

Healthcare of Things and Big Data for Healthcare Engineering

Guest Editors: Shah Nazir, Iván García-Magariño, Sara Shahzad, and Shaukat Ali





Healthcare of Things and Big Data for Healthcare Engineering

Journal of Healthcare Engineering

Healthcare of Things and Big Data for Healthcare Engineering

Guest Editors: Shah Nazir, Iván García-Magariño,
Sara Shahzad, and Shaukat Ali



Copyright © 2021 Hindawi Limited. All rights reserved.

This is a special issue published in "Journal of Healthcare Engineering." 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.

Associate Editors

Xiao-Jun Chen , China
Feng-Huei Lin , Taiwan
Maria Lindén, Sweden

Academic Editors









Cherif Adnen, Tunisia
Saverio Affatato , Italy
Óscar Belmonte Fernández, Spain
Sweta Bhattacharya , India
Prabadevi Boopathy , India
Weiwei Cai, USA
Gin-Shin Chen , Taiwan
Hongwei Chen, USA
Daniel H.K. Chow, Hong Kong
Gianluca Ciardelli , Italy
Olawande Daramola, South Africa
Elena De Momi, Italy
Costantino Del Gaudio , Italy
Ayush Dogra , India
Luobing Dong, China
Daniel Espino , United Kingdom
Sadiq Fareed , China
Mostafa Fatemi, USA
Jesus Favela , Mexico
Jesus Fontecha , Spain
Agostino Forestiero , Italy
Jean-Luc Gennisson, France
Badicu Georgian , Romania
Mehdi Gheisari , China
Luca Giancardo , USA
Antonio Gloria , Italy
Kheng Lim Goh , Singapore
Carlos Gómez , Spain
Philippe Gorce, France
Vincenzo Guarino , Italy
Muhammet Gul, Turkey
Valentina Hartwig , Italy
David Hewson , United Kingdom
Yan Chai Hum, Malaysia
Ernesto Iadanza , Italy
Cosimo Ieracitano, Italy

Giovanni Improta , Italy
Norio Iriguchi , Japan
Mihajlo Jakovljevic , Japan
Rutvij Jhaveri, India
Yizhang Jiang , China
Zhongwei Jiang , Japan
Rajesh Kaluri , India
Venkatachalam Kandasamy , Czech Republic
Pushpendu Kar , India
Rashed Karim , United Kingdom
Pasi A. Karjalainen , Finland
John S. Katsanis, Greece
Smith Khare , United Kingdom
Terry K.K. Koo , USA
Srinivas Koppu, India
Jui-Yang Lai , Taiwan
Kuruva Lakshmanna , India
Xiang Li, USA
Lun-De Liao, Singapore
Qiu-Hua Lin , China
Aiping Liu , China
Zufu Lu , Australia
Basem M. ElHalawany , Egypt
Praveen Kumar Reddy Maddikunta , India
Ilias Maglogiannis, Greece
Saverio Maietta , Italy
M.Sabarimalai Manikandan, India
Mehran Moazen , United Kingdom
Senthilkumar Mohan, India
Sanjay Mohapatra, India
Rafael Morales , Spain
Mehrbakhsh Nilashi , Malaysia
Sharnil Pandya, India
Jialin Peng , China
Vincenzo Positano , Italy
Saeed Mian Qaisar , Saudi Arabia
Alessandro Ramalli , Italy
Alessandro Reali , Italy
Vito Ricotta, Italy
Jose Joaquin Rieta , Spain
Emanuele Rizzuto , Italy




Dinesh Rokaya, Thailand
Sébastien Roth, France
Simo Saarakkala , Finland
Mangal Sain , Republic of Korea
Nadeem Sarwar, Pakistan
Emiliano Schena , Italy
Prof. Asadullah Shaikh, Saudi Arabia
Jiann-Shing Shieh , Taiwan
Tiago H. Silva , Portugal
Sharan Srinivas , USA
Kathiravan Srinivasan , India
Neelakandan Subramani, India
Le Sun, China
Fabrizio Taffoni , Italy
Jinshan Tang, USA
Ioannis G. Tollis, Greece
Ikram Ud Din, Pakistan
Sathishkumar V E , Republic of Korea
Cesare F. Valenti , Italy
Qiang Wang, China
Uche Wejinya, USA
Yuxiang Wu , China
Ying Yang , United Kingdom
Elisabetta Zanetti , Italy
Haihong Zhang, Singapore
Ping Zhou , USA

Contents






Deep Learning Approach for Discovery of In Silico Drugs for Combating COVID-19

Nishant Jha , Deepak Prashar , Mamoon Rashid , Mohammad Shafiq , Razaullah Khan , Catalin I. Pruncu , Shams Tabrez Siddiqui , and M. Saravana Kumar 
Research Article (13 pages), Article ID 6668985, Volume 2021 (2021)

Multiconstraint-Aware Routing Mechanism for Wireless Body Sensor Networks

Javed Iqbal Bangash , Abdul Waheed Khan, Asfandyar Khan, Atif Khan , M. Irfan Uddin, and Qiaozhi Hua 
Research Article (15 pages), Article ID 5560809, Volume 2021 (2021)


Software Defect Prediction for Healthcare Big Data: An Empirical Evaluation of Machine Learning Techniques

Bilal Khan, Rashid Naseem , Muhammad Arif Shah , Karzan Wakil, Atif Khan , M. Irfan Uddin , and Marwan Mahmoud 
Research Article (16 pages), Article ID 8899263, Volume 2021 (2021)




Research on Human Sports Rehabilitation Design Based on Object-Oriented Technology

Dandan Cao, Junyan Wang, and Naihong Liu 
Research Article (9 pages), Article ID 6626957, Volume 2021 (2021)


Human Gait Analysis and Prediction Using the Levenberg-Marquardt Method

Abdullah Alharbi, Kamran Eqbal, Sultan Ahmad , Haseeb Ur Rahman, and Hashem Alyami
Research Article (11 pages), Article ID 5541255, Volume 2021 (2021)



Big Data, Extracting Insights, Comprehension, and Analytics in Cardiology: An Overview

Hui Xiao , Sikandar Ali , Zhen Zhang, Muhammad Shahzad Sarfraz, Fang Zhang, and Mohammad Faisal 
Review Article (14 pages), Article ID 6635463, Volume 2021 (2021)


Segmentation and Classification of Heart Angiographic Images Using Machine Learning Techniques

Abdullah, Muhammad Hameed Siddiqi, Yousef Salamah Alhwaiti, Ibrahim Alrashdi, Amjad Ali, and Mohammad Faisal 
Research Article (9 pages), Article ID 6666458, Volume 2021 (2021)





Augmentation in Healthcare: Augmented Biosignal Using Deep Learning and Tensor Representation

Marwa Ibrahim, Mohammad Wedyan , Ryan Alturki , Muazzam A. Khan, and Adel Al-Jumaily 
Research Article (9 pages), Article ID 6624764, Volume 2021 (2021)

Protein-Protein Interaction Analysis through Network Topology (Oral Cancer)




Fazal Wahab Khattak, Yousef Salamah Alhwaiti, Amjad Ali, Mohammad Faisal , and Muhammad Hameed Siddiqi
Research Article (9 pages), Article ID 6623904, Volume 2021 (2021)

An Online-Offline Certificateless Signature Scheme for Internet of Health Things

Muhammad Asghar Khan , Sajjad Ur Rehman , M. Irfan Uddin , Shibli Nisar, Fazal Noor, Ali Alzahrani, and Insaf Ullah 

Research Article (10 pages), Article ID 6654063, Volume 2020 (2020)

Performance Assessment of Classification Algorithms on Early Detection of Liver Syndrome

Rashid Naseem , Bilal Khan, Muhammad Arif Shah, Karzan Wakil, Atif Khan , Wael Alosaimi, M. Irfan Uddin , and Badar Alouffi

Research Article (13 pages), Article ID 6680002, Volume 2020 (2020)

IoMT-Based Automated Detection and Classification of Leukemia Using Deep Learning

Nighat Bibi , Misba Sikandar , Ikram Ud Din , Ahmad Almogren , and Sikandar Ali 

Research Article (12 pages), Article ID 6648574, Volume 2020 (2020)

Big Data-Enabled Analysis of DRGs-Based Payment on Stroke Patients in Jiaozuo, China

Dawei Qiao , Yanru Zhang , Ateeq ur Rehman , and Mohammad R. Khosravi 





Research Article (9 pages), Article ID 6690019, Volume 2020 (2020)

Wireless Sensor Network Applications in Healthcare and Precision Agriculture

Naila Nawaz Malik, Wael Alosaimi, M. Irfan Uddin , Bader Alouffi, and Hashem Alyami


Research Article (9 pages), Article ID 8836613, Volume 2020 (2020)

Future Location Prediction for Emergency Vehicles Using Big Data: A Case Study of Healthcare Engineering

Muhammad Daud Kamal , Ali Tahir , Muhammad Babar Kamal , and M. Asif Naeem 

Research Article (11 pages), Article ID 6641571, Volume 2020 (2020)









Enabling Clustering for Privacy-Aware Data Dissemination Based on Medical Healthcare-IoTs (MH-IoTs) for Wireless Body Area Network

Fasee Ullah, Izhar Ullah, Atif Khan , M. Irfan Uddin , Hashem Alyami, and Wael Alosaimi

Research Article (10 pages), Article ID 8824907, Volume 2020 (2020)

Research Article

Deep Learning Approach for Discovery of In Silico Drugs for Combating COVID-19

Nishant Jha ¹, **Deepak Prashar** ¹, **Mamoon Rashid** ², **Mohammad Shafiq** ³,
Razaullah Khan ⁴, **Catalin I. Pruncu** ^{5,6}, **Shams Tabrez Siddiqui** ⁷,
and **M. Saravana Kumar** ⁸

¹School of Computer Science & Engineering, Lovely Professional University, Phagwara, India

²Department of Computer Engineering, Faculty of Science and Technology, Vishwakarma University, Pune, India

³Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China

⁴Department of Engineering Management, University of Engineering and Applied Sciences, Swat 19060, Pakistan

⁵Design, Manufacturing & Engineering Management, University of Strathclyde, Glasgow G1 1XJ, UK

⁶Mechanical Engineering, Imperial College London, Exhibition Road South Kensington, London, UK

⁷College of Computer Science and Information Technology, Jazan University, Jazan, Saudi Arabia

⁸Department of Mechanical Engineering, Mount Zion College of Engineering and Technology, Pudukkottai, India

Correspondence should be addressed to Catalin I. Pruncu; c.pruncu@imperial.ac.uk

Received 30 December 2020; Accepted 8 July 2021; Published 23 July 2021

Academic Editor: Daniel Espino

Copyright © 2021 Nishant Jha 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.

Early diagnosis of pandemic diseases such as COVID-19 can prove beneficial in dealing with difficult situations and helping radiologists and other experts manage staffing more effectively. The application of deep learning techniques for genetics, microscopy, and drug discovery has created a global impact. It can enhance and speed up the process of medical research and development of vaccines, which is required for pandemics such as COVID-19. However, current drugs such as remdesivir and clinical trials of other chemical compounds have not shown many impressive results. Therefore, it can take more time to provide effective treatment or drugs. In this paper, a deep learning approach based on logistic regression, SVM, Random Forest, and QSAR modeling is suggested. QSAR modeling is done to find the drug targets with protein interaction along with the calculation of binding affinities. Then deep learning models were used for training the molecular descriptor dataset for the robust discovery of drugs and feature extraction for combating COVID-19. Results have shown more significant binding affinities (greater than -18) for many molecules that can be used to block the multiplication of SARS-CoV-2, responsible for COVID-19.

1. Introduction

The first case of COVID-19 was detected in December 2019, and from then, it has overgrown, affecting millions of people around the globe. More than 2 million cases have been confirmed, with over 0.15 million deaths globally [1, 2]. Drug repurposing is defined as discovering and identifying newer applications for existing drugs in the treatment of various diseases [3]. Recent advancements in drug discovery using deep learning have made it possible to speed up identifying and developing new pharmaceuticals [4]. Various drugs, such as Arbidol,

remdesivir, and favipiravir, have been tested to cure COVID-19 patients and many others are in the testing phase [4]. Biomedical researchers are investigating drugs for treating the patients, with an attempt to develop a vaccine for preventing the virus [5]. On the other hand, computer scientists have developed early detection models for COVID-19 from CT scans and X-ray images [5]. These techniques are a subset of deep learning and have been applied successfully in various fields [5]. Over the past few years, a significant increase in the quantity of biomedical data has resulted in the emergence of new technologies such as parallel synthesis and HTS (high-

throughput screening), to mining large-scale chemical data [6]. Since COVID-19 is transmitted from person to person, electronic devices based on artificial intelligence may play a crucial role in preventing the spread of this virus. With the expansion of the role of health epidemiologists, the pervasiveness of electronic health data has also increased [7]. The increasing availability of electronic health data provides a massive opportunity for healthcare to enhance healthcare for both discoveries and practical applications [7]. For training machine learning algorithms, these data can be used to improve their decision-making in terms of disease prediction [7].

As the increase in the number of cases infected by coronavirus rapidly outnumbered the medical services available in hospitals, a significant burden on healthcare systems was imposed [7]. Because of the limited supply of hospital services and the delay in time for diagnostic test results, it is common for health professionals to provide patients with sufficient medical care. However, since the number of cases tested for coronavirus is growing increasingly day by day, testing is not feasible due to time and cost factors [7]. This paper aims at suggesting a technique based on deep learning which would be helpful in rapidly finding the drugs for combating the pandemic. Deep learning is currently an area that is quickly emerging and constantly expanding. To optimize its performance, it programs computers using data. Using the training data or its previous encounters, it learns the parameters to optimize the computer programs. It can also forecast the future using the data. Deep learning also lets us operate the statistics of the data to construct a mathematical model. The main goal of deep learning is that it learns without any human intervention from the feed data, and it automatically learns from the data (experience) provided and gives us the desired output where it searches the data trends/patterns [8]. Deep learning techniques have achieved greater efficiency in various tasks, including drug development, prediction of properties, and drug target forecasting. As drug development is a complex task, the deep learning approach makes this process faster and cheaper.

The challenges with COVID-19 at present make it necessary to look for some alternatives in medicine or drugs to combat the rise of cases due to COVID-19 infection. One of the significant challenges is the processing delay for the finalization of the drugs for vaccine formulation. However, many pharmaceuticals companies have achieved success to some extent after passing through different trials. Hence, predicting the most probable drugs for the vaccination formulation can speed up vaccine formulation and thus save many human lives. Another challenge is that most of the testing for vaccine formulation is done on a clinical basis where all the drug combinations are tried to get the desired selection of drugs. Still, there is less utilization of computational techniques for the same at present. Thus, there is an hour to look after some alternatives using some machine intelligence techniques to provide some solutions with more accuracy and at a faster note.

Based on the above challenges, the main contributions of the paper are as follows:

- (1) Deep learning approach based on logistic regression, SVM, and Random Forest along with QSAR modeling is proposed to discover some drugs for the treatment of COVID-19
- (2) QSAR modeling is done to find the drug targets with protein interaction along with the calculation of binding affinities
- (3) Deep learning models are used for training the molecular descriptors dataset for the robust discovery of drugs and feature extraction for combating COVID-19

The rest of the article is organized as follows. Section 2 deals with the literature reviewed. Section 3 deals with the significance of work. Section 4 deals with the suggested methodology followed by Section 5, dealing with results, and the paper is concluded in Section 6.

2. Literature Review

Artificial intelligence techniques have been utilized in various areas of drug and vaccine development [9]. This utilization and further advancements are essential for immediately discovering a cure for the current pandemic. Many studies have been done previously, and many are ongoing to find a less complex and easy-to-use technique that would speed up the drug discovery process. In [10], the authors have trained a model based on LSTM (long short-term memory) for reading the SMILE fingerprints of a molecule for predicting IC₅₀, binding to RdRp. The authors in [11] have suggested a B5G framework, which supports the diagnosis of COVID-19 through low latency and 5G. Choi et al. [12] proposed the MT-DTI model for predicting the drugs approved by FDA having solid affinities for the ACE2 receptor with TMPRSS2. The authors in [13] have reviewed all state-of-the-art research studies related to medical imaging and deep learning. Deep learning techniques and feature engineering were compared in order to efficiently diagnose COVID-19 from CT images [14]. Various neural network architectures and generative models such as RNN, autoencoders with adversarial learning, and reinforcement learning are suggested for ligand-based drug discovery [15]. Classification performance of DNN on imbalance compound datasets is explored by applying data balancing techniques in [16]. A novel approach for deep docking large numbers of molecular structures accurately is suggested in [17]. The effects of deep learning in drug design and complimentary tools were reviewed [18].

In [19], a systematic review of the application of deep learning techniques for predicting drug response in cancer cell lines has been done. A QSAR model (quantitative structure-activity relationship) is developed [20], which implements deep learning to predict antiplasmodial activity and cytotoxicity of untested compounds for screening malaria. In [21], the authors have built a multitask DNN model and compared the results with a single-task DNN model. In [22], various machine learning and deep learning algorithms used for drug discovery are reviewed, and their applications were discussed. However, various studies

suggest deep learning for drug discovery or detecting COVID-19 lacks proper practical implementation with results. Most studies have just reviewed different deep learning techniques to be used for the development of drugs. This paper will give a practical implementation on various datasets available online with efficient results. Upon analyzing various studies, we found that various studies claim HCS (high content screening) as an efficient technique for screening chemical compounds for discovering drugs. At present, deep learning techniques have been producing faster and efficient results.

The basic idea of the screening process is that the cells are exposed to various compounds, and automated optical microscopy is done to see what happens, creating thriving images of cells. A quantitative and qualitative analysis of the result can be done by using an automated HCS pipeline. HCS branches out from microscopy, and Giuliano et al. first coined the terminology in the 1990s [23]. HCS research can cover several fields, such as discovering drugs that can be defined as a form of cell phenotypic screen. It includes methods of analysis that produce simultaneous readouts of multiple parameters considering cells or cell compounds. In this phase, the screening aspect is an early discovery stage in a series of various steps needed to identify new drugs. It acts as a filter to target potential applicants that can be used for further development. Small molecules classified as a low molecular weight organic compound, e.g., proteins, peptides, or antibodies, can be the substances used for this purpose [24].

3. Significance of the Work

Hospitals are using trial and error techniques for COVID-19 drug discovery [9]. It results in an emergence of virtual screening to discover chemical compounds due to the inefficiency of the lab-based HTS technique (high-throughput screening) [9]. Also, drug discovery and development is a complex and time-consuming process [25]. It is estimated that the preapproval cost of production of new drugs has increased at the rate of 8.5% annually from 802 million USD to 2870 million USD [26, 27]. Finding molecules with the required characteristics is one of the significant challenges in drug discovery. A practical and quality drug needs to be balanced regarding safety and potency against its target and other properties such as ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) and physicochemical properties [25]. This paper aims to increase the speed of discovering new molecules using deep learning, thereby reducing the cost of producing new drugs. Deep learning techniques will help us navigate large chemical spaces to find new chemical compounds [25]. The significance of using deep learning techniques for combating COVID-19 [1] is summarized in Table 1.

4. Suggested Methodology

This section includes a description of the proposed methodology.

4.1. Dataset Preparation and Preprocessing. We have used the combination of the datasets from the sources [29–31]. Each of the datasets contains a set of chemical compounds with respective binding activity to a target protein calculated by $pIC_{50} = -\log_{10}(IC_{50})$ [32]. Preprocessing is done for removing the invalid and replicated compounds. The entries with IC_{50} measurements with filtered out compounds having suspicious measures are depicted by the “DATA VALIDITY COMMENT” column. For repeated records groups, if the standard deviation (SD) of the activity is found more significant than 1 log unit, then these datasets are deleted from the dataset, and a single entry is kept with the median of the activity [32]. Data preprocessing is one of the significant phases in data mining as it helps in achieving data integrity. Before preprocessing, data cleaning needs to be done as raw data contain abnormalities and errors affecting the results [33]. After preprocessing, conversion of SMILES [34] representations to molecular representations takes place. These are open datasets that contain the binding, ADMET, and functional information for various drugs like bioactive compounds [35]. The database containing the datasets has over 5 million bioactivity measurements for over 1 million compounds and over 5000 target proteins [35].

A minor challenge may occur in data mining algorithms due to variation in range and distribution of every variable in the large datasets due to distance measurements; also, these may contain noisy variables, which makes the learning of the algorithms more difficult [33]. These challenges can be handled by min-max normalization where the value of each variable is adjusted in a uniform range of 0 to 1 [33]. It is given in the following equation:

$$Y_{\text{normalised}} = \frac{Y_x - Y_{\text{minimum}}}{Y_{\text{maximum}} - Y_{\text{minimum}}}, \quad (1)$$

where $Y_{\text{normalised}}$ is the normalised value, Y_x is the value of interest, Y_{minimum} is the minimum value, and Y_{maximum} is the maximum value.

Apart from the dataset, the system used for performing the experiments has UBUNTU 20.04 LTS OS installed with 16 GB RAM and Intel Core i7-8700 processor. The language used for building the model is Python 3.7 with NumPy, pandas, TensorFlow, Bunch, tqdm, Matplotlib, scikit-learn, NVIDIA GPU, CUDA 9.0, Pytorch 0.4.1, Mordred, and RDkit. For evaluating the binding affinities, PyRx is used. We have used the regression model and QSAR techniques as regression models help us define relationships between dependent and independent variables and show the strength of the impact of various independent variables on dependent variables. QSAR helps in maintaining the quantitative structural relationships in molecular predictions.

4.2. Model Development and Evaluation Parameters. As mentioned above, developing a QSAR model can help us in defining the relationship between the chemical

TABLE 1: Summary of applications of deep learning for combating COVID-19.

S. no.	Application	Explanation
1	Pandemic tracking [1]	(i) Bidirectional GRU along with attentional techniques are used for analyzing patterns in respiratory images for mass scale screening of COVID-19 (ii) Application of deep learning (DL) techniques for identification of geographical hazards and spreading at the community level
2	Predicting the structure of proteins [2]	(i) CNN, DNN, and deep ResNet architecture are utilized for the identification of characteristics of proteins (ii) Virus-host prediction and early prevention of virus infectivity can be done using DL architectures
3.	Drug discovery [25]	(i) GAN and reinforcement learning techniques should be implemented for discovering the chemical compounds inhibiting COVID-19
4.	Medical imaging[28]	(i) DL architecture should be used for extraction of features and prediction of possible cases of COVID-19 from CT scan or chest X-ray images

structures and their endpoints by using various statistical methods for the construction of predictive models for revealing the origin of bioactivity [36]. Generally, a QSAR model is depicted by the equation of the form $X = m(X) + \text{Err}$ that can be utilized or prediction of endpoints or new compounds in terms of time-consuming and cost approaches. In order to derive the global molecular features for the SMILES, some notations are there [36], which are given in the following equation:

$$\begin{aligned}
 pqrstu &\rightarrow p + q + r + s + t + u(X_m), \\
 pqrstu &\rightarrow pq + qr + rs + st + tu(XX_m), \\
 pqrst &\rightarrow pqr + qrs + rst + stu(XXX_m).
 \end{aligned} \quad (2)$$

Also, these global descriptors are described as follows [36]:

- (1) BOND is defined as the presence or absence of double (=), triple (#), and stereochemical (@) bond in SMILES
- (2) PAIR is defined as the coincidence of I, N, O, P, S, Br, Cl, F, #, @, and =
- (3) NOSP is defined as the presence or absence of P, S, O, and N
- (4) HALO is defined as the presence and absence of halogens

The optimal attributes for the SMILES are calculated by the following equation [36]:

$$\begin{aligned}
 W(X_{\text{epoch}}, \text{Threshold}) &= \sum TW(X_m) + \sum TW(XX_m) \\
 &+ \sum TW(XXX_m) + \sum TW(\text{NOSP}) \\
 &+ \sum TW(\text{BOND}) + \sum TW(\text{HALO}) \\
 &\cdot \sum TW(\text{PAIR}).
 \end{aligned} \quad (3)$$

The chemical endpoints [36] can be given in the following equation:

$$\text{End} = T_0 + T_1 \times W(X_{\text{epoch}}, \text{Threshold}), \quad (4)$$

where T_0 is the intercept and T_1 is the correlation coefficient.

The development of the QSAR model consists of two significant steps: (i) describing the molecular structure and (ii) the multivariate analysis for correlation of molecular descriptors with observable characteristics [33]. Successful development of the model also includes data preprocessing and statistical evaluations. For evaluating the performance of the QSAR model, the statistical method suggested in [33] is used in the following equation:

$$\begin{aligned}
 x^2 &> 0.5, \\
 Y^2 &> 0.6, \\
 \frac{Y^2 - Y_0^2}{Y^2} &< 0.1, \\
 \text{or } \frac{Y^2 - Y_0'^2}{Y^2} &< 0.1, \\
 0.85 &\leq z \leq 1.15, \\
 \text{or } 0.85 &\leq z'' a \leq 1.15,
 \end{aligned} \quad (5)$$

where x^2 is the cross-validated explained variance, Y^2 is the coefficient of determination, Y_0^2 and $Y_0'^2$ are the predicted vs. observed activities and vice versa, respectively, and x^2 is calculated by the following equation:

$$X^2 = \frac{\sum_{j=1}^{\text{training}} (P_j - \hat{P}_j)^2}{\sum_{j=1}^{\text{training}} (P_j - \bar{P})^2}, \quad (6)$$

where P_j are the measured values, \hat{P}_j are the predicted values, and \bar{P} is the mean value of the entire dataset. This equation is also used for the calculation of external x^2 , i.e., the compounds that are not used in the QSAR model development earlier and are given in the following equation:

$$X_{\text{external}}^2 = 1 - \frac{\sum_{j=1}^{\text{training}} (P_j - \hat{P}_j)^2}{\sum_{j=1}^{\text{training}} (P_j - \bar{P}_j)^2}. \quad (7)$$

For measuring the internal chemical diversity [28], let x and y be two molecules having Z_X and Z_Y as their Morgan fingerprints [28]. The number of common fingerprints is

defined as $Z_x \cap Z_y \vee$ and the total number of fingerprints is defined as $Z_x \cup Z_y \vee$. The Tanimoto similarity [28] between x and y is defined in the following equation:

$$S(x, y) = \frac{|Z_x \cap Z_y|}{|Z_x \cup Z_y|}. \quad (8)$$

And the Tanimoto distance [28] is given by

$$S_d(x, y) = 1 - S(x, y). \quad (9)$$

We have used RDKit [28] for the implementation of Tanimoto distance. In earlier studies, the QSAR models were developed for small compounds that used limited quantitative characteristics [32]. Various algorithms were suggested for covering significant features, including hundreds or thousands of molecular descriptors. We have used the OPLRAreg algorithm suggested in [32] to illustrate the flexibility of mathematical modeling and show how the division of characteristics and regions helps enhance the features of OSAR datasets. The OPLRAreg is given in Algorithm 1.

Due to advancements in deep learning techniques, there has been an increase in the use of neural networks in a variety of applications including healthcare [25]. A neural network can be defined as a group of layers consisting of perceptrons called multilayer perceptron (MLP) or simply a neuron [25]. The perceptrons are the main building blocks of a perceptron and consist of three parts, weights, $v = [v_1, v_2, \dots, v_n]$, $v_j \in R$, biases, $b \in R$, and an activation function, $f(n)$ [25]. Let the input vector given to a perceptron be defined as, $x = [x_1, x_2, \dots, x_n]^Q$. Then, the output is given in the following equation:

$$f(vx + b) = f. \quad (10)$$

Both v and x should be in the same direction. Furthermore, for enabling the matrix multiplication, b and x_1 should be appended to the weight and input vector, respectively [25] so that $v = [v_1 v_2 \dots v_n b]$ and $x = [x_1 x_2 \dots x_n 1]^Q$.

And the output is given by

$$f(vx) = f(v_1 x_1 + v_2 x_2 + \dots + v_n x_n + b). \quad (11)$$

Due to an increase in the efficiency of computation, matrix multiplication is required for training larger networks with forward passing and backpropagation for optimizing the network parameters [25]. The different types of classification methods are given in the following sections.

4.2.1. Logistic Regression. Logistic regression is the most used method of modeling for the prediction of risk [37]. A logistic regression model uses a role to render the model range output between zero and one and should therefore be used for classification. The logistic function is defined in [37] as follows:

$$Y(x = 1) = \frac{1}{1 + \exp(-(\alpha r + s))}, \quad (12)$$

where r is the input and α and s are called as model parameters. The output given is the modeled probability of the input belonging to a class [37]. For interpreting the meaning of the weights, rearrange the above equation as follows [37]:

$$\log\log \alpha r + s. \quad (13)$$

$Y(x = 1)/Y(x = 0)$ is called as the odds. The modeling of odds is done through a linear equation [37]. Like most of the ML (machine learning) models, optimization of the parameters is done w.r.t. loss function [37]. Consider a given set of data points $\{(p_j, q_j)\}_p$, where p_j is defined as the input and q_j is the true output. Let \hat{q}_j denote the output of the logistic regressor. Then $\alpha \wedge s$ are selected according to [37] in the following equation:

$$\alpha^*, \quad (14)$$

$$s^* \operatorname{argmin}_{\alpha s}.$$

This is also known as the log-loss function. The problem of minimization is solved iteratively until the convergence of parameters, using a coordinate descent algorithm [37].

4.2.2. Random Forest. Random Forest is an ensemble approach that combines several decision trees to make predictions. More reliable and precise predictions can be made by combining several poor learners. In addition, ensemble techniques decrease variance and are less vulnerable to overfitting [37]. The Random Forest algorithm [38] is given in Algorithm 2.

As a sequence of questions, a decision tree is best defined. The principle is that questions are asked, and new questions are asked based on the responses, thus creating a tree. Data points are identified using the leaf nodes in the tree [37] by following the trajectory of the questions and answers. The tree is designed by determining which question to ask at each node and determined based on the information obtained from each possible query or the degree to which the uncertainty in the dataset [37] is reduced. The uncertainty in the dataset [37] is defined in the following equation:

$$\operatorname{Entropy}(X) = - \sum_{|XzX|} y(X) \log_2 y(X). \quad (15)$$

The information is acquired by knowing the value of certain feature F and is given in the following equation:

$$\operatorname{Gain}(F) = \operatorname{Entropy}(X) - \sum_{z \text{ values}} \frac{|Xz|}{|X|} \operatorname{Entropy}(X_z), \quad (16)$$

where X_z is defined as the subset where the feature F takes z value. Therefore, during the construction of a decision tree, a feature is to decide each node as explained in [37]. Here, the construction is either terminated once the entropy of the subset has reached zero or the tree has reached its maximum depth [37]. Upon evaluation of a sample, the tree's trajectory is decided until the leaf node is reached. An approximate probability can also be given as output by comparing the class sizes found in the leaf node [37].

```

(1) OPLRAreg is evaluated for  $P=1$ //linear regression
 $x \rightarrow$  currenterror
olderror  $\rightarrow \infty$ 
temperror  $\rightarrow \infty$ 
 $Y_{\text{best}} \rightarrow$ 
(2)  $Z \rightarrow \{Y \in y \vee F_{q_1, Y} \neq 0\}$ //choosing implicit features
 $P \rightarrow 2$ 
(3) For  $j \rightarrow 1; j \rightarrow j + 1; j \leq Z$  do//choose the best partition feature in the region
(4) Evaluate OPLRAreg having two regions and partition feature  $y_j$ 
(5) If  $k < \text{temperror}$  then
(6)  $k \rightarrow \text{temperror}$ 
(7)  $Y_{\text{best}} \rightarrow y_j$ 
(8) End if
(9) End for
olderror  $\leftarrow$  currenterror
temperror  $\rightarrow$  currenterror
(10)  $Y^* \leftarrow Y_{\text{best}}$ 
(11) While currenterror  $< (1 - \alpha)\text{olderror}$  do//increase the number of regions
 $P + 1 \leftarrow P$ 
(12) Evaluate OPLRAreg with  $P$  regions and partition feature  $y_j$ 
(13) olderror  $\leftarrow$  currenterror
(14)  $k \rightarrow$  currenterror
(15) End While
(16) Return  $y_j$ , Breakpoints as  $B_q y$ , and regression coefficients as  $F_{q_1, Y}$ 

```

ALGORITHM 1: OPLRAreg algorithm.

```

for  $j = 1$  to  $X$  do
  generation of random samples  $\Phi$ 
  while stopping criteria  $\neq$  true do
    select randomly  $f$  of all features
    training of tree on  $f$ 
  end
   $f_{\text{RF}}(n) = (1/X) \sum_{j=1}^X f_{\text{Tree}}(n \vee \Phi)$ 

```

ALGORITHM 2: Random Forest algorithm.

4.2.3. Support Vector Machine (SVM). The support vector machine (SVM) is an algorithm for classification that involves creating a hyperplane. A set of features is used in order to classify an object. Thus, the hyperplane will lie in p -dimensional space if there are p features [39]. The hyperplane is generated through SVM optimization, which in turn maximizes the distance from the nearest points, also known as support vectors [39]. Let $y_j = [y_{j1}, \dots, y_{jm}]^N$ be an arbitrary observation feature vector in the training set, x_j corresponding label to y_j , with a weight vector $v = [v_1, \dots, v_q]^N$ with $\forall v \vee v^2 = 1$ and T be the threshold. The constraints defined for the classification problem [39] are given in equations (17) to (20):

$$vN y_j + T > 0 \text{ for } \underline{x_j} = +1, \quad (17)$$

$$vN y_j + T < 0 \text{ for } \underline{x_j} = -1. \quad (18)$$

Let $f(y_j) = vN y_j + T$, then the output of the model \hat{x}_j can be given as follows:

$$\hat{x}_j = \begin{cases} 1 & \text{for } f(y_j) \geq 0, \\ 0 & \text{for } f(y_j) < 0. \end{cases} \quad (19)$$

Instead of using $\|v\|^2 = 1$, for margin maximization, the lower bound on the margin along with the optimization problem can be defined for minimization of $\|v\|^2$ [39]. The constraints for the optimization problem can be derived from equations (17) and (18), respectively, [39] as follows:

$$\hat{x}_j vN y_j + T \geq 1. \quad (20)$$

In some of the cases, it is required to implement a soft margin, allowing some points to lie on the wrong side of the hyperplane [39] in order to provide an efficient model. A cost parameter M is introduced, which plays a major role in the assignment of penalties to errors, where $M > 0$ [39]. Then, the minimized objective function [39] is defined as follows:

$$v \vee v^2 + M \sum_j \beta_j, \quad (21)$$

where β_j = slack variable. The constraints to the optimization problems [39] are now modified in the following equation:

$$\underline{x_j} N y_j + T \geq 1 - \beta_j, \quad \beta_j \geq 0. \quad (22)$$

Most of the datasets are not linearly separable. But through a nonlinear transformation into a high-dimensional space, a dataset is more likely to be linearly separable [37]. Therefore, each sample is transformed using a nonlinear function [37] so that

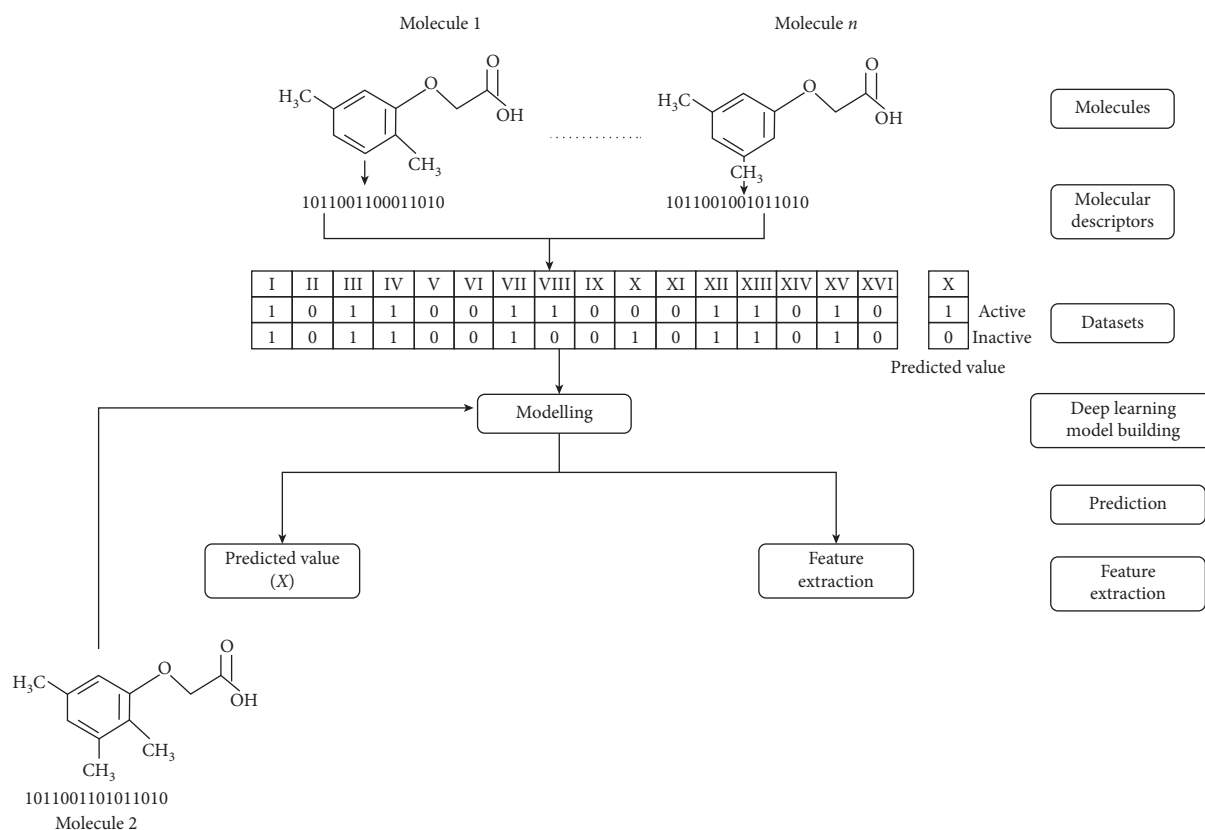


FIGURE 1: Overall workflow of the suggested methodology.

$$f: R^X \longrightarrow R^Y, \quad x > y. \quad (23)$$

And then the problem is considered using $m_j = f(y_j)$ [37]. Furthermore, using Lagrange optimization, the dual problem of maximizing [37] is defined as follows:

$$\sum_j \left[\delta_j - \frac{1}{2} \sum_i \delta_j \delta_i T_j T_i \lambda y_i \right], \quad \lambda = y_j t, \quad (24)$$

subject to the condition

$$\sum_j \delta_j T_j = 0, \quad \delta_j \geq 0 \forall j. \quad (25)$$

The overall structure of the workflow and QSAR modeling [36, 40] is explained in Figure 1. First, we have to select the number of molecules. It can be of any number. Each molecule has its molecular descriptors that describe the molecules' physical and chemical properties that help us differentiate between the molecules. Here, 1 and 0 are the binary descriptors that show the presence/absence of the molecular descriptors. A collection of these descriptors constitutes the dataset. Values of X (active/inactive) show the biological activity we want to predict. This dataset is now used for training the deep learning model, which therefore gives our results. The working of the proposed approach is represented in a flowchart, as depicted in Figure 2.

5. Results

Our goal is to develop a deep learning model to suggest novel and effective drugs for combating SARS-CoV-2 or combating COVID-19. Our regression-based models and Random Forest model were trained on a dataset of approximately 1.5 million drug-like molecules from the data sources [29–31]. The molecules were represented in Simplified Molecular Input Line Entry System (SMILES) format helping our model learn the required features for designing novel drug-like molecules. SMILES are defined as the character strings for representing drug molecules. For example, an atom of carbon can be represented as C, oxygen atom as O, double bond as =, and CO₂ molecule can be represented as C(=O)=O. The maximum length of the string can be taken as 25 [41]. SMILES grammar's learning problem and reproducing it for generating novel small molecules is considered a classification problem [42]. The SMILES strings should be considered a time series, where every symbol is considered a time point. At a given point, the model was trained for predicting the class of the next symbols in the time series.

We will only retrieve the coronavirus proteinase during preprocessing of the bioactivity data that can be reported as IC₅₀ values in nM (nanomolar) units [43]. The data for bioactivity is in the IC₅₀ unit. Compounds with less than 1000 nM values will be considered active, whereas compounds with values greater than 10,000 nM will be considered inactive. As for such values, the intermediate value is

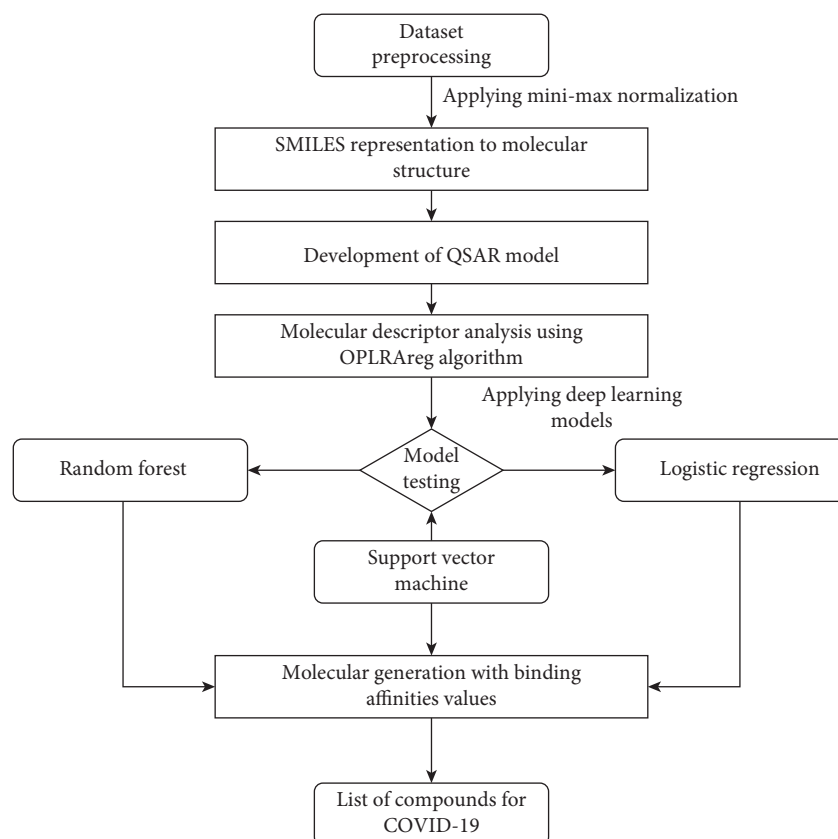


FIGURE 2: Flowchart depicting the complete working of the proposed approach.

TABLE 2: Calculated values of Lipinski descriptors.

MW	LogP	NumH donors	NumH acceptors
281.3	1.90	0.0	5.0
416.5	3.82	0.0	2.0
422.2	2.67	0.0	3.0
294.3	3.63	0.0	4.0
339.3	3.54	0.0	5.0
338.4	3.41	0.0	5.0
297.0	3.45	0.0	3.0
277.2	4.10	0.0	3.0
278.3	3.30	0.0	3.0
282.4	4.11	0.0	2.0

between 1,000 and 10,000 nM [43]. To evaluate the model, Lipinski descriptors [43] were used as given in Table 2.

Upon analyzing the pIC₅₀ values, the actives and inactives have shown a significant difference, which is expected as the values of IC < 1000nM = active, IC₅₀ > 10000nM = inactive, corresponding to pIC₅₀ >6 = active and pIC₅₀ <5 = inactive. Out of the 4 Lipinski descriptors [43], only logP showed no difference between the actives and inactives, while the other three descriptors showed significant differences between the actives and inactives. This can be better understood by Figures 3–7, respectively. A scatter plot has also been drawn to show that the two bioactivity classes (active/inactive) are spanning similar chemical spaces.

Figures 3–7 show that our model can explore the chemical spaces that are further adapted for generating

the smaller molecules specific to a target of interest. The SARS-CoV-2 contains the proteins responsible for the cation and replication of the virus [44]. The functioning of the proteins can be stopped by introducing the drug molecules capable of blocking the protein. Therefore, we have to find the molecules with a high binding affinity to bind the protein effectively. Various drugs/compounds have been tested for finding a high binding relationship, but the results are not very good. We have created novel molecules for binding with the coronavirus, using deep learning and QSAR modeling. After the generation of the molecules, PyRx was used for evaluating the binding affinities. We have also build a regression model using a Random Forest algorithm for acetylcholinesterase inhibitors, as shown in Figure 8. The binding affinities for

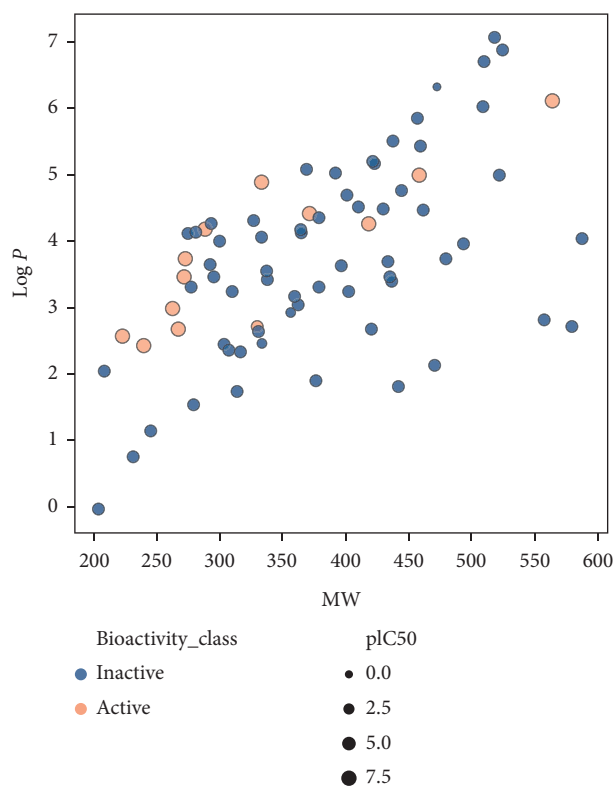


FIGURE 3: Scatter plot of MW vs. logP.

