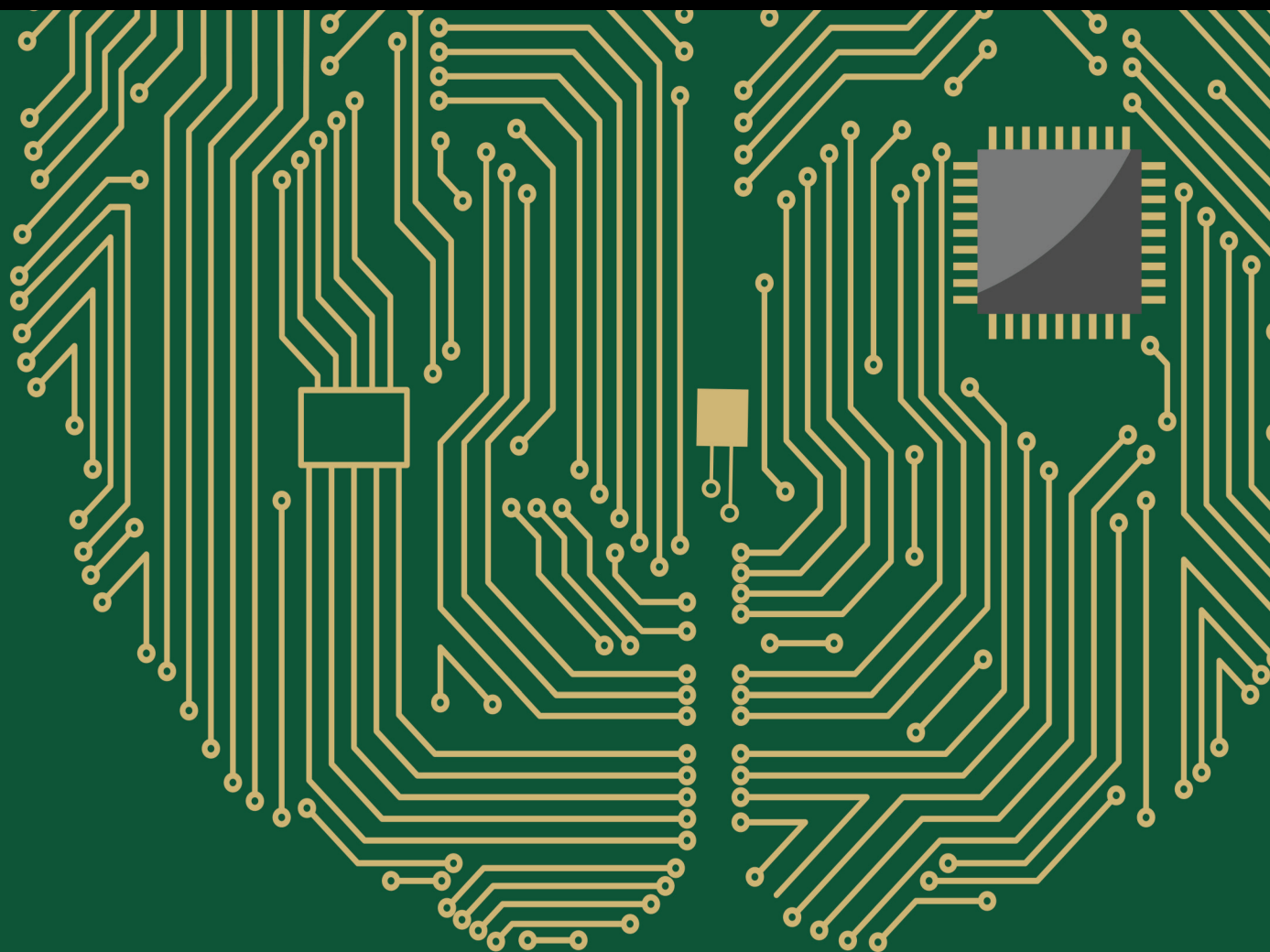


Multitask Deep Learning and Semantic Knowledge in Intelligent Healthcare

Lead Guest Editor: Farman Ali

Guest Editors: Tamer Abuhmed, Shaker El Sappagh, and Muhammad Imran





Multitask Deep Learning and Semantic Knowledge in Intelligent Healthcare

Computational Intelligence and Neuroscience

Multitask Deep Learning and Semantic Knowledge in Intelligent Healthcare

Lead Guest Editor: Farman Ali

Guest Editors: Tamer Abuhmed, Shaker EI
Sappagh, and Muhammad Imran



Copyright © 2023 Hindawi Limited. All rights reserved.

This is a special issue published in "Computational Intelligence and Neuroscience." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chief Editor

Andrzej Cichocki, Poland

Associate Editors

Arnaud Delorme, France
Cheng-Jian Lin , Taiwan
Saeid Sanei, United Kingdom

Academic Editors

Mohamed Abd Elaziz , Egypt
Tariq Ahanger , Saudi Arabia
Muhammad Ahmad, Pakistan
Ricardo Aler , Spain
Nouman Ali, Pakistan
Pietro Aricò , Italy
Lerina Aversano , Italy
Ümit Ağbulut , Turkey
Najib Ben Aoun , Saudi Arabia
Surbhi Bhatia , Saudi Arabia
Daniele Bibbo , Italy
Vince D. Calhoun , USA
Francesco Camastra, Italy
Zhicheng Cao, China
Hubert Cecotti , USA
Jyotir Moy Chatterjee , Nepal
Rupesh Chikara, USA
Marta Cimitile, Italy
Silvia Conforto , Italy
Paolo Crippa , Italy
Christian W. Dawson, United Kingdom
Carmen De Maio , Italy
Thomas DeMarse , USA
Maria Jose Del Jesus, Spain
Arnaud Delorme , France
Anastasios D. Doulamis, Greece
António Dourado , Portugal
Sheng Du , China
Said El Kafhali , Morocco
Mohammad Reza Feizi Derakhshi , Iran
Quanxi Feng, China
Zhong-kai Feng, China
Steven L. Fernandes, USA
Agostino Forestiero , Italy
Piotr Franaszczuk , USA
Thippa Reddy Gadekallu , India
Paolo Gastaldo , Italy
Samanwoy Ghosh-Dastidar, USA

Manuel Graña , Spain
Alberto Guillén , Spain
Gaurav Gupta, India
Rodolfo E. Haber , Spain
Usman Habib , Pakistan
Anandakumar Haldorai , India
José Alfredo Hernández-Pérez , Mexico
Luis Javier Herrera , Spain
Alexander Hošovský , Slovakia
Etienne Hugues, USA
Nadeem Iqbal , Pakistan
Sajad Jafari, Iran
Abdul Rehman Javed , Pakistan
Jing Jin , China
Li Jin, United Kingdom
Kanak Kalita, India
Ryotaro Kamimura , Japan
Pasi A. Karjalainen , Finland
Anitha Karthikeyan, Saint Vincent and the Grenadines
Elpida Keravnou , Cyprus
Asif Irshad Khan , Saudi Arabia
Muhammad Adnan Khan , Republic of Korea
Abbas Khosravi, Australia
Tai-hoon Kim, Republic of Korea
Li-Wei Ko , Taiwan
Raşit Köker , Turkey
Deepika Koundal , India
Sunil Kumar , India
Fabio La Foresta, Italy
Kuruva Lakshmana , India
Maciej Lawrynczuk , Poland
Jianli Liu , China
Giosuè Lo Bosco , Italy
Andrea Loddo , Italy
Kezhi Mao, Singapore
Paolo Massobrio , Italy
Gerard McKee, Nigeria
Mohit Mittal , France
Paulo Moura Oliveira , Portugal
Debajyoti Mukhopadhyay , India
Xin Ning , China
Nasimul Noman , Australia
Fivos Panetsos , Spain

Evgeniya Pankratova , Russia
Rocío Pérez de Prado , Spain
Francesco Pistolesi , Italy
Alessandro Sebastian Podda , Italy
David M Powers, Australia
Radu-Emil Precup, Romania
Lorenzo Putzu, Italy
S P Raja, India
Dr.Anand Singh Rajawat , India
Simone Ranaldi , Italy
Upaka Rathnayake, Sri Lanka
Navid Razmjooy, Iran
Carlo Ricciardi, Italy
Jatinderkumar R. Saini , India
Sandhya Samarasinghe , New Zealand
Friedhelm Schwenker, Germany
Mijanur Rahaman Seikh, India
Tapan Senapati , China
Mohammed Shuaib , Malaysia
Kamran Siddique , USA
Gaurav Singal, India
Akansha Singh , India
Chiranjibi Sitaula , Australia
Neelakandan Subramani, India
Le Sun, China
Rawia Tahrir , Iraq
Binhua Tang , China
Carlos M. Travieso-González , Spain
Vinh Truong Hoang , Vietnam
Fath U Min Ullah , Republic of Korea
Pablo Varona , Spain
Roberto A. Vazquez , Mexico
Mario Versaci, Italy
Gennaro Vessio , Italy
Ivan Volosyak , Germany
Leyi Wei , China
Jianghui Wen, China
Lingwei Xu , China
Cornelio Yáñez-Márquez, Mexico
Zaher Mundher Yaseen, Iraq
Yugen Yi , China
Qiangqiang Yuan , China
Miaolei Zhou , China
Michal Zochowski, USA
Rodolfo Zunino, Italy




Contents

Retracted: An Intelligent Classification System for Cancer Detection Based on DNA Methylation Using ML and Semantic Knowledge in Healthcare

Computational Intelligence and Neuroscience

Retraction (1 page), Article ID 9797060, Volume 2023 (2023)

Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease

Hira Khalid, Ajab Khan , Muhammad Zahid Khan , Gulzar Mehmood , and Muhammad Shuaib Qureshi 










Research Article (14 pages), Article ID 9266889, Volume 2023 (2023)

Next Generation Infectious Diseases Monitoring Gages via Incremental Federated Learning: Current Trends and Future Possibilities

Iqra Javed, Uzair Iqbal , Muhammad Bilal, Basit Shahzad, Tae-Sun Chung , and Muhammad Attique 






Review Article (12 pages), Article ID 1102715, Volume 2023 (2023)

A Multimodal Network Security Framework for Healthcare Based on Deep Learning

Qiang Qiang Chen , Jian Ping Li , Amin ul Haq , Bless Lord Y. Agbley , Arif Hussain , Inayat Khan , Riaz Ullah Khan , Jalaluddin Khan , and Ijaz Ali 





Research Article (18 pages), Article ID 9041355, Volume 2023 (2023)

[Retracted] An Intelligent Classification System for Cancer Detection Based on DNA Methylation Using ML and Semantic Knowledge in Healthcare

Anuradha Thakare , Manisha Bhende , Mulugeta Tesema , Mohammed Dighiri , R. Bhavani , and Amena Mahmoud 

Research Article (9 pages), Article ID 4334852, Volume 2022 (2022)

Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods

Zahid Ullah , Farrukh Saleem , Mona Jamjoom , Bahjat Fakieh , Faris Kateb , Abdullah Marish Ali , and Babar Shah 

Research Article (10 pages), Article ID 2557795, Volume 2022 (2022)

Blockchain-Based Optimization Model for Evaluating Psychological Mental Disease and Mental Fitness

Jayashree Rajesh Prasad , Shashikant V. Athawale , Roshani Raut , Sonali Patil , Sheetal U. Bhandari , and Mohd Asif Shah 

Research Article (9 pages), Article ID 8657313, Volume 2022 (2022)

Retraction

Retracted: An Intelligent Classification System for Cancer Detection Based on DNA Methylation Using ML and Semantic Knowledge in Healthcare

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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

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

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

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

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

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

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

References

- [1] A. Thakare, M. Bhende, M. Tesema, M. Dighriri, R. Bhavani, and A. Mahmoud, "An Intelligent Classification System for Cancer Detection Based on DNA Methylation Using ML and Semantic Knowledge in Healthcare," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 4334852, 9 pages, 2022.

Research Article

Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease

Hira Khalid,¹ Ajab Khan ,¹ Muhammad Zahid Khan ,² Gulzar Mehmood ,³
and Muhammad Shuaib Qureshi ⁴

¹Department of Information Technology, Abbottabad University of Science and Technology, Havelian 22500, Abbottabad, Pakistan

²Department of Computer Science and I.T, Network Systems and Security Research Group, University of Malakand, Chakdara 18800, Khyber Pakhtunkhwa, Pakistan

³Department of Computer Science, IQRA National University, Swat Campus 19220, Peshawar, Pakistan

⁴Department of Computer Science, School of Arts and Sciences, University of Central Asia, Bishkek, Kyrgyzstan

Correspondence should be addressed to Muhammad Shuaib Qureshi; muhhammad.queshi@ucentralasia.org

Received 25 July 2022; Revised 6 September 2022; Accepted 19 September 2022; Published 14 March 2023

Academic Editor: Farman Ali

Copyright © 2023 Hira Khalid et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To diagnose an illness in healthcare, doctors typically conduct physical exams and review the patient's medical history, followed by diagnostic tests and procedures to determine the underlying cause of symptoms. Chronic kidney disease (CKD) is currently the leading cause of death, with a rapidly increasing number of patients, resulting in 1.7 million deaths annually. While various diagnostic methods are available, this study utilizes machine learning due to its high accuracy. In this study, we have used the hybrid technique to build our proposed model. In our proposed model, we have used the Pearson correlation for feature selection. In the first step, the best models were selected on the basis of critical literature analysis. In the second step, the combination of these models is used in our proposed hybrid model. Gaussian Naïve Bayes, gradient boosting, and decision tree classifier are used as a base classifier, and the random forest classifier is used as a meta-classifier in the proposed hybrid model. The objective of this study is to evaluate the best machine learning classification techniques and identify the best-used machine learning classifier in terms of accuracy. This provides a solution for overfitting and achieves the highest accuracy. It also highlights some of the challenges that affect the result of better performance. In this study, we critically review the existing available machine learning classification techniques. We evaluate in terms of accuracy, and a comprehensive analytical evaluation of the related work is presented with a tabular system. In implementation, we have used the top four models and built a hybrid model using UCI chronic kidney disease dataset for prediction. Gradient boosting achieves around 99% accuracy, random forest achieves 98%, decision tree classifier achieves 96% accuracy, and our proposed hybrid model performs best getting 100% accuracy on the same dataset. Some of the main machine learning algorithms used to predict the occurrence of CKD are Naïve Bayes, decision tree, K-nearest neighbor, random forest, support vector machine, LDA, GB, and neural network. In this study, we apply GB (gradient boosting), Gaussian Naïve Bayes, and decision tree along with random forest on the same set of features and compare the accuracy score.

1. Introduction

Nowadays, chronic kidney disease (CKD) is a rapidly growing disease, and millions of people die due to lack of timely affordable treatment. Chronic kidney disease patients belong to low-class and middle-class income-generating countries [1, 2].

In 2013, about one million people died due to chronic kidney disease [3]. The developing world suffers more from the chronic kidney disease, and low to average income countries contain a total of 387.5 million CKD patients where 177.4 million patients are male and 210.1 million patients are female [4]. These figures show that a large number of people in developing countries suffer from

chronic kidney disease, and this ratio is increasing day by day. A lot of work has been done for the early diagnosis of chronic kidney disease so that the disease could be treated at an early stage. In this article, we are focusing on machine learning prediction models for chronic kidney disease and giving importance to accuracy.

Chronic kidney disease is a common type of kidney disease that occurs when both kidneys are damaged, and the CKD patients suffer from this condition for a long term. Here, the term kidney damage means any kidney condition that can cause improper functioning of the kidney. This could be caused by any disorder or due to lack of essentials like the glomerular filtration rate (GFR) reduction [5]. Our proposed prediction model takes the clinical symptoms as input and predicts the results using the stacking classifier with the random forest algorithm as a base classifier.

Machine learning is gaining significance in healthcare diagnosis as it enables intricate analysis, thereby minimizing human errors and enhancing the precision of predictions. Machine learning algorithms and classifiers are now considered the most reliable techniques for the diagnosis of different diseases like heart disease, diabetes, tumors disease, and liver disease predictions [6].

Different machine learning algorithms used the Naïve Bayes, SVM, and the decision tree for the classification purpose, while random forest, logistic regression, and linear regression were used for the regression purpose in the medical fields for the prediction. With the efficient use of these algorithms, the death rate can be minimized due to early-stage diagnosis and patients can be treated timely. Along with maintaining the clinical symptoms, chronic kidney disease patients should include physical activities in daily life. They should exercise, drink water, and avoid junk food. The common symptoms of chronic kidney disease are shown in Figure 1.

This article delivers an overview and analysis subsequently followed by an implementation and evaluation of the machine learning classifiers used in CKD diagnosis. Further, this article discusses the importance of machine learning classifiers in healthcare and explains how these can make more accurate predictions. Figure 2 represents the block diagram of the chronic kidney disease prediction model.

The core objective of this article is to propose and implement a hybrid machine learning prediction model for chronic kidney disease where due importance is given to accuracy. In this article, we have analyzed the accuracy of same dataset with respect to different machine learning algorithms and compared their accuracy score so as to get a better model. Our focus remains on the solution of overfitting problem using cross-validation while achieving the highest accuracy to build a best hybrid model from the combination of available popular machine learning classifiers such as decision tree, gradient boosting, Gaussian Naïve Bayes, and gradient boosting. The ultimate goal is to deliver an accurate and effective treatment to CKD patients at a reduced cost. Before we proceed further, we need to know little more about common diseases of the kidney. In Table 1,

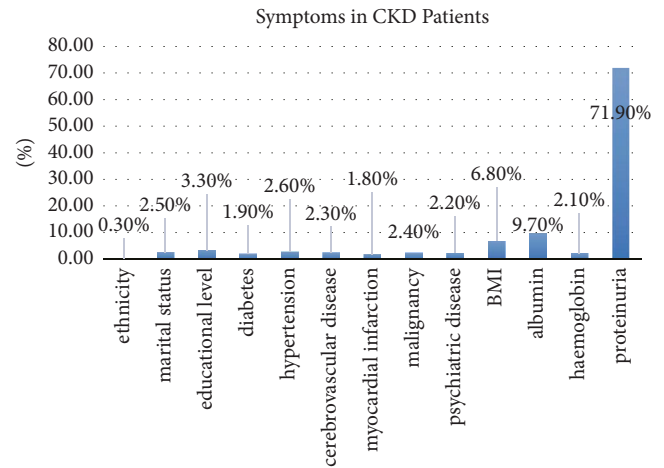


FIGURE 1: Symptoms in CKD patients [7].

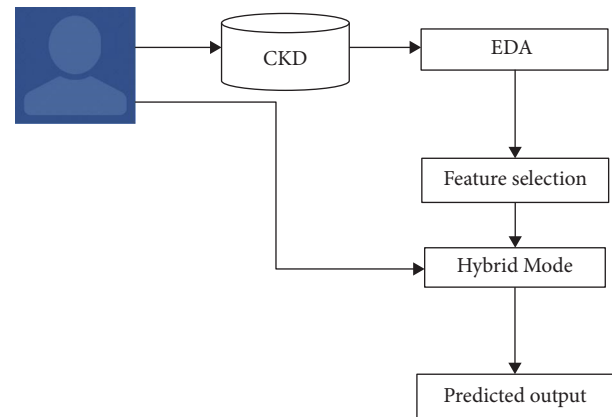


FIGURE 2: Block diagram of the machine learning hybrid model.

there is a list of some of the most common kidney diseases (Table 2).

The remaining portion of the article is organized as follows. Section 2 contains the literature survey along with the tabular comparison of the different machine learning algorithms used and an analysis of the results. Section 3 contains the proposed methodology. Section 4 contains the dataset details. Section 5 contains results and discussion. Section 6 contains conclusion and future work.

2. Literature Review

This section covers research work related to algorithms and assesses some algorithms based on their accuracy. In research work [7], the data mining technique applied to specific analysis of clinical records is a good method. The performance of the decision tree method was 91% (accuracy) compared to the Naïve Bayesian method. The classification algorithm for diabetes dataset had 94% specificity and 95% sensitivity. They also found that mining helps retrieve correlations of attributes that are no longer direct indicators of the type they are trying to predict. Similar work still needs to be done to improve the

TABLE 1: Description of common diseases of the kidney.

Diseases	Description
CKD	Chronic kidney disease (CKD) can occur when a disease or condition damages kidney function, causing kidney damage to deteriorate over a few months or years.
Kidney stones	Kidney stones (also called renal calculi) are hard pledges made of salts and minerals that form inside your kidney.
Glomerulonephritis	Glomerulonephritis causes infection and damage to the filtering part of the kidneys (glomerulus). It can occur quickly or could be over a longer period. Poisons, metabolic wastes, and surplus fluid are not properly strained into the urine. Instead, they build up in the body producing inflammation and fatigue.
Polycystic kidney disease	Polycystic kidney disease (PKD) is a genetic disorder that can produce many cysts filled with fluid and they grow inside your kidneys. Usually, they are harmless. The cysts can change the shape of the kidneys while making them much bigger.

TABLE 2: Equations for accuracy measurement.

S. no	Authors	Accuracy equations
1	Padmanaban and Parthiban [8]	Precision $i = TP_i / TP_i + FP_i$
2	Charleonnann et al. [9]	ACC = $(TP + TN) / (P + N)$
3	Ghosh et al. [7]	The results of performance degree indices are dependent on TP, TN, FP, and FN
4	Fu et al. [10]	Ext. values = points $> Q3 + 1.5$ (IQR) points $< Q1 - 1.5$ (IQR)
5	Devika et al. [11]	Accuracy = number of properly classified samples/total variety of samples
6	Revathy et al. [12]	Accuracy = $(TP + TN) / (TP + TN + FP + FN)$ Accuracy = $TP + TN / TP + TN + FP + FN$
7	Nishat et al. [14]	Accuracy = $(TP + TN) / (TP + TN + FP + FN)$ Accuracy = $TP + TN / TP + TN + FP + FN$
8	Rabby et al. [13]	Descriptive analysis of the data as well as the experimental results
9	Pouriyeh et al. [15]	Finding most significant feature using chi-square test
10	Jabbar et al. [16]	Experimental results only

True positive (TP) = list contains stated cases that are correctly categorized with CKD. False positive (FP) = list contains set that is inaccurately categorized with CKD. True negative (TN) = list contains stated instances that are correctly categorized with CKD. False negative (FN) = list contains set of instances that are exactly categorized with CKD.

overall performance of prediction engine accuracy in the statistical analysis of neural networks and clustering algorithms.

In [8], the authors described the prediction models using machine learning techniques including K-nearest neighbor (KNN), support vector machine (SVM), logistic regression (LR), and decision tree classifiers for CKD prediction. From the experiment, it was concluded that the SVM classifier provides the highest accuracy, 98.3%. SVM has the absolute best sensitivity after training and testing performed with the proposed method. Therefore, according to this comparison, it could be concluded that an SVM classifier is used to predict persistent kidney disease.

In the paper [9], they chose four different algorithms and compared them to get an accurate expectation rate over the dataset. Unlike all approaches that were presented, they got the best results from the gradient boosting classifier. The models effectively achieve an accuracy rate of 99.80%, whereas AdaBoost and LDA achieve 97.91% at a low value. Also, the gradient boosting ML classifier takes much time to make the prediction compared to others and has a higher predictable value in both the curves (ROC and AUC). Hence, an accurate expectation undoubtedly depends on the preprocessing strategy, and the methods of preprocessing must be approached cautiously to precisely achieve recognized results.

In [7], the authors investigated the machine learning ability, which is supported by predictive analysis so as to predict CKD early. An experimental procedure was performed by considering a dataset of 400 cases collected by Apollo Hospitals India. In this article, two labels were used as output/targets in this hybrid model (i.e., patients having CKD and others who are healthy) and four different machine learning classifiers were implemented. On the comparison of these classifiers, the classification along with regression tree, and the RPART classification model, showed remarkably better results in terms of accuracy. They used the information gain quotient for excruciating criterion, and here the optimum spilling reduces the noise of the resulting feature subsets. In this study, the RPART limited value of criterion for the splitting was five, meaning that splits repeatedly occur for the five instances present in the leaf node. In addition, they identified an equivalent previous probability for the class attributes. Here, the RPART prediction model used seven terminal nodes for the earlier predictions of CKD. The experimental results showed that the highest AUC and TPR were obtained with the machine learning prediction model, whereas the highest TNR (1.00) was achieved with the model RPART. The RPART model could be described as a set of rules for making the decision. However, the major drawback of RPART is the consideration of the single factor as a parameter in every division

procedure, while considering different parameter combinations could result in better CKD predictions. However, the machine learning prediction model gives the lowest error rate. The major reason is that the MLP could adopt and handle complex predictions. The complex relationships require hidden nodes and they are useful as they allow neural networks to model between parameters while sometimes deal with nonlinearity in data. The overall results indicate that the algorithms of machine learning give an inspiring and a feasible methodology for earlier CKD prediction.

As we have already seen, there are different machine learning prediction models and learning programs available to assist practitioners. In [5], they used a new selection guide for predicting CKD. In this work, CKD is predicted by using specific classifiers and a reasonable study of overall performance. In this study, they performed the evaluation of the Naïve Bayes classifier, random forest, and artificial neural network classifiers and concluded that the random forest classifier performs better as compared to other classifiers. The worth of forecasting CKD has been progressive. Several sustainable evolutionary policies can be used to improve the outcomes of the suggested classifiers. Here, Naïve Bayes, random forest, and KNN were applied to predict CKD. Early diagnosis of CKD helps to treat those affected well in time and prevent the disease from progressing to worse stage. The early detection of this type of disease and well-timed treatment is one of the main objectives of the medical field.

In [10], a machine learning prediction model was developed for the early prediction of CKD. The dataset gives input features gathered from the CKD dataset and the models were tested and validated for the given input features. Machine learning decision tree classifier, random forest classifier, and support vector classifier were constructed for the diagnosis of CKD. The performance analysis of the models was assessed on the basis of the accuracy score of the prediction model. On comparison, the results of the research showed that the random forest classifier model performs much better at predicting CKD as compared to decision tree and support vector classifiers.

The kidneys play a vital role in maintaining the body's blood pressure, acid-base sense of balance, and electrolyte sense of balance, not only needed to filter toxins from the body. Malfunction is accountable for insignificant to mortal illnesses, in addition to dysfunction in the other body organs. Therefore, researchers all over the world have dedicated themselves for finding techniques to accurately diagnose and effectively treat chronic kidney disease. As machine learning classifiers are increasingly used in the medical field for diagnosis, now CKD is also included in the list of diseases that could be predicted using machine learning classifiers. The research to detect CKD with ML algorithms has enhanced the procedure and consequence accuracy progressively. They proposed the random forest classifier (99.75% accuracy) as the maximum efficient classifier among all other classifiers. The study demonstrates the effective handling of missing values in data through four techniques, namely, mode, mean, median, and zero-point methods. It also evaluates the performance of machine

learning models under two scenarios, with and without tuning the hyperparameters, and observes significant improvement in the classifiers' performance, which is visually presented through graphs [11].

Overall, the motive of the study is to examine the applicability of specific supervised machine learning classifiers in the field of bioinformatics and offer their compatibility in detecting several serious diseases such as the diagnosis of CKD at an early stage [12].

They built an updated and proficient machine learning (ML) application that can perceptually perceive and predict the state of chronic kidney disease. In this work, the ten most important machine learning methods for predicting permanent kidney disease were considered. The level of accuracy of the classification algorithm we used in our project is as good as we wanted.

For the prediction of disease, the first most essential step is to detect the disease that is costly in developing countries like Pakistan and Bangladesh. The people of these countries mostly suffer from this. Currently, CKD patient proportion is increasing rapidly in Pakistan and Bangladesh. So, in that article, the authors tried to develop a system that helps in predicting the risk of CKD. In the proposed model, they used and processed UCI datasets and real-time datasets and tried to deal with missing data and trained the model using random forest and ANN classifiers. Then, they implemented these two algorithms in the Python language. The accuracy they got with the random forest algorithm is 97.12% and that with ANN is 94.5%, which is relatively very good. By use of this proposed method, risk prediction of CKD at an early stage is possible.

In [13], the authors predicted CKD based on sugar levels, aluminum levels, and red blood cell percentage. In this perception, five classifiers were applied, namely, Naïve Bayes, logistic regression, decision table, random tree, and random forest, and for each classifier, the results were noted based on (i) without preprocessing, (ii) SMOTE with resampling, and (iii) class equalizer. Random forest classifier has been observed to give the highest accuracy at 98.93% in SMOTE with resampling.

2.1. Comparison of Machine Learning Classifiers for CKD.

In this section, a comprehensive comparison of the state of the art is presented in the form of a table. The evaluation is formed in the aspect of accuracy, which can be comprehended in Table 3. The table has eight features that are described below:

Author: this contains the names of the authors of each article along with the reference.

Year: this column provides the year of the paper's publication.

Input data: this column shows the type of dataset that was used as input for the machine learning classifiers.

Disease type: This section shows the type of disease that was predicted by using different classifiers. It shows the best classifier found in the research paper, which is the classifier with the maximum accuracy.

TABLE 3: Comparison of classifiers for CKD.

S. no	Authors	Year	Input data	Disease type	Tools	Classifiers	Cross-validation	Accuracy
1	Padmanaban and Parthiban [8]	2016	Diabetic patients UCI machine learning	CKD	WEKA, YALE	Naïve Bayes Decision tree	10 folds	86% 91%
2	Charleonnann et al. [9]	2016	Clinical data	CKD	WEKA, MATLAB	SVM Logistic regression Decision tree KNN	5 folds	98.3% 96.55% 94.81% 98.1%
3	Ghosh et al. [7]	2020	Apollo Hospitals India	CKD	Python	SVM AB LDA GB	5 folds	99.56% 97.91% 97.91% 99.80%
4	Fu et al.. [10]	2018	UCI repository (CKD dataset)	CKD	Python	RPART SVM LOGR MLP	No cross-validation	98.2% 97.3% 99.4% 99.5%
5	Devika et al. [11]	2019	UCI repository (CKD dataset)	Chronic renal disorder	C Sharp	Naïve Bayes KNN Random forest	No cross-validation	99.63% 87.78% 99.84%
6	Revathy et al. [12]	2019	UCI repository (CKD dataset)	CKD	Python	Decision tree SVM Random forest	No cross-validation	94.16% 98.33% 99.16%
7	Nishat et al. [14]	2021	Learning repository of University of California, Irvine	CKD	Python	CNN LR DT RF SVM NB MLP QDA	No cross-validation	78% 98.25% 99% 99.75% 85% 96.5% 81.25% 37.5%
8	Rabby et al. [13]	2019	UCI repository (CKD dataset)	CKD	Python	K-nearest neighbor RF SVM GNB AB DT LDA GB LR ANN	No cross-validation	71.25% 98.75% 97.50 100% 98.75% 100% 97.50% 98.75 97.50% 65%
9	Pouriyeh et al. [15]	2020	UCI repository (CKD dataset)	CKD	Python	RF ANN	10 folds	97.12% 94.5%

TABLE 3: Continued.

S. no	Authors	Year	Input data	Disease type	Tools	Classifiers	Cross-validation	Accuracy
10	Jabber et al. [16]	2020	UCI repository (CKD dataset)	CKD	Python	Decision tree Logistic regression Naïve Bayes Random forest	No cross-validation	96.79% 97.86% 97.33% 98.93%
11	Bmc [17]	2013	UCI repository	Diabetic kidney disease	MATLAB	SVM PLS FFNN RPART Random forest Naïve Bayes C5.0	No cross-validation	0.91 0.83 0.85 0.87 0.91 0.86 0.90
12	Ramya and Radha [18]	2016	UCI repository	Chronic kidney disease	R	BP RBF Random forest (RF)	No cross-validation	80.4 85.3 78.6
13	Kumar [19]	2016	UCI repository	CKD	MATLAB	RF SMO Naïve Bayes RBF MLPC SLG	No cross-validation	95.67 90 87.64 83.78 89 87
14	Basarslan and Kayaalp [20]	2019	UCI repository	Chronic kidney disease	MATLAB	K-nearest neighbor Naïve Bayes LR RF	No cross-validation	97 96.5 97.56 99
15	Dowluru and Rayavarapu [21]	2012	UCI repository	Kidney stone	WEKA tool Orange tool	Naïve Bayes classification Logistic regression J48 algorithm Random forest Naïve Bayes KNN Classification tree C4.5 SVM Random forest	No cross-validation	0.99 1.00 0.97 0.98 0.79 0.7377 0.9352 0.9352 0.9198 0.9352

Bold values represent the highest accuracy in the relevant paper.

Classifiers: this column signifies the different machine learning classifiers that were used in the research and the comparison between them.

Tool: The column represents the programming language or the framework that was used in building the model. The researchers used these tools to preprocess the input data, then create a prediction model, and finally go to the testing stage.

Cross-validation: this column gives information about the validation of the classifiers and makes a comparison of different research papers regarding folds of cross-validation used.

Accuracy: The accuracy of the outcomes of the recommended model is represented in this column. If the article crisscrosses a comparison, the accuracy column only contains the accuracy percent of the best classifier confirmed by the author.

2.2. ML Classifier with Highest Accuracy. The machine learning algorithms that we analyzed from the above literature are listed in Table 4 and Figure 3.

3. Proposed Methodology

The proposed hybrid model is implemented in Python with pandas, sklearn, Matplotlib, Plotly, and other essential libraries. We have downloaded the CKD dataset from the UCI repository. The dataset contains two groups (CKD represented by 1 and non-CKD represented by 0) of chronic kidney disease in the downloaded information. The machine learning algorithm that has best accuracy is selected for analysis and implementation so that repeated results are produced. We have also developed a hybrid model based on knowledge that we gained during the analysis and implementation. The hybrid model consists of Gaussian Naïve Bayes, gradient boosting, and decision tree as base classifiers and random forest as a meta classifier. We have selected the tree-based machine learning algorithms for achieving the highest accuracy, while at the same time, it can handle the overfitting problem. In this paper, we detect the outliers with the violin plot as shown in Figure 4. As a solution of this problem, we implement the k -fold technique and design our model in such a way that it can reduce the problem of overfitting along with achieving the highest accuracy. The classifiers are discussed as under.

3.1. Naïve Bayes (NB). The NB classifier is related to the group of probabilistic classifiers and is constructed on the basis of the Naïve Bayes (NB) theorem. It takes up vigorous independence between the component's/features, and it contains the most crucial part of how this classifier creates forecasts. It can be built easily and is appropriately used in the medical field for the prediction of different diseases [15].

3.2. Decision Tree (DT). The decision tree classifier has a tree-like configuration or flowchart-like construction. It consists of subdivisions, leaves/child nodes, and a root/parent node. Here inner nodes comprise the features,

TABLE 4: Machine learning algorithms and classifiers.

Articles	Classifiers	Highest accuracy (%)
1	Decision tree	91
2	SVM	98.3
3	GB	99.80
4	MLP	99.5
5	Random forest	99.84
6	Random forest	99.16
7	Random forest	99.75
8	GNB Decision tree	100 100
9	Random forest	97.12
10	Random forest	98.93

Bold values represent the highest accuracy in the literature.

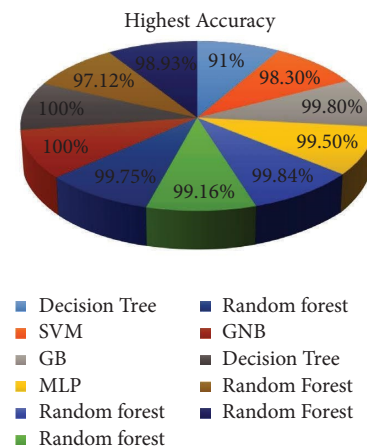
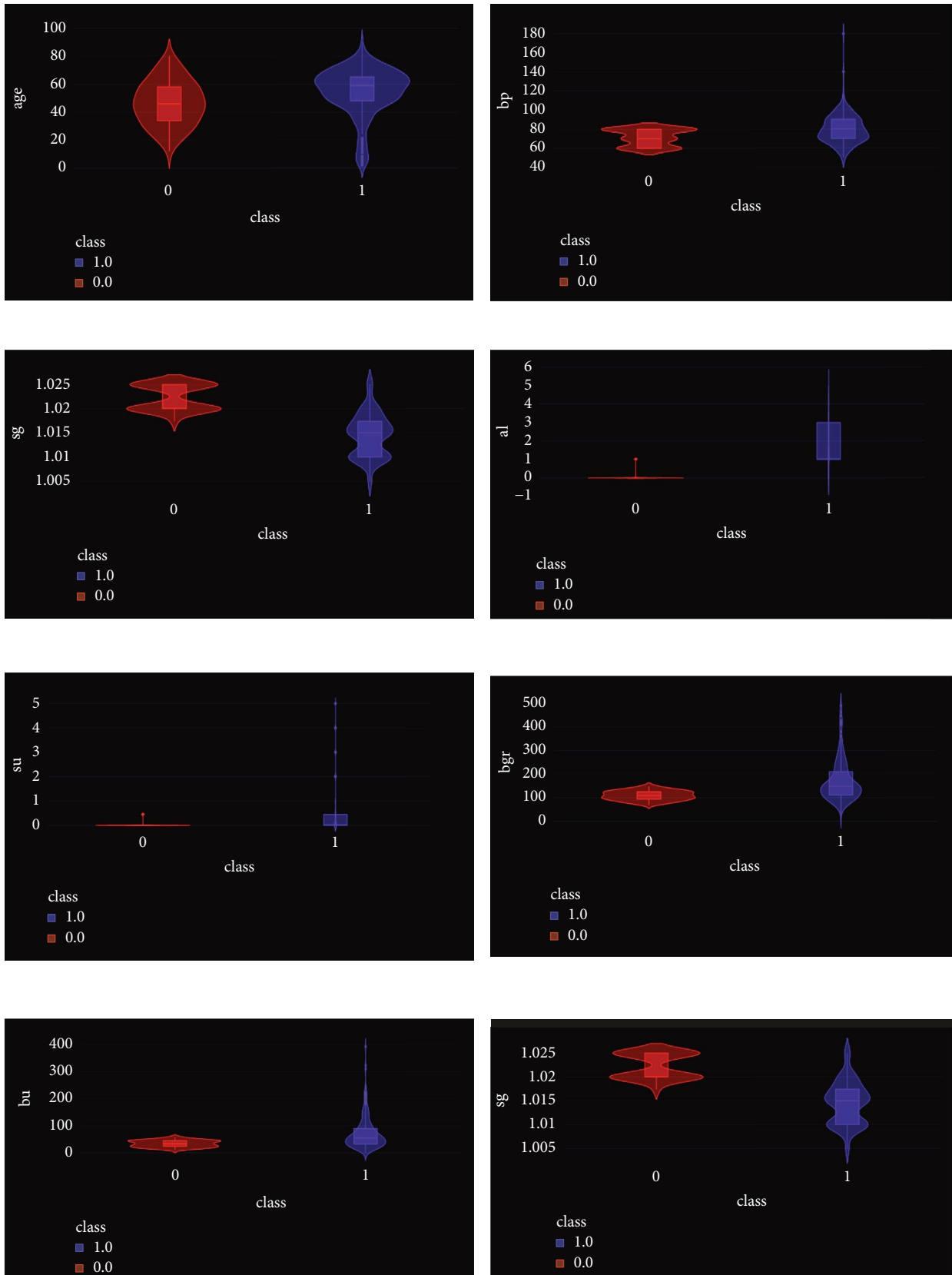


FIGURE 3: Comparison of machine learning classifiers.

whereas the subdivisions epitomize the outcome of every check on every node. Decision tree is one of the commonly used classifiers for classification determination because it does not need abundant information in the field or place constraints for it to work [15].

