telecommunications Technologies*, vol. 27, no. 5, pp. 648–661, 2016.

[26] T. G. Rodrigues, K. Suto, H. Nishiyama, and N. Kato, "Hybrid method for minimizing service delay in edge cloud computing through VM migration and transmission power control," *IEEE Transactions on Computers*, vol. 66, no. 5, pp. 810–819, 2017.

[27] T. G. Rodrigues, K. Suto, H. Nishiyama, N. Kato, and K. Temma, "Cloudlets activation scheme for scalable mobile

edge computing with transmission power control and virtual machine migration," *IEEE Transactions on Computers*, vol. 67, no. 9, pp. 1287–1300, 2018.

[28] S. Wan, A. Gu, and Q. Ni, "Cognitive computing and wireless communications on the edge for healthcare service robots," *Computer Communications*, vol. 149, pp. 99–106, 2020.

[29] Y. Zhang, C. Yin, Q. Wu, Q. He, and H. Zhu, "Location-aware deep collaborative filtering for service recommendation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–12, 2020.

[30] R. Fernandes and R. D'Souza, "A new approach to predict user mobility using semantic analysis and machine learning," *Journal of medical systems*, vol. 41, no. 12, 2017.

[31] S. Roy, R. Bose, and D. Sarddar, "Fuzzy based dynamic load balancing scheme for efficient edge server selection in cloud-oriented content delivery network using Voronoi diagram," in *2015 IEEE International Advance Computing Conference (IACC)*, pp. 828–833, Banglore, India, 2015.

[32] W. Su, S. J. Lee, and M. Gerla, "Mobility prediction in wireless networks," in *MILCOM 2000 Proceedings. 21st Century Military Communications. Architectures and Technologies for Information Superiority (Cat. No.00CH37155)*, vol. 1, pp. 491–495, Los Angeles, CA, USA, 2000.

[33] F. Calabrese, G. Di Lorenzo, and C. Ratti, "Human mobility prediction based on individual and collective geographical preferences," in *13th international IEEE conference on intelligent transportation systems*, pp. 312–317, Funchal, Portugal, 2010.

[34] C. Chen, B. Liu, S. Wan, P. Qiao, and Q. Pei, "An edge traffic flow detection scheme based on deep learning in an intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2020.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6670591, 16 pages
https://doi.org/10.1155/2021/6670591

Research Article

Edge Computing- and H_∞-Switching-Based Networked Control for Frequency Control in Multi-Microgrids with Time Delays

Peng Yang[1], **Wei Guo**[2], **Guanghua Wu**[2], **Cong Wang**[3], **Kai Zhang**[1] and **Ran Zhang**[1]

[1]*State Grid Hebei Electric Power Co., LTD., Shijiazhuang 050021, China*
[2]*Market Service Center, State Grid Hebei Electric Power Co., LTD., Shijiazhuang 050021, China*
[3]*Shijiazhuang Power Supply Company, State Grid Hebei Electric Power Co., LTD., Shijiazhuang 050021, China*

Correspondence should be addressed to Wei Guo; yxzx_guow@he.sgcc.com.cn

Received 14 October 2020; Revised 26 November 2020; Accepted 10 December 2020; Published 6 January 2021

Academic Editor: Mohammad R. Khosravi

The frequency stability of multi-microgrids is easily affected by random load fluctuations and intermittent renewable resources. Additionally, geographically distributed generation equipment usually cannot adopt the "point-to-point" dedicated communication scheme to realize the information exchange considering the construction and computation costs. Therefore, a H_∞-switching frequency control strategy for multi-microgrids based on edge computing framework is proposed in this paper. Firstly, an edge computing device is set up in each microgrid to collect the operation statuses of local participating equipment and generate the control instructions to ensure the real-time local frequency stability. Secondly, the multihop data transmission process in edge computing environment is described as a cascade queuing model. Then, the frequency control system in each microgrid is described as a switching model dependent on the varying time delays. Finally, via constructing a Lyapunov function, the constraints of the controller gains ensuring the H_∞-damping performance for external load demands and the renewable outputs are derived at the same time. Simulation results show that compared with the traditional centralized control schemes, the peak value of our proposed edge computing framework is reduced by 32.51% compared with the traditional centralized control scheme. Moreover, under the same edge computing framework, the integral of absolute error (IAE) of frequency with the proposed H_∞ control strategy can be reduced by 37.19% at least. Therefore, a better transient performance can be obtained with our proposed method.

1. Introduction

The microgrid is usually used to supply power for the rural areas, islands, and so on. Integrating multi-microgrids into the power systems can effectively address the contradiction between the load growth and the power infrastructure expansion [1]. The main feature of the multi-microgrids is that each microgrid integrates the high-proportional renewable generators and exchanges active power with the neighbouring microgrids through tie lines [2, 3]. However, the uncertainties in renewable outputs and load demands usually cause the active power mismatch between the supply and demand in the power systems. Correspondingly, the frequency in each microgrid will deviate from the rated value [4]. To maintain the real-time balance of active power supply and demand at the rated frequency point, the global operation statuses of each microgrid are usually collected in traditional centralized control strategies while the output adjustment instructions are sent out to the participating equipment including synchronous generators and energy storage systems in each microgrid [5, 6]. However, because there is only one centralized control center in the whole power system, the computing burden in the control center is usually heavy. Besides, if the number of interconnected microgrids is too large, there will be significant disadvantages of the centralized control strategies in reliability and construction costs of

communication facilities [7]. In addition, too many interconnected microgrids will cause the dimensional disaster in traditional centralized frequency control schemes, which may lead to no solution for the controller gains [8].

With the rapid development of Internet of things and edge computing technologies, using decentralized control structure and sharing communication networks have attracted considerable attentions. Different from the traditional centralized control scheme, in edge computing framework, the edge computing device with data storing, processing, and analysing capabilities is set up in each microgrid which is considered as the local control center to realize the local power balance at the rated frequency point [9, 10]. Since the edge computing device is closer to the participating equipment, the corresponding transmission delay can be shorten compared with the traditional centralized control schemes. However, it still should be noted that the influences of stochastic time delays and packet losses cannot be ignored in the controller design process. The numerical results in [11, 12] show that even the millisecond-level time delay may cause the frequency deviation exceeding the allowable ranges or even instability risk. Hence, the literatures concerning the delay/packet loss-dependent frequency control in multi-microgrids can be divided into the following two categories.

1.1. Delay Margin Calculation. In this context, the controller parameters for frequency stabilization are firstly determined. Then, the maximum allowable time delay that guarantees frequency deviation within the permitted ranges is calculated. For example, the analytical relationship between delay margin and controller gain is derived by constructing a Lyapunov function in [13]. Similarly, in [14], a stability criterion of frequency control system concerning the controller parameters and delay margin is proposed on the basis of the regular polynomial method. In [15], an event-triggered communication mechanism is proposed for the frequency control in power systems. Only when the frequency deviation exceeds the preset threshold, the update of control instructions can be triggered. The authors in [15] also strictly proved the analytical relationship among controller parameters, delay margin, and the triggering threshold. However, the disadvantage of above studies is that the controller parameters must be given in advance. Therefore, the optimal dynamic performance of frequency control systems cannot be guaranteed when the power system suffers from the varying time delays. In addition, when the actual transmission delays of the packets exceed the allowable delay margin, the frequency control system will be instable.

1.2. Delay/Packet Loss-Dependent Controller Design. In this context, the controller parameters are obtained according to the actual transmission delays. Therefore, the control performance of the frequency control system can be improved effectively. For example, in [16], queuing theory is adopted to calculate the delay ranges in power systems. Besides, a decentralized control strategy based on linear matrix inequality-linear quadratic regulator (LMI-LQR) is proposed to guarantee the closed-loop asymptotic stability in the maximum delay case. The effects of packet losses on the control performance have not been discussed in [16]. In [17], a robust model predictive control (MPC) method is proposed to tackle the frequency stability problem in power systems with stochastic time delays. However, the MPC method needs to store continuous historical data to generate the control instructions. Moreover, the instruction calculation in MPC also introduces additional delays. In [18], a decentralized robust sliding control strategy is proposed to damping the frequency deviations in power systems with time delays. However, the external active power distributions from load demands and renewable outputs are assumed to be known in advance during the sliding surface construction process. In fact, it is the uncertainties of load demand and renewable outputs that lead to the frequency deviation from the rated value, so the practicability of the method in [18] is indeed open to debate. Furthermore, the current studies mainly pay close attention to the closed-loop asymptotic stability in the maximum transmission delay case. The dynamic performance during frequency restoration process is sacrificed to some extent.

Based on the above analysis, a H_∞-switching control strategy for frequency stability in multi-microgrids based on edge computing framework is proposed in this paper. The main contributions are given as follows:

(1) An edge computing framework for frequency stability in multi-microgrids is proposed. Rather than the traditional centralized control schemes which require global operation statuses of all the microgrids, the edge computing device is set up in each microgrid to be responsible for maintaining the local power balance between supply and demand sides at the rated frequency point. Correspondingly, the lighter communication burden and lower computation cost can be realized

(2) Based on the queuing theory, the analytical relationship between the transmission delay and the network parameters such as packet size, transmission rate, and transmission hops in the process of multihop data transmission process under the edge computing framework is calculated. Then, the dynamics of frequency control system in each microgrid are further described as a switching model which depends on the time-varying delay. Hence, the mapping relationship between time-varying delay and dynamic frequency deviation response can be revealed more clearly

(3) By constructing a Lyapunov function, a stability criterion of the closed-loop delay-dependent frequency control system with H_∞-damping performance for external power disturbances is strictly derived. Furthermore, by aiming at minimizing the integral of absolute error (IAE) of the frequency in each microgrid, a constrained controller optimization algorithm is proposed. Then, the dynamic performance during frequency restoration process can be improved

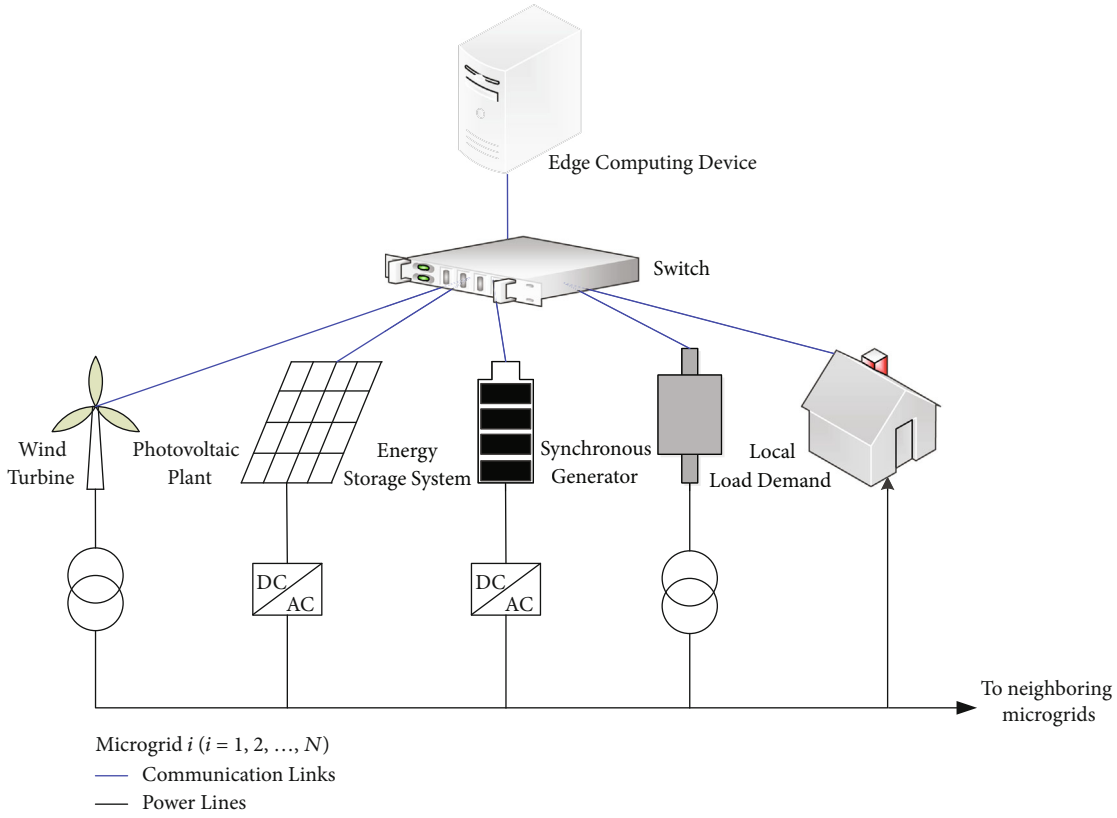


Figure 1: Framework of multi-microgrids with edge computing framework.

The rest is organized as follows. State-space model for frequency control system in each microgrid is established in Section 2. In Section 3, the relationships among transmission delay and key network parameters are investigated. The dynamics of closed-loop frequency control system with time-varying delays are discussed in Section 4. In Section 5, an optimization method combined with iterative linear matrix inequality and heuristic algorithm is proposed. Simulations are discussed in Section 6. Finally, conclusions are given in Section 7.

2. State-Space Model of Frequency Control in Multi-Microgrids Based on Edge Computing Framework

The power resources in multi-microgrids mainly contain synchronous generators, renewable generators (such as wind turbines and photovoltaic cells), and energy storage systems. Due to the uncertain fluctuations of renewable outputs and load demands, the frequency may deviate from the rated value. Hence, it is required to timely adjust the outputs of synchronous generators and energy storage systems to maintain real-time power balance between supply and demand at the rated frequency point [19, 20]. Considering that there exist the heavy communication and computation burdens in the traditional centralized control scheme which requires global operation statuses of all the microgrids, in this paper, a frequency control strategy based on edge computing framework is proposed. As shown in Figure 1, the local operating statuses are transmitted to the edge computing device installed in each microgrid. In other words, in our proposed edge computing framework, an edge computing device considered as the local control center is set up in each microgrid. Since the local controller is closer to the participating equipment and the corresponding control structure is simpler, lighter communication burden and lower computation cost can be realized. Similar to the centralized scheme, the function of the edge computing device is sending the active power output commands of synchronous generators and energy storage systems and then restoring the frequency to the rated value. Note that the edge computing devices and devices participating in frequency control are geographically dispersed; the uploading of operation statuses usually relays on the sharing communication network. Hence, there exist time delay and packet loss problems during the data transmission process. In this section, the state-space model of frequency control system in each microgrid is firstly discussed. The effects of time delay and packet loss on the dynamics of frequency control system will be analysed in the next sections.

Let the number of microgrids in a multi-microgrid be N. For the i-th ($i = 1, 2, \cdots, N$) microgrid, the equivalent block diagram of the frequency control system is shown in Figure 2. Moreover, let the number of synchronous generators and energy storage systems in i-th microgrid be M_{SG_i} and M_{ESS_i}, respectively. Therefore, the dynamics of the frequency control system in i-th microgrid satisfy [4]

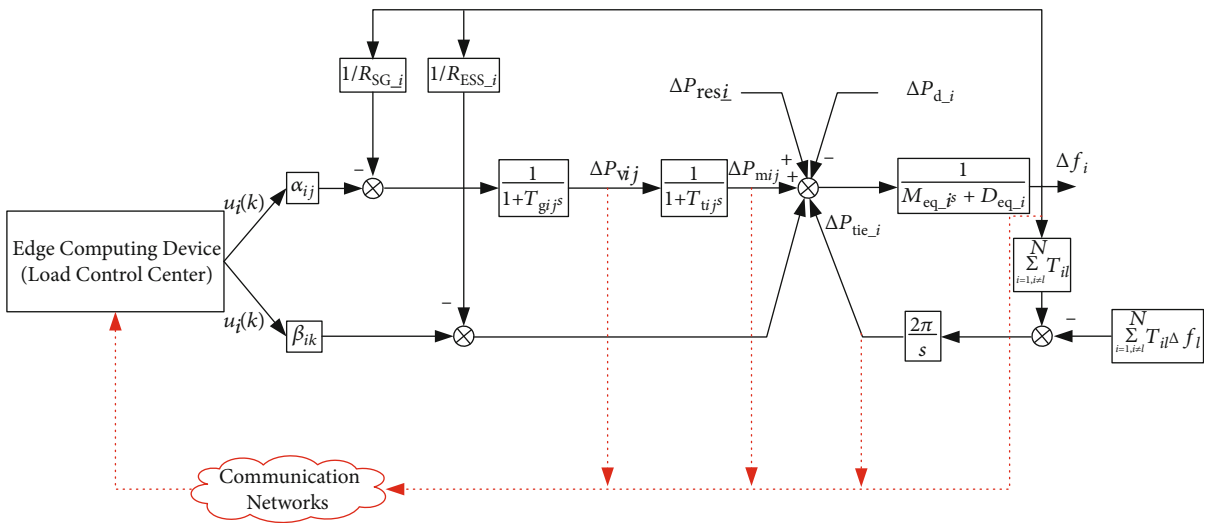


Figure 2: Control block of networked frequency control system.

$$\begin{cases} \dfrac{d}{dt}\Delta P_{vij} = -\dfrac{1}{T_{gij}}\Delta P_{vij} + \dfrac{1}{T_{gij}}\left(\alpha_{ij}u_i - \dfrac{1}{R_{ij}}\Delta f_i\right), \\ \dfrac{d}{dt}\Delta P_{mij} = -\dfrac{1}{T_{tij}}\Delta P_{mij} + \dfrac{1}{T_{tij}}\Delta P_{vij}, \\ \dfrac{d}{dt}\Delta P_{tie\_i} = 2\pi\sum\limits_{i=1,i\neq l}^{N} T_{il}(\Delta f_i - \Delta f_l), \\ \dfrac{d}{dt}\Delta f_i = -\dfrac{D_{eq\_i}}{M_{eq\_i}}\Delta f_i + \dfrac{1}{M_{eq\_i}}\left(\sum\limits_{j=1}^{M_{SG\_i}}\Delta P_{mij} + \Delta P_{res\_i} + \sum\limits_{k=1}^{M_{ESS\_i}}\Delta P_{ESS\_ik} - \Delta P_{di} - \Delta P_{tie\_i}\right), \\ \Delta P_{ESS\_ik} = \beta_{ik}u_i - \dfrac{1}{R_{ESS\_ik}}\Delta f_i, \end{cases} \tag{1}$$

where ΔP_{vij} is the valve opening of the j-th synchronous generator; ΔP_{mij} is the mechanical output power of the j-th synchronous generator; ΔP_{tie_i} is the tie line power fluctuation in i-th microgrid; Δf_i and Δf_l are the frequency deviations in the i-th and l-th microgrids, respectively; T_{il} is the synchronization coefficient of the tie line between the i-th and l-th microgrids; T_{gij} and T_{tij} are the time constants of governor and turbine in the j-th synchronous generator; R_{ij} and R_{ESS_ik} are the drooping coefficients of the j-th synchronous generator and k-th energy storage system, respectively; ΔP_{ESS_ik} is the active output power of the k-th energy storage system; ΔP_{di} and ΔP_{res_i} are the fluctuations of load demand and renewable output in the i-th microgrids, respectively; M_{eq_i} and D_{eq_i} are the equivalent inertia and damping coefficient of the i-th microgrids, respectively; α_{ij} and β_{ik} are the participation factors of the j-th synchronous generator and k-th energy storage system, respectively, and satisfy

$$\sum_{j=1}^{M_{SG\_i}}\alpha_{ij} + \sum_{k=1}^{M_{ESS\_i}}\beta_{ik} = 1. \tag{2}$$