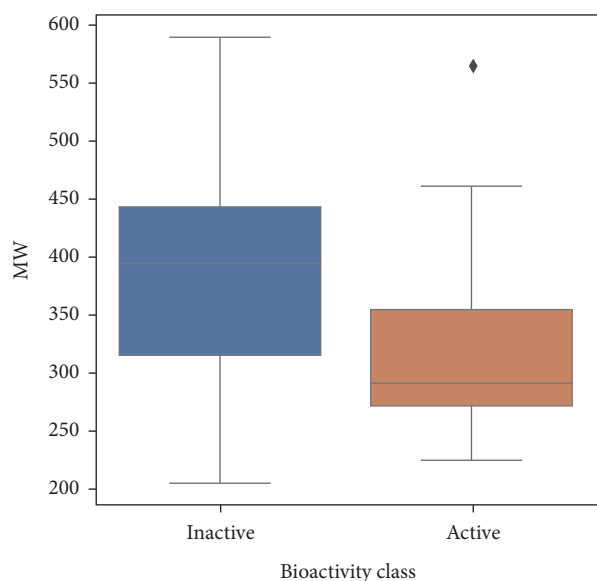


FIGURE 4: Box plot of MW.

leading drugs for other diseases such as HIV inhibitors range from -10 to -11 . Also, the most recent drug remdesivir, which is clinically tested, has the binding affinity of -13 . By convention, the more negative the scores are, the more effective the drugs would be. QSAR modeling, docking analysis, and use of regression model generate a list of bioactive compounds from which top 100

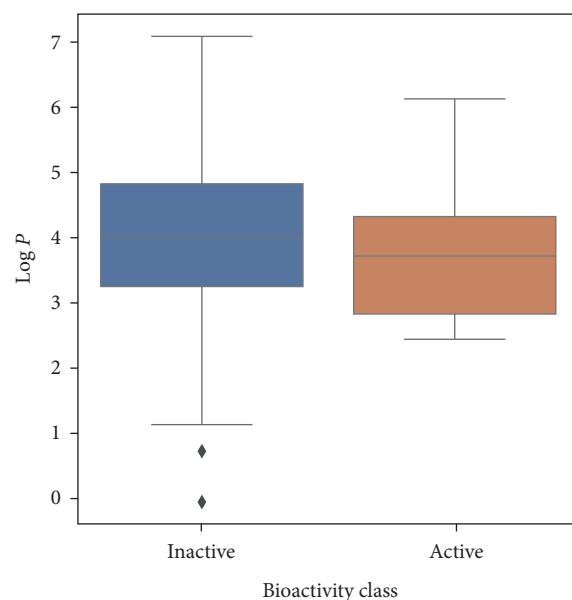


FIGURE 5: Box plot of logP.

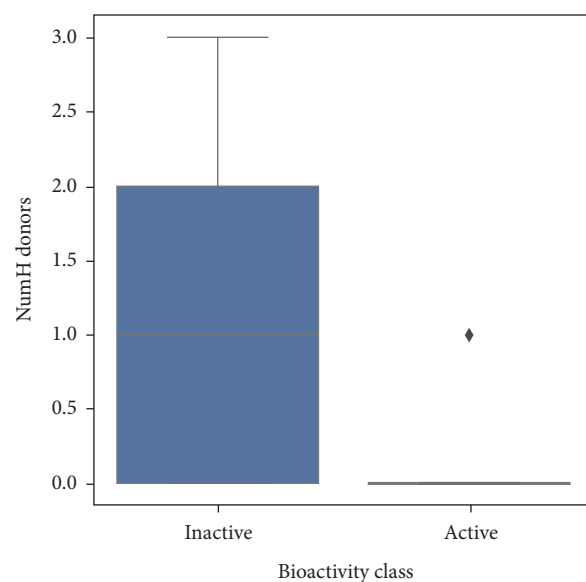


FIGURE 6: Box plot of NumH donors.

compounds were selected, which may have the potential to be effective against SARS-CoV-2. The methodology suggested in this paper is easy to use and can be a possible technique for the discovery of anti-COVID-19 drugs and also shortening the clinical development period required for drug repositioning. Our proposed methodology can give the binding affinity more than the present drugs being tested, making our approach efficient. The proposed list of top 100 chemical structures or molecules generated using our proposed approach through SMILES software is shown in Table 3.

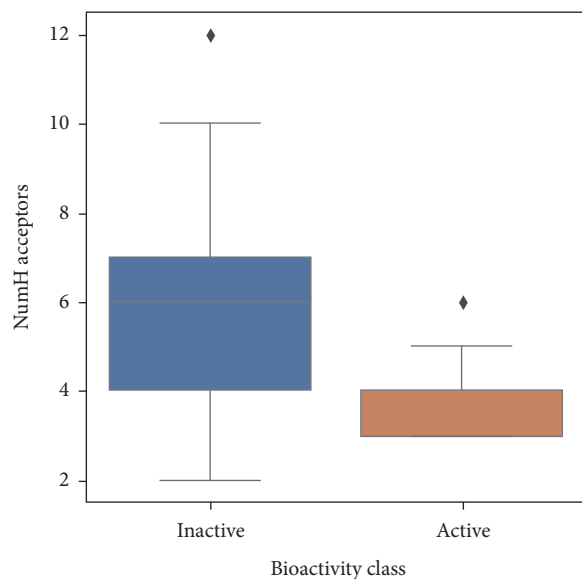


FIGURE 7: Box plot of NumH acceptors.

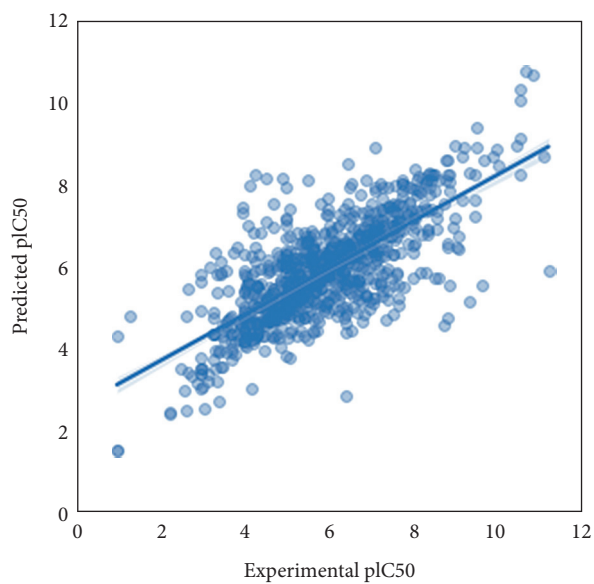


FIGURE 8: Scatter plot for experimental vs. predicted values of pIC50 for regression model developed for acetylcholinesterase inhibitors.

TABLE 3: Top 100 compounds generated using the proposed approach.

Serial no. of the chemical structure generated	SMILES generated chemical structure generated through the proposed approach	Binding affinity value (kcal/mol)
1	<chem>Cc1ccc(C2CNCCN2C)cc1</chem>	-23.1
2	<chem>CCOC(CO)c1ccccc1</chem>	-15.2
3	<chem>CC(=O)Nc1cnn(C)n1</chem>	-24.6
4	<chem>CCC(C)NCc1ncccn1</chem>	-21.5
5	<chem>CC(C)=C1CC(N)C1</chem>	-20.4
6	<chem>CN1CCCc2cc(CON)ccc21</chem>	-18.9
7	<chem>CC12CNCC1CN(CC(N)=O)C2</chem>	-28.9
8	<chem>CCNC(C)C(C)c1enccc1C</chem>	-19.5
9	<chem>CCN(Cc1ccccc1)C(C)CCCNC</chem>	-18.1

TABLE 3: Continued.

Serial no. of the chemical structure generated	SMILES generated chemical structure generated through the proposed approach	Binding affinity value (kcal/mol)
10	<chem>CCC(=O)c1cc(C)ccn1</chem>	-18.3
11	<chem>C=CC(O)c1cc(C)ccn1</chem>	-21.5
12	<chem>C#CCCOc1cnccc1C</chem>	-16.8
13	<chem>Cn1nc2cccc2c1S(N)(=O)=O</chem>	-19.8
14	<chem>Cn1cnn(CC(N)=O)c1=O</chem>	-23.1
15	<chem>CC(NCCSc1cccc1)c1ccncc1</chem>	-21.6
16	<chem>Cc1ccsc1-c1ccc(O)nc1</chem>	-21.9
17	<chem>N#Cc1ncccc1N1CC2CC1CN2</chem>	-19.6
18	<chem>N#Cc1cnccc1SCC(N)=O</chem>	-23.6
19	<chem>N#Cc1ccc(C2NCCCCC2=O)cn1</chem>	-23.5
20	<chem>CC(C)C(C)Sc1ccc(C#N)cn1</chem>	-18.6
21	<chem>Cc1ccnc(C=CCCN)c1</chem>	-24.2
22	<chem>CCOC(CC)C(=O)c1cnccc1C</chem>	-15.9
23	<chem>Cc1ccncc1C(O)CNCC(C)C</chem>	-22.2
24	<chem>CS(=O)(=O)c1ncc(N)cn1</chem>	-21.1
25	<chem>OCC(O)CCSCc1cccc1</chem>	-19.8
26	<chem>COC(=O)CNCc1cc(C)ccn1</chem>	-19.5
27	<chem>CCOC(c1cccc1)C(CC)NN</chem>	-18.0
28	<chem>Cc1ccncc1C(=O)CCCN(C)C</chem>	-19.3
29	<chem>C=CCSCCNc1cc(C)ccn1</chem>	-21.2
30	<chem>CCNC(=S)NNC(=O)Cc1cccc1</chem>	-23.6
31	<chem>OC(CCCc1cccc1)c1ccncc1</chem>	-20.4
32	<chem>CC(=O)CC(C)c1cnccc1C</chem>	-17.3
33	<chem>CN1CCC(O)(c2ccoc2)CC1</chem>	-18.1
34	<chem>Cc1ccnc(NC(=O)C#CCN)c1</chem>	-24.1
35	<chem>N#Cc1cnccc1NCCCO</chem>	-21.0
36	<chem>CCSCc1cncc(C#N)c1</chem>	-19.4
37	<chem>NC1=CCOC1=O</chem>	-16.4
38	<chem>CNC(CSC1CCCC1)Cc1ccncc1</chem>	-18.7
39	<chem>COC(=O)c1ccc(C(C)C=O)cc1</chem>	-14.3
40	<chem>CC(=O)CC(O)c1cnccc1C</chem>	-21.0
41	<chem>CCCNc1cccc1S(N)(=O)=O</chem>	-20.8
42	<chem>N#Cc1ncnc1N1CCCOC1</chem>	-22.0
43	<chem>CCC(CC)Oc1ncccc1C#N</chem>	-16.8
44	<chem>CC(C)(C)C(C)(N)c1cccc1</chem>	-17.0
45	<chem>CN(C)NCc1cccc1</chem>	-20.0
46	<chem>NC12CCCC1CNC2</chem>	-24.3
47	<chem>C(=Cc1cccc1)CNCc1ccncc1</chem>	-23.8
48	<chem>CCNCCNc1ncccc1C#N</chem>	-26.6
49	<chem>CC(C)OCc1ccc(C#N)cn1</chem>	-18.3
50	<chem>NC1Cc2csnc2C1</chem>	-26.5
51	<chem>Cc1ccsc1C1NCCCCC1O</chem>	-21.3
52	<chem>N#CCCNc1cncc1</chem>	-20.8
53	<chem>COC(=O)c1cccc1C#CCO</chem>	-15.5
54	<chem>N#CC1CN(CCN)C(=O)O1</chem>	-19.4
55	<chem>CC(CCO)Nc1ccc(C#N)cn1</chem>	-22.6
56	<chem>NC1CC2(CCNc2=O)C1</chem>	-21.8
57	<chem>C#CC(CO)NCc1cnccc1C</chem>	-22.6
58	<chem>CN1CCCc2cccc(OCC#N)c21</chem>	-16.2
59	<chem>NNC(c1ccncc1)C1CCCC1</chem>	-23.8
60	<chem>C#CCSc1ncccc1</chem>	-17.2
61	<chem>Cc1ccncc1C(C)(N)C(C)C</chem>	-22.6
62	<chem>NS(=O)(=O)c1ccc(SCCO)cc1</chem>	-21.0
63	<chem>Cc1ccnc(CC(=O)C(=O)O)c1</chem>	-18.8
64	<chem>CN1CC2CCN(CC(N)=O)C2C1</chem>	-25.9
65	<chem>O=C=NCc1ccncc1</chem>	-20.9
66	<chem>Cc1esc1C1CC(O)CN1</chem>	-19.2
67	<chem>O=C(CC1CCCCC1)NC1CCNCC1</chem>	-22.4

TABLE 3: Continued.

Serial no. of the chemical structure generated	SMILES generated chemical structure generated through the proposed approach	Binding affinity value (kcal/mol)
68	<chem>CC(O)Cc1cncnc1</chem>	-20.8
69	<chem>CCC(CC)Oc1ccc(C#N)cn1</chem>	-16.1
70	<chem>Cc1ccnc(NN=CC(C)C)c1</chem>	-19.7
71	<chem>COC(CNCCCOc1cccc1)OC</chem>	-12.3
72	<chem>N#Cc1ncccc1C1CCCCC1</chem>	-18.3
73	<chem>NC1COC2COCC12</chem>	-19.9
74	<chem>COC(=O)c1cccc1C=CCCO</chem>	-18.6
75	<chem>CCCC(C)Sc1ncccc1</chem>	-16.9
76	<chem>CC(C)CC(=O)NCCCc1cccc1</chem>	-16.8
77	<chem>CCC(CC#N)Nc1ccc(C#N)cn1</chem>	-21.8
78	<chem>CCCC(C)C(=O)c1cc(C)ccn1</chem>	-19.0
79	<chem>CCOc1cncnc1</chem>	-18.6
80	<chem>NCCCCC(O)c1cccc1</chem>	-21.0
81	<chem>N#CCNc1ccncc1C#N</chem>	-21.6
82	<chem>N#Cc1cncccc1NCC=CCN</chem>	-27.2
83	<chem>CCCOCC(NC)c1cc(C)ccn1</chem>	-18.6
84	<chem>Nc1ccc(S(N)(=O)=O)cc1</chem>	-22.4
85	<chem>c1cnc(OCCNC2CCCCC2)c1</chem>	-20.8
86	<chem>CSCC(C)CNc1ncccc1C#N</chem>	-21.1
87	<chem>CC(N)CNc1cncnc1</chem>	-26.8
88	<chem>CC(C)(N)CNC(=O)Cc1cccc1</chem>	-22.2
89	<chem>NC(CO)c1ccnnc1</chem>	-26.9
90	<chem>CC(=O)OCSc1ncccc1</chem>	-19.3
91	<chem>CN1CCCc2cccc(C=O)c21</chem>	-16.4
92	<chem>CCNc1cc(NCC(C)(C)O)ccn1</chem>	-25.6
93	<chem>CCC(CC)CC(=O)COCc1cccc1</chem>	-13.0
94	<chem>C=CCCC(=O)OCc1cccc1</chem>	-13.9
95	<chem>CN(CCCO)C(=O)Oc1cccc1</chem>	-18.8
96	<chem>CSCCC(=O)c1cncnc1</chem>	-19.6
97	<chem>CC(C)CCCC(O)CCOCc1cccc1</chem>	-13.9
98	<chem>COc1cncnc1C#N</chem>	-18.1
99	<chem>CNc1nc(N)ncc1N</chem>	-28.4
100	<chem>c1ccc(CONCCNc2ccncc2)cc1</chem>	-25.4

6. Conclusion

Drug development is a time-consuming and expensive process. Deep learning has achieved excellent performance in a lot of tasks. Drug discovery is one of the areas that can be benefitted from this. The use of deep learning techniques has made the process of drug development more manageable and cheaper. Deep learning-based models can learn the feature representations based on present drugs that can be used to explore the chemical spaces in search of more drug-like molecules. The available data for automating the processes and better predictions are what deep learning techniques promise for efficient drug discovery. These techniques have proven effective in scanning peptides or detecting COVID-19 from the CT scan or X-ray images. These techniques can speed up the drug development process but require clinical testing for more validation and accuracy [45].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] A. Asraf, M. Z. Islam, M. R. Haque, and M. M. Islam, "Deep learning applications to combat novel coronavirus (COVID-19) pandemic," *SN Computer Science*, vol. 1, no. 6, pp. 363–367, 2020.
- [2] X. Zeng, X. Song, T. Ma et al., "Repurpose open data to discover therapeutics for COVID-19 using deep learning," *Journal of Proteome Research*, vol. 19, no. 11, pp. 4624–4636, 2020.
- [3] S. Pushpakom, F. Iorio, P. A. Eyers et al., "Drug repurposing: progress, challenges and recommendations," *Nature Reviews Drug Discovery*, vol. 18, no. 1, pp. 41–58, 2019.
- [4] K. Arora and A. S. Bist, "Artificial intelligence based drug discovery techniques for covid-19 detection," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 2, no. 2, pp. 120–126, 2020.
- [5] T. T. Nguyen, "Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions," 2020, <https://arxiv.org/abs/2008.07343>.

- [6] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discovery Today*, vol. 23, 2018.
- [7] D. M. Matta and M. K. Saraf, "Prediction of COVID-sing-machine learning techniques," Dissertation, Blekinge Institute of Technology, Karlskrona, Sweden, 2020, <http://urn.kb.se/resolve?urn=urn:nbn:se:bth-20232>.
- [8] X. Yang, S. Nazir, H. U. Khan, M. Shafiq, and N. Mukhtar, "Parallel computing for efficient and intelligent industrial internet of health things: an overview," *Complexity*, vol. 2021, Article ID 6636898, 11 pages, 2021.
- [9] A. Keshavarzi Arshadi, J. Webb, M. Salem et al., "Artificial intelligence for COVID-19 drug discovery and vaccine development," *Frontiers in Artificial Intelligence*, vol. 3, p. 65, 2020.
- [10] S. Patankar, *Deep Learning-Based Computational Drug Discovery to Inhibit the RNA Dependent RNA Polymerase: Application to SARS-CoV and COVID-19*, Science Open, Berlin, Germany, 2020.
- [11] M. A. Rahman, M. S. Hossain, N. A. Alrajeh, and N. Guizani, "B5G and explainable deep learning assisted healthcare vertical at the edge: COVID-19 perspective," *IEEE Network*, vol. 34, no. 4, pp. 98–105, 2020.
- [12] Y. Choi, B. Shin, K. Kang, S. Park, and B. R. Beck, "Target-centered drug repurposing predictions of human angiotensin-converting enzyme 2 (ACE2) and transmembrane protease serine subtype 2 (TMPRSS2) interacting approved drugs for coronavirus disease 2019 (COVID-19) treatment through a drug-target interaction deep learning model," *Viruses*, vol. 12, no. 11, p. 1325, 2020.
- [13] S. Bhattacharya, P. K. R. Maddikunta, Q. V. Pham et al., "Deep learning and medical image processing for coronavirus (COVID-19) pandemic: a survey," *Sustainable cities and society*, vol. 65, Article ID 102589, 2020.
- [14] H. Wang, L. Wang, E. H. Lee et al., "Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 48, 2020.
- [15] I. I. Baskin, "The power of deep learning to ligand-based novel drug discovery," *Expert Opinion on Drug Discovery*, vol. 15, pp. 1–10, 2020.
- [16] S. Korkmaz, "Deep learning-based imbalanced data classification for drug discovery," *Journal of Chemical Information and Modeling*, vol. 60, no. 9, pp. 4180–4190, 2020.
- [17] F. Gentile, V. Agrawal, M. Hsing et al., "Deep docking: a deep learning platform for augmentation of structure based drug discovery," *ACS Central Science*, vol. 6, 2020.
- [18] F. Piroozmand, F. Mohammadipanah, and H. Sajedi, "Spectrum of deep learning algorithms in drug discovery," *Chemical Biology & Drug Design*, vol. 96, no. 3, pp. 886–901, 2020.
- [19] D. Baptista, P. G. Ferreira, and M. Rocha, "Deep learning for drug response prediction in cancer," *Briefings in Bioinformatics*, vol. 22, 2020.
- [20] B. J. Neves, R. C. Braga, V. M. Alves et al., "Deep learning-driven research for drug discovery: tackling malaria," *PLoS Computational Biology*, vol. 16, no. 2, Article ID e1007025, 2020.
- [21] B. Ramsundar, B. Liu, Z. Wu et al., "Is multitask deep learning practical for pharma?" *Journal of Chemical Information and Modeling*, vol. 57, no. 8, pp. 2068–2076, 2017.
- [22] L. Patel, T. Shukla, X. Huang, D. W. Ussery, and S. Wang, "Machine learning methods in drug discovery," *Molecules*, vol. 25, no. 22, p. 5277, 2020.
- [23] K. A. Giuliano, R. L. DeBiasio, R. T. Dunlay et al., "High-content screening: a new approach to easing key bottlenecks in the drug discovery process," *Journal of Biomolecular Screening*, vol. 2, no. 4, pp. 249–259, 1997.
- [24] S. Bergström and O. Ivarsson, *Automation of a Data Analysis Pipeline for High-Content Screening Data*, Linköping University, Linköping, Sweden, 2015.
- [25] E. Sandström, *Molecular Optimization Using Graph-To-Graph Translation*, Umeå University, Umeå, Sweden, 2020.
- [26] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: new estimates of drug development costs," *Journal of Health Economics*, vol. 22, no. 2, pp. 151–185, 2003.
- [27] M. Shafiq, X. Yu, A. K. Bashir, H. N. Chaudhry, and D. Wang, "A machine learning approach for feature selection traffic classification using security analysis," *The Journal of Supercomputing*, vol. 74, no. 10, pp. 4867–4892, 2018.
- [28] M. Benhenda, "ChemGAN challenge for drug discovery: can ai reproduce natural chemical diversity?" 2017, <https://arxiv.org/abs/1708.08227>.
- [29] ChEMBL Database, 2020, https://www.ebi.ac.uk/chembl/g/#search_results/targets/query=coronavirus.
- [30] Molecularsets/Moses, 2020, <https://github.com/molecularsets/moses>.
- [31] <https://datascience.nih.gov/covid-19-open-access-resources%20>.
- [32] J. Cardoso-Silva, G. Papadatos, L. G. Papageorgiou, and S. Tsoka, "Optimal piecewise linear regression algorithm for QSAR modelling," *Molecular informatics*, vol. 38, no. 3, Article ID 1800028, 2019.
- [33] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, and V. Prachayasittikul, "A practical overview of quantitative structure-activity relationship," *EXCLI Journal*, vol. 8, 2009.
- [34] <https://pubchem.ncbi.nlm.nih.gov%20>.
- [35] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, and A. Hersey, "Yvonne light, shaun McGlinchey, david michalovich, bissan Al-lazikani, john P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, no. D1, pp. D1100–D1107, 2012.
- [36] W. Shoombuatong, P. Prathipati, W. Owasirikul et al., "Towards the revival of interpretable QSAR models," in *Advances in QSAR Modeling* Springer, Berlin, Germany, 2017.
- [37] L. Abrahamsson Kwetczar, *Hospital Readmission Risk Prediction Using Machine Learning*, KTH, Stockholm, Sweden, 2020.
- [38] S. Tober, *Tree-based Machine Learning Models with Applications in Insurance Frequency Modelling*, KTH, Stockholm, Sweden, 2020.
- [39] A. Tahir, F. Chen, H. U. Khan et al., "A systematic review on cloud storage mechanisms concerning e-healthcare systems," *Sensors*, vol. 20, no. 18, p. 5392, 2020.
- [40] C. Nantasenamat, "Best practices for constructing reproducible QSAR models," in *Ecotoxicological QSARs*, pp. 55–75, Humana, New York, NY, USA, 2020.
- [41] Wikipedia contributors, "Simplified molecular-input line-entry system," 2020.
- [42] AI Speeds Drug Discovery to Fight COVID-19, 2020, <https://towardsdatascience.com/ai-speeds-drug-discovery-to-fight-covid-19-b853a3f93e82>.
- [43] "Computational drug discovery," 2020, <https://github.com/dataprofessor%20>.
- [44] "COVID-drug discovery for COVID-19," 2020, <https://github.com/AshishKempwad%20>.
- [45] "SARS-CoV-2 drug discovery using genetic algorithm and deep learning," 2020, <https://github.com/Skyquek/fch-drug-discovery%20>.

Research Article

Multiconstraint-Aware Routing Mechanism for Wireless Body Sensor Networks

Javed Iqbal Bangash ¹, Abdul Waheed Khan,² Asfandyar Khan,¹ Atif Khan ³,
M. Irfan Uddin,⁴ and Qiaozhi Hua ⁵

¹Institute of Computer Sciences and IT, The University of Agriculture, Peshawar 25000, Pakistan

²Department of IT and Computer Science, Pak-Austria Fachhochschule-Institute of Applied Sciences and Technology, Haripur, Pakistan

³Department of Computer Science, Islamia College Peshawar, Peshawar, Pakistan

⁴Institute of Computing, Kohat University of Science and Technology, Kohat 26000, Pakistan

⁵Computer School, Hubei University of Arts and Science, Xiangyang 441000, China

Correspondence should be addressed to Qiaozhi Hua; 11722@hbuas.edu.cn

Received 8 January 2021; Revised 1 March 2021; Accepted 21 March 2021; Published 1 April 2021

Academic Editor: Shah Nazir

Copyright © 2021 Javed Iqbal Bangash 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 merger of wireless sensor technologies, pervasive computing, and biomedical engineering has resulted in the emergence of wireless body sensor network (WBSN). WBSNs assist human beings in various monitoring applications such as health-care, entertainment, rehabilitation systems, and sports. Life-critical health-care applications of WBSNs consider both reliability and delay as major Quality of Service (QoS) parameters. In addition to the common limitations and challenges of wireless sensor networks (WSNs), WBSNs pose distinct constraints due to the behavior and chemistry of the human body. The biomedical sensor nodes (BMSNs) adopt multihop communication while reporting the heterogeneous natured physiological parameters to the nearby base station also called local coordinator. Routing in WBSNs becomes a challenging job due to the necessary QoS considerations, overheated in-body BMSNs, and high and dynamic path loss. To the best of our knowledge, none of the existing routing protocols integrate the aforementioned issues in their designs. In this research work, a multiconstraint-aware routing mechanism (modular-based) is proposed which considers the QoS parameters, dynamic and high path loss, and the overheated nodes issue. Two types of network frameworks, with and without relay/forwarder nodes, are being used. The data packets containing physiological parameters of the human body are categorized into delay-constrained, reliability-constrained, critical (both delay- and reliability-constrained), and unconstrained data packets. NS-2 is being used to carry out the simulations of the proposed mechanism. The simulation results reveal that the proposed mechanism has improved the QoS-aware routing for WBSNs by adopting the proposed multiconstraint-aware strategy.

1. Introduction

It can be observed from the history of human beings that getting older was an exception. Now, this trend is changing by the rapid increase in the elderly population living with chronic diseases and thus requires continuous monitoring [1]. According to the World Population Ageing 2019, the worldwide elderly population (65+ aged people) is expected to be increased from 703 million to 1500 million between 2019 and 2050 [2]. Similarly, according to the World Health

Organization (WHO), the world's population of 60+ aged people between 2015 and 2050 will be almost doubled (12%–22%) [3]. The rate of growth in the elderly population is high in developing countries as compared to developed countries [4]. Besides the people suffering from chronic diseases, the patients inside the hospitals also require various levels of monitoring—ranging from a couple of times a day to continuous monitoring. The continuous and on-and-off health monitoring require a huge amount of additional medical and health-care costs [3]. WBSN has

emerged to provide continuous and unsupervised physiological parameters monitoring of the human body. It may be helpful to solve the issues of chronic diseases, increased elderly population, and continuous and on-and-off in-hospital monitoring [5].

In WBSNs, the tiny, lightweight, cost-effective, and low-power BMSNs are implanted inside the human body to capture and observe the physiological parameters [6]. The heterogeneous nature of BMSNs generates various kinds of data packets that require different QoS parameters among which delay and reliability are of key importance [7]. There may be some data packets that require the shortest delay and highest possible reliability and others can allow some losses but need to assure the delivery with the shortest delay. Some data packets should be delivered with no or minimum losses but not within a specific time frame while others containing routine readings of physiological parameters do not have any such constraints.

The electromagnetic waves are absorbed by the human tissues during wireless communication among different in-body BMSNs as they are saline water in nature. The electromagnetic waves absorption and the energy consumed by the implanted BMSNs to carry out their routine operations are the two main reasons that may overheat the in-body BMSNs [8]. These overheated nodes may harm or affect the growth of human tissues [8]. Furthermore, in conventional wireless communication, path loss occurs due to two main reasons: multipath fading and free-space wave propagation. As WBSNs deal with the human body thereby resulting in high and dynamic path loss, therefore the conventional models used for path loss are not directly applicable. The reasons behind this dynamic and high path loss are the wireless communications among the different in-body BMSNs being through the human body and the human body movement [9].

The aforementioned issues of the WBSNs make routing a challenging task [10]. During the last decade, a number of routing protocols have been proposed to address the aforementioned issues that may be categorized based on QoS parameters, postural movement, and temperature rise. It is observed that most of the existing routing protocols are designed to address a single issue while few of them are designed to handle two of these issues. To the best of the authors' information, none of the existing routing protocols integrate the demanding QoS data, the human body movement, in-body wireless communications issues, and the overheating issue of the in-body BMSNs in their designs. In our previous research articles, critical data routing (CDR) [11] focuses on critical data, reliability aware routing (RAR) [12] considers reliability conscious data, and data-centric routing (DCR) [13] works on delay as well as reliability conscious data.

In this paper, the multiconstraint-aware routing mechanism is proposed which offers a more realistic solution that takes into consideration the various traits of the human body. It ensures the provision of the required QoS parameters by classifying the data packets into four categories: delay-constrained, reliability-constrained, critical, and unconstrained. The routing decisions also incorporate

the human body movement and in-body wireless communications issues. To mitigate the in-body overheating issue caused by antenna radiation absorption and energy consumption by nodes' circuitry, the routing mechanism takes into consideration the temperature rise of neighbor nodes during next-hop selection towards the body coordinator. It is a modular-based mechanism where various required tasks are performed by different modules.

The remainder of this paper is structured as follows: Section 2 presents the related literature of the existing routing mechanisms for WBSNs. Section 3 provides the design and development details of the proposed routing mechanism. The performance assessment based on the simulation results of the proposed mechanism is discussed in Section 4. In the end, Section 5 concludes the paper and provides the possible future directions.

2. Related Literature

Due to the numerous applications of WBSNs, they have attained a tremendous focus of the research society. Recently the researchers have proposed various routing algorithms for WBSNs that might be categorized based on QoS parameters, postural movement, and temperature rise. It is observed that most of the existing routing protocols are designed to address a single issue while a few of them are designed to handle two of these issues. The captured physiological parameters demand different QoS parameters and can be classified as critical data (CD), reliability-conscious data (RCD), delay-conscious data (DCD), and non-conscious data (NCD) [7,13–15]. QPRD [14] uses RCD and NCD and QPRR [15] uses DCD and NCD, while ZEQoS [16] uses DCD, RCD, and NCD classes of data. All these routing schemes are designed considering a hospital-based scenario where the physiological parameters are displayed. PARA [17] classifies the captured data into emergency, on-demand, and periodic classes. All these routing schemes take care of the demanding QoS parameter and are not considering the overheated nodes and human body movement issues of WBSNs. The routing mechanisms considering the demanding QoS data have shortcomings in their decision making while selecting a suitable next hop. Some of them such as QPRD, QPRR, and ZEQoS consider the demanding QoS parameter on the node level. Others such as TQMoS use a minimum hop-count strategy in the selection of suitable next hop for all types of data; even the RCD packets can tolerate some delays. Secondly, it uses redundant transmission for CD packets.

Both TMQoS [7] and TLQoS [18] categorize the captured data into four classes and also try to minimize the temperature rise of the in-body BMSNs by avoiding the overheated nodes as forwarder nodes. All the aforementioned QoS-based routing protocols are based on modular approach where every task is performed using a separate module. Besides the demanding QoS, these schemes also take care of the energy consumption being one of the important issues. TMQoS and TLQoS consider both the demanding QoS and overheated nodes issue but overlook the human body movement.

TARA [19], being the first routing protocol of WBSNs addressing the overheated nodes issue, looks at the activities of the neighbors to evaluate the level of the temperature rise. It is based on the withdrawn policy to forward the data using nonoverheated nodes. LTR [20] is another routing scheme where the next-hop selection decision is made by not considering the overheated nodes. On the other hand, LTRT [21] based on Dijkstra's algorithm evaluates the end-to-end path temperature level and follows the path having less temperature level. The routing protocols addressing the overheated nodes issue overlook the demanding QoS of the captured data and human body movement. TTRP [22] is another routing protocol that considers the trust and overheated nodes while selecting the next-hop node. MTR [23] is the only routing protocol that considers both the overheated nodes issue along with the human body movement but it overlooks the high path loss due to the in-body wireless communication and demanding QoS data.

ATEAR [24] is a temperature- and energy-aware routing scheme that uses a block-chain to reduce the temperature rise and energy consumption. CEPRAN [25] uses a cooperative approach for communication to enhance energy efficiency and reliable communication. EHCRP [26] consider several parameters for routing decisions to achieve the desired goal, i.e., energy efficiency. Similarly, the authors in [27] also aim to efficiently use the energy of the sensor nodes, while [28] aims to do the same but considers only critical data (CD).

To cope with the human body movement, different routing protocols such as [9,29–32] have been proposed. All these routing protocols consider the human body movement and its effects. Furthermore, most of them have also worked on energy efficiency being among the key issues of WBSNs. All of them overlooked the demanding QoS data and overheated nodes issue. Moreover, they are also not considering the high path loss due to in-body wireless communication. It is the main reason that the normal path loss models cannot be used with WBSNs. Authors in [33–35] have another interesting concept of using relay nodes along with BMSNs. The BMSNs are used only to capture the required physiological parameters and send them to the nearby relay node while the relay/forwarder nodes are used to forward the received data. Some researchers have used the relay/forwarder nodes to utilize the energy of the BMSNs efficiently while others have used it to address the path loss issue.

To the best of our information, none of the existing routing protocols integrate the demanding QoS data, human body movement, high path loss due to in-body wireless communication, and the overheating issue of the in-body BMSNs in their designs. In our previous research articles, critical data routing (CDR) [11] focuses on critical data, reliability aware routing (RAR) [12] considers reliability conscious data, and data-centric routing (CDR) [13] works on both delay- and reliability-conscious data. All these schemes also consider the human body movement, high path loss due to in-body wireless communication, and the overheating nodes issue.

2.1. Proposed Mechanism. The aforementioned research gap is addressed by the proposed routing mechanism, which considers the demanding QoS data, overheating nodes, human body movement, and wireless communication through the human body.

2.2. Network Frameworks. Two types of network frameworks, with and without relay/forwarder nodes, are being used in the proposed routing mechanism, which is discussed as follows.

Multiconstraint-aware routing mechanism without relay/forwarder nodes (MCARM): the scanned images are in this type of network framework as shown in Figure 1; different in-body BMSNs and on-body local coordinator (LC) can be grouped together using graph theory as in [13]

$$G = (V, E_d), \quad (1)$$

where V is the combination of both S and LC as in (2) [13] and S is the set of N in-body BMSNs as in (3) [13]:

$$V = \{S\}U\{LC\}, \quad (2)$$

$$S = \{s_1, s_2, s_3, \dots, s_n\}. \quad (3)$$

Similarly, E_d denotes the set of M possible in-body wireless communication connections, connecting two BMSNs or a BMSN and LC as in [14]

$$E_d = \{e_1, e_2, e_3, \dots, e_m\}. \quad (4)$$

Multiconstraint-aware routing mechanism with relay/forwarder nodes (MCARMR): in this type of network framework as shown in Figure 2, the job of the BMSNs is to capture only the physiological parameters of the human body while they are forwarded using a different type of nodes called relay/forwarder nodes. The concept of the relay/forwarder nodes is already being used in [33–35].

This type of network framework can be modeled as in (1). V is the combination of S , RN , and LC as in (5) [13]. Similarly, E_d denotes the set of M possible wireless connections, connecting two relay/forwarder nodes, or a relay/forwarder node, and a BMSN same as in (4).

$$V = \{S\}U\{RN\}U\{LC\}. \quad (5)$$

S is the set of N in-body BMSNs same as in (3) and RN is the set of M wearable relay nodes as in [13]

$$RN = \{r_1, r_2, r_3, \dots, r_m\}. \quad (6)$$

2.3. Classification of Captured Data. The captured data packets containing the physiological parameters of the human body are different in terms of the demanding QoS parameters. In this research work, the data packets are classified into four different categories same as in [7,13–15]. These four types of data packets, shown in Figure 3, are discussed below.

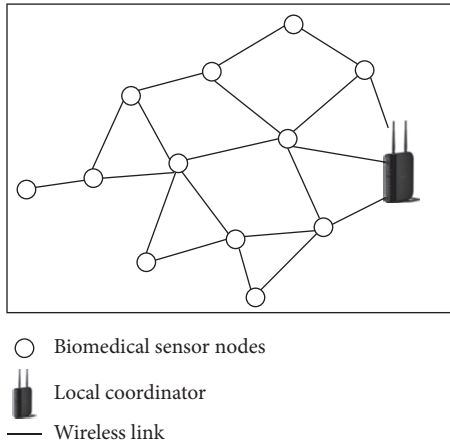


FIGURE 1: Network framework without relay/forwarder nodes.

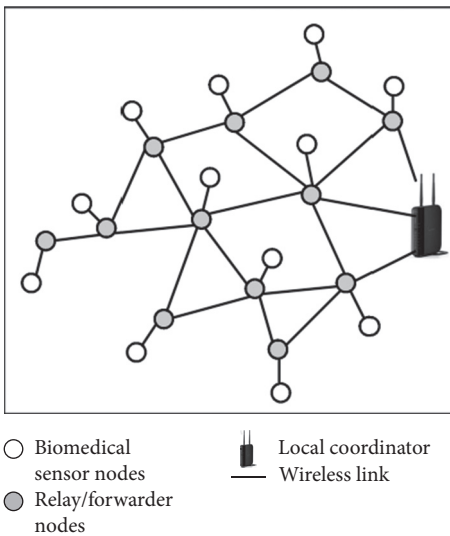


FIGURE 2: Network framework relay/forwarder nodes.

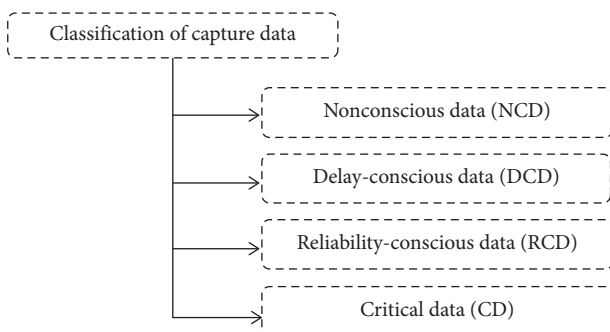


FIGURE 3: Classification of captured data.

2.3.1. *Nonconscious Data (NCD)*. This type of data packet reflects the normal and routine reading and does not enforce any time-deadline and/or reliability constraint. Body temperature, blood pressure, heartbeat, etc. are the examples of NCD packets.

2.3.2. *Delay-Conscious Data (DCD)*. This type of data packets is time-critical imposing delay constraint and reasonable packet losses are acceptable. Video imaging, telemedicine, EMG, and motion sensing are examples of DCD packets.

2.3.3. *Reliability-Conscious Data (RCD)*. This type of data packet needs to be transmitted with minimum or no packet losses and can tolerate some delays. Respiration monitoring and pH-level are examples of RCD packets.

2.3.4. *Critical Data (CD)*. This type of data packet is the most important and reflects the life-critical physiological parameters of the patients. The critical data (CD) packets impose strict delay as well as reliability constraints. This type of data packet is the most important and reflect the life-critical physiological parameters of the patients. The CD packets impose strict constraints in terms of both delay and reliability. ECG and EEG monitoring in a critical situation such as medical surgery, brain stroke, and heart attack, and other physiological parameters that indicate the critical value require real-time and reliable monitoring.

3. Proposed Multiconstraint-Aware Routing Mechanism

This section discusses the proposed routing mechanism that considers the demanding QoS data, overheated nodes, human body movements, and in-body wireless communication. It ensures selecting the best suitable route based on the data packet types by considering end-to-end delay and reliability. It takes care of the high path loss due to in-body wireless communication and dynamic path loss caused by human body movement and tries to avoid the overheated nodes while deciding the next-hop node. It is a cross-layer modular approach, where each module is assigned its duty.

The block diagram shown in Figure 4 consists of Packets Divider (PD), Data Packets Divider (DPD), MAC Receiver (MAC-R), Delay Calculator (DC), Reliability Calculator (RC), Link Quality Calculator (LQC), Temperature Calculator (TC), Routing Unit (RU), QoS-Conscious Next-Hop Selector (QoS-CNHS), QoS-Conscious Queues (QoS-CQs), and MAC Transmitter (MAC-T). The packets either Hello Packets (HPs) or Data Packets (DPs) transmitted by the neighborhood node or LC are received at MAC-R, and it is the job of the PD to divide the Hello and Data Packets using Algorithm 1. The HPs are forwarded towards RU while DPs are forwarded towards DPD. Similarly, it is the job of the MAC-T to transmit the HPs as well as DPs (either generated or received) towards the neighborhood nodes and/or LC. The DPs received from PC, and the DPD has to divide them as critical data (CD), reliability-conscious data (RCD), delay-conscious data (DCD), and nonconscious data (NCD) using Algorithm 2 and forward them towards QoS-ANHS. The other units of the proposed routing mechanism are discussed as follows.

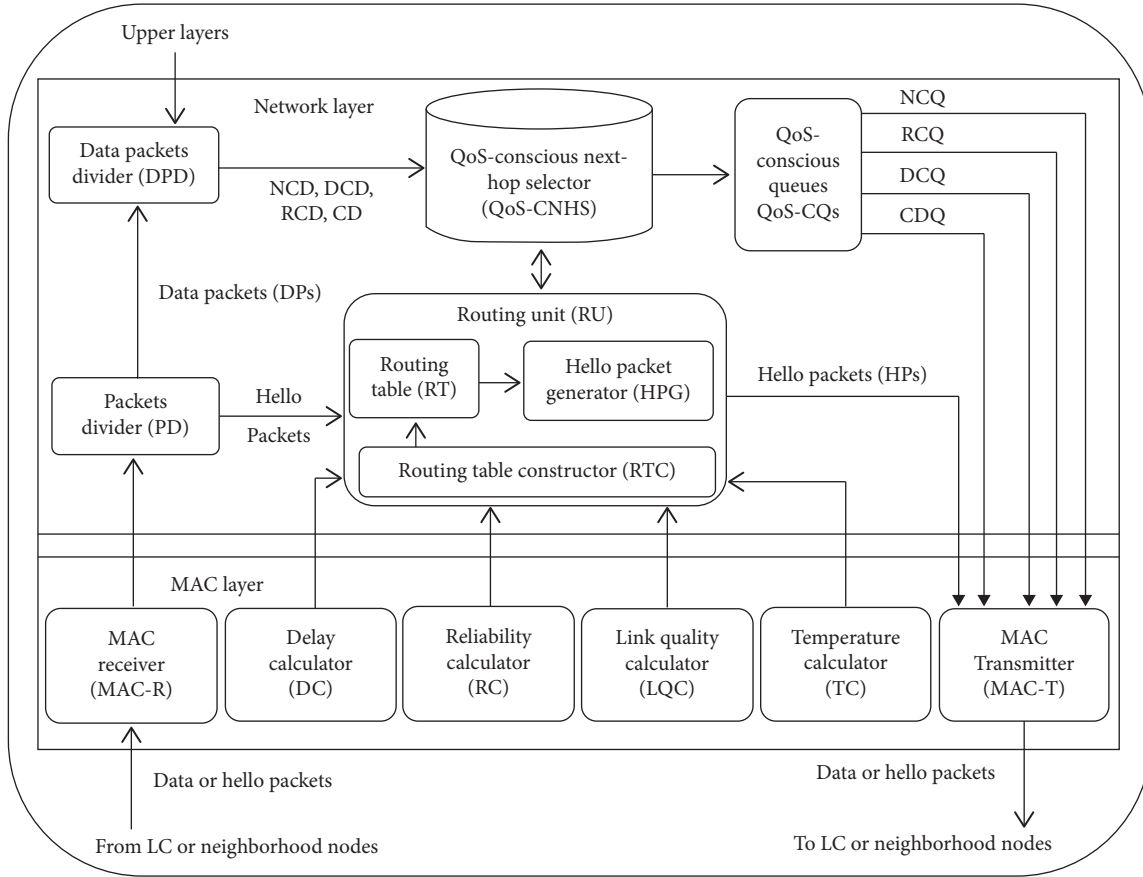


FIGURE 4: Block diagram of proposed multiconstraint-aware routing mechanism.