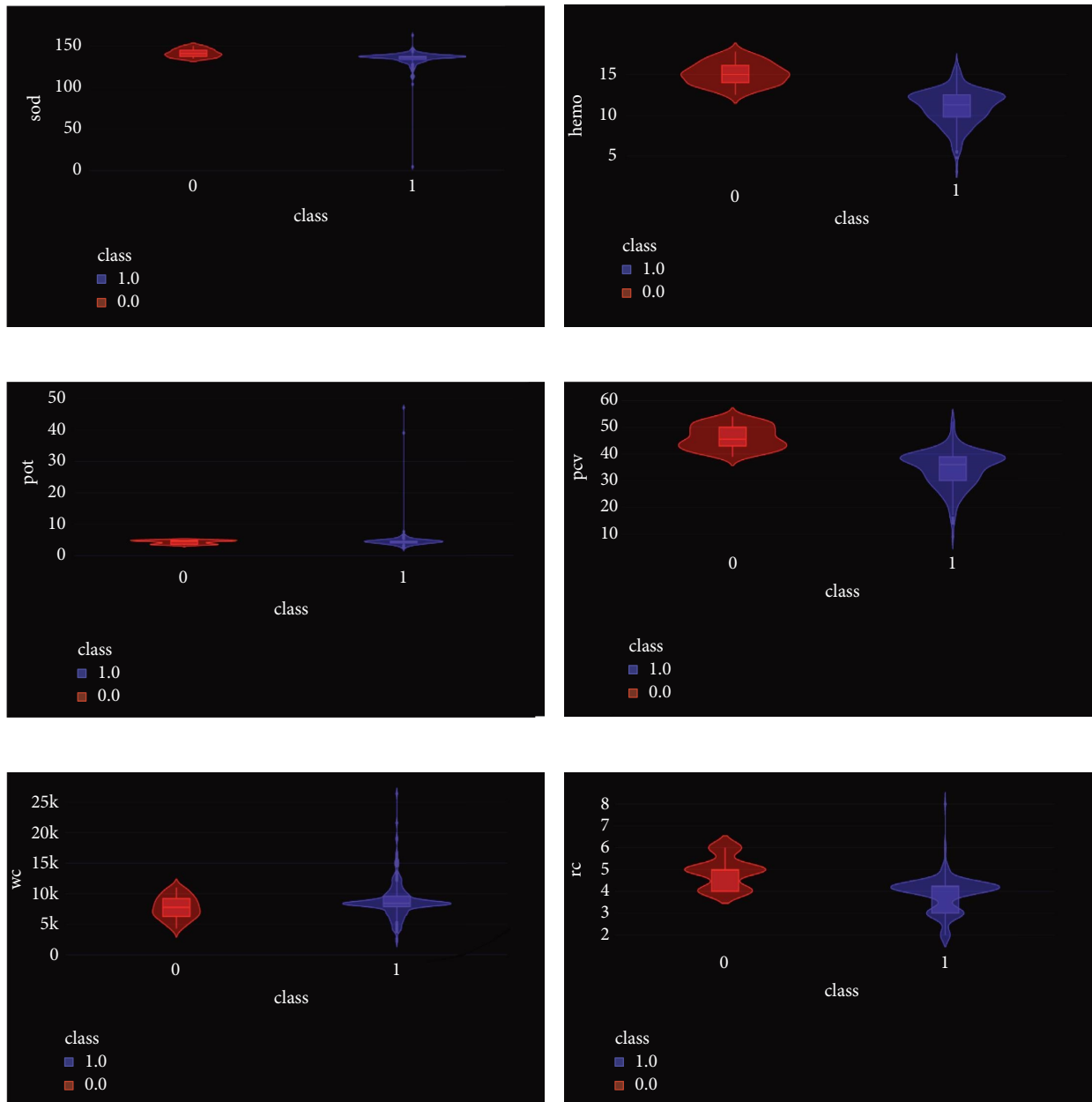
3.3. Random Forest (RF). In the ensemble and stacking classification approach, the random forest (RF) is the most effective algorithm among the other machine learning algorithms. In prediction and probability estimations, random forest (RF) algorithm has been used. Random forest (RF) classifier consists of many decision trees. Tin Kam Ho of Bell Labs introduced the concept of random forest in 1995, where each decision tree casts a vote to determine the object's class. The RF method is the combination of both bagging and random selection of attributes. Random forest classifier has the three hyperparameter tuning values [16].

- (i) Number of decision trees (n tree) used by the random forest classifier
- (ii) Size of the minimum node in the trees
- (iii) Number of attributes employed in splitting every node for every tree (m try). Here, m is the number of attributes.



(a)

FIGURE 4: Continued.



(b)

FIGURE 4: Violin plot of attributes.

Some of the advantages of the random forest classifier are listed as follows.

- (i) For ensemble learning algorithms, the random forest is the most appropriate choice
- (ii) For large datasets, random forest classifier performs well
- (iii) Random forest (RF) is able to handle hundreds of input attributes
- (iv) Random forest can estimate which attributes are more important in classification

(v) Missing value can be handled by using random forest classifier

(vi) Random forest handles the balancing error for class in unbalanced datasets

3.4. Gaussian Naïve Bayes (GNB). Gaussian Naïve Bayes (GNB) calculated the mean and standard deviation of each attribute at the training stage. To calculate the probabilities for the test data, mean and standard deviation were used. Due to this reason, some values of attributes are too big or too small from the value of the mean calculated. It affects the classifier

performance when testing data patterns have those attribute values and gives sometimes wrong output labels [22].

3.5. Hybrid Model. We use the concept of stacking for our hybrid model. As a type of ensemble technique in stacking, multiple classification models were combined with a main/meta classifier. One after the other, multiple layers were placed, where the models pass their predictions, and the upper most layer model makes decisions on the base of the combination of different models as a base model. The models in the low layer get attributes as input from the original data. The topmost layer of the model gets output from the lower layers and gives the results as a final prediction. The stacking technique involves using multiple independent machine-learning models as input to process the original data. After that, the meta classifier is used to predict the input along with the output of each machine learning model and individual algorithm's weights are estimated. The algorithms that are performing best are selected, and others having low performance are removed. In this technique, multiple classifiers as base model are combined and then, by using different machine learning algorithms, are trained on the same dataset through the use of a meta-classifier [23]. Figure 5 shows the flow diagram for the proposed hybrid model.

The execution of the model with the sequence of the steps is given below:

- (i) Collect the data of CKD from UCI repository
- (ii) Exploratory data analysis (EDA) is performed on that dataset
- (iii) This dataset is split into two parts: test data and train data
- (iv) Apply the cross-validation of 10 folds
- (v) Train the base models Gaussian Naïve Bayes, gradient boosting, and decision tree with the train set giving the predictions as M1, M2, and M3, respectively
- (vi) The output of the base models M1, M2, and M3 and test set data serve as input for random forest as input for training
- (vii) Once the random forest gets trained, it gives the prediction on the basis of training dataset and the output predictions of the base models

In this study, we have considered the UCI CKD dataset, and this dataset is split into two parts. 80% of data is used for training purposes as an input to the machine learning algorithms. We exploited the Gaussian Naïve Bayes, gradient boosting, decision tree, and stacking classifier with random forest algorithm which was used to predict the chronic kidney disease for 20% test data as input and plotted the predicted values and compared their values. Our proposed methodology has the following advantages.

- (i) We implemented four machine learning algorithms that are decision tree, gradient boosting, Gaussian Naïve Bayes, and random forest. We applied stacking classifiers to build the hybrid model that combines these four algorithms.

- (ii) We analyzed the accuracy of the same dataset with respect to different machine learning algorithms and compared their accuracy score to get the best model

- (iii) We implemented a stacking classifier technique to build a new model with improved accuracy

4. Dataset Details

We selected 14 attributes from the dataset that we are using from the UCI repository dataset of chronic kidney disease as input features as shown in Table 5 where age attribute shows the patient's age, bp indicates the blood pressure, sg indicates the specific gravity of the urine, al indicates the level of aluminum in the patient urine, bgr (blood glucose random) indicates the blood sugar level glucose tolerance, su represents the sugar level, bu indicates the blood urea, sod indicates the amount of sodium, sc indicates the serum creatinine, pot indicates the amount of potassium, hemo indicates the hemoglobin, and pcv indicates the packed cell volume. Further, wc indicates the white blood cell count, and rc indicates the red blood cell count.

To identify the number of chronic kidney disease patients and the number of healthy ones, we performed the visualization on the CKD dataset, which can be seen in the histogram plot in Figure 6. Here 0.0 represents the healthy cases, while 1.0 represents the chronic kidney disease patients. In this dataset, there are 250 chronic kidney disease patients, while 150 are healthy people.

The Pearson correlation feature selection method is used to get the best combination of features for the prediction of chronic kidney disease. The correlation of the 14 attributes and 1 output label is presented in Figure 7.

When we go from the exploratory data analysis stage to the pair plot visualization, it is observed to be very helpful as it gives the data that can be used to find the relationship between attributes for both the categorical and continuous variables. We import the Seaborn library to get pair plot. The information about all the attributes is in one picture and is clear. The statistical information is in attractive format represented with pair plot as shown in Figure 8.

The violin plots are used for all the attributes in exploratory data analysis that are used in the hybrid model. These can give additional useful information like density trace and distribution of the dataset. The violin plots give the whole range of dataset which cannot be shown by box plot. The violin plots of all 14 attributes are given in Figure 4. Figure 9 shows the comparison of different models' accuracy scores in the form of a chart.

5. Results and Discussion

Machine learning algorithms such as gradient boosting, Gaussian Naïve Bayes, decision tree, and random forest classifier were used in the proposed hybrid model. These different machine learning classifiers were used as a combination for the chronic kidney disease predictions. This also overcomes the overfitting problem and results in higher

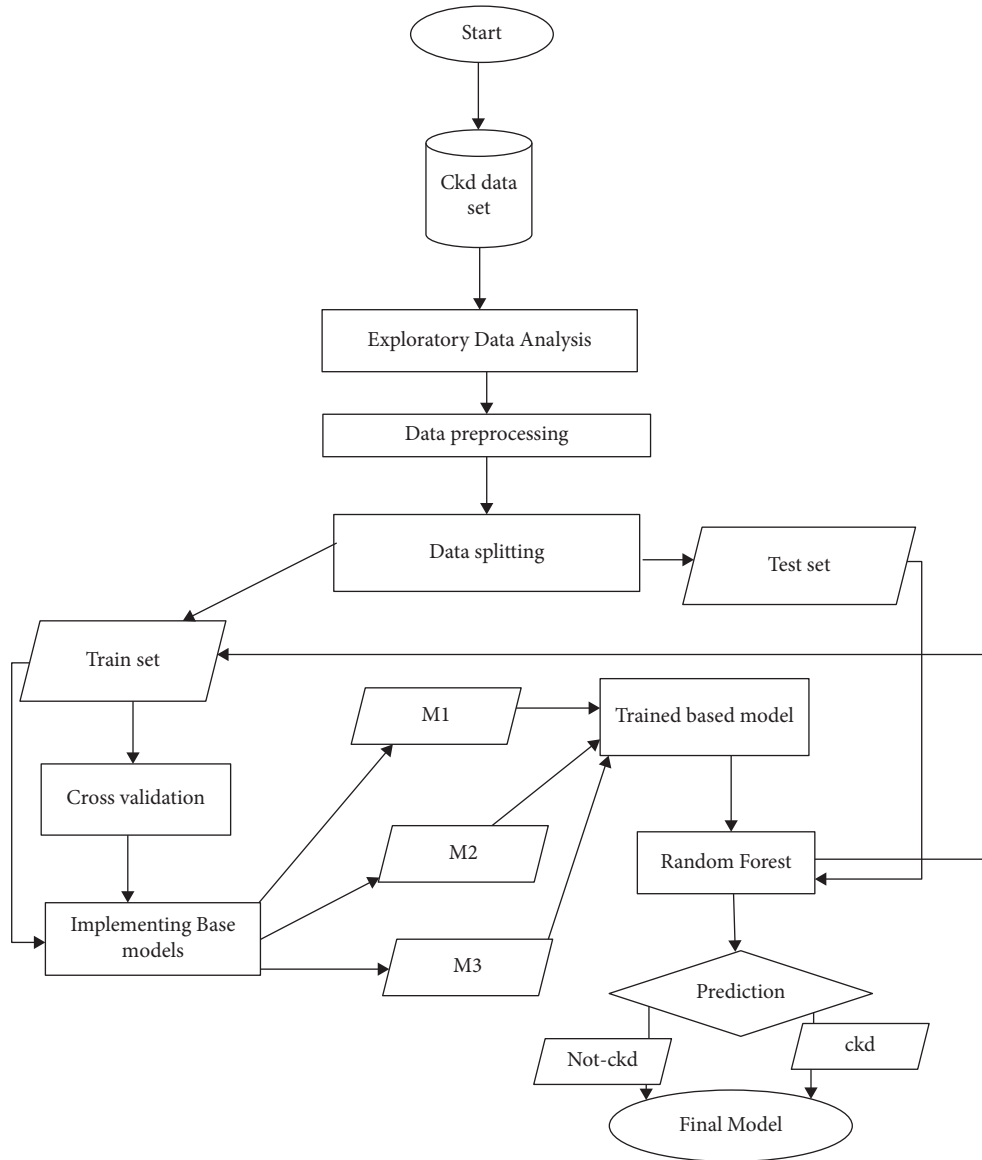


FIGURE 5: Flowchart for the proposed model.

TABLE 5: The attribute set with their data types.

#	Attributes	Full form	Data type	Nonempty value	Missing values
0	age	Age	float64	400	0
1	bp	Blood pressure	float64	400	0
2	sg	Specific gravity of urine	float64	400	0
3	al	Level of aluminum	float64	400	0
4	su	Sugar level	float64	400	0
5	bgr	Blood glucose random	float64	400	0
6	bu	Blood urea	float64	400	0
7	sc	Sugar level	float64	400	0
8	sod	Amount of sodium	float64	400	0
9	pot	Amount of potassium	float64	400	0
10	hemo	Hemoglobin	float64	400	0
11	pcv	Packed cell volume	float64	400	0
12	wc	White cell	float64	400	0
13	rc	Red cell	float64	400	0

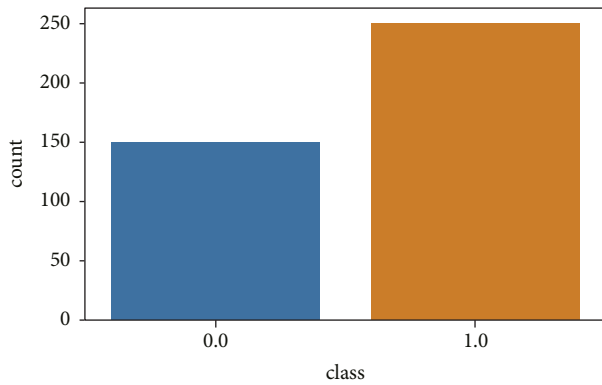


FIGURE 6: Histogram plot.

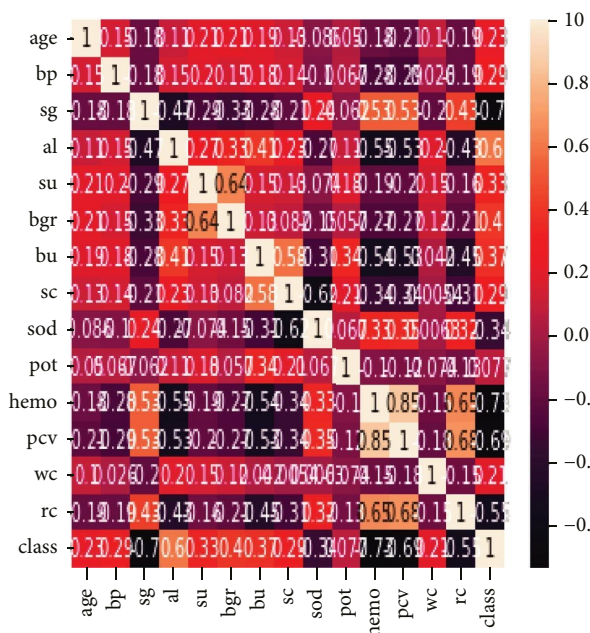


FIGURE 7: Heat map of chosen attributes.

accuracy. In order to improve accuracy and to come up with a novel approach as compared to the existing work, we have implemented the proposed hybrid model with the best combination of GB, GNB, and decision tree, along with the random forest classifiers [24–27]. The results described in Table 6 show that diagnosis of chronic kidney disease is effective using the random forest with combination as a stacking technique in the hybrid model. Gradient boosting achieves 99% accuracy, random forest achieves 98% accuracy, and our hybrid model achieves 100% accuracy, and at the same time, it has reduced the chances of overfitting.

In order to find the contributions to the development of prediction models for chronic kidney disease, a regional basis analysis is performed. As discussed in the Introduction section that the developing countries’ population suffers more from chronic kidney disease, it was observed that most of the research work is performed in developing countries. A summary of this region-wise contribution is presented in Figure 10.

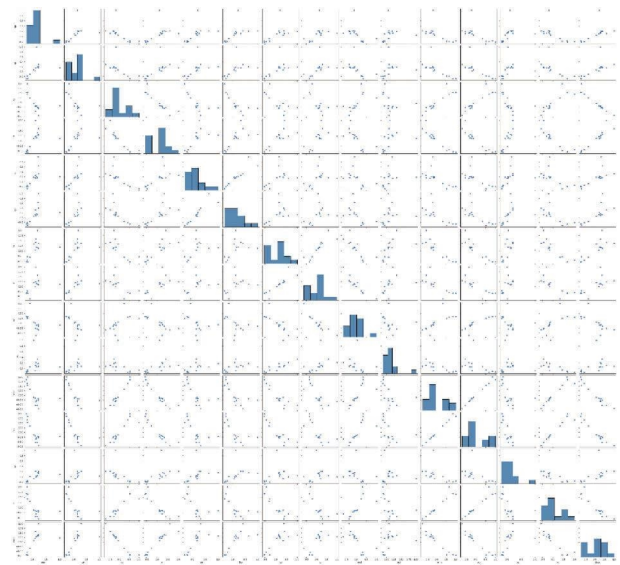


FIGURE 8: Pair plot of each attribute.

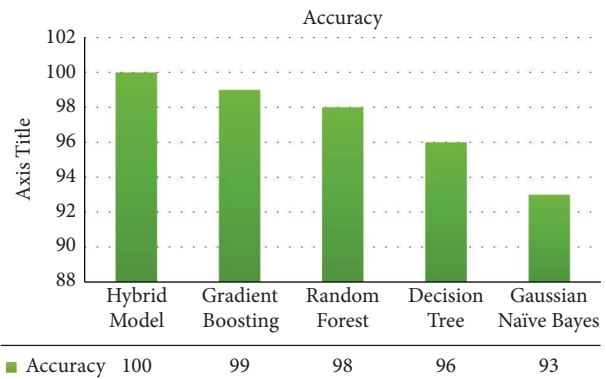


FIGURE 9: Accuracy score of implemented machine learning classifiers.

TABLE 6: Accuracy score of implemented machine learning classifiers.

ML algorithms	Accuracy (%)
Gradient boosting	99
Gaussian Naïve Bayes	93
Decision tree	96
Random forest	98
Hybrid model	100

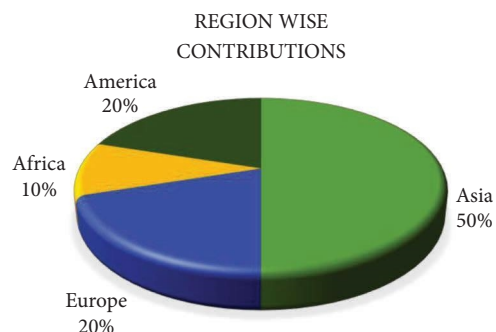


FIGURE 10: Region-wise contributions.

6. Conclusion

Chronic kidney disease is considered as one of the prominent life-threatening diseases in the developing world. The most obvious cause seems to be lack of physical exercise. The medical practitioners used a number of diagnosis processes and procedures, where machine learning is the recent development. In this paper, we have selected machine learning because in terms of accuracy, it performs better as compared to other available approaches. In this article, we have used the Pearson correlation feature selection method and applied the same on machine learning classifier. GB, GNB, decision tree, and random forest are the base classifiers for the stacking algorithm, whereas these are implemented with the cross-validation on the basis of accuracy score. In this study, we evaluated these algorithms on the same dataset. Furthermore, we have used dataset of CKD from the UCI directory that contains 14 attributes and 400 instances. On the basis of these attributes, our proposed stacking model is able to predict whether the person is a CKD patient or not with 100% accuracy. Best features are selected using the Pearson correlation method, and the stacking algorithm is implemented with the best machine learning classifiers. The cross-validation enhances the performance of the stacking model. As we have worked on the chronic kidney disease data of the binary group, the stacking algorithm performs better with these combinations of algorithms. We can implement the stacking technique for the prediction of other diseases to get better accuracy score.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] V. Jha, G. Garcia-Garcia, K. Iseki et al., "Chronic kidney disease: global dimension and perspectives," *The Lancet*, vol. 382, no. 9888, pp. 260–272, 2013.
- [2] R. Ruiz-Arenas, "A summary of worldwide national activities in chronic kidney disease (CKD) testing, the electronic journal of the international federation of," *Clinical Chemistry and Laboratory Medicine*, vol. 28, no. 4, pp. 302–314, 2017.
- [3] Thedailystar, "Over 35,000 develop kidney failure in Bangladesh every year," 2019, <https://www.thedailystar.net/city/news/18m-kidney-patients-bangladesh-every-year-1703665>.
- [4] Prothomalo, "Women more affected by kidney diseases," 2018, <https://en.prothomalo.com/bangladesh/Womenmore-affected-by-kidney-diseases>.
- [5] Scottish Intercollegiate Guidelines Network (SIGN), *Diagnosis and Management of Chronic Kidney Disease: A National Clinical Guideline*, SIGN, Victoria, Australia, 2008.
- [6] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in *Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, January 2021.
- [7] P. Ghosh, F. M. Javed Mehedi Shamrat, S. Shultana, S. Afrin, A. A. Anjum, and A. A. Khan, "Optimization of prediction method of chronic kidney disease using machine learning algorithm," in *Proceedings of the 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Bangkok, Thailand, November 2020.
- [8] K. R. A. Padmanaban and G. Parthiban, "Applying machine learning techniques for predicting the risk of chronic kidney disease," *Indian Journal of Science and Technology*, vol. 9, no. 29, 2016.
- [9] A. Charleonnann, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *Proceedings of the 2016 Management and Innovation Technology International Conference (MITicon)*, Bang-San, Thailand, October 2016.
- [10] G.-S. Fu, Y. Levin-Schwartz, Q.-H. Lin, and D. Zhang, "Machine learning for medical imaging," *Journal of healthcare engineering*, vol. 2019, pp. 1–2, 2019.
- [11] R. Devika, S. V. Avilala, and V. Subramaniaswamy, "Comparative study of classifier for chronic kidney disease prediction using naive Bayes, KNN and random forest," in *Proceedings of the 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, March 2019.
- [12] S. Revathy, B. Bharathi, P. Jeyanthi, and M. Ramesh, "Chronic kidney disease prediction using machine learning models," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1, pp. 6364–6367, 2019.
- [13] A. S. A. Rabby, R. Mamata, M. A. Laboni, Ohidujjaman, and S. Abujar, "Machine learning applied to kidney disease prediction: comparison study," in *Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, July 2019.
- [14] M. Nishat, F. Faisal, R. Dip et al., "A comprehensive analysis on detecting chronic kidney disease by employing machine learning algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, Article ID 170671, 2018.
- [15] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," in *Proceedings of the 2017 IEEE Symposium on Computers and Communications (ISCC)*, Heraklion, Greece, July 2017.
- [16] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach," *Journal of network and innovative computing*, vol. 4, pp. 175–184, 2016.
- [17] Bmc, "Biomedcentral," 2022.
- [18] S. Ramya and N. Radha, "Diagnosis of chronic kidney disease using machine learning algorithms," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 1, 2016.
- [19] M. Kumar, "Prediction of chronic kidney disease using random forest machine learning algorithm," *International Journal of Computer Science and Mobile Computing*, vol. 5, pp. 24–33, 2016.
- [20] M. S. Basarslan and F. Kayaalp, "Performance analysis of fuzzy rough set-based and correlation-based attribute selection methods on detection of chronic kidney disease with various classifiers," in *Proceedings of the 2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science (EBBT)*, April 2019.

- [21] S. K. Dowluru and A. K. Rayavarapu, "Statistical and data mining aspects on kidney stones: a systematic review and metza-analysis," *Open Access Scientific Reports*, vol. 1, no. 12, 2012.
- [22] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, and M. A. Hossain, "Comparative analysis of classification approaches for heart disease prediction," in *Proceedings of the 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pp. 1–4, Rajshahi, Bangladesh, February 2018.
- [23] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, Article ID 100203, 2019.
- [24] A. J. Aljaaf, A.-J. Dhiya, H. M. Hussein et al., "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in *Proceedings of the 2018 IEEE Congress on Evolutionary Computation (CEC)*, Rio de Janeiro, Brazil, July 2018.
- [25] S. Khan, M. Z. Khan, P. Khan, G. Mehmood, A. Khan, and M. Fayaz, "An ant-hocnet routing protocol based on optimized fuzzy logic for swarm of UAVs in FANET," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 6783777, 12 pages, 2022.
- [26] M. Fayaz, G. Mehmood, A. Khan, S. Abbas, M. Fayaz, and J. Gwak, "Counteracting selfish nodes using reputation based system in mobile ad hoc networks," *Electronics*, vol. 11, no. 2, p. 185, 2022.
- [27] M. Z. U. Haq, M. Z. Khan, H. U. Rehman et al., "An adaptive topology management scheme to maintain network connectivity in Wireless Sensor Networks," *Sensors*, vol. 22, no. 8, p. 2855, 2022.

Review Article

Next Generation Infectious Diseases Monitoring Gages via Incremental Federated Learning: Current Trends and Future Possibilities

Iqra Javed,¹ Uzair Iqbal ,² Muhammad Bilal,³ Basit Shahzad,¹ Tae-Sun Chung ,⁴ and Muhammad Attique ⁵

¹Department of Software Engineering, National University of Modern Languages, Islamabad 44000, Pakistan

²Department of Artificial Intelligence and Data Science, National University of Computer and Emerging Sciences, Islamabad 44000, Pakistan

³Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad Chiniot-Faisalabad Campus, Chiniot 35400, Pakistan

⁴Department of Artificial Intelligence, Ajou University, Suwon-Si 16499, Republic of Korea

⁵Department of Software, Sejong University, Seoul 05006, Republic of Korea

Correspondence should be addressed to Uzair Iqbal; uzairiqbal13@gmail.com, Tae-Sun Chung; tschung@ajou.ac.kr, and Muhammad Attique; attique@sejong.ac.kr

Received 10 June 2022; Revised 29 July 2022; Accepted 27 September 2022; Published 1 March 2023

Academic Editor: Muhammad Imran

Copyright © 2023 Iqra Javed et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Infectious diseases are always alarming for the survival of human life and are a key concern in the public health domain. Therefore, early diagnosis of these infectious diseases is a high demand for modern-era healthcare systems. Novel general infectious diseases such as coronavirus are infectious diseases that cause millions of human deaths across the globe in 2020. Therefore, early, robust recognition of general infectious diseases is the desirable requirement of modern intelligent healthcare systems. This systematic study is designed under Kitchenham guidelines and sets different RQs (research questions) for robust recognition of general infectious diseases. From 2018 to 2021, four electronic databases, IEEE, ACM, Springer, and ScienceDirect, are used for the extraction of research work. These extracted studies delivered different schemes for the accurate recognition of general infectious diseases through different machine learning techniques with the inclusion of deep learning and federated learning models. A framework is also introduced to share the process of detection of infectious diseases by using machine learning models. After the filtration process, 21 studies are extracted and mapped to defined RQs. In the future, early diagnosis of infectious diseases will be possible through wearable health monitoring cages. Moreover, these cages will help to reduce the time and death rate by detection of severe diseases at starting stage.

1. Introduction

At the end of 2019, the infectious disease, coronavirus, broke out in China and spread across the globe in a few months. The World Health Organization (WHO) declared that COVID-19 (Coronavirus Disease-19) is a deathly pandemic and resulted in different sorts of challenges around the world [1]. Although the patterns are still clear, studies indicate that this major issue will continue to exist over the next few years. COVID-19 is a general infectious disease that affects the

human respiratory system. One of the general infectious diseases is SARS (severe acute respiratory syndrome), influenza, and cold viruses, which are well-known. Furthermore, despite being exposed to these diseases, only a small percentage of the population produces antibodies, according to surveys conducted in various nations. This proves that most patients will regularly require examinations by a limited number of doctors in short intervals due to resource constraints. Infectious diseases are usually diagnosed by using at least one of these three tests: chest X-ray, RT-PCR

(reverse-transcriptase polymerase chain reaction), and computed tomography.

In sputum or a nasopharyngeal sample, the RT-PCR assay detects viral RNA (ribonucleic acid). It requires the use of specialist materials and equipment that are not widely available, and it typically inconveniently takes 12 hours because patients with an infectious disease must be identified and monitored as quickly as possible. Tests that use RT-PCR to determine results performed on the same patients at different times throughout the illness were found to be inconsistent, resulting in a high false-negative rate [2]. CT scan and 3D radiography images from intelligent diagnostic devices are used in a variety of clinical perspectives. Most hospitals lack the necessary equipment for this process. Patients are observed and treated on the base of clinical history. The equipment required for this examination in CXR (chest X-ray) is less cumbersome and easily adjusted. These resources are, for the most part, effortlessly accessible [3].

With the rapid evolution of electronic health records, it is now easier to use data for predictive modelling and subsequent advancements. Different applications and approaches in healthcare involve distributed machine learning, including electronic health records and chatbots, to detect a pattern in clinical status, detect the type of cancer treatment, and identify unusual diseases or infections and pathology. Contactless COVID-19 patient identification is carried out through the classification of COVID-19 cough samples, and the detection of these symptoms is accomplished by using advanced algorithms and procedures, resulting in more relevant, tailored, and accurate patient care. In addition, sensors are introduced that both monitor the temperature with facial recognition and upload each person's record to a directory [4]. Organizations are increasingly focusing on developing more efficient algorithms and using the potential of deep learning to build acceptable solutions in tackling exact, real-world challenges in the health sector.

To overcome the challenges of patients who are unaware of their symptoms at the first stage of the disease or who cannot go for a regular check-up for many reasons, DL can be used to analyse electronic health records. Due to its transformative potential, DL is a subset of ML (machine learning) and AI (artificial intelligence) that adds a new layer of complexity to medical technology solutions. The healthcare industry is using DL efficient records with efficiency and exceptional speed [5]. The modern healthcare system is extremely helpful, which makes prediction processes fast, efficient, and accurate with good learning ability, and more benefits lie within the neural networks formed by using AI and ML. The design and working of DL neural networks are like the system of the human brain. Because of multilayer networks and technology, it can be easily managed and sifted through vast quantities of data that would be lost or missed. Networks in deep learning can solve complex problems and can handle reams of data, which is very helpful in the profession of healthcare and federated learning [6].

Deep learning is currently used in the electronic health record to anticipate healthcare-associated illnesses and to

minimize administrative load [7]. Medical practitioners focused on healthcare concerns as a result of reducing administrative difficulties and enhancing access to essential patient records [8]. The use of biomedical data in deep learning is becoming increasingly important in the age of healthcare. The use of electronic health records helps to make sure that the proper medication and prescription are provided to the persistent environment and molecular traits [9]. By learning about all infectious diseases and their cure, the right treatment can be given to the target patient. It is difficult to examine the symptoms of infection and identify which kind of infection the patient is suffering from. Deep learning can work for the detection of these diseases by using an efficient framework with the help of its effective learning feature [10]. Figure 1 shows the impact of using federated learning-based monitoring gages for the detection of infectious diseases.

This systematic study is designed to highlight different machine learning approaches, especially federated learning, for accurate detection. It highlights some future possibilities, which help to design different wearable gages for the early diagnosis of different infectious diseases. Different social media platforms are used for the detection of location of infectious diseases [11]. Through social media platforms, infectious diseases can be detected easily. For instance, messages from Weibo, Facebook, Instagram, WhatsApp, and Twitter have demonstrated their use as data sources for detecting and evaluating infectious illnesses [12]. Moreover, it thoroughly overlooks the architecture view of federated learning, which plays a vital role in mapping the local training data to centralized training master data [13]. For the execution of a systematic study, different research questions are designed to investigate general infectious disease monitoring games using a federated learning scheme. In this study, four electronic databases, ACM, IEEE Access, Springer, and ScienceDirect, are used to extract recent studies from 2018 to 2021. The extracted studies answer the RQs and how machine learning approaches are used for the recognition of different infectious diseases.

2. Materials and Methods

To detect infectious diseases with more effectiveness and accuracy, a systematic literature review is carried out. The best possible research questions are highlighted to support the research problem.

RQ1: How do different machine learning algorithms play a vital role in the early identification of infectious diseases?

RQ2: What is the robust impact of smart healthcare systems in recognition of different infectious diseases through distributed machine learning and deep learning models?

RQ3: What is the influence of different federated learning models on the inclusion of the CNN (convolutional neural network) in the detection of infectious diseases?

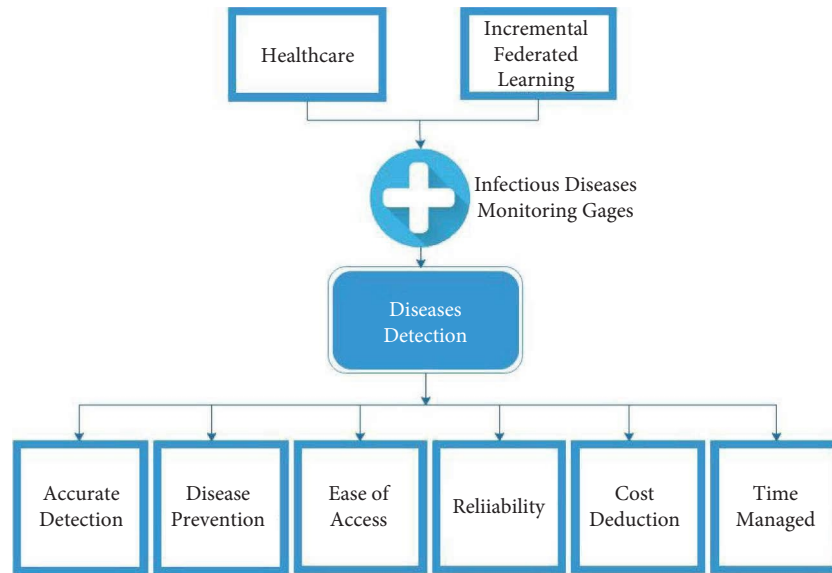


FIGURE 1: Use of infectious diseases monitoring gages.

2.1. Search Process. Different search strings are produced to search related studies. These search strings are then applied to find related results. These factors can then be used to improve healthcare systems by using the search string as a guide. Machine learning algorithms and smart healthcare systems can also be identified by using search queries. Various digital platforms are used to search for related studies. Google Scholar, IEEE digital library, and ACM are a few of them. To improve the accuracy of the search process, innovative strategies are implemented into the search strings.

("Infectious diseases detection" OR "COVID-19") AND ("Infectious disease recognition" OR "Infectious diseases classification") AND (Machine learning algorithms") AND ("Intelligent healthcare systems") AND ("Distributed machine learning") AND ("Federated learning in healthcare").

2.2. Inclusion and Exclusion Criteria. Our study is primarily focused on healthcare and improving it by using machine learning techniques. To achieve this, the inclusion/exclusion criteria are established to obtain results that are related to the research problem. Table 1 highlights the inclusion scheme of the collected studies, and Table 2 represents the exclusion scheme that supports the cleaning process.

2.3. Data Collection and Cleaning. There are plenty of ways to collect data, but electronic databases are the most used in the extraction of data. The data were extracted from four main electronic databases from the relevant literature. These electronic databases are IEEE, ACM, ScienceDirect, and Springer. Research questions are focused on data collection, with only relevant research studies added to support the questions. After applying the inclusion and exclusion criteria, the extracted studies are used to do a systematic literature review. The extracted literature supports our

research problem, while Figures 2 and 3 support the data collection and cleaning process.

After different filtration schemes, 21 articles were extracted from databases and mapped to defined RQs, Table 3. Moreover, the highlighted mapping of fetched articles to RQs declares those parameters of the federated learning scheme, which help to design in the future in terms of monitoring infectious diseases wearable gages.

3. Discussion on Current Trends

In this section, a mapping of related work is carried out to discover how many selected papers are related to the research questions. These selected studies are discussed in the bibliometric analysis. The selected study covers all research questions about how machine learning is used for the recognition of infectious diseases.

3.1. Architectural View of Centralized Machine Learning Techniques. Deep learning models consist of increased volumes of unsupervised data to produce complex representations with greater accuracy than machine learning traditional approaches. Hierarchical learning is simulated by using artificial multilayer neural networks. This allows all layers to generate various attributes by using raw information. High-end machines are required for DL algorithms because they work with a large amount of data and provide advanced solutions [14]. As a result, deep learning relies heavily on the graphics processing unit. The feature extraction improves performance and decreases the data complexity in ML. Learning high-level functions and data without the manual input of domain experts is possible with deep learning algorithms [15]. In regard to the test phase, the deep learning algorithm is much faster than machine learning algorithms and provides more accurate results [16]. To identify solutions to complex health issues and provide

TABLE 1: Inclusion scheme in the study collection.

No.	Inclusion criteria
1	Discuss the optimized methods of machine learning
2	Discuss the limitations of the use of distributed machine learning with the comparison of federated learning
3	Papers discuss the flow of federated learning in biomedical application
4	Papers discuss the detection of COVID-19 via deep learning models

TABLE 2: Exclusion scheme for the cleaning process.

No.	Exclusion criteria
1	Not the English language scholarly article.
2	Parameters of distributed machine learning are not defined clearly
3	Results are not clearly defined in biomedical applications
4	Parameters of federated learning are not defined clearly

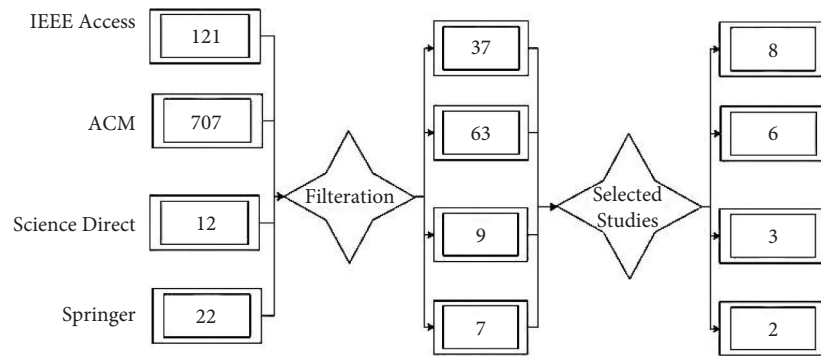


FIGURE 2: Study filtration and selection process.