By defining the state vector as $X_i = [\Delta f_i, \Delta P_{tie\_i}, \Delta P_{mi1}, \cdots, \Delta P_{miM_{SG\_i}}, \Delta P_{vi1}, \cdots, \Delta P_{viM_{SG\_i}}]^T$, the state-space model of the frequency control system in the i-th microgrid is given by

$$\begin{cases} \dfrac{d}{dt}X_i(t) = A_iX_i(t) + B_iu_i(t) + H_iw_i(t), \\ Y_i(t) = C_iX_i(t), \end{cases} \tag{3}$$

where

$$A_i = \begin{bmatrix} -\frac{D_{\text{eq\_}i}}{M_{\text{eq\_}i}} - \frac{1}{M_{\text{eq\_}i}} \sum_{k=1}^{M_{\text{ESS\_}i}} \frac{1}{R_{\text{ESS\_}ik}} & -\frac{1}{M_{\text{eq\_}i}} & \frac{1}{M_{\text{eq\_}i}} & \cdots & \frac{1}{M_{\text{eq\_}i}} & 0 & \cdots & 0 \\ 2\pi \sum_{i=1, i\neq l}^{N} T_{il} \Delta f_i & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & -\frac{1}{T_{ti1}} & \cdots & 0 & \frac{1}{T_{ti1}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{1}{T_{tiM_{\text{SG\_}i}}} & 0 & \cdots & \frac{1}{T_{tiM_{\text{SG\_}i}}} \\ \frac{1}{T_{gi1} R_{i1}} & 0 & 0 & \cdots & 0 & -\frac{1}{T_{gi1}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{T_{giM_{\text{SG\_}i}} R_{iM_{\text{SG\_}i}}} & 0 & 0 & \cdots & 0 & 0 & \cdots & -\frac{1}{T_{giM_{\text{SG\_}i}}} \end{bmatrix},$$

$$B_i = \begin{bmatrix} \frac{\sum_{k=1}^{M_{\text{ESS\_}i}} \beta_{ik}}{M_{\text{eq\_}i}} \\ 0 \\ 0 \\ \vdots \\ 0 \\ \frac{\alpha_{i1}}{T_{gi1} R_{i1}} \\ \vdots \\ \frac{\alpha_{iM_{\text{SG\_}i}}}{T_{giM_{\text{SG\_}i}}} \end{bmatrix}, H_i = \begin{bmatrix} -\frac{1}{M_{\text{eq\_}i}} & 0 \\ 0 & -2\pi \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}, C_i = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^{\text{T}}, w_i = \begin{bmatrix} \Delta P_{\text{d}i} - \Delta P_{\text{res\_}i} \\ \sum_{i=1, i\neq l}^{N} T_{il} \Delta f_l \end{bmatrix}. \tag{4}$$

Moreover, considering that the frequency control system is a sampling control system under the edge computing framework essentially, the respective discrete-time model with the sampling time T_{s} is given by

$$\begin{cases} X_i(k+1) = E_i X_i(k) + F_i u_i(k) + G_i w_i(k), \\ Y_i(k) = C_i X_i(k), \end{cases} \tag{5}$$

where $E_i = e^{A_i T_{\text{s}}}$, $F_i = \int_0^{T_{\text{s}}} e^{A_i t} dt B_i$, and $G_i = \int_0^{T_{\text{s}}} e^{A_i t} dt H_i$.

3. Transmission Delay Analysis in Edge Computing Environment

Generally, the total time delay during the transmission process from the underlying participating equipment to the edge computing device includes the following three parts: serial delay, propagation delay, and routing delay [16]. Hence, the following equation holds.

$$\tau_{\Sigma} = \tau_{\text{serial}} + \tau_{\text{propagation}} + \tau_{\text{routing}}, \tag{6}$$

where τ_{Σ} is the total time delay. Besides, the serial delay (τ_{serial}) is proportional to the size of packets (denoted as P_{size}) and inversely proportional to the transmission rate (denoted as $D_{\text{transmission}}$). The propagation delay is proportional to the transmission distance (denoted as Length_{Σ}) and inversely proportional to the propagation velocity (denoted as V) in a certain physical media. Obviously, both of the serial delay and propagation delay are constants and can be calculated by the following equations.

$$\begin{aligned} \tau_{\text{serial}} &= \frac{P_{\text{size}}}{D_{\text{transmission}}}, \\ \tau_{\text{propagation}} &= \frac{\text{Length}_{\Sigma}}{V}. \end{aligned} \tag{7}$$

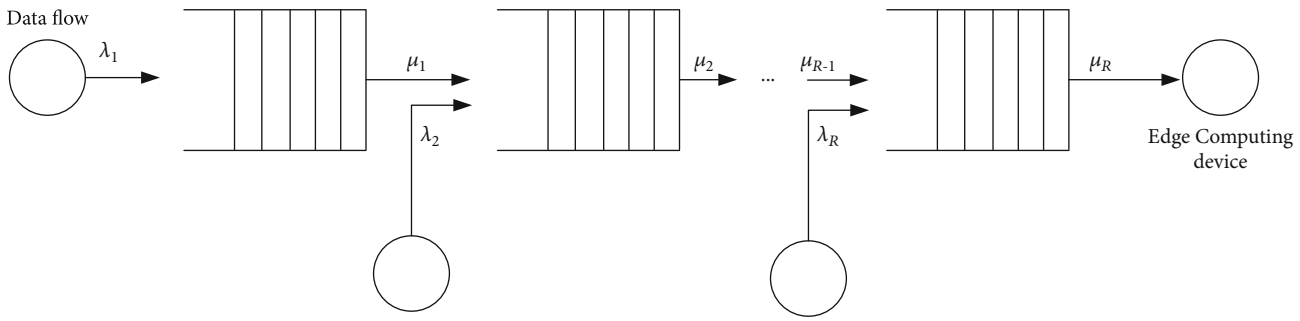


Figure 3: Multihop transmission model in edge computing environment.

In particular, the uncertain routing delay is caused by the queuing and forwarding of multiple data flows in the switches. Specifically, as shown in Figure 3, the operating statuses transmitted from the underlying equipment to the edge computing device need to pass through multiple switches with limited forward capacities. Therefore, at the entrance of a switch, the data flows from different receivers form a queue. Such queuing process can be described as a cascade M/M/1 model based on the queuing theory [21].

Let μ_r denote the forwarding rate of the r-th ($r = 1, 2, \cdots, R$) switch provided for the data flow and τ_r denote the queuing delay of the data flow transmitted by the r-th switch. Assume that a data flow passes through the switches numbered as $1, 2, \cdots, R$ in turn and the data arrival process at the r-th switch obeys the Poisson process with parameter λ_r. Based on the queuing theory, the probability density function (PDF) of the queuing delay at the r-th switch satisfies

$$g(\tau_r) = (\mu_r - \lambda_r)e^{-(\mu_r-\lambda_r)\tau_r}, \quad \tau_r > 0. \tag{8}$$

The corresponding cumulative distribution function (CDF) is given by

$$\bar{g}(\tau_r) = \int_0^{\tau_r} g(t)dt = 1 - e^{-(\mu_r-\lambda_r)\tau_r}, \quad \tau_r > 0. \tag{9}$$

It can be seen that the queuing delay in r-th switch satisfies an exponential distribution with the parameter $\mu_r - \lambda_r$. Furthermore, when the switches $1 - R$ are independent to each other, the total routing delay (τ_{routing}) satisfies

$$\tau_{\text{routing}} = \sum_{r=1}^{R} \tau_r. \tag{10}$$

The corresponding PDF and CDF are given by Equations (11) and (12), respectively.

$$g(\tau_{\text{routing}}) = \prod_{r=1}^{R}(\mu_r - \lambda_r)e^{-\sum_{r=1}^{R}(\mu_r-\lambda_r)\tau_r}, \quad \tau_r > 0, \tag{11}$$

$$\bar{g}(\tau_{\text{routing}}) = \prod_{r=1}^{R}\left(1 - e^{-(\mu_r-\lambda_r)\tau_r}\right), \quad \tau_r > 0. \tag{12}$$

Finally, considering that the serial delay and propagation delay are constants, the PDF and CDF of the total transmission delay (τ_Σ) in the proposed edge computing environment equate to those of the total routing delay (τ_{routing}).

4. Dynamic Characteristic Analysis of Closed-Loop Frequency Control System with Time Delay and Packet Loss

In this paper, to economize the memory space of edge computing devices, each microgrid adopts the memoryless state-feedback control mode, as shown in

$$u_i(k) = K_i X_i^{\text{newest}}(k), \tag{13}$$

where K_i is the controller gain and $X_i^{\text{newest}}(k)$ is the newest packet arrived at the edge computing device.

Noting that the total transmission delay (τ_Σ) always satisfies $\tau_\Sigma \le LT_s (L \in \mathbf{Z}^+)$, in this paper, LT_s is defined as the preset maximum transmission delay. In other words, if the transmission delay of a packet is less than LT_s, then this packet is called an effective packet and will be used to generate the control instructions $u_i(k)$. Otherwise, if the transmission delay of a packet exceeds LT_s, this packet is viewed as a dropped one and will not be used to generate the control instruction $u_i(k)$. Obviously, the different transmission delay may cause that the edge computing device use different packets to generate the control instructions. Hence, the corresponding dynamic characteristics of closed-loop frequency control system will switch with the varying time delays.

Without loss of generality, let p_v and p_{v+1} be the serial number of the two neighbouring effective packets sampled at $t = p_v T_s$ and $t = p_{v+1} T_s$. Both the two packets are used to generate the control instructions due to the fact that their transmission delays are within $[0, LT_s]$. Obviously, the packets sampled within $((p_v + 1)T_s, p_v T_s)$ are dropped due to the transmission delays exceeding LT_s. The maximum number of consecutive dropped packets can be denoted as $D_{\max} = \max(p_{v+1} - p_v)$. Then, the dynamics of closed-loop frequency control system during any two neighbouring effective packets (i.e., $t \in [p_v T_s, p_{v+1} T_s)$) can be described as a switching model dependent on the time-varying delays. In order to clearly illustrate the modelling process for the closed-loop frequency control system with time delay and packet loss, in this section, we take $L = 1$ as an example. As shown in Figure 4, there are two possible scenarios of the closed-loop frequency control system with varying time delays.

Scenario 1: $\tau_\Sigma = 0$, i.e., the packets sampled at $t = p_v T_s$ arrive at the edge computing device without any delay. Therefore, the newest packet satisfies $X_i^{\text{newest}}(k) = X_i(p_v)$, t

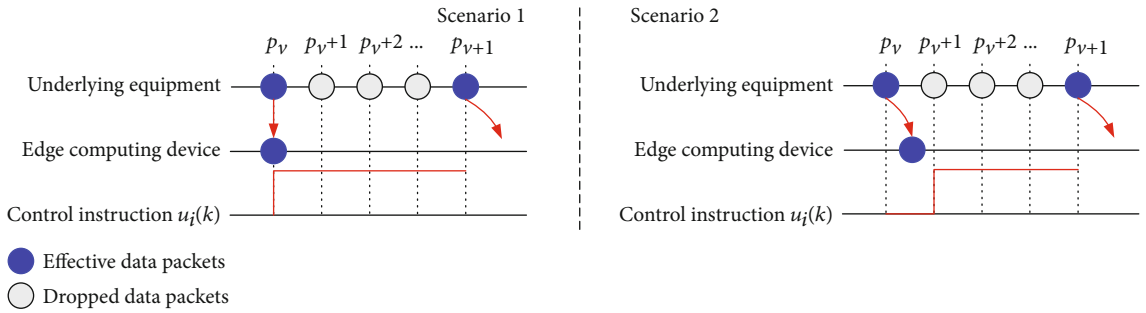


Figure 4: Effects of time delays on the closed-loop frequency control system.

$\in [p_v T_s, p_{v+1} T_s)$. According to the discrete-time model of frequency control system (i.e., Equation (5)), the dynamics of close-loop frequency control systems during the time interval $[p_v T_s, p_{v+1} T_s)$ can be derived as follows:

$$\begin{cases} X_i(p_v+1) = (E_i + F_i K_i) X_i(p_v) + G_i w_i(p_v), \\ X_i(p_v+2) = (E_i + F_i K_i) X_i(p_v+1) + G_i w_i(p_v+1) = \left(E_i^2 + E_i F_i K_i + F_i K_i\right) X_i(p_v) + E_i G_i w_i(p_v) + G_i w_i(p_v+1), \\ \vdots \\ X_i(p_{v+1}) = \left(E_i^{D_{\max}+1} + \sum_{\xi=0}^{D_{\max}} E_i^{\xi} F_i K_i\right) X_i(p_v) + \sum_{\xi=0}^{D_{\max}} E_i^{D_{\max}-\xi} G_i w_i(p_v+\xi). \end{cases} \tag{14}$$

Scenario 2: $\tau_\Sigma \le 1 \times T_s$, i.e., the packet sampled at $t = p_v T_s$ is transmitted to the edge computing device with a delay less than $1 \times T_s$. In this scenario, the control instruction during the time interval $[p_v T_s, p_{v+1} T_s)$ satisfies the following piecewise function

$$X_i^{\text{newest}}(k) = \begin{cases} X_i(p_{v-1}), p_v T_s \le t < (p_v+1) T_s, \\ X_i(p_v), (p_v+1) T_s \le t < p_{v+1} T_s. \end{cases} \tag{15}$$

Similar to Scenario 1, the dynamics of close-loop frequency control systems during the time interval $[p_v T_s, p_{v+1} T_s)$ can be derived as follows:

$$X(p_{v+1}) = \left(E_i^{D_{\max}+1} + \sum_{\xi=0}^{D_{\max}-1} E_i^{\xi} F_i K_i\right) X_i(p_v) + E_i^{D_{\max}} F_i K_i X_i(p_{v-1}) + \sum_{\xi=0}^{D_{\max}} E_i^{D_{\max}-\xi} G_i w_i(p_v+\xi). \tag{16}$$

More generally, the above derivation process in this section can be extended to case of $L \in \mathbf{N}^+$. Define the following two augmented vectors:

$$\begin{aligned} \bar{X}_i(p_v) &= \left[X_i^{\mathrm{T}}(p_v), X_i^{\mathrm{T}}(p_{v-1}), \cdots, X_i^{\mathrm{T}}(p_{v-L})\right]^{\mathrm{T}}, \\ W_i(p_v) &= \left[w_i^{\mathrm{T}}(p_v), w_i^{\mathrm{T}}(p_v+1), \cdots, w_i^{\mathrm{T}}(p_v+L)\right]^{\mathrm{T}}. \end{aligned} \tag{17}$$

Hence, there exist 2^L possible scenarios of the closed-loop frequency control system in the i-th microgrid, uniformly described by

$$\begin{cases} \bar{X}_i(p_{v+1}) = \underbrace{\begin{bmatrix} E_i^{D_{\max}+1} + \sum_{\xi=0}^{D_{\max}-\zeta} E_i^{\xi} F_i K_i & \sum_{\xi=D_{\max}-\zeta+1}^{\psi_1} E_i^{\xi} F_i K_i & \cdots & \sum_{\xi=D_{\max}-\zeta+l-1}^{\psi_{L-1}} E_i^{\xi} F_i K_i & \sum_{\xi=D_{\max}-\zeta+L}^{\psi_L} E_i^{\xi} F_i K_i \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix}}_{\Phi_i^{\xi}} \bar{X}_i(p_v) + \underbrace{\begin{bmatrix} E_i^{D_{\max}} G_i & E_i^{D_{\max}-1} G_i & \cdots & G_i \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}}_{\Lambda_i} W_i(p_v), \\ Y(p_v) = \underbrace{\left[C_i, \underbrace{0, 0, \cdots, 0}_{L}\right]}_{\bar{C}_i} \bar{X}_i(p_v), \end{cases} \tag{18}$$

where $\zeta = \{0, 1, \cdots, L\}$; the values of $\{\Psi_1, \Psi_2, \cdots, \Psi_L\}$ are shown in Equation (19) where the arrow "$\longrightarrow$" represents $\{\Psi_1, \Psi_2, \cdots, \Psi_L\}$ take values from the same row.

$$\begin{array}{cccc} \Psi_1 & & \Psi_2 & \cdots \quad \Psi_L \\ \left[\begin{array}{ccccc} D_{\max}-\zeta+1 & \rightarrow & D_{\max}-\zeta+2 & \rightarrow\cdots & D_{\max}-\zeta+L \\ D_{\max}-\zeta+2 & \rightarrow & D_{\max}-\zeta+3 & \rightarrow\cdots & D_{\max}-\zeta+L+1 \\ \vdots & \rightarrow & & \rightarrow\cdots & \vdots \\ D_{\max}-2 & \rightarrow & D_{\max}-1 & \rightarrow\cdots & D_{\max}-1 \\ & & & & 0 \\ & & & & 0 \\ D_{\max}-1 & \rightarrow & 0 & \rightarrow\cdots & \vdots \quad L-1 \\ & & & & 0 \end{array}\right] \end{array} \tag{19}$$

5. H_∞-Controller Optimization considering Dynamic Performance Improvement

The controller gain K_i not only needs to ensure that the closed-loop frequency control systems are asymptotically stable during any two neighbouring effective packets but also require to have a robust attenuation ability against the external power disturbances from the load demands and renewable outputs. In other words, there exists the following inequality between the output $Y_i(p_v)$ and the external power disturbances $W_i(p_v)$ at any time $t = p_v T_s$.