3.1. Delay Calculator (LC). At each node N_i , it is the job of the DC to calculate the Node Delay ND_{N_i} using (7) [13]. QD_{N_i} is the delay that occurred in queue and $TL_{i,j}$ is the delay that occurred during transmission of the DPs from N_i to N_j using Wireless Link $WL_{i,j}$. The other types of delays, i.e., propagation and processing, are small enough to be ignored.

$$ND_{N_i} = QD_{N_i} + TD_{i,j}, \quad (7)$$

where QD_{N_i} given in (8) [36] is the time that a DP spends in waiting for transmission, where α , the constant factor value, ranges from zero to one and in most cases such as [7,13,14,16] it is equal to 0.2. The queue delay QD_{N_i} occurred once the first delay-conscious or critical data packet is transmitted.

$$QL_{N_i} = \alpha QL_{N_i} + (1 - \alpha) QL_{N_i}. \quad (8)$$

$TL_{i,j}$ given in (9) [16] is the time that a DP spends in waiting at the MAC layer, where NP represents the number of DPs transmitted, DR_{bits} is the generated data rate (bits), and SP_{bits} is the packet size (bits).

$$TL_{i,j} = \left(\frac{1}{DR_{\text{bits}}} \right) x \left(\frac{\sum_{z=1}^{NP} SP_{\text{bits}}(Z)}{NP} \right). \quad (9)$$

3.2. Reliability Calculator (RC). RC is used to calculate the reliability of wireless link $WL_{i,j}$ denoted by $LR_{i,j}$ from N_i to N_j using (10) [37]. β represents the weighting factor with values from zero to one and β equal to 0.4 is being used to simulate the proposed routing mechanism same as in [6,10–12,14,15] and P_{ave} is given in (11) [13], where NP_{succ} is the number of successfully transmitted packets and NP_{total} represents the total transmitted packets.

$$LR_{i,j} = \beta LR_{i,j} + (1 - \beta) \times P_{\text{ave}}, \quad (10)$$

$$P_{\text{ave}} = \frac{NP_{\text{succ}}}{NP_{\text{total}}}. \quad (11)$$

3.3. Link Quality Calculator (LQC). The job of LQC is to calculate the quality of wireless link $WL_{i,j}$ represented by $WLQ_{i,j}$ from N_i to N_j . Equation (12) [38] is being used which is based on a semiempirical formula to calculate the path loss $PL_{i,j}$ in terms of the distance $d_{i,j}$ (the distance of N_i from N_j). The path loss exponent is denoted by n and PL_0 is the reference link quality at distance d_0 .

$$PL_{i,j} = PL_0 + 10n \log \frac{d_{i,j}}{d_0}. \quad (12)$$

```

Inputs: RT and DP
(1)  START
(2)  for DP received at QoS-CNHS do
(3)    for each NN  $\in$  RT do
(4)      if  $WLQ_{i,j} \geq WLQ_{thre}$  then
(5)        List NN into NHNWLQ
(6)      end if
(7)    end for
(8)  if NHNWLQ = NULL then
(9)    drop DP
(10) else if DP  $\in$  CD !! DC  $\in$  DCD then
(11)   call delay-conscious procedure with inputs DP and NHNWLQ
(12) else if DP  $\in$  RCD then
(13)   call reliability-conscious procedure with inputs DP and NHNWLQ
(14) else
(15)   SNH  $\in$  NHNWLQ with minimum  $PT_{i,j,LC}$ 
(16)   forward DP towards NCQ
(17) end if
(18) end for

  Delay-Conscious Procedure
Inputs: NHNWLQ and DP
(19)  for each NN  $\in$  NHNWLQ do
(20)    if  $PD_{i,j}, LC \leq PD_{thre}$  then
(21)      List NN into NHNPD
(22)    end if
(23)  if NHNPD = NULL then
(24)    drop DP
(25)  else if NHNPD = 1 then
(26)    SNH  $\in$  NHNPD
(27)    if CP  $\in$  DCD then
(28)      forward DP towards DCQ
(29)    else
(30) forward DP towards CDQ
(31)    end if
(32)  else if DP  $\in$  DCD hen
(33)    SNH  $\in$  NHNPD with minimum  $PT_{i,j,LC}$ 
(34)    forward DP toward DCQ
(35)  else
(36)    call reliability-conscious procedure with inputs DP and  $NHN_{PD}$ 
(37)  end if
(38) end for

  Reliability-Conscious Procedure
Inputs: NHNWLQ or NHNPD or and DP
(39)  for each NN  $\in$  NHNWLQ !! NN  $\in$  NHNPD do
(40)    if  $PR_{i,j}, LC \geq PR_{thre}$  then
(41)      List NN into NHNPR
(42)    end if
(43)  if NHNPR = NULL then
(44)    SNH  $\in$  NHNWLQ !! SNH  $\in$  NHNPD with maximum  $PR_{i,j,LC}$ 
(45)    if DP  $\in$  RCD then
(46)      forward DP towards RCQ
(47)    else
(48)      forward DP towards CDQ
(49)    end if
(50)  else if NHNPR = 1 then
(51)    SNH  $\in$  NHNPR
(52)    if DP  $\in$  RCD then
(53) forward DP towards RCQ
(54)  else
(55)    forward DP towards CDQ
(56)  end if

```

```

(57)         else
(58)         SNH ε NHNPR with minimum PTi,j,LC
(59)         if DP ε RCD then
(60)             forward DP towards RCQ
(61)         else
(62)             forward DP towards CDQ
(63)         end if
(64)     end if
(65) end for
(66) END

```

ALGORITHM 1: QoS-conscious next-hop selector algorithm.

To accommodate the dynamic human body movements, “Zero-Mean Gaussian Random Variable X_{∂} having Standard Deviation ∂ ” is being used to formulate (13) [13] from

$$PL_{i,j} = PL_0 + 10n \log \frac{d_{i,j}}{d_0} - X_{\partial}. \quad (13)$$

Equation (14) is to calculate the link quality $WLQ_{i,j}$ of wireless link $WL_{i,j}$ from N_i to N_j can be calculated using (14) [38] derived from (13), where P_{trans} is the transmission power and WLQ_{thre} represents the threshold level of the link quality.

$$WLQ_{i,j} = \frac{1}{2} - \frac{1}{2} \operatorname{erf} \left(\frac{-P_{trans} + PL_{i,j} + WLQ_{thre}}{\sqrt{2\pi\partial}} \right). \quad (14)$$

3.4. Temperature Calculator (TC). The task assigned to TC is to calculate the increase in the temperature of any in-body BMSN N_i . The rate at which the electromagnetic waves are absorbed by the human tissues during wireless communication, known as specific absorption rate (SAR), is given in (15) [19], where σ refers to the electric conductivity, the induced electric field is represented by E , and the density is denoted by ρ .

$$SAR = \frac{\sigma|E|^2}{\rho}. \quad (15)$$

Similarly, the in-body BMSNs consume energy to perform the various tasks, which is the second reason that causes an increase in their temperature. Pennes Bioheat formula [39] given in (16) can be used to measure the rate of temperature increase dT/dt due to energy consumption, where the temperature increase (TI) due to thermal conductivity is denoted by $K\Delta^2T$, $b(T - T_b)$ refers to TI caused by blood perfusion, ρSAR represents the TI due to electromagnetic waves absorption, P_c is the TI due to energy consumption of the BMSNs’ circuitry, ρ refers to the mass density, and C_p represents the specific heat of the human tissue. The aforementioned parameters of (16) are assigned the values provided by [35]

$$\frac{dT}{dt} = \frac{K\Delta^2T - b(T - T_b) + \rho SAR + P_c}{\rho C_p}. \quad (16)$$

3.5. Routing Unit (RU). RU is further divided into three subunits, namely, Routing Table (RT), Routing Table Constructor (RTC), and Hello Packets Generator (HPG). The job of RTC is to create and/or update the RT periodically using the data provided by various parameter calculators and neighborhood nodes through Hello Packets (HPs). Once an HP is received from a neighborhood node N_j , the node N_i compared the temperature increase TIN_j to a predefined level known as Temperature Increase Threshold TI_{thre} . The RT is not updated and the HP is dropped if the $TIN_j \geq TI_{thre}$ and the entry of N_j are removed from the RT if any. Based on the received data, path delay $PD_{i,j,LC}$, path reliability $PR_{i,j,LC}$, and path temperature $PT_{i,j,LC}$ from source N_i to the destination (LC) through N_j are calculated using (17)–(19) same as in [13], respectively.

$$PD_{i,j,LC} = PD_{i,j,LC} + ND_{N_i}, \quad (17)$$

$$PR_{i,j,LC} = PR_{i,j,LC} + NR_{N_i}, \quad (18)$$

$$PT_{i,j,LC} = PT_{i,j,LC} + TI_{N_j}. \quad (19)$$

Figure 5 shows the organization of RT for the proposed routing mechanism, containing destination (LC) address and location, neighborhood node N_j address and location, wireless link quality $WLQ_{i,j}$ (between N_i and N_j), path (from N_i to LC using N_j as next-hop) delay, path (from N_i to LC using N_j as next-hop) reliability, and path (from N_i to LC using N_j as next-hop) temperature.

Once the RT is created and/or updated periodically, Hello Packet Generator (HPG) is responsible for constructing the HP based on the available information. The HP is forwarded towards MAC-transmitter which broadcasts it among the neighborhood nodes.

3.6. QoS-Conscious Next Hop Selector (QoS-CNHS). The responsibility of the QoS-CNHS is to choose the suitable next-hop as required by the demanding QoS data packets. Once the DPs are classified as CD, DCD, RCD, and NCD packets in Data Packets Divider (DPD), the proposed Algorithm 1 for QoS-CNHS examines the RT and neighbor nodes (NNs) having $WLQ_{i,j} \geq WLQ_{thre}$ are selected among all neighborhood nodes and placed in NHNWLQ (Next-Hop Neighbors with acceptable wireless link quality) (lines 3–5). If

Destination address (add _{LC})	Destination location (LOC _{LC})	Neighbor node address (NN _{Add})	Neighbor node location (NN _{Loc})	Wireless link quality (WLQ _{i,j})	Path delay (PD _{i,j,LC})	Path reliability (PR _{i,j,LC})	Path temperature (PT _{i,j})
0	150,200	6	85,204	0.93	122	0.95	0.027
0	150,200	9	102,171	0.89	87	0.92	0.042

FIGURE 5: Routing table organization of the proposed routing mechanism.

NHNWLQ is empty, then the DP is dropped (lines 8–9). If it is not empty, then DP is examined for its type. In case the DP is either CD or DCD packet, then DP and NHNWLQ are sent to the delay-conscious procedure (lines 10–11). If DP is RCD packet, then DP and NHNWLQ are sent to the reliability-conscious procedure (lines 12–13). Suitable Next-Hop (SNH) is the NN with minimum $PT_{i,j}$, LC in NHNWLQ for NCD packet, and the DP is forwarded to the NCQ (lines 14–16).

The delay-conscious procedure is responsible for the CD and DCD packets and after receiving DP and NHNWLQ, it looks at NHNWLQ, and NNs with $PD_{i,j}$, $LC \leq PD_{thre}$ are listed into NHNPD (Next-Hop Neighbors with acceptable path delay) (lines 19–22). DP is dropped if NHNPD is empty (lines 23–24) and if there is only one NN in NHNPD, then it is selected as SNH (lines 25–26). DP is sent towards DCQ if it is DCD packets (lines 27–28); otherwise it is sent towards CDQ (lines 29–31). In case of more than one NN in NHNPD, then the NN with minimum $PT_{i,j}$, LC is selected as SNH for DCD packet and DP is sent towards DCQ (lines 32–34), while DP and NHNPD are sent to the reliability-conscious procedure for CD packet (35–36).

The reliability-conscious procedure is called for CD and RCD packets and after receiving DP and NHNWLQ or NHNPD, it looks at the received list, and NNs with $PR_{i,j}$, $LC \geq PR_{thre}$ are recorded into NHNPR (Next-Hop Neighbors with acceptable path reliability) (lines 39–42). SNH is the NN with the highest $PR_{i,j}$, LC if NHNPR is empty (lines 43–45) and RCD packet is sent towards RCQ (lines 46–47) while CD packet is sent towards CDQ (lines 48–39). In case of having only one NN in NHNPR, it is selected as SNH (lines 50–51) and RCD packet is sent towards RCQ (lines 52–53) while CD packet towards CDQ (lines 53–56). In case of more than one NNs in NHNPR then NN with minimum $PT_{i,j}$, LC is selected as SNH (lines 57–58). RCD packet is sent towards RCQ (lines 59–60) while CD packet is sent towards CDQ (lines 61–62). Flowchart for the proposed QoS-CNHS algorithm is given in Figure 6.

QoS-Conscious Queues (QoS-CQs): after selecting the suitable next-hop node as required by the demanding QoS data packets, they are forwarded towards QoS-CQs. Four types of QoS-CQs are being used, where Critical Data Queue (CDQ) is at the highest priority, next is Delay-Conscious Queue (DCQ), then comes Reliability-Conscious Queue (RCQ), and finally Nonconscious Queue (NCQ) is having the lowest priority. The CD packets are placed in CDQ while DCD packets are placed in DCQ until the MAC-transmitter

sends them towards the selected next-hop. Similarly, the RCD and NCD packets are retained in RCQ and NCQ before being transmitted by the MAC-transmitter, respectively. To cope with indefinite waiting, the data packets in low-priority queues are moved into the high-priority queues same as in [7,11–16].

3.7. Simulation and Performance Assessment. In this section, the simulation of the proposed routing mechanism is discussed along with its performance assessment against other related and recent mechanisms.

3.8. Simulation Setup. Network Simulator version 2 (NS2) [35] is used to carry out the simulation and performance evaluation of the proposed routing mechanism for WBSNs same as in [11–13]. It is an open-source, event-driven discrete-time simulator, which is designed to facilitate the research activities of networking and communication. It supports simulating TCP, multihop routing, and multicasting algorithms by having complete models for physical, data-link, and MAC layers.

Two types of network frameworks with and without relay/forwarder nodes (RNs) denoted as MCARM and MCARMR, discussed in Section 3.1, are being used in order to assess the performance of the proposed mechanism with other recent and related mechanisms. Some of the BMSNs are used to generate conscious (either RCD, DCD, or CD) packets and others nonconscious (NCD) packets. The proposed mechanism is implemented in such a way that every BMSN is used to generate all types of data packets discussed in Section 3.2 to get the average results. The performance of the proposed mechanism is assessed against TQMoS [6] and LTRT [20]. TQMoS considers both the demanding QoS data and the temperature increase of the in-body BMSNs. Similarly, LTRT is designed to address the temperature increase issue that uses path temperature while selecting the next-hop node. The proposed mechanism is assessed in terms of average on-time packet delivery ratio for CD and DCD packets, packet loss ratio due to in-body wireless communication and human body movements (path loss), average end-to-end-delay for DCD packets, average packet delivery ratio for CD and RCD packets, maximum temperature increase, and average energy consumed. The simulation results reveal that the proposed mechanism has improved the QoS-aware routing for WBSNs by adopting a

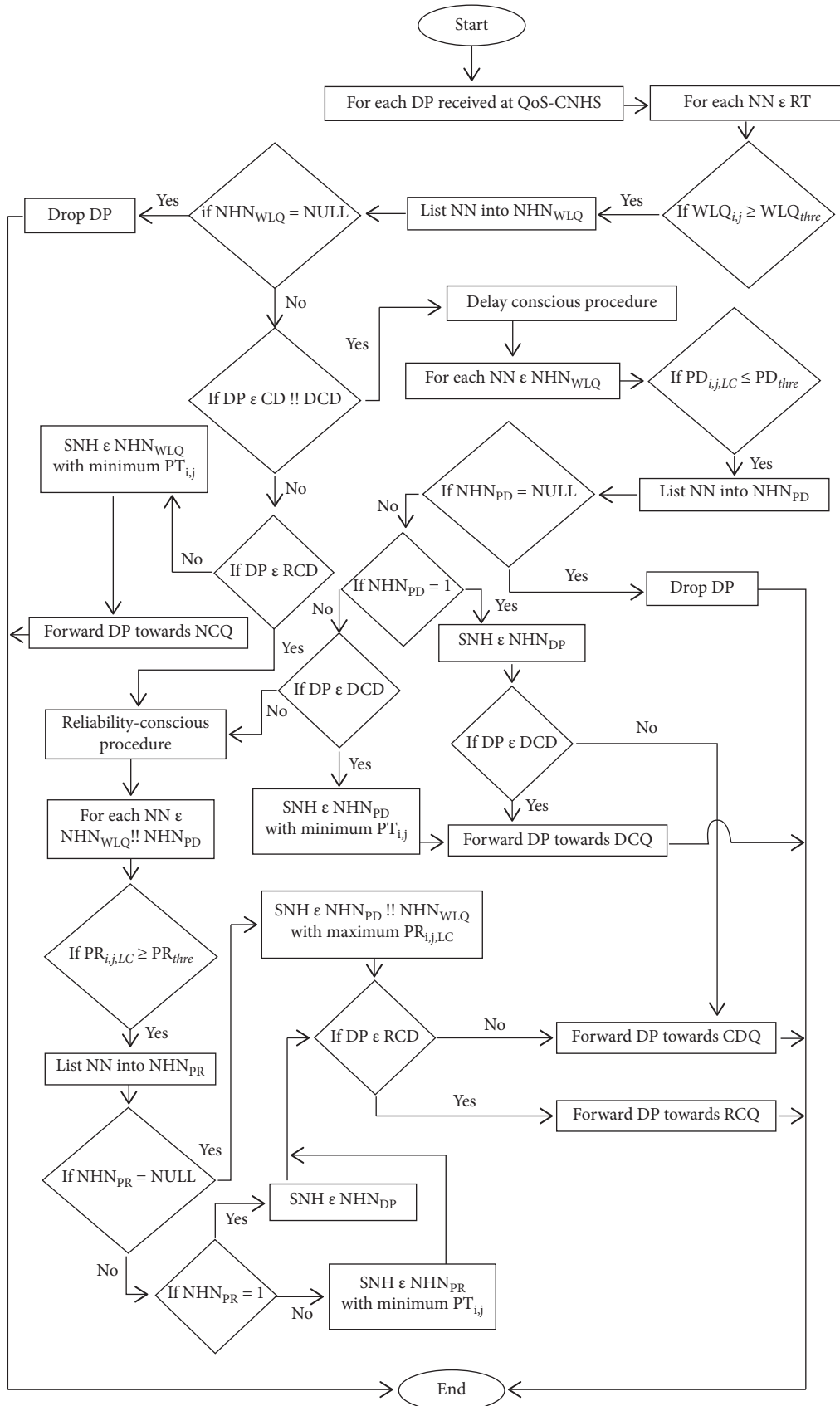


FIGURE 6: Flowchart of the proposed QoS-CNHS algorithm.

multiconstraint-aware strategy. The network parameters used in simulating the proposed routing mechanism for WBSNs are as in Table 1.

4. Simulation Results and Discussion

The performance evaluation of the proposed MCARM and MCARMR against other mechanisms in terms of the aforementioned parameters is shown and discussed in the following sections.

4.1. Packet Loss Ratio (PLR). The average PLR against wireless link qualities (WLQs) considering different data generation rates by averaging the results is given in Figure 7. It shows that the PLR is high at a very tight WLQ level for TQMoS, LTRT, and MCARM and is decreasing as its threshold level is becoming low. However, it remains almost consistent for MCARMR at different WLQ [40] threshold levels. Moreover, MCARM results in slightly poor performance compared to MCARMR [41] and significantly good performance when compared with TQMoS and LTRT. TQMoS and LTRT are not considering the in-body wireless communication and human body movements, which are the reasons for their low performance. Furthermore, TQMoS performs well when compared with LTRT because of the provision of the demanding QoS data.

4.2. Average Packet Delivery Ratio (APDR). In this section, the performance of both MCARM and MCARMR is assessed considering both the reliability-conscious and critical data in terms of APDR.

4.3. Reliability-Conscious Data (RCD). Figure 8 illustrates the APDR of RCD packets against data generation rates (DGRs) at different wireless link qualities by averaging their results. It is observed from the figure that for mechanisms the APDR is slightly reducing as the DGR is growing high which is due to the increased network congestion. The figure shows that, at high DGR, MCARM performs well when compared with all three but at medium and low DGRs it is replaced by MCARMR [42]. The reason is the increased traffic congestion on the RNs at high DGR. Furthermore, it is also observed that TQMoS shows better results when compared with LTRT at all DGRs.

TQMoS considers the temperature increase issue of the in-body BMSNs along with the provision of the demanding QoS data but completely ignores in-body wireless communication and human body movements issues. Secondly, it uses a minimum hop-count strategy while selecting the suitable next-hop; even RCD packets can tolerate some delays. Similarly, the aim of LTRT is to address the temperature increase issue of in-body BMSNs and completely overlooks the in-body wireless communication and human body movement's issues along with the provision of the demanding QoS data.

TABLE 1: Network parameters used in simulation.

Parameters	Value
Nodes quantity (MCARM)	14 (BMSNs) + 1 (LC)
Nodes quantity (MCARMR)	14 (BMSNs) + 14 (RN) + 1 (LC)
Communication range (MCARM)	40 cm
Communication range (MCARMR)	20 cm (BMSNs) and 40 cm (RN)
Initial energy	100 joules
Bit error rate (BER)	$10^{-2} - 10^{-4}$
Communication power (MCARM)	$8.5872e^{-4}$
Communication power (MCARMR)	$8.5872e^{-4}$ (BMSNs) and $1.0872e^{-4}$ (RN)
Propagation model	TwoRayGround
Buffer size	60 packets
Application type	Event-driven
Type of traffic	Constant bit rate (CBR)
MAC layer protocol	IEEE 802.15.4
Type of network interface	WirelessPhy
Simulation time	1000 seconds

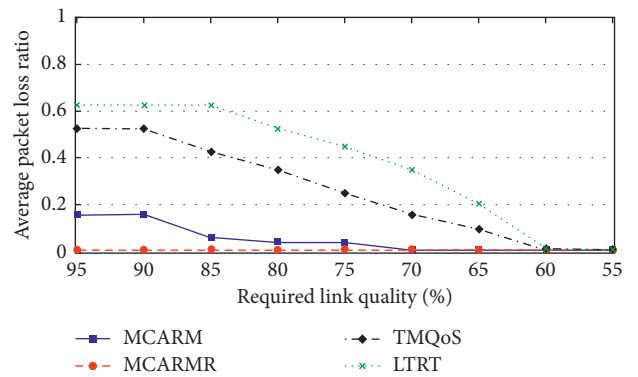


FIGURE 7: Average PLR vs. WLQ thresholds at different DGRs.

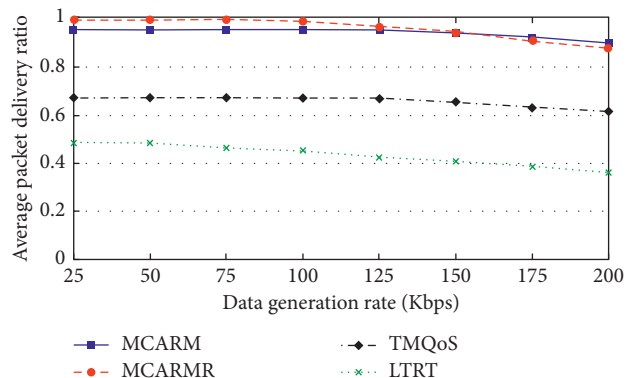


FIGURE 8: APDR vs. DGRs at different WLQs for RCD packets.

4.4. Critical Data (CD). The APDR of the critical data (CD) packets is given against DGRs at different wireless link qualities by taking an average of their results in Figure 9. By comparing Figures 8 and 9, it is observed that the performances of MCARM and MCARMR are almost the same for

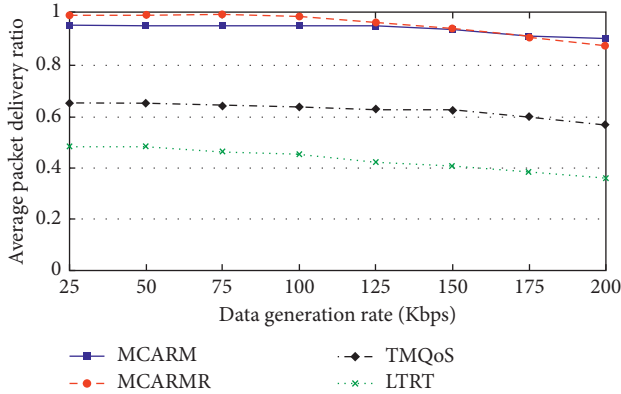


FIGURE 9: APDR vs. DGRs at different WLQs thresholds for CD packets.

both RCD and CD packets in terms of APDR for all DGRs. TMQoS results in a slightly low packet delivery ratio at low DGR but with the increase in the DGRs, its performance is becoming poorer in CD packets compared to RCD packets. It uses redundant transmission of CD packets [42], causing high network congestion which results in comparatively low APDR. Furthermore, there is no effect on the performance of LTRT as it does not consider the provision of the demanding QoS data.

4.5. Average End-to-End Delay (AEED). Figure 10 shows the AEED of the delay-conscious data (DCD) packets against different DGRs at various wireless link quality threshold levels by averaging their results. The figure illustrates that the TMQoS performs slightly better than MCARM, MCARMR, and LTRT because the suitable next-hop selection procedure of TMQoS uses hop-counts. MCARM results in slightly high AEED as compared to TMQoS but outperforms the MCARMR and LTRT. In both MCARM and MCARMR, the selection of suitable next-hop is based on the path delay and wireless link quality level for DCD packets [43]. Furthermore, LTRT results in high AEED among all due to its ignorance about the provision of the demanding QoS data.

4.6. On-Time Average Packets Success Ratio (OTAPSR). This section presents the performance assessments of both MCARM and MCARMR in terms of OTAPSR for delay-conscious data (DCD) and critical data (CD) packets.

4.7. Delay-Conscious Data (DCD). The OTAPSR of DCD packets against demanded Time-To-Leave (TTL) deadline and considering different wireless link quality threshold levels by taking the average of their results at high and low DGR is shown in Figures 11(a) and 11(b), respectively. The figures clarify that MCARM performs well when compared with other mechanisms considering both high and low DGRs at very tight TTL deadlines. As the TTL deadline is becoming relaxed, MCARMR results in improved performance when compared with others at low DGR. For high DGR, its performance remains below MCARM because of

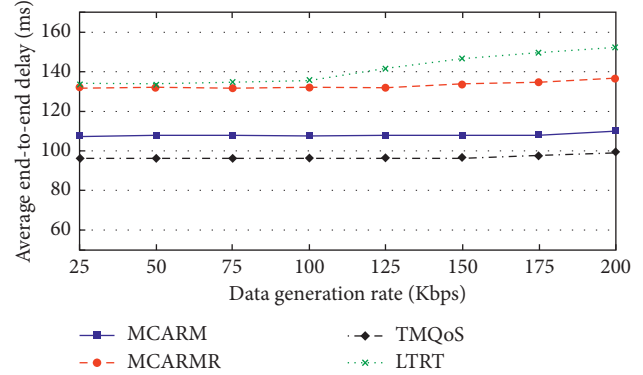


FIGURE 10: AEED vs. DGRs at different WLQ thresholds for DCD packets.

high network congestion on the RNs caused by high DGR [44]. At tight TTL deadlines, the slightly poor performance of MCARMR compared with MCARM considering both high and low DGRs is because of the delays at the RNs.

The performance of TMQoS and LTRT in terms of OTAPSR for DCD packets is improving as the TTL deadline becomes relaxed; however, they are still poorly performing when compared to MCARM and MCARMR due to not considering the in-body wireless communication and human body movements. TMQoS results in improved performance compared to LTRT because of the provision of demanding QoS data which is not considered by LTRT.

4.8. Critical Data (CD). Figures 12(a) and 12(b) illustrate the OTAPSR for CD packets against demanded TTL at different wireless link quality threshold levels by averaging their results at high and low GDGRs, respectively. By comparing the results of TMQoS for DCD packets with CD packets, it is clear that it gives comparatively better results for DCD packets because of the redundant transmission in the case of the CD packets. Moreover, there is no difference in the results of other mechanisms.

4.9. Maximum Temperature Increase. Figures 13(a) and 13(b) present the maximum temperature increase of the in-body BMSNs for DCD/RCD/NCD packets and CD packets against DGRs at different wireless link quality threshold levels by averaging their results, respectively. It is clear that the temperature increase is becoming more with the rise in the DGRs for all mechanisms. More communications occur at high DGR that results in a temperature increase of the in-body BMSNs [45]. MCARMR results in a lower temperature increase when compared with other mechanisms because the RNs are used to forward the captured data of the BMSNs. LTRT outperforms MCARM and TMQoS as its main aim is to address the temperature increase issue of the in-body BMSNs.

In the case of DCD/RCD/NCD packets, TMQoS is poorly performing as compared to MCARM because of its minimum hop-count based on suitable next-hop selection strategy. Furthermore, by comparing the maximum

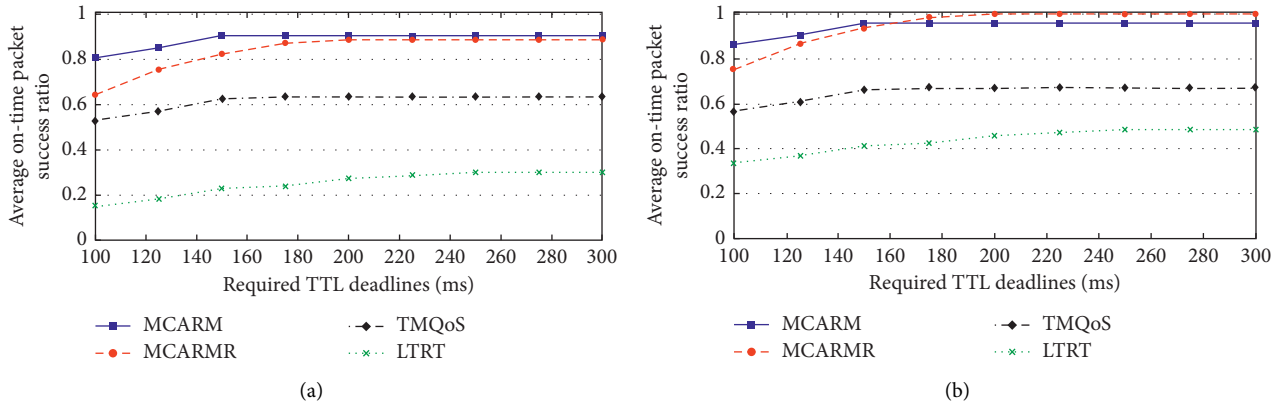


FIGURE 11: OTAPSR vs. required TTL deadlines for DCD packets at (a) high DGR and (b) low DGR.

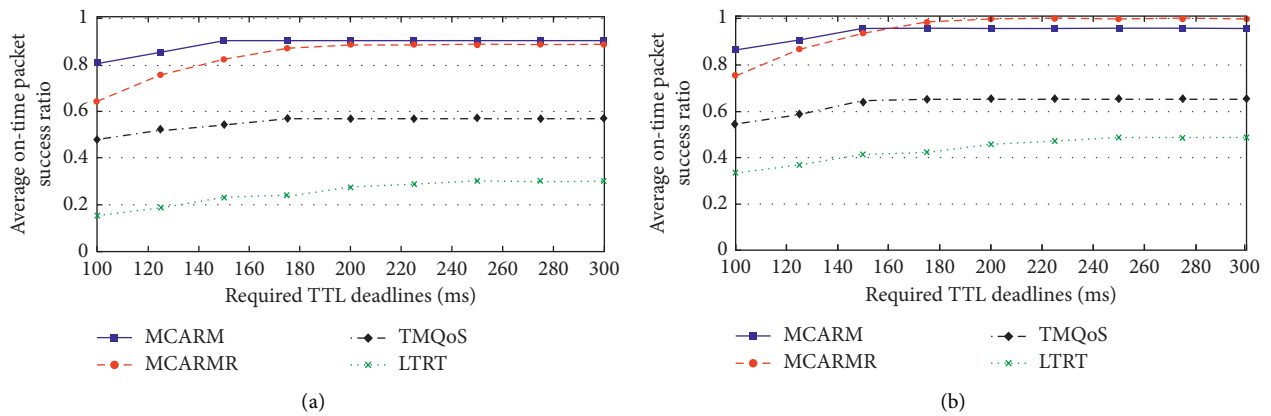


FIGURE 12: OTAPSR vs. required TTL deadlines for CD packets at (a) high DGR and (b) low DGR.

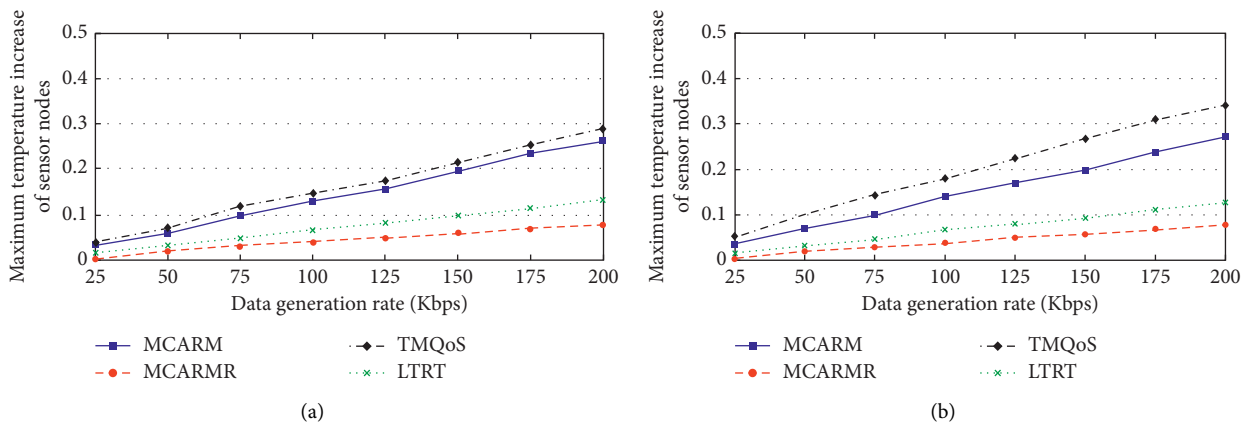


FIGURE 13: Maximum temperature increase of BMSNs vs. DGRs for (a) DCD/RCD/NCD and (b) CD packets.

temperature for TQMoS in the case of DCD/RCD/NCD packets and CD packets, it is observed that it results in more temperature increase in the case of CD packets when compared with DCD/RCD/NCD packets. This is because of the redundant transmission of CD packets resulting in more communication and more communication means high temperature increase. The results of LTRT, MCARM, and

MCARMR remain the same considering both DCD/RCD/NCD packets and CD packets.

4.10. Average Energy Consumption (AEC). The AEC of the in-body BMSNs for DCD/RCD/NCD packets and CD packets against DGRs at different wireless link quality threshold levels by taking the average of their results is

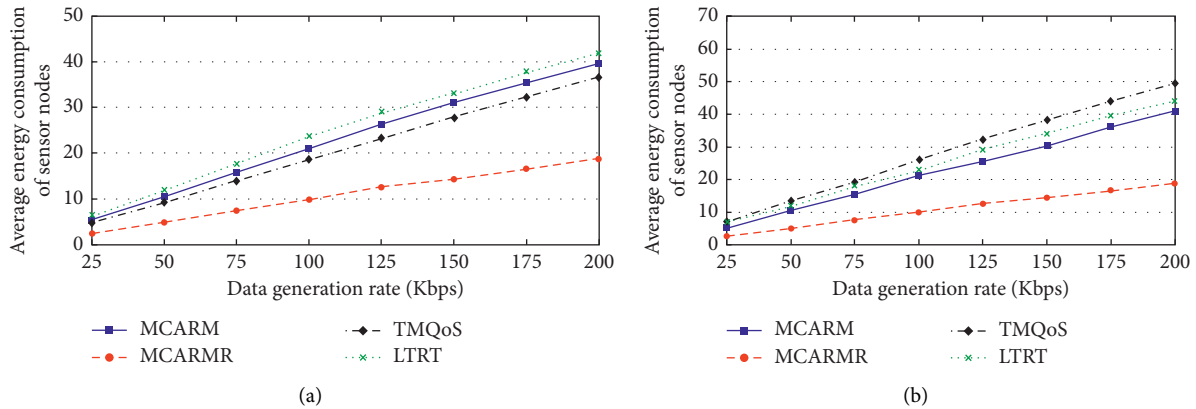


FIGURE 14: AEC of BMSNs vs. DGRs for (a) DCD/RCD/NCD and (b) CD packets.

shown in Figures 14(a) and 14(b), respectively. It is clear that the increase in DGRs increases the AEC in both cases for all mechanisms. The high energy consumption is due to more communication caused by high DGR. It is observed that MCARMR outperforms the others as the BMSNs are not involved in communicating the data of other BMSNs which consumes more energy among all tasks performed by the sensor node. LTRT consumes more energy when compared with other mechanisms because of its main aim of addressing the temperature increase issue and not considering the energy consumption. TMQoS is poorly performing in the case of CD packets when compared to MCARM because of the redundant transmission of CD packets. On the other hand, TMQoS shows improved performance when compared with MCARM because of its minimum hop-count suitable next-hop selection strategy.

5. Conclusion and Future Directions

WBSN is the medical and health-care application of WSNs offering continuous remote monitoring of different vital-signs information of the human body. Along with inherited limitations and challenges of WSNs, WBSNs pose distinct constraints due to the behavior and chemistry of the human body. The diversity of the generated data from BMSNs demands different QoS parameters in the delivery of the data to the local coordinator (LC). In addition to the demanding QoS data, the routing mechanisms need to be aware of the temperature increase of in-body BMSNs, in-body wireless communication, and human body movement issues. The existing routing mechanisms in this domain have partially addressed these issues. This research work has integrated the aforementioned issues by adopting a multiconstraint-aware strategy. Two types of network frameworks with and without relay/forwarder nodes are being used. The data packets containing physiological parameters of the human body are categorized into delay-constrained, reliability-constrained, critical (both delay- and reliability-constrained), and nonconstrained data packets. The proposed routing mechanism offers a more realistic solution with the dynamics of the human body. The contributions of the proposed routing mechanism have

demonstrated better results in terms of latency, reliability, temperature increase, and energy efficiency when compared with the existing work.

The possible future directions could be the integration of the proposed mechanism with inter-WBSNs, ensuring the privacy of the patients' vital-signs information, the optimal number of RNs and their placement, prolonging networks' lifetime, and assessment using test-bed through real-world implementation.

Data Availability

The data used to support the findings of this study are included within the article (see Table 1).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Counterpart Service for the Construction of Xiangyang Science and Technology Innovation China Innovative Pilot City.

References

- [1] Y.-Y. Shih, P.-C. Hsiu, and A.-C. Pang, "A data parasitizing scheme for effective health monitoring in wireless body area networks," *IEEE Transactions on Mobile Computing*, vol. 18, no. 1, pp. 13–27, 2019.
- [2] World Population Aging, *Population Division*, Department of Economic and Social Affairs, New York, NY, USA, 2019, <https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf>.
- [3] World Health Organization (WHO), Ageing and Health <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>, 2018.
- [4] World Population Aging, *Population Division*, Department of Economic and Social Affairs, New York, NY, USA, 2017, https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2017_Highlights.pdf.
- [5] Y. Qu, G. Zheng, H. Ma, X. Wang, B. Ji, and H. Wu, "A survey of routing protocols in WBAN for healthcare applications," *Sensors*, vol. 19, no. 7, pp. 1–24, 2019.

- [6] M. Salayma, A. Al-Dubai, I. Romdhani, and Y. Nasser, "Reliability and energy efficiency enhancement for emergency-aware wireless body area networks (WBANs)," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 3, pp. 804–816, 2018.
- [7] M. M. Monowar, M. M. Hassan, F. Bajaber, M. A. Hamid, and A. Alamri, "Thermal-aware multiconstrained intrabody QoS routing for wireless body area networks," *International Journal of Distributed Sensor Networks*, vol. 2014, Article ID 676312, 14 pages, 2014.
- [8] W. Jiang, Z. Wang, M. Feng, and T. Miao, "A survey of thermal-aware routing protocols in wireless body area networks," *IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, vol. 2, pp. 17–21, 2017.
- [9] A. Samanta and S. Misra, "Energy-efficient and distributed network management cost minimization in opportunistic wireless body area networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 2, pp. 376–389, 2018.
- [10] J. Bangash, A. Abdullah, M. Anisi, and A. Khan, "A survey of routing protocols in wireless body sensor networks," *Sensors*, vol. 14, no. 1, pp. 1322–1357, 2014.
- [11] J. I. Bangash, A. H. Abdullah, M. A. Razzaque, and A. W. Khan, "Reliability aware routing for intra-wireless body sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2014, Article ID 786537, 10 pages, 2014.
- [12] J. I. Bangash, A. H. Abdullah, M. A. Razzaque, A. W. Khan, and R. Yusof, "Critical data routing (CDR) for intra wireless body sensor networks," *Telkommika (Telecommunication Computing Electronics and Control)*, vol. 13, no. 1, pp. 181–192, 2015.
- [13] J. I. Bangash, A. W. Khan, and A. H. Abdullah, "Data-centric routing for intra wireless body sensor networks," *Journal of Medical Systems*, vol. 39, no. 91, pp. 1–13, 2015.
- [14] Z. Khan, S. Sivakumar, W. Phillips, B. Robertson, and Q. P. R. D. "QoS-aware peering routing protocol for delay sensitive data in hospital body area network communication," in *Proceedings of the 7th IEEE International Conference on Broadband, Wireless Computing, Communication and Applications (BWCCA)*, pp. 178–185, British Columbia, Canada, November 2012.
- [15] Z. A. Khan, S. Sivakumar, W. Phillips, and B. Robertson, "A QoS-aware routing protocol for reliability sensitive data in hospital body area networks," *Procedia Computer Science*, vol. 19, pp. 171–179, 2013.
- [16] Z. A. Khan, S. Sivakumar, W. Phillips, and B. Robertson, "ZEQoS: a new energy and QoS-aware routing protocol for communication of sensor devices in healthcare system," *International Journal of Distributed Sensor Networks*, vol. 2014, Article ID 627689, 10 pages, 2014.
- [17] G. Ahmed, Z. Jianhua, and M. M. S. Fareed, "PERA: priority-based energy-efficient routing algorithm for WBANs," *Wireless Personal Communications*, vol. 96, no. 3, pp. 4737–4753, 2017.
- [18] M. Monowar and F. Bajaber, "On designing thermal-aware localized QoS routing protocol for in-vivo sensor nodes in wireless body area networks," *Sensors*, vol. 15, no. 6, pp. 14016–14044, 2015.
- [19] D. Hansen, N. Tummala, S. K. S. Gupta, and L. Schwieber, "Evidence of a gastrin-like substance in *Rhinobatus productus*," *Comparative Biochemistry and Physiology. C: Comparative Pharmacology*, vol. 52, no. 1, pp. 61–63, 1975.
- [20] A. Bag and M. A. Bassiouni, "Energy efficient thermal aware routing algorithms for embedded biomedical sensor networks," in *Proceedings of the 2006 IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pp. 604–609, Vancouver, BC, October 2006.
- [21] D. Takahashi, Y. Xiao, F. Hu, J. Chen, and Y. Sun, "Temperature-aware routing for telemedicine applications in embedded biomedical sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, Article ID 572636, 11 pages, 2008.
- [22] A. R. Bhangwar, P. Kumar, A. Ahmed, and M. I. Channa, "Trust and thermal aware routing protocol (TTRP) for wireless body area networks," *Wireless Personal Communications*, vol. 97, no. 1, pp. 349–364, 2017.
- [23] B. S. Kim, S. Kang, J. Lim, K. H. Kim, and K. I. Kim, "A mobility-based temperature-aware routing protocol for wireless body sensor networks," in *Proceedings of the IEEE International Conference on Information Networking (ICOIN)*, pp. 63–66, Jeju Island, Korea, January 2017.
- [24] Z. Shahbazi and Y.-C. Byun, "Towards a secure thermal-energy aware routing protocol in wireless body area network based on blockchain Technology," *Sensors*, vol. 20, no. 12, pp. 1–26, 2020.
- [25] M. Geeta and R. Ganesan, "CEPRAN-cooperative energy efficient and priority based reliable routing protocol with network coding for WBAN," *Wireless Personal Communications*, vol. 19, 2020.
- [26] M. D. Khan, Z. Ullah, A. Ahmad et al., "Energy harvested and cooperative enabled efficient routing protocol (EHCRP) for IoT-WBAN," *Sensors*, vol. 20, no. 21, pp. 1–23, 2020.
- [27] R. Saha, S. Biswas, S. Sarma, S. Karmakar, and P. Das, "Design and Implementation of Routing Algorithm to Enhance Network Lifetime in WBAN," *Wireless Personal Communications*, vol. 1, pp. 1–38, 2021.
- [28] A. K. Sagar, S. Singh, and A. Kumar, "Energy-aware WBAN for Health Monitoring Using Critical Data Routing (CDR)," *Wireless Personal Communications*, pp. 1–30, 2020.
- [29] A. Sangwan and P. P. Bhattacharya, "Revised EECBSR for energy efficient and reliable routing in WBAN," *Majlesi Journal of Telecommunication Devices*, vol. 7, no. 2, pp. 33–40, 2018.
- [30] O. Smail, A. Kerrar, Y. Zetili, and B. Cousin, "ESR: energy aware and Stable Routing protocol for WBAN networks," in *Proceedings of the IEEE International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 452–457, Paphos, Cyprus, September 2016.
- [31] M. Anwar, A. H. Abdullah, A. Altameem et al., "Green communication for wireless body area networks: energy aware link efficient routing approach," *Sensors (Basel)*, vol. 18, no. 10, p. 17, 2018.
- [32] N. Javaid, A. Ahmad, Q. Nadeem, M. Imran, and N. Haider, "SIMPLE: Improved stable increased-throughput multi-hop link efficient routing protocol for Wireless Body Area Networks," *Computers in Human Behavior*, vol. 51, pp. 1003–1011, 2015.
- [33] F. D'Andreagiovanni, D. Nace, A. Nardin, and E. Natalizio, "Robust relay node placement in body area networks by heuristic min-max regret," *IEEE Balkan Conference on Communications and Networking (BALKANCOM)*, vol. 5, 2017.
- [34] A. Maskooki, C. B. Soh, E. Gunawan, and K. S. Low, "Adaptive routing for dynamic on-body wireless sensor networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 2, pp. 549–558, 2015.
- [35] N. Javaid, A. Ahmad, Y. Khan, Z. A. Khan, and T. A. Alghamdi, "A relay based routing protocol for wireless in-body sensor networks," *Wireless Personal Communications*, vol. 80, no. 3, pp. 1063–1078, 2015.

- [36] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: properties and enhancements," *Technometrics*, vol. 32, no. 1, p. 12, 1990.
- [37] A. Woo and D. Culler, "Evaluation of Efficient Link Reliability Estimators for Low-Power Wireless Networks," *Computer Science Division*, vol. 20, 2003.
- [38] E. Reusens, W. Joseph, B. Latre et al., "Characterization of on-body communication channel and energy efficient topology design for wireless body area networks," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 6, pp. 933–945, 2009.
- [39] H. H. Pennes, "Analysis of tissue and arterial blood temperatures in the resting human forearm," *Journal of Applied Physiology*, vol. 1, no. 2, pp. 93–122, 1948.
- [40] J. Zhang, K. Yu, Z. Wen, X. Qi, and A. Kumar Paul, "3D reconstruction for motion blurred images using deep learning-based intelligent systems," *Computers, Materials & Continua*, vol. 66, no. 2, pp. 2087–2104, 2021.
- [41] K.-P. Yu, L. Tan, M. Aloqaily, H. Yang, and Y. Jararweh, "Blockchain-enhanced data sharing with traceable and direct revocation in IIoT," *IEEE Transactions on Industrial Informatics*, vol. 1, 2021.
- [42] K. Yu, L. Lin, M. Alazab, L. Tan, and B. Gu, "Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, 2020.
- [43] K. Yu, L. Tan, X. Shang, J. Huang, G. Srivastava, and P. Chatterjee, "Efficient and privacy-preserving medical research support platform against COVID-19: a blockchain-based approach," *IEEE Consumer Electronics Magazine*, vol. 10, no. 2, pp. 111–120, 2021.
- [44] C. Feng, K. Yu, A. K. Bashir et al., "Efficient and secure data sharing for 5G flying drones: a blockchain-enabled approach," *IEEE Network*, vol. 35, no. 1, pp. 130–137, 2021.
- [45] L. Zhen, A. K. Bashir, K. Yu, Y. D. Al-Otaibi, C. H. Foh, and P. Xiao, "Energy-Efficient random access for LEO satellite-assisted 6G internet of remote things," *IEEE Internet of Things Journal*, vol. 1, 2020.