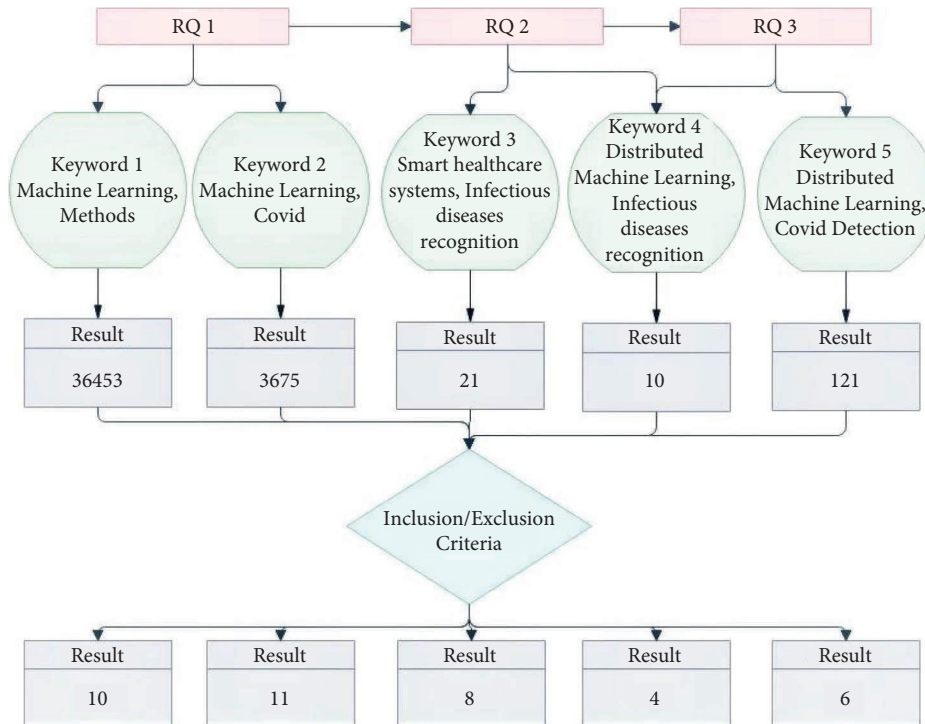


FIGURE 3: Structural diagram of the study extraction process.

patients with long-term treatment, the algorithms of ML and DL are applied [17]. Providers of healthcare can benefit from medical images by merging them with demographic data

[18]. In addition to DNNs and RNNs, there are also probabilistic neural networks (PNNs) and feed-forward neural networks (FFNNs). Most DL systems use CNNs

TABLE 3: Bibliometric measurement.

Ref#	Key factors	Merits	Demerits	Mapping	Year
[2]	EHRs (electronic health records) are supplemented with hierarchical information from medical ontologies by using a GRAM (graph-based attention model)	GRAM is performing excellently. When the data are inadequate, it works well.	Improvement is needed in the way this method incorporates knowledge DAG (directed acyclic graph) into neural networks.	RQ2	2017
[7]	Restricted Boltzmann machines and autoencoder stacked units' networks are implemented	Comparing the results of deep learning methods, which have highly precise values	Experiment with a broader scope of preprocessing methods is needed.	RQ3	2016
[8]	To identify infectious disease host genes, a machine learning classification technique is developed	Wide-scale host gene prediction connected to infectious diseases is made possible.	There are no major benefits of not being able to use a small-scale dataset.	RQ1	2019
[12]	Comprehensive review study for the diagnosis of COVID-19 via deep learning models	A detailed review of different diagnosis methods of COVID-19 by using the different structures of the CNN and ResNet-50 (residual network-50) model.	Computational complexity factor requires to highlight clearly	RQ2	2019
[13]	Added recent research on the use of DL (deep learning) to improve the domain of health care.	Big biomedical data could be translated into improved human health by using deep learning techniques.	The development of applications needs to be improved.	RQ2, RQ3	2018
[10]	CNN is a perfect model to use for the analysis of applications and challenges of medical images.	It can detect infectious disease outbreaks, among other applications.	System inconsistencies include heterogeneity of data quality and security.	RQ2	2021
[14]	Use of neural networks in the prediction of diseases.	Helps to identify how neural networks can be helpful in detecting infectious diseases.	Results and technical parts are missing, which would be helpful in implementing the framework	RQ2	2019
[15]	Medical, e-healthcare, and bioinformatics applications of DL are discussed.	Contains effective DL methods for biomedical and health-related applications.	In healthcare, distinctions between deep learning technologies and techniques need to be improved.	RQ2	2020
[16]	SAPS II and SOFA ratings (severity scores) ML ensembles were compared for quality check.	As per the results, the DL model defeated most other techniques.	Current data must be added.	RQ2	2018
[17]	Privacy concerns are highlighted in the flow of EHR through federated learning.	A unique federated learning framework proposed for efficient diagnosis of different human diseases	At least discuss the computational complexity in the flow of HER through federated learning.	RQ3	2018
[18]	The fusion-based federated learning model for accurate detection of COVID-19.	Medical image analysis for detection of COVID-19 for better communication and performance if federated learning model.	Along with the accuracy factor, the robustness parameter is missed in the proposed model.	RQ3	2021
[19]	Deep learning techniques are used which are working in healthcare	Exposed a few key areas of medicine where DL computational methods can have a positive impact.	Some other techniques of deep learning are not discussed	RQ2	2019
[20]	Driver drowsiness is predicted by using a deep CNN model.	Helps to create an improved system that detects driver drowsiness by using the deep CNN	Needs further improvement in eye detection speed.	RQ1	2019
[21]	DeepSol, a novel protein solubility predictor based on deep learning, has been proposed by researchers.	DeepSol has overcome the limitations of its feature selection step and two-stage classifier.	It can be projected with DeepSol to lower costs.	RQ2	2018
[22]	FML (federated machine learning) thoroughly discusses the different parameters of training and testing the ML models.	A comprehensive review of the concepts of vertical and horizontal federated learning models. Moreover, we thoroughly discussed the applications of FML inclusion in healthcare applications.	Compromises detailed discussion on security protocols when electronic health records move from one node to another node.	RQ3	2019

TABLE 3: Continued.

Ref#	Key factors	Merits	Demerits	Mapping	Year
[23]	An evolutionary algorithm is proposed for training a DNN (deep neural network) model for the estimation of morbidity of gastrointestinal infections.	Compared to the extensively used ANN (artificial neural network) and MLR (multiple linear regression) models, this model is much more accurate at predicting disease morbidity.	Further samples should be collected, and pollutants should be determined.	RQ2	2017
[24]	On the MovieQA question answering dataset, a model is presented.	Models are learning matching patterns for the selection of the right response.	To improve machine reading comprehension, the system should include entailments and answers.	RQ1	2018
[25]	This study introduced the independently recurrent neural network.	By learning long-term dependencies, IndRNN (independently recurrent neural network) helps to prevent gradient explosion and disappearance.	It is not possible to improve the performance of the LSTM (long short-term memory) by raising the size of parameters or layers.	RQ1	2018
[26]	The performance of ML networks is compared to that of feed-forward neural networks, also with logistic regression.	The XGB (gradient-boosted trees) model, which was found to be the most accurate, outperformed the logistic regression in terms of calibration.	There is a need for further research to improve the prediction of administrative information.	RQ1	2020
[27]	The RNN technique can be formally developed for differential equations by using the RNN canonical formulation.	Signal processing-based analysis of RNNs and vanilla LSTMs and comprehensive treatment of the RNN concepts using descriptive and meaningful notation are presented.	The augmented LSTM system is effective, but it needs to be enhanced with more techniques.	RQ1	2020
[28]	Developed a wearable body sensor fusion data-driven deep RNN activity recognition system.	A human's functionality and lifestyle can be determined based on physical actions by using body sensors.	A human behaviour monitoring system can further be evaluated in real-time on overly complex datasets.	RQ1	2020

(ResNet and GoogleNet) and recurrent neural networks (RNNs) (LSTM and GRU (gated recurrent unit)). Stacking autoencoders is also popular in machine learning [19].

3.2. Convolutional Neural Network. The input, hidden, and output layers are regular neural network layers. This is because every layer contains neurons, and each neuron of the present layer is connected to a neuron of the previous layer, so all neurons are of high weightage. This method is effective in predicting simple and small data but fails when dealing with complex data objects and translations. Cells are only connected to their nearest neighbours in the convolutional layer, and all cells have the same weight. Figure 3 highlights the structure of the CNN with the inclusion of input, output, and hidden layers.

In the figure of the CNN, we will treat eyes as a separate object in image detection; it will not find eyes all over the image. The CNN requires images of a fixed size as an input, and preprocessing is required to achieve output. These key features are then stored in a database for preprocessing before they are sent to an application. Features of these images are detected and used for further image detection and classification. Figure 4 shows the flow of the CNN. Layers such as the convolutional pooling and ReLU (rectified linear activation function) functions, as well as a fully connected layer, are all used to build the network. It is divided into several layers of kernels. Each kernel covers a specific feature

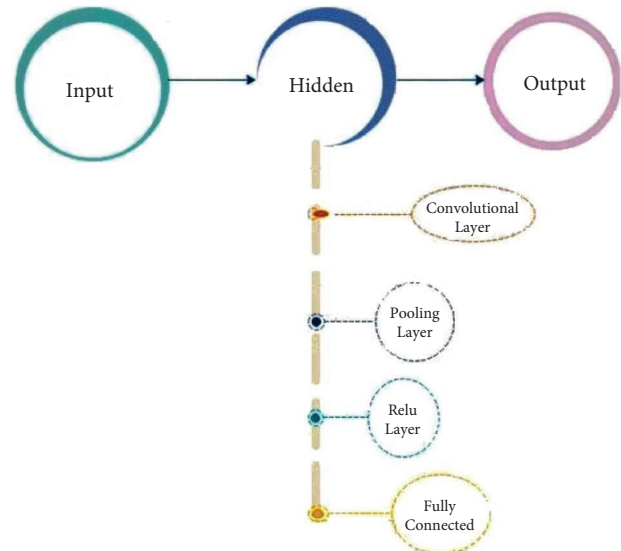


FIGURE 4: Structure of the convolutional neural network (feed-forward).

of the object with specific dimensions. Kernel 1 will detect the eyes of the object, kernel 2 will detect the nose, kernel 3 will detect its lips, and kernel 4 will detect the shape of the object. Next-layer classification and accurate prediction will be based on these vectors [20].

Max or average algorithm is used for the feature map to decrease its range. This algorithm increases the speed of the pooling layer. The supreme area of a particular feature map is taken as input and places in the same area are returned as output in the max-pooling process. When using average pooling, a feature map of average size is used as input. Negative values are converted to zero in the ReLU layer. Using activation, classify the input into a fully connected layer and assign it a class score.

Infectious disease instances are detected with the help of an extremely basic CNN model. This model contains a single convolutional layer with sixteen filters. These filters are followed by the batch normalization layer, the ReLU layer, two fully connected layers, and the final layer, the SoftMax layer. A preprocessed picture dataset is read into the input layer of the model. These images are subjected to a separate preprocessing phase. Images are cropped and resized during the preprocessing stage. Primarily, the purpose of convolutional is to extract features from a picture dataset and establish a spatial connection between image pixels in the image. To decrease the number of training epochs required for deep network stabilization and training, a batch normalization layer is used. As a result of the use of the ReLU layer, the negative pixels in the convolved features are replaced by zeros. A nonlinearity map of CNN's features is generated by using this function. The primary job of the fully connected layer is to classify the recovered features from picture datasets into classes. The function of the Softmax layer is purely for determining the activation function results from the probability values of the preceding layer. In the diagnosis of infectious diseases, the values can be classified into two classes: "0" and "1." In the last output layer of the CNN model, results from the previous layer can be labelled. Therefore, for instance, a COVID-19 value of "1" indicates a positive case, while a non-COVID-19 value of "0" indicates that the chest X-ray or CT was normal [21].

3.3. Recurrent Neural Networks. Because of its memory, the RNN can analyse data sequences of variable length and store them in its database. In addition, it takes into account the previous input state [22]. When making predictions, it uses information from its past, and an infinite number of steps are repeated indefinitely to propagate information through its hidden state over time [23]. Figure 5 shows the structure representation of the RNN.

It manipulates current and recent past states to produce a new data output [24]. The output is used to determine the previous state for the next time step. RNNs have short-term memory because of this role. In addition to language generation and DNA sequence analysis, it is also used in text assessment, sound analysis, time string analysis, and many other applications because it is extremely efficient for data sequences that occur in time. A simple and robust RNN is a good model to use [25]. Figure 6 describes the internal flow of the RNN model. (See Figure 7)

Because the CNN only focuses on the current input state, it has no memory and is unable to handle sequential data [26]. It is, therefore, essential to employ an RNN model for

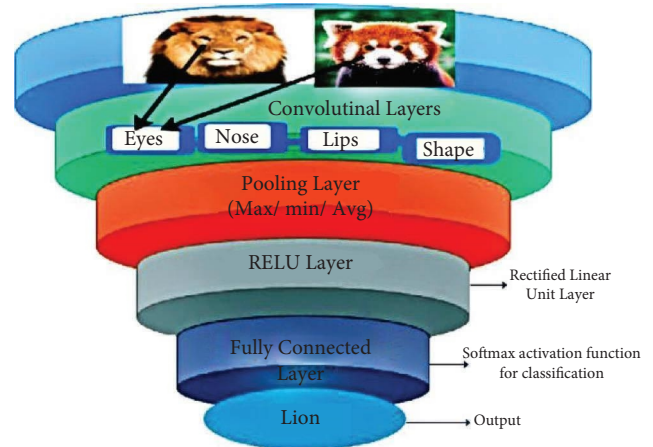


FIGURE 5: The internal flow of the CNN model.

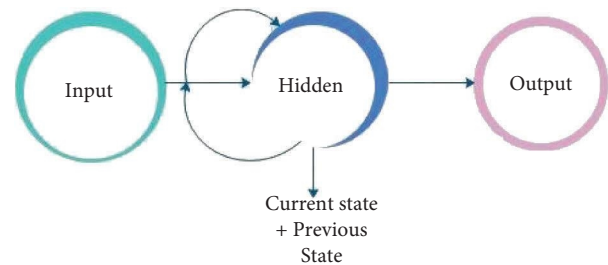


FIGURE 6: RNN structural diagram.

the improvement of the prediction and to manage sequential records. Then, the RNN model feeds itself data by using the output as a previous state for the next time step. Data can be checked over time using RNNs [27].

3.4. Deep Neural Network. The layered architecture of advanced systems is used in DNN's architecture and implementation. Processing power and hardware performance are required for performing complex tasks. Models such as the DNN are used for classification and regression purposes. Classification results are more precise in complex classifications than the method itself [28]. For several years, DNNs were deemed impractical because they required too much computational power to train and process, for instance, real-time applications [29]. Due to advancements in hardware and synchronization by GPUs (graphics processing units) and big data, DNNs are now considered a major technological innovation in the field [30].

3.5. Probabilistic Neural Network. Feed-forward neural networks, such as PNNs, are commonly used to solve classification and pattern recognition concerns. A non-parametric function and a Parzen window approximate the PDF function for each class in the PNN. A PNN structure consists of 4 layers, an input pattern layer, as well as a summation and output layer.

The greatest operational advantage of the PNN is that the training is quick and easy. As soon as a pattern from each category is recognized, the network can begin generalizing to

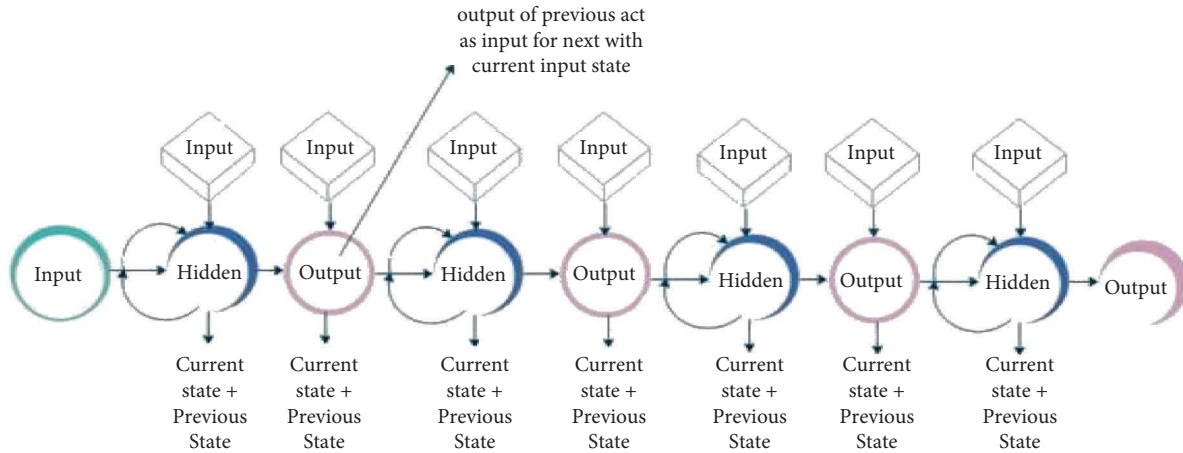


FIGURE 7: Sequential RNN model.

new patterns. As more patterns are discovered and saved in the network, the generalization improves, and the decision boundary becomes more complex.

3.6. Reinforcement Learning. In reinforcement learning, there is no way to predict the outcome, so the system must choose the best course of action. Reward-based learning is also called a behaviour-based process. In the reinforcement learning system, you receive a reward based on behaviour. Critics point out that the current situation is better than it was in the past. Figure 8 represents the environment. Agent, reward, state, and action are the five components of a reinforcement learning agent [31].

To maximize the positive reward, reinforcement learning focuses on agents' intelligence. Reinforcement learning differs from supervised learning because, in supervised learning, there is no need for input or output labels. As such, it aims to strike a balance between previous and current information. Using techniques from dynamic programming, the environment acts like a Markov decision process [32].

In reinforcement learning, there is no way to predict the outcome, so the system must choose the best course of action. Reinforcement learning is behaviour based. In the reinforcement learning system, get the reward according to the behaviour of the object. Critic information shows the current state rewards concerning the past. There are five elements of reinforcement learning: agent, environment, reward, state, and action [33].

3.7. Architectural View of Federated Learning. A "federated learning" technique involves training an algorithm without exchanging information between servers containing local data samples or other clustered edge gadgets as compared to conventional centralized machine learning methods, in which all local datasets are transferred to a single server and trained using the master model that will further globally train the peer nodes [34]. Data access rights, data privacy, heterogeneous data access, and security are factors that can be addressed with the help of federated learning. Pharmaceuticals, telecommunications, and IoT (Internet of Things) are among the industries where federated learning is used in

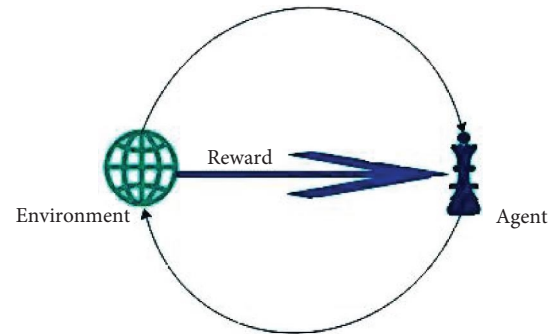


FIGURE 8: Reinforcement learning basic diagram.

effective applications [35]. Figure 9 represents the architecture of federated learning, which highlights the training of local data and synchronizes it with the master model of the ANN.

Without unambiguously trading samples of data, the goal of federated learning is multiple datasets stored in local nodes used to train machine learning algorithms. To create a linear model that is shared by all endpoints at some frequency, the models are trained locally using data samples collected locally [36].

More effective machine learning approaches can be used to improve smart healthcare systems. Using a distributed machine learning model to detect infectious diseases will provide more accurate and justified outcomes [37]. The disease detection systems or devices are lacking in quality and reliability; there is room for future research in distributed machine learning approaches to improve disease detection technology [38]. This will benefit the healthcare business as well as human health. Human life will be safeguarded by accurate predictions made at the appropriate moment and with good medical records.

4. Future Work: Incremental Federated Learning Model

In contrast to distributed learning, which maximizes computing power, federated learning focuses on training a dataset that is heterogeneous [39]. A widely known

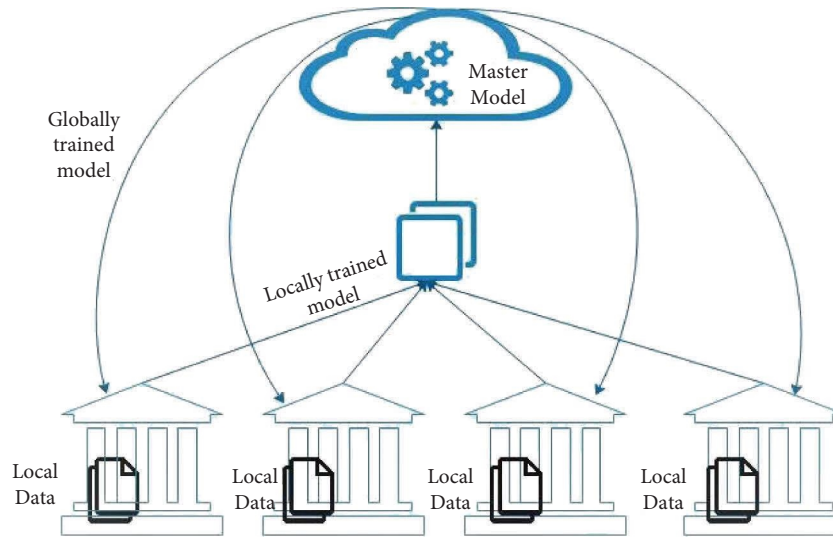


FIGURE 9: The architecture of federated learning.

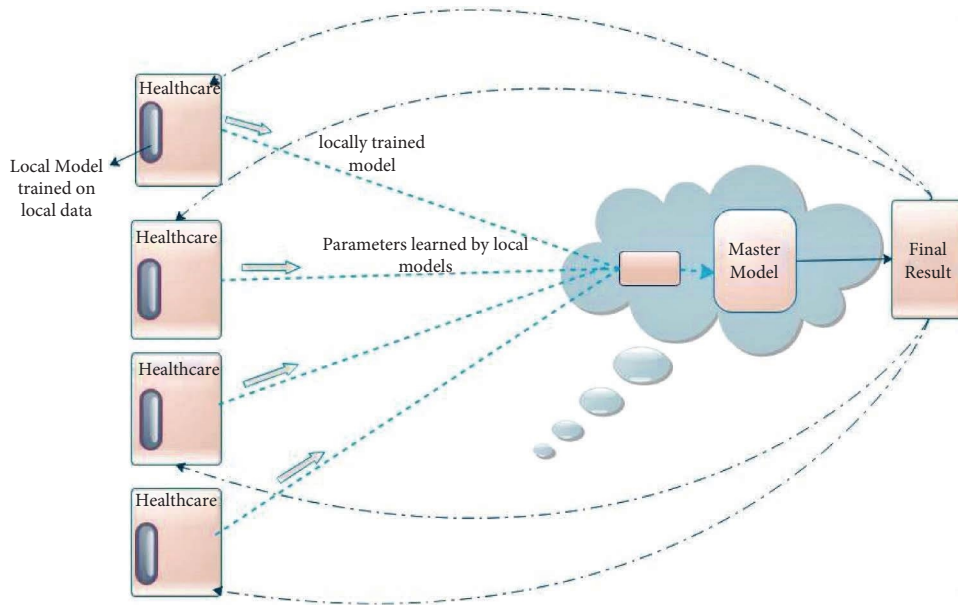


FIGURE 10: Incremental federated learning integrates into current digital healthcare systems.

underlying assumption in distributed learning is whether the local datasets are identically distributed and the same size, even though it also aims to train a single model on multiple servers. For federated learning, these hypotheses are not applicable; rather than homogeneity, datasets tend to be heterogeneous and have a range in size. As a result of their dependence on ineffective communication media, clients who are participating in federated learning could be unpredictable battery-powered systems and wireless technology (IoT devices and smartphones), but in distributed learning, all nodes are used as data centers with advanced computing capabilities and high-speed network connections. Federated learning is a smarter model with a lower legacy and less power consumption.

These machine learning approaches are very efficient in detecting infectious diseases more accurately with their efficient algorithms and frameworks. Smart healthcare systems can further be upgraded by implementing more effective machine learning approaches. The detection of infectious diseases will give more accurate and justified results by using distributed machine learning approaches. These infectious diseases can include the detection of hepatitis (B or C), malaria, dengue, tuberculosis, and COVID-19 as well. The use of decentralized learning can make detection and prediction accurate and will be able to work with the latest data as well as old data. The framework of federated learning can be helpful in learning about decentralized data.

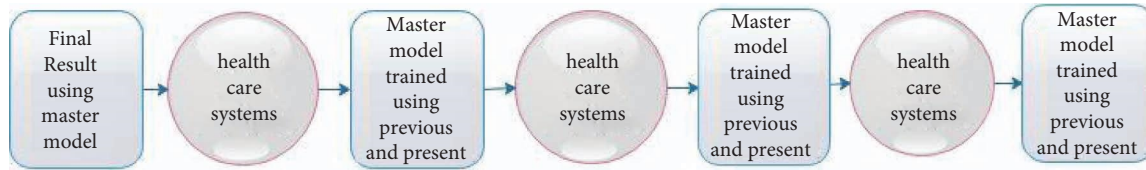


FIGURE 11: Healthcare model overview using distributed federated learning.

In the future, smart healthcare systems can be upgraded for the recognition of different infectious diseases by using distributed federated learning clusters. Figure 10 displays the next generation healthcare systems that will help to robustly recognize different infectious diseases. In distributed federated learning clusters, every smart healthcare system has locally trained a model for the prediction and recognition of different diseases. Moreover, the distributed federated learning clusters will take all parameters from these smart healthcare systems and generate a master model [40]. Such a master model will not take data for learning. Instead, it will take all parameters of smart healthcare systems and train itself through these parameters to generate a master model.

Furthermore, the master model will be the initial model of the next round, and at every round of training, the master model will learn more. This master model will have the training experience of models of all healthcare systems, so it will predict more accurately. Figure 11 is a representation of the master model increment after every round.

The above-highlighted model can be improved with time and will predict more accurately. This distributed technology will get parameters from multiple healthcare system models. These systems will have a local model, and that local model will work with machine learning algorithms to predict the results. The parameters of these local models will be transferred to a decentralized master model. This master model will learn from all parameters and predict accordingly. This model will help to learn from the present and previous models. The local model will learn from the new data at every round, and then, the master model will learn from the parameters of the local model. The master model will also learn from the parameters of previous local models. Therefore, the use of a decentralized learning approach will be helpful in improving the performance of smart healthcare systems and the recognition of infectious diseases.

5. Conclusions

With the rapid advancement in the modern healthcare system, machine learning is used for the detection of infectious diseases. These healthcare systems play a vital role in the detection of infectious diseases, maintaining healthcare records, and in communication with doctors. The healthcare systems are giving the healthcare industry easy and more effective ways to cure and identify diseases. A systematic literature review is carried out to identify upgrades in smart healthcare systems. Kitchenham guidelines are followed to extract the literature from the study by using four electronic databases. Different technologies and machine learning

algorithms are used in the detection of infectious diseases. These algorithms are working on centralized data for prediction, due to which it is difficult for healthcare systems to learn the latest data and to deal with the latest technologies with innovations. These machine learning approaches are very efficient in the more accurate detection of infectious diseases with their efficient algorithms and frameworks. Smart healthcare systems can further be upgraded by implementing more effective machine learning approaches. The use of decentralized learning can make detection and prediction accurate and will be able to work with the latest data as well as the old. As a result, a framework based on federated machine learning is introduced in this study. Wearable devices will be used to assist in the earlier detection of infectious diseases through federated learning. Federated learning is a smarter model with a lower legacy and less power consumption. Federated learning will be helpful in the precise detection of infectious diseases, which will also reduce the chance of death. The healthcare community will also be able to use it for the detection of COVID-19 and will work with the software industry to further improve it. The detection of infectious diseases will give more accurate and justified results by using distributed machine learning. These infectious diseases can include the detection of hepatitis (B or C), malaria, dengue, tuberculosis, and COVID-19 as well. Moreover, these gages will help to reduce the time and death rate by detection of severe diseases at starting stage. The accuracy and sustainability of the healthcare gadgets will be carried out by using these algorithms.

Abbreviations

RNN: Recurrent neural network
 CNN: Convolutional neural network
 PNN: Probabilistic neural network
 DNN: Deep neural network
 ML: Machine learning
 DL: Deep learning
 RL: Reinforcement learning
 FL: Federated learning.

Data Availability

There is no data involved in the composition of this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2021-0-02051) supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP) and the BK21 FOUR program of the National Research Foundation of Korea funded by the Ministry of Education (NRF5199991014091). Muhammad Attique was supported by the National Research Foundation of Korea, funded by the Ministry of Education (2020R1G1A1013221). Moreover, Uzair Iqbal would like to thank FAST National University of Computer and Emerging Sciences, Pakistan, for supporting this research with a Faculty Research Support Grant (Fall-2021) under letter ID: “11-71-13/NU-R/21.”

References

- [1] A. A. Reshi, F. Rustam, A. Mehmood et al., “An efficient CNN model for COVID-19 disease detection based on X-ray image classification,” *Complexity*, vol. 2021, pp. 2021–12.
- [2] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, “GRAM: graph-based attention model for healthcare representation learning,” in *Proceedings of the KDD: Proceedings. International Conference on Knowledge Discovery & Data Mining*, pp. 787–795, Halifax NS Canada, August 2017.
- [3] Q. U. A. Mastoi, T. Y. Wah, and U. Iqbal, “Automated diagnosis of coronary artery disease: a review and workflow,” *Cardiology Research and Practice*, vol. 2018, Article ID 2016282, 9 pages, 2018.
- [4] S. Felsenstein, J. A. Herbert, P. S. McNamara, and C. M. Hedrich, “COVID-19: immunology and treatment options,” *Clinical Immunology*, vol. 215, Article ID 108448, 2020.
- [5] A. C. Cunningham, H. P. Goh, and D. Koh, “Treatment of COVID-19: old tricks for new challenges,” *Critical Care*, vol. 24, no. 1, pp. 91–97, 2020.
- [6] B. Shahzad, I. Javed, A. Shaikh, A. Sulaiman, A. Abro, and M. A. Memon, “Reliable requirements engineering practices for COVID-19 using blockchain,” *Sustainability*, vol. 13, no. 12, p. 6748, 2021.
- [7] O. Jacobson and H. Dalianis, “Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections,” in *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pp. 191–195, Berlin, Germany, August 2016.
- [8] R. K. Barman, A. Mukhopadhyay, U. Maulik, and S. Das, “Identification of infectious disease-associated host genes using machine learning techniques,” *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019.
- [9] J. P. Van Der Heijden, N. F. De Keizer, J. D. Bos, P. I. Spuls, and L. Witkamp, “Teledermatology applied following patient selection by general practitioners in daily practice improves efficiency and quality of care at lower cost,” *British Journal of Dermatology*, vol. 165, no. 5, pp. 1058–1065, 2011.
- [10] P. S. Mathew and A. S. Pillai, “Boosting traditional healthcare-analytics with deep learning AI: techniques, frameworks and challenges,” *Enabling AI Applications in Data Science. Studies in Computational Intelligence*, vol. 911, pp. 335–365, 2021.
- [11] M. S. Nawaz, M. Bilal, M. I. Lali, R. Ul Mustafa, W. Aslam, and S. Jajja, “Effectiveness of social media data in healthcare communication,” *Journal of Medical Imaging and Health Informatics*, vol. 7, no. 6, pp. 1365–1371, 2017.
- [12] M. Bilal, A. Gani, M. I. U. Lali, M. Marjani, and N. Malik, “Social profiling: a review, taxonomy, and challenges,” *Cyberpsychology, Behavior, and Social Networking*, vol. 22, no. 7, pp. 433–450, 2019.
- [13] X. Ye, S. Li, X. Yang, and C. Qin, “Use of social media for the detection and analysis of infectious diseases in China,” *ISPRS International Journal of Geo-Information*, vol. 5, no. 9, p. 156, 2016.
- [14] M. Z. Uddin, M. M. Hassan, A. Alsanad, and C. Savaglio, “A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare,” *Information Fusion*, vol. 55, pp. 105–115, 2020.
- [15] U. Iqbal, T. Y. Wah, M. Habib ur Rehman, G. Mujtaba, M. Imran, and M. Shoaib, “Deep deterministic learning for pattern recognition of different cardiac diseases through the internet of medical Things,” *Journal of Medical Systems*, vol. 42, no. 12, p. 252, 2018.
- [16] A. Keles, M. B. Keles, and A. Keles, “COVID-19-CNNNet and COVID-19-ResNet: diagnostic inference engines for early detection of COVID-19,” *Cognit. Comput.*, pp. 1–11, Article ID 0123456789, 2021.
- [17] M. F. Aslan, M. F. Unlarsen, K. Sabanci, and A. Durdu, “CNN-based transfer learning-BiLSTM network: a novel approach for COVID-19 infection detection,” *Applied Soft Computing*, vol. 98, Article ID 106912, 2021.
- [18] S. Tabik, A. Gomez-Rios, J. L. Martin-Rodriguez et al., “COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3595–3605, 2020.
- [19] G. Hinton, “Deep learning—a technology with the potential to transform health care,” *The Journal of the American Medical Association*, vol. 320, no. 11, pp. 1101–1102, 2018.
- [20] V. R. Chirra, S. R. Uyyala, V. K. Kishore, and C. N. N. Deep, “A machine learning approach for driver drowsiness detection based on eye state,” *Rev. d’Intelligence Artif.* vol. 33, no. 6, pp. 461–466, 2019.
- [21] H. Maghdid, A. T. Asaad, K. Z. G. Ghafoor, A. S. Sadiq, S. Mirjalili, and M. K. K. Khan, “Diagnosing COVID-19 pneumonia from x-ray and CT images using deep learning and transfer learning algorithms,” vol. 11734, pp. 99–110, 2021.
- [22] E. Irmak, “COVID-19 disease severity assessment using CNN model,” *IET Image Processing*, vol. 15, no. 8, pp. 1814–1824, 2021.
- [23] M. Sanderson, A. G. Bulloch, J. L. Wang, T. Williamson, and S. B. Patten, “Predicting death by suicide using administrative health care system data: can recurrent neural network, one-dimensional convolutional neural network, and gradient boosted trees models improve prediction performance?” *Journal of Affective Disorders*, vol. 264, pp. 107–114, 2020.
- [24] M. Blohm, G. Jagfeld, E. Sood, X. Yu, and N. T. Vu, “Comparing attention-based convolutional and recurrent neural networks: success and limitations in machine reading comprehension,” in *Proceedings of the CoNLL 2018 - 22nd Conf. Comput. Nat. Lang. Learn. Proc.*, pp. 108–118, Brussels, Belgium, October 2018.
- [25] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, “Independently recurrent neural network (IndRNN): building A longer and deeper RNN,” *of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5457–5466, Salt Lake City, UT, USA, June 2018.

- [26] A. Sherstinsky, “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network,” *Physica D: Nonlinear Phenomena*, vol. 404, Article ID 132306, 2020.
- [27] Q. Song, M. R. Zhao, X. H. Zhou, Y. Xue, and Y. J. Zheng, “Predicting gastrointestinal infection morbidity based on environmental pollutants: deep learning versus traditional models,” *Ecological Indicators*, vol. 82, no. June, pp. 76–81, 2017.
- [28] I. Banerjee, Y. Ling, M. C. Chen et al., “Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification,” *Artificial Intelligence in Medicine*, vol. 97, no. November, pp. 79–88, 2019.
- [29] T. Pham, T. Tran, D. Phung, and S. Venkatesh, “Predicting healthcare trajectories from medical records: a deep learning approach,” *Journal of Biomedical Informatics*, vol. 69, pp. 218–229, 2017.
- [30] S. Khurana, R. Rawi, K. Kunji, G. Y. Chuang, H. Bensmail, and R. Mall, “DeepSol: a deep learning framework for sequence-based protein solubility prediction,” *Bioinformatics*, vol. 34, no. 15, pp. 2605–2613, 2018.
- [31] L. Canese, G. C. Cardarilli, L. Di Nunzio et al., “Multi-agent reinforcement learning: a review of challenges and applications,” *Applied Sciences*, vol. 11, no. 11, p. 4948, 2021.
- [32] A. Mehrotra, K. K. Singh, M. J. Nigam, and K. Pal, “Detection of tsunami-induced changes using generalized improved fuzzy radial basis function neural network,” *Natural Hazards*, vol. 77, no. 1, pp. 367–381, 2015.
- [33] M. Field, N. Hardcastle, M. Jameson, N. Aherne, and L. Holloway, “Machine learning applications in radiation oncology,” *Physics and Imaging in Radiation Oncology*, vol. 19, no. May, pp. 13–24, 2021.
- [34] K. Zeberga, M. Attique, B. Shah, F. Ali, Y. Z. Jembre, and T. S. Chung, “A novel text mining approach for mental health prediction using Bi-LSTM and BERT model,” *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 7893775, 18 pages, 2022.
- [35] A. Esteva, A. Robicquet, B. Ramsundar et al., “A guide to deep learning in healthcare,” *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.
- [36] S. Chae, S. Kwon, and D. Lee, “Predicting infectious disease using deep learning and big data,” *International Journal of Environmental Research and Public Health*, vol. 15, no. 8, p. 1596, 2018.
- [37] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: review, opportunities and challenges,” *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2017.
- [38] S. Purushotham, C. Meng, Z. Che, and Y. Liu, “Benchmarking deep learning models on large healthcare datasets,” *Journal of Biomedical Informatics*, vol. 83, no. April, pp. 112–134, 2018.
- [39] C. D. Naylor, “On the prospects for a (Deep) learning health care system,” *The Journal of the American Medical Association*, vol. 320, no. 11, pp. 1099–1100, 2018.
- [40] R. Zhou, W. Yin, W. Li et al., “Prediction Model for Infectious Disease Health Literacy Based on Synthetic Minority Over-sampling Technique Algorithm,” *Computational and Mathematical Methods in Medicine*, vol. 2022, Article ID 8498159, 6 pages, 2022.

Research Article

A Multimodal Network Security Framework for Healthcare Based on Deep Learning

Qiang Qiang Chen ¹, Jian Ping Li ¹, Amin ul Haq ¹, Bless Lord Y. Agbley ¹,
Arif Hussain ², Inayat Khan ³, Riaz Ullah Khan ⁴, Jalaluddin Khan ⁵ and Ijaz Ali ⁶

¹School of Computer Science and Engineering, University of Electronic Science and Technology China, Chengdu 611731, China

²Abdul Wali Khan University Mardan, Mardan 23200, KPK, Pakistan

³Department of Computer Science, University of Buner, Buner 19290, Pakistan

⁴Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Huzhou 313001, China

⁵Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh 522502, India

⁶Iqra National University Swat Campus Odigram, Department of Computer Science, Swat 19130, Pakistan

Correspondence should be addressed to Jian Ping Li; jpli2222@uestc.edu.cn

Received 13 May 2022; Accepted 16 August 2022; Published 20 February 2023

Academic Editor: Farman Ali

Copyright © 2023 Qiang Qiang Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As the network is closely related to people's daily life, network security has become an important factor affecting the physical and mental health of human beings. Network flow classification is the foundation of network security. It is the basis for providing various network services such as network security maintenance, network monitoring, and network quality of service (QoS). Therefore, this field has always been a hot spot of academic and industrial research. Existing studies have shown that through appropriate data preprocessing techniques, machine learning methods can be used to classify network flows, most of which, however, are based on manually and expert-originated feature sets; it is a time-consuming and laborious work. Moreover, only features extracted by a single model can be used in classification tasks, which can easily make the model inefficient and prone to overfitting. In order to solve the abovementioned problems, this study proposes a multimodal automatic analysis framework based on spatial and sequential features. The framework is completely based on the deep learning method and realizes automatic extraction of two types of features, which is very suitable for processing large-flow information; this improves the efficiency of network flow classification. There are two types of frameworks based on pretraining and joint-training, respectively, with analyzing the advantages and disadvantages of them in practice. In terms of evaluation, compared with the previous methods, the experimental results show that the framework has good performance in both accuracy and stability.

1. Introduction

The rapid development of the Internet makes the Internet technology penetrate into all aspects of people's lives. The quality of the network environment is closely related to the physical and mental health of human beings. At present, various types of bad website traffic or apps will push bad information or use network viruses to infringe on privacy, so it is very important to classify various network applications in a timely and effective way. Network flow classification refers to the use of a certain algorithm to construct a

classification model, which can be used to classify network flow of various applications. It is a fundamental work for providing various network services such as network security, network monitoring, and quality of service (QoS). Therefore, this field has always been a hot spot in academic and industrial research [1]. With the continuous development of the Internet of Things, many devices are connected to the network, and network flow classification has become an important part of this scenario [2]. New network applications based on different devices are emerging with the mutual information interaction between applications which

has prompted the network to face the status quo of large throughput and difficulty of network flow classification. It is urgent to deal with large and complex types of network flow and improve the efficiency of classification.