$$\|Y_i(p_v)\|_2 \le \gamma^2 \|W_i(p_v)\|_2, \tag{20}$$

where γ is the attenuation factor and $\|\cdot\|_2$ is the 2-norm operator. Let $k(p_v)$ and $k(p_{v+1}) \in \{1, 2, \cdots, 2^L\}$ denote the serial numbers of the possible operation scenarios at time $t = p_v T_s$ and $t = p_{v+1} T_s$. A candidate Lyapunov function is constructed by

$$V_i(p_v) = \bar{X}_i^{\mathrm{T}}(p_v) \Omega_{k(p_v)} \bar{X}_i(p_v), \tag{21}$$

where $\Omega_{k(p_v)}$ is a symmetric positive definite matrix. When the state vector of closed-loop frequency control system changes from $\bar{X}_i(p_v)$ to $\bar{X}_i(p_{v+1})$, the increment of Equation (21) is given by

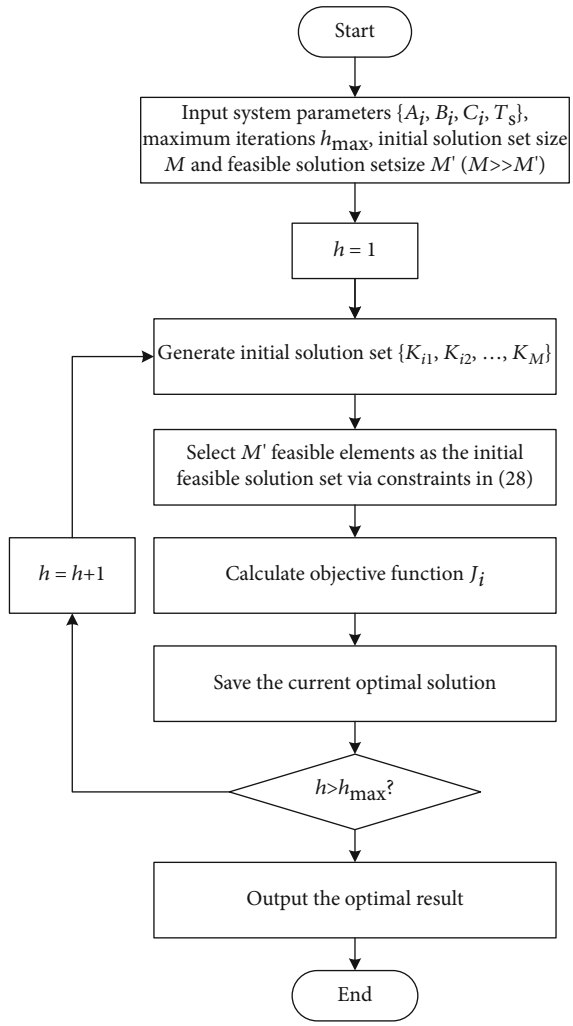


Figure 5: Flowchart for controller gain optimization.

$$\begin{aligned} \Delta V_i(p_v) = V_i(p_{v+1}) - V_i(p_v) = \bar{X}_i^{\mathrm{T}}(p_{v+1}) \Omega_{k(p_{v+1})} \bar{X}_i(p_{v+1}) \\ - \bar{X}_i^{\mathrm{T}}(p_v) \Omega_{k(p_v)} \bar{X}_i(p_v). \end{aligned} \tag{22}$$

Substituting Equation (18) and inequality (20) in Equation (22) results in

$$\begin{aligned} \Delta V_i(p_v) &\le \begin{bmatrix} \bar{X}_i(p_v) \\ W_i(p_v) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \left(\Phi_i^{\zeta}\right)^{\mathrm{T}} \Omega_{k(p_v)} \Phi_i^{\zeta} - \Omega_{k(p_{v+1})} & \left(\Phi_i^{\zeta}\right)^{\mathrm{T}} \Omega_{k(p_v)} \Lambda_i \\ * & \Lambda_i^{\mathrm{T}} \Omega_{k(p_v)} \Lambda_i \end{bmatrix} \begin{bmatrix} \bar{X}_i(p_v) \\ W_i(p_v) \end{bmatrix} - Y_i^{\mathrm{T}}(p_v) Y_i(p_v) + \gamma^2 W_i^{\mathrm{T}}(p_v) W_i(p_v) \\ &= \begin{bmatrix} \bar{X}_i(p_v) \\ W_i(p_v) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \begin{pmatrix} \left(\Phi_i^{\zeta}\right)^{\mathrm{T}} \Omega_{k(p_v)} \Phi_i^{\zeta} - \Omega_{k(p_{v+1})} \\ + \bar{C}_i^{\mathrm{T}} \bar{C}_i \end{pmatrix} & \left(\Phi_i^{\zeta}\right)^{\mathrm{T}} \Omega_{k(p_v)} \Lambda_i \\ * & \Lambda_i^{\mathrm{T}} \Omega_{k(p_v)} \Lambda_i - \gamma^2 I \end{bmatrix} \begin{bmatrix} \bar{X}_i(p_v) \\ W_i(p_v) \end{bmatrix}, \end{aligned} \tag{23}$$

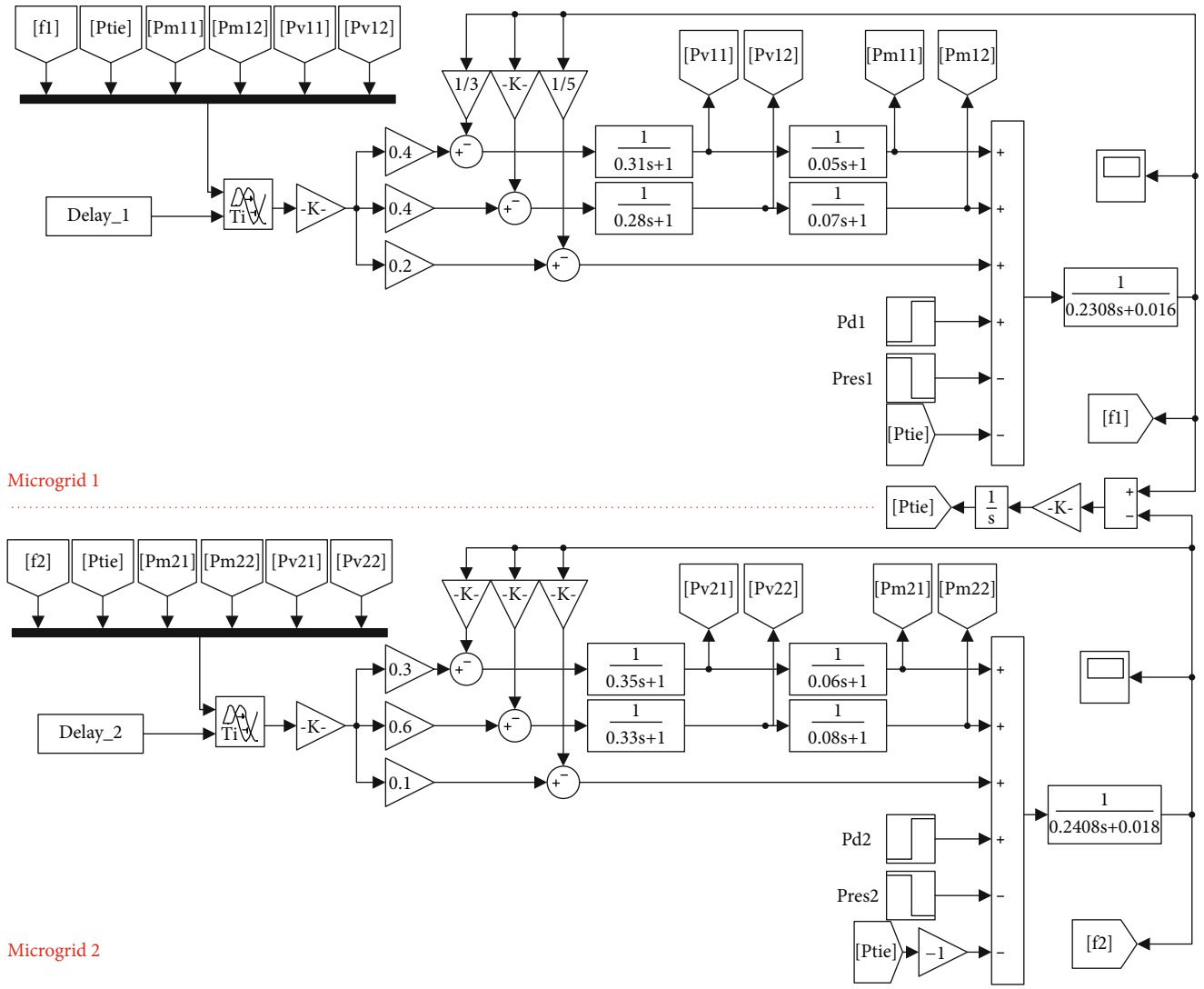


FIGURE 6: Simulation diagram of multi-microgrids.

where I is an identity matrix with suitable dimensions. When the following matrix inequality

$$\begin{bmatrix} \begin{pmatrix} \left(\Phi_i^{\zeta}\right)^{\mathrm{T}} \Omega_{k(p_v)} \Phi_i^{\zeta} - \Omega_{k(p_{v+1})} \\ +\bar{C}_i^{\mathrm{T}} \bar{C}_i \end{pmatrix} & \left(\Phi_i^{\zeta}\right)^{\mathrm{T}} \Omega_{k(p_v)} \Lambda_i \\ * & \Lambda_i^{\mathrm{T}} \Omega_{k(p_v)} \Lambda_i - \gamma^2 I \end{bmatrix} < 0 \tag{24}$$

holds, the integral of Lyapunov function on $[0, +\infty)$ with zero initial conditions satisfies

$$\begin{aligned} 0 \le V_i(+\infty) - V_i(0) &\le \sum_{p_v=0}^{+\infty} \left(-Y_i^{\mathrm{T}}(p_v) Y_i(p_v) + \gamma^2 W_i^{\mathrm{T}}(p_v) W_i(p_v)\right) \\ &= -\|Y_i(p_v)\|_2 + \gamma^2 \|W_i(p_v)\|_2. \end{aligned} \tag{25}$$

Overall, if the matrix inequality (24) holds, the control gain K_i can guarantee the closed-loop frequency control system has the asymptotic stability in the case of stochastic communication changes and H_∞-robust attenuation performance against external power disturbances simultaneously.

Besides, in order to improve the dynamic performance during frequency restoration process, the integral of absolute error (IAE) of the frequency is chosen as the objective function, as shown in Equation (26). The IAE of frequency reflects the transient response performance of the frequency control system to external power disturbances

$$J_i = \sum_{k=0}^{\infty} |\Delta f_i(k)|. \tag{26}$$

In conclusion, a constrained optimization model for the controller gain K_i is established, as follows.

$$\begin{cases} \text{objective}: \ \min \ J_i \\ \text{s.t.}: \ \begin{cases} \begin{bmatrix} \left(\left(\Phi_i^{\zeta}\right)^{\mathrm{T}}\Omega_{k(p_v)}\Phi_i^{\zeta} - \Omega_{k(p_{v+1})}\right) + \bar{C}_i^{\mathrm{T}}\bar{C}_i & \left(\Phi_i^{\zeta}\right)^{\mathrm{T}}\Omega_{k(p_v)}\Lambda_i \\ * & \Lambda_i^{\mathrm{T}}\Omega_{k(p_v)}\Lambda_i - \gamma^2 I \end{bmatrix} < 0 \\ \Omega_{k(p_r)} = \Omega_{k(p_r)}^{\mathrm{T}} > 0, \Omega_{k(p_{r+1})} = \Omega_{k(p_{r+1})}^{\mathrm{T}} > 0. \end{cases} \end{cases} \tag{27}$$

According to Equation (18), the matrix Φ_i^{ζ} contains the controller gain K_i which needs to be optimized. Besides, there exists a multiplicative relation between matrix Φ_i^{ζ} and unknown matrix $\Omega_{k(p_v)}$. Therefore, the constraints (27) do not satisfy the linear matrix inequality (LMI) forms and cannot be solved by directly using the robust control toolbox of MATLAB directly. In this paper, an iterative relaxation technology combined with heuristic search method is proposed to optimize the controller gain K_i. The flowchart is shown in Figure 5. Firstly, an initial solution set containing considerable matrixes with the same dimensions of K_i is generated randomly. Then, the initial feasible solution set is selected from the initial solution set according to the constraints in (27). Secondly, the initial solution set is updated via the operations such as mutation and pheromone update. The constraints (27) are used again to select the feasible solution set for the next iteration. Since the candidates in initial solution set for K_i are fixed during the whole optimization process, the matrix nonlinear matrix inequality in (27) is relaxed into the LMIs.

6. Simulations and Discussions

In this section, the feasibility of the proposed cascade M/M/1 model for communication network in edge computing framework is verified by using OPNET Modeler 14.5. In addition, assume that the number of microgrids in the multi-microgrid is 2 and each microgrid contains two synchronous generators and one energy storage system. The frequency control in multi-microgrid is simulated in MATLAB/Simulink environment, as shown in Figure 6. The time delays during the simulation process adopt the simulation results obtained from the OPNET. The controller gain is obtained according to the optimization algorithm proposed in Section 5 where the particle swarm optimization method is used. The synchronization coefficient between the two microgrids satisfies $T_{12} = T_{21} = 0.52$ p.u./Hz. Besides, the communication system configurations in both two microgrids are assumed to be the same. To evaluate the dynamic performance of the proposed control strategy, the step changes of renewable output and load demand with 0.1 p.u. and -0.1 p.u. occur at $t = 0$ s in each microgrid. Other simulation parameters are demonstrated in Table 1.

6.1. Transmission Delays in Edge Computing Framework. As shown in Figure 7, a multihop data transmission network is established in OPNET environment to verify the proposed cascade M/M/1 model for time delay calculation under the edge computing framework. It should be noted that since the number of transmission hops is equal to the number of forwarding nodes, a chain topology is adopted in the simulation. Figure 8 shows the statistical results of transmission delays of 100 consecutive packets and the theoretical mean value of time delays calculated by the proposed M/M/1 model.

According to Figure 8, the simulation mean value of the transmission delays is 10.15 ms. Meanwhile, the theoretical mean value calculated by the proposed cascade M/M/1 model is 11.41 ms. The relative error between theoretical and simulation values is 11.04%. Therefore, the simulation results illustrate that the proposed cascade M/M/1 model can effectively reflect the transmission characteristics in edge computing environment. In addition, it can be found that the number of continuous packets with transmission delays exceeding 20 ms is less than 3. Therefore, in controller design process, the preset maximum transmission delay is assumed

Table 1: Simulation parameters.

Parameters	Values	Illustrations
T_{g11} (s)	0.31	Microgrid 1
T_{t11} (s)	0.05	
T_{g12} (s)	0.28	
T_{t12} (s)	0.07	
R_{11} (Hz/p.u.)	3	
R_{12} (Hz/p.u.)	2.8	
R_{ESS_11} (Hz/p.u.)	5	
M_{eq_1} (p.u. s)	0.2308	
D_{eq_1} (p.u./Hz)	0.016	
$(\alpha_{11}, \alpha_{12}, \beta_{11})$	(0.4, 0.4, 0.2)	
T_{g21} (s)	0.35	Microgrid 2
T_{t21} (s)	0.06	
T_{g22} (s)	0.33	
T_{t22} (s)	0.08	
R_{21} (Hz/p.u.)	2.87	
R_{22} (Hz/p.u.)	2.5	
R_{ESS_21} (Hz/p.u.)	4.5	
M_{eq_2} (p.u. s)	0.2408	
D_{eq_2} (p.u./Hz)	0.018	
$(\alpha_{21}, \alpha_{22}, \beta_{21})$	(0.3, 0.6, 0.1)	
T_s (ms)	10	Communication networks
P_{size} (bits)	200	
$D_{transmission}$ (Mbps)	10	
Length_{Σ} (km)	50	
V (km/s)	1.8×10^5	
λ_r (packet/s)	50	
μ_r (packet/s)	500	
R	5	

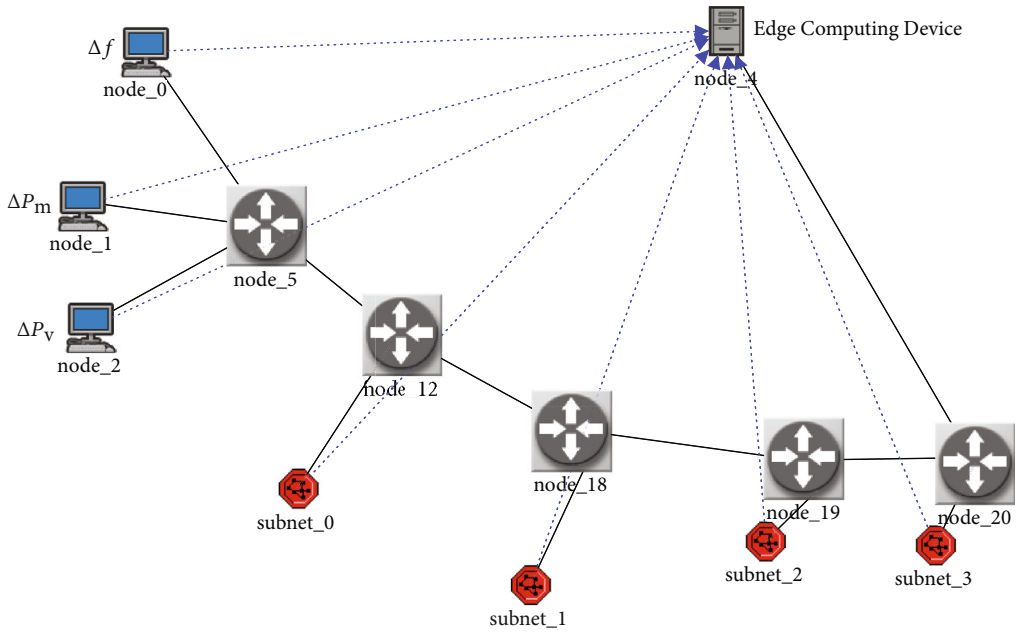


Figure 7: Schematic diagram of communication network in edge computing framework.