Research Article

Software Defect Prediction for Healthcare Big Data: An Empirical Evaluation of Machine Learning Techniques

Bilal Khan,¹ Rashid Naseem ,² Muhammad Arif Shah ,² Karzan Wakil,³ Atif Khan ,⁴ M. Irfan Uddin ,⁵ and Marwan Mahmoud ⁶

¹Department of Computer Science, City University of Science and Information Technology, Peshawar 25000, Pakistan

²Department of IT and Computer Science, Pak-Austria Fachhochschule Institute of Applied Sciences and Technology, Haripur, Pakistan

³Research Center, Sulaimani Polytechnic University, Sulimani 46001, Kurdistan Region, Iraq

⁴Department of Computer Science, Islamia College, Peshawar 2500, Pakistan

⁵Institute of Computing Kohat University of Science and Technology, Kohat, Pakistan

⁶Faculty of Applied Studies, King Abdulaziz University, Jeddah, Saudi Arabia

Correspondence should be addressed to Rashid Naseem; rnsqau@gmail.com and Muhammad Arif Shah; arif.websol@gmail.com

Received 9 September 2020; Revised 29 September 2020; Accepted 24 February 2021; Published 16 March 2021

Academic Editor: Nazir Shah

Copyright © 2021 Bilal Khan 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.

Software defect prediction (SDP) in the initial period of the software development life cycle (SDLC) remains a critical and important assignment. SDP is essentially studied during few last decades as it leads to assure the quality of software systems. The quick forecast of defective or imperfect artifacts in software development may serve the development team to use the existing assets competently and more effectively to provide extraordinary software products in the given or narrow time. Previously, several canvassers have industrialized models for defect prediction utilizing machine learning (ML) and statistical techniques. ML methods are considered as an operative and operational approach to pinpoint the defective modules, in which moving parts through mining concealed patterns amid software metrics (attributes). ML techniques are also utilized by several researchers on healthcare datasets. This study utilizes different ML techniques software defect prediction using seven broadly used datasets. The ML techniques include the multilayer perceptron (MLP), support vector machine (SVM), decision tree (J48), radial basis function (RBF), random forest (RF), hidden Markov model (HMM), credal decision tree (CDT), *K*-nearest neighbor (KNN), average one dependency estimator (A1DE), and Naïve Bayes (NB). The performance of each technique is evaluated using different measures, for instance, relative absolute error (RAE), mean absolute error (MAE), root mean squared error (RMSE), root relative squared error (RRSE), recall, and accuracy. The inclusive outcome shows the best performance of RF with 88.32% average accuracy and 2.96 rank value, second-best performance is achieved by SVM with 87.99% average accuracy and 3.83 rank values. Moreover, CDT also shows 87.88% average accuracy and 3.62 rank values, placed on the third position. The comprehensive outcomes of research can be utilized as a reference point for new research in the SDP domain, and therefore, any assertion concerning the enhancement in prediction over any new technique or model can be benchmarked and proved.

1. Introduction

Software engineering (SE) is a discipline that is worrisome with all qualities of software development from the beginning of software specification over to keeping up to the software maintenance after it has gone into practice [1]. In the domain of SE, software defect prediction (SDP) is the utmost significant and dynamic research zone that assumes a significant job in the software quality assurance (SQA) [2, 3]. The rising convolutions as well

dependencies of software systems have expanded the difficulty to deliver software with minimal effort, high caliber, and maintainability as well increase the chances of making software defects (SDs) [4, 5]. SD is a flaw or insufficiency in a software system that roots the development of a spontaneous result. An SD can moreover be the situation when the last software product does not meet the client's desire or client prerequisite [6]. SD's can cause the diminution of the software product quality and increase the development cost.

SDP is a momentous commotion to assure the substances of a software system that leads to adequate development cost and recover the quality by identifying defect-prone instances before testing [4]. It moreover embraces categorizing software components in different varieties of a software system that constructs the testing progression supplementary by concentrating on testing as well as evaluating the components classified as defective [7]. Defects adversely affect software reliability and quality [8].

SDP in the primary period of the software development life cycle (SDLC) is measured as an utmost thought-provoking aspect of SQA [9]. In SE, bug fixing and testing are very costly which also require a massive amount of resources. Forecasting the software defects in software development has been observed by numerous studies in the last decades. Amid all these studies, machine learning (ML) techniques are considered as the best approach toward SDPs [7, 10, 11].

Keeping the above issue related to SDP, various researchers evaluated and built SDP models utilizing diverse classification techniques. Still, it is quite challenging to sort any broad-spectrum preparation to inaugurate the usability of these techniques. Inclusively, it was originated that notwithstanding some dissimilarities in the studies, no particular SDP technique delivers higher to the other techniques diagonally different datasets. The researchers have utilized different evaluation measures to assess the projected models to find the best model for SDP [12, 13].

However, this study focuses on the empirical analysis of ten ML techniques amid which some are proposes as new solutions for SDP. ML techniques include the multilayer perceptron (MLP), radial basis function (RBF), support vector machine (SVM), decision tree (J48), random forest (RF), hidden Markov model (HMM), credal decision tree (CDT), K -nearest neighbor (KNN), average one dependency estimator (A1DE), and Naïve Bayes (NB) for SDP. Amid all these techniques, HMM and A1DE are proposed aimed for the first time for SDP. These techniques are employed on seven different datasets including AR1, AR3, CM1, JM1, KC2, KC3, and MC1. All the experiments are validated using relative absolute error (RAE), mean absolute error (MAE), root relative squared error (RRSE), root mean squared error (RMSE), recall, and accuracy.

Following is a list of the contributions of this research:

- (1) To benchmark ten different ML techniques (MLP, J48, SVM, RF, RBF, HMM, CDT, A1DE, KNN, and NB) for SDP
- (2) To demeanor a series of try-outs on different datasets such as AR1, AR3, CM1, JM1, KC2, KC3, and MC1
- (3) To reveal insight into the experimental outcomes, evaluation is accomplished using MAE, RAE, RMSE, RRSE, recall, and accuracy
- (4) To show that experimental outcomes are significantly different and comparable with verifying the best results, Friedman two-way examination of difference by ranks is performed

Hereinafter, Section 2 presents the literature survey, Section 3 comprises the methodology and techniques, while

experimental outcomes are discussed in Sections 4, and Section 5 covers the inclusive conclusion.

2. Literature Survey

This section delivers an ephemeral study about existing techniques in the field of SDP. Several researchers have employed ML techniques for SDP at the initial phase of software development. Several particular studies converse here. Czibula et al. [11] presented a model grounded on relational association discovery (RAD) for SDP. They apply all investigations on NASA dataset including KC1, KC3, MC2, MW1, JM1, PC3, PC4, PC1, PC2, and CM1. To assess the model as compared to other models, use accuracy, precision, specificity, probability of detection (PD), and area under cover (ROC) assessment measure. The acquired outcomes present that RAD perform well rather than other employed techniques.

A framework for SDP named the Defect Prediction through Convolutional Neural Network (DP-CNN) has been recommended by Li et al. [14]. The authors evaluated the DP-CNN on seven different open source projects such as Camel, jEdit, Lucene, Xalam, Xerces, Synapse, and Poi in terms of F -measure in defect predictions. Overall outcomes illustrate that on average, the DP-CNN enhanced the up-to-the-minute technique by 12%.

Jacob and Raju [15] introduced a hybrid feature selection (HFS) method for SDP. They also perform their analysis on NASA datasets including PC1, PC2, PC3, PC4, CM1, JM1, KC3, and MW1. The outcomes of HFS are benchmarked with Naïve Bayes (NB), neural networks (NN), RF, random tree (RT), and J48. Benchmarking is carried out using accuracy, specificity, sensitivity, and Matthew's correlation coefficient (MCC). The analyzed outcome shows that HFS outperform while improving classification accuracy from 82% to 98%.

Bashir et al. [16] presented a joined framework to improve the SDP model using Ranker feature selection (RFS), data sampling (DS), and iterative partition filter (IPF) techniques to conquest class imbalance, noisy correspondingly, and high dimensionality. Seven ML techniques including NB, RF, KNN, MLP, SVM, J48, and decision stump are employed on CM1, JM1, KC2, MC1, PC1, and PC5 datasets for evaluations. The outcomes are carried out utilizing receiver operating characteristic (ROC) performance evaluation. Overall experimental outcomes of the proposed model outperformed other models.

A new approach for SDP utilizing a hybridized gradual relational association (HyGRAR) and artificial neural network (ANN) to classify the defective and nondefective objects is projected in [7]. Experiments were achieved based on ten different open source datasets such as Tomcat 6.0, Anr 1.7, jEdit 4.0, jEdit 4.2, jEdit 4.3, AR1, AR3, AR4, AR5, and AR6. For module evaluation, accuracy, sensitivity, specificity, and precision measures were utilized. The author concluded that HyGRAR achieved better outcomes as compared to most of the foregoing projected approaches.

Alsaedi and Khan [8] performed the comparison on supervised learning techniques including bagging, SVM,

decision tree (DT), and RF and ensemble classifiers on different NASA datasets such as CM1, MC1, MC2, PC1, PC3, PC4, PC5, KC2, KC3, and JM1. The basic learning and ensemble classifiers are evaluated using *G*-measure, specificity, F-score, recall, precision, and accuracy. The experimental results conducted show that RF, AdaBoost with RF, and DS with bagging outperform than other employed techniques.

The author in [9] performed comparative exploration of several ML techniques for SDP on twelve NASA datasets such as MW1, CM1, JM1, PC1, PC2, PC3, PC4, PC5, KC1, KC3, MC1, and MC2, while the classification techniques include one rule (OneR), NB, MLP, DT, RBF, kStar (K*), SVM, KNN, PART, and RF. The performance of each technique is assessed using MCC, ROC area, recall, precision, *F*-measure, and accuracy.

Malhotra and Kamal [6] evaluated the efficiency of ML classifiers for SDP on twelve excessive datasets taken from the NASA repository by employing sampling approaches and cost-sensitive classifiers. They examine five prevailing methods including J48, RF, NB, AdaBoost, and bagging, as well as suggest the SPIDER3 method for SDP. They have compared the performance based on accuracy, sensitivity, specificity, and precision.

Manjula and Florence [17] developed a hybrid model of the genetic algorithm (GA) and the deep neural network (DNN). GA is utilized for feature optimization while DNN is for classification. The enactment of the projected technique is benchmarked with NB, RF, DT, Immunos, ANN-artificial bee colony (ABC), SVM, majority vote, AntMiner+, and KNN. All the performances are carried out on a dataset that includes KC1, KC2, CM1, PC1, and JM1 and assessed via recall, F-score, sensitivity, precision, specificity, and accuracy. The tentative results show that the recommended technique beats other techniques in terms of achieving better accuracy.

Researchers have used various techniques to incredulous the boundaries of SDP on a variety of datasets. In each study, different evaluation measures are accomplished to evaluate and benchmark the proposed techniques. The overall summary of the literature discussed above is listed in Table 1, where the first column represents the authors who conducted research studies utilizing various ML techniques. The second column of the table shows techniques utilized by an individual study, while the third and fourth columns represent dataset and evaluation measures utilized in different studies. As shown in Table 1, each study has used different evaluation measures to achieve higher accuracy, but none affects decreasing error rate which is a significant feature.

Moreover, the ML techniques are also utilized by many researchers in healthcare engineering and the development of medical data analyzing software [1]. Khan et al. [2] utilized machine learning techniques for the prophecy of chronic kidney disease (CKD) to suggest the best model of early prediction of CKD. The study of Makumba et al. [3] on heart disease prediction using data mining (DM)/ML techniques can also be the baseline for new researchers. They have employed the DM/ML techniques on heart disease datasets. Hence, many researchers have utilized ML techniques on different healthcare datasets for early prediction of disease. However, the most important task is that

when they propose an optimal solution for any kind of disease, they also have to give the assurance for the quality of software that will be developed using their optimal solution. To ensure this, we have to predict the defect that may occur in the software which leads towards decreasing the quality of the software system. Those are the reasons behind this research study.

3. Methodology and Techniques

This study objects to present the performance analysis of ML techniques for SDP on various datasets including AR1, AR3, CM1, JM1, KC2, KC3, and MC1. All these datasets can be found on the UCI ML repository (<https://archive.ics.uci.edu/>). The experimentation is performed using the open source ML and DM tool Weka version 3.9 (<https://machinelearningmastery.com/use-ensemble-machine-learning-algorithms-weka/>). As per the information presented in Table 1, AR1 and AR3 are reported in the literature single time; as shown in Figure 1, CM1 and JM1 reported 6 times, KC2 and MC1 reported 1 time, while KC3 reported 4 times. Each dataset is consisting of some attributes along with known output class. Respectively, datasets contain numerical data, while the total numbers of attributes and instances are different as presented in Table 2. In Table 2, the first column shows the datasets and second and third columns present number of metrics (attributes) and several cases (instances) correspondingly. The fourth and fifth columns represent the number of defective modules and the number of nondefective modules correspondingly, while the last column shows the type of data in each dataset. However, Table 3 shows the list of all attributes (software metrics) according to each dataset utilized in this research. The experimental setup for SDP is shown in Figure 2, which explains how each task is performed in this research. After training the datasets, the preprocessing step is taken only on the class attribute of each dataset that is solitary to change the type of data from numerical to categorical due to some of the ML techniques unable to work on numerical type class attributes. After all, when ML techniques apply to each dataset, the outcome is assessed using different assessment measures to show the better performance of an individual technique. Therefore, six assessment measure named MAE [13, 18, 19], RMSE [8, 20, 21], RAE [16, 22, 23], RRSE [22, 24], recall [9, 10, 25], and accuracy [26–28] are utilized to evaluate the performance of ML techniques on SDP datasets. We have used error-based assessment measures which are not reported in the literature, while recall and accuracy have been used 3 and 7 times, respectively (Figure 3).

Table 4 shows the calculation mechanism and a description of each evaluation measure. The second column of Table 4 shows the list of evaluation measures, while the third column represents the equation of each measure, where, $|y_i - y|$ is the absolute error, n is the number of errors, T_j is the goal value for record j , P_{ij} is the prediction value by the particular technique I for record j (beyond n records), TP is the quantity of true-positive classification, FN is the amount of false-negative classification, TN is the amount of true-negative classification, and FP is the quantity of false-positive classifications.

TABLE 1: Summary of the literature survey.

Author	Technique/Model	Datasets	Evaluation measures
Czibula et al. [11]	RAD	MW1, JM1, PC1, PC2, PC3, PC4, KC1, KC3, MC2, and CM1	Accuracy, specificity, precision, PD, and ROC
Li et al. [14]	DP-CNN	Camel, jEdit, Lucene, Xalam, Xerces, Synapse, and Poi	<i>F</i> -measure
Jacob and Raju [15]	HFS, NB, NN, RF, RT, J48	PC1, PC2, PC3, PC4, CM1, MW1, KC3, and JM1	Specificity, sensitivity, MCC, and accuracy
Bashir et al. [16]	NB, RF, KNN, MLP, SVM, J48, and decision stump	CM1, JM1, KC2, MC1, PC1, and PC5	ROC
Miholca et al. [7]	HyGRAR	Tomcat 6.0, Anr 1.7, jEdit 4.0, AR1, jEdit 4.2, AR3, jEdit 4.3, AR5, AR4, and AR6	Accuracy, sensitivity, specificity, and precision
Alsaeedi and Khan [8]	Bagging, SVM, DT, and RF	PC1, PC3, PC4, PC5, JM1, KC2, KC3, MC1, MC2, and CM1	<i>G</i> -measure, specificity, <i>F</i> -score, recall, precision, and accuracy
Iqbal et al. [9]	OneR, NB, K*, MLP, SVM, RBF, RF, KNN, DT, and PART	JM1, MW1, CM1, MC1, PC1, MC2, PC4, PC3, PC2, PC5, KC3, and KC1	MCC, ROC area, <i>F</i> -measure, recall, precision, and accuracy
Malhotra and Kamal [6]	J48, RF, NB, AdaBoost, and bagging, and SPIDER3	NASA datasets	Accuracy, sensitivity, specificity, and precision
Manjula and Florence [17]	GA, DNN, NB, RF, DT, ABC, SVM, and KNN	KC1, KC2, CM1, PC1, and JM1	Precision, sensitivity, specificity, recall, <i>F</i> -score, and accuracy

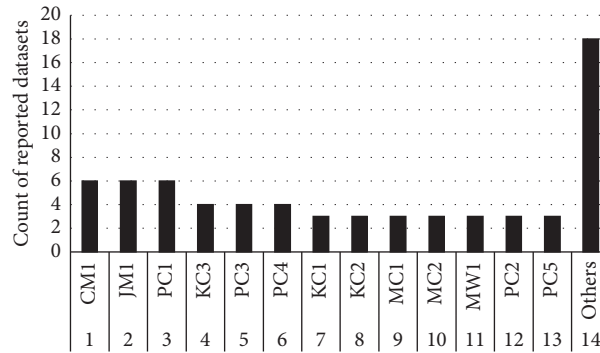


FIGURE 1: Count of reported datasets.

TABLE 2: Attributes, instances, defective, and nondefective modules of each utilized dataset.

S. No.	Datasets	No. of attributes	No. of instances	No. of defective modules	No. of nondefective modules	Data type
1	AR1	30	121	9 (7.4%)	112 (92.6%)	Numerical
2	AR3	30	63	8 (12.7%)	55 (87.3%)	Numerical
3	CM1	22	498	49 (9.8%)	449 (90.2%)	Numerical
4	JM1	22	9593	1759 (18.3%)	7834 (81.7%)	Numerical
5	KC2	22	522	107 (20.5%)	415 (79.5%)	Numerical
6	KC3	40	194	36 (18.6%)	158 (81.4%)	Numerical
7	MC1	40	9466	68 (0.7%)	9398 (99.3%)	Numerical

4. Techniques Employed

ML techniques are currently extensively used to excerpt significant knowledge commencing massive volumes of data in diverse areas. ML applications embrace numerous real-world situations such as cyber-security, bioinformatics, detecting communities in social networks, and software process enhancement to harvest high-quality software systems [7]. ML-based solutions for SDP have also been investigated [6, 10, 29]. From which, we have selected the top seven techniques as reported in Table 1, and the count of

each technique is given in Figure 4. RBF is selected randomly, while the other two, i.e., HMM and A1DE, are new explorations for SDP. All of the ten selected techniques are briefly discussed in the following subsections.

4.1. Support Vector Machine. SVM has numerous uses in the field of classification, biophotonics, and pattern recognition [8, 25]. First, it was developed for binary classification; however, it can also be used for multiple classes [30]. In binary classification, the core impartial of SVM is to describe a line among

TABLE 3: List of attributes according to datasets.

Attributes	Datasets							
	AR1	AR3	CM1	JM1	KC2	KC3	MC1	
Halstead attributes	Halstead content	✓	✓	—	✓	-	✓	✓
	Halstead difficulty	✓	✓	✓	✓	✓	✓	✓
	Halstead effort	✓	✓	✓	✓	✓	✓	✓
	Halstead error estimator	✓	✓	—	✓	-	✓	✓
	Halstead length	✓	✓	✓	✓	✓	✓	✓
	Halstead level	✓	✓	✓	✓	✓	✓	✓
	Halstead program time	✓	✓	✓	✓	✓	✓	✓
	Halstead volume	✓	✓	✓	✓	✓	✓	✓
	Number of operands	✓	✓	✓	✓	✓	✓	✓
	Number of operators	✓	✓	✓	✓	✓	✓	✓
	Number of unique operands	✓	✓	✓	✓	✓	✓	✓
	Number of unique operators	✓	✓	✓	✓	✓	✓	✓
McCabe attributes	Essential complexity	—	—	✓	✓	✓	✓	✓
	Cyclomatic complexity	✓	✓	✓	✓	✓	✓	✓
	Design complexity	✓	✓	✓	✓	✓	✓	✓
	Cyclomatic density	✓	✓	—	—	—	✓	✓
Size attributes	Number of lines	—	—	✓	-	✓	✓	✓
	LOC total	✓	✓	✓	✓	✓	✓	✓
	LOC executable	✓	✓	—	✓	—	✓	✓
	LOC comments	✓	✓	✓	✓	✓	✓	✓
	LOC code and comments	✓	✓	✓	✓	✓	✓	✓
	LOC blank	✓	✓	✓	✓	✓	✓	✓
	Branch count	✓	✓	✓	✓	✓	✓	✓
	Condition count	✓	✓	—	—	—	✓	✓
	EDGE count	—	—	—	—	—	✓	✓
	Parameter count	✓	✓	—	—	—	✓	✓
	Modified condition count	—	—	—	—	—	✓	✓
	Multiple condition count	✓	✓	—	—	—	✓	✓
Others attributes	Node count	—	—	—	—	—	✓	✓
	Design density	✓	✓	—	—	—	✓	✓
	Essential density	—	—	—	—	—	✓	✓
	Decision count	✓	✓	—	—	—	✓	✓
	Decision density	✓	✓	—	—	—	✓	-
	Call pairs	✓	✓	—	—	—	✓	✓
	Global data complexity	—	—	—	—	—	✓	✓
	Global data density	—	—	—	—	—	✓	✓
	Maintenance severity	—	—	✓	—	✓	✓	✓
	Normalized cyclomatic complexity	✓	✓	—	—	—	✓	✓
Class attribute	Pathological complexity	—	—	—	—	—	—	✓
	Percent comments	—	—	✓	—	✓	✓	✓
	Defective	✓	✓	✓	✓	✓	✓	✓

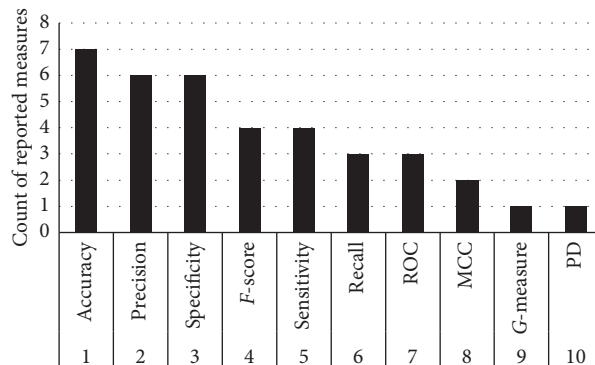


FIGURE 2: Software defect prediction model.

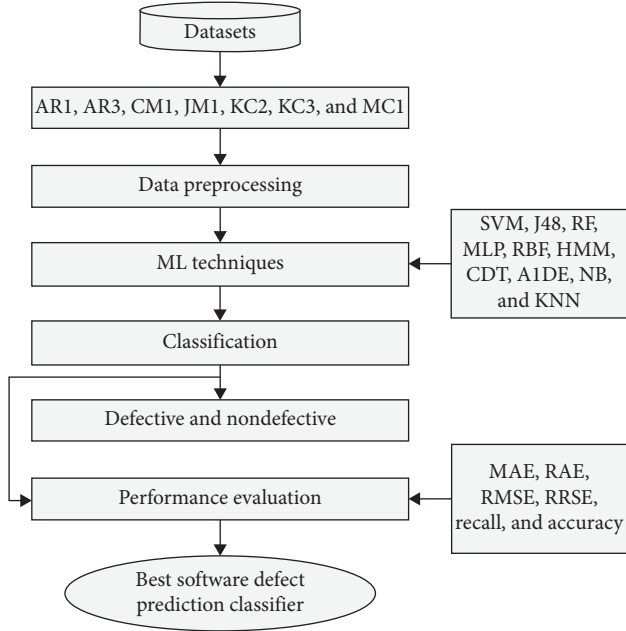


FIGURE 3: Count of reported measures.

TABLE 4: Measurements to evaluate the experimental results.

S. No.	Measure	Equation
1	MAE	$MAE = (1/2) \sum_{j=1}^n y_i - y $
2	RMSE	$RMSE = \sqrt{(1/2) \sum_{j=1}^n (y_i - 1)^2}$
3	RAE	$RAE = ((\sum_{j=1}^n P_{ij} - T) / (\sum_{j=1}^n T_j - \bar{T}))$
4	RRSE	$RRSE = \sqrt{((\sum_{j=1}^n (P_{ij} - T_j)^2) / (\sum_{j=1}^n (T_j - T)^2))}$
5	Recall	$Recall = TP / (TP + FN)$
6	Accuracy	$Accuracy = (TP + TN) / (TP + TN + FP + FN)$

classes of data to exploit the remoteness of edge line from data points lying neighboring to it. In that case, if data are linearly inseparable, a mathematical function is utilized to transmute the data to a higher-attribute space, so that it may become linear divisible in the new space. The function used is called kernel function, and the equation of a linear SVM can be written as

$$f(x) = \sum_{i=0}^N \alpha_i y_i x_i^T \cdot x + \beta_0, \quad (1)$$

where x_i is the prompt with label y_i , α is the Lagrange multiplier, and β_0 is the partiality, while N signifies the number of support vectors. For nonlinearly divisible issue, the overhead equation can be improved for kernel SVM as

$$f(x) = \sum_{i=0}^N \alpha_i y_i K(x_i, x) \cdot x + \beta_0, \quad (2)$$

where $K(x_i, x)$ is the kernel function.

4.2. Decision Tree (J48). This is the basic C4.5 decision tree (DT) used for classification problems [26]. It is the deviation of information gain (IG), usually utilized to stun the result of

unfairness. An attribute with a maximum gain ration is nominated in direction to shape a tree as a splitting attribute. Gain ratio- (GR-) based DT performs well as compared to IG [31], in terms of accuracy. GR is defined as

$$Gain_{ratio(D.A)} = \frac{Entropy(D) \sum_{j=1}^l (P_j \cdot entropy(P_j))}{Splitting_{info}}. \quad (3)$$

4.3. Random Forest. It produces a set of techniques that involve constructing an ensemble or termed as a forest of decision trees from a randomized variation of tree induction techniques [32]. RF works by forming a mass of decision trees at the training period and harvesting the class in the approach of the class output by a single tree [33]. It is deliberated as one of the utmost techniques which is extremely proficient for both classification and regression problems.

4.4. Multilayer Perceptron. MLPs are deliberated as the utmost momentous classes of the neural network including an input layer, output layer, and least one hidden layer [34–36]. The techniques behind the neural network are that when data are accessible as the input layer, the network neurons start calculation in the sequential layer until an output value is gained at each of the output neurons. A threshold node is moreover added in the input layer which identifies the weight function. The resultant calculations are used to gain the activity of the neurons by smearing a sigmoid activation function that can be defined as

$$P_j = \sum_{i=1}^n w_{j,i} x_i + \theta_j, m_j = f_j(p_j), \quad (4)$$

where P_j is the linear combination of inputs x_1, x_2, \dots, x_n , θ_j is the threshold, $w_{j,i}$ is the connection weight between x_i and neuron j , f_j is the activation function of the j^{th} neuron, and m_j is the output. A sigmoid function is a mutual choice of activation function that can be described as

$$f(t) = \frac{1}{1 + e^{-t}}. \quad (5)$$

4.5. Radial Basis Function. It is also a neural network model that needs a very few computational time for training a network [37, 38]. Likewise, MLP also contains input, hidden, and output layers. The input variables in the input layer permit straight to the hidden layer deprived of weights. The transfer functions of the hidden knobs are RBFs, which factors are elevated throughout the training. The process of appropriating RBFs to data, for function of rough calculation, is thoroughly associated with space-weighted regression.

4.6. Hidden Markov Model. HMM is a probabilistic or [39] a statistical Markov model where the scheme being modeled is probable to be a Markov procedure using unobservable

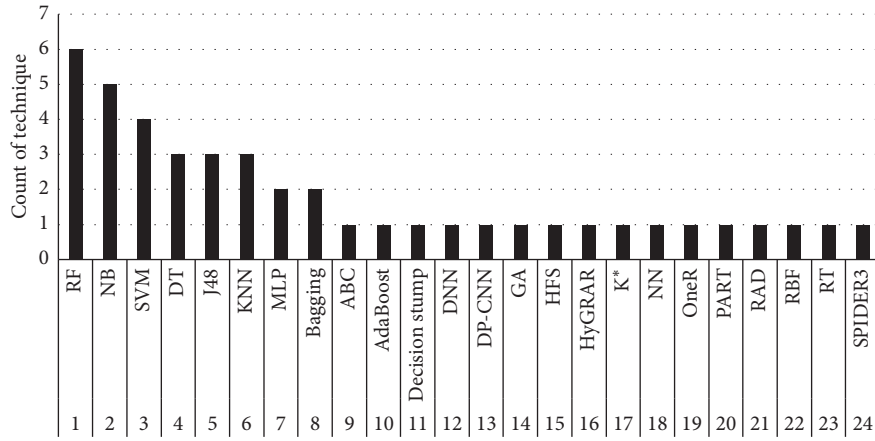


FIGURE 4: Count of each technique.

states or hidden statuses. It can be epitomized as the gentlest dynamic Bayesian network. It is reliant on splitting large data into the smallest sequences of data using a fewer sensitive pairwise sequence comparison method [40]. This model can be reflected in the generality of a combination model where the hidden variables that control the combination section to be nominated for every statement are connected through a Markov process moderately than liberated from each other. HMMs are particularly identified for their use in reinforcement learning and chronological pattern recognition such as speech, handwriting, part-of-speech tagging, gesture recognition, partial discharges, musical score following, and bioinformatics [39, 41].

4.7. Credal Decision Tree. Credal decision trees (CDTs) are algorithms to design classifiers grounded on inexact possibilities and improbability measures [42]. Throughout the creation procedure of a CDT, to sidestep producing a very problematical decision tree, a new standard was presented: stay once the total improbability rises due to splitting of the

decision tree. The function utilized in the total hesitation dimension can be fleetingly articulated as [43, 44]

$$TU(\xi) = IG(\xi) + GG(\xi), \quad (6)$$

where ξ is a Credal fixed on frame X , TU is the value of total hesitation, IG represents a common function of non-specificity on the resultant Credal set, and GG is a common function of arbitrariness for a Credal set.

4.8. Average One Dependency Estimator. A1DE is a probabilistic technique used for mostly classification problems. It succeeds extremely precise classification by averaging inclusive of a minor space of different NB-like models that have punier independence suppositions than NB. A1DE was designed to address the attribute-independence issues of a popular NB technique. It was designed to address the attribute-independence issues of the prevalent naive Bayes classifier. A1DE pursues to estimate the possibility of every class y assumed a quantified set of features $x_1, x_2, \dots, x_n, P(y|x_1, \dots, x_n)$ [45]. This can be calculated as

$$\hat{P}(y|x_1, x_2, \dots, x_n) = \frac{\sum_i: 1 \leq n \wedge F(x_i) \geq m \hat{P}(y, x_i) \prod_{j=1}^n \hat{P}(y, x_j)}{\sum_{y' \in Y} \sum_i: 1 \leq i \leq n \wedge F(x_i) \geq m \hat{P}(y', x_i) \prod_{j=1}^n \hat{P}(x_i|y', x_j)}, \quad (7)$$

where $\hat{P}(\cdot)$ represents an assessment of $P(\cdot)$, $F(\cdot)$ is the frequency through which the influences seem in the trial data, and m is a user quantified least frequency by which a term essentially seems in direction to be utilized in outer summation. Currently, m is the habitually set at 1.

4.9. Naïve Bayes. NB is a kinfolk of modest probabilistic technique grounded on Bayes theorem with unconventionality suppositions amid the predictors [46, 47]. The NB model is precise simple to construct and can be executed for any dataset containing a large amount of data. The posterior probability, $P(c|x)$, is taken from $P(c)$, $P(x)$, and $P(x|c)$.

The consequence of the value of a forecaster (x) on assumed class (c) is independent of the value of other forecasters.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \text{ or} \quad (8)$$

$$\text{Posterior} = \frac{\text{Prior}^* \text{likelihood}}{\text{Evidence}}$$

4.10. K-Nearest Neighbor. KNN is a supervised learning technique where the preparation of features attributes to forecast the class of new test data. KNN classifies first-hand

data grounded on the least distance from the new data to the K -nearest neighbors [48, 49]. The nearest distance can be found using different distance functions such as Euclidean distance (ED), Manhattan distance (MD), and Minkowski distance (MkD). Here, in this study, ED is used that can be formulated as

$$d(X, Y) = \sqrt{\sum_{i=1}^k (x_i - y_i)^2}, \quad (9)$$

where $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_3)$.

5. Experimental Results

5.1. Results and Analysis. This section provides an experimental study for SDP employing ten ML techniques using a standard approach of the 10-fold cross-validation process for assessment [34]. This process splits the complete data into ten subgroups of equal sizes; one subgroup is used for testing, whereas the rest of the subgroups are used for training. This process is continuing until each subgroup has been used for testing.

In this work, we considered seven different software defect datasets named AR1, AR3, CM1, JM1, KC2, KC3, and MC1. Using these datasets, we apply a software defect prediction system where the performance of all employed ML techniques is compared with each other based on correctly and incorrectly classified instances, true-positive and false-positive rates, MAE, RAE, RMSE, RRSE, recall, and accuracy. Table 5 presents the benchmark analysis of correctly classified instances (CCI), while Table 6 presents the benchmark analysis of incorrectly classified instances (ICI) using ML techniques. In both tables, the first column represents techniques employed, while the rest of the columns show details of each dataset concerning CCI and ICI. Figure 5 shows the inclusive performance CCI and ICI evaluation of each employed ML technique.

Table 7 illustrates the true-positive rate (TPR) and false-positive rate (FPR) of each technique on different hired datasets. TPR reveals the probability of the positive modules correctly classified, while FPR defines the probability of the negative modules incorrectly classified as the positive modules [5]. The first column of the table shows the list of datasets used, while the second column represents the TPR and FPR on the respective dataset. Apart from this, each row represents the achieved TPR and FPR concerning the individual dataset.

Tables 8 and 9 show the outcomes of absolute errors that are MAE and RAE, respectively. In each table, the first column represents the list of techniques, while the rest of the columns represent the error rate of each dataset concerning techniques employed. As shown in Table 8, while calculating MAE, SVM performs well in reducing the error rate as associated to other utilized techniques. SVM produces better results on five datasets, while MLP and NB produce better results only on two datasets. In the case of calculating RAE, SVM creates better results utilizing four datasets, while A1DE and NB do the same only for one dataset individually.

TABLE 5: Comparative analysis of correctly classified instances.

S. No.	Technique	AR1	AR3	CM1	JM1	KC2	KC3	MC1
1	SVM	111	56	446	7842	432	159	9398
2	J48	109	55	438	7668	425	154	9406
3	RF	109	58	444	7930	435	158	9417
4	MLP	109	59	436	7863	442	150	9411
5	RBF	111	55	446	7869	437	155	9398
6	HMM	112	55	449	1759	415	36	9398
7	CDT	112	55	445	7833	433	159	9407
8	A1DE	110	58	430	7816	435	155	9297
9	NB	103	57	425	7810	436	153	8913
10	KNN	109	54	422	7395	420	140	9418

TABLE 6: Comparative analysis of incorrectly classified instances.

S. No.	Technique	AR1	AR3	CM1	JM1	KC2	KC3	MC1
1	SVM	10	7	52	1751	90	35	68
2	J48	12	8	60	1925	97	40	60
3	RF	12	5	54	1663	87	36	49
4	MLP	12	4	62	1730	80	44	55
5	RBF	10	8	52	1724	85	39	68
6	HMM	9	8	49	7834	107	158	68
7	CDT	9	8	53	1760	89	35	59
8	A1DE	11	5	68	1777	87	39	169
9	NB	18	6	73	1783	86	41	553
10	KNN	12	9	76	2198	102	54	48

This determines to calculate the absolute error, and SVM outperforms other techniques.

However, Tables 10 and 11 present the outcomes of each squared error that are RMSE and RRSE individually. Here, the outcomes of squared error are different than outcomes of absolute error. While calculating RMSE or RRSE in both cases, RF produces better results for three datasets that are JM1, KC3, and MC1, RBF for two datasets that are CM1 and KC2, whereas MLP and CDT for only one dataset separately that are AR3 and AR1, respectively. Although, this analysis shows the best performance of RF as compared to other employed ML techniques.

Table 12 shows the outcomes achieved using recall assessment measures. In this table, the first row represents the list of datasets, while the first column represents the list of employed techniques. The rest of the rows concerning individual techniques shows the outcomes utilizing each dataset. This table shows that calculating recall using the AR1 dataset, HMM, and CDT performs well and produces the same results of 0.926. Proceeding utilizing AR3 and KC2 datasets, MLP outperforms other techniques generating 0.937 and 0.847 correspondingly, while on CM1 and AR1 datasets, HMM and on KC3 and AR1 datasets CDT performs well while producing 0.926 and 0.902 results. Moreover, on MC1 and JM1 datasets, the results of RF are better as compared to other techniques that are 0.827 and 0.995 accordingly; while, on the KC3 dataset, SVM performance is better, that is, 0.82. Figure 6 presents the overall recall performance of ML techniques for datasets. It can be concluded that RF, MLP, HMM, and CDT have better performed in terms of recall.

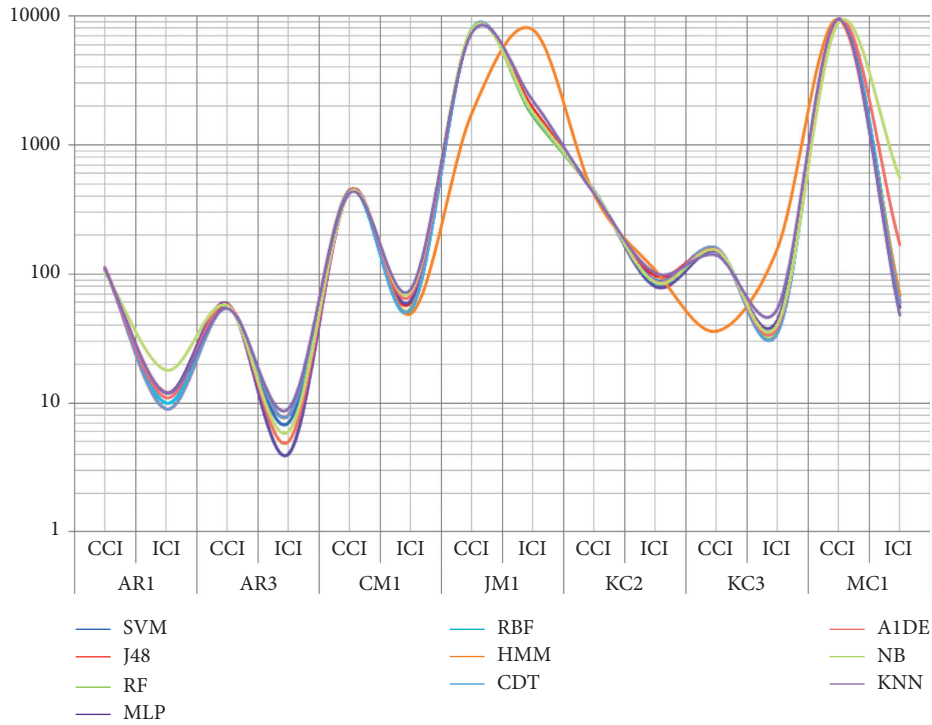


FIGURE 5: ML techniques performance comparison on CCI and ICI using SDP datasets.

TABLE 7: Comparative analysis of TPR and FPR of ML technique on different datasets.

Dataset		SVM	J48	RF	MLP	RBF	HMM	CDT	AIDE	NB	KNN
AR1	TPR	0.917	0.901	0.901	0.901	0.917	0.926	0.926	0.909	0.851	0.901
	FPR	0.926	0.723	0.928	0.723	0.926	0.926	0.926	0.927	0.523	0.621
AR3	TPR	0.889	0.873	0.921	0.937	0.873	0.873	0.873	0.921	0.905	0.857
	FPR	0.657	0.446	0.332	0.33	0.766	0.873	0.873	0.332	0.227	0.555
CM1	TPR	0.896	0.88	0.892	0.876	0.896	0.902	0.894	0.863	0.853	0.847
	FPR	0.902	0.849	0.848	0.886	0.902	0.902	0.902	0.869	0.616	0.762
JM1	TPR	0.817	0.799	0.827	0.82	0.82	0.183	0.817	0.815	0.814	0.771
	FPR	0.812	0.631	0.635	0.77	0.757	0.183	0.695	0.662	0.658	0.551
KC2	TPR	0.828	0.814	0.833	0.847	0.837	0.795	0.83	0.833	0.835	0.805
	FPR	0.634	0.422	0.431	0.435	0.472	0.795	0.439	0.424	0.473	0.432
KC3	TPR	0.82	0.794	0.814	0.773	0.799	0.186	0.82	0.789	0.789	0.722
	FPR	0.792	0.562	0.707	0.609	0.797	0.186	0.663	0.561	0.52	0.728
MC1	TPR	0.993	0.994	0.995	0.994	0.993	0.993	0.994	0.982	0.942	0.995
	FPR	0.993	0.701	0.657	0.73	0.993	0.993	0.774	0.628	0.38	0.496

TABLE 8: Comparative analysis of MAE.

S. No.	Technique	AR1	AR3	CM1	JM1	KC2	KC3	MC1
1	SVM	0.826	0.1111	0.1044	0.1825	0.1724	0.1804	0.0072
2	J48	0.127	0.1606	0.1757	0.2573	0.2374	0.2372	0.01
3	RF	0.127	0.1479	0.1631	0.2479	0.2205	0.257	0.0083
4	MLP	0.1037	0.1101	0.1568	0.2569	0.2259	0.2371	0.0072
5	RBF	0.1556	0.1812	0.1816	0.2773	0.2395	0.2995	0.025
6	HMM	0.5	0.5	0.5	0.5	0.5	0.5	0.5
7	CDT	0.1378	0.209	0.1745	0.2633	0.2296	0.2802	0.0112
8	AIDE	0.157	0.105	0.1886	0.2591	0.197	0.2708	0.0258
9	NB	0.1519	0.1085	0.1524	0.1863	0.1638	0.2162	0.059
10	KNN	0.1044	0.155	0.155	0.2319	0.2114	0.2809	0.0063

The bold values in the table indicate the reduced error rate.

TABLE 9: Comparative analysis of RAE.

S. No.	Technique	AR1	AR3	CM1	JM1	KC2	KC3	MC1
1	SVM	57.249	47.908	58.379	60.9388	52.7775	59.2313	49.9624
2	J48	87.9769	69.248	98.2132	85.9021	72.675	77.8877	69.4547
3	RF	87.9611	63.4368	91.1945	82.7532	67.4929	84.3792	57.6946
4	MLP	71.8266	47.4611	87.6482	85.7559	69.135	77.8521	49.9963
5	RBF	107.8175	78.1281	100.974	92.5753	73.3103	98.3279	174.0763
6	HMM	346.3562	215.586	279.5455	166.9291	153.0549	164.1553	3477.5284
7	CDT	95.4752	90.1037	97.5893	87.9121	70.29	91.9819	78.072
8	A1DE	108.7465	43.7714	105.4305	86.5138	60.3	88.8947	179.4312
9	NB	105.249	46.7686	85.2218	62.2139	50.1471	70.9899	410.217
10	KNN	72.3151	66.846	86.6364	77.4311	64.7095	92.209	44.0253

The bold values in the table indicate the reduced error rate.

TABLE 10: Comparative analysis of RMSE.

S. No.	Technique	AR1	AR3	CM1	JM1	KC2	KC3	MC1
1	SVM	0.2875	0.3333	0.3231	0.4272	0.4152	0.4247	0.0848
2	J48	0.2997	0.3424	0.3301	0.4053	0.3968	0.43	0.0779
3	RF	0.2856	0.2724	0.2951	0.3577	0.349	0.3667	0.0669
4	MLP	0.2882	0.256	0.3121	0.3706	0.3419	0.4414	0.0754
5	RBF	0.2664	0.2939	0.2919	0.3683	0.3413	0.3879	0.0837
6	HMM	0.5	0.5	0.5	0.5	0.5	0.5	0.5
7	CDT	0.2627	0.3377	0.3046	0.3752	0.3627	0.3818	0.0772
8	A1DE	0.2931	0.2925	0.3183	0.3754	0.3554	0.4034	0.1184
9	NB	0.3733	0.3176	0.38	0.4291	0.4019	0.4546	0.24
10	KNN	0.3122	0.3719	0.3905	0.475	0.4427	0.5246	0.0712

The bold values in the table indicate the reduced error rate.

TABLE 11: Comparative analysis of RRSE.

S. No.	Technique	AR1	AR3	CM1	JM1	KC2	KC3	MC1
1	SVM	109.405	99.6674	108.4851	110.4067	102.8529	109.2121	100.3607
2	J48	114.0496	102.3912	110.822	104.7491	98.2924	110.5573	92.2254
3	RF	108.6878	81.4368	99.0872	92.4278	86.4543	94.2824	79.2174
4	MLP	109.6657	76.5306	104.785	95.7816	84.6955	113.4827	89.3325
5	RBF	101.3862	87.8669	97.9878	95.1733	84.5435	99.7454	99.0846
6	HMM	190.2829	149.5011	167.8622	129.2111	123.8513	128.5606	592.0558
7	CDT	99.9593	100.9585	102.2522	96.9708	89.8344	98.1628	91.4393
8	A1DE	111.5568	87.4683	106.8639	97.0031	88.0436	103.724	140.1728
9	NB	142.0683	94.9565	127.572	110.8776	99.5502	116.8883	284.2012
10	KNN	118.7971	111.1859	131.084	122.7563	109.6576	134.8914	84.3477

The bold values in the table indicate the reduced error rate.

TABLE 12: Comparative analysis of recall.