So far, scholars have proposed many different network flow classification techniques. These technologies are mainly divided into four categories: port-based, deep packets inspection (DPI)-based [3], machine learning (ML) [4–6], and deep learning methods (different from traditional machine learning methods, deep learning is listed separately for better discussion). On the one hand, due to the development of network technology itself, classification techniques used previously such as port detection are no longer adequate for the current network flow classification. Along with the importance of data privacy, deep packet inspection is no longer favored by researchers and engineers. With the rise and vigorous development of artificial intelligence technology, intelligent classification technology has become an important direction for researchers. Network flow classification technology based on machine learning and deep learning has emerged as the main method of current classification research. This study summarizes the different methods based on machine learning and analyzes the main methods of deep learning to propose a multimodal framework, which not only improves the classification accuracy but also enhances the stability of the model. The main contributions of this study are as follows:

- (i) The network flow classification based on spatial features and time series features is studied by using visualization methods
- (ii) A multimodal network flow classification method is proposed, which integrates different network flow features to improve the stability and accuracy of the classification model
- (iii) a comprehensive analysis of the differences in structure and training methods of the two types of models in multimodal framework and their advantages and disadvantages and solid experiments on the ISCXVPN2016 dataset

The study is structured as follows. The second part summarizes the related work in the field of network flow classification, and the third part discusses the research methodology, including the data processing, the structure of the framework, and differences between them. The fourth part is the comparison of experimental results and analysis. The fifth part is the conclusion and future work. For the convenience of writing, the acronyms used in this article are listed in Table 1.

2. Related Work

As the basic task of many network services, network flow classification has always been the focus of research in academic and engineering fields. So far, network flow classification technologies are mainly divided into four categories, namely, port-based, deep packet inspection, machine learning, and deep learning method.

TABLE 1: List of the acronyms used in this study.

Acronym	Definition
QoS	Quality of service
RNN	Recurrent neural network
DPI	Deep packets inspection
GRU	Gated recurrent unit
FCBF	Fast correlation-based filter
CV	Cross validation
LDA	Linear discriminate analysis
CONV	Convolutional layers
SAE	Sparse autoencoder
NN	Neural network
FC	Fully connected layers
CNN	Convolutional neural network
ML	Machine learning
LSTM	Long and short-term memory
SFS	Sequential forward selection
RTT	Round-trip time

The earliest network flow classification technology is to use port number of UDP or TCP protocol at the transport layer. This method is easy to implement with lower algorithm complexity, so it is often used to detect certain specified port applications. However, with the diversification of applications and protocols, as well as the emergence of port hopping and port masquerading technologies, this method is no longer reliable and can only be used as an auxiliary method. Many current network applications use port numbers that are different from common ports to bypass operating system access control permissions [7], while some other network applications encapsulate different services into well-known streams such as HTTP protocol-based streams or conversation, these operations usually reduce the accuracy of port-based network flow classification [7]. In fact, Madhukar and Williamson [8] proved that nearly 70% of network flow cannot be classified correctly using the port-based identification method.

DPI refers to the identification of the unique fingerprint characteristics reflected in the payload of each packet and then the detection of specific network flow [9, 10]. If the payload of the network flow and the known application or protocol can match in certain features, then it can be considered that this network flow is the known application or protocol with a high probability. For example, some traditional load data fingerprint features include: “\ GET”-http, “0×13Bit”-Bit torrent p2p, “PNG”0×0d0a-MSN messenger, “USERHOST”-IRC, “ARTICLE”-nntp, and “SSH”-ssh Internet traffic [11]. Compared with port-based method, its accuracy is greatly improved. Although the DPI method is very accurate for the network flow classification, it also requires scientific research staff to extract characteristic fingerprints of network flow, and to maintain and update the existing fingerprint database from time to time, which is a very resource-consuming task.

In recent years, researchers tried to use the statistical characteristics of network flow and machine learning algorithms to classify network flow. Different network flow will produce different traffic characteristics that can be used to distinguish, such as the distribution of data packet size,

data packet arrival time, flow length, and flow duration. Among them, Moore and Zuev [12] proposed a method based on the naive Bayes principle which builds a Bayesian classifier for supervised learning, combining with the fast correlation-based filter (FCBF) algorithm and kernel estimation technology, the method can achieve 95% accuracy. In [13], authors used nearest neighbour and linear discriminate analysis (LDA) to classify various applications. The experimental results showed that supervised ML algorithms are also able to separate traffic into classes with encouraging accuracy. In [4], authors used Bayesian network, C4.5 decision tree, naïve Bayes, and naïve Bayes tree methods to give common features set for classification with different feature selection algorithms. In [14] authors used a variety of different algorithms to filter the wrong label data in the original data set to obtain a more accurate training data set, and used machine learning to retrain the filtered data to obtain a more accurate and stable classifier. In [15], authors extracted the unique characteristics of various application during the information commutation, and realized a lightweight classification method for network application. Although the classic machine learning has solved the problems that cannot be settled with methods based on the port or DPI, it also faces many new problems. The first problem is feature selection, as machine learning methods rely on manually and expert-originated feature sets, which requires a lot of manpower to choose a feature set by themselves. The second problem is feature extraction, as the feature set that performs well for a specific data set does not have universal applicability in practice.

With the rapid development of deep learning in the field of artificial intelligence, researchers have tried to transfer deep learning methods that shine in computer vision processing, natural language recognition to the field of network flow classification. In [16], authors used neural network (NN) and sparse autoencoder (SAE) network to classify specific network protocol traffic and achieved a high accuracy. In [17], authors explored online traffic detection methods. In this study, the basic idea is to employ a compact nonparametric kernel embedding based method to convert early flow sequences into images which can be trained in convolutional neural network (CNN) and its accuracy exceeds 99%. Classification tasks can also be accomplished using network flow sequential information [18]. In [19], authors investigated the classification and prediction performances of LSTM networks, using real server-generated traffic streams, experiment result showed that LSTM is able to classify and predict the occurrence of highly intensive traffic flows accurately. In [20], authors used CNN LSTM network and their various combinations to detect network flows, which the classification accuracy for applications reached 96%. In [21], CNN and CNN & LSTM was used to classify mobile applications where the payload of the first few packets were mainly used to achieve high accuracy. In [22], authors focused on three practical problems which are network bandwidth, network flow duration, and network flow detection and then proposed a multitask training method, that is, first used the CNN network to train the network bandwidth and duration tasks and then trained the

network flow classification task. Based on this training method, it had achieved better result than the previous CNN & RNN method. In [23], authors proposed to introduce the capsule network into the field of network flow classification, and combined the advantages of CNN & LSTM network to achieve high accuracy of network classification. Giuseppe Aceto etc. in [24] provided a wide experiment analysis based on multimodal framework (CNN + LSTM) for classification of encrypted mobile traffic. This work provides guidance for the subsequent exploration of multimodel fusion. Then, in [25], authors further proposed a novel multimodel multitask deep learning approach and DISTILLER classifier, it can solve different traffic classification simultaneously. Liu et al. [26] proposed a method which applied RNN to encrypted traffic classification. Moreover, the framework added a multilayer structure which can explore sequential characteristics deeply and experiment results outperformed the state-of-the-art methods. In [27], authors tried to use explainable artificial intelligence to improve multimodel behavior, the experiment results showed that the proposed method provide global interpretation, rather than sample-based ones. Table 2 lists an overview of the above literature citations.

3. Research Methodology

3.1. Dataset. In this article, we used two different datasets including the USTC dataset provided in [28] and the ISCXVPN2016 dataset [29] provided by the Canadian Institute for Cybersecurity. The USTC dataset has 10 categories of normal traffic such as FaceTime and Gmail generated using IXIA BPS (Professional Traffic Simulator); this study will use this dataset for feature analysis of network flow classification. The ISCXVPN2016 dataset is captured at the University of New Brunswick which contains raw pcap files of several traffic types. The dataset provides labels with different categorization, such as AIM chat, Gmail, Facebook, chat, and streaming . The ISCXVPN2016 dataset is publicly available for researchers, and this study will use this dataset to conduct experiments and compare the experimental results. For more details on the captured traffic and the traffic generation process, refer to [29].

3.2. Method Background. In the following sections, the background of the proposed framework is presented.

3.2.1. Feature Selection and Classification. Network flow has an obvious hierarchical architecture: according to the general TCP/IP system structure, the network flow is packaged into data units in different layers which is unique to each layer. The frame of the data link layer is the lower-level data that can be studied which receive the data frame from the upper layer and disassemble it into data in the form of bit stream. Therefore, the frame contains different types of features in network flow that can be distinguished, so it is very important to take the frame of data link layer as the basic research object for network flow classification. In practice, the frame is easy to obtain, and all the protocol packets can be directly captured which are passing through the

TABLE 2: Summary of methods in the literature employing machine and deep learning models.

Ref.	Model	Data	Eva metrics	Contribution
[12]	Naive bayes	—	Accuracy	A vast improvement over traditional techniques
[13]	Nearest neighbour (NN) and linear discriminant analysis(LDA)	WAND	Error rate	Using traffic traces from a variety of network locations, demonstrate the feasibility, and potential of the approach
[14]	Noise elimination, random forest(RF)	ToN and ISP	Accuracy, $F1$	The framework delivers consistently superior performance to other traffic classification schemes in the presence of unclean training data
[15]	A classifier based on Weka’s classifiers library	—	Recall, precision, accuracy	Authors suggest a fingerprint that is based on zero-length packets, hence enabling a highly efficient sampling strategy
[16]	NN and SAE	TCP flow data	Precision, recall	The approach solves the problem of nonautomation and poor adaptation in traditional ways
[17]	CNN	Data including 5 protocol and 5 application	Accuracy	Propose a nearly end-to-end framework for online IP traffic classification
[19]	LSTM	Real server-generated traffic	Accuracy	The LSTM NNs prove to be a highly efficient computational model capable of solving real server-generated traffic
[20]	CNN and RNN	Internet of things traffic	Recall, precision, accuracy, $F1$	The study shows the performance of CNN and RNN models and a combination of them
[21]	CNN and LSTM	Mobile traffic	Recall, precision, accuracy, $F1$	Introduce two deep learning models for mobile app identification
[22]	RF, CNN, RNN, multitask learning	QUIC and ISCX	Accuracy	Multitask learning approach out-performs, or performs as accurately as the transfer learning
[23]	Capsule network	UTSC-2016	Recall, precision, accuracy, $F1$	This study proposed an end-to-end traffic classification method and used the capsule network model for traffic classification
[24]	CNN, LSTM, SAE	Encrypted mobile traffic	G-mean, accuracy, $F1$	This study provided a wide experiment analysis based on multimodel framework (CNN + LSTM) for classification of encrypted mobile traffic
[26]	RNN, autoencoder	Encrypted traffic	True positive rate, false positive rate, FTF	This study provided the framework containing a multilayer structure which can explore sequential characteristics deeply and import the reconstruction mechanism which can enhance the effectiveness of features
[25]	CNN, RNN	ISCX VPN-nonVPN	Accuracy, $F1$	This study proposed a novel multimodal multitask deep learning approach and DISTILLER classifier, it can solve different traffic classification simultaneously
[27]	CNN, RNN	MIRAGE-2019	Accuracy, $F1$, G-mean, precision	This study used explainable artificial intelligence to improve multimodel behavior, the experiment results showed that the proposed method provide global interpretation, rather than sample-based ones

network card. For example, using wireshark and tcpdump can capture any data packet of interest. In the field of traffic classification, the usage of machine learning methods based on traffic characteristics has greatly improved the accuracy of classification compared with the previous methods [12, 30]. Research on this type of method shows that the key to improve the accuracy of network flow classification lies in the usage of a suitable classifier and the ability to design a flow feature set which is based on different types of traffic that can meet the classification specifications as shown in Figure 1.

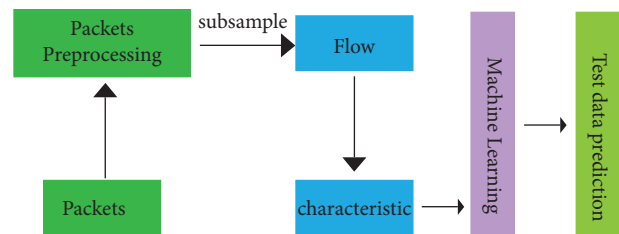


FIGURE 1: The processing of machine learning.

3.2.2. *Spatial and Sequence Features.* This part mainly analyzes the spatial and sequence features of network flow. In [16], the author proposed that the application of deep

learning methods can realize the automatic extraction of network flow features, which is more suitable for the classification requirements than the manually and expert-

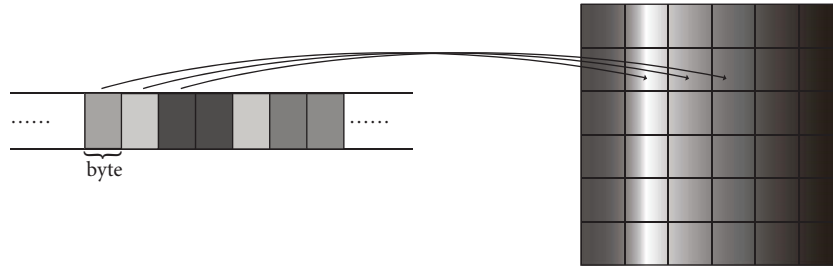


FIGURE 2: Schematic diagram of network flow transformed into a grayscale graph.

originated features. In the article, authors used NN and SAE network to extract and classify the features of the processed network flow. Compared with the previous machine learning algorithms, the accuracy has been greatly improved. In the data preprocessing stage, the data packet obtained by the data link layer is used as the processing object, and packet is represented in the form of byte stream with the usage of the data packet extraction tool. After preprocessing, the data is sent to the neural network for training with the purpose of automatic extraction of feature.

Inspired by the successful application of CNN in the field of image processing, the authors in [28] stated that the network data can be represented by a matrix as shown in Figure 2, that is, the flow sequence $F_p(x)$, transformed into a matrix M^P , can be expressed as

$$F_p(x) \longrightarrow M_{ij}^P (n \times n), x = n \times i + j. \quad (1)$$

In this way, there is a one-to-one correspondence between the specific matrix M^P and the network flow $F^P(x)$. For each node M_{ij}^P of the matrix, the value range is (0–255), which is the same as the range of each pixel value of the grayscale image, so there is a one-to-one correspondence between the matrix M^P and the gray image T_p . The network flow classification is transformed into grayscale image classification and network flow feature extraction is transformed into grayscale image spatial feature extraction.

It should be noted that although the abovementioned network flow classification task can be transformed into image classification, the extracted features are only the feature representation in the network flow graph, not the characteristics of the network byte stream, however, graph structure information is still very useful for feature analysis in this paper. On the one hand, if the feature is a unique feature of the network flow, it must be expressed in the form of a specific pixel in the map to form a specific spatial structure and this is the basis for the CNN to extract the spatial features of the network flow. On the other hand, the area formed by the feature also reflects the focus of the model in the classification task, which provides a reference for the analysis of classification features in this paper.

(1) *CNN Model Construction.* CNN is widely used in the field of image pattern recognition. It is a kind of deep learning model which contains a large number of convolution operations. A complete CNN includes several convolutional layers (CONV), pooling layers (POOL), and fully connected layers (FC). The common architecture patterns are shown as follows.

$$\text{INPUT} \longrightarrow [(\text{CONV}) \times N \longrightarrow \text{POOL}] \times M \longrightarrow [\text{FC}] \times K. \quad (2)$$

The parameters of the convolutional layer are composed of some learnable filter sets (convolution kernels), which can capture the image features of the previous output layer, and another layer, pooling layer which is mainly responsible for subsampling. At last, a set of fully connected layers are often used to capture high-level features of an input.

This article uses the above architecture to learn features and classify the processed network flow which contains four convolutional layers (conv2d), two pooling layers (max pooling), three dense connection layers (dense) and one flattening layer (flatten), each data transformation in the model uses a normalization process (batch normalization).

(2) *HAN Model Construction.* Compared with converting the network flow to the graph and extracting the spatial feature information to complete the classification task, it is more straightforward to use the recurrent network to classify the network flow and extract the sequence information features of the network flow. The experiments in this section refer to the processing ideas of [31], mainly classify the network flow through the hierarchical attention network (HAN) [32], and display the feature distribution characteristics in the classification process through visualization technology.

This section adopts the byte-packet-stream processing mode, that is, a network flow label corresponds to a three-level data flow, this obvious hierarchical data structure is similar to the structure of token-sentence-article. The processing mode of network flow classification is shown in Figure 3.

This experiment uses the USTC dataset and divides the dataset into training set, validation set, and test set, which account for 60%, 20%, and 20% of the resampling data set, respectively.

Table 3 lists the classification results of the two models on the USTC test set. It can be seen from table that the two types of models have achieved high accuracy on the classification task, and the models perform well.

This study randomly selects 120 samples in the test data for testing and uses Grad-CAM [33] technology to visualize spatial features. The visualization results are shown in Figure 4; among them, red represents the feature with higher activation degree and blue with lower activation degree, and

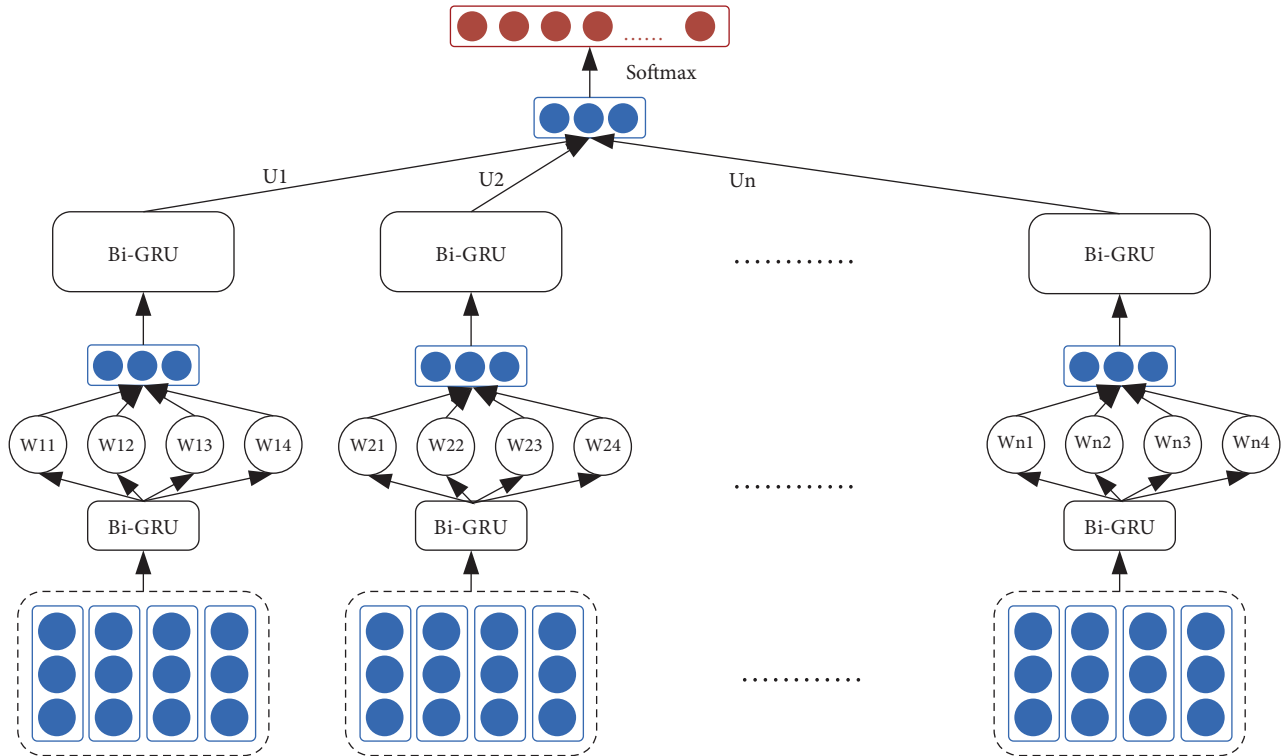


FIGURE 3: HAN for network flow classification.

TABLE 3: Accuracy (%) of models on the test set.

Type	HAN	CNN
BitTorrent	99.92	99.75
FaceTime	99.92	100
FTP	100	99.92
Gmail	99.92	99.33
MySQL	100	100
Outlook	100	99.92
Skype	100	100
SMB	100	100
Weibo	100	99.92
WoW	100	100
Average	99.98	99.88

in the grayscale image, 0 represents black and 255 represents white. Through visualization, we can see the distribution of important features of network flow classification in the CNN model.

From the class activation map, it can be seen that most of the features involved in network map classification are concentrated in the network flow header information, some of the features used for classification are concentrated in the tail of the data, and a small part of the map information also includes features in the middle. From the comparison of the original grayscale images, it can be seen that for MySQL, the black tail indicates that there is no network flow data in the current area, and the reason why the data information is not used for feature extraction is that MySQL does not have a unified representation in the grayscale image space., that is,

feature extraction is more difficult. Compared with the information in the data, the structural information displayed by the map itself is more prominent and obvious, so the features used for classification are concentrated in the tail. A small number of features in the middle also show this feature distribution. For example, the middle features of Gmail are mostly concentrated on the boundary between the blank data area and the data area, which also reflects the unique feature structure information of the graph. Furthermore, in the WoW class, although the network graph also presents obvious structural features, the structural features of the information features of the WoW class are more obvious than those of the blank data area.

Next experiment randomly selects 10 samples of each type of network flow for attention visual analysis of HAN model. Table 4 lists the visualization results of 10 types of network flow in the data set. The packet attention column represents the weight value of the packets in the network flow (the shade of red indicates the value), and the byte attention column represents the internal data of each network stream in hexadecimal, and identifies the weight value corresponding to each byte, in which blue indicates that the weight is larger, and green indicates less weight.

It can be seen from the table that when using sequence features to classify network flows, it shows different characteristics from spatial features in convolutional networks. When there are multiple data packets in the same flow, there is a specific packet that has a greater impact on the final classification, but the sequence characteristics of other packets with less impact are similar to those of packets with

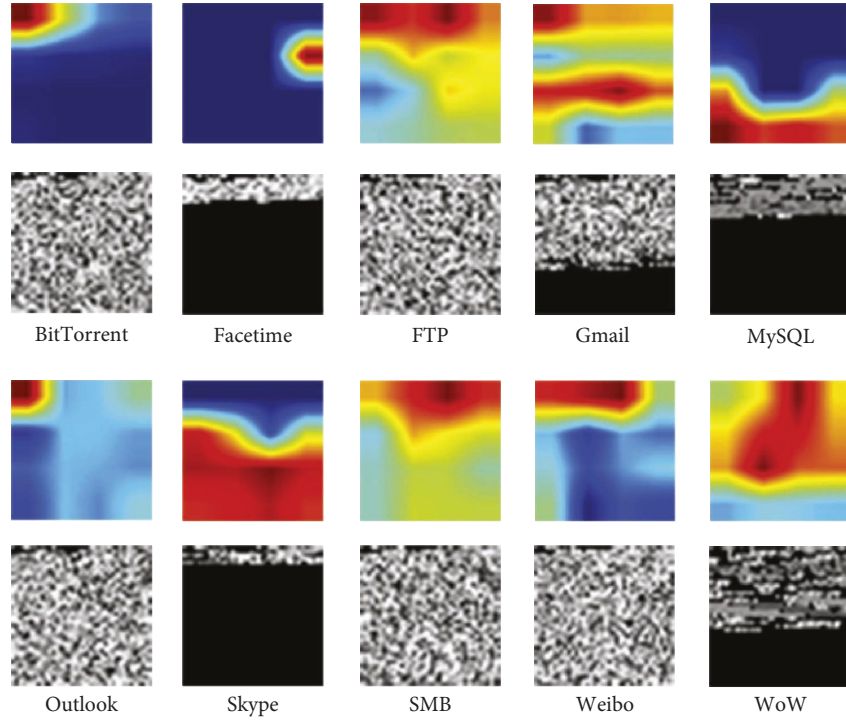


FIGURE 4: Class activation map for network flow classification.

TABLE 4: HAN attention visualization.

Type	Packet No.	Packet attention	Byteattention									Information
BitTorrent	1	1	...	0x1	0x1	0x8	0xa	0x11	0xf6	0x54	...	tcp.option_kind, tcp.option_len
Facetime	1	1	...	0x45	0x0	0x1	0x12	0xd8	0xb	0x40	...	ip.len, ip.hdr_len, data
FTP	3	0.98	...	0x0	0x4	0xe4	0xe	0x84	0x40	0x0	...	ip.len, ip.id
Gmail	2	1	...	0x1	0x1	0x8	0xa	0x5	0x4a	0x62	...	ip.len
MySQL	1	1	...	0x65	0x72	0x73	0x8	0x70	0x61	0x73	...	data
Outlook	1	1	...	0x1	0x1	0x8	0xa	0x1e	0xb6	0xc3	...	timestampvalue, tcp.option_kind
Skype	1	1	...	0x1	0x1	0x8	0xa	0x10	0x6d	0x15	...	tcp.option_kind, tcp.option_len, tcp.dstport
SMB	1	0.74	...	0x1	0x1	0x8	0xa	0x8e	0xab	0xf5	...	tcp.option_kind, tcp.option_len
Weibo	3	0.87	...	0x45	0x0	0x5	0xdc	0xc0	0x40	0x40	...	ip.hdr_len, ip.len, tcp.option_kind
WoW	1	1	...	0x1	0x1	0x8	0xa	0x6	0xc4	0x5f	...	tcp.option_kind, tcp.option_len, ip.hdr_len

greater impact. As shown in the table, the information column lists the bytes with larger weights. The top ones include tcp.option_kind, tcp.option_len, ip.hdr_len, ip.len and other bytes. These features are related to network flow environment variables.

By visualizing the features of the two types of models, it can be found that the spatial features based on the network flow graph are similar to the network flow sequence features. From the perspective of feature distribution, the features with high weights are located in the first half of the network flow data. At the same time, there are differences between the two types of features. When the spatial features of the data are not enough to distinguish the types of network flows, the model prefers to use the spatial structure information of the graph, but the sequence information can only be extracted from the network

flow itself. Therefore, both spatial features and sequence features can be used for network flow classification tasks, and the two types of features are distinguished from each other.

3.3. Proposed Classification Method. After obtaining the spatial and sequential features of the network flow, a natural idea is whether the classification accuracy can be improved if both types of features are used in the classification task. Based on the above two types of models, this study proposes a multimodal framework that uses the two types of features to classify the various types of network flow.

In [22, 34, 35], the authors proposed that for the same network, finding the best set of auxiliary tasks will improve the traffic classification which should be treated similar to

hyper-parameter tuning. The abovementioned idea of multitask training can be expressed as: for the same network, using similar tasks to train separately, which can improve the efficiency of the network to complete the final task. Then, on the contrary, for the same task, there is a way to use different combination of methods to improve training efficiency of the target task.

3.3.1. Multimodal Network Flow Classification Model. Through the abovementioned analysis, we can extract spatial features through the CNN model and sequential features through the GRU model to build the multimodal framework. There are two main ways to build the framework.

(1) Model Pretraining (Model A): The Framework Based on Pretraining. The framework based on pretraining refers to: train the networks, respectively, and select the characteristics preliminary, then integrate the selected features of each network and send them to the secondary network for further screening, as shown in Figure 5.

The model in Figure 5(a) shows the structure of the multimodal classification model based on pretraining. First, the pcap file is segmented, and the data is pre-processed to form the input data format file (image and byte stream), and then sent to the convolutional layer and the downsampling layer to extract and simplify the data features. The spatial features and sequence features of the network flow are extracted through the GAP layer and the GRU layer, respectively, and the extracted two types of features are then sent to the feature fusion module to form fusion features. The dense layer and softmax are used for output of classification.

Figure 5(b) shows that the features used in Figure 5(a) are extracted from the two submodels through pretraining. Figure 5(c) shows that the model parameters in the first half of the model in Figure 5(a) are actually frozen and do not participate in the training of the entire model. The training part is mainly the parameters of the feature fusion part.

Strategy of combination: during the training processing, the features of the same type will be completely extracted layer by layer, at last, it will focus on the features useful for the task. For example, in the field of image recognition and classification, with the usage of heat map [33], it is easy to find the important feature which will be helpful for the final task. However, this type of feature set is still redundant for the entire training task. Each feature in the mixed feature set does not necessarily contribute to the final classification task and same feature may have different weights in different models. Therefore, it is necessary to further filter the extracted features. In order to filter the combined features, this article adds a layer of weight learning to the second step to realize the automatic assignment of the weight of each feature.

Assuming that the feature set is θ , and each feature is represented as θ_i , the weight of θ_i can be calculated by the following equations:

$$w_i = \tanh(u_i \cdot \theta + b_w), \quad (3)$$

$$\alpha_i = \frac{\exp(w_i)}{\sum_t \exp(w_t)}. \quad (4)$$

We calculate the dot product of the weight and the original feature and recombine it into a new feature set θ^r :

$$\theta^r = \text{concatenate}[\alpha_i \cdot \theta_i]. \quad (5)$$

The classification task is implemented through a fully connected layer and softmax layer.

Based on the abovementioned the framework design, on the one hand, it is beneficial for the framework to filter feature sets automatically, which meets the processing requirements to allocate different feature sets for different types of network flow. On the other hand, by analyzing the attention of each feature, we can further study the importance of each feature to the classification task.

In summary, two types of features are extracted from the trained network previously, after integrating the extracted features, they are sent to the second learner for training again to complete the final task. This training method needs to be divided into two steps. The performance of the first step learner directly affects the second.

(2) Model Joint-Training (Model B): The Framework Based on Joint-Training. The main idea of this method lies in the combination of models, that is, the characteristics learned by the two types of models are directly combined in one network to construct a wide and deep large-scale network model as shown in Figure 6. The extracted features are more diverse and accurate than that of a single network, and can be directly used for classification tasks. The framework only needs a single-step training to obtain a useable model.

(3) Comparison of Two Types of Frameworks. Although the abovementioned models are based on the idea of integration of mixed feature, they are very different in the way of framework construction and training method.

Framework Construction. Model pretraining can actually be divided into three models, including two basic models, namely, the CNN model for extracting spatial features and the GRU model for extracting sequential features, and a secondary model. In the secondary model, it is necessary to design a proper extraction strategy of input data. In this study, the attention-like mechanism is used to realize the automatic learning of feature weights. Model joint-training is one model essentially. The characteristic of model joint-training is wide and deep, that is, in terms of width, the integration of multiple models is adopted to expand the longitudinal direction of the framework, and for the depth, the feature extracted by the basic model is relearned and trained to expand the horizontal direction of it.

Training Method: Model pretraining is divided into two steps in the training method. The first step is to train basic models to complete the preliminary feature extraction. The second step is to send the extracted features to the secondary model to complete the final training task. In fact, the above

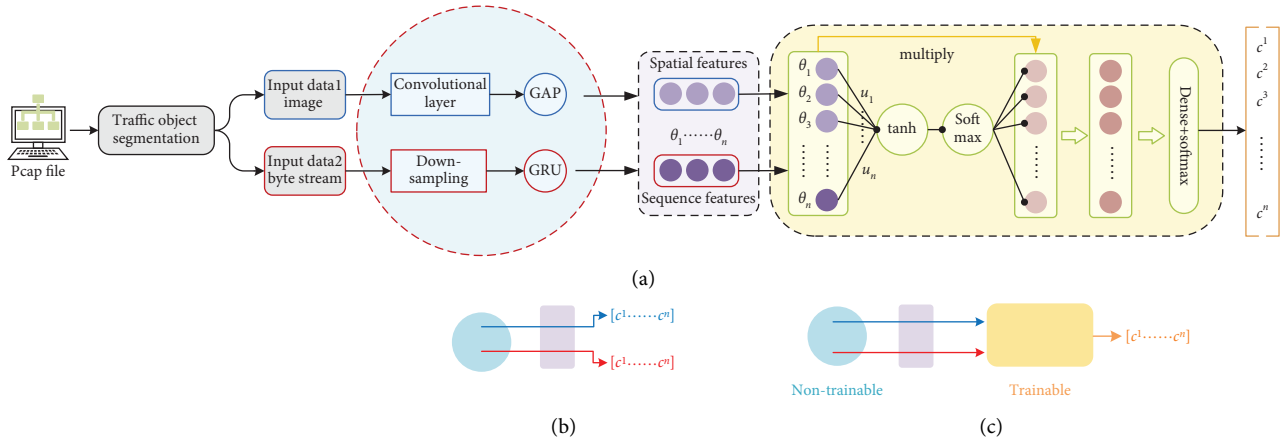


FIGURE 5: The framework based on pretraining. (a) Pretraining model, (b) pretraining, and (c) parameter training.

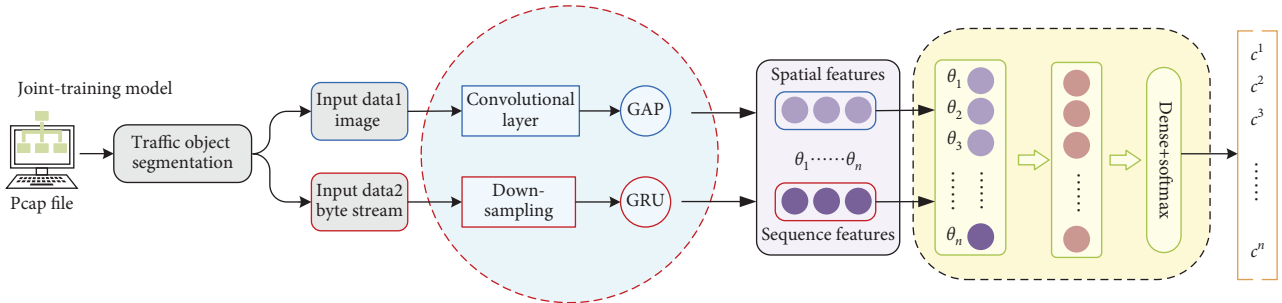


FIGURE 6: The framework based on joint-training.

training can be referred to as one training step, as the training task is completed on the secondary model finally, not on the basic model. Model joint-training only needs one step for training, that is, sending the data to the network for training directly, which is an end-to-end training model actually. Considering training task in practice, although model pretraining needs to complete the training of three models, the basic model can be trained in parallel and separately, which is more flexible for actual operations. In contrast, it seems that model joint-training only needs one step to train, actually, the time and hardware parameters required in the training environment are higher than those of model pretraining due to the high complexity of joint training and the inability to perform parallel processing.

Notably, it is worth to mention that the multimodal framework idea is different from the ensemble strategies, which is widely adopted in machine learning competitions. Ensemble strategies improve the efficiency of the ensemble model through reducing the deviation and variance between basic models by adjusting the data set or combination of training result, like boosting integration [36, 37], bagging integration [38] and stacking integration. The multimodal framework is the integration of extracted patterns from different dimensions to produce the final result. It is to integrate the data from different perspectives to make the collected information more comprehensive with assigning different weights

according to different features automatically that making the framework more robust and efficient.

3.3.2. The Framework Cross-Validation Criteria. In order to verify the reliability of the mentioned framework, we used k -fold cross validation on the training data to conduct experiments on different data sets. Specific algorithm 1 implementation is as follows.

3.3.3. The Framework Evaluation Criteria. To evaluate the performance of the multimodal framework, we have used accuracy (Ac), recall (Rc), precision (Pr), and F_1 score (F_1) metrics [39–42]. The abovementioned metrics are described mathematically as follows:

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} 100\%, \quad (6)$$

$$Rc = \frac{TP}{TP + FN} 100\%, \quad (7)$$

$$Pr = \frac{TP}{TP + FP} 100\%, \quad (8)$$

$$F_1 = \frac{2 \cdot Rc \cdot Pr}{Rc + Pr} 100\%, \quad (9)$$

```

(1) Begin:
(2) Break the data into  $k$  folds. Data capacity per fold is  $N$ ;
(3) Construct model joint-training based on CNN model and GRU model;
(4) for each  $i$  in  $[0, k]$  do
(5)   validation data = data  $[N: (i + 1) \times N]$ ;
(6)   training data = rest of data;
(7)   train CNN model  $[i]$  on the training data, observe it on the validation data;
(8)   train GRU model  $[i]$  on the training data, observe it on the validation data;
(9)   construct model pretraining  $[i]$  based on CNN model  $[i]$  and GRU model  $[i]$ ;
(10)  train model pretraining  $[i]$  and model joint-training;
(11)  Model pretraining  $[i]$  prediction on the test data;
(12)  Model joint-training prediction on the test data;
(13) end for
(14) End

```

ALGORITHM 1: k -fold cross validation used to test the stability.

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

4. Experiment

The experiment is divided into four steps.

- (i) Converting raw data to trainable data
- (ii) Network flow spatial feature learning, mainly using CNN to classify the grayscale images
- (iii) Network flow sequential feature learning, mainly using GRU network to classify network flow digital sequences
- (iv) Hybrid feature learning, which uses the multimodal framework to classify network flow

The TensorFlow [43] is used as an experiment software framework that runs on Windows 10 home edition with Intel(R) Core (TM) i5-9300H CPU @ 2.40 GHz and 8 GB memory. An Nvidia GeForce GTX1650 GPU is used as an accelerator. The mini-batch size is 256 and the cost function is categorical cross-entropy. Adam optimizer built-in TensorFlow is used as an optimizer, training time is about 70 epochs.

4.1. Dataset and Preprocessing. In this article, we used the ISCXVPN2016 dataset [29] mentioned in A. In order to better compare the experimental results, this study relabels the dataset according to the classification method of the literature [44]. Under-sampling is also applied according to the number of data set. Sampling is a simple method to tackle this problem. Hence, to train the proposed framework, using the under-sampling method, we randomly select samples of major classes until the classes are relatively balanced.

The dataset above is obtained from the data link layer. From hierarchical perspective, at the data link layer, the frame header information contains physical connection information, such as MAC address and other protocol content. The network transmission layer also contains IP address information. These data play a key role in network

stream transmission, but they cannot provide any valuable information in the field of network flow classification and even training networks will use the address information to classify the network flow, which is ridiculous in practice. Therefore, in the data preprocessing part, the MAC and IP addresses are directly removed to eliminate the impact on the training task due to different addresses.

The second step is file cleaning, which is to clean up duplicate network stream files and empty files to avoid bias when training the network.

Finally, due to the need of using deep learning network to train the data, the data length standard needs to be unified. TCP and UDP protocol headers have of different length. In order to unify the length of the transport layer, we inject zeros to the end of UDP segment's headers to make them equal with TCP headers. Finally, according to the literature [16], most of the valuable information is at the header of payload data. In this article, the first 1225 bytes (35×35) are intercepted as the research object in the ISCXVPN2016 dataset to balance the accuracy and simplicity of the experiment.

Through the analysis of the number of data samples, we found that it is very different with the number of samples of various types of data. In the literature [45, 46], experiment results show that the performance of the learner will decrease due to the unbalanced number of samples, and this study will adopt a random sampling method until the number of various flows is relatively balanced. According to the network flow generated by different application types, this study relabels the data set and divides the network flow into 17 categories, as shown in Table 5.

The dataset is divided into the training set, validation set, and test set, which respectively account for 60 %, 20 %, and 20 % of the resampled dataset.

4.2. Results of the Multimodal Framework. Since the framework combines CNN and GRU, it is necessary to convert the network flow into trainable grayscale images and digital sequences, respectively. In the spatial and sequential features section above, we discussed how to convert the

TABLE 5: Type of network flow in ISCXVPN2016.