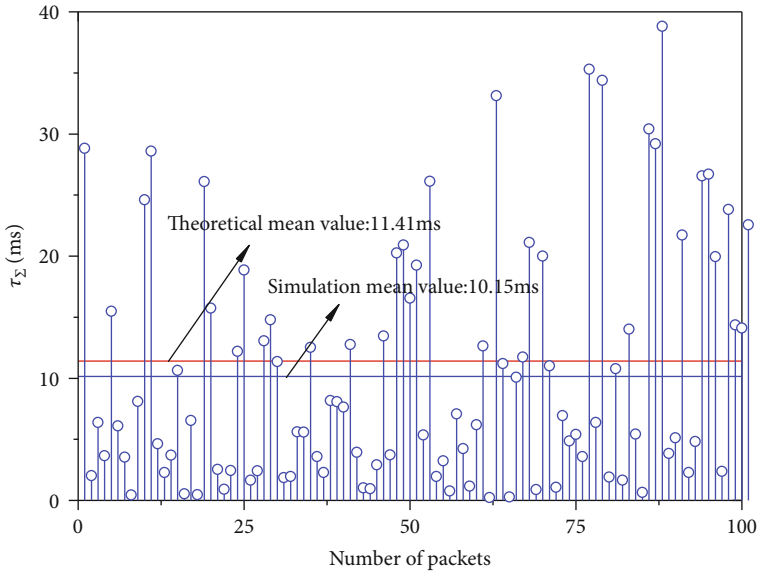


Figure 8: Statistical results of transmission delays of 100 consecutive packets in edge computing environment.

to be 20 ms (i.e., 2×10 ms) while the maximum number of consecutive dropped packets (i.e., $D_{\max}$) is assumed to be 3.

6.2. Control Performance Analysis with the Proposed Method

6.2.1. Comparisons with the Traditional Centralized Control Scheme. In the traditional centralized control schemes, all the operation statuses of the microgrids should be transmitted to the only control center in the whole power system. Without loss of generality, the simplest case that there is only one switch between the edge computing device and the control center is considered in this subsection, as shown in Figure 9. Figure 10 shows the statistical results of transmission delays of 100 consecutive packets and the theoretical mean value of time delays calculated by the proposed M/M/1 model. According to Figure 10, the number of continuous packets with transmission delays which exceeds 20 ms is less than 6. Therefore, the preset maximum transmission delay is assumed to be 20 ms (i.e., 2×10 ms) while the maximum number of consecutive dropped packets (i.e., $D_{\max}$) is assumed to be 6. In addition, in order to quantitatively evaluate the control performance, the following indexes are selected:

(1) Peak value of frequency deviation: the maximum value of the absolute value of frequency deviation when step changes of external power disturbances occur

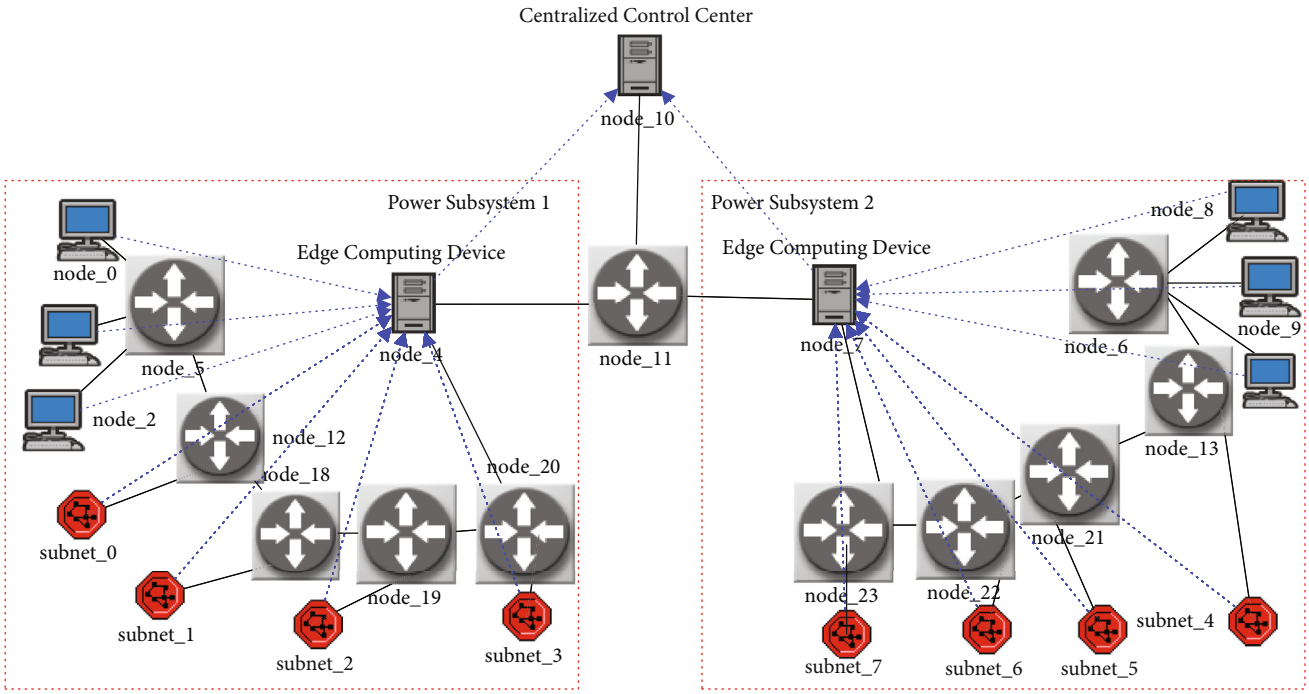


Figure 9: Schematic diagram of communication network in traditional centralized framework.

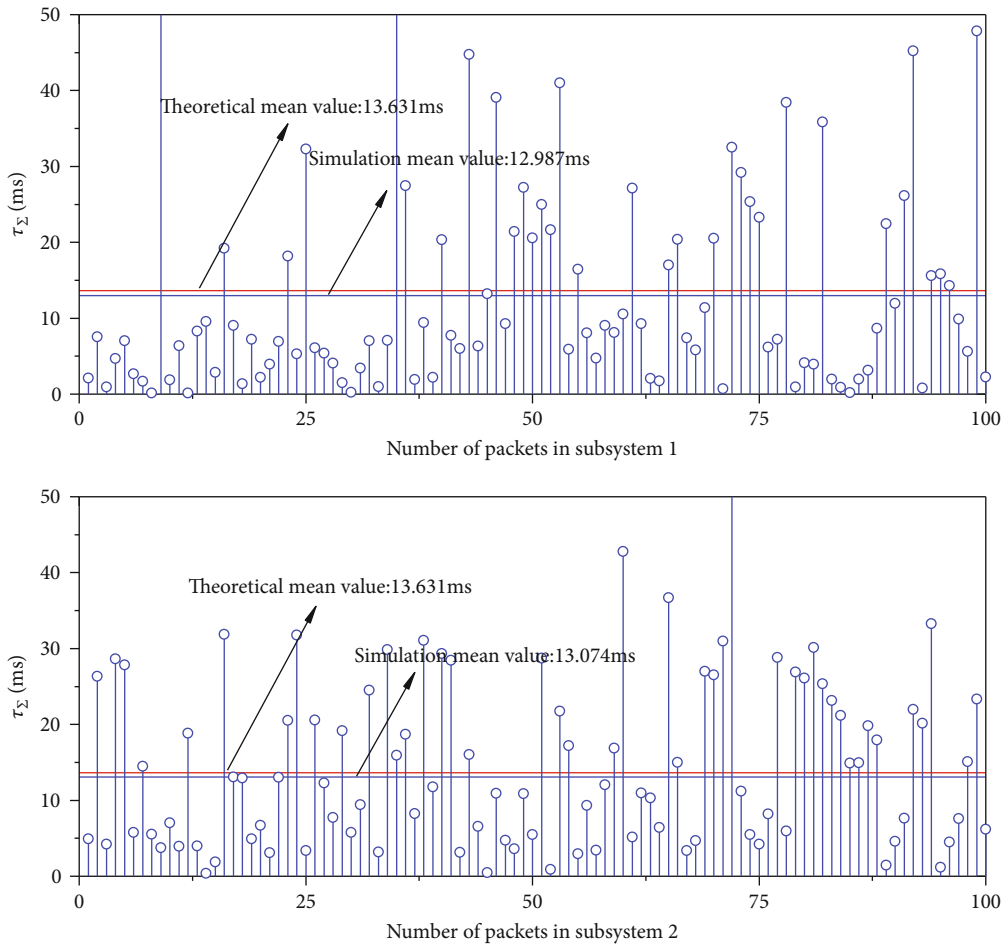


Figure 10: Statistical results of transmission delays of 100 consecutive packets in traditional centralized framework.

Table 2: Performance indexes with traditional centralized framework and proposed edge computing framework.

Microgrids	Control strategies	Peak value (Hz)	Restoration time (s)	IAE (Hz)
1	This paper	0.0164	1.36	0.4182
	Centralized framework	0.0243	1.14	0.6092
2	This paper	0.0209	2.01	0.8920
	Centralized framework	0.0370	4.27	3.5346

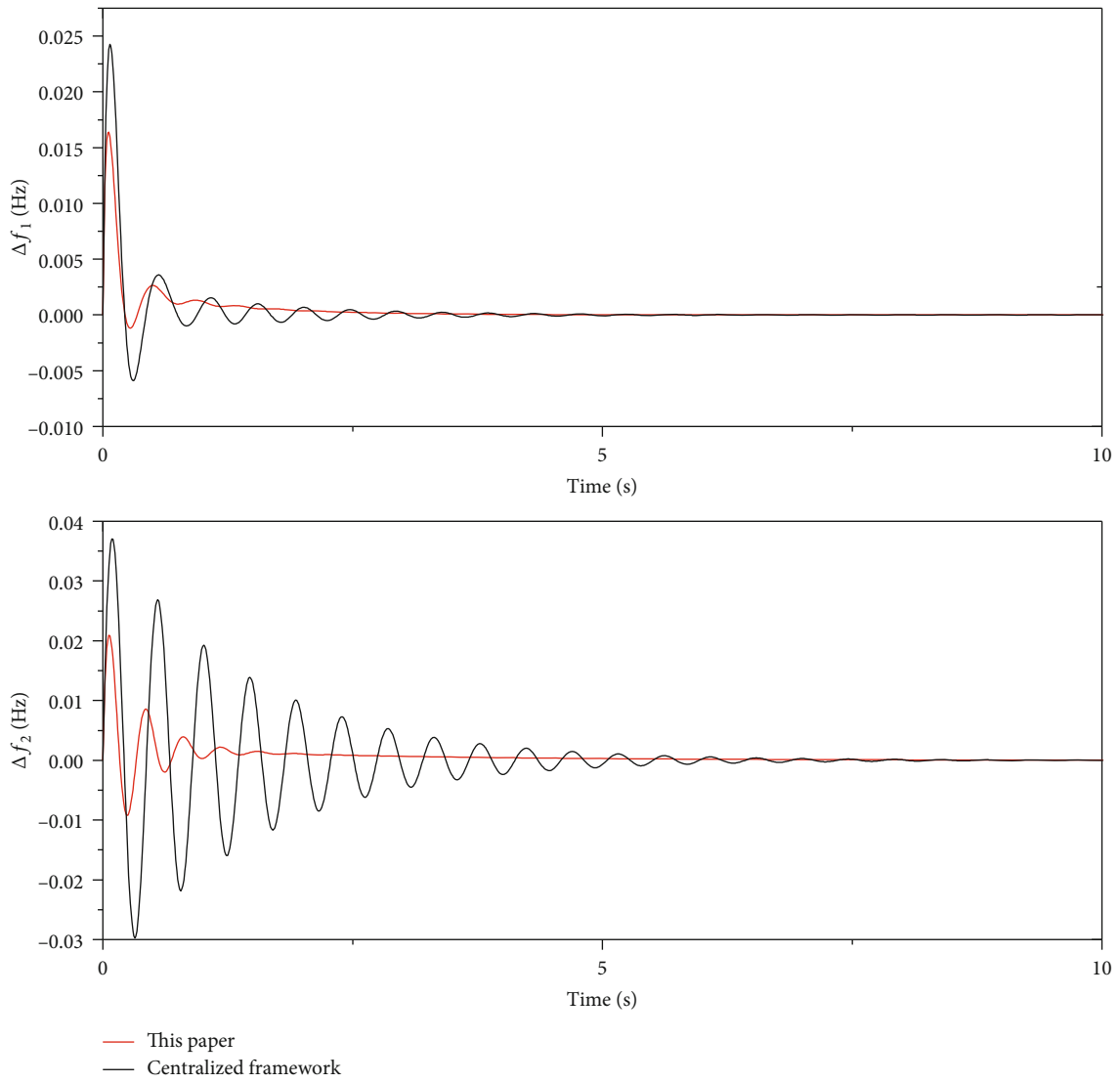


Figure 11: Frequency deviations with traditional centralized framework and proposed edge computing framework.

(2) Restoration time of frequency deviation: the starting time when the frequency deviation restores and maintains within ±5% of the peak value

(3) IAE of the frequency: the result calculated according to Equation (26)

In this subsection, the controller gains of both centralized scheme and edge computing framework are obtained with the proposed H_∞-switching control method in Section 5. Figure 10 shows the frequency deviation responses with the traditional centralized and the proposed edge computing framework. Table 2 demonstrates the corresponding performance indexes. For the first microgrid, according to Figures 10 and 11 and Table 2, although the restoration times of the two control schemes are close, the peak value of our proposed edge computing framework is reduced by 32.51% compared with the traditional centralized control scheme. Meanwhile, for the second microgrid, not only the peak value of our proposed edge computing framework is reduced by

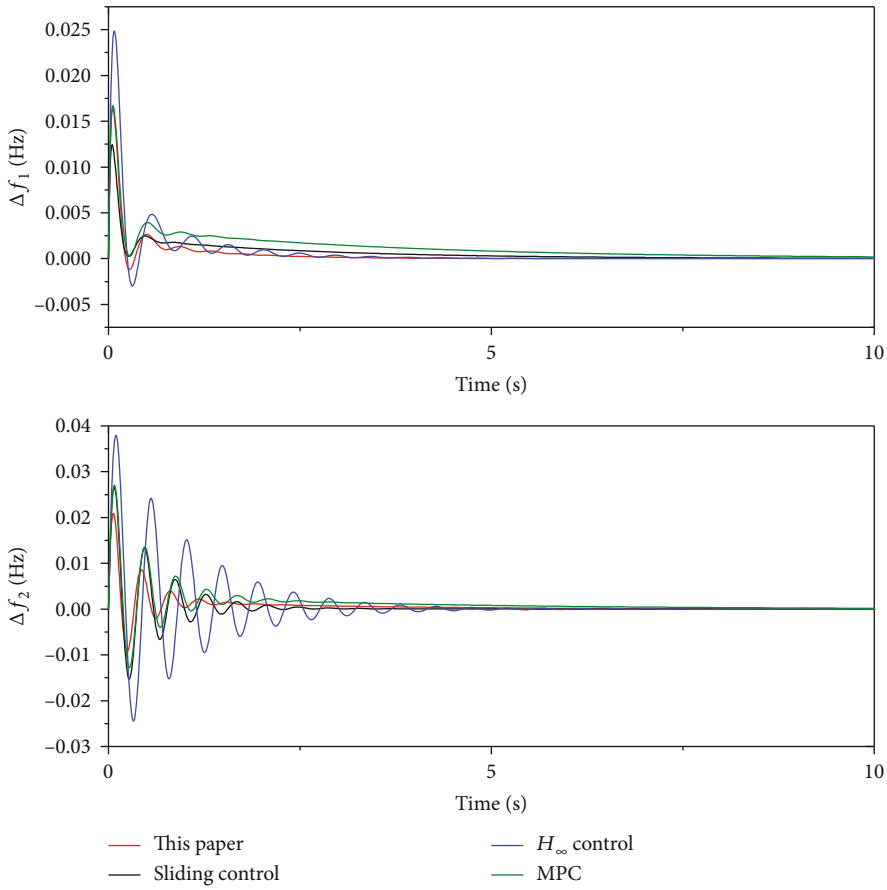


Figure 12: Frequency deviations with different control strategies.

Table 3: Performance indexes with different control strategies.

Microgrids	Control strategies	Peak value (Hz)	Restoration time (s)	IAE (Hz)
1	This paper	0.0164	1.36	0.4182
	Sidling control	0.0124	3.26	0.6658
	H_∞ control	0.0248	1.65	0.7259
	MPC	0.0167	4.95	1.3294
2	This paper	0.0209	2.01	0.8920
	Sidling control	0.0266	1.73	1.0061
	H_∞ control	0.0380	2.93	2.4598
	MPC	0.0270	3.31	1.6974

43.51% but also the corresponding restoration time can be shortened by 52.93% compared with the centralized control scheme. In addition, the IAEs in both two microgrids using the proposed edge computing framework are less than the results using the centralized scheme. The simulation results show that the proposed edge computing framework can provide a better frequency deviation damping performance. This is because the proposed control strategy based on edge computing framework only requires the local operation statuses and the edge computing device which is closer to the underlying participating equipment. Hence, compared with the traditional centralized control scheme, the communication network with edge computing framework can provide a higher quality of service (QoS) for the data flows and has the less consecutive dropped packets.

6.2.2. Comparisons with Different Controller Design Methods in Edge Computing Framework. In this subsection, the superiority of dynamic performance with our proposed H_∞-switching control strategy compared with other design methods is discussed. The comparison methods are as follows: (1) H_∞ control strategy in [9]; (2) MPC strategy in [17]; (3) sliding control strategy in [18]. All the control strategies in this subsection adopt the decentralized scheme and can be applied into the edge computing framework.

Figure 12 shows the frequency deviation responses with different controller design methods. Table 3 demonstrates corresponding dynamic performance indexes. For the first microgrid, according to Figure 12 and Table 3, compared with the sliding control strategy, the maximum frequency deviation with our proposed H_∞-switching control is 32.3% larger while the restoration time is shortened by 58.28%. Compared with H_∞ control, the maximum frequency deviation and restoration time are reduced by 33.87% and 17.58%, respectively. Compared with the MPC method, the maximum frequency deviation and restoration time are reduced by 1.80% and 72.53%, respectively. Note that the IAE of frequency with the proposed H_∞-switching control is reduced by 37.19%, 42.40%, and 68.54%, respectively, compared with the sliding control, H_∞ control, and MPC. Similar results can be obtained for the second microgrid. The reason is that our proposed controller design method optimizes the dynamic performance during frequency restoration process on the premise of ensuring the asymptotical stability of the closed-loop frequency control systems. Therefore, the shorter restoration time and better transition process can be obtained.

7. Conclusions

In this paper, an edge computing-based control scheme is proposed to deal with the frequency stability problem in multi-microgrids. Different from the traditional centralized scheme, the edge computing device is set up in each microgrid to realize the local frequency stability with higher QoS for data transmission and the lower computation burden for controller design. Via describing the closed-loop frequency control system as a delay-dependent switching model, the constraints of controller gain guaranteeing asymptotic stability and H_∞-attenuation performance are strictly derived. Moreover, an optimization algorithm is proposed aiming at improving the transition performance during frequency restoration process. Simulation results show that our proposed method has shorter restoration time in stochastic time delay and packet loss cases. Further researches will focus on the frequency stability problem in edge computing framework suffering from malicious network attacks.