S. No.	Technique	AR1	AR3	CM1	JM1	KC2	KC3	MC1
1	SVM	0.917	0.889	0.896	0.817	0.828	0.82	0.993
2	J48	0.901	0.873	0.88	0.799	0.814	0.794	0.994
3	RF	0.901	0.921	0.892	0.827	0.833	0.814	0.995
4	MLP	0.901	0.937	0.876	0.82	0.847	0.773	0.994
5	RBF	0.917	0.873	0.896	0.82	0.837	0.799	0.993
6	HMM	0.926	0.873	0.902	0.183	0.795	0.186	0.933
7	CDT	0.926	0.873	0.894	0.817	0.83	0.82	0.994
8	A1DE	0.909	0.921	0.863	0.815	0.833	0.799	0.982
9	NB	0.851	0.905	0.853	0.814	0.835	0.789	0.942
10	KNN	0.901	0.857	0.847	0.771	0.805	0.722	0.995

The bold values in the table indicate the highest recall in each column.

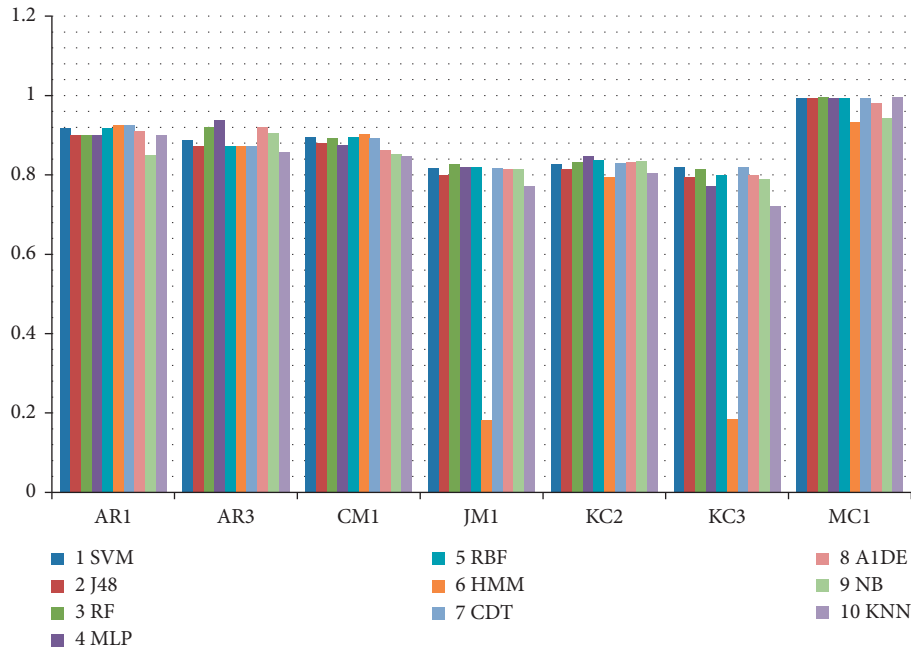


FIGURE 6: Recall comparison of ML technique using an individual dataset.

TABLE 13: The results of different techniques in terms of accuracy along with the rank values (it ranks the technique for each dataset separately, the best performing algorithm getting the rank of 1 and the second-best rank 2. Last two columns present the sum and average of ranks for each technique.).

	SVM	J48	RF	MLP	RBF	HMM	CDT	A1DE	NB	KNN
AR1	91.73 (2.5)	90.08 (4.25)	90.08 (4.25)	90.08 (4.25)	91.73 (2.5)	92.56 (1.5)	92.56 (1.5)	90.90 (3)	85.12 (5)	90.08 (4.25)
AR3	88.88 (5)	87.30 (6.33)	92.06 (3.5)	93.65 (2)	87.30 (6.33)	97.30 (1)	87.30 (6.33)	92.06 (3.5)	90.47 (4)	85.71 (7)
CM1	89.55 (2.5)	87.95 (5)	89.15 (4)	87.55 (6)	89.55 (2.5)	90.16 (1)	89.35 (3)	86.34 (7)	85.34 (8)	84.39 (9)
JM1	81.74 (4)	79.93 (8)	82.66 (1)	81.96 (3)	82.02 (2)	18.33 (10)	81.65 (5)	81.47 (6)	81.41 (7)	77.08 (9)
KC2	82.75 (6)	81.41 (7)	83.33 (4.5)	84.67 (1)	83.71 (2)	79.50 (9)	82.95 (5)	83.33 (4.5)	83.52 (3)	80.45 (8)
KC3	81.95 (1.5)	79.38 (4)	81.44 (2)	77.31 (7)	79.89 (3.5)	18.55 (9)	81.95 (1.5)	79.89 (3.5)	78.86 (6)	72.16 (8)
MC1	99.28 (5.33)	99.36 (4)	99.48 (1.5)	99.41 (2)	99.28 (5.33)	99.28 (5.33)	99.37 (3)	98.21 (6)	94.15 (7)	99.49 (1.5)
Sum (accuracy)	615.93	605.43	618.23	614.66	613.52	495.70	615.16	612.24	598.90	589.39
Average (accuracy)	87.99	86.49	88.32	87.81	87.65	70.81	87.88	87.46	85.56	84.20
Sum (rank)	26.83	38.58	20.75	25.25	24.16	36.83	25.33	33.50	40.00	46.75
Average (rank)	3.83	5.51	2.96	3.61	3.45	5.26	3.62	4.79	5.71	6.68

Table 13 shows the accuracy performance of each employed technique using different datasets. In this table, the first column represents the list of techniques, whereas the first row represents the list of datasets. The rest of the columns and rows show the outcome of each technique utilizing every dataset. Amid all the outcomes, the better performance of each technique under the individual dataset is listed in bold as shown in Table 13. This analysis shows that HMM produces better accuracy on three datasets, namely, AR1, AR3, and CM1, and outcomes are 92.562%, 97.3016%, and 90.1606%, respectively. RF harvests better accuracy on JM1 and near to best on MC1, that is,

82.6644% and 99.4824%, while SVM and MLP create better accuracy for KC3 and KC2, that is, 81.9588% and 84.6743%, respectively. Utilizing the MC1 dataset, A1DE outperforms other techniques achieving the accuracy of 99.4929%. The clinched performance of all techniques on individual datasets is presented in Figure 7.

Our outcomes suggest that there is uncertainty in the ML techniques. No individual technique performs well on every dataset. Different assessment measures are utilized to test the performance of each ML techniques on every dataset. Table 14 also presents the ranking of each technique, where we

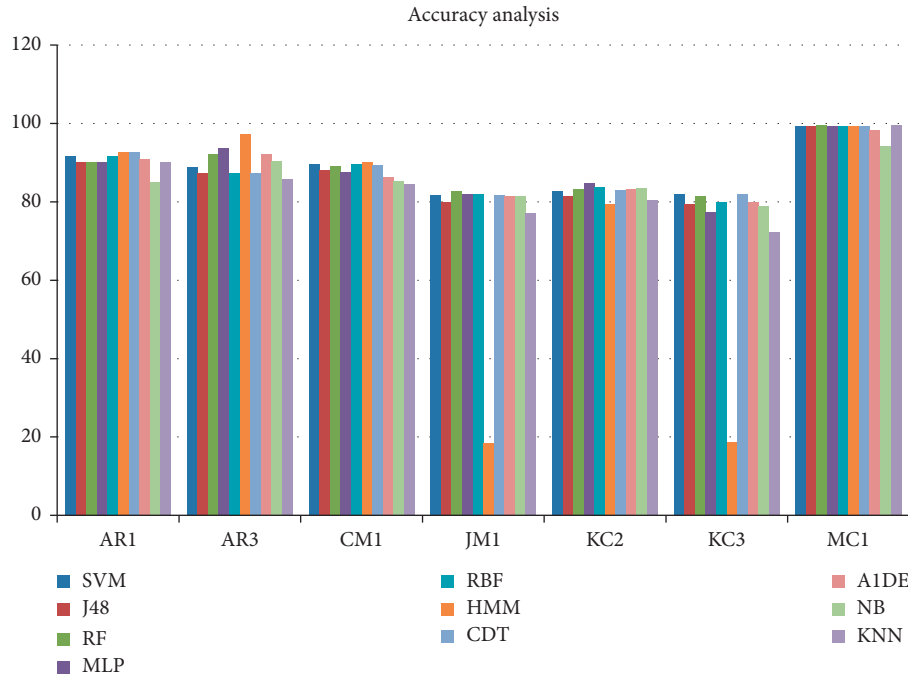


FIGURE 7: Accuracy comparison of ML technique using an individual dataset.

can see that HMM produces better results on 3 datasets; this number is maximum from the better results produced by any other techniques. However, on average, RF produces better results (average rank = 2.96), and the KNN produced poor results (average rank = 6.68). This is due to RF produces the forest with several trees [33, 50]. Overall, the more trees in the forest, the more forceful the forest resembles. Likewise in the RF classifier, the large amount of trees in the forest causes to give higher accuracy results [51, 52].

To get insight into the number, Table 13 shows the overall decision for SDP utilizing ML techniques on AR1, AR3, CM1, JM1, KC2, KC3, and MC1 datasets. This table concludes that which technique performs well on an individual dataset to a specific assessment criterion.

A standard approach to benchmark the performances of classifiers is to count (w) the number of datasets on which an algorithm is an overall subjugator, also known as the Count of Wins test. We have used 7 datasets, and no technique has given the best results for at least 7 datasets at $\alpha = 0.05$, according to the critical values in Table 3 of [53]. Since the Count of Wins test is also considered to be a weak testing procedure, therefore, we have a detailed matrix Table 14. As it can be observed from the very first dataset from Table 14, that is AR1, CDT outperforms other techniques in terms of increasing accuracy and reducing squared error while reducing absolute errors; MLP and SVM also perform well. On second and third datasets such as AR3 and CM1, HMM outperforms other techniques in terms of increasing accuracy; however, reducing the error rate on the AR3 dataset, MLP and A1DE produces better results, and utilizing the CM1 dataset, SVM and RBF performs well. Moreover, using JM1 and MC1, RF and KNN produce better results in terms of increasing accuracy and decreasing squared error rate, while decreasing absolute error SVM and KNN outperform well. Furthermore, on the KC2 dataset, MLP

performs well in increasing accuracy, and using the KC3 dataset, SVM performs well. However, on KC2 and KC3, SVM, RF, RBF, and NB performance is better in terms of reducing error rates.

All the employed techniques perform well certain in terms of reducing error rate, while some in terms of increasing accuracy, excluding J48. J48 is an insecure technique, for data containing categorical variables with a diverse number of altitudes as we have in employed datasets, and information gain in the decision tree is unfair in service of those metrics with more levels and fairly imprecise [54]. The performance of every individual technique is different on each singular dataset, which is due to the change of population in each dataset as well as differences between the values range and a number of attributes.

5.2. Friedman Two-Way Analysis of Variance by Ranks.

To compare all applied ML techniques on numerous datasets, we have smeared the statistical technique as defined by Sheskin [55] and García [56]. The Friedman two-way analysis of difference by ranks (Friedman) [57] is adopted with rank-order data in a hypothesis testing condition. A significant test specifies that there is a significant variance amid at least two of the techniques in the set of k techniques. The Friedman test checks whether the measured average ranks are significantly different from the mean rank (in our case, $R_j = 4.54$). The chi-square (χ^2) distribution is used to approximate the Friedman test statistic [55]. Friedman's statistic is

$$\chi^2 = 63.218. \quad (10)$$

To throw away the null hypothesis, the workout value must be equal to or greater than χ^2 , the tabled (table of the

TABLE 14: Decision table.

Datasets	Evaluation measures					
	MAE	RAE	RMSE	RRSE	Recall	Accuracy
AR1	MLP	SVM	CDT	CDT	HMM, CDT	HMM, CDT
AR3	A1DE, NB	A1DE	MLP	MLP	MLP	HMM
CM1	SVM	SVM	RBF	RBF	HMM	HMM
JM1	SVM	SVM	RF	RF	RF	RF
KC2	NB, SVM	NB	RBF	RBF	MLP	MLP
KC3	SVM	SVM	RF	RF	SVM	SVM
MC1	KNN, SVM, MLP	KNN	RF	RF	RF	KNN, RF

TABLE 15: Family of hypotheses ordered by the P value and adjusting α by Nemenyi and Holm's procedures, considering an initial $\alpha = 0.05$.

S. No.	Algo versus algo		z	P	NM (0.05)	Holm	$R_i - R_j$	CD
1	RF	KNN	14.8740	$6.07E-08$	0.001	0.0011	3.7143	>
2	RBF	KNN	12.9232	$2.04E-07$	0.001	0.0011	3.2271	>
3	MLP	KNN	12.2997	$3.12E-07$	0.001	0.0012	3.0714	>
4	CDT	KNN	12.2539	$3.22E-07$	0.001	0.0012	3.0600	>
5	SVM	KNN	11.3958	$5.97E-07$	0.001	0.0012	2.8457	>
6	RF	NB	11.0125	$7.97E-07$	0.001	0.0013	2.7500	>
7	J48	RF	10.2001	$1.52E-06$	0.001	0.0013	2.5471	>
8	RF	HMM	9.1990	$3.57E-06$	0.001	0.0013	2.2971	>
9	RBF	NB	9.0617	$4.04E-06$	0.001	0.0014	2.2629	>
10	MLP	NB	8.4381	$7.21E-06$	0.001	0.0014	2.1071	>
11	CDT	NB	8.3924	$7.53E-06$	0.001	0.0014	2.0957	>
12	J48	RBF	8.2494	$8.65E-06$	0.001	0.0015	2.0600	>
13	J48	MLP	7.6258	$1.62E-05$	0.001	0.0015	1.9043	>
14	J48	CDT	7.5800	$1.7E-05$	0.001	0.0016	1.8929	>
15	A1DE	KNN	7.5800	$1.7E-05$	0.001	0.0016	1.8929	>
16	SVM	NB	7.5343	$1.78E-05$	0.001	0.0017	1.8814	>
17	RF	A1DE	7.2940	$2.3E-05$	0.001	0.0017	1.8214	>
18	RBF	HMM	7.2482	$2.41E-05$	0.001	0.0018	1.8100	>
19	SVM	J48	6.7219	$4.32E-05$	0.001	0.0019	1.6786	>
20	MLP	HMM	6.6247	$4.83E-05$	0.001	0.0019	1.6543	>
21	HMM	CDT	6.5789	$5.09E-05$	0.001	0.0020	1.6429	>
22	SVM	HMM	5.7208	0.000143	0.001	0.0021	1.4286	>
23	HMM	KNN	5.6750	0.000152	0.001	0.0022	1.4171	>
24	RBF	A1DE	5.3432	0.000233	0.001	0.0023	1.3343	>
25	MLP	A1DE	4.7196	0.000545	0.001	0.0024	1.1786	>
26	J48	KNN	4.6739	0.000581	0.001	0.0025	1.1671	>
27	CDT	A1DE	4.6739	0.000581	0.001	0.0026	1.1671	>
28	NB	KNN	3.8615	0.001919	0.001	0.0028	0.9643	>
29	SVM	A1DE	3.8158	0.002058	0.001	0.0029	0.9529	>
30	A1DE	NB	3.7185	0.002391	0.001	0.0031	0.9286	>
31	SVM	RF	3.4782	0.003479	0.001	0.0033	0.8686	>
32	J48	A1DE	2.9062	0.00871	0.001	0.0036	0.7257	>
33	RF	CDT	2.6201	0.013903	0.001	0.0038	0.6543	<
34	RF	MLP	2.5743	0.014987	0.001	0.0042	0.6429	<
35	RF	RBF	1.9508	0.041431	0.001	0.0045	0.4871	<
36	HMM	A1DE	1.9050	0.044585	0.001	0.0050	0.4757	<
37	HMM	NB	1.8135	0.051582	0.001	0.0056	0.4529	<
38	SVM	RBF	1.5274	0.080499	0.001	0.0063	0.3814	<
39	J48	HMM	1.0011	0.171458	0.001	0.0071	0.2500	<
40	SVM	MLP	0.9039	0.194805	0.001	0.0083	0.2257	<
41	SVM	CDT	0.8581	0.206548	0.001	0.0100	0.2143	<
42	J48	NB	0.8124	0.218775	0.001	0.0125	0.2029	<
43	RBF	CDT	0.6693	0.260042	0.001	0.0167	0.1671	<
44	MLP	RBF	0.6236	0.274195	0.001	0.0250	0.1557	<
45	MLP	CDT	0.0458	0.482248	0.001	0.0500	0.0114	<

chi-square distribution) precarious chi-square value at the prespecified level of significance [55]. The number of degrees of freedom $df = k - 1$. Thus, $df = 10 - 1 = 9$. For $df = 9$, the tabled critical $\alpha = 0.05$ and chi-square value $\chi^2 = 16.92$. Since the computed value = 63.218 is greater than $\chi^2_{0.05} = 16.92$, the alternative hypothesis is supported at $\alpha = 0.05$. It can be concluded that there is a significant difference among at least nine of the ten ML techniques. This result can be summarized as follows: $\chi^2_{0.05} (9) = 63.218, P < 0.05$.

Since the critical value is lower than χ^2 , we can continue with posthoc tests to spot the significant pairwise differences among all the techniques. The results are shown in Table 15, where z is the corresponding statistics and P values are for each hypothesis. Z is computed using the following equation:

$$z = \frac{(R_i - R_j)}{SE}, \quad (11)$$

where R_i is the i^{th} technique, and the standard error is $SE = \sqrt{(k(k+1))/6n} = 0.249$. Columns 5 and 6 represent Nemenyi's and Holm's static procedure. The second last column lists the differences between the average ranks of i^{th} and j^{th} techniques. While, the last column shows the critical difference (CD), and it states that the performance of the two techniques is expressively diverse if the consistent average ranks differ by at least the CD. CD can be assessed using

$$CD = q\alpha \sqrt{\frac{k(k+1)}{6n}}, \quad (12)$$

where critical values $q\alpha$ is given in (Table 5(b), Demsar 2006) [53]. The notations ">" and "<" represent whether the difference of the average rank ($R_i - R_j$) is greater or less than the value of CD, respectively. Greater means a significant difference between two means. Here, the value of CD is 0.692.

In Table 15, the family of hypotheses is ordered by their P values. As can be seen, Nemenyi's procedure rejects the first 27 hypotheses, whereas Holm's procedure also rejects the next 4 hypotheses; meanwhile, the corresponding P values are lesser than the adjusted NM- α 's and Holm. Hence, we conclude that the performance of MLP and CDT is comparable, and KNN has a lower performance. Besides, the obtained value $CD = 0.692$ specifies that any variance amid the average ranks of two techniques that is equal to or greater than 0.692 is significant. Concerning the pairwise comparisons in Table 15, the difference between the average ranks of two techniques which are greater than $CD = 0.692$ is the first 32. Thus, it can be concluded that there is a momentous alteration among the average ranks of the first 32 pairs of techniques.

6. Threats to Validity

This section converses the effects that could anguish the validity of this research work.

6.1. Internal Validity. The exploration of this study is grounded on diverse very familiar valuation standards that are used in the past in various studies. Amid these standards,

several are used to assess the error rate while certain used to assess the accuracy. So, the treat can be that the renewal of new valuation standards as a replacement for utilized standards may deteriorate the accuracy. Furthermore, the machine learning techniques used in this study may be replaced with other existing techniques and can be merged that can harvest enhanced outcomes than the employed techniques.

6.2. External Validity. We piloted investigations on various datasets. A threat to validity may arise if the projected techniques are related in the other actual data composed from the diverse software development organizations using surveys or replace these datasets with some other datasets, which may distress the outcomes while growing the error rates. Likewise, the projected technique might not be capable to harvest improved forecast in outcomes utilizing several other SDP datasets. Hence, this study concentrated on AR1, AR3, CM1, JM1, KC2, KC3, and MC1 datasets to measure the performance of the utilized techniques.

6.3. Construct Validity. Diverse ML techniques are benchmarked with each on various datasets on the base of several valuation standards. The assortment of techniques utilized in this study is on the canter of their progressive features over other techniques that ought to exploit by the researchers in the last decades. Though the threat can be that we put on several new techniques, at that point, it can be the probability that these new techniques can exhaust the projected techniques. Furthermore, the training and testing method is applied or we change the number of folds validation (increase or decrease) for the experimentations that can decrease the error rate. It moreover can be promising that using the newest valuation standards creates improved outcomes that can beat the current accomplished outcomes.

7. Conclusions

Nowadays SDP using ML techniques is dignified as one of the developing research zones. The identification of software defects at the primary phase of SDLS is a challenging task, as well it can subsidize the provision of high-quality software systems. This study focused on comparing seven famous ML techniques that are broadly used for SDP, on seven extensively used openly available datasets. The ML techniques include SVM, J48, RF, MLP, RBF, HMM, and CDT. The performance is evaluated utilizing different measures such as MAE, RAE, RMSE, RRSE, recall, and accuracy.

The experimental results have illustrated that NB and SVM produced fewer MAE and RAE, respectively. However, experimental results using RMSE, RRSE, recall, and accuracy showed that an average RF performed better. Friedman's two-way analysis of variance by ranks has performed on experimental results using accuracy. The Friedman test indicates that results are significant at $P < 0.05$. We also performed a pairwise statistical test which revealed that several pairs are significant. Moreover, a critical difference test showed that RF and KNN produced significantly different results at $P < 0.05$, where RF produced better while

KNN the poorest. The outcomes obtainable in this study may be recycled as the reference point for other studies and researchers, in such a way that the outcomes of any projected technique, model, or framework can be benchmarked and simply confirmed. For future works, class imbalance matters ought to be committed to these datasets. Furthermore, to increase the enactment, ensemble learning and feature selection techniques could also be explored.

Data Availability

The datasets used in this research are taken from UCI ML Learning Repository available at <https://archive.ics.uci.edu/>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] A. O. Balogun, A. O. Bajeh, V. A. Orie, and A. W. Yusuf-asaju, "Software defect prediction using ensemble learning: an ANP based evaluation method," *Journal of Engineering Technology*, vol. 3, no. 2, pp. 50–55, 2018.
- [2] T. Menzies, Z. Milton, B. Turhan, B. Cukic, Y. Jiang, and A. Bener, "Defect prediction from static code features: current results, limitations, new approaches," *Automated Software Engineering*, vol. 17, no. 4, pp. 375–407, 2010.
- [3] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, "A systematic literature review on fault prediction performance in software engineering," *IEEE Transactions on Software Engineering*, vol. 38, no. 6, pp. 1276–1304, 2012.
- [4] Z. Li, X.-Y. Jing, and X. Zhu, "Progress on approaches to software defect prediction," *IET Software*, vol. 12, no. 3, pp. 161–175, 2018.
- [5] H. Wang, "Software defects classification prediction based on mining software repository," 2014.
- [6] R. Malhotra and S. Kamal, "An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data," *Neurocomputing*, vol. 343, pp. 120–140, 2019.
- [7] D.-L. Miholca, G. Czibula, and I. G. Czibula, "A novel approach for software defect prediction through hybridizing gradual relational association rules with artificial neural networks," *Information Sciences*, vol. 441, pp. 152–170, 2018.
- [8] A. Alsaeedi and M. Z. Khan, "Software defect prediction using supervised machine learning and ensemble techniques: a comparative study," *Journal of Software Engineering and Applications*, vol. 12, no. 05, pp. 85–100, 2019.
- [9] A. Iqbal et al., "Performance analysis of machine learning techniques on software defect prediction using NASA datasets," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 300–308, 2019.
- [10] T. Menzies, A. Dekhtyar, J. Distefano, and J. Greenwald, "Problems with precision: a response to "comments on data mining static code attributes to learn defect predictors"," *IEEE Transactions on Software Engineering*, vol. 33, no. 9, pp. 637–640, 2007.
- [11] G. Czibula, Z. Marian, and I. G. Czibula, "Software defect prediction using relational association rule mining," *Information Sciences*, vol. 264, pp. 260–278, 2014.
- [12] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.
- [13] H. Alasker, S. Alharkan, W. Alharkan, A. Zaki, and L. S. Riza, "Detection of kidney disease using various intelligent classifiers," in *Proceedings of the 2017 3rd International Conference on Science in Information Technology: "Theory and Application of IT for Education, Industry, and Society in Big Data Era"*, Bandung, Indonesia, October 2017.
- [14] J. Li, P. He, J. Zhu, and M. R. Lyu, "Software defect prediction via convolutional neural network," in *Proceedings of the 2017 IEEE International Conference on Software Quality, Reliability & Security. QRS*, pp. 318–328, Prague, Czech Republic, July 2017.
- [15] S. Jacob and G. Raju, "Software defect prediction in large space systems through hybrid feature selection and classification," *International Arab Journal of Information Technology*, vol. 14, no. 2, pp. 208–214, 2017.
- [16] K. Bashir, T. Li, C. W. Yohannese, and Y. Mahama, "Enhancing software defect prediction using supervised-learning based framework," in *Proceedings of the 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Nanjing, China, November 2017.
- [17] C. Manjula and L. Florence, "Deep neural network based hybrid approach for software defect prediction using software metrics," *Cluster Computing*, vol. 22, no. S4, pp. 9847–9863, 2019.
- [18] B. Khan, R. Naseem, M. Ali, M. Arshad, and N. Jan, "Machine learning approaches for liver disease diagnosing," *International Journal of Data Science and Advanced Analytics*, vol. 1, no. 1, pp. 27–31, 2019.
- [19] S. A. Lauer, K. Sakrejda, E. L. Ray et al., "Prospective forecasts of annual dengue hemorrhagic fever incidence in Thailand, 2010–2014," *Proceedings of the National Academy of Sciences*, vol. 115, no. 10, pp. E2175–E2182, 2018.
- [20] K. Balasarayanan and M. Prakash, "Detection of dengue disease using artificial neural network based classification technique," *International Journal of Engineering and Technology*, vol. 7, no. 1, pp. 13–15, 2018.
- [21] 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, 2018.
- [22] C. G. Raji and S. S. Vinod Chandra, "Graft survival prediction in liver transplantation using artificial neural network models," *Journal of Computational Science*, vol. 16, pp. 72–78, 2016.
- [23] K. Morik, "Medicine: applications of machine learning," *Encyclopedia of Machine Learning and Data Mining*, Springer, Berlin, Germany, pp. 809–817, 2017.
- [24] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An empirical evaluation of machine learning techniques for chronic kidney disease prophecy," *IEEE Access*, vol. 8, pp. 55012–55022, 2020.
- [25] C. Davi, A. Pastor, T. Oliveira et al., "Severe dengue prognosis using human genome data and machine learning," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2861–2868, 2019.
- [26] M. M. Saritas, "Performance analysis of ANN and naive Bayes classification algorithm for data classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, 2019.
- [27] C. Wu, S.-C. Kao, C.-H. Shih, and M.-H. Kan, "Open data mining for Taiwan's dengue epidemic," *Acta Tropica*, vol. 183, pp. 1–7, 2018.

- [28] N. Nahar and F. Ara, "Liver disease prediction by using different decision tree techniques," *International Journal of Data Mining & Knowledge Management Process*, vol. 8, no. 2, pp. 01–09, 2018.
- [29] J. Chen, Y. Yang, K. Hu, Q. Xuan, Y. Liu, and C. Yang, "Multiview transfer learning for software defect prediction," *IEEE Access*, vol. 7, pp. 8901–8916, 2019.
- [30] S. Khan, R. Ullah, A. Khan, N. Wahab, M. Bilal, and M. Ahmed, "Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM)," *Biomedical Optics Express*, vol. 7, no. 6, p. 2249, 2016.
- [31] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "A systematic machine learning based approach for the diagnosis of non-alcoholic fatty liver disease risk and progression," *Scientific Reports*, vol. 8, no. 1, pp. 1–12, 2018.
- [32] A. N. Arbain and B. Y. P. Balakrishnan, "A comparison of data mining algorithms for liver disease prediction on imbalanced data," *International Journal of Data Science and Analytics*, vol. 1, no. 1, 2019.
- [33] A. Gulia, R. Vohra, and P. Rani, "Liver patient classification using intelligent techniques," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 4, pp. 5110–5115, 2014.
- [34] K. A. Otunaiya and G. Muhammad, "Performance of data-mining techniques in the prediction of chronic kidney disease," *Computer Science and Information Technology*, vol. 7, no. 2, pp. 48–53, 2019.
- [35] S. Chatterjee, N. Dey, F. Shi, A. S. Ashour, S. J. Fong, and S. Sen, "Clinical application of modified bag-of-features coupled with hybrid neural-based classifier in dengue fever classification using gene expression data," *Medical & Biological Engineering & Computing*, vol. 56, no. 4, pp. 709–720, 2018.
- [36] A. B. Nassif, D. Ho, and L. F. Capretz, "Towards an early software estimation using log-linear regression and a multi-layer perceptron model," *Journal of Systems and Software*, vol. 86, no. 1, pp. 144–160, 2013.
- [37] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, Article ID 100178, 2019.
- [38] K. Kesorn, P. Ongruk, J. Chompoosri et al., "Morbidity rate prediction of dengue hemorrhagic fever (DHF) using the support vector machine and the *Aedes aegypti* infection rate in similar climates and geographical areas," *PLoS One*, vol. 10, no. 5, pp. 1–16, 2015.
- [39] M. Stanke and S. Waack, "Gene prediction with a hidden Markov model and a new intron submodel," *Bioinformatics*, vol. 19, no. SUPPL. 2, pp. 215–225, 2003.
- [40] L. S. Johnson, S. R. Eddy, and E. Portugaly, "Hidden Markov model speed heuristic and iterative HMM search procedure," *BMC Bioinformatics*, vol. 11, 2010.
- [41] S. Z. Yu and H. Kobayashi, "An efficient forward-backward algorithm for an explicit-duration hidden Markov model," *IEEE Signal Processing Letters*, vol. 10, no. 1, pp. 11–14, 2003.
- [42] C. J. Mantas and J. Abellán, "Credal decision trees in noisy domains," in *Proceedings of the 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. ESANN 2014*, pp. 683–688, Bruges, Belgium, April 2014.
- [43] Q. He et al., "Novel entropy and rotation forest-based credal decision tree classifier for landslide susceptibility modeling," *Entropy*, vol. 21, no. 2, 2019.
- [44] J. Abellán and A. R. Masegosa, "An ensemble method using credal decision trees," *European Journal of Operational Research*, vol. 205, no. 1, pp. 218–226, 2010.
- [45] S. Picek, A. Heuser, and S. Guilley, "Template attack versus Bayes classifier," *Journal of Cryptographic Engineering*, vol. 7, no. 4, pp. 343–351, 2017.
- [46] A. Naik and L. Samant, "Correlation review of classification algorithm using data mining tool: WEKA, rapidminer, tanagra, orange and knime," *Procedia Computer Science*, vol. 85, pp. 662–668, 2016.
- [47] T. R. Baitharu and S. K. Pani, "Analysis of data mining techniques for healthcare decision support system using liver disorder dataset," *Procedia Computer Science*, vol. 85, pp. 862–870, 2016.
- [48] U. R. Acharya, H. Fujita, S. Bhat et al., "Decision support system for fatty liver disease using GLST descriptors extracted from ultrasound images," *Information Fusion*, vol. 29, pp. 32–39, 2016.
- [49] E. K. Hashi, M. S. Uz Zaman, and M. R. Hasan, "An expert clinical decision support system to predict disease using classification techniques," in *Proceedings of the ECCE 2017-International Conference on Electrical, Computer and Communication Engineering*, pp. 396–400, Cox's Bazar, Bangladesh, February 2017.
- [50] T. M. Carvajal, K. M. Viacrusis, L. F. T. Hernandez, H. T. Ho, D. M. Amalin, and K. Watanabe, "Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines," *BMC Infectious Diseases*, vol. 18, no. 1, pp. 1–15, 2018.
- [51] L. Lau, Y. Kankanige, B. Rubinstein et al., "Machine-learning algorithms predict graft failure after liver transplantation," *Transplantation*, vol. 101, no. 4, pp. e125–e132, 2017.
- [52] H. Jin, S. Kim, and J. Kim, "Decision factors on effective liver patient data prediction," *International Journal of Bio-Science and Bio-Technology*, vol. 6, no. 4, pp. 167–178, 2014.
- [53] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
- [54] H. Deng, G. Runger, and E. Tuv, "Bias of importance measures for multi-valued attributes and solutions," *Lecture Notes in Computer Science*, Springer, vol. 6792, pp. 293–300, Berlin, Germany, 2011.
- [55] D. J. Sheskin, "Parametric and nonparametric statistical procedures," 2000.
- [56] S. Garcia, "An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons," *Journal of Machine Learning Research*, vol. 9, pp. 2677–2694, 2008.
- [57] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, pp. 37–41, 2012.

Research Article

Research on Human Sports Rehabilitation Design Based on Object-Oriented Technology

Dandan Cao, Junyan Wang, and Naihong Liu 

Taiyuan University of Science and Technology, Taiyuan 030024, China

Correspondence should be addressed to Naihong Liu; liunaihong@tyust.edu.cn

Received 30 December 2020; Revised 16 January 2021; Accepted 26 February 2021; Published 4 March 2021

Academic Editor: Shah Nazir

Copyright © 2021 Dandan Cao 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.

In order to improve the effect of human motion rehabilitation, a design model of human motion rehabilitation based on object-oriented technology is proposed. The entire model design process includes the following steps. First, a visual dynamic tracking model for human motion rehabilitation is established, and then a fuzzy PID (Proportion Integration Differentiation) super-heterodyne control method is used to design the bone training control for human motion rehabilitation. The bone tracking control and adaptive training are under the control of object-oriented technology; it is analyzed by collecting human activity data during training. The 6-DOF kinematics problem of human movement rehabilitation is decomposed into the bone training control problem in the subspace. Combining object-oriented technology, visual blur recognition of human sports rehabilitation training, and adopting an adaptive kinematics model to design sports rehabilitation can improve the control convergence and global stability of the human sports rehabilitation process. The simulation results show that the method has a good overall steady state and the sports rehabilitation training effect is obvious.

1. Introduction

With the improvement of sports infrastructure and the development of sports, Chinese sports enthusiasm has been improved. However, in the process of rehabilitation training for injuries caused by various accidental conditions, the rehabilitation cycle of sports injuries is long due to the lack of systematic theoretical guidance and the corresponding knowledge base and it is easy to cause adverse physical injury sequelae. In order to promote the development of sports, it is necessary to systematize the theory of sports rehabilitation training. People do not know enough about sports injuries due to various accidents, especially the lack of systematic exercise rehabilitation training guidance; it is easy to cause the rehabilitation cycle to be too long and even leave the danger of the sequelae of physical injury [1]. Especially for students in school, the lively nature of young people makes them too keen on high-load sports such as football, basketball, boxing, and so on, which may cause a series of physical injuries in the process of exercise. There are many possibilities for sports injury, especially for the daily exercise

of students in school at present, the main reasons are overload exercise, non-standard exercise methods and inadequate sports protection. Especially for high-intensity sports, such as basketball, football, and boxing, young people's excessive passion can easily cause physical damage in the process of release, such as wear, dislocation, viscera damage, and so on [2].

Rehabilitation training for physical injury has become the consensus of sports protection. This is mainly due to the fact that a large number of practices have proved that the traditional "bed rest" method for sports injury cannot complete the rehabilitation effect of sports injury, or cause the recovery of sports injury to be slower [3]. The body will leave sequelae during rest, which will affect the body's functions. However, there are many problems in sports rehabilitation training, mainly due to the imperfection of its sports rehabilitation training system and the lack of basic theoretical knowledge of corresponding sports rehabilitation training, such as overload and discontinuity in its practical use. Lack of systematicness, poor pertinence, and too much dependence on objective requirements lead to poor effect of

sports rehabilitation training and even secondary damage to body function. Therefore, some systematic principles should be followed in the exercise rehabilitation training [4].

The design process of human motion rehabilitation is an object-oriented control process. The control of human body motion rehabilitation and human-machine motion planning are typical motion planning problems with multiple constraints in high-dimensional C-space. In order to improve the effect of human sports rehabilitation, this paper presents a design model of human sports rehabilitation based on object-oriented technology and constructs a visual dynamic tracking model of human sports rehabilitation [5]. The fuzzy PID hyperheterodyne control method is used to design the bone training control of human body movement rehabilitation, and the bone tracking control and adaptive training are carried out under the control of object-oriented technology [6]. The 6-DOF (degree of freedom) kinematics problem of human motion rehabilitation is decomposed into the skeletal training control problem in subspace, and the visual fuzzy recognition of human sports rehabilitation training is realized by combining the object-oriented technology. Adaptive kinematics model is used to design sports rehabilitation to improve the control convergence and global stability of human sports rehabilitation. Finally, the simulation results show the effectiveness of this method in improving the control stability of human sports rehabilitation process [7].

The research contributions of the thesis mainly include the following:

- (1) A design model of human motion rehabilitation based on object-oriented technology is proposed
- (2) Fuzzy PID (proportional integral derivative) superheterodyne control method is used to design the bone training control system for human movement rehabilitation
- (3) Combining object-oriented technology, visual blur recognition of human sports rehabilitation training, and adopting adaptive kinematics model to design sports rehabilitation

The rest of this paper is organized as follows. Section 2 discusses kinematics model of human sports rehabilitation and control constraint parameter analysis, followed by the control model optimization discussed in Section 3. Simulation experiment and result analysis are discussed in Section 4. Section 5 concludes the paper with summary and future research directions.

2. Kinematics Model of Human Sports Rehabilitation and Control Constraint Parameter Analysis

2.1. Skeletal System Modeling. When the man-machine system model was established in the AnyBody, the complexity of the organization was negatively correlated with the efficiency of the operation. The organization model should be simplified as much as possible. The exoskeleton model is simplified as a form in which a plurality of rods are

connected. The length of the rods is a fixed length and corresponds to the body segment parameters of the experimental subject, and a pedal is added to the exoskeleton to drive the ankle joint movement [8]. Draw a simplified 3D model and import it into AnyBody in STL format. Because the simulation mainly analyzes the muscle-related parameters of the lower extremities as the object of study, in order to improve the speed of operation, when setting up the mannequin, the upper extremity ignores and removes most of the trunk muscles. Adjust the initial position of the human model to contact the exoskeleton model. With AnyBody's criteria for degrees of freedom and constraints to adjust the connection between the human model and the exoskeleton model, it is connected in a manner close to the real situation [9]. The original skeletal system model and the improved skeletal system model are shown in Figures 1 and 2.

It can be seen from Figures 1 and 2 that the establishment of a simplified 3D model can well simulate the motion parameters of human body, which has a good guiding significance for the subsequent collection of human motion data.

2.2. Kinematics Model of Human Sports Rehabilitation.

Visual dynamic tracking model of human sports rehabilitation is built, and the kinematics model of human sports rehabilitation is established [10]. This paper assumes that the longitudinal movement of human body sport rehabilitation training process is symmetrical, and that the longitudinal movement is symmetrical in the process of longitudinal movement, in which the object-oriented technology is used to optimize the control design of human sports rehabilitation training [11]. The tilt control mechanism and the yaw operation mechanism of human body movement rehabilitation have no action. Given the initial configuration of human sports rehabilitation training model $\theta_{\text{start}} \in C_{\text{free}}$ (free C- space), object pose \mathbf{p}_{obj} , and feasible grab set \mathbf{g}_c , let the movement chain of human sports rehabilitation training model composed of waist and left (right) arm be described as $\{A^0, A^1\}$; the homogeneous matrix ${}^{i-1}T_i(q_i)$ between $i-1$ and $i-1$ can be expressed as [12]

$${}^{i-1}T_i(q_i) = \begin{bmatrix} c_i & -c_{\alpha_i}s_i & s_{\alpha_i}s_i & a_i c_i \\ s_i & c_{\alpha_i}c_i & s_{\alpha_i}c_i & a_i s_i \\ 0 & s_{\alpha_i} & c_{\alpha_i} & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (1)$$

The equations of longitudinal motion for human sports rehabilitation are obtained as follows:

$$m \frac{dV}{dt} = P \cos \alpha - X - mg \sin \theta, \quad (2)$$

$$mV \frac{d\theta}{dt} = P \sin \alpha + Y - mg \cos \theta, \quad (3)$$

$$J_z \frac{d\omega_z}{dt} + (J_y - J_x)\omega_y\omega_x + J_{xy}(\omega_y^2 - \omega_x^2) = M_z, \quad (4)$$

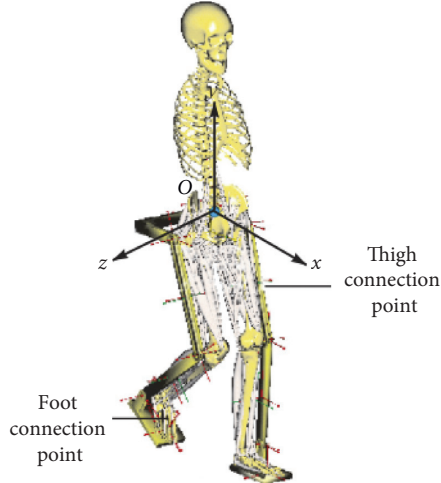


FIGURE 1: Skeletal system modeling.

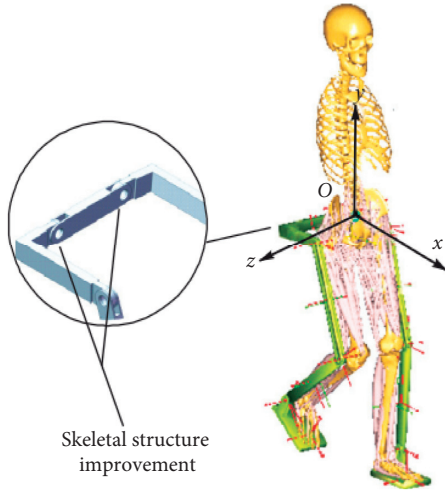


FIGURE 2: Exoskeletal structure after improvement.

$$\frac{dx}{dt} = V \cos \theta, \quad (5)$$

$$\frac{dy}{dt} = V \sin \theta, \quad (6)$$

$$\frac{d\vartheta}{dt} = \omega_z, \quad (7)$$

$$\alpha = \vartheta - \theta, \quad (8)$$

$$\delta_z = f(e_1). \quad (9)$$

Here, x , y are the position of centers of mass, ω_x , ω_y indicate that the angular velocity of XOY axis in the body coordinate system Ox_1 , Oy_1 mean the pitch angle of rehabilitation of human body, δ_z means the error of control system, and the quality of human body is indicated by the error of the control system and the error of the control system is indicated by the position of center of mass. C

indicates the resistance, lift, and lateral force acting on the human body for rehabilitation training, in which the waist X, Y are regarded as the root of the movement chain, and the forward kinematics equation of the right arm of the human body can be used for intelligent feature extraction by the right hand. It includes six degrees of freedom of rotation, such as deflection angle α_0 , pitch angle β_0 , and rolling angle γ_0 , which describe the motion of waist joint [13]. A six-degree-of-freedom control model of human sports rehabilitation training is constructed, which can be expressed as follows:

$$\mathbf{q}_0 = [\alpha_0, \beta_0, \gamma_0]^T \equiv [\theta_1, \theta_2, \theta_3]^T. \quad (10)$$

It can be seen that the longitudinal motion equation of human exercise rehabilitation training model is a group of dynamic systems composed of nonlinear differential equations. The elbow joint center of human sports rehabilitation training model is driven by mechanics through the wrist joint. The present kinematics model is constructed [14].

2.3. Analysis of Control Constraint Parameters for Human Sports Rehabilitation. Given the initial configuration of human sports rehabilitation training model $\theta_{\text{start}} \in \mathbf{C}_{\text{free}}$ (free C-space), pose \mathbf{p}_{obj} of objects, and feasible grab set \mathbf{g}_c , the equations corresponding to the elements of both sides of the two matrices of human sports rehabilitation around the arm can be solved as [15]

$$q_5 \equiv \theta_8 = a \tan 2(\pm o_{ey}, \pm o_{ex}), \quad (11)$$

$$q_6 \equiv \theta_9 = a \tan 2(-o_{ez}, -c_5 o_{ex} - s_5 o_{ey}), \quad (12)$$

$$q_7 \equiv \theta_{10} = a \tan 2(-s_5 n_{ex} + c_5 n_{ey}, s_5 a_{ex} - c_5 a_{ey}). \quad (13)$$

This paper deduces the analytical form of inverse kinematics of arm in the model of human exercise rehabilitation training and obtains six rotational degrees of freedom of the left (or right) arm A^1 including shoulder, elbow, and wrist, which are expressed as $\mathbf{q}_1 = [q_1, \dots, q_7]^T \equiv [\theta_4, \dots, \theta_{10}]^T$, and then obtains the exercise rehabilitation training of human body. The IK analytic equation of model control is used to realize the sixth-degree control model design of rehabilitation human sports rehabilitation training model. Combined with kinematics model [16], the constraint parameter model of skeletal training control for human exercise rehabilitation is constructed as follows:

$$\frac{\partial L}{\partial C_i} = 2MC_i + \eta \mathbf{1}, \quad (14)$$

where $M = \lambda S_i^T S_i + (X_i - D_i)^T (X_i - D_i) + \theta \mathbf{I}$.

In order to find the extremum, set $\partial L / \partial C_i = 0$; then

$$C_i = -\frac{1}{2} \eta M^{-1} \mathbf{1}. \quad (15)$$

Based on the object-oriented technology, the machine vision tracking recognition method is used to recognize the dynamic process of rehabilitation, and the visual dynamic

tracking model of human body movement rehabilitation is constructed to improve the dynamic control ability of rehabilitation training [17].