AIMchat	E-mail	Facebook	ftps	Gmail
Hangouts	icq	Netflix	Scp	Sftp
Skype	Spotify	tor	Torrent	Vimeo
Voipbuster	Youtube			

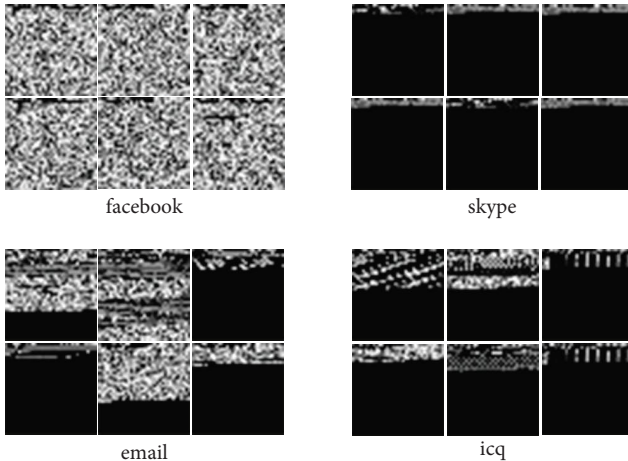


FIGURE 7: Grayscale images of different types of network flows.

network flow into a matrix and then into a grayscale image in [28], which is the input into CNN for training. Figure 7 shows part of the grayscale images of ISCXVPN2016 data set after network flow conversion. Figure 7 lists visual pictures of part of the network flows which shows vividly that for different types of network flow, it is distinguishable for texture characteristics of the picture. The above analysis can infer that after the network flow is converted to grayscale image, the different network flows have distinguishable features, and the same types have consistent ones.

In the part of input data of GRU, similar to the text processing method, the value corresponding to each byte of the network flow is similar to the token after the text tokenization which can be called network flow token, then associating it with vector by using embedding method. These vectors are combined into a sequence tensor, which is input into the GRU.

Table 6 shows the performance of four types of models (CNN, GRU, model pretraining, model joint-training) on the test set. Experiments show that the multimodal framework has entirely extracted network flow characteristics to distinguish each application accurately.

In order to show the experimental results of the model on the test set in more detail, this study draws a heat map based on the prediction results of each type of network flow. At the same time, hierarchical clustering is used to analyze the relationship of each type of network flow [47]. This method uses Euclidean distance as the distance metric, average as the agglomeration method, and reveals the different relationships among the different types of network flows. Here, we just show the result of the model pretraining and joint-training, as shown in Figures 8, 9. The figures indicate four

TABLE 6: Four models' performance for the network flow classification.

Model	Ac (%)	Pr (%)	Rc (%)	F_1 (%)
CNN	98.52	98.55	98.50	98.52
GRU	99.21	99.24	99.19	99.22
Model A	99.41	99.46	99.38	99.42
Model B	99.45	99.47	99.45	99.46

models have achieved high classification accuracy. In particular, model pretraining and model joint-training have similar results in accuracy and are better than the basic model. The clustering reveals that there are similarities between AIMchat and ICQ, skype and e-mail. Furthermore, there are similarities between Gmail, AIMchat, and ICQ. This is consistent in practice, as both AIM and ICQ provide online chat functions, and skype and e-mail also provide chat services based on the online. The conclusion is similar to [44] but more accurate which reveals the true relationships among the different types of network flows. Network flow classification shows a certain relationship because of the functional similarity of the applications abovementioned, which just shows that the network flows are classified on the basis of extracting features accurately.

4.3. Comparison. In this section, we compare the experimental results of network flow classification based on the ISCXVPN2016 dataset in recent years. Among them, Yamansavascular et al. [48] used the k -NN method to classification, Lotfollahi et al. [44] used a method called "Deep Packet," namely SAE and CNN to classify network flow. It can be found from Table 7 that both model pretraining and model joint-training outperform two methods above.

It should be noted that Yamansavascular et al. used machine learning methods to implement classification based on manually and expert-originated feature sets. Lotfollahi M et al., model pretraining and model joint-training use deep learning methods. This article has analyzed the shortcomings of classification based on manually and expert-originated features, and automatic extraction of network flow features based on deep learning is more suitable for practical applications with the development of intelligence.

According to [44], we compared the results of the ISCXVPN2016 data set based on different machine learning methods, where the decision tree depth parameter is 2, random forests is four, regression (with $c=0.1$), and naive Bayes with default parameters. As shown in Table 8, combined with Table 6, we can find that the classification based on deep learning is better than that of various machine learning. This shows the power of deep learning tools, especially in processing big data tasks and is consistent with the analysis results of the III-B1.

Figure 10 further shows the experimental results based on the deep learning on the test set. It indicates that the two types of the multimodal framework outperform the method based on "Deep Packet" in network flow classification.

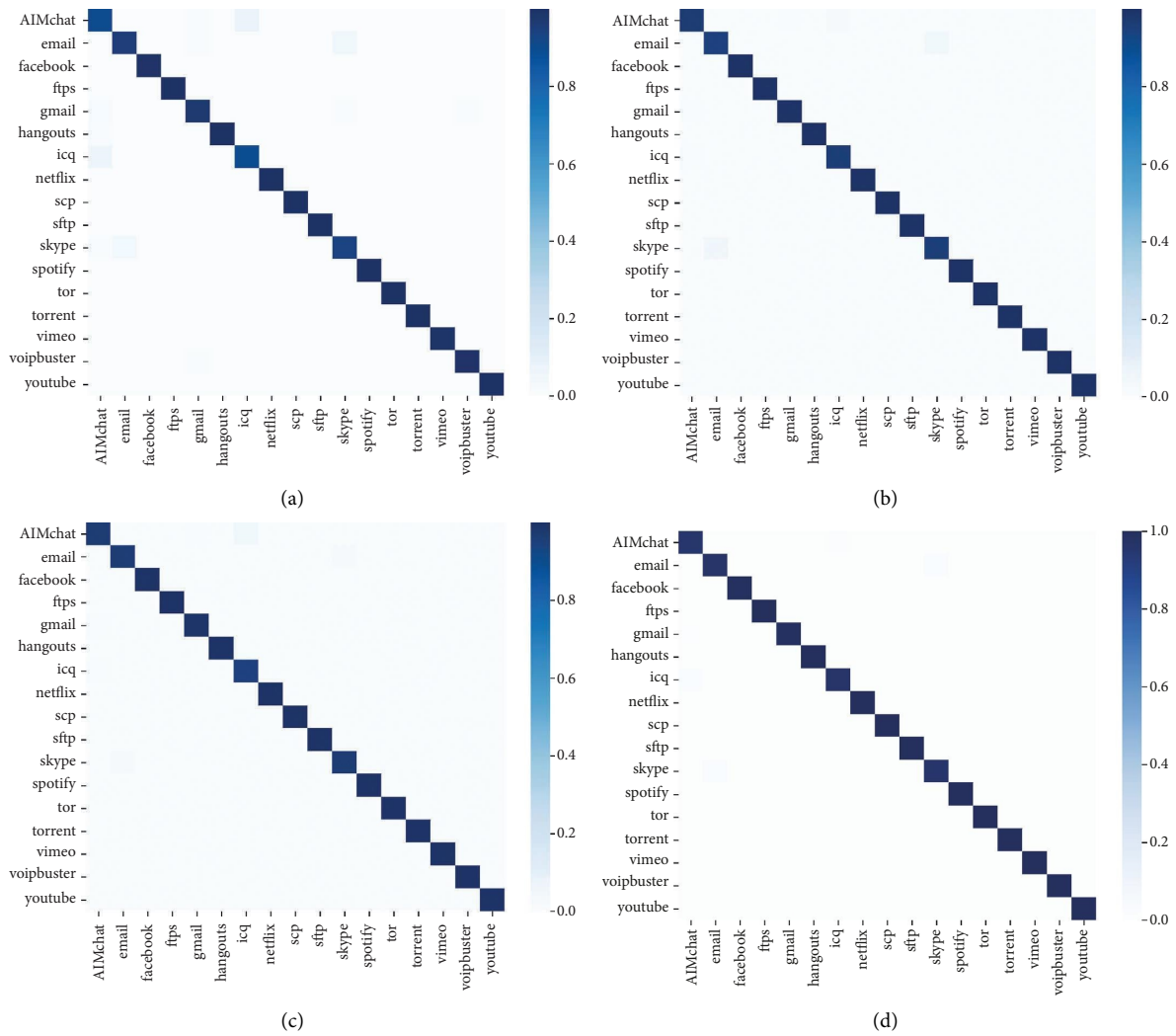


FIGURE 8: Heat map of test results for each model; the abscissa represents the true label of each type of network flow, and the ordinate represents the predicted label. The color represents the predicted probability. (a) CNN model, (b) GRU model, (c) model pre-training, (d) model joint-training.

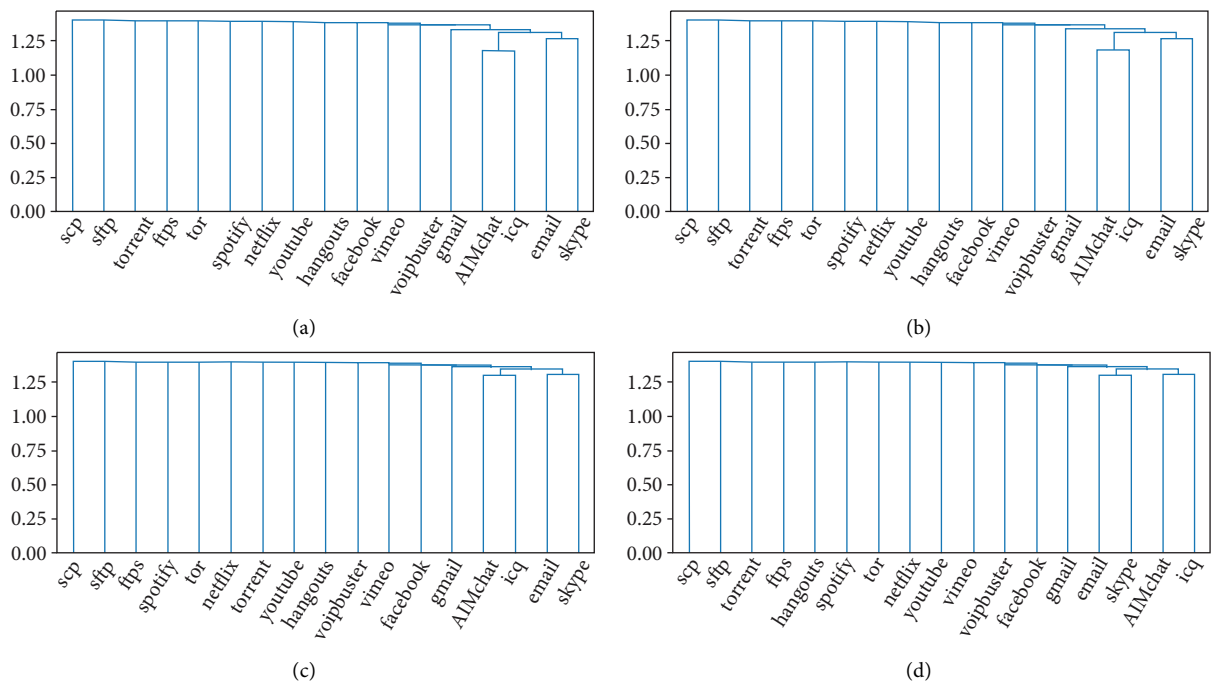


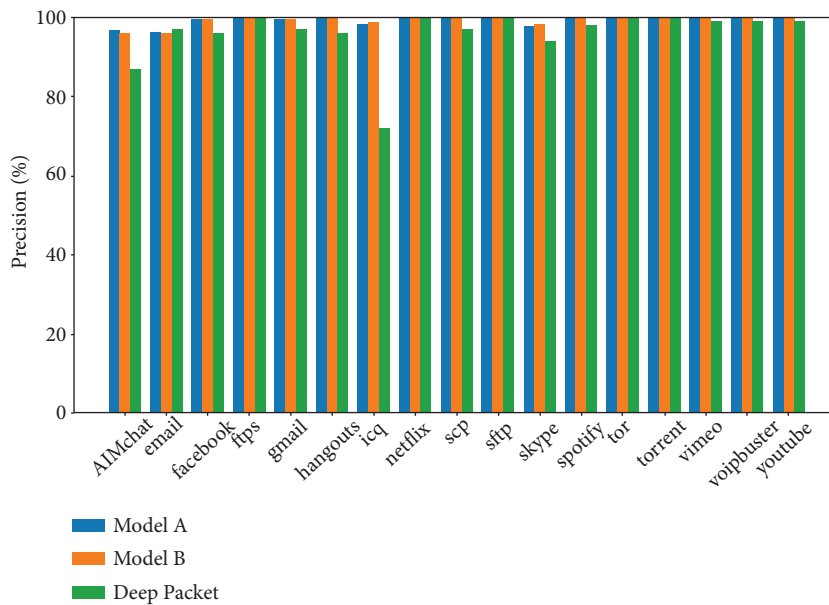
FIGURE 9: Hierarchical cluster diagram; the tree structure reveals the distribution of network flow clustering. (a) CNN model, (b) GRU model, (c) model pre-training, (d) model joint-training.

TABLE 7: The comparison between the multimodal framework and other methods on the “ISCXVPN2016” dataset.

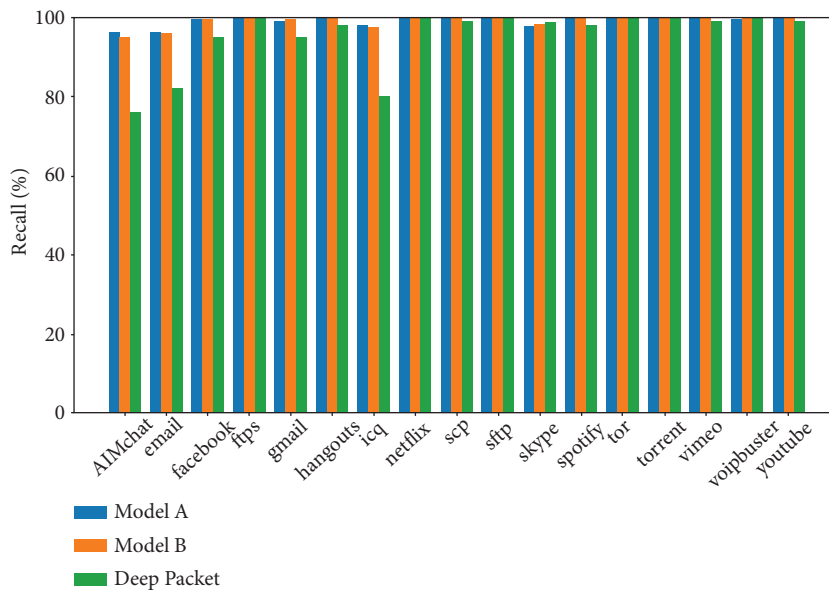
Study	Metric	Result (%)	Alg.
Yamansavascular et al.[48]		94	k -NN
Deep Packet (2020)		98%	CN
The multimodal framework	Accuracy	99.41	Model pretraining
The multimodal framework		99.45%	Model joint-training

TABLE 8: The comparison between the multimodal framework and other machine learning methods on the “ISCXVPN2016” dataset.

Method	Pc	Pr	F1
Decision tree	0.90	0.90	0.90
Random forests	0.91	0.90	0.90
Logistic regression	0.91	0.91	0.91
Naive Bayes	0.40	0.34	0.26

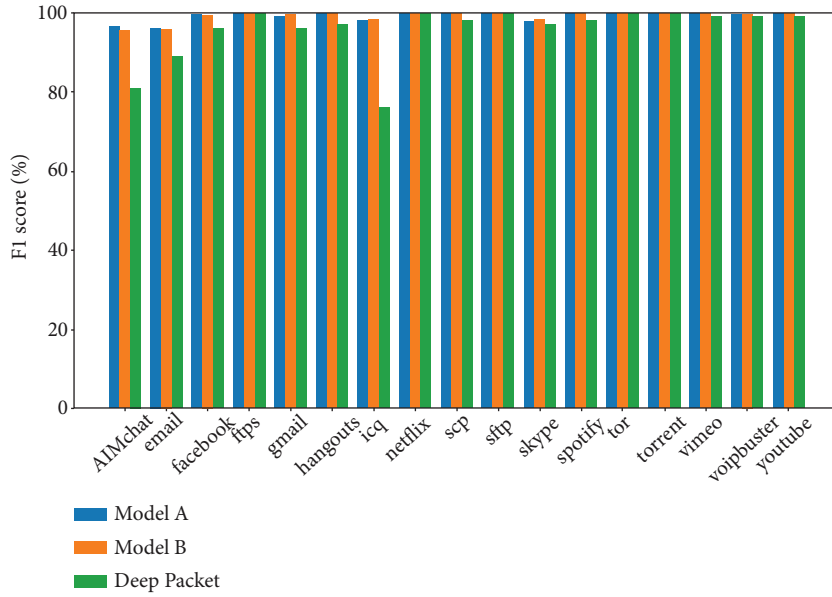


(a)



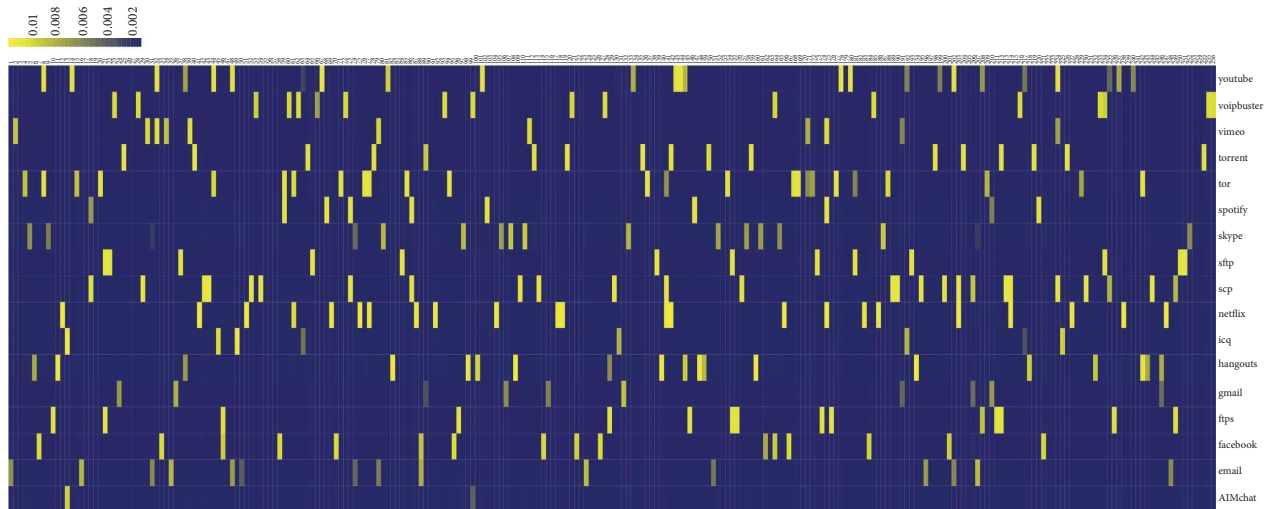
(b)

FIGURE 10: Continued.



(c)

FIGURE 10: Comparison of each class experiment result with the benchmark model. (a) Precision of models, (b) recall of models, and (c) F1 score of models.



(a)

FIGURE 11: Continued.

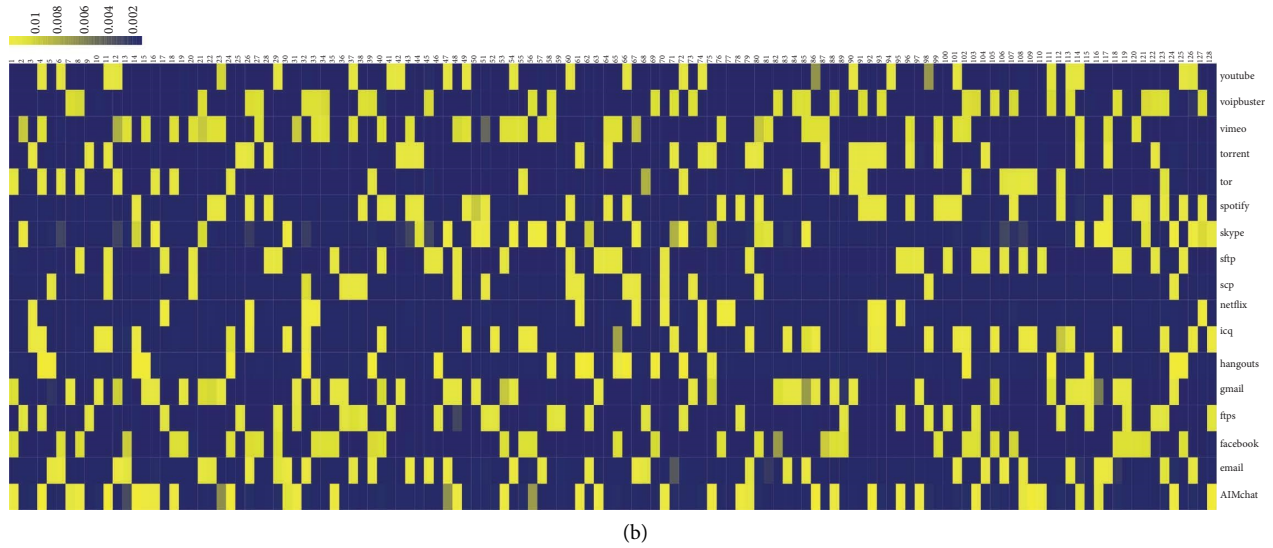


FIGURE 11: Distribution map of attention value of each type of network flow; the ordinate is the network flow type, and the abscissa is the number of extracted feature. (a) The spatial feature extracted by CNN; (b) the sequential feature extracted by GRU; the intensity of the color indicates the contribution of a certain feature to the final classification result.

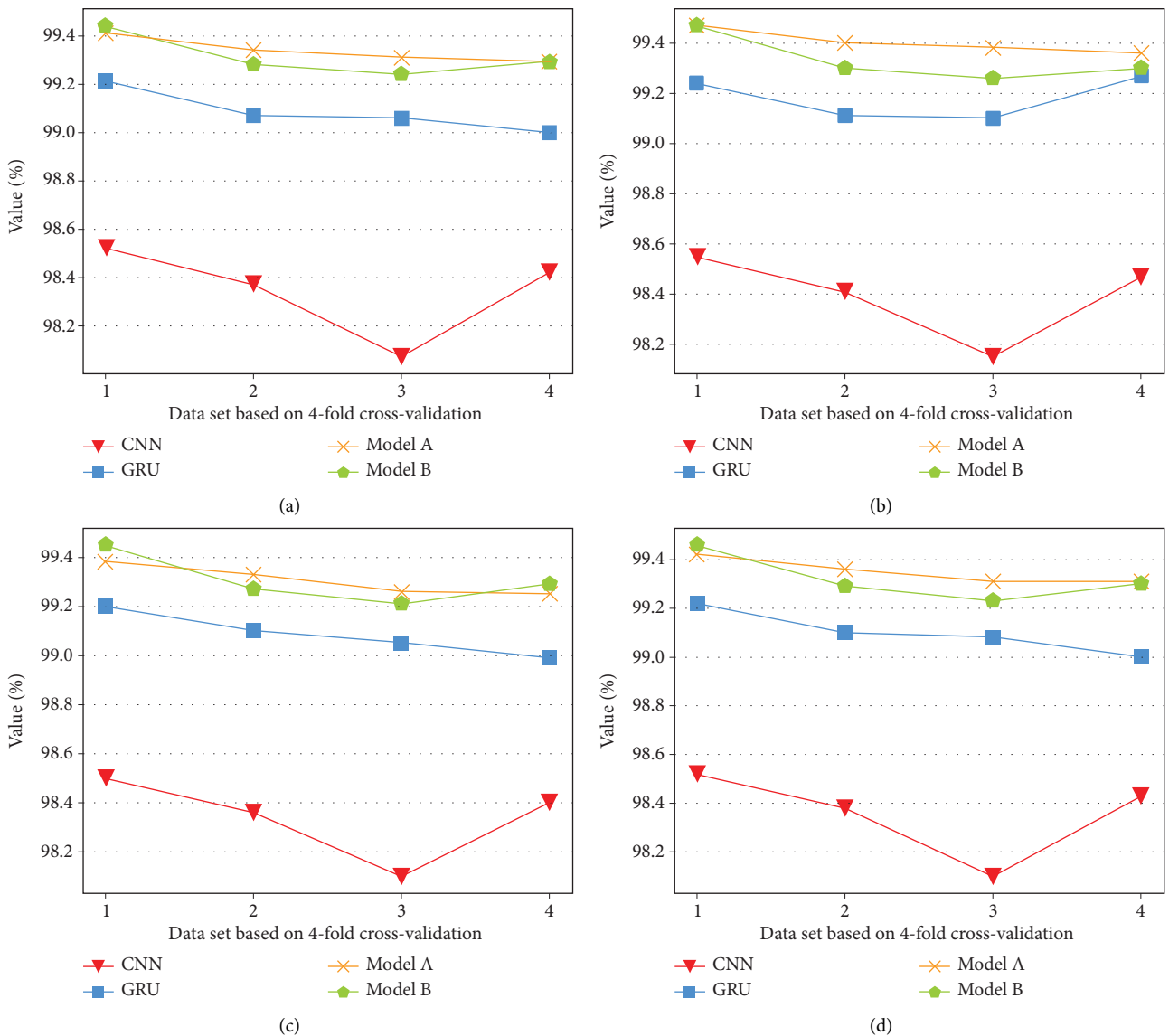


FIGURE 12: Experiment results based on the method 4-fold cross validation. (a) Accuracy of models, (b) Precision of models, (c) recall of models, (d) F1 score of model.

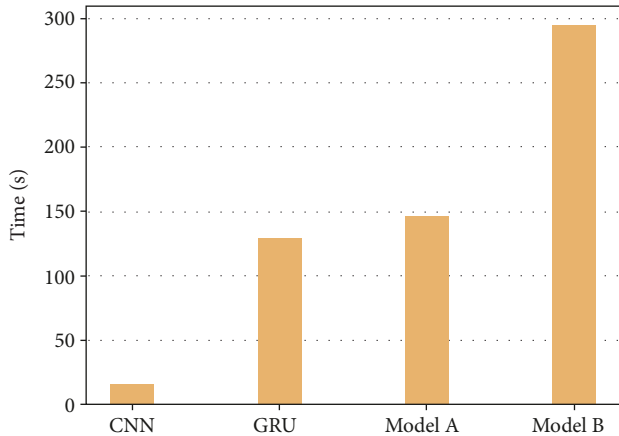


FIGURE 13: Time consumption of different models.

4.4. Discussion. In this part, we try to explore why the multimodal framework has better performance. Basic learners with relatively good performance observe the data from different dimensions and obtain part of the information of the truth, but not all of the information. Only by gathering all kinds of information together can we get a more accurate description of the data. In order to ensure the effectiveness of the multimodal framework, the key lies in the diversity of the learners. The deviation of different learners is from different perspectives; after integrating the different learners, the framework will cancel each other in these deviations, so the result is more stable and accurate.

The abovementioned two models are the CNN and GRU model for extracting spatial and sequential features. The two types of models are different, so the deviations in the classification can be offset by each other, so as to achieve more stable and accurate classification results.

Figure 11 is our attention analysis based on the features extracted by model pretraining. The two types of basic models are integrated into input data after extracting 256 and 128 features of each training object, and then sent to the secondary learner for training. As can be seen from Figure 11, on the one hand, the proportion of sequence features used by secondary classifiers is much larger than that of spatial features, because the classification accuracy of GRU network using sequence features is higher than that of CNN network classification. On the other hand, from the perspective of usage of the overall feature, in the secondary classifier, both spatial and sequence features participate in the final classification task, which indicates that the above two types of features both contribute to the classification task and promote the classification efficiency of the model.

Figure 12 shows the result of models trained on the different data based on k -fold cross-validation. It can be seen from the figure that the Acc scores of the result of the CNN model vary from 98.07 % to 98.52 %, and the scores of the GRU model are (99.01 %-99.21 %). The scores of model pretraining are (99.29 %-99.41 %), and the values of model joint-training are (99.23 %-99.45 %). Although the performance of the basic model is slightly different on different data sets, the multimodal framework has improved the performance of the basic model.

TABLE 9: Comparison of model parameters.

Model	CNN	GRU	Model A	Model B
Params	424,145	150,673	346,769	589,713

In III-C1, we compared the two types of models and found that compared with model pretraining, model joint-training requires higher training conditions and more time. Table 9 shows the specific parameters of each model. Figure 13 shows how much time it needs to build a model on the training set of 113004 flows and validation set of 37668 flows in each epoch based on the same training environment mentioned above. As shown in the figure, model pretraining requires less training time than model joint-training, and the basic models in model pretraining can be processed in parallel. It seems that the model pretraining has more advantages in practice. But the latter one is an end-to-end model, if the size of sample data is relatively insufficient and data enhancement is allowed for training tasks, model joint-training is more powerful, and model pretraining is limited to the dataset used by the basic model. From this perspective, it is more flexible in selection of data set comparing model joint-training with model pretraining. Therefore, in practice, it is necessary to select the proper model according to the specific training task.

5. Conclusions

To solve network security problems, this study proposes a new method of network flow classification based on deep learning: the multimodal framework which is based on pretraining and joint-training, respectively. To the best of our knowledge, this is the first time that such a framework has been proposed in the field of network flow classification. Experimental results show that the framework outperform similar work done in recent years based on the data set ISCXVPN2016. At the same time, the multimodal framework is a deep learning network, which is handy in processing big data. More data is conducive to the improvement of the framework performance, moreover it can realize the automatic extraction of network flow features, saving a lot of manpower and time which has good practical significance. We believe that the multimodal framework is a meaningful attempt in the field of network security, and it is also very useful for the construction of a human physical and mental health system.

In the next step, we can adopt more methods suitable for network flow classification and expand the framework to build a better classification model. At the same time, it is possible to explore new network for network flow classification tasks. For example, in [23], the author proposed that the capsule network can improve the efficiency of classification tasks, which can be considered as a new direction for future research. With the continuous improvement and use of transformer in the field of feature extraction, it also has great inspiration for the feature extraction of network flow.

In the experiment, we extracted the high-level information features of the network flow and studied the impact of feature extraction on the classification task. In order to fit

the experimental results, high-level information may lose some important information, causing the extracted features to not fully reflect the unique information of the network flow. At the same time, due to the usage of CNN, it is difficult to extract the global information of the features, and it may also affect the classification results. New technologies, such as the Conformer structure [49], can be applied to the field of network flow classification; moreover, we can explore other feasible technologies.

Finally, this article mainly discusses the feasibility of the framework using unencrypted datasets to conduct experiments. In practice, as information security has received much attention, network encryption has become the mainstream. The next step is to conduct experiments based on the encrypted network flow for better application in practice.

Data Availability

The datasets used in this study are available on public repositories.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

All authors contributed equally to the article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 61370073), the National High Technology Research and Development Program of China, the Project of Science and Technology Department of Sichuan Province (Grant no. 2021YFG0322), the Project of Science and Technology Department of Chongqing Municipality, the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJZD-K202114401), Chongqing Qinchengxing Technology Co., Ltd., Chengdu Haitian Digital Technology Co., Ltd., Chengdu Chengdian Network Technology Co., Ltd., Chengdu Civil-military Integration Project Management Co., Ltd., and Sichuan Yin Ten Gu Technology Co., Ltd.

References

- [1] P. Velan, M. Čermák, P. Čeleda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," *International Journal of Network Management*, vol. 25, no. 5, pp. 355–374, 2015.
- [2] A. Sivanathan, D. Sherratt, H. H. Gharakheili et al., "Characterizing and classifying iot traffic in smart cities and campuses," in *Proceedings of the IEEE Conference on Computer Communications Workshops*, pp. 559–564, IEEE, Atlanta, GA, USA, May 2017.
- [3] M. Finsterbusch, C. Richter, E. Rocha, J.-A. Muller, and K. Hanssgen, "A survey of payload-based traffic classification approaches," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 1135–1156, 2013.
- [4] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification," *ACM SIGCOMM - Computer Communication Review*, vol. 36, no. 5, pp. 5–16, 2006.
- [5] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, and J. Aguilar, "Towards the deployment of machine learning solutions in network traffic classification: a systematic survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, p. 1988, 2018.
- [6] A. U. Haq and J. P. Li, *A Survey of Deep Learning Techniques Based Parkinson's Disease Recognition Methods Employing Clinical Data*, Expert Systems With Applications, Amsterdam, Netherlands, 2022.
- [7] H. Dahmouni, S. Vaton, and D. Rossé, "A Markovian signature-based approach to ip traffic classification," in *Proceedings of the 3rd Annual ACM Workshop on Mining Network Data*, pp. 29–34, New York, NY, USA, June 2007.
- [8] A. Madhukar and C. Williamson, "A longitudinal study of p2p traffic classification," in *Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation*, pp. 179–188, IEEE, Monterey, CA, USA, September 2006.
- [9] T. Karagiannis, A. Broido, N. Brownlee, K. C. Claffy, and M. Faloutsos, "Is p2p dying or just hiding?[p2p traffic measurement]," in *Proceedings of the IEEE Global Telecommunications Conference*, pp. 1532–1538, IEEE, Dallas, TX, USA, December 2004.
- [10] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," in *Proceedings of the 13th International Conference on World Wide Web*, pp. 512–521, New York, NY, USA, May 2004.
- [11] Y. Dhote, S. Agrawal, and A. J. Deen, "A survey on feature selection techniques for internet traffic classification," in *Proceedings of the International Conference on Computational Intelligence and Communication Networks*, pp. 1375–1380, IEEE, Jabalpur, India, December 2015.
- [12] A. W. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 50–60, New York, NY, USA, June 2005.
- [13] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for qos: a statistical signature-based approach to ip traffic classification," in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, pp. 135–148, New York, NY, USA, October 2004.
- [14] B. Wang, J. Zhang, Z. Zhang, L. Pan, Y. Xiang, and D. Xia, "Noise-resistant statistical traffic classification," *IEEE Transactions on Big Data*, vol. 5, no. 4, pp. 454–466, 2017.
- [15] J. Kampeas, A. Cohen, and O. Gurewitz, "Traffic classification based on zero-length packets," *IEEE Transactions on Network and Service Management*, vol. 15, no. 3, pp. 1049–1062, 2018.
- [16] Z. Wang, "The applications of deep learning on traffic identification," *BlackHat USA*, vol. 24, no. 11, pp. 1–10, 2015.
- [17] Z. Chen, K. He, J. Li, and Y. Geng, "Seq2img: a sequence-to-image based approach towards ip traffic classification using convolutional neural networks," in *Proceedings of the IEEE International Conference on Big Data*, pp. 1271–1276, IEEE, Boston, MA, USA, December 2017.
- [18] C. Q. Qiang, L. J. Ping, A. U. Haq, L. He, and A. Haq, "Net traffic classification based on gru network using sequential features," in *Proceedings of the 18th International Computer Conference on Wavelet Active Media Technology and*

- Information Processing*, pp. 460–465, IEEE, Chengdu, China, December 2021.
- [19] M. Balanici and S. Pachnicke, “Classification and forecasting of real-time server traffic flows employing long short-term memory for hybrid e/o data center networks,” *Journal of Optical Communications and Networking*, vol. 13, no. 5, pp. 85–93, 2021.
- [20] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, “Network traffic classifier with convolutional and recurrent neural networks for internet of things,” *IEEE Access*, vol. 5, p. 18, 2017.
- [21] S. Rezaei, B. Kroencke, and X. Liu, “Large-scale mobile app identification using deep learning,” *IEEE Access*, vol. 8, pp. 348–362, 2019.
- [22] S. Rezaei and X. Liu, “Multitask learning for network traffic classification,” in *Proceedings of the 29th International Conference on Computer Communications and Networks*, pp. 1–9, IEEE, Honolulu, HI, USA, August 2020.
- [23] H. Yao, P. Gao, J. Wang, P. Zhang, C. Jiang, and Z. Han, “Capsule network assisted iot traffic classification mechanism for smart cities,” *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7515–7525, 2019.
- [24] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, “Mobile encrypted traffic classification using deep learning: experimental evaluation, lessons learned, and challenges,” *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 445–458, 2019.
- [25] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, “Distiller: encrypted traffic classification via multimodal multitask deep learning,” *Journal of Network and Computer Applications*, vol. 183, Article ID 102985, 2021.
- [26] C. Liu, L. He, G. Xiong, Z. Cao, and Z. Li, “Fs-net: a flow sequence network for encrypted traffic classification,” in *Proceedings of the IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 1171–1179, IEEE, Paris, France, April 2019.
- [27] A. Nascita, A. Montieri, G. Aceto, D. Ciuonzo, V. Persico, and A. Pescapé, “Xai Meets mobile Traffic Classification: Understanding and Improving Multimodal Deep Learning Architectures,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, 2021.
- [28] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, “Malware traffic classification using convolutional neural network for representation learning,” in *Proceedings of the International Conference on Information Networking*, pp. 712–717, IEEE, Da Nang, Vietnam, January 2017.
- [29] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, “Characterization of encrypted and vpn traffic using time-related,” in *Proceedings of the 2nd International Conference on Information Systems Security and Privacy*, pp. 407–414, Italy, February 2016.
- [30] S. Zander, T. Nguyen, and G. Armitage, “Automated traffic classification and application identification using machine learning,” in *Proceedings of the IEEE Conference on Local Computer Networks 30th Anniversary*, pp. 250–257, IEEE, Sydney, NSW, Australia, November 2005.
- [31] K.-H. Lee, S.-H. Lee, and H.-C. Kim, “Traffic classification using deep learning: being highly accurate is not enough,” in *Proceedings of the SIGCOMM’20 Poster and Demo Sessions*, pp. 1–2, New York, NY, USA, August 2020.
- [32] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, San Diego, CA, USA, January 2016.
- [33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: visual explanations for deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, Venice, Italy, October 2017.
- [34] H. Sun, Y. Xiao, J. Wang et al., “Common knowledge based and one-shot learning enabled multi-task traffic classification,” *IEEE Access*, vol. 7, pp. 39485–39495, 2019.
- [35] A. Rago, G. Piro, G. Boggia, and P. Dini, “Multi-task learning at the mobile edge: an effective way to combine traffic classification and prediction,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, 10 374 pages, 2020.
- [36] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm,” in *Icml*, vol. 96, pp. 148–156, Citeseer, 1996.
- [37] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
- [38] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [39] A. Ul Haq, J. Li, M. H. Memon, J. Khan, and S. Ud Din, “A novel integrated diagnosis method for breast cancer detection,” *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 2, pp. 2383–2398, 2020.
- [40] A. U. Haq, J. P. Li, Z. Ali et al., “Stacking approach for accurate invasive ductal carcinoma classification,” *Computers & Electrical Engineering*, vol. 100, Article ID 107937, 2022.
- [41] A. U. H. Haq, J. P. L. Li, B. L. Y. Agbley et al., “Iimfcbm: intelligent integrated model for feature extraction and classification of brain tumors using mri clinical imaging data in iot-healthcare,” *IEEE Journal of Biomedical and Health Informatics*, 2022.
- [42] A. U. Haq, J. P. Li, A. Saboor et al., “Detection of breast cancer through clinical data using supervised and unsupervised feature selection techniques,” *IEEE Access*, vol. 9, 2021.
- [43] M. Abadi, A. Agarwal, P. Barham et al., S. Ghemawat, Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,” 2016, <https://arxiv.org/abs/1603.04467>.
- [44] M. Lotfollahi, M. J. Siavoshani, R. S. H. Zade, and M. Saberian, “Deep packet: a novel approach for encrypted traffic classification using deep learning,” *Soft Computing*, vol. 24, no. 3, p. 1999, 2020.
- [45] G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning,” *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.
- [46] B. Juba and H. S. Le, “Precision-recall versus accuracy and the role of large data sets,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 4039–4048, 2019.
- [47] D. Müllner, “Modern hierarchical, agglomerative clustering algorithms,” 2011, <https://arxiv.org/abs/1109.2378>.
- [48] B. Yamansavascular, M. A. Guvensan, A. G. Yavuz, and M. E. Karşlıgil, “Application identification via network traffic classification,” in *Proceedings of the International Conference on Computing, Networking and Communications*, pp. 843–848, IEEE, Silicon Valley, CA, USA, January 2017.
- [49] Z. Peng, W. Huang, S. Gu et al., “Conformer: Local Features Coupling Global Representations for Visual Recognition,” 2021, <https://arxiv.org/abs/2105.03889>.