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 is no conflict of interest regarding the publication of this paper.

Acknowledgments

This paper was supported by the Science and Technology Project of State Grid Hebei Electric Power Co. LTD. (B104DY200299).

References

[1] A. Jafari, H. G. Ganjehlou, T. Khalili, and A. Bidram, "A fair electricity market strategy for energy management and reliability enhancement of islanded multi-microgrids," *Applied Energy*, vol. 270, p. 115170, 2020.

[2] B. Wang, Q. Sun, R. Han, and D. Ma, "Consensus-based secondary frequency control under denial-of-service attacks of distributed generations for microgrids," *Journal of the Franklin Institute*, vol. 358, no. 1, pp. 114–130, 2021.

[3] M. M. Esfahani, A. Hariri, and O. A. Mohammed, "A multiagent-based game-theoretic and optimization approach for market operation of multimicrogrid systems," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 280–292, 2019.

[4] T. Yang, Y. Zhang, Z. Wang, and H. Pen, "Secondary frequency stochastic optimal control in independent microgrids with virtual synchronous generator-controlled energy storage systems," *Energies*, vol. 11, no. 9, p. 2388, 2018.

[5] D. E. Olivares, A. Mehrizi-Sani, A. H. Etemadi et al., "Trends in microgrid control," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1905–1919, 2014.

[6] L. Yin, T. Yu, B. Yang, and X. Zhang, "Adaptive deep dynamic programming for integrated frequency control of multi-area multi-microgrid systems," *Neurocoputing*, vol. 344, pp. 49–60, 2018.

[7] W. Liu, W. Gu, W. Sheng, X. Meng, Z. Wu, and W. Chen, "Decentralized multi-agent system-based cooperative frequency control for autonomous microgrids with communication constraints," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 2, pp. 446–456, 2014.

[8] X. Liu, H. Nong, K. Xi, and X. Yao, "Robust distributed model predictive load frequency control of interconnected power system," *Mathematical Problems in Engineering*, vol. 2013, 10 pages, 2013.

[9] H. R. Baghaee, M. Mirsalim, G. B. Gharehpetian, and H. A. Talebi, "A generalized descriptor-system robust H_∞ control of autonomous microgrids to improve small and large signal stability considering communication delays and load nonlinearities," *International Journal of Electrical Power & Energy Systems*, vol. 92, pp. 63–82, 2017.

[10] J. Pahasa and I. Ngamroo, "PHEVs bidirectional charging/-discharging and SoC control for microgrid frequency stabilization using multiple MPC," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 526–533, 2015.

[11] T. Yang, Y. Zhang, W. Li, and A. Y. Zomaya, "Decentralized networked load frequency control in interconnected power systems based on stochastic jump system theory," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4427–4439, 2020.

[12] K. S. Ko and D. K. Sung, "The effect of EV aggregators with time-varying delays on the stability of a load frequency control system," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 669–680, 2018.

[13] L. Xiong, H. Li, and J. Wang, "LMI based robust load frequency control for time delayed power system via delay margin estimation," *International Journal of Electrical Power & Energy Systems*, vol. 100, no. 1, pp. 91–103, 2018.

[14] Ş. Sönmez, S. Ayasun, and C. O. Nwankpa, "An exact method for computing delay margin for stability of load frequency control systems with constant communication delays," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 370–377, 2016.

[15] S. Wen, X. Yu, Z. Zeng, and J. Wang, "Event-triggering load frequency control for multiarea power systems with communication delays," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 2, pp. 1308–1317, 2016.

[16] V. P. Singh, N. Kishor, and P. Samuel, "Load frequency control with communication topology changes in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 5, pp. 1943–1952, 2016.

[17] P. Ojaghi and M. Rahmani, "LMI-based robust predictive load frequency control for power systems with communication delays," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 4091–4100, 2017.

[18] X. Su, X. Liu, and Y. Song, "Event-triggered sliding-mode control for multi-area power systems," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 8, pp. 6732–6741, 2017.

[19] G. Zhang, C. Li, D. Qi, and H. Xin, "Distributed estimation and secondary control of autonomous microgrid," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 989–998, 2017.

[20] M. Varzaneh, A. Rajaei, A. Jolfaei, and M. Khosravi, "A high step-up dual-source three-phase inverter topology with decoupled and reliable control algorithm," *IEEE Transactions on Industry Applications*, vol. 56, no. 4, pp. 4501–4509, 2020.

[21] J. W. Stahlhut, T. J. Browne, G. T. Heydt, and V. Vittal, "Latency viewed as a stochastic process and its impact on wide area power system control signals," *IEEE Transactions on Power Systems*, vol. 23, no. 1, pp. 84–91, 2008.

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6694729, 15 pages
https://doi.org/10.1155/2021/6694729

Research Article

Data Security Storage Method for Power Distribution Internet of Things in Cyber-Physical Energy Systems

Jiayong Zhong and Xiaofu Xiong

State Key Laboratory of Power Transmission Equipment & System Security and New Technology (Chongqing University), China

Correspondence should be addressed to Jiayong Zhong; 20071102080@cqu.edu.cn

Received 30 October 2020; Revised 4 December 2020; Accepted 15 December 2020; Published 5 January 2021

Academic Editor: Mohammad R. Khosravi

The existing cloud storage methods cannot meet the delay requirements of intelligent devices in the power distribution Internet of Things (IoT), and it is difficult to ensure the data security in the complex network environment. Therefore, a data Security Storage method for the power distribution IoT is proposed. Firstly, based on the "cloud tube edge end" power distribution IoT structure, a cloud edge collaborative centralized distributed joint control mode is proposed, which makes full use of the collaborative advantages of cloud computing and edge computing to meet the real-time requirements. Then, a distributed data storage method based on the Kademlia algorithm is proposed, and the homomorphic encryption and secret sharing algorithm are used to store the data in the cloud as ciphertext and perform data query directly on the ciphertext. Finally, considering the heterogeneity of edge nodes, the security protection model of edge nodes based on noncooperative differential game is established, and the algorithm of optimal defense strategy of edge nodes is designed to ensure the security of edge nodes. The experimental results show that the proposed method obtained excellent query performance, and the ability to resist network attacks is better than other comparison methods. It can reduce the data storage and query delay and ensure the data security of the system.

1. Introduction

As the core manifestation of the application in the field of power Internet of Things (IoT), the power distribution IoT in cyber-physical energy systems is responsible for the visual perception of the state of the distribution network, the IoT to manage and control the distribution network equipment, the opening of the distribution service capabilities, and the sharing of distribution network data [1]. On the one hand, a large number of sensor and complex communication networks were used to turn the distribution network into a multidimensional and heterogeneous complex network capable of real-time perception, dynamic control, and information query by the power distribution IoT in cyber-physical energy systems; its massive external data can affect the distribution network. The control decision of the electrical system increases the complexity of operation and control [2, 3].

With the development of cloud computing technology, more and more power grid companies are accustomed to using various services provided by cloud service providers to meet the needs of power business application development and data storage [4]. In recent years, applications such as IoT, artificial intelligence, and big data have also developed rapidly. However, because cloud computing is located at the upper layer of the network and is far away from the actual physical equipment, it cannot achieve good support for low-latency power business applications and cannot meet certain requirements. Some power applications must rely on local equipment to perform a large number of calculations [5, 6]. Edge computing allows devices to complete data collection and preprocessing in the local network by deploying edge computing devices close to the data source, thereby overcoming the problems of low processing speed and large transmission delay for massive native data in cloud computing [7]. The edge computing nodes in the power distribution IoT use edge intelligent terminals to complete the collection, aggregation, and model processing of IoT device data to meet the response requirements of low-latency applications [8].

The cooperation of cloud edge collaboration overcomes the problems of cloud computing for distributed data collection, transmission delay, and data analysis efficiency [9]. And in the power distribution IoT, the edge intelligent terminals are deployed near the power grid line data source to provide computing services, which has the advantages of real-time and efficiency [10, 11]. However, with the introduction of edge computing, a large amount of data is stored in the local edge intelligent terminal, which brings serious security risks. Moreover, edge computing involves the interaction between the edge intelligent terminal and the downstream terminal device, the interaction between the edge intelligent terminal and the upstream cloud platform, the interaction between the edge intelligent terminal, etc., which will lead to the security threats from the end devices, the edge intelligent terminal itself, the edge network infrastructure, and the cloud platform [12, 13]. At the same time, the development of the network security standards of the power distribution IoT is uneven, resulting in greater difficulties in protecting data storage from external threats [14]. Therefore, it is meaningful to study the security protection of the distributed storage of the power distribution IoT to ensure the safety and reliability of the grid data.

2. Related Research

In the existing research methods, most methods are based on the topology model to establish the power grid information model by centralized storage, which is mainly divided into three types based on the adjacency matrix, the correlation characteristic matrix [15], and the graph theory [16]. Ref. [17] studies the storage architecture of mobile edge computing, which explores the potential of mobile edge computing to enhance data analysis of IoT applications. The experiment results show that the data security and computing efficiency were achieved. Ref. [18] proposed an efficient and secure encrypted search architecture based on mobile cloud storage. In architecture, mobile devices can offload intensive computing tasks to edge servers to improve efficiency. In addition, in order to protect data security, the correlation between query keywords and search results from the cloud is hidden to reduce the information acquisition of untrusted cloud. However, the architecture model has the defect of a large amount of data, which requires a lot of memory resources for calculation, which is not suitable for a large power grid [19]. Ref. [20] proposed a nontechnical loss (NTL) detection scheme supported by edge computing and big data analysis tools to solve the problem of big data NTL fraud detection in a smart grid, providing experience for the development of big data security solutions in smart grid. However, it only focuses on the topological connection relationship, and the data interaction relationship is over conceptualized and unable to correspond with the actual system components [21]. Ref. [22] proposed a data exchange architecture for energy Internet that takes into account edge computing efficiency and data security. In this architecture, edge computing is applied to solve the challenges related to data exchange and data security at the same time. However, due to the lack of topological structure caused by the complete formulation, the model cannot reflect the actual structural characteristics of the system.

Due to the large amount of data, the above control mode model is difficult to ensure the real-time control and information security, and the energy consumption of cloud computing is too high [23]. Based on the concept of edge computing, Ref. [24] proposes an efficient and privacy-preserving data download scheme for VANET. By analyzing the encrypted requests from nearby vehicles, the road-side unit can find popular data without sacrificing the privacy of its download request. The results of the security analysis show that the scheme can resist various security attacks and improve the download efficiency of the system. Ref. [25] proposed an effective ciphertext policy attribute based on the encryption scheme, which introduced the concept of partial hiding policy to protect private information in the access policy. From the perspective of distributed control, Ref. [26] constructs a cloud edge collaborative computing framework and proposes data token and energy token inspired by blockchain and security solutions for protecting vehicle data interaction. However, the introduction of edge computing into the cyber-physical system storage data security modeling is still lack of research. Based on the existing research, this paper constructs a cloud edge collaborative data processing structure model of the power distribution IoT based on the existing research and studies the data Security Storage methods of the Distribution IoT.

Aiming at the data security problem in cloud edge collaboration of power distribution IoT in cyber-physical energy systems, a data Security Storage method is proposed. The innovation of the proposed method is as follows:

(1) In view of the fact that the distribution cloud master station cannot meet the demand of massive terminal data request delay, the proposed method is based on the "cloud-tube-edge-end" power distribution IoT structure in cyber-physical energy systems and proposes a cloud edge collaborative control mode, which makes full use of the coordination of cloud and edge computing to improve the efficiency

(2) Aiming at improving the data storage security of the edge intelligent terminal, a distributed data storage method based on the Kademlia algorithm is proposed, and the improved homomorphic encryption and secret sharing algorithm are used to make all the edge intelligent terminal data stored and queried in the ciphertext

(3) Because of the heterogeneous and distributed characteristics of edge intelligent terminals, it is easier for network attackers to launch malicious attacks. Therefore, the proposed method establishes an intrusion prevention model of edge intelligent terminals based on the stochastic differential game, which provides the optimal defense strategy for each edge intelligent

terminal, so as to ensure the data security of power distribution IoT

3. System Architecture

3.1. Hierarchical Structure of Power Distribution IoT in Cyber-Physical Energy Systems. The power distribution IoT is the embodiment of the application of the power IoT in the field of distribution. It undertakes the functions of perceiving the status of the visual distribution network, controlling the distribution network equipment, opening the service ability of the distribution network, and sharing the data of the distribution network, so as to realize the internal support of the grid operation, customer service, enterprise operation, and other businesses, and the external business supports the resource commercial operation, energy finance, comprehensive energy service, and virtual power plant and other businesses [27]. The power distribution IoT overall structure in cyber-physical energy systems is shown in Figure 1.

The structure of power distribution IoT can be divided into four core levels of "cloud-management-edge-device," and each level is described as follows.

(1) *Cloud*: as the distribution cloud master station platform, it adopts cloud computing, big data, artificial intelligence, and other technologies to realize the comprehensive cloud and microservice of the master station under the IoT architecture. The first mock exam platform of distribution cloud can satisfy the business requirements of massive devices such as plug and play, data integration, and cloud collaboration. It supports the business requirements such as low voltage unified model management, plug and play, data cloud synchronization, and IoT management. The main station needs to have flexible Internet of Things cloud service and cloud edge collaboration ability, which could meet the requirements of rapid response, dynamic allocation of resources, intensive operation, and maintenance of the system at the same time. "Cloud" layer includes the platform as a service, infrastructure as a service, and software as a service layer

(2) *Management*: as a data transmission channel of "cloud," "edge," and "end," it is used to complete the efficient transmission of massive information in the power grid. It can be divided into two main parts: remote communication network and local communication network, where the remote communication network provides the data communication channel between the distribution cloud master station and the edge intelligent terminal, and the local communication network provides the data communication channel between the edge intelligent terminal and the terminal unit

(3) *Edge*: the edge intelligent terminal, with "edge cloud, cloud gateway" as the main landing form, and "cloud edge collaboration, edge intelligence" as the core feature, which is an open platform for data aggregation and computing. In the power distribution IoT system structure, the edge intelligent terminal is the carrier and key link of terminal data self-organization and end cloud business self-coordination, which realizes the decoupling of terminal hardware and software functions. For the "end" end, the data exchange and intelligent sensing equipment are used to complete the edge end collaboration to achieve full data acquisition, full perception, and full control; for the "cloud" end, the edge intelligent terminal and the distribution cloud master station interact in real-time and full-duplex mode with key operation data to complete edge cloud collaboration, give full play to the expertise of cloud computing and edge computing, and realize reasonable division of labor

(4) *Devices*: terminal device (various types of sensor units), as the sensing layer and execution layer in the power distribution IoT architecture; "end" refers to the source of basic data such as operation status, environmental status, and equipment environmental status of the distribution network to "edge" or "cloud," and the terminal for executing decision-making command or local control

3.2. Cloud Edge Collaboration for Power Distribution IoT in Cyber-Physical Energy Systems. Different from the centralized storage structure where the distribution cloud master station completes all the computing tasks, the edge intelligent terminal is added to the edge side of the cloud edge collaborative structure near the data source, as shown in Figure 2. The distributed collaboration theory divides the distribution network terminal devices and the edge intelligent terminals into multiple distributed collaboration according to the region and operation state. All the power and information components in each distributed collaboration and the edge intelligent terminal jointly constitute a distributed open-edge service platform integrating the core functions of the network, storage, computing, and application, providing the edge intelligent services in the regional distributed collaboration, shorten the information transmission link, and realize the communication and regional interconnection with the cloud computing center through the backbone network [28].

The control mode of cloud edge collaboration can make full use of the collaborative advantages of cloud computing and edge computing, realize unified scheduling, and meet the security and real-time requirements [29, 30]. The business in the local area is uploaded to the edge intelligent terminal after the data is collected by the terminal devices, which is executed locally by the edge intelligent terminal or completed by the cooperation of multiple edge intelligent terminals through the local area networks, such as plug and play application and application localization management. The data information of the edge intelligent terminal and the terminal devices is stored in the edge intelligent terminal in modular form. Some advanced applications, such as distribution network

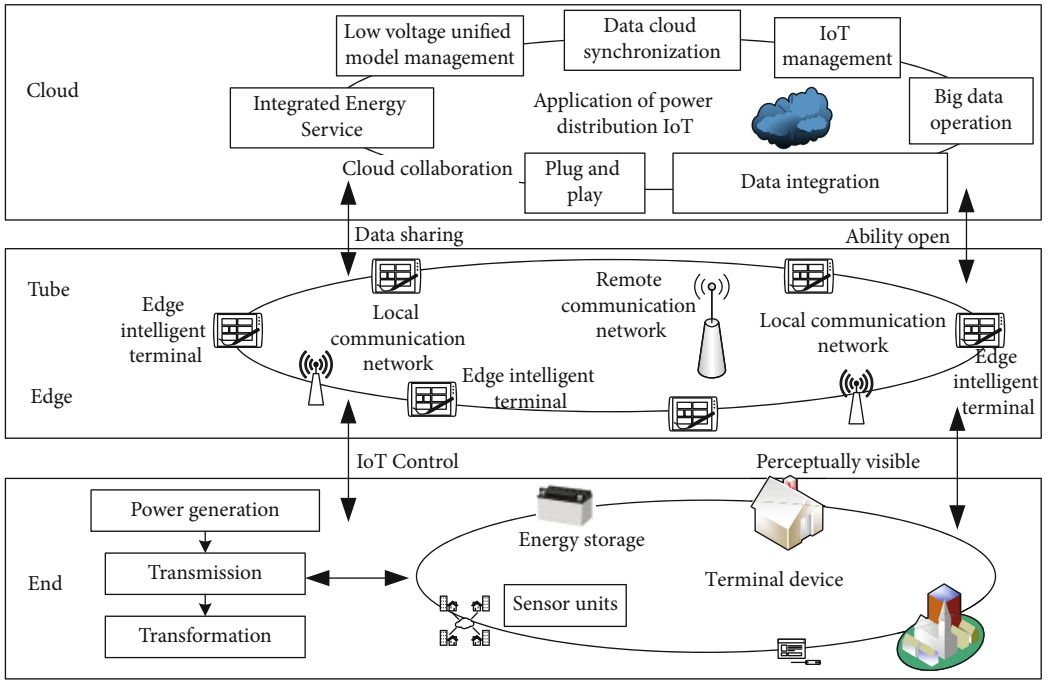


Figure 1: Overall structure of power distribution IoT in cyber-physical energy systems.