3. Control Model Optimization

3.1. Control Design of Bone Training for Human Body Movement Rehabilitation. The fuzzy PID superheterodyne control method is used to design the bone training control of human body movement rehabilitation, the bone tracking control and adaptive training of human body movement rehabilitation are carried out under the control of object-oriented technology, and the human body movement rehabilitation training is carried out [18]. The equivalent control law of model control is

$$u_{\text{eqx}} = \frac{\lambda \left(-\hat{f}_x - \lambda_x \dot{e}_x - \alpha e_x + \ddot{x} \right)}{(\lambda g_x + g_\theta)} \quad (16)$$

Some parameters of the control system of human sports rehabilitation training model are measured, and then the model is modified according to the measurement [19]. Several constraints of the model system are obtained as follows:

$$X_{\text{RL}} = R \times \theta_{\text{RL}}, \quad (17)$$

$$X_{\text{RR}} = R \times \theta_{\text{RR}}, \quad (18)$$

$$X_{\text{RL}} - X_{\text{RR}} = D \times \delta, \quad (19)$$

$$\dot{X}_p = \dot{\theta}_p L \cos \theta_p + \dot{X}_{\text{RM}}, \quad (20)$$

$$Y_p = L \cos \theta_p, \quad (21)$$

$$\dot{Y}_p = -\dot{\theta}_p L \sin \theta_p, \quad (22)$$

$$X_{\text{RR}} + X_{\text{RL}} = 2X_{\text{RM}}. \quad (23)$$

In order to eliminate the effect of parameter estimation on the stability of skeletal training for human exercise rehabilitation, the third Lyapunov function is chosen as follows:

$$V_3 = V_2 + \frac{\lambda_1 \zeta_1^2}{2} + \frac{\lambda_2 \zeta_2^2}{2} + \frac{\bar{\delta}^2}{2\varepsilon_1 \delta}. \quad (24)$$

In order to obtain the desired stability characteristics, the skeleton training process of human motion rehabilitation is controlled [20], and the nonlinear integral substitution control law is designed to limit the steady-state prediction error:

$$u = u_{\text{eqx}} + u_{\text{eq}\theta} + u_{\text{sw}}. \quad (25)$$

The inverse kinematics problem of 6-DOF human exercise rehabilitation is decomposed into two sub-inverse

kinematics problems with smaller dimensions. The adaptive regulation of skeletal training for human sports rehabilitation is chosen as follows:

$$V_1 = \frac{1}{2} e_1^2. \quad (26)$$

The derivation of the Lyapunov function for the skeletal training of human sports rehabilitation is

$$\dot{V}_1 = e_1 e_2 - c_1 e_1^2 - e_1 \lambda_1 \zeta_1. \quad (27)$$

By using the inverse design method and the combination of fuzzy control and adaptive control, the corresponding Lyapunov function is found and the derivative is obtained [21]:

$$\dot{e}_2 = \dot{\omega}_2 - \dot{\omega}_{2r}, \quad (28)$$

$$e_2 = \alpha V^2 + mg(\sin \vartheta + V \omega_2) + m(\cos \vartheta + V \omega_2) + c_1 e_2 + \lambda_1 e_1 - c_1^2 e_2 - c_1 \lambda_1 \zeta_1 - \ddot{\vartheta}_r. \quad (29)$$

3.2. Process Control Design of Rehabilitation Training. The control function of the mechanical tracking controller for human sports rehabilitation training is obtained as follows:

$$V_2 = V_1 + \frac{1}{2} e_2^2. \quad (30)$$

Derivative:

$$\dot{V}_2 = \dot{V}_1 + e_1 \dot{e}_2. \quad (31)$$

The inverse kinematics problem of 7-DOF right arm is decomposed and the control model of two degrees of freedom is obtained. The total degrees of freedom of skeletal training for human sports rehabilitation are 10. The configuration of human rehabilitation training can be expressed as $\theta = [\mathbf{q}_0^T, \mathbf{q}_1^T]^T \equiv [\theta_1, \dots, \theta_{10}]^T$. Set $\mathbf{q}_1 = [q_1, \dots, q_7]^T$. $\sin q_i$ and $\cos q_i$ recorded as s_{q_i} and c_{q_i} , respectively. The initial motion mechanics error compensation of the human body movement rehabilitation training model is S and the body posture error compensation of the training is $\theta_{\text{start}} = [\theta_{1\text{start}}, \dots, \theta_{10\text{start}}]^T$.

The design of control error compensation based on Lyapunov method is realized. According to Barbalat's theorem,

$$\lim_{t \rightarrow \infty} e_1 = \lim_{t \rightarrow \infty} e_2 = 0. \quad (32)$$

Above all, the adaptive kinematics model is used to design the sports rehabilitation; the control convergence and global stability are improved.

Through the above-mentioned design, the algorithm of the optimized control model of the lower extremity exoskeleton rehabilitation robot using the Lyapunov method and the inversion technique adaptive nonlinear tracking is obtained as follows.

Robot optimization control model algorithm:


```

(1) RRT.SetSConfig ( $\theta_{start}$ )//Set initial position
(2) do { //Perform a single tree RRT exercise planning cycle
(3)   NormalExtend = true//Set random expansion tags
(4)    $f_r = \text{rand} () * 1.0 / (0x7fff)$ //Generate 0~1 random sampling probability
(5)   if ( $f_r \leq f_{Extend}$ ){//Extension to body target pose
(6)     NormalExtend = false//Indicates the expansion to the target pose
(7)     ExtendStatus = ExtendToGrasp (RRT  $p_{obj}$   $g_c$ )
(8)   if (ExtendStatus == RRT_ERROR) { //Failed to expand
(9)     StopSearch = true//Set stop plan marker
(10)    if (ExtendStatus == RRT_REACHED) { //Achieve goals
(11)    RRT.GetConfig ( $\theta_{goal}$ )//Get the target configuration
(12)    RRT.GetSolutionGrasp ()//Get the lower limbs bone movement goal solution
(13)    FoundSolution = true//Search path solution}}
(14)    if (NormalExtnd) { //Random sampling configuration expansion
(15)    RRT.RandomConfig ( $\theta_{rand}$ )//Generate random patterns
(16)     $\theta_{near} = \text{RRT.NearestNeighbor} (\theta_{rand})$ //Recently shaped
(17)    ExtendStatus = RRT.Connect ( $\theta_{near}, \theta_{rand}, \theta_{new}$ )
(18)  if (ExtendStatus == RRT_ERROR){ //Failed to expand StopSearch = true} //Set stop plan marker
(19)    Cycles++//Number of searches plus 1
(20)  } while (!StopSearch && Cycles < MaxCycles && !FoundSolution)
(21)  return RRT.GetSolutionPath (); }

```

4. Simulation Experiment and Result Analysis

In order to test the application performance of this method in the control of human exercise rehabilitation training, the simulation experiment is carried out. The hardware design part of the system selects ARM11 CPU as the central processor. Select ARM11 CPU S3 C6410 as the hardware core. DM9000 network card chip is used. DM9000 supports 8-bit, 16-bit, and 32-bit interface to access internal memory, and the design of control algorithm is taken. In this paper, the control parameter is selected as $\lambda_1 = 1, \lambda_2 = 1, c_1 = 2, c_2 = 2$, the initial value of parameter adaptive estimation is $\hat{\delta}_0 = -15$, the adaptive parameter is $\varepsilon_1 = 0.1$, and the human

body is designed. The actual model parameters of sports rehabilitation training are as follows:

$$M_p = 1.6 \times 10^4 \text{kg}, m_r = 1.13 \times 10^4 \text{kg}, R = 2.05 \text{m}, l = 1.87 \text{m}, \quad (33)$$

$$K_m = 0.0508 \text{N.m/V}, K_e = 0.5732 \text{V}_s/\text{rad}, \quad (34)$$

$$J_p = 0.804(1 \pm 0.5) \text{kg} \cdot \text{m}^2, J_r = 0.00623(1 \pm 0.5) \text{kg} \cdot \text{m}^2. \quad (35)$$

The state of the human skeleton is affected by many factors. In order to be closer to the actual situation, the subject must be allowed to walk in advance before the experiment is conducted, and the subject can start the experiment until it reaches a state of relaxation. As the human body walks, the left and right are basically symmetrical. This experiment only uses the right lower limb of the human body as the object of analysis.

Through experiments, the joints of the hips, knees, and ankles of the right leg can be measured with time and other motion information at different speeds. The hip angle refers to the angle between the thigh and the horizontal plane, and the knee angle refers to the lower leg. Between the thighs, the angle of the ankle refers to the angle between the foot support surface and the lower leg. Taking a walking pace of 3.6 km/h as an example, after processing the data, the angle of each joint angle changes over time, as shown in Figure 3:

As you can see from Figure 3, the elevation angle of the robot's behavior in the assisted rehabilitation process is used as the evaluation index. The pitch tracking performance designed in this paper is shown in the figure. Using this algorithm, it has better robot control performance, and the control system can quickly track the input signal within 2 seconds without any control error. It has excellent anti-interference and robustness. Pitch elevation tracking of skeletal rehabilitation process is shown in Figure 4.

The human body model is mainly composed of muscles, bones, and ligaments. By introducing the acquired action capture data files into the AnyBody, the human walking motion can be simulated and analyzed by its walking simulation module. The simulation of the walking state is shown in Figure 5.

In the AnyBody simulation analysis, the kinematic analysis and the reverse dynamics analysis are used to obtain the kinematic information of the human body model and the information about the parameters of the muscles. The force of all muscles in the right leg during walking is obtained, as shown in Figure 6.

In the process of walking, not all bones are subjected to a large force. In order to study the force of the bones better and analyze the force of the muscles, only a part of the major muscles can be selected as the analysis object. From Figure 5, we can see that, in the course of walking in the last two

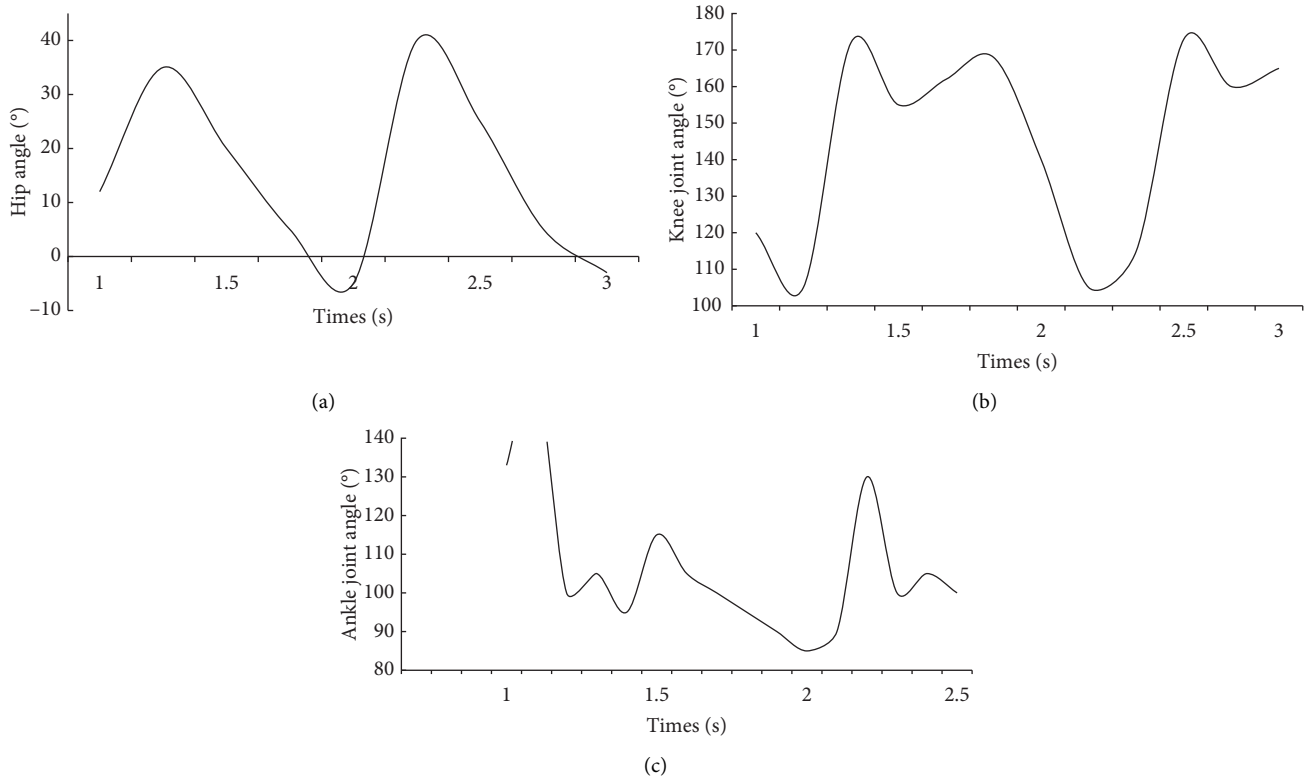


FIGURE 3: Motion state setting of bone. (a) Hip joints. (b) Knee joint. (c) Ankle joint.

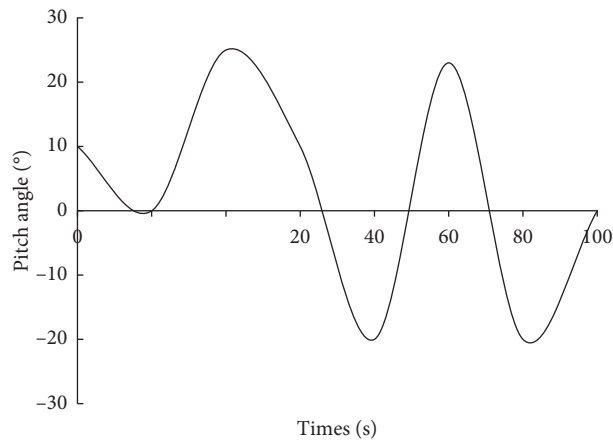


FIGURE 4: Pitch elevation tracking of skeletal rehabilitation process.

cycles, the muscle force of the right leg peaked at 2.15s and 2.90s at two time nodes, respectively, and the muscle recruitment was larger at this time.

On the basis of the design of the simulation environment, the control system is designed to train the rehabilitation training of human body movement, and the visual recognition output of the body movement rehabilitation training is shown in Figure 7.

Figure 7 shows that the visual recognition ability of this method is good, and it has a good object-oriented ability. The control stability of human sports rehabilitation process is further tested, and the test results are obtained. Control performance test is shown in Figure 8.

As shown in Figure 8, the control system can track the input signal quickly within 2 seconds, and there is no control error at the same time. The control stability of this method is

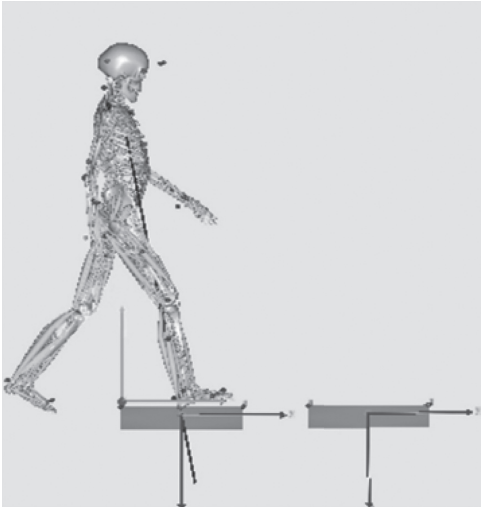


FIGURE 5: Walking state simulation.

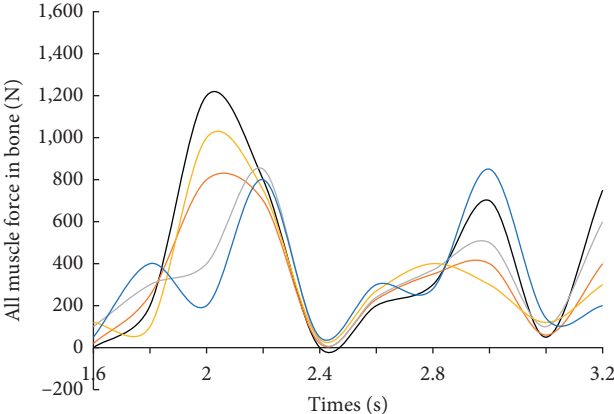


FIGURE 6: All muscle force of human's right leg.



FIGURE 7: Visual recognition of human sports rehabilitation training based on object-oriented technology.

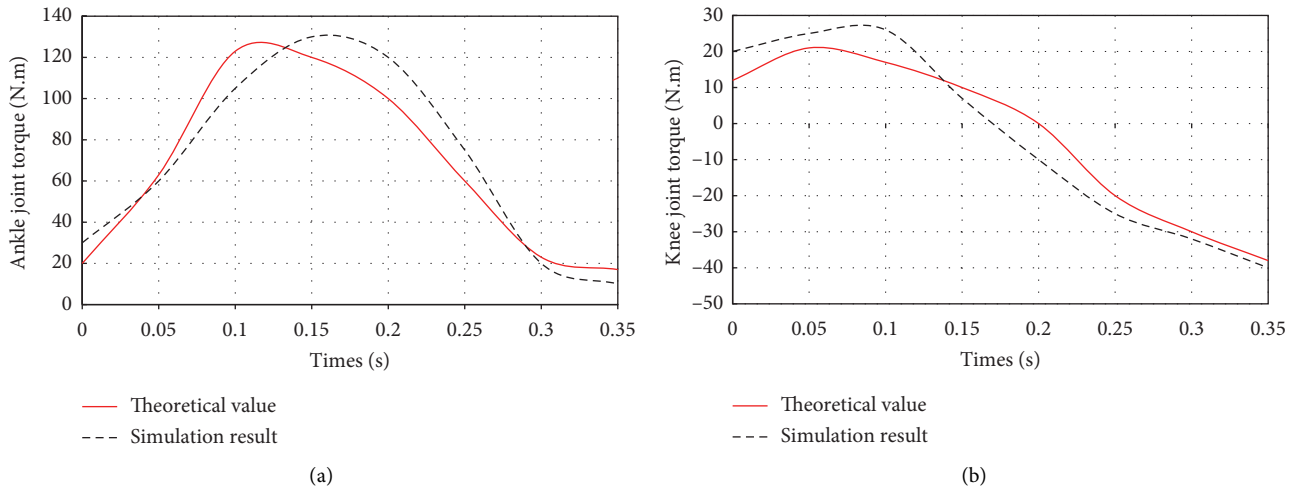


FIGURE 8: Control performance test. (a) Control simulation value of ankle joint rehabilitation training. (b) Control Simulation value of knee joint rehabilitation training.

good for the rehabilitation training of human body movement.

5. Conclusions

This paper presents a design model of human motion rehabilitation based on object-oriented technology. The research contributions of the thesis mainly include the following aspects:

- (1) A visual dynamic tracking model for human movement rehabilitation is established.
- (2) Fuzzy PID superheterodyne control method is used to design the bone training control of human movement rehabilitation.
- (3) Bone tracking control and adaptive training are carried out under the control of object-oriented technology, and relevant data are collected for research.

Combined with object-oriented technology, the visual blur recognition of human sports rehabilitation training and the use of adaptive kinematics model to design the sports rehabilitation process can improve the control convergence and global stability of the human sports rehabilitation process. The simulation results show that the method has a good overall steady state and the sports rehabilitation training effect is obvious. This method has good application value in guiding sports rehabilitation training.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Taiyuan University of Science and Technology (No. XJ2020136).

References

- [1] C. Hou, Z. Wang, Y. Zhazo, and G. Song, "Load adaptive force-free control for the direct teaching of robots," *ROBOT*, vol. 39, no. 4, pp. 439–448, 2017.
- [2] Y. You, Y. Zhang, and C. G. Li, "Force-free control for the direct teaching of robots," *Journal of Mechanical Engineering*, vol. 50, no. 3, pp. 10–17, 2014.
- [3] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.
- [4] D. Zhang, B. Li, and L. Wang, "Tracking control method of the centre-of-mass velocity for a Snake-like robot based on the continuum model," *ROBOT*, vol. 39, no. 6, pp. 829–837, 2017.
- [5] K. Li and M. I. Jie, "Research on mechanical and electrical control algorithm of bionic robot based on variable structure PID," *Journal of Henan University of Engineering (Natural Science Edition)*, vol. 28, no. 2, pp. 32–37, 2016.
- [6] X. Du, Y. Cai, T. Lu, S. Wang, and Z. Yan, "A robotic grasping method based on deep learning," *ROBOT*, vol. 39, no. 6, pp. 820–828, 2017.
- [7] Z. Ren, Q. G. Zhu, and R. Xiong, "A joint physical constraints avoidance method for inverse kinematics problem of redundant humanoid manipulator," *Journal of Mechanical Engineering*, vol. 50, no. 19, pp. 58–65, 2014.
- [8] F. Jing, C. Yang, and G. Yang, "TAN min. Robot trajectory rectification control methods," *ROBOT*, vol. 39, no. 3, pp. 292–297, 2017.
- [9] M. F. A. Abdullah, M. S. Sayeed, K. Sonai Muthu, H. K. Bashier, A. Azman, and S. Z. Ibrahim, "Face recognition with symmetric local graph structure (SLGS)," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6131–6137, 2014.
- [10] S. Xu, G. Li, J. Liu, and H. A. O. Jie, "Inverse kinematics solution of deformable manipulator for point touching task," *ROBOT*, vol. 39, no. 4, pp. 405–414, 2017.

- [11] Y. Hwang, T.-Y. Yu, V. Lakshmanan, D. M. Kingfield, D.-I. Lee, and C.-H. You, "Neuro-fuzzy gust front detection algorithm with S-band polarimetric radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 3, pp. 1618–1628, 2017.
- [12] En Shi, L. I. Qian, D. Gu, and Z. Zhao, "Weather radar echo extrapolation method based on convolutional neural networks," *Journal of Computer Applications*, vol. 38, no. 3, pp. 661–665, 2018.
- [13] T. D. Fletcher, H. Andrieu, and P. Hamel, "Understanding, management and modelling of urban hydrology and its consequences for receiving waters: a state of the art," *Advances in Water Resources*, vol. 51, no. 1, pp. 261–279, 2013.
- [14] B. Ma, X. Xie, and P. M. Psho-Hf-, "An efficient proactive spectrum handover mechanism in cognitive radio networks," *Wireless Personal Communications*, vol. 79, no. 3, pp. 1–23, 2014.
- [15] Y. Ji, Y. Li, and C. Shi, "Aspect rating prediction based on heterogeneous network and topic model," *Journal of Computer Applications*, vol. 37, no. 11, pp. 3201–3206, 2017.
- [16] R. C. Wade and A. S. Gorgey, "Skeletal muscle conditioning may be an effective rehabilitation intervention preceding functional electrical stimulation cycling," *Neural Regeneration Research*, vol. 11, no. 8, p. 1232, 2016.
- [17] K. Węgrzynowska-Teodorczyk, A. Siennicka, K. Josiak et al., "Evaluation of skeletal muscle function and effects of early rehabilitation during acute heart failure: rationale and study design," *Biomed Research International*, vol. 10, no. 5, pp. 1–8, 2018.
- [18] A. Boukerche and V. Soto, "Computation offloading and retrieval for vehicular edge computing," *Acm Computing Surveys*, vol. 6, 2020.
- [19] A. Beaufort, F. Moatar, E. Sauquet, P. Loicq, and D. M. Hannah, "Influence of landscape and hydrological factors on stream–air temperature relationships at regional scale," *Hydrological Processes*, vol. 34, 2020.
- [20] S. M. Alvarado and H. Feng, "Representation of dark skin images of common dermatologic conditions in educational resources: a cross-sectional analysis," *Journal of the American Academy of Dermatology*, vol. 8, 2020.
- [21] C. Alberto and K. Bartosz, "Evolving rule-based classifiers with genetic programming on GPUs for drifting data streams," *Pattern Recognition*, vol. 87, pp. 248–268, 2019.

Research Article

Human Gait Analysis and Prediction Using the Levenberg-Marquardt Method

Abdullah Alharbi,¹ Kamran Eqbal,² Sultan Ahmad ,³ Haseeb Ur Rahman,⁴ and Hashem Alyami⁵

¹Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia

²Biomedical Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India

³Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia

⁴Department of Computer Science & Information Technology, University of Malakand, Chakdara Dir Lower, Pakistan

⁵Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia

Correspondence should be addressed to Sultan Ahmad; s.alisher@psau.edu.sa

Received 6 January 2021; Revised 2 February 2021; Accepted 9 February 2021; Published 18 February 2021

Academic Editor: Shaukat Ali

Copyright © 2021 Abdullah Alharbi 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.

A high-accuracy gait data prediction model can be used to design prosthesis and orthosis for people having amputations or ailments of the lower limb. The objective of this study is to observe the gait data of different subjects and design a neural network to predict future gait angles for fixed speeds. The data were recorded via a Biometrics goniometer, while the subjects were walking on a treadmill for 20 seconds each at 2.4 kmph, 3.6 kmph, and 5.4 kmph. The data were then imported into Matlab, filtered to remove movement artifacts, and then used to design a neural network with 60% data for training, 20% for validation, and remaining 20% for testing using the LevenbergMarquardt method. The mean-squared error for all the cases was in the order of 10^{-3} or lower confirming that our method is correct. For further comparison, we randomly tested the neural network function with untrained data and compared the expected output with actual output of the neural network function using Pearson's correlation coefficient and correlation plots. We conclude that our framework can be successfully used to design prosthesis and orthosis for lower limb. It can also be used to validate gait data and compare it to expected data in rehabilitation engineering.

1. Introduction

In the entire world, there is an ever-increasing count of amputees. Spoden et al. established that in Germany alone, the number of lower leg amputation cases was 52,096 in 2005 and 55,595 in 2015, confirming that a significant population suffers from lower leg amputations and it is on the increase [1]. A study by Manickum et al. suggests that 53.1% of lower leg amputations are caused by Diabetes Mellitus alone [2]. In India, Pooja et al. performed a study of 155 amputees and observed trauma to be the primary cause of amputation (70%) followed by vascular diseases. They also reported 94.8% of all amputations were of the lower limb [3]. The people having amputations have a very peculiar and different

gait cycle as compared to an average human, which makes them unconfident about themselves, driving them towards depression and anxiety [4] because of pain and pity. The primary target of any commercial prosthesis or orthoses is to reduce the gap between the gait data of its user and that of a healthy adult of the same physiology. Human gait refers to the walking pattern of Homo sapiens. It is similar for all healthy human adults and comprises different continuous phases in order, such as Heel Strike, Loading Response, Mid Stance, Terminal Response, and Swing Phase (Preswing and Swing). While designing and constructing a prosthesis (lower or upper limb), the primary aim is always to make it as similar as possible to that of a healthy person, which is why the data recorded from a healthy human still act as a

reference for such purposes. Nowadays, there is always the demand for a more and more intelligent and active prosthesis (working on an energy source) rather than a passive prosthesis (without an energy source). In general, the training time required by the patient to adjust to an active prosthesis is higher than that needed for a passive prosthesis. Yet, they are of higher demand because of their better performance aspect. The only other major drawback with active or adaptive prostheses/orthoses is their design complications. To simplify this, Chan et al. came up with an EMG based gait phase prediction model for prosthesis control [5]. As an extension of their work, Lee and Lee proposed the technique for predicting the posture angles of the patient's orthosis for a single lower limb [6]. Moissenet et al. tried the regression model approach to determine deviations for physiological and pathological subjects; however, we kept our study limited to perfectly healthy subjects [7]. Recently, Kim et al. used Doppler radar data for human gait analysis [8]. In this paper, we propose a neural network-based prediction control scheme for intelligent orthoses and prostheses, where the gait data for the next 20 seconds are predicted for given constant speed. In the case of Lee and Lee, the error was approx. 0.5° , but in our neural network model, the error (MSE) is in the order of 10^{-4} . The oldest research work that could be traced in the field of gait pattern identification was done by Murray et al. in 1970 when they compared the gait data of women [9]. Muro et al. compared the accuracy of nonwearable and wearable sensors [10]. In a nutshell, it can be concluded that the gait data can be measured via wearable goniometric sensors as well as nonwearable image processing alternatives. The recent advancement in technology allows both these approaches to have a very good amount of accuracy. Yao et al. have already confirmed that gait data angles are similar for subjects, so long as they are walking, regardless of whether on a treadmill or over-ground [11]. Pardasani et al. presented an algorithm for using the Kinect-based mocap procedure for comparative analysis of gait data using a nonwearable sensor [12]. There have been several other recent published works on Kinect and other nonwearable sensors [13, 14], but the scope of wearable sensors which can give much higher accuracy compared to nonwearable sensors is very limited. Internet of Things and its Sensors Devices are playing a great role in the new smart healthcare sector [15]. In this paper, first, we have presented basic information about human gait and its various stages and then discussed out data recording protocols and algorithm for designing neural network. The flow of the processes is as given in Figure 1.

2. Features of Human Gait

The human gait is a cycle of alternative forms of the lower limb, e.g., both the arms will always have the opposite form of each other while the subject is walking. This cycle is initiated when the foot of either leg hits the ground after a swing, called heel strike, and ends at the next heel strike of the same foot [16]. The various forms in between are loading response, midstance, terminal stance, preswing, and swing phase (preswing phase is mostly clubbed with the swing

phase [17]). The loading response, midstance, and terminal stance can be collectively termed as support phase or stance phase. Gait comprises of alternative repetition of these two (swing and stance) phases, one after the other, tracing a continuous cycle. It is evident from Figure 2, and it has been demonstrated by Kour et al. [18] and Vaughan et al. [19] in their work that human gait continuously changes in angles of the hip joint, knee joint as well as ankle joint. These changes are brought about by the energy gradient in the human skeletal muscles of the lower limb [20, 21].

3. Data Recording

Data were recorded for 5 different healthy subjects, able-bodied, without any known physiology or pathology at the time of recording. Sampling frequency was set at 100 Hz, i.e., 100 samples/second. The subjects were made to walk on a treadmill at following three constant velocities: 2.4 kmph (0.67 m/s), 3.6 kmph (1 m/s), and 5.4 kmph (1.5 m/s). Data were recorded at each of these speeds for 20 seconds each, keeping a buffer of 20 s at the start and between two consecutive recordings allowing sufficient transition time for the treadmill to gain the required velocity and the subject to acclimatize to the speed. The same steps were repeated for three instances per individual to ensure good quality of data is collected for analysis. So, a total of 45 different datasets were derived overall, which were then bifurcated into their respective speeds. Gait data were collected using the Biometrics goniometer. Data were recorded for each of the joints of both the lower limbs, i.e., right hip, left hip, right knee, and left knee. These subjects were aged 19–23 and weighed around 65–85 kgs, i.e., they were neither too obese nor too malnourished. A random subject during data recording can be seen in Figure 3. The location of the sensors is marked in black.

4. Methodology

As mentioned above, the data were recorded for five different healthy subjects, thrice for three different speeds, making it a total of 45 instances. Recorded data were first split based upon respective speeds because the device used to record the angle data using a Biometrics goniometer [22] exports the recorded data in a single file with each joint as a separate channel. A periodogram showed that they all had high-frequency noises in them, which was clear from the plot [23]. This noise can be attributed to the electronics involved and minute shifting of sensors while in motion. The actual data are only in the range of 0.5–5 Hz at maximum. So, appropriate low pass Butterworth filters were designed and applied to ensure the data can be made as much noise-free as possible. The most observable noises were the high-frequency noise observed from the spectrogram, as well as some aperiodic low-frequency noise, which was affecting the peaks of the signal. So, Butterworth low pass filters were used to filter out the high-frequency noise and preserve the low-frequency components, which too consisted of some noise as peak distortion was still observable in some cases. We have processed the different stages as the flow of the processes, available in Figure 1. So,

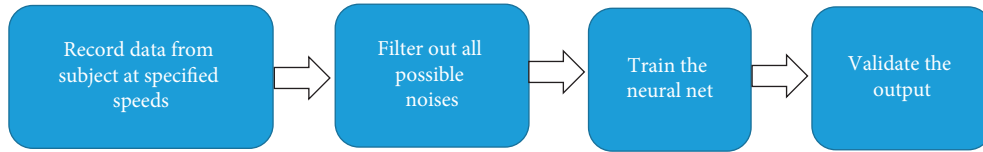


FIGURE 1: Flow of the processes.

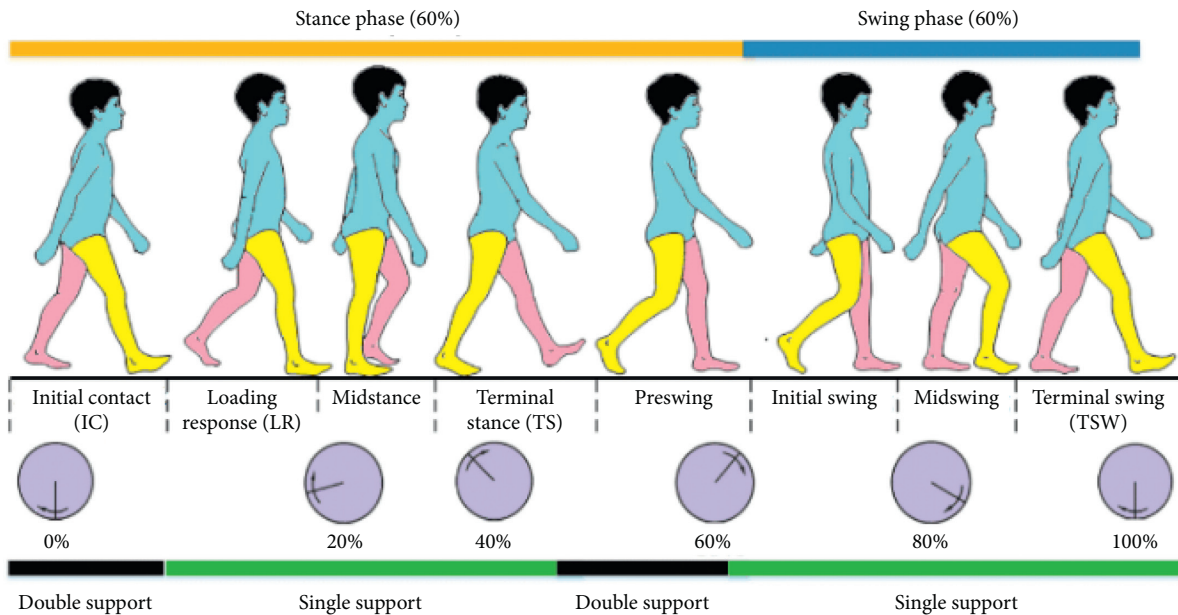


FIGURE 2: Various stages of human gait [18].

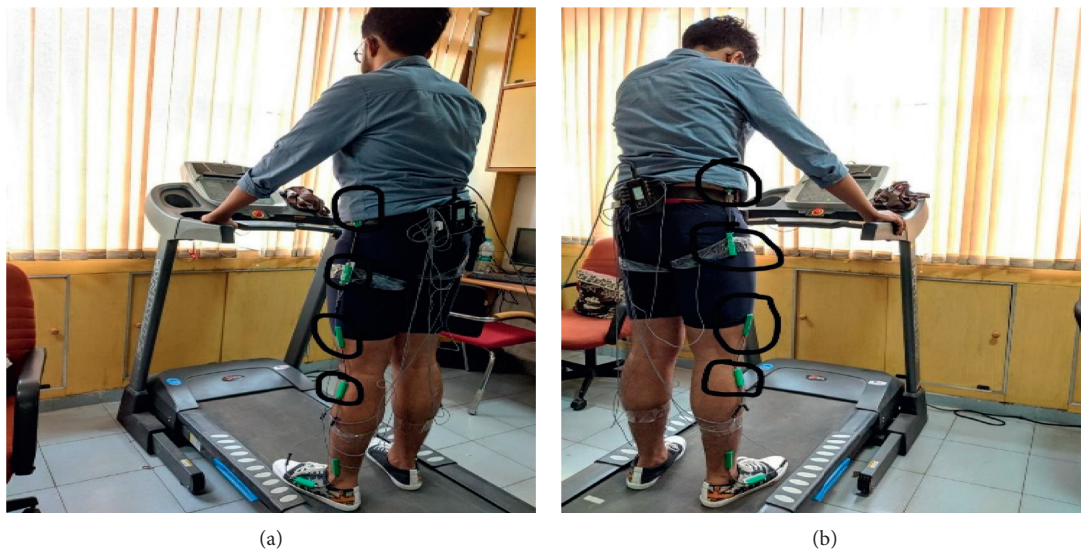


FIGURE 3: A random subject during data recording.

to take care of that, a moving average filter was used with varying window sizes based upon the speed at which the subject is walking. But, for two different subjects, the window size applied was kept the same to ensure uniformity. Similarly, the low pass filters were designed separately for each speed, although it was kept the same for different

individuals at the same speed. The several sets of figures (Figures 4–9) show before (b) and after filtration data (a) for 7 seconds for visualization purposes. Figures 4 and 5 for hip and knee, respectively, show the data at 2.4 kmph, Figures 6 and 7 at 3.6 kmph, and Figures 8 and 9 at 5.4 kmph.

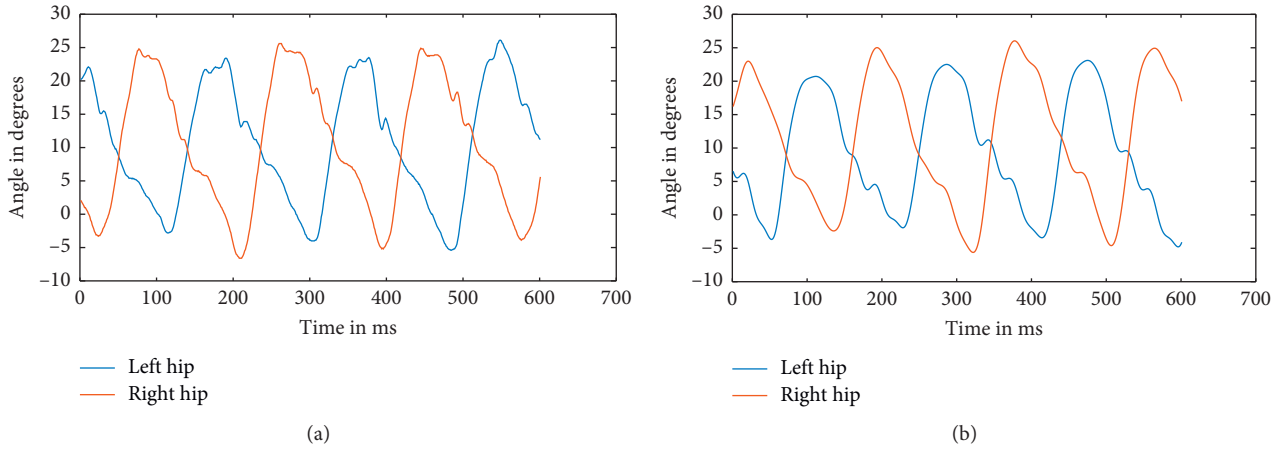


FIGURE 4: A Raw data of hips at 2.4 kmph (a) and filtered data of hips at 2.4 kmph (b).

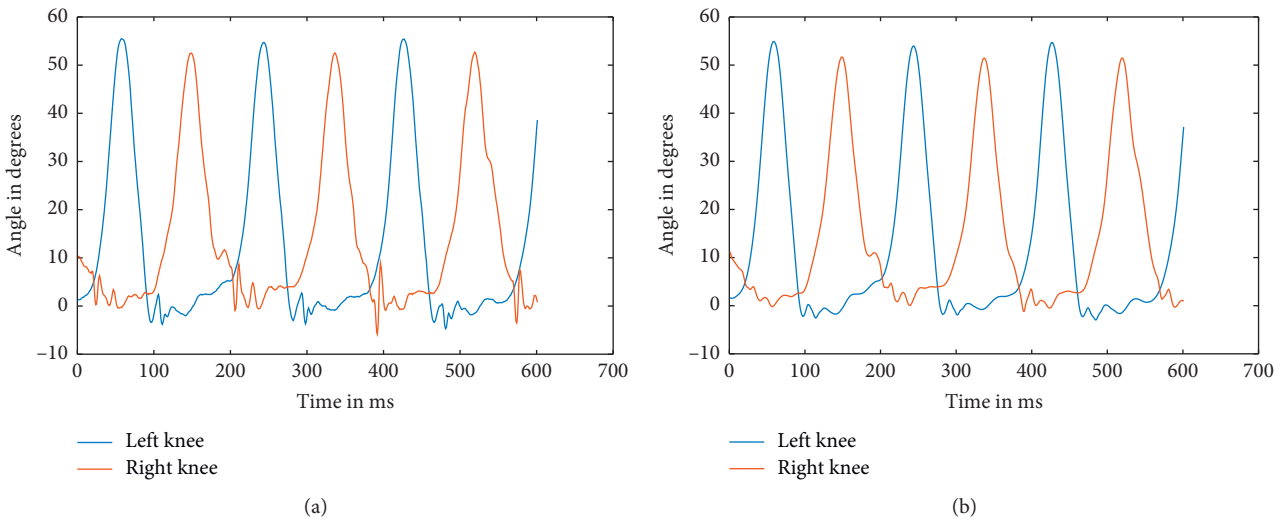


FIGURE 5: Raw data of knees at 2.4 kmph (a) and filtered data of hips at 2.4 kmph (b).

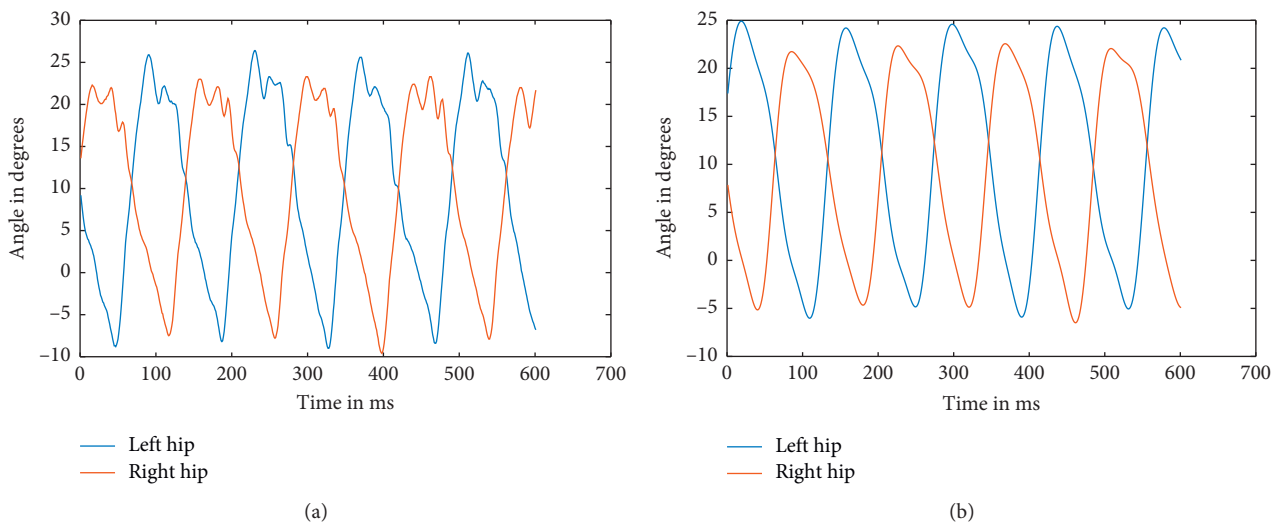


FIGURE 6: Raw data of hips at 3.6 kmph (a) and filtered data of hips at 3.6 kmph (b).

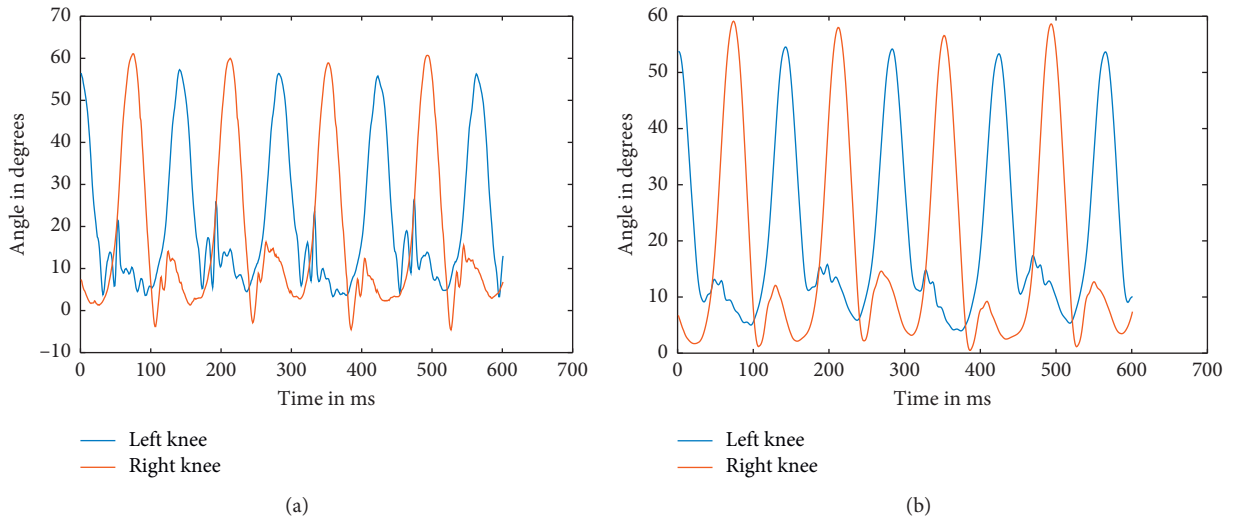


FIGURE 7: Raw data of knees at 3.6 kmph (a) and filtered data of knees at 3.6 kmph (b).

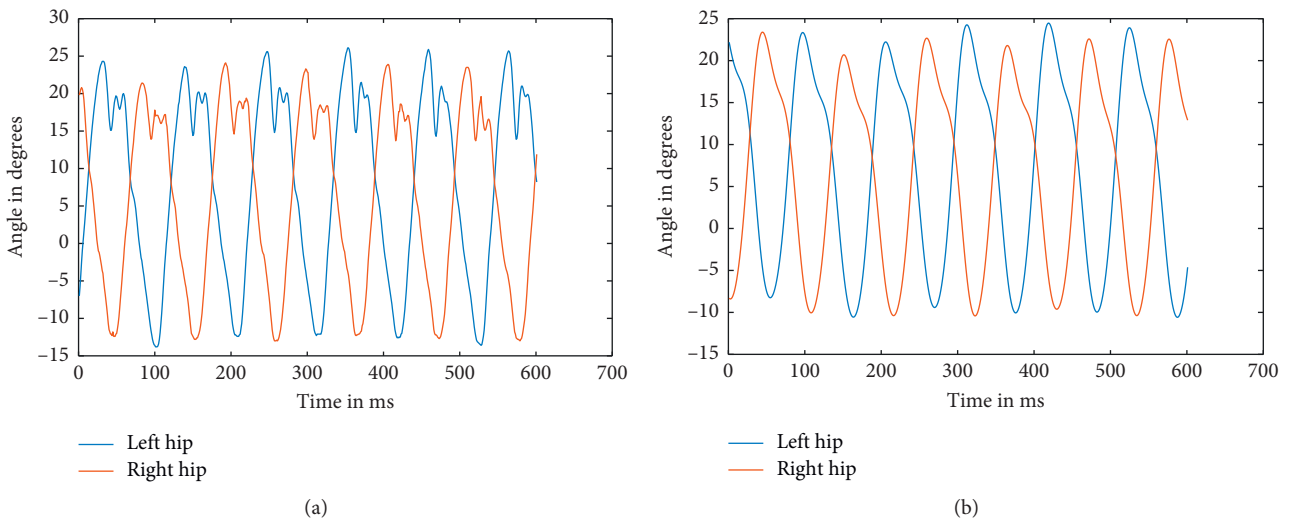


FIGURE 8: Raw data of hips at 5.4 kmph (a) and filtered data of knees at 5.4 kmph (b).