Retraction

Retracted: An Intelligent Classification System for Cancer Detection Based on DNA Methylation Using ML and Semantic Knowledge in Healthcare

Computational Intelligence and Neuroscience

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

Copyright © 2023 Computational Intelligence and Neuroscience. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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

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

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

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

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

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

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

References

- [1] A. Thakare, M. Bhende, M. Tesema, M. Dighriri, R. Bhavani, and A. Mahmoud, "An Intelligent Classification System for Cancer Detection Based on DNA Methylation Using ML and Semantic Knowledge in Healthcare," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 4334852, 9 pages, 2022.

Research Article

An Intelligent Classification System for Cancer Detection Based on DNA Methylation Using ML and Semantic Knowledge in Healthcare

Anuradha Thakare ¹, Manisha Bhende ², Mulugeta Tesema ³,
Mohammed Dighriri ⁴, R. Bhavani ⁵ and Amena Mahmoud ⁶

¹Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India

²Marathwada Mitra Mandal's Institute of Technology, Pune, India

³Department of Chemistry (Analytical), College of Natural and Computational Sciences, Dambi Dollo University, Dambi Dollo, Oromia Region, Ethiopia

⁴Department of Basic Sciences and General Requirements -IT skills, Fakeeh College for Medical Sciences (FCMS), Jeddah, Saudi Arabia

⁵Department of CSE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai 602105, India

⁶Computer Science Department, Faculty of Computers and Information, Kafrelsheikh University, Kafr El Sheikh, Egypt

Correspondence should be addressed to Mulugeta Tesema; mulugeta@dadu.edu.et

Received 24 July 2022; Revised 1 September 2022; Accepted 10 September 2022; Published 10 October 2022

Academic Editor: Farman Ali

Copyright © 2022 Anuradha Thakare et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To consistently assess a patient's internal and external wellness and diagnose chronic conditions like cancer, Alzheimer's disease, and cardiovascular disease, wearable sensing devices are being used. Wearable technologies and networking websites have become incredibly common in the medical sector in recent times. The condition of a patient's health can be influenced by a number of factors, including psychological response, emotional stability, and anxiety levels, which can be evaluated using social network analysis based on graph theory-based techniques and these ideas, known as "social network analysis" (SNA) are used to study relationship phenomena. Therefore, numerous uses for SNA in health research are possible, ranging from social science to exact science. For example, it can be used to research cooperative networks of healthcare providers and hazard-prone behaviors, infectious disease transmission, and the spread of initiatives for health promotion and prevention. Recently, a number of machine learning-based healthcare solutions have been proposed to track chronic illnesses utilizing data from social networks and wearable monitoring devices. In our suggested approach, we are using an intelligent system with the assistance of wearable sensors for the classification of cancer based on DNA methylation, an important epigenetic process in the human genome that controls gene expression and has been connected to a number of health issues. A mixed-sampling imbalanced data ensemble classification technique is created with the help of biomedical sensors to address the problem of class imbalance and high dimensionality in the Cancer Genome Atlas (TCGA) massive data. This technique is based on the Intelligent Synthetic Minority Oversampling (SMOTE) algorithm. The false-negative rate significantly rises as a result of this, to give a larger data set, a new minority class sample will be first obtained. The noise created during the sample expansion process is actually any data that has been acquired, preserved, or altered in a way that prevents the system that initially conceived it from accessing or utilizing it. Noisy data boosts the amount of space needed excessively and can also drastically influence the findings of any data collection investigation and therefore can also affect the sample sets of one or the other class, resulting in the class imbalance which acts as a common problem in ML datasets. The Tomek Link method is then used to eliminate this noise, producing a reasonably balanced data set. Each layer selects two random forest structures using the cascading forest structure of the deep forest (GC-Forest) algorithm to increase the generalization ability of the model and create the final classification model. Experiments using DNA methylation data collected by employing biosensors from six tumor patients reveal that the mixed-sampling unbalanced data ensemble classification technique may increase the sensitivity to the minority class while maintaining the majority class's classification accuracy.

1. Introduction

The manufacturing of therapeutic devices has advanced much in the last 20 years, with attention to the significance of sustaining human health. Biomedical sensors are being utilized more extensively as wearable devices that enable real information monitoring, such as fitness trackers, wristbands, and watches. Possessing smart materials integrated into them that track real-time data (heart rate, blood glucose, plasma levels, etc.) to guide healthcare professionals. As promising tools for online human research, devices have thus been in surge demand. Current method followed cancer monitoring and other diseases that ought to be attached to critical ports for machine learning and deep forest approach neural networks for ailment detection. In recent years, predictive models of cancer classification combined with biological and genetic data have enabled a more accurate assessment of cancer risk [1]. DNA methylation has become one of the most important epigenetic modifications in cancer research, with studies showing abnormal DNA methylation patterns in “tumor” tissues compared to “normal” tissues [2]. Using machine learning (ML), massive and difficult data sets can be incorporated. The patient experience and outcomes might be optimized by employing these data sets. The creation of functional genomic is tightly linked to a specific treatment approach. Genetic code collection, for instance, may rise by double factors every two years. In contrast, the speed of innovation in a virtual machine has been exceeded by the rise in computational power, linked with the quick reduction in the expense of genotyping. Thus it is only happening with the miracles of ML. Therefore, a new line of research in the field of biological information involves applying machine learning theories and techniques to locate oncogene-related DNA methylation regulatory sites, examine the mechanisms behind the development and incidence of cancer, and discover fresh cancer indicators [3].

The Cancer Genome Atlas (TCGA) is currently one of the most comprehensive cancer sequencing databases, and the rich cancer sample data provides a prospect for developing cancer classification models [4]. The TCGA is a research that employs genetic sequencing and bioinformatics to assemble a list of genetic alterations that cause cancer and thereby plays a significant role in DNA sequencing. The key aim was to implement increased DNA sequencing approaches to improve the diagnosis of cancer, management, and control through a profound understanding of the genomics of the ailment. Like most data, the data in TCGA is inherently imbalanced, which means one or more classes have significantly lower proportions in the training data than the other classes. There is an imbalance resulting in the wrong classification in the detection and identification of cancer sequencing, and this issue can also be termed as the high dimensional and class imbalance data. The classification of these highly imbalanced data suffers from the majority class, resulting in increased false negative rates [5]. A mixed-sampling imbalanced data ensemble classification technique based on the

Intelligent Synthetic Minority Oversampling (SMOTE) algorithm is developed with the help of biomedical sensors to address the problem of class imbalance and high dimensionality in The Cancer Genome Atlas (TCGA) massive data.

Hence, to solve the problems of class imbalance and high dimensionality issues in the data set of cancer classification model, the main contribution of our study is to propose an integrated intelligent classification model embedded with biomedical sensors and mixed sampling. The minority sample set is expanded using the intelligent SMOTE method, and the boundary and noise data are removed using the Tomek Link algorithm, resulting in generally balanced training data. On the basis of ensuring the classification accuracy of the majority class, it also imports the training data into the Gcforest model and successfully improves the classification accuracy of cancer minority class samples.

The ML and DL techniques employed in the analysis of cancer development are explored in this work. The bulk of predictions mentioned is associated with particular ML inputs and targeted sample management [6]. To improve academic approaches and prepare the way for information and analyse of medical research, we focused on analyzing and evaluating countless research AI and machine learning approaches, strategies, and perspectives in this study [7]. To categorize the various cancer kinds according to the tissue from which they emerged, we employed SVM, Naive-Bayes, Extreme-gradient-boosting, and RF machine learning models. RF outperformed the other predictors, achieving 99% reliability. In fact, we employed local interpretable model-agnostic explanations to assess relevant methylated patterns to identify specific disease classifications [8]. The vision of medical guidance will move toward speedier modeling of a new medication for each patient via medical application of machine learning and artificial intelligence in cancer diagnosis and therapy. Experts may work together in real-time and disseminate expertise digitally using the AI-based systematic approach, which has the power to heal millions of citizens. By fusing genetics and intelligent systems, the study presented game-changing medical innovations in this study and highlighted how oncologists might gain from intelligence support for focused cancer care [9].

1.1. Organization. The study is organized into several modules where the first module provides the introduction to the problem statement followed by the 2nd section which states about the various methods involved in the study. Section 3rd discusses about the analysis and discussions regarding the experiments conducted and investigations performed, followed by the ultimate section which provides the conclusion of the study.

2. Methods

The methods here, are separated into three stages: data preparation, feature selection, and model training and validation. In the preprocessing step, the intelligent SMOTE

algorithm is employed to maintain a balanced class distribution, and the Tomek Link under-sampling approach is utilized to remove noise from the data which is the main parameter that is considered in the gene sequencing because noisy data boosts the amount of space needed excessively and can also drastically influence the findings of any data collection investigation, therefore, affects the sample sets of one or the other class resulting in the class imbalance which acts as a common problem in ML datasets. Thus, only genes with cancer-causing mutations were examined to limit the data's feature space. COSMIC and CIVic Internet database resources were used to collect data. Create a classification model using the Gcforest technique, the model was tested on six distinct forms of cancer obtained from biomedical sensors embedded on the patient's body [6, 10]. The data on DNA methylation came from <https://portal.gdc.cancer.gov/repository>. Figure 1 displays the technical flow chart of the research in this paper.

2.1. Data Preprocessing

2.1.1. Data Processing. DNA methylation data for 28 cancer types was released by the TCGA study. Raw data (0×1) may be downloaded from the TCGA website and mapped to particular data spots or ranges (eg, chr19:19033575 indicates location 19033575 on chromosome 19). The Broad Institute's FireBrowse, which maps numerical values to particular human genes labelled using HGNC nomenclature, is used to preprocess DNA methylation data in this research [7, 8]. Each sample file has a TCGA identification number that specifies whether it is a tumor tissue or a normal tissue (e.g., TCGA-2F-A9KW-01: tumor type: 0109 (category 1), normal type: 10 19). (Category 0) [9]. Table 1 shows the statistics of six tumor types from the TCGA database that has quite extensive sample data.

2.1.2. Sampling. The data from TCGA is substantially skewed, as seen in Table 1, due to the nonuniform distribution of the target classes. For cancer samples, current classification algorithms offer good accuracy, but limited sensitivity for normal samples [11]. As a result, this research provides a mixed sampling approach that is used when a sample strategy calls for the use of two or more fundamental sampling techniques. These approaches are employed for evaluating and modifying processes that influence the execution of evidence-based solutions. These techniques further optimizing the normal sample sensitivity while maintaining excellent accuracy.

(1) *Technique of Intelligent Synthetic Minority Sampling (ISMOTE).* The author has presented Intelligent SMOTE (Intelligent Synthetic Minority Oversampling Technique), an enhanced approach based on the random oversampling algorithm [12]. To balance the dataset, fresh samples are inserted into a limited number of comparable samples. Rather than using a random oversampling approach that just copies the sample, the SMOTE algorithm creates a fresh sample from scratch, bypassing some categorization

filtering. The SMOTE algorithm works on the following principle:

- (1) Calculate the distance between each sample x in the minority class and all samples in the minority class sample set using the Euclidean distance as the standard, and determine its k closest neighbors.
- (2) Determine the sampling ratio N based on the sample imbalance ratio, then randomly choose multiple samples from the k -nearest neighbors for each minority class sample x .
- (3) Create a new sample from the old sample using the procedure for each randomly picked neighbor (1).

$$p_i = x + \text{rand}(0, 1) \times (y_i - x), i = 1, 2, \dots, N, \quad (1)$$

where x is the sample, $\text{rand}(0,1)$ represents a random number in the interval $(0,1)$, and y_i is the k -nearest neighbors.

(2) *Tomek Link.* The concern is that while the Intelligent SMOTE approach extends the sample space of the minority class while balancing the class distribution, the space initially belonging to the majority class sample may be "invaded" by the minority class, resulting in model overfitting. To overcome this issue, the Tomek Link method [13] is used to remove noise points or boundary points, which effectively solves the "intrusion" problem. The Tomek Link algorithm is based on the following principle: assume that the sample points x_i and x_j belong to separate categories, and that the distance between them is represented by $d(x_i, x_j)$. If there is no third sample point x_l such that $d(x_l, x_i) < d(x_i, x_j)$ or $d(x_l, x_j) < d(x_i, x_j)$ holds, call (x_i, x_j) a Tomek Link pair. If two sample points are Tomek Link pairs, one of the samples is either noise (too much deviation from the normal distribution) or both samples are on the border between the two classes. It means that these assumptions are necessary to make separate categories of the data to analyse the noise and the normal data set. These assumptions are mandatory for the removal of ambiguity. Furthermore, by inserting the Euclidean distance between the sample point and the original sample point and its neighbors, the research in this article ensures that the inserted data has a fair resemblance with the original sample. The Tomek Link technique is employed after the SMOTE algorithm has extended the minority samples. The Euclidean distance is calculated and sample points with low similarity, referred to as noise points or boundary points in the text, are discarded.

2.1.3. Blood Pressure Measurement Using Biomedical Sensors. Blood pressure is one of the four vital signs of the human body, which can reflect the systolic function of the heart. The pulse transit time (PTT) is the core principle of noncontact blood pressure measurement. It was initially estimated by ECG and PPG jointly, and then the author measured it by two rPPG signals, which opened the rPPG noncontact

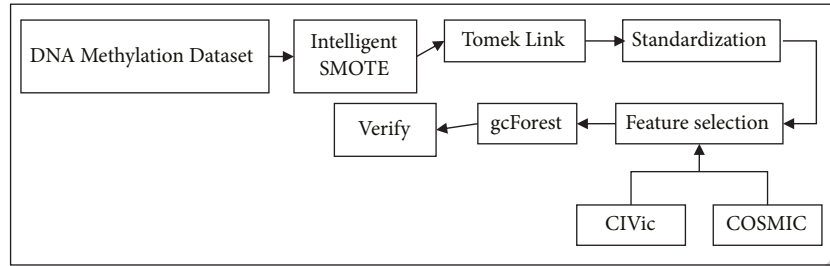


FIGURE 1: The technical flow chart of the research in this paper.

TABLE 1: DNA methylation data used in this paper.

Tumor type	Abbreviation	#patients	Tumor-1	Normal-0
Breast invasive carcinoma	BRCA	885	790	95
Lung adenocarcinoma	LUAD	490	463	34
Urothelial bladder carcinoma	BLCA	435	410	22
Prostate adenocarcinoma	PRAD	546	497	53
Lung squamous cell carcinoma	LUSC	416	370	44
Thyroid cancer	THCA	562	504	59

measurement of blood pressure prelude. From the literature, the calculation formula of BP estimated by PTT can be known as follows:

$$BP = b + c * PTT. \quad (2)$$

B , c are related to the elasticity of human blood vessel walls. Based on this concept, the author proved for the first time that the value of multipoint PTT of the body can be calculated in a noncontact way and developed a noncontact multiparameter measurement system based on this. The author has designed a framework for adaptively selecting rPPG modules based on the Gaussian model and proved the high correlation between PTT and BP by analyzing the characteristics between rPPG signals [13]. The quality of the signal pulse plays a vital role in estimating PTT based on rPPG. Authors improved the Kalman filter to improve the signal-to-noise ratio of the rPPG signal and show more apparent peaks to improve the estimation accuracy of PPT. In addition, the blood pressure monitoring method based on multipoint pulse wave phase difference has also been proved to have good measurement accuracy in addition to the PTT estimated by the single-point signal peak due to the influence of the body's voluntary movement. The author collects signals from the radial artery of the left hand and the end of the finger to calculate the PPT. The experiment proves that the correlation between the calculated PTT and blood pressure reaches 0.79, which is higher than that of the single-point pulse wave phase difference calculation method that only uses a single signal to calculate the PTT. However, the author also pointed out that the multipoint measurement method has higher requirements on the camera's frame rate.

2.1.4. Heart Rate Variability Measurement Using Biomedical Sensor. HRV, a parameter closely related to heart disease, is an essential indicator of whether the heart rate is abnormal. ECG has always been the standard equipment for HRV

detection, and the characteristics of QRS complexes analyse the difference between heartbeat cycles in terms of clinical use. Studies have shown that the pulse wave and HRV signal have an equivalent relationship. Still, the time-domain parts of rPPG movement are easily affected by noise, and pulse wave signal characteristics (64) have become an effective method. Each skin patch provides a pulse signal, which is selected from the time domain and frequency domain features of multiple passwords and combined with practical information to improve the discriminability of rPPG for abnormal heart rate detection under noise and unnatural interference. In addition, atria fibrillation can lead to abnormal PPG signals. Therefore, Pereira et al. proposed a dual-window support vector machine classification model based on this feature. After testing, the model showed good performance on a dataset consisting of many patients. Generalization performance and test performance; also using the dual-window detection strategy, authors used the periodic variance maximization algorithm to extract the rPPG signal. Periodic Variance Maximization also is a newly developed technique used to extract the cardiac signal embedded within the RGB temporal patterns in remote-photo-plethysmography-signal (rPPG). By integrating the two strategies, the PVM algorithm seeks to determine the required signal's unknown period. Two procedures are used: first, an incremental subdomain dissection process that creates a periodicity-maximizing basis for a particular frequency, then secondly, a global optimization tabu search algorithm is employed to identify the frequency with the highest global periodicity across the search space. For any type of biosensor measuring scenarios without vibration, the suggested technique is utilized to retrieve any desired signal of deviations from a blend of data and can adaptively detect the peak through the dual-window, which successfully improved the detection effect of rPPG on HRV. In the frequency domain, the power information of high frequency and low frequency is another indicator of whether the heart

rate information is abnormal. Still, it is also easily affected by noise. The author separated the noise and signal into independent components based on ICA, showing better experimental results than EVM. HRV analysis based on rPPG is still in the laboratory stage, and the clinical use, and diagnosis of other arrhythmia-related physiological diseases based on HRV will be the focus of future research.

2.2. Data Preprocessing. The TCGA DNA methylation data in diverse cancer types include about 20 000 protein-coding genes as distinctive characteristics. Feature selection is critical in this instance [14]. As a result, only those genes that have been scientifically recognized as having cancer mutational importance are targeted by the research. The Cancer Gene Census (COSMIC) and Clinical Interpretation of Variants in Cancer (CIVC) were used to find these genes (CIVic). The COSMIC Cancer Gene Census (CGC) is a benchmark in cancer genetics used in fundamental research, medical reporting, and pharmaceutical development. It is an elite description of the genomes creating human cancer. While as (CIVic) describes the therapeutic, predictive, analytical, and inducing relevance of hereditary and physiological variations of all types. CIVic is an elite aspect of learning for Clinical-Interpretation-Variants in cancer. To facilitate the transparency and open generation of current and reliable variant analyses for use in cancer targeted therapies, CIVic is dedicated to accessible code, increased samples, accessible app programming interfaces (APIs), and traceability of substantiating evidence.

2.3. Intelligent Classification Model. Authors devised the Gcforest technique, a decision tree-based ensemble algorithm [15]. The two essential elements that make up the core of Gcforest are Cascade Forest and MultiGrained Scanning. The makeup of the Cascade Forest is as follows: The decision trees that make up each forest in the cascade forest are composed of a number of random and utterly random forests. Random forests at each layer and overall ensure the model's heterogeneity. Figure 2 depicts the particular cascade forest structure.

Two full random forests (black) and two random forests (red) make up each layer of the cascade forest in Figure 3 (blue). Each random forest also contains 30 entirely random decision trees, each of which randomly chooses a feature for splitting until the examples contained in each leaf node belong to the same class. The best base value for splitting is picked for each decision tree by selecting \sqrt{d} features (the sum of the features of d inputs) at random. When the effect cannot be further enhanced, the cascade forest iteration comes to an end.

Each forest contains many decision trees, each of which will determine a class vector result (for example, three classes, as shown below), then combine all decision tree results, and then take the mean to generate the forest's results. The final decision result is a three-dimensional class vector, and Figure 4 depicts the decision process for each forest. Each forest will choose a three-dimensional class vector in this manner. Returning to Figure 3, each of the four

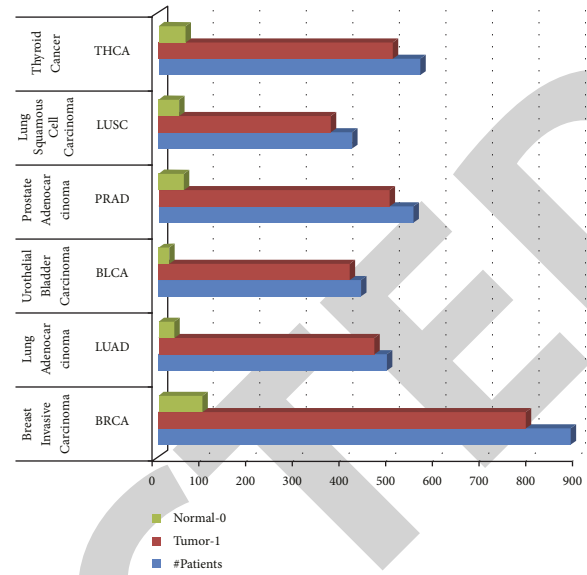


FIGURE 2: DNA methylation data.

forests in the cascade forest can choose a three-dimensional class vector, then average the four class vectors, and finally take the highest value. The final classification result is the category that corresponds to the value.

2.4. Evaluation Indicators. Recall/Sensitivity:- The larger the value of Sen/Rec, the larger the disease is judged to be diseased, and the smaller the missed detection (FN).

$$\text{Rec} = \text{Sen} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}. \quad (3)$$

Precision:- Precision, that is, the proportion of all positive predictions that are correctly predicted.

$$\text{Prec} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}. \quad (4)$$

F_1 is the ratio of the arithmetic mean to the geometric mean, the bigger the better.

$$F_1 = 2 \times \frac{\text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}. \quad (5)$$

The response sensitivity and specificity ROC curve is a comprehensive measure of continuous variables. It allows for a natural comparison of various trials on the same scale. The bigger the diagnostic value, the more convex and closer the ROC curve is to the top left corner, which is useful for comparing various indicators the area under the curve may be used to assess the diagnostic accuracy.

3. Analysis and Discussion

Training set: test set ratio of the DNA methylation data using biomedical sensor received from TCGA is 7:3. Figure 5 illustrates the PCA 2D plot of the training data, which demonstrates that the sample data distribution is extremely imbalanced.

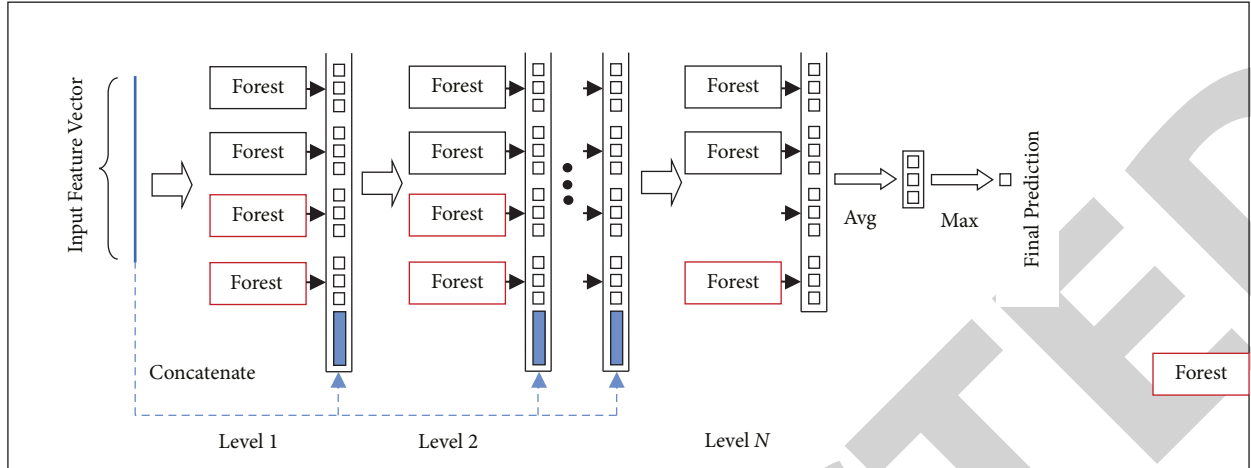


FIGURE 3: Cascading forest structure diagram.

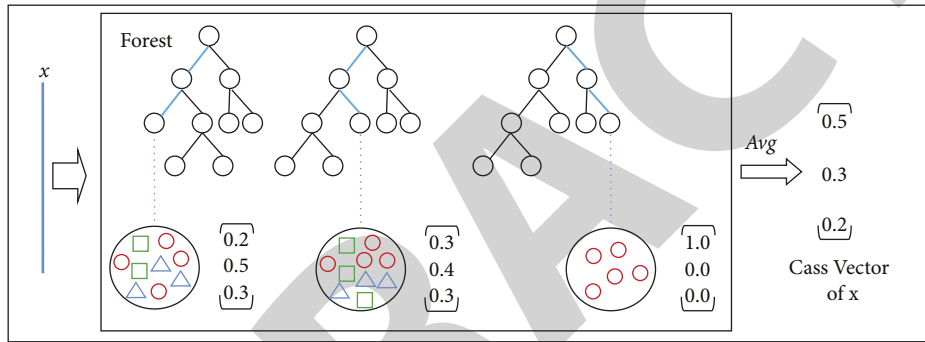


FIGURE 4: Decision making process for each forest.

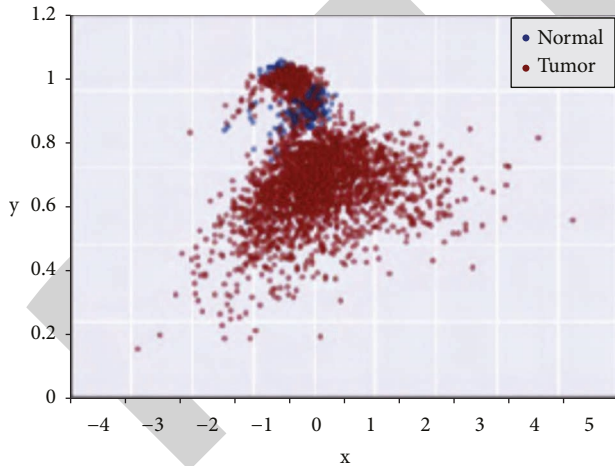


FIGURE 5: Distribution before sampling.

Table 2 shows the model performance comparison of four classification methods, CcForest, Logistic Regression (LR), Random Forest (RF), and Deep Belief Network (DBN), and the same indicators has been represented in Figure 6 [16, 17]. It can be seen from Table 2 that the four classification algorithms have high accuracy for the majority class samples, but poor sensitivity on the minority class, which is caused by the imbalance within the data [18].

TABLE 2: Performance indicators of the four models before mixed sampling.

Method	Sen/Rec		Pre		F_1	
	0	1	0	1	0	1
LR	0.705	0.974	0.790	0.976	0.750	0.977
RF	0.795	0.989	0.845	0.990	0.824	0.989
DBN	0.757	0.980	0.825	0.984	0.795	0.977
GcForest	0.834	0.989	0.846	0.994	0.843	0.991

To solve the above problems, the SMOTE algorithm proposed in this paper is combined with the mixed sampling model of the TomekLink algorithm to preprocess the DNA methylation data. The PCA two-dimensional map of the processed DNA methylation data is shown in Figure 7, and the data distribution is relatively balanced [19].

After the data obtained from bio medical sensors are standardized, the four classification models are compared again. As shown in Table 3, after using the mixed sampling model proposed in this paper, the evaluation indicators Sen/Rec, Pre, and F_1 of the four classification models for the minority class have been greatly improved.

Comparing Table 2 and Table 3, it can also be found that among the four classification models and Figure 8 demonstrates the performance indicators of the four models after mixed sampling, whether before or after sampling, the

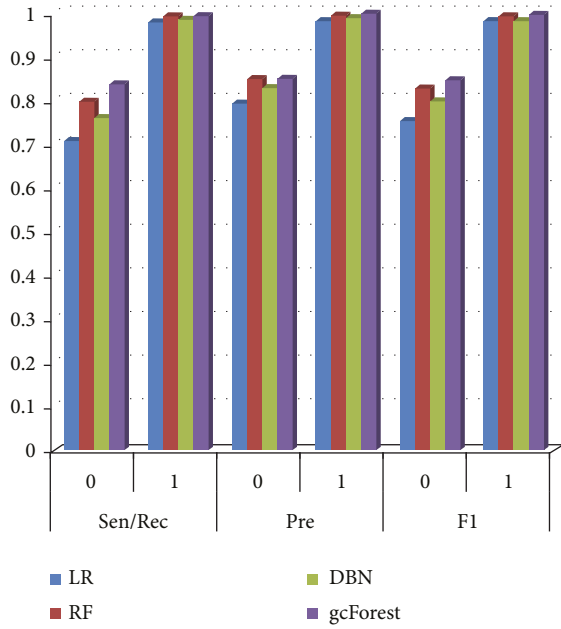


FIGURE 6: Performance indicators of the four models.

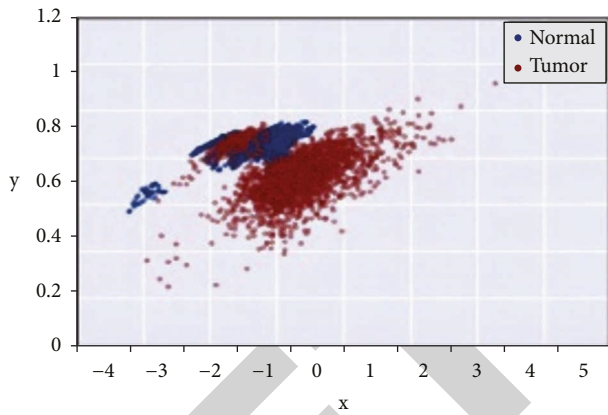


FIGURE 7: Distribution sampling.

TABLE 3: Performance indicators of the four models after mixed sampling.

Method	Sen/Rec		Pre		F_1	
	0	1	0	1	0	1
LR	0.863	0.980	0.871	0.983	0.868	0.980
RF	0.915	0.989	0.915	0.993	0.919	0.990
DBN	0.895	0.985	0.901	0.987	0.903	0.981
gcForest	0.939	0.987	0.940	0.994	0.936	0.993

Gcforest algorithm has the best classification effect. To clearly and intuitively compare the performance of the four classification models as shown in Figures 9 and 10, shown are the ROC curves of the four classification models, and the comparison shows that the deep forest Gcforest algorithm has the best performance [20]. This is due to the high dimensionality of the DNA methylation sequencing data using biomedical sensors in this study, and the multi-granularity scanning structure in the Gcforest algorithm uses a sliding

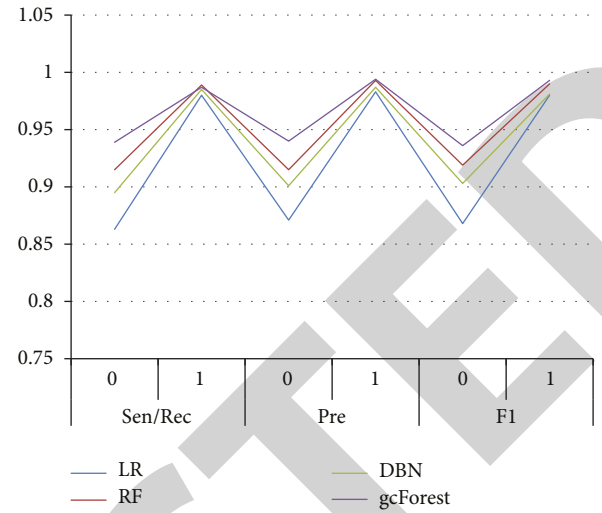


FIGURE 8: Performance indicators of the four models after mixed sampling.

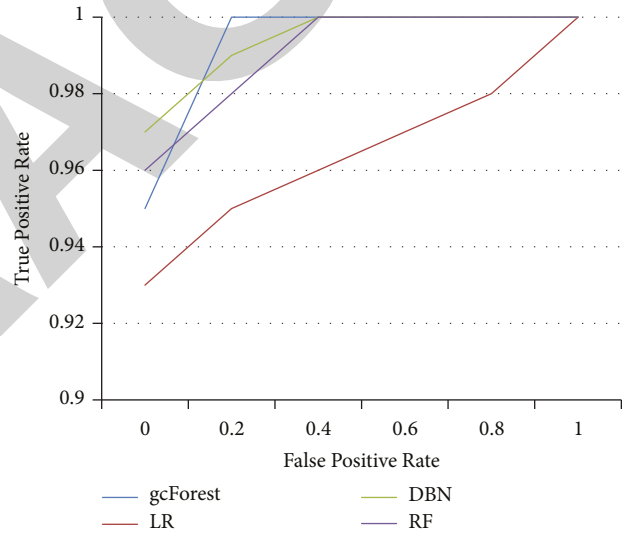


FIGURE 9: ROC curves of four classification models.

window to preprocess the input data features, and its representation learning ability is further improved. Secondly, the obtained features are input into the cascaded forest of the Gcforest algorithm for training. The cascaded forest combines the input features with the original features. Through the learning of random forests and complete random forests in two-level cascaded forests, compared with logistic regression, Random Forest, Deep Belief Network, and the correlation between features can be learned more fully, so the best performance is obtained. In addition, compared with the deep belief network, the Gcforest algorithm has fewer model parameters and is easy to train, which is more advantageous in small datasets in cancer classification research [21].

In addition, in this study, a comparative analysis of the influence of different neighbor's k and sampling ratio N on the comprehensive evaluation index F_1 in the Gcforest classification model is also carried out, the best performance

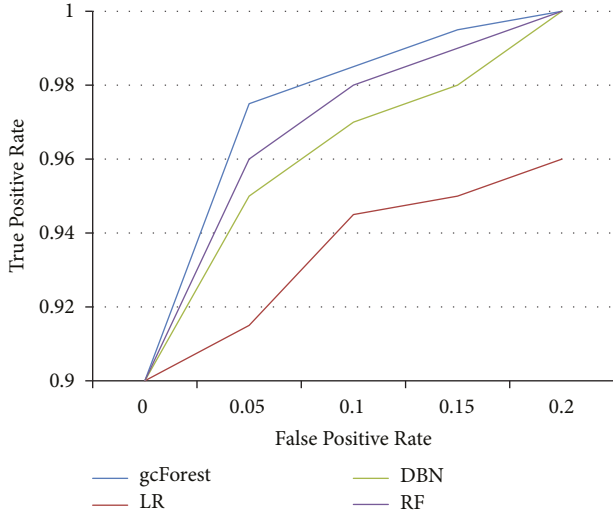
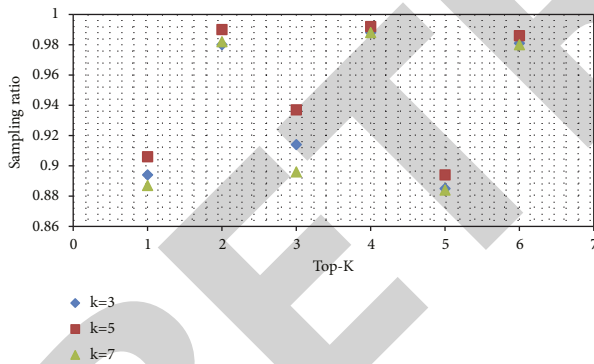


FIGURE 10: ROC curve graph top left detailed view.

TABLE 4: Influence of different neighbor k and sampling ratio N on F_1 .

Experimental program	$N = 100$		$N = 200$		$N = 300$	
	0	1	0	1	0	1
$k = 3$	0.894	0.980	0.914	0.988	0.885	0.981
$k = 5$	0.906	0.990	0.937	0.992	0.894	0.986
$k = 7$	0.887	0.982	0.896	0.988	0.884	0.980

FIGURE 11: Influence of different neighbor k and sampling ratio N on F_1 .

is when $k = 5$. Table 4 shows the influence of different neighbor k and sampling ratio N on F_1 .

There are two main reasons for the analysis:

- (1) When the sampling ratio is $N = 100$, the balanced positive and negative sample data still have a large imbalance, which makes the experimental results insignificant.

When the sampling ratio is $N = 300$, the number of samples expanded after balancing is much larger than the original samples. Since various over-sampling operations such as the SMOTE algorithm are essentially “out of nothing,” the performance of the model after balancing is not obvious as

demonstrated in Figure 11. The statement implies a comparative analysis of the influence of different neighbor’s k and sampling ratio N on the comprehensive evaluation index F_1 in the Gcforest classification model. However, in the main study, to address the issue of class imbalance and high dimensionality in The Cancer Genome Atlas (TCGA) massive data, a mixed-sampling imbalanced data ensemble classification technique based on the Intelligent Synthetic Minority Oversampling (SMOTE) algorithm with the aid of biomedical sensors is developed and is essentially a significant model. This leads to a significant increase in the false-negative rate and is used to expand the minority sample set, which effectively improves the classification accuracy of cancer minority class samples under the assumption that the majority class classification accuracy will be maintained.

- (2) Regarding the selection of the nearest neighbor k , when $k = 3$, the model complexity is high, overfitting is easy to occur, and the learning estimation error increases; when $k = 7$, although the learning error is reduced, due to the sample the data set is small, and when k is 7, the data far from the sample will also affect the classification result of the model, increasing the approximation error of the model learning.

4. Conclusion

It can be difficult to extract relevant information from the vast amount of healthcare data that wearable computing devices collect and to accurately analyse that data to make an effective diagnosis. To successfully analyse the data that has been taken from biomedical data and analyse it to uncover unrecognized chronic disease signs and forecast a patient’s care, artificial intelligence systems and semantic knowledge are required. Additionally, for intelligent healthcare, multitasking deep learning models like Deep Forest that can analyse sensor data are required. This research proposes an integrated intelligent classification model for cancer diagnosis that is embedded with biomedical sensors and uses mixed sampling to overcome the aforementioned problems with the unbalanced data set. The minority sample set is expanded using the intelligent SMOTE technique, and the boundary and noise data are removed using the Tomek Link algorithm. The training data is utilized to significantly increase the classification accuracy of cancer minority class samples after being imported into the Gcforest model, assuming that the classification accuracy for the majority class will be preserved. The experimental findings show that the imbalanced data ensemble classification model embedded with biomedical sensors based on mixed sampling proposed in this paper can significantly increase the classification accuracy of the majority class. This is based on the comparison of models such as Logistic Regression, Random Forest, and Deep Belief Network DBN sensitivity to class. Additionally, when applied to small, unbalanced datasets, the Gcforest classification model using the intelligent SWORT algorithm outperforms the deep belief network

Research Article

Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods

Zahid Ullah ¹, Farrukh Saleem ¹, Mona Jamjoom ², Bahjat Fakhieh ¹, Faris Kateb ³,
Abdullah Marish Ali ⁴ and Babar Shah ⁵

¹Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

²Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia

³Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

⁴Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

⁵College of Technological Innovation, Zayed University, Abu Dhabi, UAE

Correspondence should be addressed to Mona Jamjoom; mmjamjoom@pnu.edu.sa

Received 27 May 2022; Revised 13 September 2022; Accepted 19 September 2022; Published 29 September 2022

Academic Editor: Abdul Rehman Javed

Copyright © 2022 Zahid Ullah et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Diabetes is a chronic disease that can cause several forms of chronic damage to the human body, including heart problems, kidney failure, depression, eye damage, and nerve damage. There are several risk factors involved in causing this disease, with some of the most common being obesity, age, insulin resistance, and hypertension. Therefore, early detection of these risk factors is vital in helping patients reverse diabetes from the early stage to live healthy lives. Machine learning (ML) is a useful tool that can easily detect diabetes from several risk factors and, based on the findings, provide a decision-based model that can help in diagnosing the disease. This study aims to detect the risk factors of diabetes using ML methods and to provide a decision support system for medical practitioners that can help them in diagnosing diabetes. Moreover, besides various other preprocessing steps, this study has used the synthetic minority over-sampling technique integrated with the edited nearest neighbor (SMOTE-ENN) method for balancing the BRFSS dataset. The SMOTE-ENN is a more powerful method than the individual SMOTE method. Several ML methods were applied to the processed BRFSS dataset and built prediction models for detecting the risk factors that can help in diagnosing diabetes patients in the early stage. The prediction models were evaluated using various measures that show the high performance of the models. The experimental results show the reliability of the proposed models, demonstrating that k-nearest neighbor (KNN) outperformed other methods with an accuracy of 98.38%, sensitivity, specificity, and ROC/AUC score of 98%. Moreover, compared with the existing state-of-the-art methods, the results confirm the efficacy of the proposed models in terms of accuracy and other evaluation measures. The use of SMOTE-ENN is more beneficial for balancing the dataset to build more accurate prediction models. This was the main reason it was possible to achieve models more accurate than the existing ones.