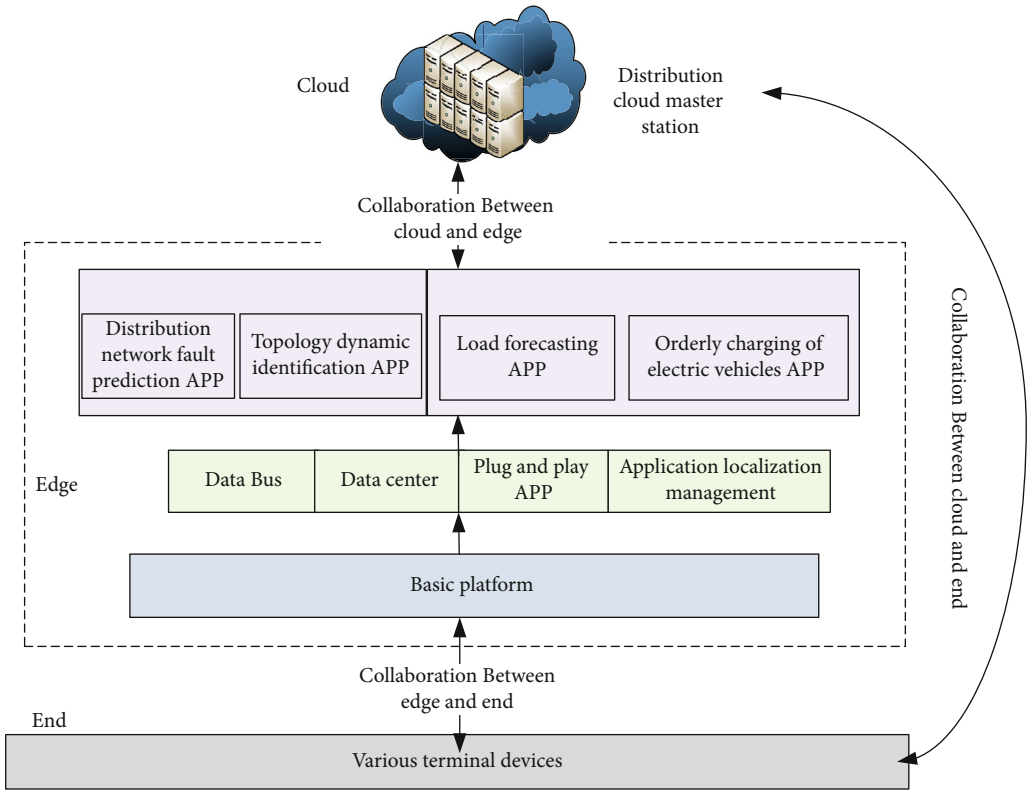


Figure 2: The structure of centralized-distributed joint control based on cloud-edge collaboration.

fault prediction, topology dynamic identification, orderly charging of electric vehicles, and load forecasting, are completed by the edge intelligent terminal and the distribution cloud master station. Among them, the edge intelligent terminal completes the optimization calculation, and then, the distribution cloud master station sends control commands to the edge intelligent terminal for partition execution [31].

4. Data Security Storage Method

4.1. Data Storage and Query Model

4.1.1. Data Storage Model. Firstly, it is necessary to create a database on the edge intelligent terminal, that is, the client application, and randomly generate two large prime numbers $p1$ and $p2$ (usually used to use prime numbers with more than 512 bits, such as 1024 bits), to obtain the product n of the two, namely:

$$\varphi(n) = (p1 - 1)(p2 - 1), \tag{1}$$

A random number p also need to be generated to represent a positive integer coprime with n. Then, a new table T is created in the distribution cloud master station database, a field column A is created, and a column key named $ck_A = \langle x_A, y_A \rangle$, x_A, and y_A are randomly generated, but $x_A, y_A < n$ is required. In addition, each row is defined as $r_i (r_i > 0)$ which is stored separately in a column named *row-id*, and the *row-id* requires additional encryption, and the value of the column can be encrypted using an improved homomorphic encryption algorithm (defined as $r_i (r_i > 0)$) that supports addition. In this way, table T has two columns (*row-id*, A). The edge intelligent terminal only needs to store $p1$, $p2$, and ck_A, and the actual value of the table is stored in the distribution cloud master station database.

In summary, the data model is built in the integer field for operation. After getting the plaintext data V to be inserted, V needs to be encrypted by ck_A and r_i. In other words, V_{key} is generated by ck_A andr_i, and V_{key} is generated as follows:

$$V_{\text{key}} = g(r, (x, y)) = xp^{ry \bmod \varphi(n)} \bmod n, \tag{2}$$

Then, V_e is generated by V_{key} and plaintext V. V_e is the encrypted ciphertext value of the data:

$$V_e = E\left(V, V_{\text{key}}\right) = VV_{\text{key}}^{-1} \bmod n, \tag{3}$$

where V_{key}^{-1} is the modular inverse of V_{key}.

The generated V_e is stored in the distribution cloud master station database, and V_{key} as the intermediate value of calculation does not need to be stored, because the value of V_{key} can be recovered through ck_A and r_i. V_{key} and V_{key} values are needed to decrypt the data when the value needs to be decrypted:

$$V = D\left(V_e, V_{\text{key}}\right) = V_e V_{\text{key}} \bmod n. \tag{4}$$

For the whole database, the edge intelligent terminal only needs to save two positive integers n and p, while for table T and column A in the database, the edge intelligent terminal only needs to save the column key ck_A of the column. In the distribution cloud master station database, the encrypted line number $E^+(r)$ and the ciphertext value A_e of the data are saved in the database. Compared with other encryption cloud data storage models, this model does not need to occupy additional database space of the distribution cloud master station to store metadata for data repair [32].

4.1.2. Query Model. The database system SHAMC can directly execute ciphertext SQL queries on the data tables created by the database layer of the power distribution cloud master station, which all rely on the improved homomorphic encryption algorithm of the model. The query algorithm is jointly implemented by the protocol stack designed and stored on the edge intelligent terminal and the power distribution cloud master station database [33, 34]. These protocols are designed and written in the User-Defined Function (UDF) of the edge intelligent terminal of the database management software (DMS).

SHAMC supports most of the operators of SQL statements and can pass all the statement tests of TPC-H. Taking the commonly used multiplication operators as an example, we will introduce the process of implementing encrypted queries.

Assuming that the data table T has two encrypted columns, column A and column B, the calculation result $A \times B$ is to be obtained. A, B keys $ck_A = \langle x_A, y_A \rangle$ and $ck_B = \langle x_B, y_B \rangle$. Assuming that the result column is column C, the column key of column C is $ck_C = \langle x_C, y_C \rangle$. To get the value of C through the values of A and B, you need to calculate C_e and ck_C. Specifically, execute the protocol edge intelligent terminal protocol *mul_cal_x* and *mul_cal_y*, get ck_c:

$$ck_C = \langle x_C, y_C \rangle = \langle x_A y_B, x_A + y_B \rangle. \tag{5}$$

Then, execute the protocol *mul_cal_*c_e on the database of the power distribution cloud master station to get c_e:

$$C_e = A_e B_e \bmod n. \tag{6}$$

can be pushed:

$$C_{\text{key}} = x_c \cdot p^{\text{ryc}} = x_A \cdot x_B \cdot p^{r(x_A + y_B)} = A_{\text{key}} \cdot B_{\text{key}} (\bmod n). \tag{7}$$

Therefore, it can be proved:

$$\begin{aligned} C &= C_e \cdot C_{\text{key}} = A_e \cdot B_e \cdot C_{\text{key}} \\ &= A \cdot A_{\text{key}}^{-1} \cdot B \cdot B_{\text{key}}^{-1} \cdot A_{\text{key}} \cdot B_{\text{key}} = A \cdot B. \end{aligned} \tag{8}$$

4.2. Data Safe Storage.

In order to avoid the problems caused by the centralized storage system, a distributed data storage system is designed based on the Kademlia algorithm by using the edge computing architecture. The Kademlia algorithm has the characteristics of simplicity, flexibility, and security. Assign a randomly generated 160-bit node identity (ID, identity) to each edge intelligent terminal joining the Kademlia network. The 160-bit hash value of the encrypted data block is used as the number, called the key, and the encrypted data block itself is used as the value, and then, the data block is stored in the form of key-value pairs on several edge intelligent terminals with ID values similar to the key. The maximum number of nodes that can be accommodated in the Kademlia network is 2,160, and its storage capacity far

exceeds the number of devices required in the actual network, thus meeting the scalability requirements of large-scale IoT applications [35].

Each edge intelligent terminal in the distributed storage system only stores a part of the encrypted data and does not store a complete data ledger. In addition, the state information of the edge intelligent terminal is stored in each node through the K-bucket mechanism. Kademlia algorithm calculates the distance between nodes through exclusive OR operation. The distributed storage structure based on edge computing is shown in Figure 3. Each edge intelligent terminal has a 160-layer K-bucket mechanism table.

For K-bucket i, the edge intelligent terminal stores the status messages of k nodes whose distance is $[2^{i-1}, 2^i)$. These messages include node ID, Internet Protocol (IP) address, and access port. k is a system-level constant, which can be set to 8 according to the dynamic setting of the storage system, such as the Kademlia algorithm used in the bit stream. The state storage method based on the K-bucket mechanism makes n edge intelligent terminals need lg$n$ queries at most to find the target information.

The distributed storage architecture based on edge computing effectively avoids the two common problems of traditional distributed systems. Firstly, the entry/exit of nodes in a distributed system is very frequent. When the node status changes, the entire network will update the broadcast address and synchronize the nodes, which leads to network congestion and greatly reduces the storage and search efficiency [36]. In the proposed secure storage solution, each node only maintains some of the messages of edge intelligent terminals, so that the impact on the entire network is minimized when any node changes its state. Then, in the traditional architecture, each node maintains the status information of the entire network. Once a node is attacked or deliberately committed evil, the status information of all nodes will be leaked. The Kademlia algorithm is used to provide partition fault tolerance for the storage system, which greatly reduces the risk of information leakage.

4.3. Data Defense Model. Edge intelligent terminals process and store data, and the separation of ownership and control rights causes edge intelligent terminals to lose physical control of their data. A large number of edge intelligent terminals, local deployment, and wide geographic distribution make it easier and more efficient for intruders in this computing mode to launch denial of service attacks [37]. If effective detection and defense mechanisms are not deployed on edge smart terminals, malicious intruders can launch attacks by consuming limited resources of computing and bandwidth. Meanwhile, it also can forge false data centers, deceive edge smart terminals and obtain users sensitive data or even tty to control the devices.

In order to establish a defense mechanism suitable for edge intelligent terminals, in this section, modeling and analysis of the interaction behavior between attack nodes and edge nodes by noncooperative differential game theory are taken into account, where the heterogeneity of distribution IoT and the ability of edge nodes are able to respond detection and defense functions autonomously.

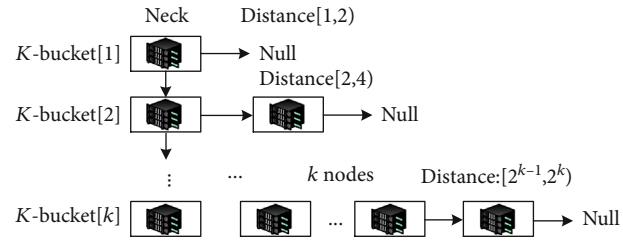


Figure 3: Distributed storage architecture based on edge computing.

In the environment of edge computing, the number of edge intelligent terminals is recorded as N, and each edge intelligent terminal is deployed with an intrusion prevention system, so that $x(t)$ is the number of intruders at t time, and represents the defense strength of the intrusion prevention system deployed at the edge intelligent terminal i at time t, where $i = 1, 2, \cdots, N$. Let $v(t)$ denote the attack frequency of the intruder at t time. When the intruder attacks any edge intelligent terminal maliciously, the change process of the number is related to the defense strength of the edge intelligent terminal's intrusion prevention system and the current attack strength. Therefore, the change process of the number of invaders can be described by the following equations:

$$\begin{cases} \dfrac{dx(t)}{dt} = ax(t) - b_i u_i(t) + cv(t), \\ x(t_0) = x_0 > 0, \end{cases} \tag{9}$$

where a represents that when the intruder's trajectory is not detected, the intruder increases the growth rate of its number by attacking the edge intelligent terminal, b_i represents that the intrusion prevention system deployed on the edge intelligent terminal successfully detects and blocks the intruder probability, c represents the probability of an intruder successfully attacking under the action of the intrusion prevention system, t_0 represents the initial time of the game, and x_0 represents the initial number of intruders.

When it is attacked maliciously in the game process, the edge intelligent terminal can detect and block the behavior of malicious intruders by deploying and responding to the intrusion prevention system. The edge nodes also could prevent the attacks by reducing the attack intensity and the number of intruders minimizing the resource consumption cost caused by its own defense measures. In the process of edge intelligent terminal being attacked, with the increase of the number and frequency of intruders, the cost of deploying defense system and reducing the number of intruders are $\alpha_i u_i^2(t)$ and $\varepsilon_i x(t)$, respectively, where α_i is the unit cost of edge intelligent terminal i, and ε_i is the unit cost of reducing the number of intruders to respond to the defense system.

In addition, the cost of computing resources consumed by each edge intelligent terminal to successfully resist malicious attacks can be expressed as a function of attack frequency, i.e., $\beta v(t) u_i(t)$, where β is the unit cost of computing resource consumption. The resource consumption caused by a false alarm attack of the intrusion prevention

system is $\chi_i u_i(t)$, and χ_i represents the unit cost caused by false alarm attack.

For any edge intelligent terminal i, try to minimize its computing resource cost during the game. According to the above analysis, the total cost function of the edge intelligent terminal i deployed with the intrusion prevention system on the game time $[t_0, T]$ is expressed as:

$$J_i^D = \min_{u_i(t)} \int_{t_0}^{T} (u_i(t)(\alpha_i u_i(t) + \chi_i) + \beta v(t) u_i(t) + \varepsilon_i x(t)) \exp[-r(t-t_0)]dt + q_i(x(T)) \exp[-r(T-t_0)], \tag{10}$$

where T represents the time when the game ends, r represents the ratio of the future cost of the edge smart terminal to the current cost, and $q_i(x(T)) \exp[-r(T-t_0)]$ represents the cost function of the edge smart terminal at the end of the game.

In the attack process, the intruder tries to achieve the maximum damage to the defense system by increase the attack intensity, which is conducted by maximizing the attack frequency and increasing the number of intruders. Therefore, the total cost function of the intruder in the game time $[t_0, T]$ is expressed as:

$$J^A = \min_{v(t)} \int_{t_0}^{T} \left(\eta v^2(t) + \kappa u_i(t) v(t) + \lambda x(t)\right) \exp[-r(t-t_0)]dt + q_i(x(T)) \exp[-r(T-t_0)]. \tag{11}$$

For the edge intelligent terminal, if there is a continuous differentiable function $U^i(t,x)$: $[t_0,\infty] \times R \to R$ for any edge intelligent terminal i, it satisfies the Isaacs Bellman equation:

$$U^i(t,x) = \exp[-r(t-t_0)] \int_0^{\infty} \left(\alpha_i [\phi_i^*(t,x)]^2 + \beta v(t) \phi_i^*(t,x) + \chi_i \phi_i^*(t,x) + \varepsilon_i x(t)\right) \times \exp[-r(t-t_0)]dt. \tag{12}$$

Then, the strategy set $\{u_i^*(t) = \phi_i^*(t,x) | i = 1, 2, \cdots, N\}$ is the feedback Nash equilibrium solution.

According to the Nash equilibrium solution process, in the infinite time domain, the optimal defense strategy of the edge intelligent terminal i is $u^*(t)$:

$$u^*(t) = \frac{(2\eta b_i \varepsilon_i + \beta c \lambda) \exp[r(t-t_0)]}{(4\alpha_i \eta - \beta\kappa)(r-a)} - \frac{2\eta\chi_i}{4\alpha_i\eta - \beta\kappa}. \tag{13}$$

Similarly, the optimal strategy for the intruder i is $v^*(t)$:

$$v^*(t) = \frac{-\varepsilon_i c \exp[r(t-t_0)]}{2\eta r - a} - \frac{2\eta\chi_i\kappa(r-a) + (\varepsilon_i\kappa\beta c + 2\eta b_i \kappa\lambda) \exp[r(t-t_0)]}{2\eta(4\alpha_i\eta - \beta\kappa)(r-a)}. \tag{14}$$

According to the above analysis, the edge intelligent terminal and the intruder adopt stochastic differential game modeling, and the optimal strategies in the finite and infinite time domain are obtained according to the equilibrium solution so that each edge intelligent terminal and intruder can consider the resources maximize revenue under limited circumstances. The data security preserving model of edge intelligent terminal based on the stochastic differential game in the edge computing environment is shown in Algorithm 1.

4.4. Cloud Edge Collaborative Storage Security Defense. The proposed cloud edge collaborative storage security defense scheme includes four stages: preparation stage, transmission stage, sharing stage, and retrieval stage.

(1) *Preparation stage*: each edge intelligent terminal inputs a security parameter 1^k to generate public-private key (PK, SK). The initialization algorithm is as follows:

$$\begin{aligned} (\mathrm{PK}_{\mathrm{PKE}}, \mathrm{SK}_{\mathrm{PKE}}) &\leftarrow \mathrm{PKE.Setup}\left(1^k\right), \\ (\mathrm{PK}_{\mathrm{PKES}}, \mathrm{SK}_{\mathrm{PKES}}) &\leftarrow \mathrm{PKES.Setup}\left(1^k\right), \\ (\mathrm{PK}_{\mathrm{DS}}, \mathrm{SK}_{\mathrm{DS}}) &\leftarrow \mathrm{DS.Setup}\left(1^k\right). \end{aligned} \tag{15}$$

The edge intelligent terminal manages the private key SK by itself and then sends the corresponding public key PK to the CA for registration. CA will use its private key $\mathrm{SK}_{\mathrm{DS}}^{\mathrm{CA}}$ to sign the identity information of the edge intelligent terminal and the public key PK of the edge intelligent terminal, so as to generate the digital certificate Cert of the edge intelligent terminal. Finally, CA sends the generated digital certificate Cert to the edge intelligent terminal.