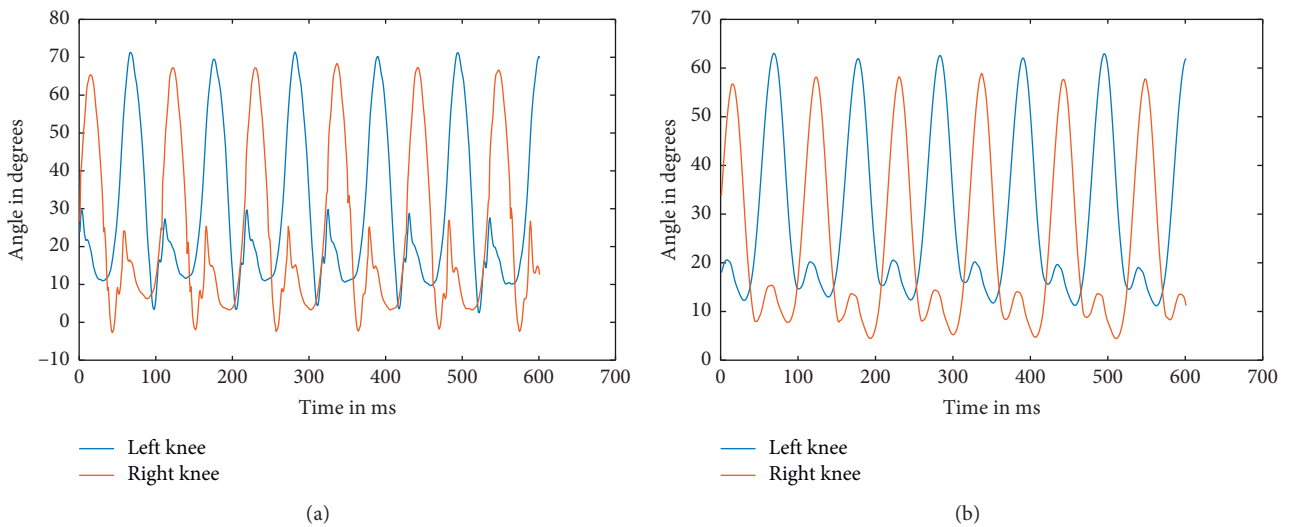


FIGURE 9: Raw data of knees at 5.4 kmph (a) and filtered data of hips at 5.4 kmph (b).

Once it was done on the data recorded from all the subjects, it was observed that, although the signal of angle measurement at different speeds was slightly similar, at a constant speed, it was pretty identical not only in the three instances of the same subject but also in case two distinct subjects. So, we tried to train a neural net to design a function in Matlab, which will give the next angle if the previous 100 aspects recorded are given to it as an input (keeping in mind $f_s = 100$ Hz). This function can be used as a reference for designing prosthesis as well as validating gait in case of biomechanical rehabilitation. So, using a for loop for each data sequence recorded, two matrices were created, one to act as input for the neural net and the second one to serve as a target. These matrices were made for each speed separately so that separate networks can be trained to result in distinct functions for each speed. In the input matrix, every set of 101 consecutive cells was passed as one element from the data recorded and the 102nd data were passed into the target matrix. This process was repeated for each individual at each speed. Afterward, the neural net was designed and trained. The algorithm used was the LevenbergMarquardt method using the Neural Network Toolbox of MatLab 2018a [24]. It was observed that after training the net only for the first subject, the functions were able to predict the data accurately for the remaining four subjects without even being trained by their data. For example, in the two plots shown in Figure 10, the one on top is the actual data recorded for the second individual and the one on the bottom is the neural net output although; it was trained only on the data of the first person. This particular data is of right hip, but a similar phenomenon has been observed in all the joints.

The uncanny similarity (as one of the plots almost superimposes the other) is further confirmed by correlation analysis calculated by Pearson's correlation coefficient, which proves that these two plots are exactly similar discussed later in the paper. The correlation coefficient, which is used to determine the similarity of two datasets scientifically, in this particular case, comes out to be 0.999, where affinity towards 1 suggests similarity and resemblance towards 0 suggests nonsimilarity of the data. In some cases, the coefficient can be negative, which suggests the two inputs being similar out of phase.

5. Neural Net Specifications

As mentioned earlier, using Matlab's Neural Net Fitting tool [24] and the abovementioned dataset and targets, separate neural nets were designed for each joint, and a Matlab code was exported using the same tool for function. The specific reason for using LM is that our data here are nonlinear, and the best approach to predict nonlinear data is to use the LM method. A total of 1600 samples were obtained for each case, out of which 1120 samples were used for training, 240 for testing, and 240 for validation in each case. Following individual functions were realized for each speed (2.4kmph, 3.6kmph, and 5.4 kmph separately):

- (1) Right hip
 - (1.1) Right hip slow, 2.4 km/h
 - (1.2) Right hip medium, 3.6 km/h
 - (1.3) Right hip fast, 5.4 km/h
- (2) Left hip
 - (2.1) Left hip slow, 2.4 km/h
 - (2.2) Left hip medium, 3.6 km/h
 - (2.3) Left hip fast, 5.4 km/h
- (3) Right knee
 - (3.1) Right knee slow, 2.4 km/h
 - (3.2) Right knee medium, 3.6 km/h
 - (3.3) Right knee fast, 5.4 km/h
- (4) Left knee
 - (4.1) Left knee slow, 2.4 km/h
 - (4.2) Left knee medium, 3.6 km/h
 - (4.3) Left knee fast, 5.4 km/h

It was then observed that the neural net showed unsupervised learning, and even if the net was trained for just one subject's data, it was able to predict the data for the remaining subjects accurately. However, they were never trained for that subject. Furthermore, the accuracy was so high that the correlation coefficient was always in the range of 0.99. This can be accredited to the low mean-squared error while training the net. Roughly, every net had 1600 sets of inputs and targets. They were trained using LevenbergMarquardt Model as mentioned earlier, and the number of hidden layers was set between 10 and 25 determined by the hit and trial method wherever it gave the best output with maximum accuracy while not requiring a very long time to train the net, although still training each net took approx. 20–25 minutes. The intention was to establish an efficient tradeoff between accuracy and time required to train the model. Still, in most cases, the MSE was very low (Figure 11(a)), so the figure was settled at ten although it may vary depending upon the initial conditions and the output accuracy. The exact time could vary based upon the initial conditions which are taken randomly in the case of the LM Model. The very low mean-squared error resulted in a very high correlation coefficient, which proves the high accuracy of the net and the model. Figure 11(b) and 11(c) show the properties of the neural net designed.

6. Results and Validation

The neural net result was validated using the correlation coefficient and correlation plot to identify the similarity between the expected output and the corresponding neural net function output. These correlation plots have been plotted for both the dataset used for training as well as the dataset not used for training. Yet, the similarity is very high in both cases, confirming unsupervised learning of the neural net. Table 1 shows the accuracy when the dataset used in training and the dataset used for validation are of same subject but from different trials. Table 2 shows the accuracy when the dataset for training and validation belongs to two different subjects. Figures 12 and 13

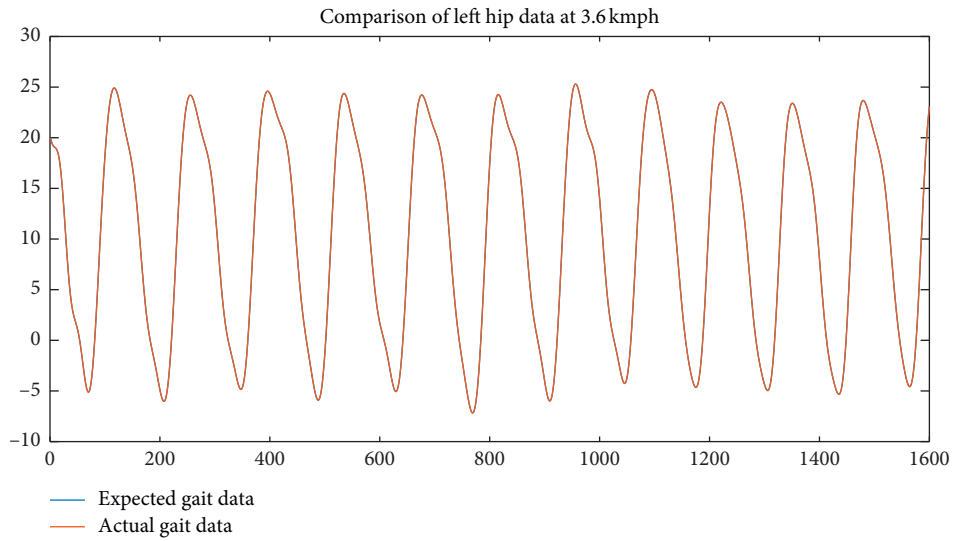
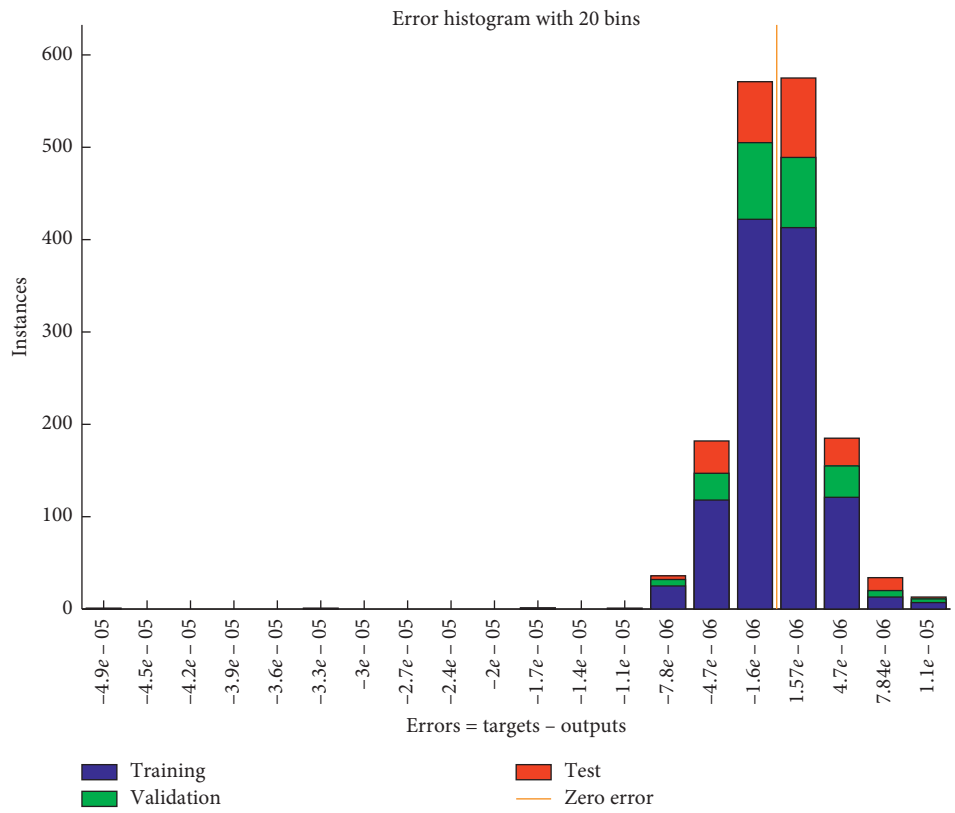


FIGURE 10: Comparison between neural net output and expected output for the hip @3.6 kmph.

Results			
	Samples	MSE	R
Training:	1120	$8.99338e - 12$	$9.99999e - 1$
Validation:	240	$1.23336e - 11$	$9.99999e - 1$
Testing:	240	$2.75970e - 11$	$9.99999e - 1$

(a)



(b)

FIGURE 11: Continued.

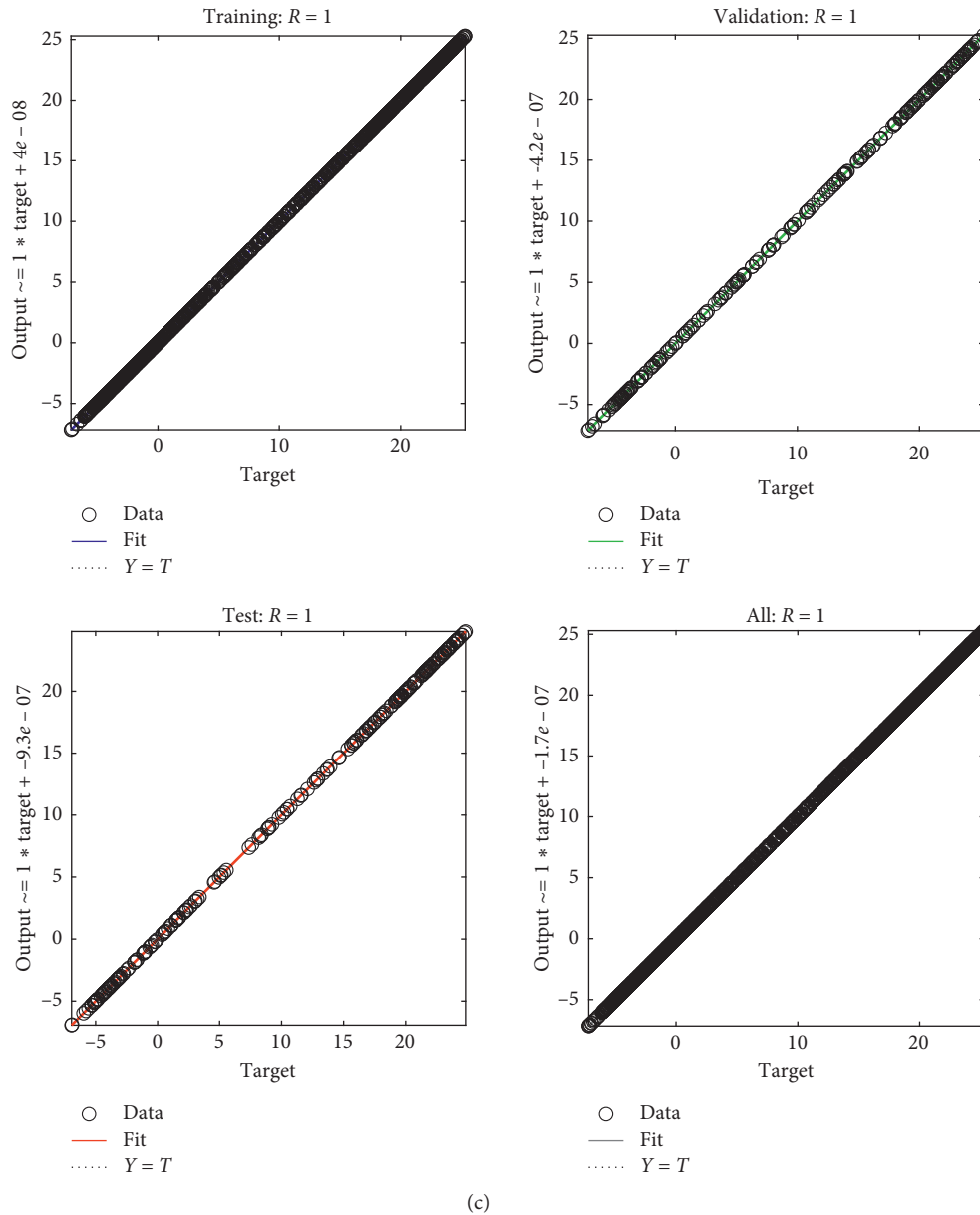


FIGURE 11: Neural net training specs for right hip @ 3.6 kmph.

TABLE 1: Percentage correlation values of output compared against the target for the dataset used for training.

Speed (kmph)	2.4	3.6	5.2
Correlation of right hip	99	100	100
Correlation of left hip	100	100	100
Correlation of right knee	99.33	97.39	98.03
Correlation of left knee	99.33	97.39	98.03

show the correlation plot between the expected and the actual output of neural net function for right hip and right knee, respectively, at 3.6 kmph when the data belong to same subject but from different trials. Figures 14 and 15 show the same plots when the dataset belongs to two different subjects.

TABLE 2: Percentage correlation values of output compared against the target for dataset not used for training.

Speed (kmph)	2.4 (%)	3.6 (%)	5.2 (%)
Correlation of right hip	99.33	97.39	98.03
Correlation of left hip	100	100	100
Correlation of right knee	98.99	98.01	99.60
Correlation of left knee	98.86	96.54	98.87

Case-I: when training dataset and validating dataset are of same subject

Case II: when training dataset and validating dataset are of different subjects

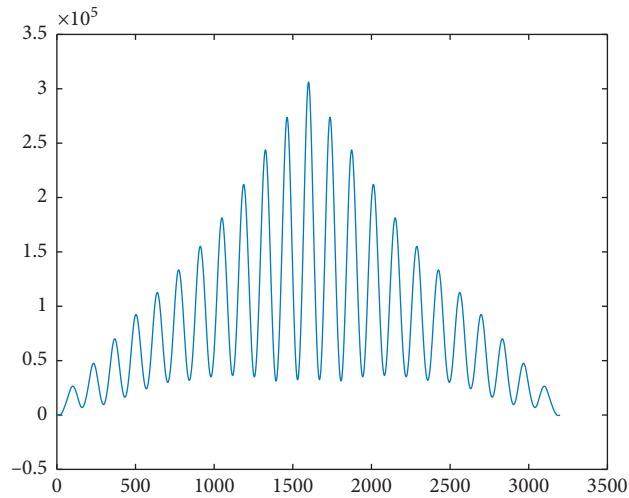


FIGURE 12: Plot of correlation between neural net output and target for right hip at 3.6 km/h. The Y-axis is in the range of 10^5 which confirms very high similarity in both the arrays passed.

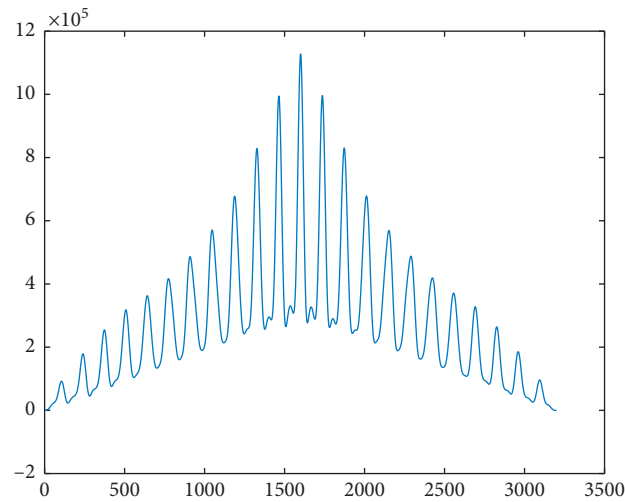


FIGURE 13: The plot of correlation between neural net output and target for right knee at 3.6 km/h. The Y-axis is in the range of 10^5 , which confirms very high similarity in both the arrays passed.

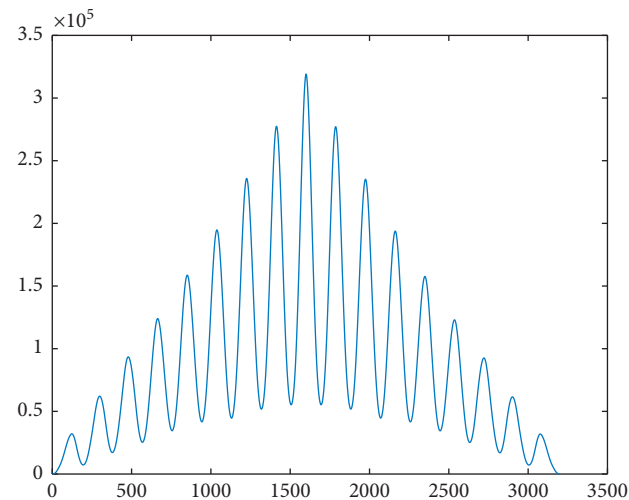


FIGURE 14: Plot of correlation between neural net output and target for right hip at 3.6 km/h. The Y-axis is in the range of 10^5 which confirms very high similarity in both the arrays passed.

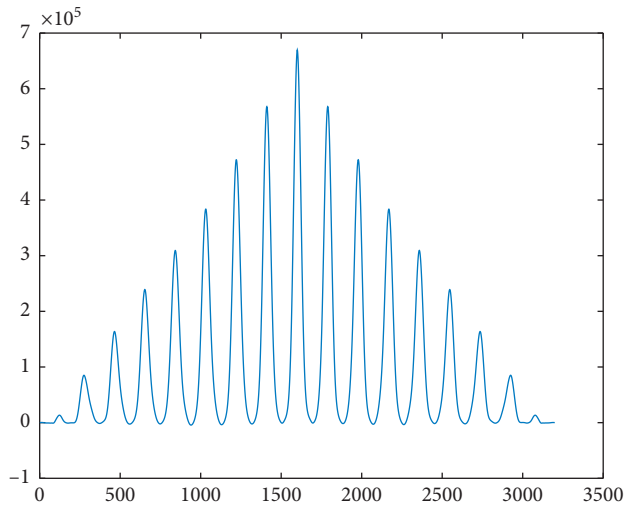


FIGURE 15: Plot of correlation between neural net output and target for right knee at 3.6 km/h. The Y-axis is in the range of 105, which confirms very high similarity in both the arrays passed.

7. Conclusion

The study aims at proposing a new approach for prediction of gait data to be used for reference while designing prostheses and orthoses as well as validation of gait data during biomechanical rehabilitation. Experimentally, the least accuracy observed was 97.39% for a subject whose data are used to frame the model, whereas 96.54% for a subject whose data were not used to frame the model. Thus, we conclude our experiment to be a success. The prostheses and orthoses using this method are expected to have a very positive effect on the life of subjects and are likely to assist in the betterment of the quality of life of an amputee.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this work.

Acknowledgments

This research was supported by the Taif University Researchers Supporting Project number (TURSP-2020/231), Taif University, Taif, Saudi Arabia.

References

- [1] M. Spoden, U. Nimptsch, and T. Mansky, "Amputation rates of the lower limb by amputation level - observational study using German national hospital discharge data from 2005 to 2015," *BMC Health Services Research*, vol. 19, no. 1, p. 8, 2019.
- [2] P. Manickum, S. S. Ramklass, and T. E. Madiba, "A five-year audit of lower limb amputations below the knee and rehabilitation outcomes: the Durban experience," *Journal of*

- Metabolism, Endocrinology and Diabetes of South Africa Volume*, vol. 24, 2019.
- [3] G. Das, "Prevalance and aetiology of amputation in Kolkata, India: a retrospective analysis," *HongKong Physiotherapy Journal*, vol. 31, no. 1, 2013.
- [4] S. Bhutani, J. Bhutani, A. Chhabra, and R. Uppal, "Living with amputation: anxiety and depression correlates," *Journal of Clinical and Diagnostic Research: JCDR*, vol. 2016, 2016.
- [5] F. H. Y. Chan, Y. S. Yong-Sheng Yang, F. K. Lam, Y. T. Yuan-Ting Zhang, and P. A. Parker, "Fuzzy EMG classification for prosthesis control," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 3, pp. 305–311, 2000.
- [6] L. Lee, "Gait angle prediction for lower limb orthotics and prostheses using an EMG signal and neural networks," *International Journal of Control, Automation, and Systems*, vol. 3, no. 2, pp. 152–158, 2005.
- [7] F. Moissenet, F. Leboeuf, and S. Armand, "Lower limb sagittal gait kinematics can be predicted based on walking speed, gender, age and BMI," *Scientific Reports*, vol. 9, no. 1, p. 9510, 2019.
- [8] I. Alnujaim and Y. Kim, "Augmentation of Doppler radar data using generative adversarial network for human motion analysis," *Healthcare Informatics Research*, vol. 25, no. 4, pp. 344–349, 2019.
- [9] M. Murray, R. Kory, and S. Sepic, "Walking patterns of normal women," *Archives of Physical Medicine and Rehabilitation*, vol. 51, no. 11, pp. 637–650, 1970.
- [10] A. Muro, B. Zapirain, and A. Mendez-Zorrilla, "Gait analysis methods: an overview of wearable and non-wearable systems, highlighting clinical applications," *Sensors (Basel, Switzerland)*, vol. 14, pp. 3362–3394, 2014.
- [11] J. Yao, "Lower limb joint motion and muscle force in treadmill and over-ground exercise," *BioMedical Engineering OnLine*, vol. 18, 2019.
- [12] J. Narayan, A. Pardasani, and S. Dwivedy, "Comparative gait analysis of healthy young male and female adults using kinect-labview setup," 2020.
- [13] X. Xu, R. W. McGorry, L.-S. Chou, J.-H. Lin, and C.-C. Chang, "Accuracy of the Microsoft Kinect for measuring gait parameters during treadmill walking," *Gait & Posture*, vol. 42, no. 2, p. 145, 2015.
- [14] Q. Li, Y. Wang, A. Sharf et al., "Classification of gait anomalies from kinect," *The Visual Computer*, vol. 34, no. 2, pp. 229–241, 2018.
- [15] S. Ahmad, M. Hasan, M. Shahabuddin, T. Tabassum, and M. W. Allvi, "IoT based pill reminder and monitoring system," *International Journal of Computer Science and Network Security*, vol. 20, no. 7, pp. 152–158, 2020.
- [16] N. Ozkaya and M. Nordin, *Fundamental Biomechanics; Equilibrium, Motion, and Deformation*, Springer, New York, NY, USA, 1998.
- [17] An Affordable Insole-Sensor-Based Trans-femoral Prosthesis for Normal Gait Sensors, 2018.
- [18] N. Kour, Sunanda, and S. Arora, "Computer-vision based diagnosis of parkinson's disease via gait: a survey," *IEEE Access*, vol. 1, 2019.
- [19] C. L. Vaughan, B. L. Davis, and J. C. Connor, *Dynamics of Human Gait*, Kiboho, Cape Town, South Africa, 1992.
- [20] Y. Koike and M. Kawato, "Trajectory formation from surface EMG signals using a neural network model," *Japan EIC*, vol. J77, no. 1, pp. 193–203, 1994.
- [21] L. Wang and T. S. Buchanan, "Prediction of joint moments using a neural network mode of muscle activations from EMG

signals,” *IEEE Trans. on Rehabilitation Engineering*, vol. 10, no. 1, pp. 30–37, 2002.

- [22] Twin axis Wireless Goniometers, *Twin-Axis Goniometers for Dynamic Joint Movement Analysis*, Biometrics Ltd., Newport, UK, 2020.
- [23] MATLAB, *MATLAB and Signal Processing Toolbox Release 2018a*, The MathWorks, Inc., Natick, MA, USA, 2018.
- [24] MATLAB, *Neural Network Toolbox Release 2018a*, The MathWorks, Inc., Natick, MA, USA, 2020.

Review Article

Big Data, Extracting Insights, Comprehension, and Analytics in Cardiology: An Overview

Hui Xiao ¹, Sikandar Ali ², Zhen Zhang,¹ Muhammad Shahzad Sarfraz,³ Fang Zhang,¹ and Mohammad Faisal ⁴

¹Zhongnan Hospital of Wuhan University, Information Center, Wuhan 430071, China

²Department of Computer Science and Technology, China University of Petroleum-Beijing, Beijing 102249, China

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

⁴Department of Computer Science and Information Technology, University of Malakand, Chakdara, Pakistan

Correspondence should be addressed to Hui Xiao; zhospitalxh@163.com and Sikandar Ali; sikandar@cup.edu.cn

Received 8 December 2020; Revised 9 January 2021; Accepted 20 January 2021; Published 31 January 2021

Academic Editor: Iván García-Magariño

Copyright © 2021 Hui Xiao 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.

Healthcare system facilitates the treatment of patients with the support of wearable, smart, and handheld devices, as well as many other devices. These devices are producing a huge bulk of data that need to be moulded for extracting meaningful insights from them for the useful use of researchers and practitioners. Various approaches, methods, and tools are in use for doing so and to extract meaningful information in the field of healthcare. This information is being used as evidence to further analyze the data for the early care of patient and to devise treatment. Early care and treatment can facilitate healthcare and the treatment of the patient and can have immense potentiality of dropping the care cost and quality refining of care and can decrease waste and chances of error. To facilitate healthcare in general and cardiology in specific, the proposed study presents an overview of the available literature associated with big data, its insights, and analytics. The presented report will help practitioners and researchers to devise new solutions for early care in healthcare and in cardiology.

1. Introduction

The increase in use of smart devices such as sensor, actuator, and wearable devices, as well as other devices, has produced massive amount of data that need to be shaped in a structure way to mine the information and useful insights for the useful use of research and practice. This increase of information can yield research issues and challenges as extracting useful information becomes a challenging task for research. The useful insights once drawn in a successful way can ultimately care the patient and provide effective treatment in the early stage. Diverse approaches, methods, and mechanisms are in practice to tackle the issues of big data and its analytics in the field of healthcare in general and cardiology in specific.

The role of data processing and information in healthcare has always been vital in healthcare for decision-making

and its provision. Medical big data is produced from the communication and digitization in healthcare. Healthcare providers and hospital industry provide a huge amount of data from other segments, such as medical equipment, medical insurance medical research, and life science. Huge amount of data exists, which grasps the potential of support of healthcare and medical tasks. The integration of machine learning, artificial intelligence, and advanced analytics offers numerous opportunities for transmuting such data into actionable and expressive insights for supporting decision-making. This can ultimately make availability of patient care at high quality and real-time situation response and can protect lives on the clinical side and develop the services and processes, improve the use of resources, and minimize the costs on the maintenance and financial front [1, 2].

With the rise of advanced approaches such as analytical techniques and approaches, the stakeholders of healthcare

can not only connect the data power for historical analysis of data but also predict future outcomes with predictive analytics for defining best accomplishment for present situation [3, 4]. Conventionally, the practitioners of clinic rely on reserved information accessible to them and their past involvement for treatment of patients. Data availability from diverse sources deals with the chance to have a complete thought of patient well-being. The use of cutting-edge technologies against such data aids access to the appropriate information at precise place and accurate time for delivering precise care [5].

The proposed study presents an overview of the available literature associated with big data, its insights, and analytics. The process of search for the proposed study was done in the popular libraries with the aim of obtaining associated materials. The presented report will help practitioners and researchers to devise new solutions for early care in healthcare and in cardiology.

The remainder of the paper is organized as follows: Section 2 presents the interrelated research to current study. Section 3 presents library-based search process for the proposed study. Section 4 concludes the paper.

2. Related Work

Several approaches have been in practice to tackle diverse issues of big data and its analytics in healthcare. Pevnick et al. [6] offered a review that discusses the current and upcoming devices intended for measuring the actions of heart rhythm, heart rate, and thoracic fluid. Various frameworks were presented, which classify and understand the wearable devices. Mehta et al. [7] presented a systematic mapping study for analyzing and identifying the research studies on analytics of big data and use of artificial intelligence in healthcare. The study identified 2421 papers for the year's ranges from 2013 to February 2019. These papers were evaluated, and the results show that the study will support the necessity in the use of technologies in healthcare. Atallah et al. [8] surveyed the literature associated with the DL and IoT applications for smart cities' developments. Initially, the basics of IoT were defined followed by the characteristics of IoT-produced big data. After that, the various structures used for analytics of IoT big data were presented. The common DL models were surveyed and reviewed the current research employing the IoT and DL for developing services and smart applications for smart cities. The existing issues and challenges encountered throughout the smart city's development were outlined. Kazmierska [9] presented a study on the needs of community in translating multisource data into clinical decision aids.

Ben-Assuli et al. [10] demonstrated power prediction of four popular algorithms and matched their accuracy in congestive heart failure predicting initial patient mortality. The results show that the current models outperform those described in the literature. The results further support the policy-makers in allocation of resources for establishment of comprehensive systems of integrated health IT aiming at simplification of analytics of ML. Dipti Itchhaporia [11] analyzed the existing application and state of machine

learning approaches and artificial intelligence in cardiovascular medicine. The effects of emerging technologies on cardiovascular medicine are emphasized for providing understanding to the clinical practice and to find probable patient assistances. Nazir et al. [12] provided a wide-ranging overview of the available big data studies in cardiology. The study followed a protocol of systematic literature review for presenting the published material from 2008 till 2018 associated with big data features, applications, and analytics in cardiology field. The authors identified 190 potential studies and analyzed them. These studies were published in conferences, books, journals, and many other online materials. The study was presented as an evidence for the researchers and practitioners to devise novel solutions in the area of interest. Nazir et al. [13] presented a comprehensive review of the 10 years from 2008 to 2018 associated with the visualization of big data in the area of cardiology. The study identified 53 prospective papers related to visualization of big data in cardiology. The study was based on protocol with defined research questions, inclusion and exclusion criteria, and quality criteria. These identified studies were analyzed according to the defined research questions. The study highlighted the increase of the number of researches in the area and focused on further research and innovations in the field. These studies were done in order to support the usage of big data in healthcare.

Bizopoulos and Koutsouris [14] surveyed applications of deep learning that uses structured data and signal and imaging modalities from cardiology. The benefits and limitations of applications of deep learning in cardiology and in medicine in general are discussed. Cannière [15] examined the developments of heart rate variability factors during short-term interval all the way through cardiac rehabilitation. Electrocardiography signals, documented with the help of wearable device in 129 patients following cardiac rehabilitation program, were analyzed. The findings of the study present appreciated insights into disease monitoring during cardiac rehabilitation in future application.

3. Library-Based Search Process

This study offers to present an overview of the existing approaches and methods for big data, its analytics, and insights in cardiology. Various popular libraries such as ScienceDirect, IEEE, Springer, and Wiley were searched with the aim of obtaining associated materials interconnected to the current study. The information gathered from these libraries was analyzed and presented from different perspectives in the form of different tables and figures. This information includes the type of article, number of publications, topics covered, subject areas, and publication titles. Initially, the library of ScienceDirect was checked and the following information was obtained. Figure 1 depicts the types of articles with publications. The figure shows that a bigger number of publications were in the form of research article.

Figure 2 presents the articles in total with the given year. More publications are shown in the year 2020, which shows the increase in number of researches.

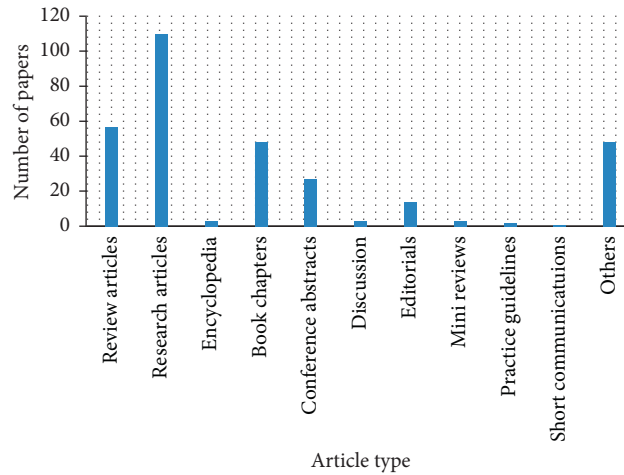


FIGURE 1: Article types.

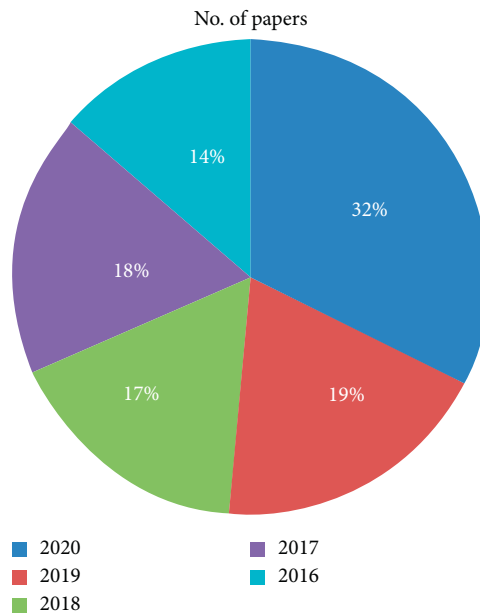


FIGURE 2: Number of papers published.

Figure 3 depicts the subject areas with the number of publications.

The library of IEEE was searched for the purpose of identifying relevant information. Figure 4 represents the information of publication topics with the total number of articles published.

The paper type and total number of publications in the same library are shown in Figure 5.

Figure 6 presents the conference location with the number of publications.

After this, the library of Springer was searched to view the information for the purpose of analysis. Figure 7 depicts the type of articles with the number of publications.

The discipline with the total number of articles is shown in Figure 8. The purpose of this search was to identify the disciplines covered by the area.

The libraries of Wiley and Taylor & Francis were also part of the proposed study. These libraries were searched for relevant information and analysis. Figure 9 depicts the publication types with total number of articles published. In the figure, it is shown that more papers are published with type journal.

Figure 10 presents the number of publications in the given years from 2016 till 2020.

After the statistics were obtained, the papers were reviewed and the details with short descriptions of the papers were given. Table 1 shows the big data and its analytics in cardiology.

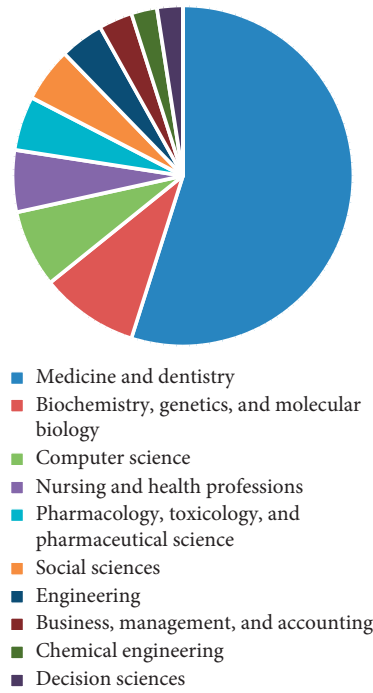


FIGURE 3: Subject area with number of articles.

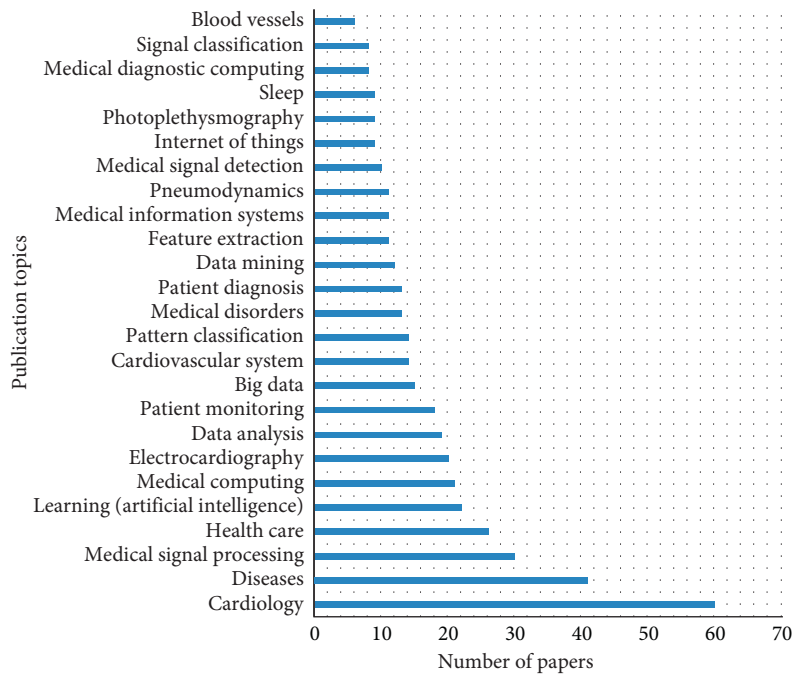


FIGURE 4: Publication topic with number of articles.

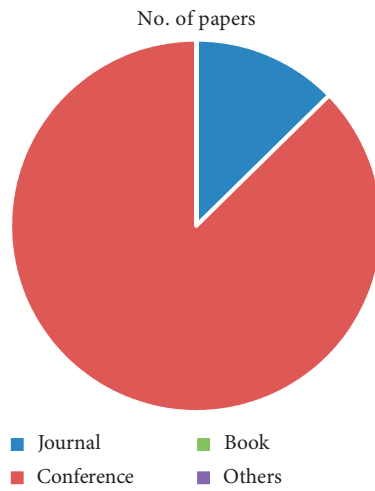


FIGURE 5: Paper type with number.

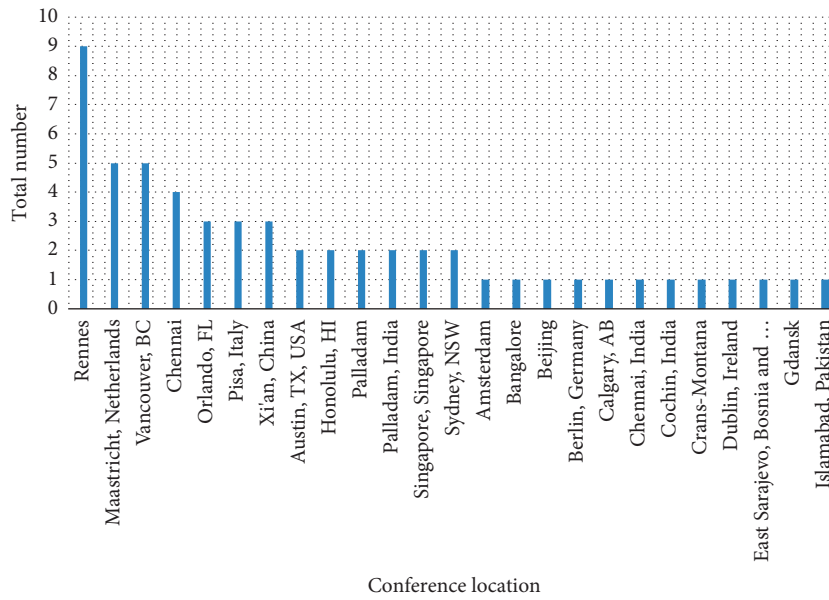


FIGURE 6: Conference location with number of articles.

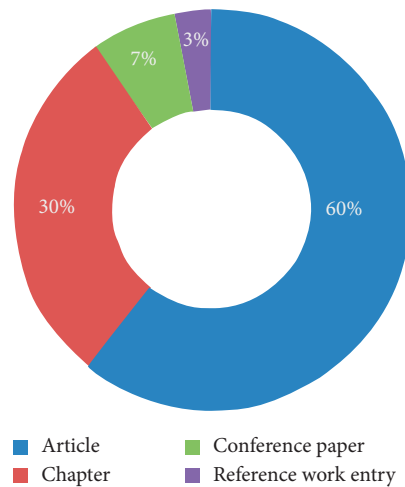
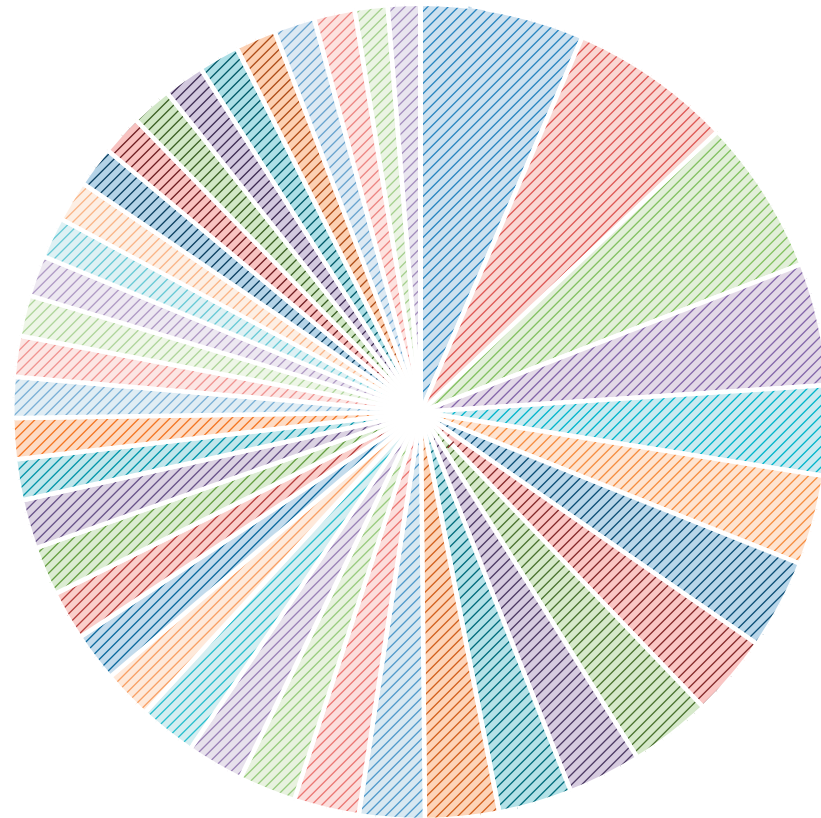


FIGURE 7: Content type.



- | | |
|-------------------------------------------------|---------------------------------------------------------|
| Cardiology | Internal medicine |
| Health informatics | Medicine/public health, general |
| Artificial intelligence | Biomedical engineering and bioengineering |
| Computational intelligence | Imaging/radiology |
| Pharmacology/toxicology | Biomedicine, general |
| Public health | Statistics for life sciences, medicine, health sciences |
| Intensive/critical care medicine | Oncology |
| Bioinformatics | Geriatrics/gerontology |
| Rheumatology | General practice/family medicine |
| Health administration | Information systems and communication service |
| Neuroradiology | Proteomics |
| Big data | Communications engineering, networks |
| Computational biology/bioinformatics | Computer communication networks |
| Diagnostic radiology | Endocrinology |
| Health policy | Health promotion and disease prevention |
| Health services research | Information systems applications (incl. internet) |
| Innovation/technology management | Interventional radiology |
| Management of computing and information systems | Pediatrics |
| Surgery | Ultrasound |
| Analytical chemistry | Big data/analytics |

FIGURE 8: Discipline with number of articles.

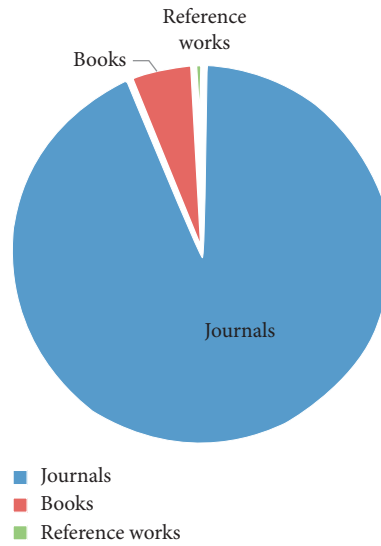


FIGURE 9: Publication type with number of papers.