1. Introduction

Diabetes mellitus is a metabolic disease caused by the presence of an excessive amount of glucose in the blood due to the inadequate secretion of insulin or insulin resistance [1]. The pancreas is the main source for producing insulin, a crucial hormone that is responsible for transferring the

converted glucose through the bloodstream to different body parts [2]. Furthermore, the inappropriate secretion of insulin causes the glucose to persist in the blood, which ultimately causes a surge in the sugar level in the blood [2]. This disease causes a huge economic burden and has attracted deep public concern globally [3]. According to [4], diabetes has hugely burdened the US economy, with a total

estimated cost of 327 billion in 2017, including the direct medical cost of 237 billion and 90 billion in reduced productivity. It is evident from several estimations and forecasts that diabetes is related to augmented mortality and has increasing prevalence [5]. As per the report of [6] discussed in [3], the worldwide prevalence of diabetes was around 9.3% in 2019 among adults, accounting for a total of around 463 million adults with diabetes; the report further predicted that this number may increase to 700 million in 2045. According to a report [7], around 422 million people have diabetes globally, of whom the majority live in low and middle income countries, and around 1.5 million mortality cases are due to diabetes every year.

Diabetes has three different types: type 1, type 2, and gestational [2, 4]. In most cases, patients recover from gestational diabetes after delivery, while prediabetes can be controlled through proper diet and exercise [2]. Type 1 diabetes is mostly detected in people under 30 years of age [8]. However, type 2 diabetes develops at a later age [4] due to obesity and insulin resistance of cells [2], high blood pressure, dyslipidemia, arteriosclerosis, and other related diseases [8]. In addition to these risk factors, recent experiments show that some environmental endocrine disturbances might cause the occurrence of diabetes [3]. Among the types of diabetes, type 2 is predictable and preventable because it occurs at a later age due to lifestyle and other risk factors [4].

Diabetes is a common disease that affects people worldwide and increases the risk of life-threatening long-term complications such as heart disease and kidney disease, among others [9]. However, if diabetes is detected at an early stage, patients can live longer and healthier. Approaches of artificial intelligence (AI) and machine learning (ML) have changed and affected every sector. Generally, the medical sector is one of the vital sectors where healthcare makes great use of such technology in terms of detecting and diagnosing some critical diseases [10, 11]. One of them is the use of ML to identify the risk factors of diabetes at the early stage and diagnose the disease before complications occur. While ML methods have increased the accuracy of medical diagnosis while reducing medical costs [12] of diagnosing and without surgical intervention. In the literature, several attempts have been made to detect and diagnose diabetes.

This study aims to develop prediction models for detecting the risk factors that cause diabetes and to provide decision-based models for diagnosing this disease at an early stage. For this purpose, several ML techniques are used to provide an accurate model that can help medical practitioners in diagnosing this disease. The experimental results show the higher performance of the proposed models in terms of accuracy and other evaluation measures. The better performance of the proposed models provides support for using these models as a decision support system to detect the risk factors of diabetes and help medical doctors in diagnosing diabetes mellitus at an early stage.

The rest of this study is organized as related work has been described in the next section, followed by a detailed methodology. Section 4 describes the experimental setup; Section 5 describes the results and discussion. Section 6 concludes this study.

2. Related Work

In this section, domain-specific studies are analyzed to understand the trends and techniques used in the existing studies for detecting the high-risk factors of diabetes using ML methods. For this purpose, several databases were explored with various keywords for searching related studies. The databases searched included Google Scholar, Science Direct, IEEE Xplore, MDPI, and several others. In the existing studies, most of the researchers have used the Pima India diabetes dataset (PIDD) for detecting, diagnosing, early diagnosing, building smart applications, and other functions for diabetes patients. For example, in [8], two datasets (i.e., a private dataset and the PIDD) were used. The authors used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) for the dimensionality reduction. Several ML algorithms were used for detecting diabetes. The results reported that RF outperformed other methods with an accuracy of 80.84% for the private dataset, while the PIDD yields an accuracy of 77.21%. Similarly, [13] attempted to detect diabetes patients using ML methods. They used the PIDD and used the PCA methods for dimensionality reduction. A bootstrapping method was used to compare the performance of the trained models. The reported results show better performance of SVM and AB classifiers after the bootstrap operation that both achieved an accuracy of 94.44%.

Reference [4] attempted to build risk prediction models for type 2 diabetes. They used the BRFSS-2014 dataset and trained several ML models. In the dataset, the class imbalance issue was handled using the SMOTE method in order to avoid bias. The experimental results showed that the overall performance of the neural network (NN) showed a higher accuracy rate of 82.41% than all other measures.

In [14], the authors proposed a comparative study of ML methods for the efficient diagnosis of five major diseases, including diabetes. The authors used the BRFSS dataset and trained logistic regression and RF models based on it. The theme of the study is to predict the percentage of chronic diseases based on the inputs via a chatbot in which suggestions are provided using modeled and interactive data visualization to lower the risk. They have attempted several experiments with different parameters and concluded that RF with 100 trees and a maximum depth of 10 achieved better results than LR, detecting diabetes with an accuracy of 86.80%.

In [15], the authors used 24 different classification algorithms for detecting diabetes in the early stage. The experiment was performed using MATLAB. The model performance was evaluated using cross-validation. The authors reported that the LR was the best fitted model of all 24 ML methods used in the study, as LR reached an accuracy rate of 77.9%.

A study conducted by [16] used the PIDD and trained 7 different ML models. In this approach, a feature selection was used in which two of the features were dropped. The highest accuracy of LR and SVM reached around 77%-78% in both split and k-fold validations. The same dataset was also used for training the NN model with different hidden

TABLE 1: Summary of related work.

S. No.	Ref.	Dataset	Preprocessing method(s)	Outperformed method(s)	Model accuracy (%)
1	[8]	Private PIDD	PCA, mRMR	RF	80.84 77.21
2	[13]	PIDD	PCA	SVM, AB, bootstrap	94.44
3	[4]	BRFSS-2014	SMOTE	NN	82.41
4	[14]	BRFSS	Different parameters used	RF	86.80
5	[15]	—	—	LR	77.9
6	[16]	PIDD	Feature selection	NN	86.6
7	[9]	PIDD Other	Label encoding, normalization	SVM DT, RF	80.26 96.81
8	[17]	PIDD	Features extraction	RF	88.31
9	[2]	Private	—	LR	96.02
10	[18]	Private	—	Bagging	97.7

layers, learning rates, and iterations. The authors concluded that NN with 2 hidden layers outperformed other methods with an accuracy rate of 86.6%.

An attempt was made by [9] to detect diabetes using ML methods. In this study, the authors used two datasets (i.e., the PIDD and another dataset) and applied several ML algorithms. Various preprocessing steps, such as label encoding and normalization, were utilized for improving the accuracy rate of the prediction models. The author reported that SVM outperformed the rest of the methods with an accuracy rate of 80.26% on the PIDD, while DT and RF outperformed the other datasets with an accuracy rate of 96.81%. Based on the prediction model, the author developed a smart web application.

The authors of [17] used the PIDD for predicting diabetes using ML methods. A total of five ML algorithms were applied to the processed data, with two additional extracted features. The models were trained using the split method, with 70% of the data used for training and the remaining 30% used for testing. The model's performance was measured using evaluation measures. The reported results reached the highest accuracy rate for the RF model at 88.31%.

The risk factors for diabetes are outlined in [2] using ML techniques. The data collection was carried out using a survey distributed randomly to Indian participants, and 251 responses were received. Three ML algorithms were used: LR, SVM, and RF. The reported results show that LR outperformed the other two methods and achieved an accuracy rate of 96.02%. Likewise, a study conducted by [18] applied various machine learning algorithms to a dataset consisting of 520 observations containing data about both new and diabetic patients. The experimental results exhibited higher accuracy achieved by the bagged method, at 97.7%.

A novel approach of hybrid firefly bat optimized fuzzy artificial neural network (FFBAT-ANN) was proposed by [19] for diagnosing diabetes. In this approach, the fuzzy rules were produced using the LPP method by identifying the features related to diabetes, and the classification was performed using the FFBAT-ANN method. The reported results show the high performance of the proposed method in that FFBAT-ANN achieved a higher accuracy rate of 74.4%. Table 1 summarizes the related work.

3. Methodology

This section will discuss the step-by-step methodology used for conducting this study. Data analysis was performed using Python. The rest of the steps will be discussed in the following subsections.

3.1. Data Collection. The data collection was carried out from the publicly available data source Kaggle [20], which was collected from the behavioral risk factor surveillance system (BRFSS) [21]. The collected data is a cleaned version of the BRFSS, which consists of a total of 253,680 records reflecting the actual responses to the survey conducted by the CDC's BRFSS2015. The dataset comprised a total of 22 features, including the class feature. The class variable (Diabetes_binary) is a binary variable indicating whether the patient has diabetes. More specifically, "0" indicates no diabetes, and "1" indicates prediabetes or diabetes. Moreover, this study used the whole feature set for training the proposed models. Figure 1 shows the features of the dataset.

3.2. Data Preprocessing. One of the challenging steps in building prediction models, and especially healthcare decision support systems, is to prepare the data in a manner conducive to the achievement of reliable results. The raw data collected from real-world scenarios is often incomplete, imbalanced, and not clean [22, 23]. Therefore, before training the model with real-world data, various preprocessing steps must be used to enhance the quality of the data [24]. ML provides several methods for cleaning the data. For example, the missing values can be handled with imputers, etc. In this study, several steps were utilized for handling the inconsistencies in the dataset.

Although the data has no missing values, the dataset was extremely imbalanced, as shown in Figure 2. In an imbalanced data scenario, the data of a certain type are fewer in number than the other types of data in a dataset [25]. Most of the time, the minority class type is of interest for investigation. In Figure 2, the class labeled "0.0" represents 86.07% of the data, while the class labeled "1" accounts for only 13.93%. To balance the class types in a dataset, researchers use various methods, such as the SMOTE [26], random

#	Column	Non-Null	Count	Dtype
0	Diabetes_binary	253680	non-null	float64
1	HighBP	253680	non-null	float64
2	HighChol	253680	non-null	float64
3	CholCheck	253680	non-null	float64
4	BMI	253680	non-null	float64
5	Smoker	253680	non-null	float64
6	Stroke	253680	non-null	float64
7	HeartDiseaseorAttack	253680	non-null	float64
8	PhysActivity	253680	non-null	float64
9	Fruits	253680	non-null	float64
10	Veggies	253680	non-null	float64
11	HvyAlcoholConsump	253680	non-null	float64
12	AnyHealthcare	253680	non-null	float64
13	NoDocbcCost	253680	non-null	float64
14	GenHlth	253680	non-null	float64
15	MentHlth	253680	non-null	float64
16	PhysHlth	253680	non-null	float64
17	Diffwalk	253680	non-null	float64
18	Sex	253680	non-null	float64
19	Age	253680	non-null	float64
20	Education	253680	non-null	float64
21	Income	253680	non-null	float64

FIGURE 1: Dataset description.

oversampling, and other subtypes. In the SMOTE method, the minority class is oversampled in which the minority class samples are considered and generate synthetic samples in the feature area based on the selected k number in the KNN [27].

In this study, the imbalanced dataset problem was handled using SMOTE-ENN. SMOTE-ENN [28] is a powerful method that merges the advantages of both SMOTE and ENN, with SMOTE oversampling the minority class and ENN undersampling the majority class samples [25]. Moreover, ENN drops any samples whose class types are different from the class of at least two of its three nearest neighbors; hence, any sample that is inaccurately classified by its three nearest neighbors is dropped from the training dataset [29]. The application of SMOTE-ENN for handling the imbalanced dataset problem achieved better performance than the single SMOTE method. Similarly, the dataset was normalized using feature scaling, in which the data were transformed between 0 and 1. Feature scaling is a useful method for enhancing model accuracy.

3.3. Prediction Models. In this study, various ML models were applied to the BRFSS dataset. For the building of each model, hyperparameter tuning was performed to choose the best fitted set of parameters that are optimal for achieving the best performance of the model. The models achieved high performance in terms of accuracy, and other evaluation measures were finalized for predicting the high-risk factors of diabetes. The following section discusses the finalized prediction model for this study.

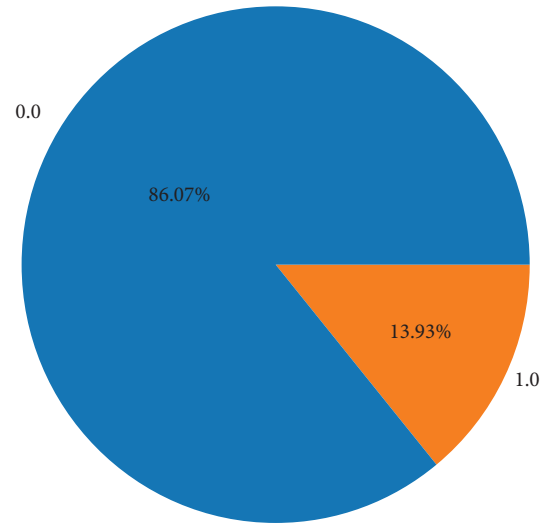


FIGURE 2: Imbalanced dataset.

3.3.1. KNN. KNN is an ML method that classifies the data based on the nearest proximity of training data in a feature set [30]. In this method, the classifier attempts to find the k number of closely similar samples from the training set for predicting the class label of a new sample. Furthermore, the k number is set to an odd number, which ensures that the majority of a class is recognized clearly [31]. In this method, the k number is set to 3 to achieve higher accuracy and other evaluation measures.

3.3.2. RF. RF is an ensemble machine learning technique that utilizes several DT to create a forest. In this method, each DT in the forest is trained using randomly selected training data and a subset of features [31]. Moreover, the main parameter for this method is the number of trees [32]. The majority of trees selected by the RF are the ultimate selection of the classification [33]. In this study, the number of trees was set at 50 for building the RF model. The model evaluation shows the higher performance of the RF model with the best-fitted parameters.

3.3.3. XGBoost. XGBoost (XGB) is a recently developed ML algorithm proposed by Chen and Guestrin [34] in 2016. This is an enhanced algorithm based on gradient boosting DT that can significantly build boosted trees and execute them in parallel [35]. In the iteration process, gradient boosting seeks to enhance the robustness by dropping the loss function of the algorithm as well as the gradient direction [25]. XGB trains multiple classifiers slowly and sequentially. Like RF, the boosting algorithm is using DT, but it depends on individuals how to utilize them [36]. In this study, the number of trees was set to 100 based on the suggested hyperparameter tuning test for building the XGB model.

3.3.4. Bagging. Bagging is an ensemble learning method combining several classifiers using training data, in which different training data are presented for learning in each

instance. Moreover, the new training set is generated by randomly selected examples with replacements from the original training set. A class achieving the majority of votes wins [37]. Moreover, in this method, several trees using a bootstrap sampling of the training set are created and integrated into their individual predictions to achieve the final classification. In this study, the number of trees per hyperparameter tuning is set to 100 with the bootstrap method. The model shows higher performance in terms of accuracy and other evaluation measures.

3.3.5. AB. AB is an ensemble ML method that aims to integrate several weak classifiers and transform them into strong ones [38]. In this method, DT is used as a default base estimator for training the model. The base estimator in AB is a weak learner in which every tree is trained to reduce the weakness by learning from the trees being trained that are boosted using weights. Moreover, this is a loop-based method in which weights are assigned to train the data in every iteration of the loop. The iteration process continues until the accurate classification of the data is confirmed [37]. Per the hyperparameter tuning, the number of trees was set to 100 for building the AB model.

3.4. Model Evaluation. Model evaluation is the practice of measuring the prediction results of the model built and then comparing those results against the real data, which is generally known as test data [39]. For model evaluation, there are several methods available, but this study utilized the percentage split method. In this method, the processed dataset was split into two sets; 70% of the whole dataset was used for training the aboveproposed models, and the remaining 30% was used for testing the efficacy of the proposed models. The model evaluation shows the higher performance of the proposed model.

4. Experimental Results

4.1. Experimental Setup. The prediction models discussed in the above sections were applied to the BRFS dataset for detecting the risk factors associated with diabetes, which can be useful for diagnosing diabetes in patients at an early age. As noted above, the dataset was initially split into two subsets; the training set comprised 70% of the total dataset, while the remaining 30% was used as the testing set. During the experiment, several attempts were made to finalize the best classifiers to accurately detect the risk factors. Therefore, a hyperparameter test was utilized to set the most suitable parameters of each classifier to maximize the likelihood of predictions in terms of selecting an accurate model that can help medical practitioners in decision-making about diabetes patients. After running several experiments with best fitted parameters on the processed data, and the best classifiers according to accuracy and other measures were used to report the results.

In the experimental phase, for building each model, a confusion matrix is computed, which provides four important values: true-positive (tp), true-negative (tn), false-

positive (fp), and false-negative (fn), as shown in Figure 3. The model evaluation was performed on the basis of these four values using the following measures:

- (i) Accuracy is the ratio of correctly identified diabetes patients to the whole number that is predicted [40]. Equation (1) shows the mathematical representation of accuracy.

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}. \quad (1)$$

- (ii) Precision, a measure calculated using equation (2), is the ratio of correctly identified patients with diabetes to all patients with diabetes [41].

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}. \quad (2)$$

- (iii) Recall or sensitivity, calculated using equation (3), is the ratio of correctly classified diabetes patients to the whole numbers in that particular class [41].

$$\text{Recall or Sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}}. \quad (3)$$

- (iv) *F*-measure is the weighted average of precision and recall [40] and is mathematically calculated using .

$$F - \text{measure} = \frac{(2 * \text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}. \quad (4)$$

- (v) Specificity is a performance measure of a model that is defined as the ratio of correctly classified patients without diabetes to all patients who do actually have diabetes [41]. Specificity is also known as true-negative rate (TNR).

- (vi) ROC is a visualized curve that measures the performance of classifiers at various thresholds, while the AUC is a measurement of separability between the class labels. A higher AUC value shows a higher performance of the model in terms of accurately differentiating between patients with and without diabetes [40].

5. Results and Discussion

Comparing the experimental results of the proposed method to the existing state-of-the-art methods in the literature, our proposed method showed high performance in terms of accuracy, precision, sensitivity, specificity, *f*-measure, and ROC/AUC score. Table 2 shows the comparison of the proposed method to prominent existing studies using the BRFS dataset. Although the proposed prediction models showed higher performance compared to the existing, Table 2 reported the KNN results in the comparison table.

On the BRFS dataset, our proposed method showed higher performance than the existing methods in that KNN achieved an average test accuracy of 98.363%; precision, sensitivity, and *f*-measures of 98%; and ROC/AUC score of 98.3%, which are the highest values so far. The reason the

Actual	Predicted	
	0	1
0	42572	1004
1	657	57247

KNN

Actual	Predicted	
	0	1
0	41201	2375
1	2097	55807

RF

Actual	Predicted	
	0	1
0	41790	1786
1	3265	54639

XGB

Actual	Predicted	
	0	1
0	41158	2418
1	2995	54909

Bagging

Actual	Predicted	
	0	1
0	41190	2386
1	3346	54558

AB

Note:
0 = No diabetes
1 = Diabetes

Diabetes types

FIGURE 3: Confusion matrix.

TABLE 2: Comparison of the proposed method with existing studies used BRFSS dataset.

Study	Dataset	Method	Accuracy (%)	Sensitivity	Specificity	AUC
[4]	BRFSS-2014	NN	82.4	0.378	0.902	0.795
[14]	BRFSS-2017	RF	86.8	—	—	—
Proposed method	BRFSS-2015	KNN	98.36	0.98	0.98	0.983

TABLE 3: Comparison of the proposed method with existing studies that used other datasets.

Study	Dataset	Method	Accuracy (%)	Precision	Sensitivity	Specificity	F-measure
[8]	Private	RF	80.84	—	0.85	0.767	—
	PIDD	RF	77.21	—	0.746	0.799	—
[13]	PIDD	SVM, AB	94.44	0.971	0.910	—	—
[16]	PIDD	LR,SVM	78.85, 77.71	0.788, 0.774	0.789, 0.777	—	0.788,0.775
		NN	88.6	—	—	—	—
[17]	PIDD	RF	88.31	0.88	0.86	—	0.87
[2]	Private	LR	96.02	0.887	0.857	—	0.871
Proposed method	BRFSS	KNN	98.36	0.98	0.98	0.98	0.98

proposed methods were able to achieve high accuracy and other evaluation measures is the use of the SMOTE-ENN method, which is used for balancing the dataset in the preprocessing step. The SMOTE method alone was also tested on the BRFSS dataset, but the performance of the proposed models was not much different from that found in the existing studies. Therefore, the use of SMOTE-ENN is more powerful than the SMOTE method alone.

Similarly, our KNN method also outperformed those of other studies that used other prominent datasets, such as PIDD and other private datasets, as shown in Table 3. This

shows the reliability of our proposed method for predicting the risk factors of diabetes.

Moreover, the individual performance of each proposed method with a detailed discussion is shown in the following tables and figures. Figure 4 shows the accuracy of the proposed methods in predicting the high-risk factors for detecting and diagnosing diabetes patients at an early stage.

Moreover, the proposed methods were also evaluated using precision, sensitivity, specificity, f-measure, and AUC scores. Precision, which is also referred to as positive predictive value (ppv), here refers to the fraction of accurately

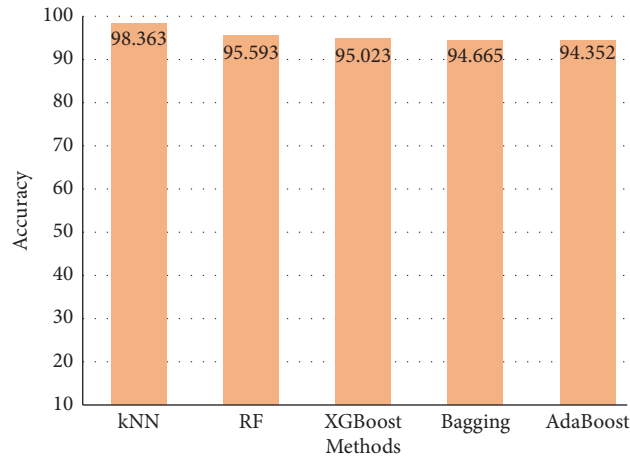


FIGURE 4: Accuracy of the proposed methods.

TABLE 4: Model evaluation measures.

Classifier	Precision	Sensitivity	Specificity	F-measure	AUC
kNN	0.98	0.98	0.98	0.98	0.983
RF	0.96	0.95	0.95	0.95	0.955
XGBoost	0.95	0.95	0.96	0.95	0.951
Bagging	0.93	0.94	0.94	0.94	0.946
AdaBoost	0.94	0.94	0.95	0.94	0.944

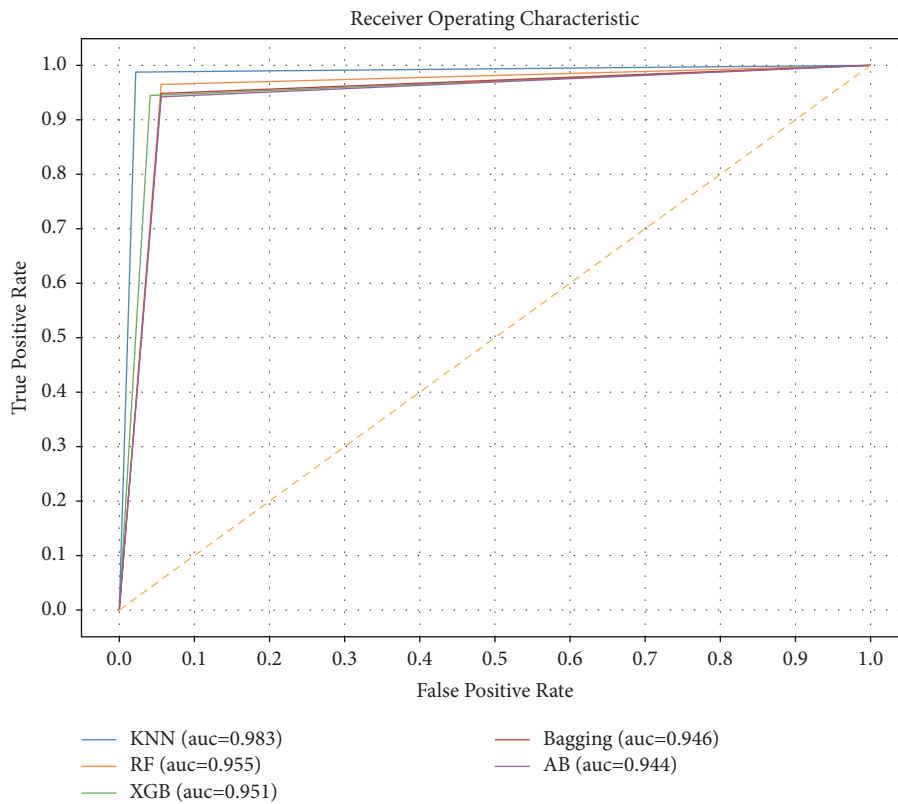


FIGURE 5: ROC curves of prediction models.

classified patients having diabetes over the total number of patients who actually have diabetes [41, 42]. The precision is also called the confidence of the prediction model.

Sensitivity is the fraction of accurately classified patients with diabetes over the total number of patients in that class [40]. The F-measure is the harmonic mean of ppv and

sensitivity [41]. Table 4 shows the model evaluation measures.

The values in Table 4 are the average measures for a model evaluation that surpasses the values in the comparison in Table 2, which shows the reliability of the proposed models in detecting diabetic patients to help medical practitioners in diagnosing the patients at an early stage.

Similarly, the model was also evaluated using the ROC curves. ROC curves are highly beneficial for creating classifiers and visualizing their performance and are commonly utilized in healthcare decision-making [37], because they envisage the whole scenario of the trade-off between sensitivity and false-positive rate across a set of thresholds and are considered a powerful measure of a diagnostic test [43]. In the ROC, the AUC values decide the performance of a model. The higher the AUC score, the higher the performance of a prediction. An AUC value close to the left upper corner shows the high performance of the model. The AUC score shown in Table 4 is high, as it is very close to the left upper corner, and this is reflected in the ROC graph, as shown in Figure 5.

To summarize the above discussion, it is essential to prepare the data in a high-quality manner, especially for prediction purposes. Predictions are actually based on historical data from which the hidden patterns are extracted to form the basis for predicting the unseen cases. Therefore, the historical data should be of high quality, especially when the predictions are made in the healthcare field, where lives are at high risk. For these reasons, several preprocessing steps must be performed to remove outliers, handle the missing values, and balance the data in a manner that allows for the building of high-quality prediction models that can help medical practitioners in deciding about a particular disease.

The dataset used in this study was preprocessed in advance but was extremely imbalanced. The data imbalance issue was handled using SMOTE-ENN, which is a more powerful method than the SMOTE method alone. Thus, several ML algorithms were applied to the processed data. For the building of each model, hyperparameter tuning was performed to choose the best fitted model architecture for detecting the high-risk factors of diabetes. After running several experiments with optimal model architecture on the processed data, and the best classifiers according to accuracy and other measures were used to report the results. In this study, the finalized classifiers for detecting the high-risk factors of diabetes are KNN, RF, XGBoost, Bagging, and AdaBoost. The results achieved by these models were also compared to the existing state-of-the-art studies, and the efficacy of our proposed methods was found to be higher in terms of testing accuracy, precision, sensitivity, f-measure, and ROC/AUC score. This shows that the proposed models can be used as a decision-making process for detecting high-risk factors for diabetes and can also help medical practitioners in diagnosing diabetes patients in the early stages.

6. Conclusion and Future Work

This study was conducted to provide a system that can automatically detect the risk factors of diabetes as well as to

provide an automatic decision-making system that can help medical practitioners in diagnosing diabetes patients based on risk factors. For that purpose, various preprocessing methods were used to prepare the data to increase the likelihood of prediction and increase the opportunity for developing reliable models. Moreover, hyperparameter tuning was performed for the building of each model to finalize the optimal parameter set that can achieve the maximum possible accuracies. Therefore, various experiments were performed on the processed BRFS dataset in which the finalized methods discussed in the above sections achieved the best possible results in terms of accuracy, precision, sensitivity, specificity, f-measure, and ROC/AUC score. Among them, KNN outperformed the best-fitted model compared to others and even the state-of-the-art methods available in the literature. The reason behind the high performance of the proposed method was the use of the SMOTE-ENN method for handling the imbalanced dataset problem. The study has also attempted to use the SMOTE method alone, but the results were not much different from those of the existing studies. The use of SMOTE-ENN made it possible to achieve higher accuracies of the proposed models compared to the existing ones. This confirms the reliability of the proposed method for detecting the risk factors of diabetes as well as for providing accurate decision support systems for diagnosing diabetes early before it becomes chronic.

In the future, our model can be tested on other datasets collected from different clinics and research centers. The model efficiency can be enhanced using other advanced methods in the future.

Data Availability

The data were taken from the publicly available data source Kaggle [20].

Conflicts of Interest

There are no conflicts of interest.

Acknowledgments

This research work was funded by Institutional Fund Project under grant no. (IFPIP-381-611-1442). Therefore, the authors gratefully acknowledge technical and financial support from the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

References

- [1] Y. Pan, M. Shao, P. Li et al., "Polyaminoglycoside-mediated cell reprogramming system for the treatment of diabetes mellitus," *Journal of Controlled Release*, vol. 343, 2022.
- [2] M. S. Akanksha, K. Vinutna, and M. N. Thippeswamy, "Analysing machine learning techniques in Python for the prediction of diabetes using the risk factors as parameters," *Lecture Notes in Electrical Engineering*, vol. 790, pp. 619–639, 2022.

- [3] H. Zhang, Y. Ben, Y. Han, Y. Zhang, Y. Li, and X. Chen, "Phthalate exposure and risk of diabetes mellitus: implications from a systematic review and meta-analysis," *Environmental Research*, vol. 204, Article ID 112109.
- [4] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building risk prediction models for type 2 diabetes using machine learning techniques," *Preventing Chronic Disease*, vol. 16, no. 9, Article ID 190109, 2019.
- [5] A. A. Motala, J. C. Mbanya, K. Ramaiya, F. J. Pirie, and K. Ekoru, "Type 2 diabetes mellitus in sub-Saharan Africa: challenges and opportunities," *Nature Reviews Endocrinology*, vol. 18, no. 4, pp. 219–229, 2022.
- [6] IDF, "International Diabetes Federation," 2019, <https://www.idf.org/aboutdiabetes/what-is-diabetes.html>.
- [7] WHO, "Diabetes," 2022, https://www.who.int/health-topics/diabetes#tab=tab_1.
- [8] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, pp. 1–10, 2018.
- [9] N. Ahmed, R. Ahammed, M. M. Islam et al., "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, 2021.
- [10] M. Rizwan, A. Shabbir, A. R. Javed et al., "Risk monitoring strategy for confidentiality of healthcare information," *Computers & Electrical Engineering*, vol. 100, Article ID 107833, 2022.
- [11] M. Jamjoom, "Data mining in healthcare to predict cesarean delivery operations using a real dataset," in *Proceedings of the First International Conference on Computing and Emerging Sciences ICCE'2020*, pp. 20–26, Erbil, Iraq, December 2020.
- [12] M. Aminul and N. Jahan, "Prediction of onset diabetes using machine learning techniques," *International Journal of Computer Application*, vol. 180, no. 5, pp. 7–11, 2017.
- [13] A. Aada and S. Tiwari, "Predicting diabetes in medical datasets using machine learning techniques," *Int. J. Sci. Res. Eng. Trends*, vol. 5, no. 2, pp. 257–267, 2019.
- [14] G. Bholra, A. Garg, and M. Kumari, "Comparative study of machine learning techniques for chronic disease prognosis," *Computer Networks and Inventive Communication Technologies*, vol. 58, pp. 131–144, 2021.
- [15] A. Al-Zebari and A. Sengur, "Performance comparison of machine learning techniques on diabetes disease detection," in *Proceedings of the 1st International Informatics and Software Engineering Conference*, pp. 1–4, UBMYK, Ankara, Turkey, November 2019.
- [16] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021.
- [17] S. Nahzat and M. Yağanoğlu, "Diabetes prediction using machine learning classification algorithms," *Eur. J. Sci. Technol.* vol. 24, pp. 53–59, 2021.
- [18] B. K. Sahu, N. Ghosh, B. Kumar Sahu, and N. Ghosh, "Early stage prediction of diabetes using machine learning techniques," in *Advances in Distributed Computing and Machine Learning*, pp. 310–317, Springer, Berlin, Germany, 2022.
- [19] G. T. Reddy and N. Neelu, "Hybrid Firefly-Bat optimized fuzzy artificial neural network based classifier for diabetes diagnosis," *International Journal of Intelligent Engineering and Systems*, vol. 10, no. 4, pp. 18–27, 2017.
- [20] A. Teboul, "Diabetes Health Indicators Dataset," 2022, <https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset>.
- [21] CDC, "Behavioral Risk Factor Surveillance System (BRFSS)," 2022, <https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system?select=2015.csv>.
- [22] A. S. Al-Mudimigh and Z. Ullah, "Prevention of Dirty Data and the Role of MADAR Project," in *Proceedings of the 2011 UKSim 5th European Symposium on Computer Modeling and Simulation*, Madrid, Spain, November 2011.
- [23] A. S. Al-Mudimigh, Z. Ullah, and T. A. Alsubaie, "A framework for portal implementation: a case for Saudi organizations," *International Journal of Information Management*, vol. 31, no. 1, pp. 38–43, 2011.
- [24] H. Ahmad, S. Ahmad, M. Asif, M. Rehman, A. Alharbi, and Z. Ullah, "Evolution-based performance prediction of star cricketers," *Computers, Materials & Continua*, vol. 69, no. 1, pp. 1215–1232, 2021.
- [25] M. Lin, X. Zhu, T. Hua, X. Tang, G. Tu, and X. Chen, "Detection of ionospheric scintillation based on xgboost model improved by smote-enn technique," *Remote Sensing*, vol. 13, p. 2577, 2021.
- [26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 2, pp. 321–357, 2002.
- [27] S. Saleem, S. S. Naqvi, T. Manzoor, A. Saeed, N. ur Rehman, and J. Mirza, "A strategy for classification of 'vaginal vs. Cesarean section' delivery: bivariate empirical mode decomposition of cardiocographic recordings," *Frontiers in Physiology*, vol. 10, pp. 1–18, 2019.
- [28] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.
- [29] J. Luengo, A. Fernández, S. García, and F. Herrera, "Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling," *Soft Computing*, vol. 15, no. 10, pp. 1909–1936, 2011.
- [30] S. Dutta and S. K. Bandyopadhyay, "Fake job recruitment detection using machine learning approach," *International Journal of Engineering Trends and Technology*, vol. 68, no. 4, pp. 48–53, 2020.
- [31] S. Mehta and K. S. Patnaik, "Improved prediction of software defects using ensemble machine learning techniques," *Neural Computing & Applications*, vol. 33, no. 16, Article ID 10551, 2021.
- [32] S. E. Seker and I. Ocak, "Performance prediction of road-heads using ensemble machine learning techniques," *Neural Computing & Applications*, vol. 31, no. 4, pp. 1103–1116, 2019.
- [33] N. K. Sinha, M. Khulal, M. Gurung, and A. Lal, "Developing A web based system for breast cancer prediction using XGboost classifier," *International Journal of Engineering Research and Technology*, vol. 9, no. 6, pp. 852–856, 2020.
- [34] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, Sydney, Australia, August 2016.
- [35] J. Liu, J. Wu, S. Liu, M. Li, K. Hu, and K. Li, "Predicting mortality of patients with acute kidney injury in the ICU using

- XGBoost model,” *PLoS One*, vol. 16, no. 2, Article ID e0246306, 2021.
- [36] H. Hu, A. J. van der Westhuysen, P. Chu, and A. Fujisaki-Manome, “Predicting Lake Erie wave heights and periods using XGBoost and LSTM,” *Ocean Modelling*, vol. 164, Article ID 101832, 2021.
- [37] Z. Ullah, F. Saleem, M. Jamjoom, and B. Fakieh, “Reliable prediction models based on enriched data for identifying the mode of childbirth by using machine learning methods: development study,” *Journal of Medical Internet Research*, vol. 23, no. 6, Article ID 288566, 2021.
- [38] H. Tabassum, G. Ghosh, A. Atika, and A. Chakrabarty, “Detecting online recruitment fraud using machine learning,” in *Proceedings of the 9th International Conference on Information and Communication Technology ICoICT*, vol. 2021, pp. 472–477, Yogyakarta, Indonesia, August 2021.
- [39] T. C. Smith and E. Frank, “Introducing machine learning concepts with WEKA,” in *Statistical Genomics*, pp. 353–378, Humana Press, New York, NY, USA, 2016.
- [40] Z. Ullah and M. Jamjoom, “An intelligent approach for Arabic handwritten letter recognition using convolutional neural network,” *PeerJ Computer Science*, vol. 8, p. e995, 2022.
- [41] Z. Ullah and M. Jamjoom, “A deep learning for alzheimer’s stages detection using brain images,” *Computers, Materials & Continua*, vol. 74, 2022.
- [42] M. Al-Sudairi, A. S. Al-Mudimigh, and Z. Ullah, “A Project management approach to service delivery model in portal implementation,” in *Proceedings of the IEEE Second International Conference on Intelligent Systems, Modelling and Simulation*, pp. 329–331, Phnom Penh, Cambodia, January 2011.
- [43] M. Rizwan Ali, F. Ahmad, M. Hasanain Chaudary et al., “Petri Net based modeling and analysis for improved resource utilization in cloud computing,” *PeerJ Computer Science*, vol. 7, pp. 1–22, 2021.

Research Article

Blockchain-Based Optimization Model for Evaluating Psychological Mental Disease and Mental Fitness

Jayashree Rajesh Prasad ¹, Shashikant V. Athawale ², Roshani Raut ³, Sonali Patil ³,
Sheetal U. Bhandari ⁴ and Mohd Asif Shah ⁵

¹School of Engineering, MIT Art, Design and Technology University, Pune 412201, Maharashtra, India

²Department of Computer Engineering, AISSMS COE, Savitribai Phule Pune University, Pune, India

³Department of Information Technology, Pimpri Chinchwad College of Engineering, Akurdi, Savitribai Phule Pune University, Pune, Maharashtra, India

⁴Department of E & T/C Engineering, Pimpri Chinchwad College of Engineering, Pune, India

⁵Kebri Dehar University, Somali, Ethiopia

Correspondence should be addressed to Mohd Asif Shah; drmhohdasifshah@kdu.edu.et

Received 7 May 2022; Revised 18 June 2022; Accepted 25 June 2022; Published 14 July 2022

Academic Editor: Farman Ali

Copyright © 2022 Jayashree Rajesh Prasad et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The current work describes a blockchain-based optimization approach that mimics the psychological mental illness evaluation procedure and evaluates mental fitness. Combining lightweight models with blockchains can give a variety of benefits in the healthcare business. This study aims to offer an improved review and learning optimization technique (SPLBO) based on the social psychology theory to overcome the biogeography-based optimization (BBO) algorithm's shortcomings of low optimization accuracy and instability. It also creates high-accuracy solutions in recognized domains quickly. To retain student individuality, students can be divided into two groups: Human psychological variables are incorporated in the algorithm's improvement: in the "teaching" step of the original BBO algorithm; the "expectation effect" theory of social psychology is combined: "field-independent" and "field-dependent" cognitive styles. As a consequence, low-weight deep neural networks have been designed in such a manner that they require fewer resources for optimal design while also improving quality. A responsive student update component is also introduced to duplicate the effect of the environment on students' learning efficiency, increase the method's global search capabilities, and avoid the problem of falling into a local optimum in the first repetition.