(2) *Transmission stage*: the data sending terminal device logs into a similar edge intelligent terminal, extracts some keywords W for the query from the data F it wants to store in the distribution cloud master station, and then uses its own private key $\mathrm{SK}_{\mathrm{DS}}^{O}$ to generate a digital signature sign for the data F

$$\mathrm{sign} \leftarrow \mathrm{DS.Sig}\left(\mathrm{SK}_{\mathrm{DS}}^{O}, F\right). \tag{16}$$

Data sending terminal device sends data F, keyword W, authorized terminal device list U, digital signature sign, and its digital certificate Cert^O to edge intelligent terminal

Pseudocode of edge node oriented security defense algorithm
Input: number of nodes N
Begin
1. The security defense model of stochastic differential game is established: $\begin{cases} dx(t)/dt = ax(t) - b_i u_i(t) + cv(t), \\ x(t_0) = x_0 > 0. \end{cases}$
2. Set parameters according to network conditions a, b_i, c, α_i, ε_i, η, κ, λ, r
3. **For** $t = 0$ to T
4. Nash equilibrium method is used to calculate the game model, and the optimal strategy is obtained:
$u^*(t) = (2\eta b_i \varepsilon_i + \beta c\lambda) \exp[r(t-t_0)]/(4\alpha_i\eta - \beta\kappa)(r-a) - 2\eta\chi_i/4\alpha_i\eta - \beta\kappa$,
$v^*(t) = -\varepsilon_i c \exp[r(t-t_0)]/2\eta r - a - 2\eta\chi_i\kappa(r-a) + (\varepsilon_i\kappa\beta c + 2\eta b_i\kappa\lambda) \exp[r(t-t_0)]/2\eta(4\alpha_i\eta - \beta\kappa)(r-a)$.
5. **End for**
6. According to the equilibrium solution structure, the number of intruders is analyzed.
End

Algorithm 1:

through a secure channel. The edge intelligent terminal will generate a unique symmetric key K according to the identifier ID of each data file after receiving the data sent by the data sending terminal device.

$$K \leftarrow \mathrm{SE.Setup}\left(1^k\right). \tag{17}$$

The edge intelligent terminal uses a symmetric key K to encrypt each data file F to generate data ciphertext C_{SE}.

$$C_{\mathrm{SE}} \leftarrow \mathrm{SE.Enc}(K, F). \tag{18}$$

The edge intelligent terminal obtains the certificates $\{\mathrm{Cert}^R | R \in U\}$ of all authorized terminal devices from CA and obtains the public key $\{PK^R_{PKE}, PK^R_{PEKS}, PK^R_{DS} | R \in U\}$ of all authorized edge intelligent terminals and encrypts symmetric key K and keyword W with $\mathrm{PK}^R_{\mathrm{PKE}}$ and $\mathrm{PK}^R_{\mathrm{PEKS}}$, respectively, to generate symmetric key ciphertext C^R_{PKE} and public key searchable ciphertext $C^{R,W}_{\mathrm{PKES}}$.

$$\begin{aligned} C^R_{\mathrm{PKE}} &\leftarrow \mathrm{PKE.Enc}\left(\mathrm{PK}^R_{\mathrm{PKE}}, K\right), \\ C^{R,W}_{\mathrm{PEKS}} &\leftarrow \mathrm{PEKS.Enc}\left(PK^R_{\mathrm{PEKS}}, W\right). \end{aligned} \tag{19}$$

The edge intelligent terminal uploads data ciphertext C_{SE}, symmetric key ciphertext C^R_{PKE}, public key searchable ciphertext $C^{R,W}_{\mathrm{PKES}}$, digital signature sign of data, and the digital certificate Cert^O of data sending terminal device to distribution cloud master station.

(1) *Sharing stage*: an authorized data receiving terminal device logs in to a neighboring edge intelligent terminal, and a sharing request is submitted to the edge intelligent terminal. The edge intelligent terminal forwards the sharing request to the distribution cloud master station, and the distribution cloud master station returns all data ciphertext C_{SE}, symmetric key ciphertext C^R_{PKE}, the digital signature sign of data, and digital certificate Cert^O of data sending terminal device to the edge intelligent terminal

The edge intelligent terminal obtains the public key $\mathrm{PK}^O_{\mathrm{DS}}$ from the digital certificate Cert^O of the data sending terminal device and sends the symmetric key ciphertext C^R_{PKE} to the authorized data receiving terminal device. The authorized data receiving terminal device uses its own private key $\mathrm{SK}^R_{\mathrm{PKE}}$ to decrypt symmetric key ciphertext C^R_{PKE} and obtain symmetric key K.

$$K \leftarrow \mathrm{PKE.Dec}\left(\mathrm{SK}^R_{\mathrm{PKE}}, C^R_{\mathrm{PKE}}\right). \tag{20}$$

The authorized data receiving terminal device returns the symmetric key K to the edge intelligent terminal through the secure channel, and the edge intelligent terminal uses K to decrypt data ciphertext C_{SE} to obtain plaintext F.

$$F \leftarrow \mathrm{SE.Dec}(K, C_{\mathrm{SE}}). \tag{21}$$

The edge intelligent terminal will return the integrity verified data F to the authorized data receiving terminal device through the secure channel.

(2) *Retrieval stage*: an authorized data receiving terminal device logs in to a neighboring edge intelligent terminal and uses its own private key $\mathrm{SK}^R_{\mathrm{PEKS}}$ to generate a search trapdoor T_W for the keyword W to be queried

$$T_W \leftarrow \mathrm{PEKS.Tra}\left(\mathrm{SK}^R_{\mathrm{PEKS}}, W\right). \tag{22}$$

The data receiving terminal device sends the retrieval request of T_W and its digital certificate Cert^R to the edge intelligent terminal, which forwards the retrieval request to the distribution cloud master station. The distribution cloud master station obtains the public key $\mathrm{PK}^R_{\mathrm{PEKS}}$ of the authorized data receiving terminal device from the digital certificate Cert^R of the authorized data receiving terminal device, and uses PK^R_{PEKS} and T_W to retrieve the matched public key

searchable ciphertext set ψ_{PEKS}^{R} generated by the data sending terminal device for the authorized data receiving terminal device, and searchable ciphertext ψ_{W} can be retrieved.

$$\psi_{W} \leftarrow \mathrm{PEKS.Search}\left(\mathrm{PK}_{\mathrm{PEKS}}^{R}, \psi_{\mathrm{PEKS}}^{R}, T_{W}\right). \tag{23}$$

The distribution cloud master station will return the retrieved public key searchable ciphertext corresponding data ciphertext C_{SE}, symmetric key ciphertext C_{PKE}^{R}, digital signature of data, and digital certificate Cert^{O} of data sending terminal device to the edge intelligent terminal. After that, the data processing process between the edge intelligent terminal and the authorized data receiving terminal device is consistent with the security defense in the sharing phase.

5. Experimental Results and Analysis

The host configuration is Intel Core i3-3240 CPU@3.4GHz and 4GB of memory and using the SHAMC encryption model. My SQL 5.5 is installed at both ends as the basic database, and all encrypted query protocols are built on the UDF of MySQL. The configuration of each distribution cloud master station database is dynamically adjustable, which is convenient for comparative experiments.

In addition, six hosts are used to simulate the edge intelligent terminal, named edge1-6. Edge1-4 is equipped with Intel Xeon CPU e3-1220 (3.00 GHz) and 32 GB random access memory (RAM), while edge5-6 is equipped with Intel Xeon CPU e5620 (2.40 GHz) and 24 GB RAM; a MacBook Pro equipped with Intel Core i9-9880h and 16 GB RAM is used as the IoT consumer.

5.1. Time Delay Analysis. In order to verify the download delay performance of the proposed method, the time delay experiment is carried out. And the result is compared with the traditional cloud storage which is shown in Figure 4.

As shown in Figure 4, when the amount of data to be allocated is very small, the delay performance of cloud storage and cloud edge collaborative storage architecture is similar, because the small amount of data brings less transmission delay, and the powerful computing power of distribution cloud master station can make up for the delayed loss caused by transmission. With the continuous increase of tasks, due to the distance between the distribution cloud master station and the terminal device, a large amount of data will cause a long transmission delay, so the service response delay of cloud storage architecture increases significantly. Compared with cloud storage architecture, because the edge computing layer is close to the end devices, it can provide services for the end devices at the network edge, so the cloud edge collaborative network architecture has better delay performance.

As for the performance of the stochastic differential game algorithm in this optimization problem, it is compared with the algorithm in Ref. [17], Ref. [20], and Ref. [25], and the results are shown in Figure 5.

As can be seen from Figure 5, the delay of the four optimization algorithms increases with the increase of the task amount. However, in the case of the same amount of data, the delay of the proposed stochastic differential game algorithm is significantly less than that of other comparative algorithms, which fully proves that it has better delay optimization performance, can quickly complete information exchange, and is suitable for high standard security protection.

5.2. Comparative Analysis of Storage Capacity. The storage capacity of the edge intelligent terminal has a great impact on the data security storage performance of the power distribution IoT. In the experiment, the storage capacity of the edge intelligent terminal changes from 100 to 200 data blocks, and the network delay is 10 ms.

In order to demonstrate the performance of the proposed data security storage method, its storage capacity is compared with Ref. [17], Ref. [20], and Ref. [25]. The average acquisition delay of data resources is shown in Figure 6.

It can be seen from Figure 6 that the average acquisition delay of the distributed storage method is significantly lower than that of other storage methods. As the storage capacity of edge intelligent terminals increases, the average acquisition delay of various storage methods has decreased. This is because the greater the storage capacity of edge intelligent terminals, the more data can be stored at the network edge, thereby reducing the average acquisition delay. With the increase of the storage capacity of edge intelligent terminals, when the storage capacity is 200 data blocks, compared with 100 data blocks, the average acquisition delay of the proposed storage method is reduced by 58.3%. Ref. [17], Ref. [20], and Ref. [25] reduced by 25.5%, 22.6%, and 40.8%, respectively. It can be seen that the average acquisition delay reduction effect of the proposed storage method is the best.

5.3. Query Performance Analysis. TPC-H performance test specification is used to analyze the query performance. TPC-H performance test includes all the commonly used query operation operators and contains complex queries. Through TPC-H, it usually means that the database can support normal use and can cope with some complex business scenarios. In the experiment, all TPC-H statements can be executed correctly.

In order to compare the usability of SHAMC with other encrypted databases, two kinds of algorithm prototypes, MONOMI and crypt dB, are implemented in the experiment. In the SHAMC model, Q4, Q11, Q12, q13, Q16, and q21 are not involved in the ciphertext operation, and q13, Q15, and Q16 are not supported by the SDB, Crypt DB, and MONOMI models. Therefore, in a comprehensive consideration, some TPC-H statements are selected for verification. The execution time of the SHAMC ciphertext query TPC-H statement and plaintext query is shown in Figure 7.

As can be seen from Figure 7, SHAMC achieved much more efficiency than Crypt DB in the execution time. In the SHAMC system, most of the computation is transferred to the database layer of the distribution cloud master station. Therefore, in order to further analyze the proportion of processing time in each layer, taking Q1, Q8, q14, and q22 statements of TPC-H as an example, the comparison of the execution time of three processing processes of distribution

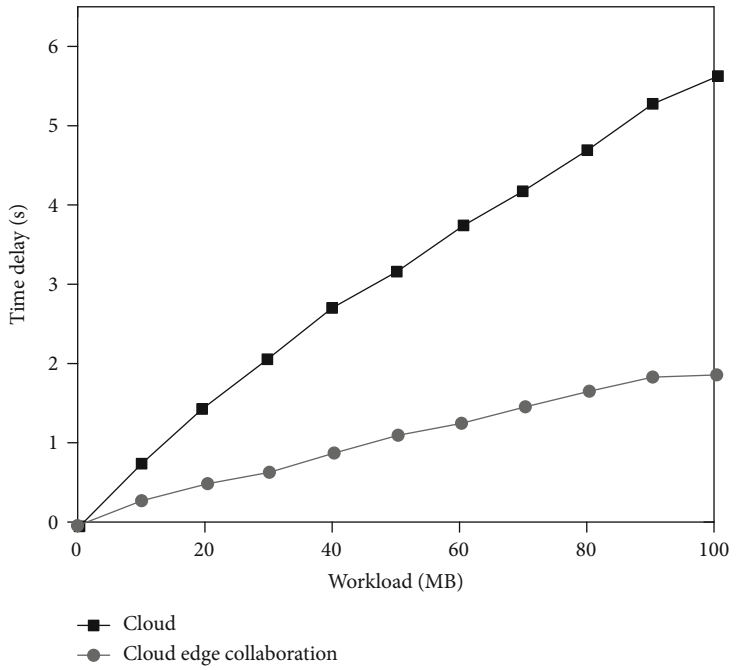


Figure 4: Delay comparison between distributed data storage and cloud storage.

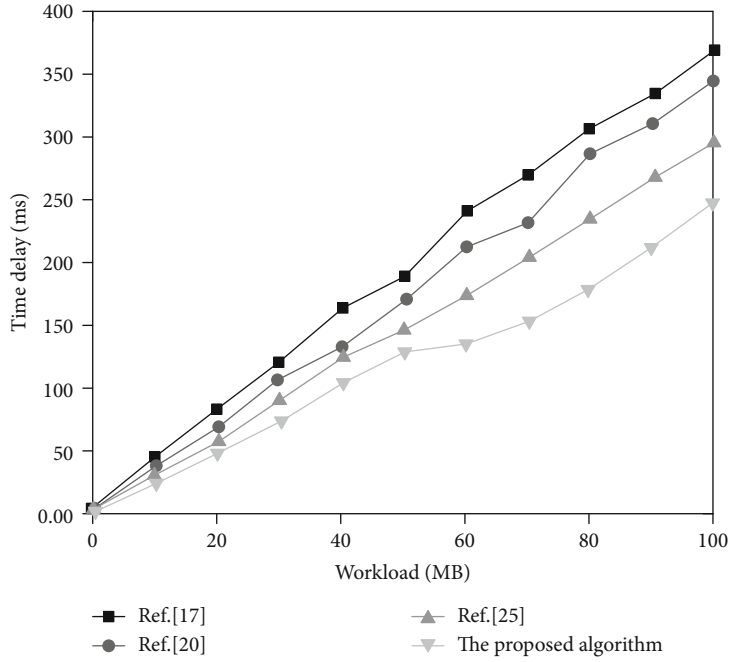


Figure 5: Delay comparison between the proposed algorithm and other algorithms.

cloud master station database layer, client application layer, and network transmission is shown in Figure 8.

As shown in Figure 8, the database protocol operation of the distribution cloud master station in the database layer of the distribution cloud takes up the vast majority of the calculation process. Compared with MONOMI, which has similar query performance, MONOMI needs to precalculate data on the client and work with the cloud to complete the query operation. In general, SHAMC has an acceptable computing overhead and transfers most of the computation to the database layer of the distribution cloud master station, which reduces the computing load of the client.

5.4. Safety Analysis. Consider that the number of edge intelligent terminals participating in the game is $N = 6$, the edge intelligent terminal and the intruder discount their future

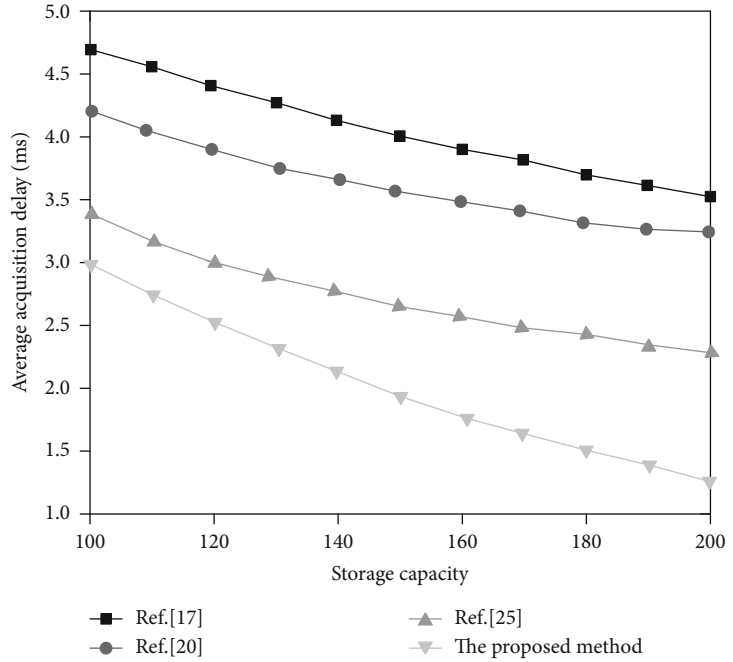


Figure 6: The effect of edge server storage capacity on cache performance.

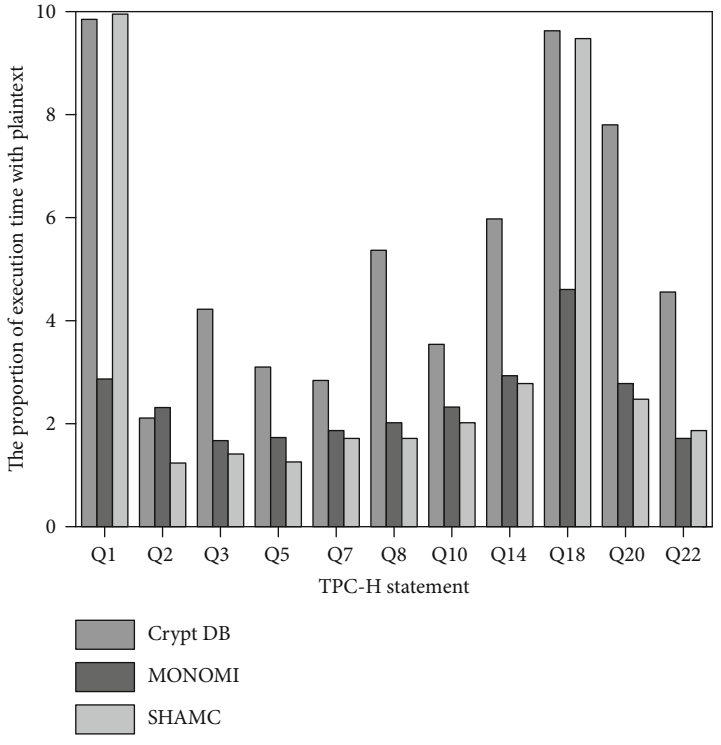


Figure 7: The ratio of execution time between the TPC-H statement and plaintext query in the encrypted database.

costs into the current cost ratio $r = 0.05$, the initial time of the game $t_0 = 0$, and the end time of $T = 20$ s.

Consider that when the probability of an intruder's successful attack is greater than or equal to 90%, the dynamic change of its attack frequency $v^*(t)$ over time is shown in Figure 9.