1. Introduction

Machine learning algorithms can effectively identify possible features from the provided data; therefore, they do not require any hand-crafted characteristics throughout the simulation. Neural network models can effectively identify possible features from the provided data; therefore, they do not need any manual features throughout the prediction phase. Teaching-learning-based optimization (TLBO) is a new heuristic algorithm proposed by author [1]. The algorithm simulates the teaching and learning process design of teachers and students. Students can acquire knowledge from teachers' teaching and understanding through interaction

between students. The TLBO algorithm has the advantages of few parameters, simple structure, and fast solution speed, and its competitiveness mainly comes from the ingenious design of the teaching and learning stages. Compared with other typically improved intelligent optimization methods, the TLBO algorithm also shows its outstanding performance and advantages. However, simultaneously, it holds a large number of drawbacks such as to take up a lot of resources like storage and memory. It is a time-consuming procedure because it requires several iterations and thus processing takes a longer time than usual, which are the main concerns of the TLBO algorithm. Furthermore, as compared with the Genetic Algorithm (GA), TLBO allows the population to

learn from the optimal individual in the teaching phase, thereby improving the convergence speed of the algorithm; compared with Particle swarm optimization (PSO), for a single operator, TLBO introduces one more learning stage than PSO, which is beneficial to improve the exploration ability of the algorithm; compared with Cuckoo Search (CS), TLBO provides interactive learning in the learning stage method [2, 3]. It is precisely because TLBO has these advantages that scholars have never stopped studying it since the algorithm was proposed. However, the TLBO algorithm also has shortcomings such as low optimization accuracy, poor stability, and slow convergence speed. Many scholars have improved it from multiple perspectives. The improvement direction is mainly divided into three aspects: improving the teaching process, introducing weights or adaptation factors, and combining with other intelligent optimization algorithms. Among them, the improvement of the teaching process refers to adding a self-study stage or introducing new learning rules based on the original teaching stage and learning stage [4]. Collaborative Learning Model (CLM) is used for the learning phase. In the CLM method, to guide the learners effectively, the teacher will adaptively update its position according to the neighborhood information in the self-learning stage [5]. A novel optimization algorithm based on autonomous learning was proposed and the authors remodeled the proposed algorithm according to the three stages of the teaching process: teacher's learning, mutual learning, and self-learning between students. The literature introduced a teaching and learning algorithm with logarithmic helical strategy and triangular mutation rule (LNTLBO) to enhance the exploration and development ability in the learning stage [6]. The literature proposed a teaching and learning optimization algorithm based on multi-reverse learning, established a hybrid reverse learning model, and added a self-learning stage based on search boundary guidance, making the algorithm more robust to global search and local detection capability [7].

The literature adopted a different feedback learning stage to speed up the convergence, further recording the previous generation teachers and communicating with the current teachers to provide comprehensive feedback to the learners and supervise the learning direction to avoid wasting previous generation computation quantity [8]. The literature let teachers perform dynamic random search algorithms in the later stage of the algorithm to improve the ability of the optimal individual to explore new solutions [9]. For improvements in introducing weights or unknown parameters, the literature has proposed Advanced Teaching Learning-Based Optimization (ATLBO). New weight parameters were introduced to improve the accuracy and speed up convergence [10]. Authors have proposed the nonlinear inertia-weighted teaching-based optimization algorithm (NIWTLBO) [11]. The algorithm introduces a nonlinear inertia weighting factor into the basic TLBO to control the learner's memory rate. It uses a dynamic inertia weighting factor to replace the original random number in the teaching and learning stages. The literature introduced the crossover operator of the difference algorithm in the "teaching" stage

and the "learning" stage, and at the same time carried out an adaptive local search according to the normal distribution around the elite individuals to improve the convergence speed and solution accuracy of the algorithm [12]. To enhance the performance of TLBO, many scholars try to integrate it with other optimization algorithms. The literature added an error correction strategy and Cauchy distribution (ECTLBO) in TLBO, where Cauchy distribution is used to expand the search space and correct wrong to avoid detours for a more accurate solution [13]. The literature combined harmony search with the teaching and learning optimization algorithm and proposed a hybrid optimization algorithm (HHSTL) based on harmony search and teaching and learning optimization, which enabled the algorithm to solve more complex problems [14]. The literature proposed an improved teaching and learning optimization algorithm (ITLBOBSO) incorporating the idea of brainstorming and introduced Cauchy mutation and a random parameter associated with the number of iterations in the operator to improve the performance of the algorithm [15]. The literature proposed a hybrid search algorithm named HSTLBO, in which HS mainly aims to explore unknown regions. In contrast, TLBO aims to rapidly develop high-accuracy solutions in known areas [16].

TLBO is a new heuristic algorithm that takes people as the main body of activities and simulates teaching phenomena. The improvement of TLBO mainly focuses on manufacturing the teaching process and combining it with other algorithms. However, very little consideration is given to people's psychological and emotional factors, such as considering the influence of psychological factors on behavior results from the perspective of people; individuals with different personalities show different states in the same environment. This improvement makes the algorithm have a specific revision in the optimization performance, but there is still room for improvement in stability and convergence speed. Figure 1 shows the mental state relation with the students.

To further improve the algorithm's performance, this paper focuses on social psychology, considers human emotions and behaviors, and simulates the impact of human psychological factors on the results in the teaching process. We apply the social psychological theory to algorithm improvement. First, the "expectation effect" theory is added to the teaching stage [17]. The theory states that in interpersonal interactions, one party has expectations of the other party. The party that has expectations will treat the other party as he expects, thereby causing changes in the other party's behavior. It is reflected in the algorithm that the individual teacher provides one-to-one teaching to the students with good fitness value, and the teacher guides the students who are taught one-to-one. They also change their learning behavior and begin to learn from other students. The theory of "field independence-field dependence" is added [18]. This theory divides people's cognitive styles into "field-independent" and "field-dependent" according to their different degrees of dependence on the external environment. Starting from the actual situation, considering the different cognitive styles of international students, we

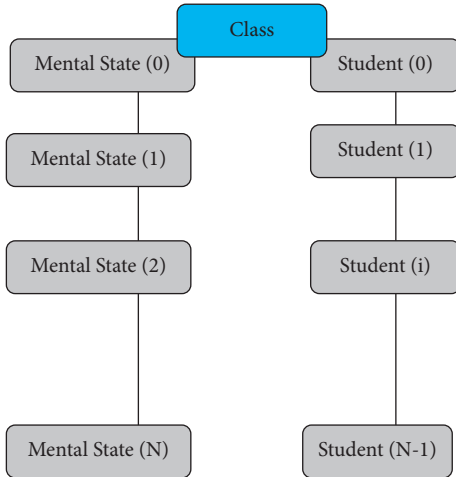


FIGURE 1: Mental state relation to student.

simulated field-independent and field-dependent students and adopted different learning strategies. After being taught by teachers and learning from each other, students need to digest knowledge points and evaluate their rankings. The learning method is extended or adjusted according to the ranking situation, and thus the self-learning method adjustment phase is added after the learning phase [19]. Bandura’s “self-regulation theory” exists in psychology, which shows that self-regulation includes three processes: self-observation, self-judgment, and self-reaction. According to this theoretical score, different self-regulation strategies are used for students in various positions to achieve a better state. Finally, a series of test functions prove that the improved algorithm has obtained better performance in terms of optimization accuracy and convergence speed. The TLBO method has a number of flaws, including poor optimization accuracy, instability, and slow convergence time. It has been improved by a number of academics from a number of views.

Adding social psychology to algorithm improvement also has specific innovations in intelligent optimization algorithm improvement. The current work describes a blockchain-based optimization approach that mimics the psychological mental illness evaluation procedure and evaluates mental fitness. Combining lightweight models with blockchains can give a variety of benefits in the healthcare business. This study aims to offer an improved review and learning optimization technique (SPLBO) based on the social psychology theory to overcome the biogeography-based optimization (BBO) algorithm’s shortcomings of low optimization accuracy and instability. Since people are a complex system affected by their psychological state, they will take timely measures to adjust their behavior to seek a sense of self-protection. Therefore, compared with the improved methods of other intelligent optimization algorithms, incorporating human psychological factors can make the algorithm improvement more flexible and allow the algorithm to balance global search and local search. This research paper, focusing on the defects of low optimization precision and slow convergence speed when solving

complex optimization problems of the teaching and learning optimization algorithm, from the perspective of social psychology, combined with the changes of people’s psychological emotions, improves the original teaching and learning optimization algorithm.

2. Blockchain-Based Teaching and Learning Optimization Algorithm (B-TLBO)

B-TLBO is an algorithm designed to simulate the two stages of teacher teaching and student learning in the process of simulating class teaching. It uses the entire population as a class, the best individuals in the population as teachers, and the other individuals as students. The concept of the B-TLBO algorithm comes from the replicated class’s teaching process; to better replicate the new state of middle school students in the teaching phase, an adaptable student updating factor is incorporated, and the method is expected to produce superior results. The algorithm is divided into the “teaching stage” and the “learning stage.”

Figure 2 depicts the B-TLBO method that relies on the teaching process of a repeated class and has two stages as discussed. The “teaching stage” refers to when the entire student body learns from the teacher, while the “learning stage” corresponds to when the students learn from one another. The total level of the population is improved by the co-evolution of these two stages. In this study, N denotes the total number of students (i.e., the population size), and d is the number of subjects studied by each student (i.e., the individual dimension). Each student is identified as $S_i = \{S_1, S_2, S_3, S_n\}$ with the fitness function $f(x_i)$ indicating the student’s grade; the higher the fitness value, the higher the grade. The specific content of the algorithm is described in two stages, namely the teaching stage and the learning stage, respectively. The best fitness value for each of the iteration has been selected to convey the knowledge to the students in the best possible way; similarly, the learning stage is the technique of combined learning of all students in a group after the accomplishment of the teaching stage and here the fitness function is chosen to select the best student among the students. Hence, we can complement that the two methods in the BTLBO algorithm are employed for the enhancement of the fitness functions. The two techniques are listed as follows:

2.1. Teaching Phase. In the teaching phase, the individual with the best fitness value for each of the iteration will be selected as the teacher T_X . The teacher imparts knowledge to the students to improve the average grade of the whole class. He hopes that the overall middle position of the class T_M is close to its T_X . Therefore, the teaching method design is given by formula (1):

$$t_{i,\text{new}} = t_i + r_i(t_x - t_f t_m), \quad (1)$$

where $t_{i,\text{new}}$ represents the new state of student i after learning in the teaching stage; t_i is the original state of student i before learning; r_i is a random number on $[0, 1]$; the influence degree of the value generally takes 1 or 2. After

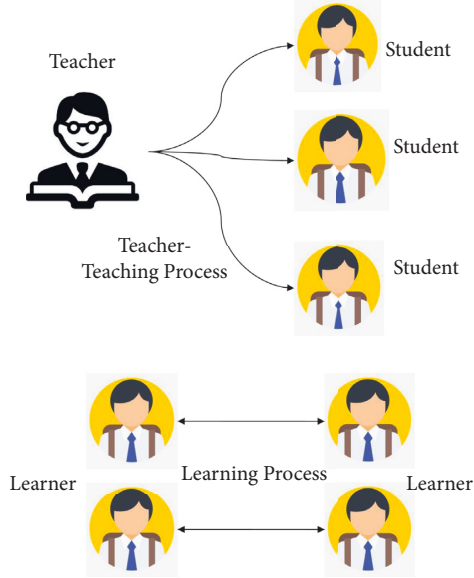


FIGURE 2: Teacher learning phenomena.

the teaching phase is completed, the students update the knowledge reserve, and each student decides whether to update according to the new state or the original state. Take the minimization problem as an example:

$$\text{if } f(t_{i,\text{new}}) < f(t_i). \quad (2)$$

2.2. Learning Stage. This stage simulates the process of mutual learning among students after the class is over. To further improve their learning level, students communicate with other individuals in the class. Student_{*i*} randomly selects student_{*j*}, compares the fitness values of the two students, and subtracts the second-best student from the position of the best student. Taking the minimum optimization problem as an example, the learning is carried out in the following way:

$$t_{i,\text{new}} = \begin{cases} t_i + \text{rand}(t_i - t_j), \\ f(t_i) < f(t_j), \\ t_i + \text{rand}_i(t_j - t_i), \end{cases} \quad (3)$$

where rand_i is a random number on (0, 1). After the learning period is over, perform the same update operation on the students as in the teaching period again.

3. Teaching and Learning Algorithms Based on the Social Psychological Theory

To further improve the algorithm's performance, this paper improves three aspects of the teaching and learning algorithm from the perspective of human emotion and psychology, combined with the social psychology theory. Firstly, the "expectation effect" is introduced in the teaching stage, and students who bear different expectations will

adopt different learning strategies. The idea of "field independence-field dependence" is presented at the learning stage to distinguish the differences in the learning styles of other students. Finally, considering the actual teaching situation, a self-learning method adjustment phase is added after the teaching and learning phases to adjust students' learning methods in time.

3.1. Introducing the "Expectation Effect" Theory to Improve the "Teaching Stage". Teachers always have higher expectations for students with relatively good grades in daily teaching. To get better grades, teachers will take one-on-one teaching or set up advanced courses and other methods. An expectation is a judgment about oneself or others that one expects to achieve a specific goal or meet a confident behavioral expectation. Students with better grades will take active measures to study harder to live up to teachers' expectations after teachers have focused on them. For example, they can improve themselves by increasing their study time and sharing their learning experiences with their classmates. This phenomenon is known in social psychology as the "Pygmalion effect" or the "expectation effect." An expectation is a judgment about oneself or others that one expects to achieve a specific goal or meet a confident behavioral expectation. The behavioral outcome that results from expectations is the expectation effect. In this statement, the author wants to express the psychological aspect about the person's expectation of the other individual, and it means that an expectation is a kind of judgment about a person or somebody that is referred as the "expectation effect." Expectation emphasizes the activity process of the individual's psychological stimulation, while the expectation effect focuses on the behavioral results produced by psychological stimulation. The literature introduced this theory into business management practice [20]. The results show that managers' expectations of subordinates and how they treat associates determine the work performance and career progress of these subordinates to a large extent. Inspired by this, the algorithm is improved: Classify students whose grades are above the class average as outstanding students, and learn by combining one-to-one teaching with teachers and learning from other students, as shown in formula (4):

$$t_{i,\text{new}} = t_i + r_i(t_x - i_x) + r_i(t_{r1} - i_{r2}). \quad (4)$$

Students whose grades are below the class average will study according to formula (5):

$$t_{i,\text{new}} = t_i + r_i(t_x - f_t * \text{mean}). \quad (5)$$

Among them, t_{r1} and t_{r2} are the status of any two students in the class, and mean is the average level of the classmates. When the fitness value of a student is higher than the average, it will bear the high expectations of the teacher. It can be seen from formula (4) that to meet the teacher's expectations in the learning process, in addition to relying on its own knowledge state t_i , the student_{*i*} also adopts one-to-one learning from the teacher.

Compared with formula (1), the student refers to the average difference between the teacher and the class in the learning process. Under this learning method, the space for excellent students to improve their grades is limited, which will reduce the speed of the algorithm converging to the optimal solution. The strategy of one-to-one learning from teachers is added to the improved formula (4), which increases the influence of teachers on students. This design allows outstanding students to approach teachers quickly and enables students to jump out of their limitations. To solve the problem of poor local search ability of the algorithm and to speed up the algorithm's convergence, a strategy of learning from classmates is added to equation (4). Students with an average score or above exchange experience with any classmate in the class, which improves themselves and helps others, improve their performance, thereby narrowing the gap between classes and accelerating the process of convergence of the entire population to the optimal value. When student's fitness value is lower than average, he does not bear the teachers' high expectations, and so his learning style will not change.

3.2. Introduce the "Field-Independent-Field-Dependent" Theory to Improve the "Learning" Stage. The concept of "field independence-field dependency" is introduced throughout the learning stage to distinguish between students' learning styles. In the learning phase of the standard teaching and learning algorithm, individual students learn in a unified way. However, in reality, students with different personalities take different learning styles. For example, some students are introverted and more independent and tend to accumulate experience in learning alone; some are extroverted, good at socializing, and like to gain knowledge in discussing and communicating with others. These two types of students are called "field-independent" and "field-dependent" types in social psychology, respectively. The two concepts of field independence and field dependence originate from the research on perception in literature. Field-independent people tend to refer to themselves when judging objective things and are not easily influenced and interfered with by external factors; field-dependent people tend to refer to the outside to process information and are less independent and easily influenced by the outside world. Due to differences in cognitive styles in learning activities, field-independent and field-dependent students tend to have different learning strategies. The knowledge sources of field-independent students are mainly composed of their knowledge accumulation and discussions with very few classmates; the knowledge of field-dependent students primarily comes from a part of their knowledge and extensive social discussions. In terms of algorithm design, considering that different students are affected by the outside world at different levels, a 0-1 matrix W_i is randomly generated to simulate field-independent (1) and field-dependent (0) students and take other learning methods for them. Strategies: Field-independent students study with the learning strategy of formula (2); field-dependent students study according to the following strategy:

$$t_{i,\text{new}} = \begin{cases} f_{ti} < f_{tj}: f_t * t_i + r_i(t_x - i_x) + r_i(t_{r1} - i_{r2}), \\ f_{ti} > f_{tj}: f_t * t_i + r_i(i_x - t_x) + r_i(t_{r3} - i_{r4}). \end{cases} \quad (6)$$

It can be seen from formula (6) that $j-X$, $3r_X$, and $4r_X$ are three randomly selected students, r_{1i} and r_{2i} are random numbers on $[0, 1]$, and t_f is a scale factor, which is used to reduce self-esteem at the previous moment. The technique works best when the $f t$ value is set to 0.3 after several iterations. When WI is 1, it means that the individual student in X is field-independent, and in the learning stage, it learns according to the learning method of the original algorithm and completely retains its own state at the previous moment. When the value of WI is 0, it means that the individual student t_i is field dependent. Field dependence-field independence is a form of learning control that has been examined. It refers to the degree to which humans are influenced by inner or environmental stimuli when organizing themselves in time and making precise discriminations of their surroundings. Individuals who are field reliable are better at learning social content and doing it in a social context. People who work in the field in an independent manner are less reliant on being given a system to follow and are more self-motivated. In addition to randomly selecting a student to study, it will also exchange experience with other students and absorb some other people's knowledge to improve their own performance. At this time, the student t_i only retains part of their own state. Compared with the middle school stage of the original algorithm, this design weakens the influence of the state at the previous moment, and at the same time enhances the communication between individuals, reduces the probability of the algorithm falling into the local optimum, and maintains the diversity of understanding. Since the other half of the particles completely retain their own state, the convergence speed of the algorithm is also guaranteed. To verify the impact of improving the learning stage on the diversity of students, the program breakpoints are set in the learning stage, and the running results are shown in Figure 3.

Select some test functions in Figure 3, draw the student position map when the algorithm iterates 30 times, and compare the improved algorithm with the original B-TLBO; we can find that the diversity of students has been greatly improved.

3.3. Join the Self-Learning Method Adjustment Stage. Students need to have a precise understanding of their learning situation after two stages of learning through teacher teaching and communication with students. Therefore, self-assessment and regulation play an indispensable role in efficient learning. Bandura proposed the theory of self-regulation in the social learning theory emphasizing the internal reinforcement process of the individual [21]. Self-regulation includes three primary functions: self-observation, self-judgment, and self-reaction. People observe self-behavior according to social activities' standards, judge the gap between self-behavior and standards, and make positive or negative evaluations of self based on self-assessment.

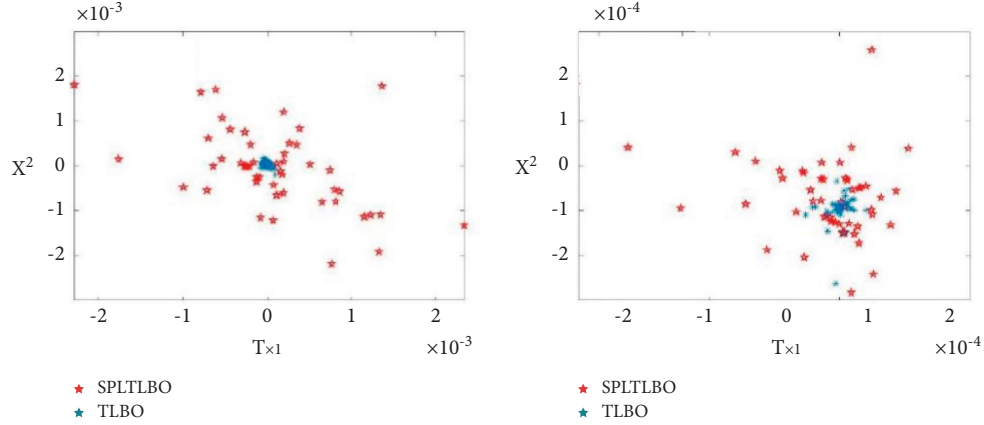


FIGURE 3: Students' diversity.

The individual will have various inner experiences, such as self-satisfaction, self-blame, and criticism, resulting in self-regulation. People can use the effects of self-regulation to adopt more appropriate strategies to achieve their goals. Based on this theory, this paper adds a self-learning method adjustment stage which means earning information or skills via one's personal endeavors rather than through official instruction; after the teaching and learning stages in self-learning, we acquire observation, judgment, and consciousness which are efficient for boosting ones morale. It also creates high accuracy solutions in recognized domains quickly, providing students with a platform for self-reflection and timely adjustment of learning strategies. The specific settings of the self-learning method adjustment stage are as follows: after the teaching stage and the learning stage, the average score of the overall students is calculated, the individuals whose scores are higher than the average score are classified as excellent individuals, and the average score of the beautiful students is calculated. The class is divided into three categories based on the overall average score and the average score of outstanding students:

When the student's grade is higher than the average score of the excellent students, the student is an excellent student, which proves that his learning method is efficient; thus, the student will continue to study with his own learning method, as shown in formula (7):

$$t_{i,\text{new}} = t_i. \quad (7)$$

When a student's grade is lower than the average grade of outstanding students, but higher than the overall average grade of the class, the student is an ordinary student, which proves that his learning method is partially effective, but there is still room for improvement, and the learning method can be fine-tuned to obtain better grades, such as formula (8) shows:

$$t_{i,\text{new}} = t_i + (t_i^{\min} + t_i^{\max})(t_{r1} - t_{r2}). \quad (8)$$

Among them, t_{r1} , t_{r2} are (0, 1) random numbers, and t_i^{\min} and t_i^{\max} are the upper and lower bounds of student i , respectively. When the student's grade is lower than the average score of the class, it proves that the learning method

is ineffective, and the learning strategy needs to be changed to a great extent. Here, the reverse learning method is used, as shown in formula (9):

$$t_{i,\text{new}} = (t_i^{\min} + t_i^{\max}) - t_i. \quad (9)$$

The self-learning method adjustment stage enables individuals to make full use of the population information and adopt a better strategy to update. t_i^{\max} and t_i^{\min} represent the sum of the upper and lower bounds of student i , which provides a greater possibility for student i to change. Therefore, a fine-tuning random number t_{r1} is added to equation (8), such that individuals can still maintain diversity while converging. In formula (9), the sum of the upper and lower bounds is used to subtract the value of the previous state of the individual, which completely changes the position of the individual, thereby increasing the efficiency of the algorithm optimization. Putting this process after the learning stage can help individuals discover their own deficiencies in time and make adjustments quickly. It reduces the probability of bad solutions appearing in the iterative process, which helps improve the optimization speed and accuracy of the algorithm. In addition, since the B-TLBO algorithm idea originates from the teaching process of the simulated class, to better simulate the new state of middle school students in the teaching stage, an adaptive student update factor is introduced to expect the algorithm to obtain better results.

4. Experimental Results and Analysis

In this part, the performance evaluation of the algorithm adopts the exact maximum fitness evaluation time to evaluate the optimization accuracy of the algorithm. The proposed theory may be used to enhance the teaching and learning methods in the early stages of learning, as well as to solve more sophisticated optimization issues such as dynamic vehicle route optimization, parameter optimization in numerous domains, and so on. For each test function, the SPB-TLBO, B-TLBO, PSO, GA, and IA algorithms are run independently 30 times to obtain the optimal solution, worst solution, mean, and standard deviation, respectively.

4.1. *Datasets.* These archives and repositories include datasets that may be used for research purposes. Read the terms of use carefully to verify that you are using the data according to the standards set out by the data originator or repository.

4.1.1. *Inter-University Consortium for Political and Social Research (ICPSR).* Data from social science research may be found in more than 500,000 digital files made available by ICPSR. Science, history, and gerontology are among the many disciplines represented. Other topics of interest include criminology, ageing, and healthcare issues in the public and in the military and foreign policy. Moreover, included are topics such as early childhood education and ethnic minorities in the United States of America. Please contact the Social Science Data Archive for help with ICPSR data.

4.1.2. *Data Archive on Substance Abuse and Mental Health.* Documentation linked to the collection, analysis, and distribution of behavioral health data is provided by the Substance Abuse & Mental Health Data Archive (SAMHDA). Various data formats, including SAS, SPSS, State, and others, may be downloaded.

4.1.3. *Data Repository for Criminal Justice Research and Analysis.* The preservation, upgrading, and sharing of computerized data resources; research based on archived data; as well as specific courses in quantitative analysis of criminal justice data are all part of the NACJD’s objective to help researchers better understand the field of criminal justice.

4.1.4. *Odum Institute’s Data Verse.* Researchers may use the Odum Institute’s data management, archiving, and preservation services. Machine-readable data accumulated over more than a century may be found at the Institute. Datasets may be browsed and searched.

Based on the above four datasets, we are going to see how well the method works in this paper. This paper looks at the standard accuracy rate (P), the recall rate (R), and the microaverage $F1$ as indicators of how well the model does; the formula for this is:

$$\text{Precision } (p) = \frac{\text{true Positive}}{\text{True Positive} + \text{false Positive}},$$

$$\text{Recall } (R) = \frac{\text{true Positive}}{\text{True positive} + \text{False negative}}, \quad (10)$$

$$F1 = \frac{2PR}{P + R}$$

In this paper, we evaluate four data on the entire evaluating matrix, as shown in Tables 1–4.

Actual cases: This stands for the number of attributes predicted by the model to be positive and the real is also

TABLE 1: Accuracy of physiological prediction.

Methods	Dataset			
	ICPSR	SAMHDA	NACJD	ODUM
SPTLBO	75.56	77.26	78.52	79.52
B-TLBO	81.56	85.05	86.35	89.64
PSO	79.52	76.52	74.23	76.25
GA	71.25	74.52	75.24	77.25

TABLE 2: F1-Score of physiological prediction.

Methods	Dataset			
	ICPSR	SAMHDA	NACJD	ODUM
SPTLBO	65.23	63.45	68.45	69.65
B-TLBO	74.52	81.56	78.56	79.52
PSO	64.52	66.85	69.45	70.56
GA	61.56	66.45	65.45	68.52

TABLE 3: Recall of physiological prediction.

Methods	Dataset			
	ICPSR	SAMHDA	NACJD	ODUM
SPTLBO	66.45	64.23	67.52	66.12
B-TLBO	75.62	80.23	79.52	78.52
PSO	61.25	64.25	67.52	69.23
GA	60.23	61.35	65.23	59.52

TABLE 4: Precision of physiological prediction.

Methods	Dataset			
	ICPSR	SAMHDA	NACJD	ODUM
SPTLBO	59.52	60.23	64.56	66.45
B-TLBO	70.23	75.56	77.52	76.52
PSO	60.41	62.23	67.45	64.12
GA	56.23	56.23	60.49	61.85

positive; TP stands for actual case, which means the number of attributes predicted by the model to be positive and real is also positive; fake positives: The number of attributes the model predicts to be both positive and negative. False negatives: This is the number of attributes the model predicts to be both positive and negative. FP stands for “false positives” [22].

The proposed blockchain-based B-TLBO methods acquire a maximum 89.64% accuracy over the ODUM data set, whereas other methods acquire a maximum 79.52% accuracy, i.e., gain by SPTLBO as shown in Figure 4.

The proposed blockchain based B-TLBO methods acquire a maximum 79.52% $F1$ -score over the ODUM dataset, whereas other methods acquire a maximum 69.65% $F1$ -score, i.e., gain by SPTLBO as shown in Figure 5.

The proposed blockchain-based B-TLBO methods acquire a maximum 78.52% recall over the ODUM dataset, whereas other methods acquire a maximum 66.12% recall, i.e., gain by SPTLBO as shown in Figure 6. The proposed

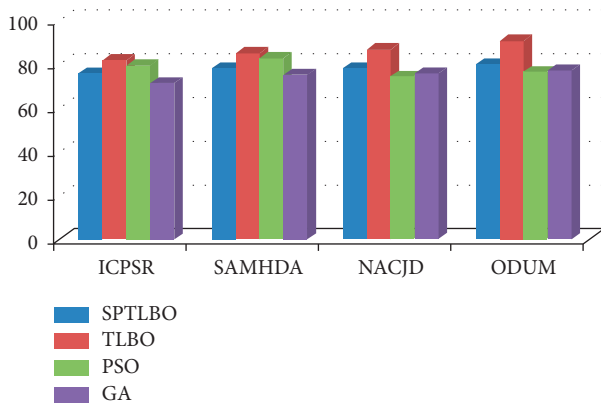


FIGURE 4: Accuracy of physiological prediction.

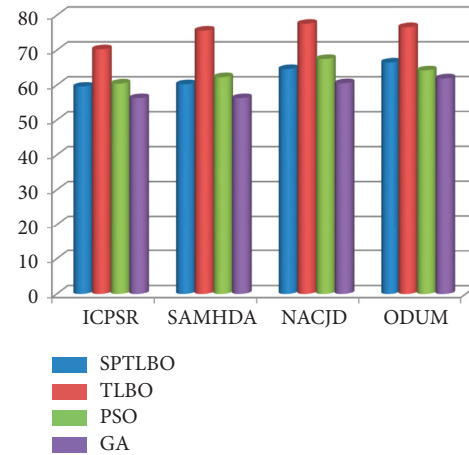


FIGURE 7: Precision of physiological prediction.

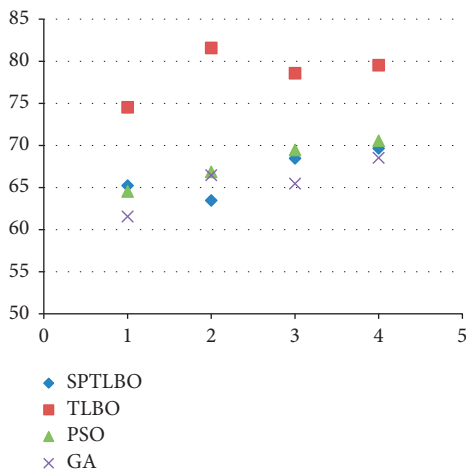


FIGURE 5: F1-score of physiological prediction.

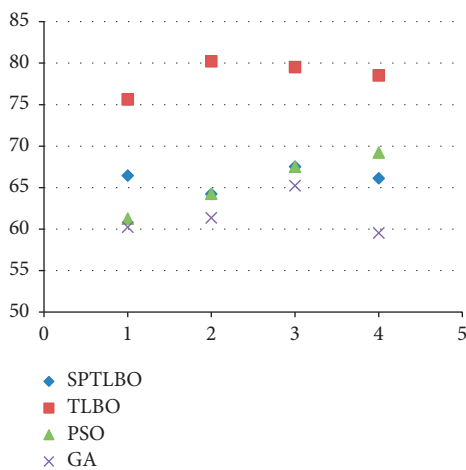


FIGURE 6: Recall of physiological prediction.

blockchain-based TLBO methods acquire a maximum 76.52% precision over the ODUM dataset, whereas other methods acquire a maximum 66.45% precision, i.e., gain by SPTLBO as shown in Figure 7.

5. Conclusion

In the teaching stage, the “expectation effect” theory in social psychology is introduced to simulate the phenomenon that teachers have higher expectations for outstanding students such that individuals with better fitness values can move closer to the optimal individual faster; in the learning stage, the theory of field dependence simulates the differences in the way students with different personalities acquire knowledge, to preserve the diversity of results better and avoid falling into local optimum; after the learning stage, a self-learning method adjustment stage is added to allow individuals to self-rank by adopting different strategies for learning, thereby effectively improving the optimization accuracy and convergence speed of the algorithm. This research paper, focusing on the defects of low optimization precision and slow convergence speed when solving complex optimization problems of the teaching and learning optimization algorithm, from the perspective of social psychology, combined with the changes of people’s psychological emotions, improved the original teaching and learning optimization algorithm. To verify the algorithm’s performance, 25 test functions are selected for numerical experiments. The results show that, compared with the original B-TLBO, PSO, GA, and IA algorithms, the SPTLBO algorithm proposed in this paper has fast convergence speed, high optimization accuracy, and stronger algorithm stability when solving low-dimensional and high-dimensional functions. This research looks at social psychology, bearing in mind human emotions and behaviors, and simulating the impact of human psychological factors on educational outcomes. It can be observed that taking human psychological elements into account while creating algorithms has a positive impact on algorithm performance. The target of TLBOs is to generate high-accuracy solutions as rapidly as feasible in recognized domains. This theory can be used to improve the teaching and learning algorithms in the early stages of learning, as well as to tackle more sophisticated optimization problems like dynamic vehicle path optimization, parameter optimization in numerous disciplines, and so on.

Data Availability

The data shall be made available on request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] S. T. Alvi, M. N. Uddin, L. Islam, and S. Ahamed, "From conventional voting to blockchain voting: categorization of different voting mechanisms," in *Proceedings of the 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, IEEE, Dhaka, Bangladesh, June 2020.
- [2] G.-B. Charla, J. Karen, H. Miller, and M. Chun, "The Human-side of Emerging Technologies and Cyber Risk: a case analysis of blockchain across different verticals," in *Proceedings of the 2021 IEEE Technology & Engineering Management Conference - Europe (TEMSCON-EUR)*, IEEE, Dubrovnik, Croatia, July 2021.
- [3] L.-W. Wong, G. W.-H. Tan, V.-H. Lee, K.-B. Ooi, and A. Sohal, "Psychological and system-related barriers to adopting blockchain for operations management: an artificial neural network approach," *IEEE Transactions on Engineering Management*, pp. 1–15, 2022.
- [4] K. Xin, S. Zhang, X. Wu, and W. Cai, "Reciprocal crowdsourcing: building cooperative game worlds on blockchain," in *Proceedings of the 2020 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, Las Vegas, NV, USA, March 2020.
- [5] Y. Kano and T. Nakajima, "An alternative approach to blockchain mining work for making blockchain technologies fit to ubiquitous and mobile computing environments," in *Proceedings of the 2017 10th International Conference on Mobile Computing and Ubiquitous Network (ICMU)*, IEEE, Toyama, Japan, October 2017.
- [6] X. Zhao, "Teaching-learning based optimization with cross-over operation," in *Proceedings of the 27th Chinese Control and Decision Conference (2015 CCDC)*, IEEE, Qingdao, China, June 2015.
- [7] X. Zhao, "Improved teaching-learning based optimization for global optimization problems," in *Proceedings of the 2015 34th Chinese Control Conference (CCC)*, IEEE, Hangzhou, China, March 2015.
- [8] D. Maity, S. Ghosal, S. Banerjee, and C. K. Chanda, "Bare bones teaching learning based optimization for combined economic emission load dispatch problem," in *Proceedings of the 3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences (EEECOS 2016)*, IEEE, Tadepalligudem, India, June 2016.
- [9] S. Tuo, "Modified teaching-learning-based optimization algorithm," in *Proceedings of the 32nd Chinese Control Conference*, pp. 7976–7981, Xi'an, China, June 2013.
- [10] H. Ouyang, Q. Wang, and X. Kong, "Modified teaching-learning based optimization for 0–1 knapsack optimization problems," in *Proceedings of the 2017 29th Chinese Control And Decision Conference (CCDC)*, IEEE, Chongqing, China, May 2017.
- [11] Q. Yao, M. Shabaz, T. K. Lohani, M. Wasim Bhatt, G. S. Panesar, and R. K. Singh, "3D modelling and visualization for vision-based vibration signal processing and measurement," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 541–553, 2021.
- [12] A. Raj and C. Venkaiah, "Optimal PMU placement by teaching-learning based optimization algorithm," in *Proceedings of the 2015 39th National Systems Conference (NSC)*, IEEE, Greater Noida, India, December 2015.
- [13] B. Crawford, R. Soto, F. A. Leiva, F. Johnson, and F. Paredes, "The set covering problem solved by the binary teaching-learning-based optimization algorithm," in *Proceedings of the 2015 10th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, Aveiro, Portugal, June 2015.
- [14] A. Dhandhia and V. Pandya, "Binary classification of static security assessment using teaching learning based optimization enhanced support vector machine," in *Proceedings of the 2019 IEEE 16th India Council International Conference (INDICON)*, IEEE, Aveiro, Portugal, December 2019.
- [15] J. Godara, I. Batra, R. Aron, and M. Shabaz, "Ensemble classification approach for sarcasm detection," in *Behavioural Neurology*, H. Lin, Ed., vol. 2021, Article ID 9731519, 13 pages, 2021.
- [16] C.-H. Chen, "Group leader dominated teaching-learning based optimization," in *Proceedings of the 2013 International Conference on Parallel and Distributed Computing, Applications and Technologies*, IEEE, Taipei, Taiwan, September 2013.
- [17] A. Tiwari, V. Dhiman, M. A. M. Iesa, H. Alsarhan, A. Mehbodniya, and M. Shabaz, "Patient behavioral analysis with smart healthcare and IoT," in *Behavioural Neurology*, H. Lin, Ed., vol. 2021, Article ID 4028761, 9 pages, 2021.
- [18] V. Chaudhary, R. Yadav, and R. Panwar, "Design and analysis of a 5G electromagnetic shielding structure using teaching learning based optimization," in *Proceedings of the 2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, IEEE, Chennai, India, December 2020.
- [19] A. Gupta and L. K. Awasthi, "Peer enterprises: possibilities, challenges and some ideas towards their realization," in *Proceedings of the On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*, pp. 1011–1020, Vilamoura, Portugal, June 2007.
- [20] H. K. Kwan and M. Zhang, "Minimax design of linear phase FIR Hilbert transformer using teaching-learning-based optimization," in *Proceedings of the 2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*, IEEE, Yangzhou, China, October 2016.
- [21] C. Sharma, A. Bagga, R. Sobti, M. Shabaz, and R. Amin, "A robust image encrypted watermarking technique for neurodegenerative disorder diagnosis and its applications," in *Computational and Mathematical Methods in Medicine*, D. Koundal, Ed., vol. 2021, Article ID 8081276, 14 pages, 2021.
- [22] S. R. Qureshi and A. Gupta, "Towards efficient big data and data analytics: a review," in *Proceedings of the 2014 Conference on IT in Business, Industry and Government (CSIBIG)*, IEEE, Indore, India, March 2014.