As can be seen from Figure 9, the attack frequency of intruders decreases with time. With the enhancement of the defense level of the edge intelligent terminal, the attack frequency of the intruder decreases gradually with the improvement of the defense level of the edge intelligent terminal. At the same time, in the process of launching

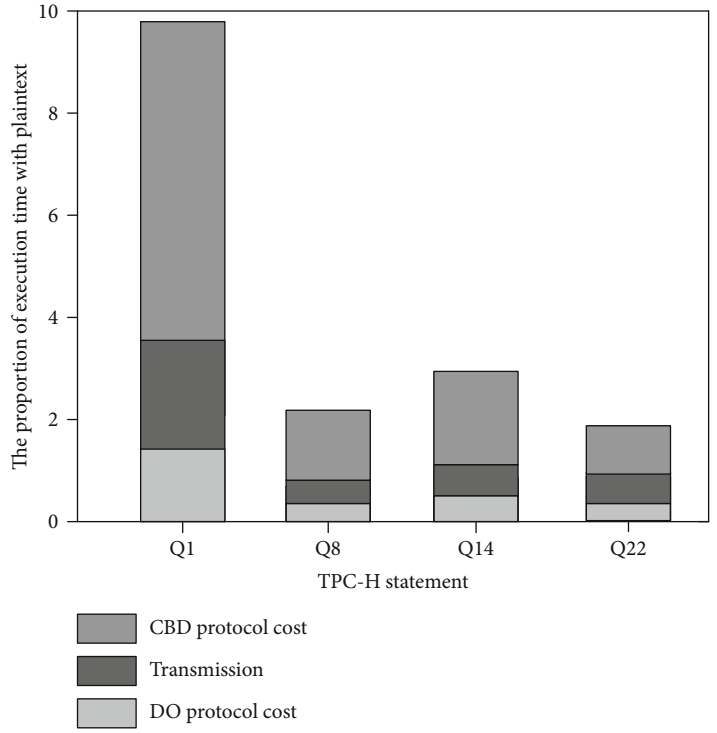


Figure 8: Comparison of execution time and plaintext query time of each process.

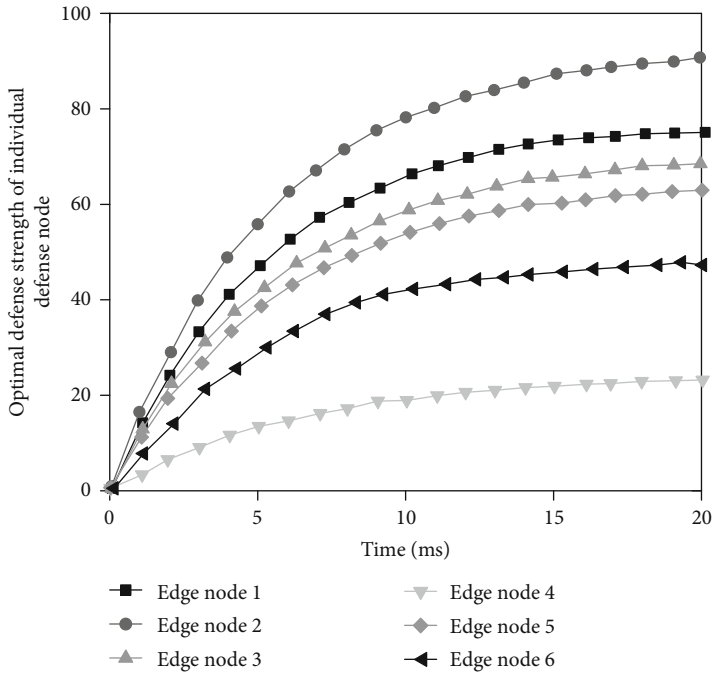


Figure 9: Change of individual optimal defense strategy of the edge node with time.

the attack, the intruder tries to maximize the illegal benefits and minimize the cost by dynamically adjusting its attack strength.

According to the above analysis, the edge intelligent terminal selects the optimal defense strategy when the intruder selects its optimal attack intensity. The change track of the number $x^*(t)$ of intruders over time is shown in Figure 10.

As shown in Figure 10, the number of intruders gradually decreases with time. Combined with the analysis in Figure 9, after $t = 10$ ms, the attack intensity of the intruders would not threaten the edge intelligent terminal security. Therefore, the proposed security protection method can effectively resist intruders while improving the security of edge smart terminals.

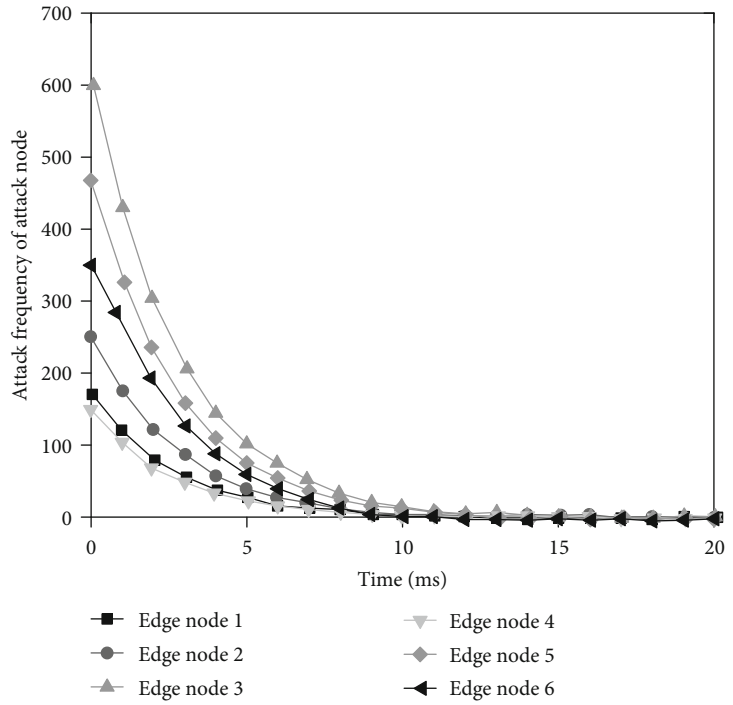


Figure 10: The change of optimal attack frequency of attack node with time.

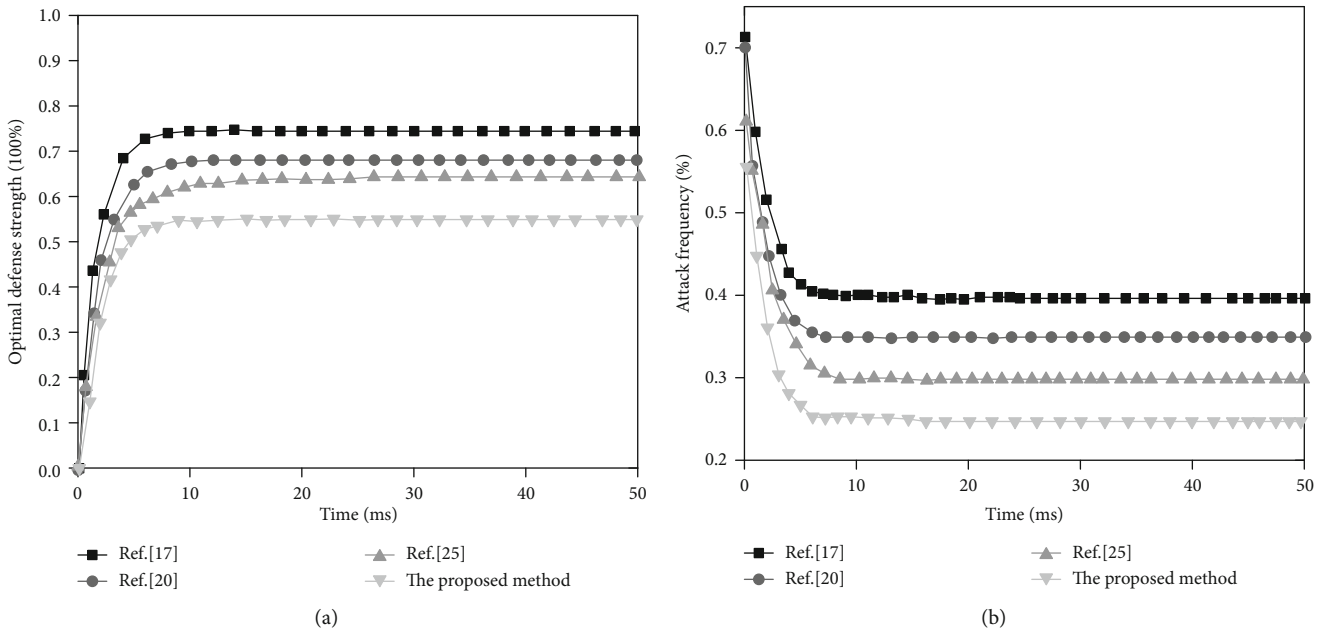


Figure 11: Safety performance comparison of various protection methods.

In order to demonstrate the proposed method security performance, the optimal defense strength and the attack frequency are compared with the defense models proposed in Ref. [17], [20], and [25], as shown in Figure 11.

As Figure 11 shows that with the change of time, the protection methods of the proposed protection method and the comparison method increase rapidly and tend to be stable, while the attack frequency of the intruder is gradually reduced and tends to be stable. However, the proposed method can control the attacker's attack frequency better when the edge intelligent terminal consumes low computing resources.

6. Conclusion

The rapid growth of the number of intelligent terminal devices at the edge of the power distribution IoT leads to the massive physical data generated at the edge of the

network. However, the big data technology based on cloud computing can not meet the low energy consumption and real-time requirements of the edge intelligent terminal for data processing. Edge computing makes up for the deficiency of cloud computing. Edge computing offloads cloud computing services to the network edge. However, the edge network environment is more complex, the heterogeneity between terminal devices and the limited resources of computing and storage make the edge intelligent terminals, and their data face a series of new security challenges. Therefore, a data Security Storage method for power distribution IoT is proposed. Based on the "cloud-tube-edge-end" power distribution IoT structure, a cloud edge collaborative centralized distributed joint control mode is proposed to meet the real-time requirements. The distributed data storage method based on the Kademlia algorithm and encryption algorithm is used to store the data in the ciphertext and execute data query directly on the ciphertext to ensure the security of data storage. In addition, the security protection model of the edge intelligent terminal based on the stochastic differential game is established to ensure the security of the edge intelligent terminal. The results show that the storage and query delay of the proposed method is low, and with the improvement of the storage capacity of the server, the data acquisition delay is less. Moreover, it has better security performance than other methods.

The proposed method assumes that the randomness of the attacker obeys the normal distribution in the process of establishing the model. However, in the actual edge network, the behavior of the attacker is more complex, and the random joining or exiting of the edge intelligent terminal will lead to the change of the edge network structure. Therefore, the edge network based on the randomness of the attacker needs further research. In addition, in order to ensure the data security, the proposed algorithm uses an encryption algorithm and game algorithm at the same time, and the structure is relatively complex. The next research will focus on the design of a data security method which takes into account the security, lightweight, and suitability for the power distribution IoT.

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 no conflicts of interest.

References

[1] M. Gusev and S. Dustdar, "Going back to the roots—the evolution of edge computing, an IoT perspective," *IEEE Internet Computing*, vol. 22, no. 2, pp. 5–15, 2018.

[2] Z. Li, M. Shahidehpour, and X. Liu, "Cyber-secure decentralized energy management for IoT-enabled active distribution networks," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 5, pp. 900–917, 2018.

[3] H. Fullmer, "Healthcare power systems may be unprepared for digital age," *Electrical Contractor*, vol. 83, no. 1, pp. 13–13, 2018.

[4] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: deep learning for the internet of things with edge computing," *IEEE Network*, vol. 32, no. 1, pp. 96–101, 2018.

[5] R. Morabito, V. Cozzolino, A. Y. Ding, N. Beijar, and J. Ott, "Consolidate IoT edge computing with lightweight virtualization," *IEEE Network*, vol. 32, no. 1, pp. 102–111, 2018.

[6] R. Dautov, S. Distefano, D. Bruneo et al., "Metropolitan intelligent surveillance systems for urban areas by harnessing IoT and edge computing paradigms," *Software: Practice and Experience*, vol. 48, no. 8, pp. 1475–1492, 2018.

[7] T. Ogino, S. Kitagami, T. Suganuma, and N. Shiratori, "A multi-agent based flexible IoT edge computing architecture harmonizing its control with cloud computing," *International Journal of Networking and Computing*, vol. 8, no. 2, pp. 218–239, 2018.

[8] F. Ud Din, A. Ahmad, H. Ullah, A. Khan, T. Umer, and S. Wan, "Efficient sizing and placement of distributed generators in cyber-physical power systems," *Journal of Systems Architecture*, vol. 97, pp. 197–207, 2019.

[9] X. Xu, Q. Liu, Y. Luo et al., "A computation offloading method over big data for IoT-enabled cloud-edge computing," *Future Generation Computer Systems*, vol. 95, no. 6, pp. 522–533, 2019.

[10] C. Pan, M. Xie, and J. Hu, "ENZYME: an energy-efficient transient computing paradigm for ultralow self-powered IoT edge devices," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2440–2450, 2018.

[11] K. Peng, H. Huang, S. Wan, and V. C. M. Leung, "End-edge-cloud collaborative computation offloading for multiple mobile users in heterogeneous edge-server environment," *Wireless Networks*, vol. 7, no. 4, pp. 2622–2629, 2020.

[12] T. Suganuma, T. Oide, S. Kitagami, K. Sugawara, and N. Shiratori, "Multiagent-based flexible edge computing architecture for IoT," *IEEE Network*, vol. 32, no. 1, pp. 16–23, 2018.

[13] L. Lei, H. Xu, X. Xiong, K. Zheng, W. Xiang, and X. Wang, "Multiuser resource control with deep reinforcement learning in IoT edge computing," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10119–10133, 2019.

[14] T. Ogino, S. Kitagami, and N. Shiratori, "A multi-agent based flexible IoT edge computing architecture and application to ITS," *Journal of Communications*, vol. 14, no. 1, pp. 47–52, 2019.

[15] J. Xue, M. Li, and J. Luo, "Modeling Method for Coupling Relations in Cyber Physical Power Systems Based on Correlation Characteristic Matrix[J]," *Dianli Xitong Zidonghua/Automation of Electric Power Systems*, vol. 42, no. 2, pp. 11–19, 2018.

[16] X. Liu, J. Yu, J. Wang, and Y. Gao, "Resource allocation with edge computing in IoT networks via machine learning," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3415–3426, 2020.

[17] J. Ni, X. Lin, and X. S. Shen, "Toward edge-assisted internet of things: from security and efficiency perspectives," *IEEE Network*, vol. 33, no. 2, pp. 50–57, 2019.

[18] Y. Guo, F. Liu, Z. Cai, N. Xiao, and Z. Zhao, "Edge-based efficient search over encrypted data mobile cloud storage," *Sensors*, vol. 18, no. 4, pp. 1189–1203, 2018.

[19] X. Kong, Y. Xu, Z. Jiao, D. Dong, X. Yuan, and S. Li, "Fault location technology for power system based on information about the power Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 6682–6692, 2020.

[20] W. Han and Y. Xiao, "Edge computing enabled non-technical loss fraud detection for big data security analytic in Smart Grid," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 4, pp. 1697–1708, 2020.

[21] Z. Lv and H. Song, "Mobile internet of things under data physical fusion technology," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4616–4624, 2020.

[22] Z. Guan, Y. Zhang, G. Si et al., "ECOSECURITY: tackling challenges related to data exchange and security: an edge-computing-enabled secure and efficient data exchange architecture for the energy internet," *IEEE Consumer Electronics Magazine*, vol. 8, no. 2, pp. 61–65, 2019.

[23] C. A. Pardue, M. L. F. Bellaredj, H. M. Torun, M. Swaminathan, P. Kohl, and A. K. Davis, "RF wireless power transfer using integrated inductor," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 9, no. 5, pp. 913–920, 2019.

[24] J. Cui, L. Wei, H. Zhong, J. Zhang, Y. Xu, and L. Liu, "Edge computing in VANETs-an efficient and privacy-preserving cooperative downloading scheme," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 6, pp. 1191–1204, 2020.

[25] H. Xiong, Y. Zhao, L. Peng, H. Zhang, and K.-H. Yeh, "Partially policy-hidden attribute-based broadcast encryption with secure delegation in edge computing," *Future Generation Computer Systems*, vol. 97, pp. 453–461, 2019.

[26] H. Liu, Y. Zhang, and T. Yang, "Blockchain-enabled security in electric vehicles cloud and edge computing," *IEEE Network*, vol. 32, no. 3, pp. 78–83, 2018.

[27] S. Kim, K. J. Han, Y. Kim, and S. Kang, "Power integrity coanalysis methodology for multi-domain high-speed memory systems," *IEEE Access*, vol. 7, no. 99, pp. 95305–95313, 2019.

[28] T. Zhuang, M. Ren, X. Gao, M. Dong, W. Huang, and C. Zhang, "Insulation condition monitoring in distribution power grid via IoT-based sensing network," *IEEE Transactions on Power Delivery*, vol. 34, no. 4, pp. 1706–1714, 2019.

[29] C. Fu, C. Peng, X.-Y. Liu, L. T. Yang, J. Yang, and L. Han, "Search engine: the social relationship driving power of Internet of Things," *Future Generation Computer Systems*, vol. 92, pp. 972–986, 2019.

[30] J. Fei and M. Xiaoping, "Fog computing perception mechanism based on throughput rate constraint in intelligent Internet of Things," *Personal and Ubiquitous Computing*, vol. 23, no. 3-4, pp. 563–571, 2019.

[31] S. Hajiheidari, K. Wakil, M. Badri, and N. J. Navimipour, "Intrusion detection systems in the Internet of Things: a comprehensive investigation," *Computer Networks*, vol. 160, no. 9, pp. 165–191, 2019.

[32] M. H. Eldefrawy, N. Pereira, and M. Gidlund, "Key distribution protocol for industrial Internet of Things without implicit certificates," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 906–917, 2018.

[33] Y. Yang, Z. Zheng, K. Bian, L. Song, and Z. Han, "Real-time profiling of fine-grained air quality index distribution using UAV sensing," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 186–198, 2018.

[34] F. Tong, Y. Sun, and S. He, "On positioning performance for the narrow-band internet of things: how participating eNBs impact?," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 423–433, 2019.

[35] D. B. Avancini, J. J. P. C. Rodrigues, S. G. B. Martins, R. A. L. Rabêlo, J. al-Muhtadi, and P. Solic, "Energy meters evolution in smart grids: a review," *Journal of Cleaner Production*, vol. 217, no. 4, pp. 702–715, 2019.

[36] T. M. Fernández-Caramés, "From pre-quantum to post-quantum IoT security: a survey on quantum-resistant cryptosystems for the Internet of Things," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6457–6480, 2020.

[37] H. Ibrahim, W. Bao, and U. T. Nguyen, "Data rate utility analysis for uplink two-hop internet of things networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3601–3619, 2019.