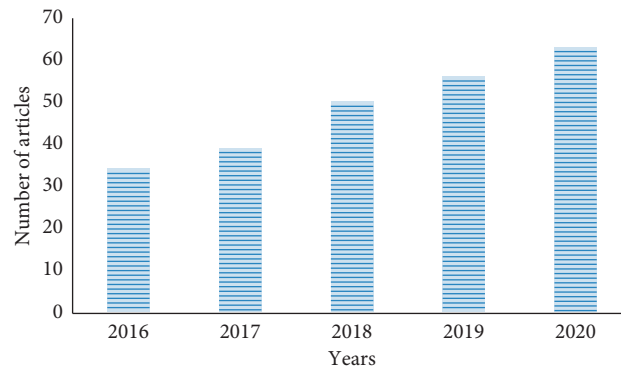


FIGURE 10: Publication in the year.

TABLE 1: Work in the area of big data and its analytics in cardiology.

Citation	Method	Description
[6]	Wearable technology for cardiology	Devices of wearable technology were intended for measuring the heart rhythm, heart rate, activity, and thoracic fluid. For classification and understanding, a framework was given to the wearable devices for improving healthcare.
[16]	Technology for cardiovascular disease patient	475 papers from PubMed library were examined for solutions of telemedicine to improve medication adherence of cardiovascular disease patient. 74 articles were assessed. The articles exhibited that suggestions associated with solutions of telemedicine are typically conflictive. The use of SMS was considered for patients regarding their medication because of forgetfulness.
[17]	Big data for monitoring of proactive healthcare of patients of chronic diseases	A system of health proactive monitoring is planned for cardiac patients. In the study, an electronic band is worn by the patient and the real-time situation of health and system of e-health are to process the obtained data from patient. The system supports the patient to take proactive process beside some of the abnormal states in their health and further supports the doctors to monitor the health of patient on a regular basis.

TABLE 1: Continued.

Citation	Method	Description
[18]	Deep learning network in left ventricular volumes identification in cardiac MRI	The research has established left ventricular volumes approach of identification without segmentation by deep learning technology and the data set form large-scale cardiac MRI from second annual data science bowl in 2016
[19]	Medical data mining and heart disease	The study offered review of the associated techniques of data mining for finding of diseases classes and heart failure. Additionally, emphasis is given on sequential mining.
[20]	ML horizon in cardiac hybrid imaging	The study presented summary of the fundamental notions in ML and its applications in standard cardiac imaging
[21]	Mobile heart rate monitoring system for patient of MI	Application of mobile for the patient of myocardial infarction (MI) is presented to preserve heart rate tracking and with stress-free pursuing for emergency support
[22]	Clinical guidelines in cardiology run-through in Sudan	Interviews in the two main cardiac hospitals of Sudan were led and an exploratory study among the hospitals' doctors was prepared for examining the perceptions of a huge population of prescribers of the subject examined
[23]	Big health data	Management of cardiac data and operative concept of remodelling to examine early symptoms of heart failure is presented
[24]	Big health data records for early and late translational cardiovascular research	The research censoriously reviewed the challenges of big data before time and after the event stages of research in translational cardiovascular disease
[25]	Factorization of tensor for precision medicine in failure of heart with preserved ejection fraction	The study examined the related woks on factorization of tensor applications in the associated biomedical field of phenotyping and genotyping
[26]	Mobile messaging applications' effect on knowledge of cardiac patients with risk of coronary artery disease and healthy lifestyle adherence	The research led a study from January to April 2017 in Klang Valley's teaching hospital for determining the effect of mobile messaging applications on patients with coronary artery disease and observation of healthy life
[27]	SVM for classification of biomedical signal on IoT platform	The study observed the signal over digital signal processor and then computed the blood oxygen saturation, heart rate, and blood pressure. SVM was considered for the purpose of showing the data and its classification into unhealthy, healthy, and very unhealthy and designates accuracy of classification prediction.
[28]	Heart failure prevention	Risk factors of heart failure and focus on prevention are specified
[29]	Statistical shape models of the heart	The study presented summary of the collected works of statistical shape models in cardiac imaging
[30]	Classifying mining techniques for the accessible clarification for heart disease prediction	Framework to use healthcare data for attributes based heart disease prediction is presented in the study
[31]	Statistical based recommendation model for the patients of heart disease	Smart system of recommendation is presented for patients with heart disease in the fields of medical informatics and e-health
[32]	Modelling 4D for rapid assessment of biventricular function in congenital heart disease	The study has quantified 4D biventricular function from analysis of standard cardiac MRI
[33]	Big database applications and forthcoming in cardiology	Applications of big database studies in cardiology were presented in the study
[34]	Approach of big data to myocyte membrane analysis	The study presented an approach for identification of particular pathological ion dynamics responsible for abnormal electrical behaviour practiced through the experiment
[35]	Distance, quality, or relationship? Interhospital transfer of patients with heart attack	The research inspected the patterns where the patients of heart attack are transferred between the hospitals. The three key factors in transferring destinations are: (i) The distance between sending and receiving in hospitals (ii) Widely reported quality measures of receiving hospitals as specified by whether they are related with the same multihospital system (iii) The relationship between sending and receiving hospitals

TABLE 1: Continued.

Citation	Method	Description
[36]	Feature analysis and coronary artery heart disease data sets	Integrate the experimental results of the examination of ML which are applied on varied data sets aiming at the coronary artery heart disease
[37]	Mapping of ventricular tachycardia in patients with heart structural disease	The paper has focused on the procedure of mapping ventricular tachycardia; the conventional and novel technique of mapping and the details of some methodological tips are given
[38]	Patients' baseline characteristics with heart failure and preserved ejection fraction during admission with acute heart failure in Saudi Arabia	Saudi Arabian patients with HFpEF were examined with acute heart failure. The clinical characteristics, signs, and indications of heart failure, echocardiographic findings, and medications during admission and at hospital discharge were determined.
[39]	Cardiovascular dysautonomias diagnosis and treatments through data mining	The authors established a cardiovascular dysautonomias identification system for the prediction of appropriate treatments and diagnosis for patients with cardiovascular dysautonomias through the data set extracted from the ANS unit of University Hospital Avicenne in Morocco
[40]	Approach of data transmission based on adaptive energy efficiency for prediction of heart disease	The research developed an adaptive energy resourceful transmission system which can recognise the important events like myocardial infarction and reduce data transmission from the devices
[41]	Analytics of big data in prediction of heart attack	The study identified the analytics uses in big data for the prevention and prediction of heart attack, privacy of the patients, and the challenges for the use of technology in big data. The study analyzed the national and international databases for the proposed study.
[42]	Cloud computing for myocardial fibre information in vivo	System of cloud-based investigation is intended for cardiac images and link services of computation for remote sharing. A method for postprocessing of image is defined as important service for obtaining information on in vivo myocardial fibres.
[43]	Using big data for assessing the risks of arrhythmia	The research presented an algorithm to involuntarily identify the R, S, and T wave peaks in epicardial electrogram signals
[44]	ML framework and imaging based big data for rapid phenotyping of left ventricular diastolic function	The study proposed that the cardiac biomechanics produce adequate information which can affect ML and framework of big data analytics for function of automated left ventricular diastolic assessment
[45]	Insights from echo reports of paediatric disease of heart	The entity site-feature values are mined in triples in the report of echo and then on the ground of this prediction of the level of risk
[46]	Framework of probabilistic data driven for scoring the preoperatives recipient-donor heart transplant survival	The technique of Bayesian belief networks is used. The approach contains four phases; the first and second phases of the data are preprocessed and a set of predictors are produced based on different variable selection method. The medically associated variables are added to the list of variables in the third phase, and in the last phase the Bayesian belief networks technique is applied.
[47]	Identification of heart arrhythmia through big data-based extraction of fuzzy partition rules	The research presented a novel semiautomatically fuzzy partition rules for facilitating an accurate and robust aspect into cardiac arrhythmia. The approach of text mining is demonstrated and applied to large data set containing freely existing articles in the PubMed library. The information is mined and then put to the experimental data and expert information for facilitating robust system to tackle the issues arising through the assessment of medical big data.
[48]	Big data for prediction of heart disease through map reduction	The research has established a central monitoring system for patients of large set of health records as input. It is intended to mine the essential information from large set of medical records through the method of map reduction. By using this approach, it can be decided whether there is patient normality or abnormality.

TABLE 1: Continued.

Citation	Method	Description
[49]	Cardiovascular risk clustering factors highlighted the coronary artery calcium as a strong clinical discriminator	The authors studied the relations between cardiovascular risk clustering and the discriminators of disease of cardiovascular factors
[50]	Mobile health initiatives for cardiovascular disease	The current technological and clinical improvements containing wearable health tracking devices, smartphone devices, and social media for supporting behaviour factors of risk for cardiovascular disease in terms of smoking, physical inactivity, and suboptimal nutrition
[51]	Sudden cardiac death with risk stratification and computational cardiology	The study defined guidelines of what is to be required for making the translational step, through the comparatively well intended cases required or drug induced long QT as a case of syndrome
[52]	Technology of smartphone in cardiology	The research presented various applications of smartphone based technologies in cardiology and gave a review of them
[53]	Visualization of cardiovascular MRI challenges and opportunities	The study offered an overview of the existing associated works of visualization approaches and emphasis on the visualizing imagery issues resulting from 2D myocardial tagging in CMR
[54]	Big data in cardiology	The article's purpose is the three encouraging big data applications in cardiovascular care, with "proof-of-concept" challenges to be met if the encouraging data is to be comprehended
[55]	Cardiovascular medicine big data, health informatics, and future	The study offered a report on cardiovascular medicine big data, health informatics, and future
[56]	Tool for the MIMIC-II database, a web-based data visualization	The objectives of the study are: (a) to build an interactive and (b) data visualization tool based on web MIMIC-II Furthermore, the research mainly offered two features of exploration and comparison. The first feature helps the patient cohort within MIMIC-II and visualized the distribution of various variables including administrative, clinical, and demographic variables within the selected cohort. The second feature helps the users in selection of two patient cohorts and visual comparison with other variables.
[57]	Connecting the dots: from big data to healthy heart	The study designated various sources of big data in cardiology followed by talk over the possibilities of building the best use of data-driven knowledge production models
[58]	Libraries implementation of open-source data visualization of web portal for patients of diabetes	A web portal is employed for improved communications of diabetes patients with doctors for the process of identification and handling of diabetes. Medical data are offered on the portal based on open-source libraries.
[52]	Technology of smartphone in cardiology	The research discusses the details of diverse applications of technologies of smartphone in cardiology
[59]	Machine learning approaches in detection of ischemic heart disease	SVM and Osuna were used for detecting the ischemic disease of heart. The principal component analysis algorithm was also used.
[60]	Paediatric cardiovascular disease in the era of transparency in healthcare using big data	The research offered a review on analytics of big data impact in paediatric cardiovascular disease and its possible issues of transparency in distribution of care
[61]	Data visualization: science on the map	A tool box for data visualization
[62]	Harnessing the heart of big data	The paper discussed the following: (i) Report on big data science research (ii) Potential of data science to support examinations of cardiovascular diseases (iii) Challenges and opportunities
[63]	4D OCT in developmental cardiology	The chapter emphasizes on numerous existing solutions and gives review of the perspective in the evaluation of 4D OCT imaging for cardiovascular system in the past several years

TABLE 1: Continued.

Citation	Method	Description
[64]	Kinect-based gesture prediction in volumetric visualization of heart from CMR imaging	The research aims to offer a virtual human heart from medical imaging data with incorporation of collaborating interface using visual 3D holographic, haptic, and sonic feedback
[65]	Feast for the eyes	The research presented the existing data visualization uses and reviewed the probable issues, benefits, and applications of libraries
[66]	Cardiac 4D ultrasound imaging	Overview of the technological developments for volumetric imaging of the heart beat with the support of ultrasound is given
[67]	Probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival	The study presented Bayesian belief network containing four phases. The data is preprocessed in the first two phases and produces a candidate set of predictors. Medical related variables are added in the third phase and, finally, the model of Bayesian belief network is applied.
[68]	Health analytics	The chapter discussed the visualization, analysis, and mining of healthcare data and concludes the way in which data can be proficiently accomplished which further improves the ability of organization to control risk, yield revenue, and cost
[69]	Electrophysiology-morphous merging of human heart based on composite visualization approach	The paper presented cardiac electrical excitation propagation model based on the data of human cardia cross-sectional to discover the cardiac electrical activities. After that, biophysical visualization method is applied for the biophysical integration of cardiac anatomy and electrophysiological properties, which provide the equivalent position, spatial relationship, and the whole process in 3D space with the context of anatomical structure for giving the details of biophysical and electrophysiological activity.
[70]	Big data for cardiology: novel discovery?	The paper determined the encouraging data sets for finding of science and the impact on the approaches used in science in general and explicitly in cardiology
[71]	Visualization of medical volume through intelligent approaches	The uses of algorithms and intelligent approaches of visualizing medical big data are presented. The article discusses the existing software and toolkits for visualization of medical volume.

Big data are considered to be the main asset of the organization for its successful operations and future endeavour [72–77].

4. Conclusion

Healthcare system facilitates the patients with the support of wearable devices, smart devices, handheld devices, and many other devices. These devices are producing a huge bulk of data that need to be moulded for extracting expressive insights from them for the useful use of researchers and practitioners. Various approaches, methods, and tools are in use for doing so and to extract meaningful information in the field of healthcare. This information is being used as evidence to further analyze the data for the early care of patients and to devise treatment. Early care and treatment can facilitate healthcare and patient and can have immense potentiality of quality refining of care and lessen care cost and can decrease waste and chances of error. To facilitate healthcare in general and cardiology in specific, the proposed study presents an overview of the existing literature associated with big data, its insights, and analytics. The presented report will help practitioners and researchers to devise new solutions for early care in healthcare and in cardiology.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, 2014.
- [2] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.
- [3] E. A. Mohammed, B. H. Far, and C. Naugler, "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends," *BioData Mining*, vol. 7, no. 1, 2014.
- [4] Y. Wang and N. Hajli, "Exploring the path to big data analytics success in healthcare," *Journal of Business Research*, vol. 70, pp. 287–299, 2017.


- [5] B. Kayyali, D. Knott, and S. V. Kuiken, "The big-data revolution in us health care: accelerating value and innovation," *McKinsey Co.*, pp. 1–6, 2013.
- [6] J. M. Pevnick, K. Birkeland, R. Zimmer, Y. Elad, and I. Kedan, "Wearable technology for cardiology: an update and framework for the future," *Trends in Cardiovascular Medicine*, vol. 28, no. 2, pp. 144–150, 2018.
- [7] N. Mehta, A. Pandit, and S. Shukla, "Transforming healthcare with big data analytics and artificial intelligence: a systematic mapping study," *Journal of Biomedical Informatics*, vol. 100, p. 103311, 2019.
- [8] S. B. Atitallah, M. Driss, W. Boulila, and H. B. Ghézala, "Leveraging Deep Learning and IoT big data analytics to support the smart cities development: review and future directions," *Computer Science Review*, vol. 38, Article ID 100303, 2020.
- [9] J. Kazmierska, "From multisource data to clinical decision aids in radiation oncology: the need for a clinical data science community," *Radiotherapy and Oncology*, vol. 153, pp. 43–54, 2020.
- [10] O. Ben-Assuli, T. Heart, N. Shlomo, and R. Klempfner, "Bringing big data analytics closer to practice: a methodological explanation and demonstration of classification algorithms," *Health Policy and Technology*, vol. 8, no. 1, pp. 7–13, 2019.
- [11] D. Itchhaporia, "Artificial intelligence in cardiology," *Trends in Cardiovascular Medicine*, 2020, In press.
- [12] S. Nazir, M. Nawaz, A. Adnan, S. Shahzad, and S. Asadi, "Big data features, applications, and analytics in cardiology-A systematic literature review," *IEEE Access*, vol. 7, pp. 143742–143771, 2019.
- [13] S. Nazir, M. Nawaz Khan, S. Anwar et al., "Big data visualization in cardiology-A systematic review and future directions," *IEEE Access*, vol. 7, pp. 115945–115958, 2019.
- [14] P. Bizopoulos and D. Koutsouris, "Deep learning in cardiology," *IEEE Reviews in Biomedical Engineering*, vol. 12, pp. 168–193, 2019.
- [15] H. D. Cannière, "The progression of heart rate variability parameters throughout cardiac rehabilitation," *Computing in Cardiology*, vol. 8–11, pp. 1–4, 2019.
- [16] R. W. Treskes, E. T. Van der Velde, J. W. Schoones, and M. J. Schalijs, "Implementation of smart technology to improve medication adherence in patients with cardiovascular disease: is it effective?" *Expert Review of Medical Devices*, vol. 15, no. 2, pp. 119–126, 2018.
- [17] A. Naseer, B. Y. Alkazemi, and E. U. Waraich, "A big data approach for proactive healthcare monitoring of chronic patients," in *Proceedings of the 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*, Vienna, Austria, July 2016.
- [18] G. Luo, G. Sun, K. Wang, S. Dong, and H. Zhang, "A novel left ventricular volumes prediction method based on deep learning network in cardiac MRI," in *Proceedings of the 2016 Computing in Cardiology Conference (CinC)*, Vancouver, Canada, 2016.
- [19] C. B. Rjeily, G. Badr, A. H. E. Hassani, and E. Andres, "Medical data mining for heart diseases and the future of sequential mining in medical field (machine learning paradigms)," *Machine Learning Paradigms, Intelligent Systems Reference Library*, vol. 149, pp. 71–99, 2018.
- [20] L. E. Juarez-Orozco, O. Martinez-Manzanera, S. V. Nesterov, S. Kajander, and A. J. Knuuti, "The machine learning horizon in cardiac hybrid imaging," *European Journal of Hybrid Imaging*, vol. 2, no. 15, pp. 1–15, 2018.
- [21] M. F. B. Mustapha and D. T. Anw, "Mobile heart rate monitor for myocardial infarction patients," in *Proceedings of the 2017 6th ICT International Student Project Conference (ICT-ISPC)*, Skudai, Malaysia, May 2017.
- [22] H. Elsadig, M. Weiss, J. Scott, and R. Laaksonen, "Use of clinical guidelines in cardiology practice in Sudan," *Journal of Evaluation in Clinical Practice*, vol. 24, no. 1, pp. 127–134, 2018.
- [23] D. Seth, N. Biswas, and D. Ghosh, "Big health data: cardiac remodelling and functional interactions of big brain based implications in body sensor networks," in *Proceedings of the 2017 7th International Conference on Communication Systems and Network Technologies*, Nagpur, India, November 2017.
- [24] H. Hemingway, F. W. Asselbergs, J. Danesh et al., "Big data from electronic health records for early and late translational cardiovascular research: challenges and potential," *European Heart Journal*, vol. 39, no. 16, pp. 1481–1495, 2018.
- [25] Y. Luo, F. S. Ahmad, and S. J. Shah, "Tensor factorization for precision medicine in heart failure with preserved ejection fraction," *Journal of Cardiovascular Translational Research*, vol. 10, no. 3, pp. 305–312, 2017.
- [26] Y. H. Tang, M. C. Chong, Y. P. Chua, P. L. Chui, L. Y. Tang, and N. Rahmat, "The effect of mobile messaging apps on cardiac patient knowledge of coronary artery disease risk factors and adherence to a healthy lifestyle," *Journal of Clinical Nursing*, vol. 27, pp. 4311–4320, 2018.
- [27] S.-W. Liou, D. Kurniadi, B.-R. Zheng, W.-Q. Xie, C.-J. Tien, and G.-J. Jong, "Classification of biomedical signal on IoT platform using support vector machine," in *Proceedings of IEEE International Conference on Applied System Innovation*, Chiba, Japan, April 2018.
- [28] Z. Taimeh, D. Duprez, and D. J. Garry, "Heart failure prevention," *Proceedings of the Congestive Heart Failure and Cardiac Transplantation*, vol. 6, no. 8, pp. 1120–1128, 2017.
- [29] C. Piazzese, M. Chiara Carminati, and E. G. Caiani, "Statistical shape models of the heart: applications to cardiac imaging," in *Statistical Shape and Deformation Analysis Methods*, pp. 445–480, Elsevier, Amsterdam, Netherlands, 2017.
- [30] R. G. Saboji and P. K. Ramesh, "A scalable solution for heart disease prediction using classification mining technique," in *Proceedings of the International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)*, Chennai, India, August 2017.
- [31] A. Mustaqeem, S. M. Anwar, A. R. Khan, and M. Majid, "A statistical analysis based recommender model for heart disease patients," *International Journal of Medical Informatics*, vol. 108, pp. 134–145, 2017.
- [32] K. Gilbert, B. Pontre, C. J. Occlshaw, B. R. Cowan, A. Suinesiaputra, and A. A. Young, "4D modelling for rapid assessment of biventricular function in congenital heart disease," *The International Journal of Cardiovascular Imaging*, vol. 34, no. 3, pp. 407–417, 2018.
- [33] K. T. Lee, A. L. Hour, B. C. Shia, and P. H. Chu, "The application and future of big database studies in cardiology: a single-center experience," *Acta Cardiologica Sinica*, vol. 33, no. 6, pp. 581–587, 2017.
- [34] C. A. Ledezma, "A big data approach to myocyte membrane analysis: using populations of models to understand the cellular causes of heart failure," in *Proceedings of the Computing in Cardiology*, Vancouver, Canada, September 2017.
- [35] L. X. Lu and S. F. Lu, "Distance, quality, or relationship? Interhospital transfer of heart attack patients," *Production and Operations Management*, vol. 27, no. 12, pp. 2251–2269, 2017.

- [36] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature analysis of coronary artery heart disease data sets," in *Proceedings of the International Conference on Communication, Management and Information Technology (ICCMIT 2015)*, Prague, Czech Republic, April 2015.
- [37] H. Mizuno, "Mapping of ventricular tachycardia in patients with structural heart disease," *Journal of Arrhythmia*, vol. 30, no. 4, pp. 283–291, 2014.
- [38] S. Abohammar, M. A. ElSaidy, D. Fathalla, and M. Aldosarri, "Baseline characteristics of patients with heart failure and preserved ejection fraction at admission with acute heart failure in Saudi Arabia," *The Egyptian Heart Journal*, vol. 69, no. 1, pp. 21–28, 2017.
- [39] A. Idri and I. Kadi, "A data mining-based approach for cardiovascular dysautonomias diagnosis and treatment," in *Proceedings of the 2017 IEEE International Conference on Computer and Information Technology*, Helsinki, Finland, August 2017.
- [40] A. B. Christian, L. Sharma, and S.-L. Wu, "AED: adaptive energy-efficient data transmission scheme for heart disease detection," in *Proceedings of the 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress*, Orlando, FL, USA, November 2017.
- [41] C. A. Alexander and a. L. Wang, "Big data analytics in heart attack prediction," *Journal of Nursing and Care*, vol. 6, no. 2, pp. 1–9, 2017.
- [42] Q. Wang, W. Xiong, Y. Zhang et al., "Remote analysis of myocardial fiber information in vivo assisted by cloud computing," *Future Generation Computer Systems*, vol. 85, pp. 146–159, 2018.
- [43] C. A. Ledezma, "Evaluating the risks of arrhythmia through big data: automatic processing and neural networks to classify epicardial electrograms," in *Proceedings of the Computing in Cardiology 2017*, Rennes, France, September 2017.
- [44] A. M. S. Omar, "Imaging based big data and machine learning framework for rapid phenotyping OF left ventricular diastolic function," *Journal of the American College of Cardiology*, vol. 67, no. 13, p. 1614, 2016.
- [45] Y. Shi, "Automatic knowledge extraction and data mining from echo reports of pediatric heart disease: application on clinical decision support," in *Proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Changsha, China, October 2015.
- [46] A. Dag, K. Topuz, A. Oztekin, S. Bulur, and F. M. Megahed, "A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival," *Decision Support Systems*, vol. 86, pp. 1–12, 2016.
- [47] O. Behadada, M. Trovati, M. Chikh, and N. Bessis, "Big data-based extraction of fuzzy partition rules for heart arrhythmia detection: a semi-automated approach," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 2, pp. 360–373, 2016.
- [48] G. Vaishali and V. Kalaivani, "Big data analysis for heart disease detection system using map reduce technique," in *Proceedings of the 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, Kovilpatti, India, January 2016.
- [49] C. M. Bucci, W. E. Legnani, and R. L. Armentano, "Clustering of cardiovascular risk factors highlighted the coronary artery calcium as a strong clinical discriminator," *Health and Technology*, vol. 6, no. 3, pp. 159–165, 2016.
- [50] B. Urrea, "Mobile health initiatives to improve outcomes in primary prevention of cardiovascular disease," *Current Treatment Options in Cardiovascular Medicine*, vol. 17, no. 12, pp. 1–12, 2015.
- [51] A. P. Hill, M. D. Perry, N. Abi-Gerges et al., "Computational cardiology and risk stratification for sudden cardiac death: one of the grand challenges for cardiology in the 21st century," *The Journal of Physiology*, vol. 594, no. 23, pp. 6893–6908, 2016.
- [52] H. H. Nguyen and J. N. A. A. Silva, "Use of smartphone technology in cardiology," *Trends in Cardiovascular Medicine*, vol. 26, no. 4, pp. 376–386, 2016.
- [53] S. Walton, K. Berger, J. Thiyagalingam et al., "Visualizing cardiovascular magnetic resonance (CMR) imagery: challenges and opportunities," *Progress in Biophysics and Molecular Biology*, vol. 115, no. 2-3, pp. 349–358, 2014.
- [54] R. U. Shah and J. S. Rumsfeld, "Big data in cardiology," *European Heart Journal*, vol. 38, no. 24, pp. 1865–1867, 2017.
- [55] J. Kim, "Big data, health informatics, and the future of cardiovascular medicine," *Journal of the American College of Cardiology*, vol. 69, no. 7, pp. 899–902, 2017.
- [56] J. Lee, E. Ribey, and J. R. Wallace, "A web-based data visualization tool for the MIMIC-II database," *BMC Medical Informatics and Decision Making*, vol. 16, no. 15, pp. 1–8, 2016.
- [57] E. Lau, K. E. Watson, and P. Ping, "Connecting the dots," *Circulation*, vol. 134, no. 5, pp. 362–364, 2016.
- [58] G. Kopanitsa, A. Karpov, G. Lakovenko, and A. Laskovenko, "Implementation of a web portal for diabetes patients using open source data visualization libraries," *Studies in Health Technology and Informatics*, vol. 224, pp. 189–194, 2016.
- [59] M. Ciecholewski, "Ischemic heart disease detection using selected machine learning methods," *International Journal of Computer Mathematics*, vol. 90, no. 8, pp. 1734–1759, 2013.
- [60] A. Asante-Korang and J. P. Jacobs, "Big Data and paediatric cardiovascular disease in the era of transparency in health-care," *Cardiology in the Young*, vol. 26, no. 8, pp. 1597–1602, 2016.
- [61] M. Zastrow, "Data visualization: science on the map, Easy-to-use mapping tools give researchers the power to create beautiful visualizations of geographic data," *Nature International Weekly Journal of Science*, vol. 519, no. 7541, pp. 1–2, 2015.
- [62] S. B. Scruggs, "Harnessing the heart of big data," *Circulation Research*, pp. 1–11, 2015.
- [63] M. W. Jenkins and A. M. Rollins, "4-D OCT in developmental cardiology," *Optical Coherence Tomography*, vol. 116, pp. 2003–2023, 2015.
- [64] A. H. Basori, M. R. B. D. A. Kadir, R. M. Ali, F. Mohamed, and S. Kadiman, "Kinect-based gesture recognition in volumetric visualisation of heart from cardiac magnetic resonance (CMR) imaging," *Virtual, Augmented Reality and Serious Games for Healthcare I, Intelligent Systems Reference Library*, vol. 68, pp. 79–92, 2014.
- [65] T. J. Brigham, "Feast for the eyes: an introduction to data visualization," *Medical Reference Services Quarterly*, vol. 35, no. 2, pp. 215–223, 2016.
- [66] J. D'hooge, "Cardiac 4D ultrasound imaging," in *Biomedical Image Processing, Biological and Medical Physics, Biomedical Engineering*, pp. 81–104, Springer, Berlin, Germany, 2011.
- [67] A. Dag, K. Topuz, A. Oztekin, S. Bulur, and F. M. Megahed, "A probabilistic data-driven framework for scoring the

- preoperative recipient-donor heart transplant survival,” *Decision Support Systems*, vol. 86, pp. 1–12, 2018.
- [68] P. M. Griffin, H. B. Nembhard, C. J. DeFlitch, N. D. Bastian, H. Kang, and D. A. Muñoz, “Health analytics,” in *Healthcare Systems Engineering* European society of cardiology, Sophia Antipolis, France, 2016.
- [69] F. Yang, “A composite visualization method for electrophysiology-morphous merging of human heart,” *BioMed Eng OnLine*, vol. 16, no. 70, pp. 1–16, 2017.
- [70] V. Mayer-Schönberger, “Big Data for cardiology: novel discovery?” *European Heart Journal*, vol. 37, no. 12, pp. 996–1001, 2016.
- [71] Y. Abdallah, A. Abdelhamid, T. Elarif, and A.-B. M. Salem, “Intelligent techniques in medical volume visualization,” *Procedia Computer Science*, vol. 65, pp. 546–555, 2015.
- [72] S. Nazir, S. Khan, H. U. K. S. Ali, I. García-Magariño, R. B. Atan, and M. Nawaz, “A comprehensive analysis of healthcare big data management, analytics and scientific programming,” *IEEE Access*, vol. 8, 2020.
- [73] A. U. Haq, “Intelligent machine learning approach for effective recognition of diabetes in the e-healthcare using clinical data,” *Sensors*, vol. 20, no. 9, p. 2649, 2020.
- [74] S. Khan, S. Nazir, I. García-Magariño, and A. Hussain, “Deep learning based urban big data fusion in smart cities: towards traffic monitoring and flow-preserving fusion,” *Computers & Electrical Engineering*, vol. 89, Article ID 106906, 2020.
- [75] X. Liao, S. Nazir, Y. Zhou, M. Shafiq, and X. Qi, “User knowledge, data modelling and visualization- handling through fuzzy logic based approach,” *Journal of Structural Geology*, vol. 141, 2020.
- [76] S. Nazir, S. Ali, M. Yang, and Q. Xu, “Deep learning algorithms and multi-criteria decision making used in big data- a systematic literature review,” *Complexity*, vol. 2020, Article ID 2836064, 18 pages, 2020.
- [77] J. Chi, Y. Li, J. Huang et al., “A secure and efficient data sharing scheme based on blockchain in industrial Internet of things,” *Journal of Network and Computer Applications*, vol. 167, p. 102710, 2020.

Research Article

Segmentation and Classification of Heart Angiographic Images Using Machine Learning Techniques

Abdullah,¹ Muhammad Hameed Siddiqi,² Yousef Salamah Alhwaiti,² Ibrahim Alrashdi,² Amjad Ali,¹ and Mohammad Faisal ³

¹Department of Computer and Software Technology, University of Swat, KPK, Mingora, Pakistan

²Department of Computer Science, Jouf University, Sakakah, AlJouf, Saudi Arabia

³Department of CS & IT, University of Malakand, Chakdara, KPK, Pakistan

Correspondence should be addressed to Mohammad Faisal; mfaisal@uom.edu.pk

Received 6 November 2020; Accepted 9 January 2021; Published 28 January 2021

Academic Editor: Shaukat Ali

Copyright © 2021 Abdullah 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.

Heart angiography is a test in which the concerned medical specialist identifies the abnormality in heart vessels. This type of diagnosis takes a lot of time by the concerned physician. In our proposed method, we segmented the interested regions of heart vessels and then classified. Segmentation and classification of heart angiography provides significant information for the physician as well as patient. Contradictorily, in the mention domain of heart angiography, the charge is prone to error, phase overwhelming, and thought-provoking task for the physician (heart specialist). An automatic segmentation and classification of heart blood vessels descriptions can improve the truthfulness and speed up the finding of heart illnesses. In this work, we recommend a computer-assisted conclusion arrangement for the localization of human heart blood vessels within heart angiographic imageries by using multiclass ensemble classification mechanism. In the proposed work, the heart blood vessels will be first segmented, and the various features according to accuracy have been extracted. Low-level features such as texture, statistical, and geometrical features were extracted in human heart blood vessels. At last, in the proposed framework, heart blood vessels have been categorized in their four respective classes including normal, block, narrow, and blood flow-reduced vessels. The proposed approach has achieved best result which provides very useful, easy, accurate, and time-saving environment to cardiologists for the diagnosis of heart-related diseases.

1. Introduction

Cardiovascular diseases caused first death many years ago in North America. Abnormality or blockage of coronary artery of the heart instigates it. In the last decade, very crucial developments were made to improve the diagnosis of patient in cardiology in order to increase the rate of survival. For the diagnosis of the coronary heart diseases, angiography is used.

There are four classes of the heart blood vessels, i.e., blood flow-reduced vessels, narrow vessels, block vessels, and normal vessels. It is very time-consuming and not so much significant to manually detect the classes of heart blood vessels in manual diagnoses. There are various artificial intelligence techniques and digital image processing

modalities (computer-aided system) which efficiently segment and categorize the human heart blood vessels in angiography images [1]. The classification system in angiography obtains a sequence of images and chooses the section of interest (SOI) to segment an image, like left ventricle [2]. Intravenously sharpens improved heart angiogram X-ray was complete throughout a single inhalation clench, and echocardiograph-matching images were reconstructed with perception. Two investigators evaluated each part and section of the coronary artery greater than or equal to 3.0 millimeter having no stents and consociated them with algebraic average heart angiography. Patients were classified on the basis of average heart rate, and standard deviations are taken in three clusters [3]. To detect significant scratches of diameter greater than or equal to 50

percent decrease, diagnostic performance of data related to heart was contrasted with data of heart angiogram X-ray's quantification [4].

Cardiology images are used a lot for analyzing and management of heart patients. A busy cardiologist is facing the increasingly unpromising task to keep informed of the current knowledge and the new changes in the analysis and treatment of diseases [5]. Techniques of imaginings, such as scoring of calcium in heart arteries and diagnosis of heart diseases with angiography, are achieved by revised dominion associated with present-day machine-learning approaches in the domain of heart diseases [6].

The framework will improve the clinical instruments and examination program in order to label and classify into images of the heart vessels of a human. When human heart vessels are classified and segmented automatically, it will help cardiologists to diagnose various heart infections. The goal of our research is to determine if the proposed image allotment methods are effective in classification and segmentation of human heart vessels. Our main attention is towards the classification of the human heart vessels in the field of practical cardiology. In various medical fields worked on, the main agony is the data which are not restful to relate to large quantities of models. The image information was taken from the research center of cardiology located in Hayatabad (HMC), Peshawar, Pakistan. There are 400 angiographic imaginings in the data set accumulated under the supervision of a heart doctor. These data consist of many conditions due to which human heart vessels are affected. For the purpose to help cardiologist in easily diagnosing various long-lasting heart diseases such as cardiac arrest, pain in the leg, and pain in the chest, the novelty of this propose work is that first time we categorize the ventricle heart vessels into four classes in supervision of a heart specialist/cardiologist. Furthermore, horizontal and vertical edge detection of heart vessels, which is explained in Segmentation, is new idea.

2. Related Study

Langley et al. created an algorithm for myocardial ischemia classification. In this paper, the proposed method will improve the algorithm. The peak value of sensitivity of the Langley classifier is 99.0 percent. However, its lower specificity value is 93.3 percent. For the improvement of the specificity, this algorithm recategorizes the ischemic events of the Langley classifier. Support vector machine is used for classification. Standard deviation means, standard deviation peak value, and the starting standard deviation are the features used. The specificity is increased from 92.3 percent to 93.3 percent by the classifier. However, there is a disadvantage, i.e., the reduction in sensitivity from 99.0 percent to 97.5 percent as a result of which the total correctness is reduced from 95.6 to 94.8. The algorithm fulfills the promise of increasing the specificity. However, more work is to be done in order to find features that do not affect the sensitivity to decrease [7].

Automatic coronary artery centerline when extracted from 3D CT angiography (CTA) has significant clinical

applications for diagnosing the atherosclerotic heart disease. Most of the work is dominated by segmentation where the complete coronary artery system is segmented like trees using a computer. However, the process, where various branches of vessels (defined by their medical semantics), is of much clinical significance and is performed manually. A hierarchical machine learning approach is proposed in this paper that will help in tackling of the tubular structure parsing problem in medical imaging. This framework will also help in parsing tasks of other tubular structures generically [8]. In this study, an intelligent system based on genetic-support vector machines is proposed. This intelligent system deals with the combination of feature extraction and classification from measured Doppler signal waveforms at the heart valve using the Doppler ultrasound. GSVM is used in this study for the diagnosis of heart valve diseases. The outcomes presented GSVM's success in detecting Doppler coronary sounds. The average rate of precise classifying rate was approximately 95 percent [9].

The categories of the infections in which coronary plaques play significant role to infect the heart with diseases were determined. Therefore, the previous focuses, either to detect specified class of coronary infections caused by plaque or to distinguish between infected and noninfected coronary vessels, ignore [10]. Various categories of coronial infections caused by plaque were detected or recognized using the data of affected patients. Typical imaging approaches such as electrocardiogram (ECG), computed tomography (CT), magnetic resonance imaging (MRI) and radiology tools of choice are used to diagnose cardiovascular diseases. All of these modalities will be impacted essentially and enduringly by machine level learning and neural network. However, the image processing context is discussed mainly with deep learning, and we present that the effect is more severe than this. Nonetheless, the procedure in which the outcomes brought to doctors are nontrivial, and we have also discussed our experience with this algorithm's deployment [11].

With the beginning and speedy presentation of artificial intelligence (AI), Internet of things, mobile phones, and devices with sensors such as smart watch, we enter into the age of automated, isolated, and portable services. As directed by the World Health Organization (WHO), cardiovascular infection is the modern disease. However, prediction rate of coronary disease sufferers can be made conceivably and made excessive with initial finding and analysis. We portray a framework that observes robotized coronary heart health making full use of cell phone and wearable sensors [12]. The proposed technique is capable of reliably and efficiently assessing the position and radius of coronary vessels (arteries) in the heart, which is based on straight facts from the heart image data collection. The framework can be trained with minimal data on the train heart image and allows for rapid automated or interactive extraction of the coronary artery from CCTA heart images once trained [13]. The key aim of this work is to establish an automated method for CT images to detect calcified coronary plaques. In comparison to the avant-garde, both native and angio data sets are processed in this technique, and this dual information is used for identification and evaluation of calcified plaques.

The success rate of the proposed method is stated by the authors as 85%. The research focuses only on the plaques being calcified [14]. The proposed method consists of pre-processing, extraction of features, and classification. Using preprocessing techniques, the vessels in the coronary image are enhanced, and then features are extracted from these images that are supplied to the CANFIS classifier. This classifier classifies the coronary image of the test given as either normal or abnormal. In addition, if the proposed method classifies the test image as irregular, the blockage is detected and segmented [15]. Coronary computed tomography angiography (CCTA) is now a well-established, noninvasive cardiovascular disease evaluation modality. The proposed framework is based on extracting data from CCTA as well as non-contrast-enhanced cardiac CT scans, and ML has been increasingly used to improve performance [16].

3. Proposed Methodology

In the current era, numerous applications curiously want the high pixel quality in sense of resolving power to get improved images by means of low eminence info over separate apparatus such as angiography device angiogram, computer tomography scanner, and MRI. Medical imaging area investigation and finding are too much tough objectives to gain from very low-slung eminence images in sense of pixel quality and color scheme. On the basis of this thoughtful matter, the high resolving images are gained using the super resolution method with automated modalities. It is very thought-provoking objective to evaluate the imageries taken with angiogram equipment with low-slung pixel and color quality and composite tree-like arrangement as well as low accuracy and resolving power. Moreover, the classification and segmentation of human heart vessels also provides an appropriate background on the basis of images processing idea, and modalities can be very easily and flexibly evaluated. The proposed scheme consists of four key steps: adjustment of contrast, segmentation, features extraction and features selection, and classification with multiclass-assembled support vector machine. The proposed hierarchal method is used to achieve best result which provides very useful and time-saving environment for diagnosis of the heart-related diseases to the cardiologist/physician.

3.1. Contrast Adjustment. The first phase of the proposed method is contrast adjustment in which we increase the sharpness intensity of heart blood vessels which is obtained from angiography as shown in Figure 1.

3.2. Segmentation. In this phase, we divided the human heart angiographic image into X -axis and Y -axis edges as shown in Figure 2. After the division of vertical and horizontal edges, we combine both the edges to refine the shape of angiography images of tree-like structure. After the concatenation of vertical and horizontal edges, we apply Laplacian filter on the targeted image. Laplacian filter has two parameters: alpha and gamma which sort out the

information below 0.5, respectively, on X -axis and Y -axis. Our proposed approach gives better results as shown in Figure 3.

3.3. Features Extraction and Selection. Here, in the proposed method, we extracted the local binary pattern (LBP) and histogram-oriented gradient (HOG). In LBP, we take a mask which is further used on a target image and take binary patterns, while in HOG, we divide the image into subsections, take histogram of every section, and then combine the section histogram. When we extracted both LBP and HOG features, we combined both the features and saved it in feature vector for classification. Examples of LBP features, pixel value of image, and selected mask are given in Tables 1 and 2, respectively.

3.4. Classification. The final step is classification when we extract features from LBP and HOG and then combine it and save it in a feature vector. On the basis of these features, we trained a multiclass-assembled classifier as support vector machine on the basis of cardiovascular images data from angiogram images. Here, in the proposed method, we select a vector which best fits the data after extraction of features from the proposed data set according to the selected framework. So, for this purpose, we select a method called principle component analysis with the help of which we reduced the redundant features. The cross validation used here is as follows: 30% data used for testing purpose and 70% for training purpose in the data set. The proposed framework for classification and segmentation of heart images is shown in Figure 4.

4. Results and Discussion

Investigation remained accepted and appropriate on the basis of human heart angiography blood carrier's data collected from two hospitals of Peshawar cardiology center of complex hospital Hayatabad and teaching hospital Khyber Peshawar. The proposed system is simulated on the MATLAB tool. It is a well-established tool used by researchers for statistical and data analysis due to diverse set of libraries and gives optimized, efficient, and accurate results. To collect the minced multiplicity of human heart blood carriers, heart specialists were desired to classify the human heart angiographic imageries into their consistent categories, i.e., normal blood carrier, blood flow-reduced carrier, block blood carrier, and narrow blood carrier. Segmentation and classification are applicable for ventricular vessels, which may be normal vessels, block vessels, blood flow-reduced vessels, and narrow vessels. The solid results were predictable to the database and are accumulated to judge the result of various catalogue procedures. The investigation was accomplished on 400 data sets of human heart blood carrier images, consisting of both defective and nondefective images of human heart vessels. From acknowledged facts storehouse of descriptions, only the heart angiography images were integrated and catalogued into their own 4 categories as shown in Figure 5. The consequences after that

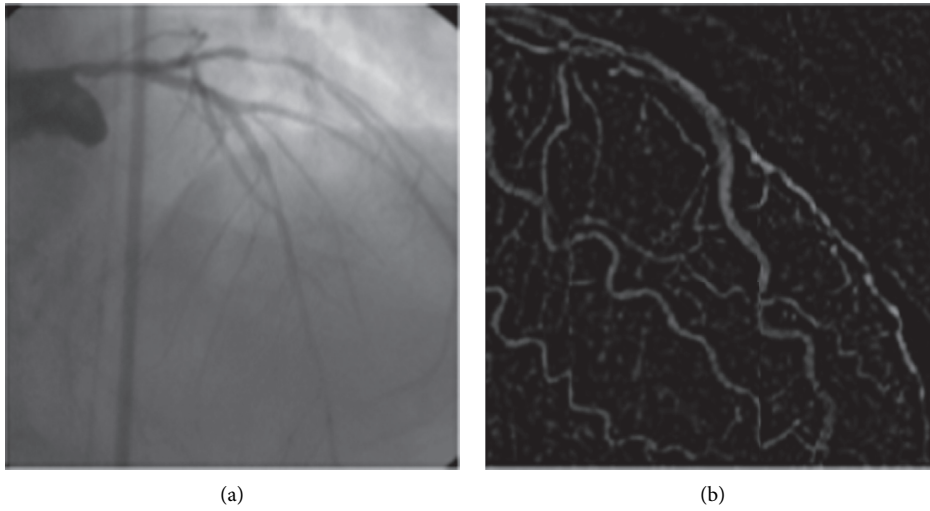


FIGURE 1: Divergence balance.

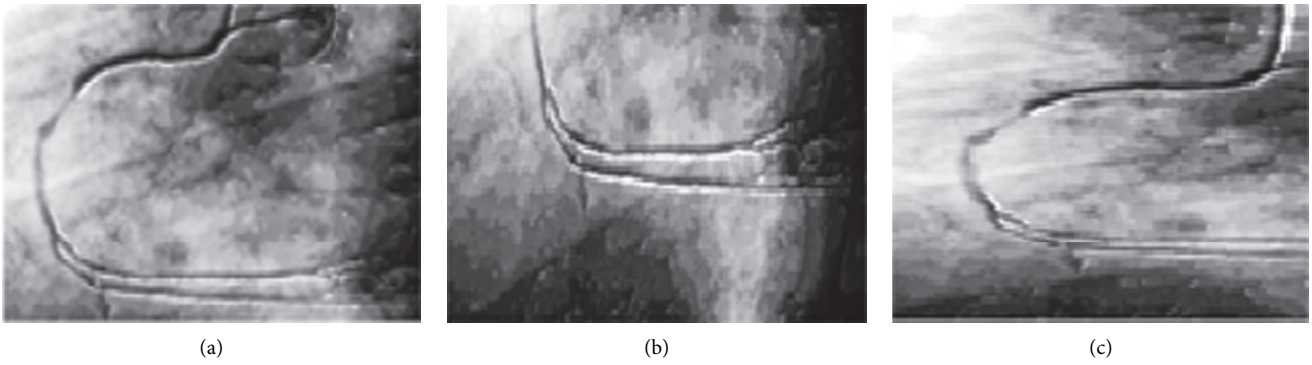


FIGURE 2: The process of segmentation.

Original image	Adaptive thresholding	Global thresholding	Proposed thresholding

FIGURE 3: Segmentation approaches [17] with the proposed system.

TABLE 1: Pixel value of an image.

55	66	45
77	60	32
52	33	88

TABLE 2: Pattern after applying mask which is 0100110001.

0	1	0
1	0	0
0	0	1

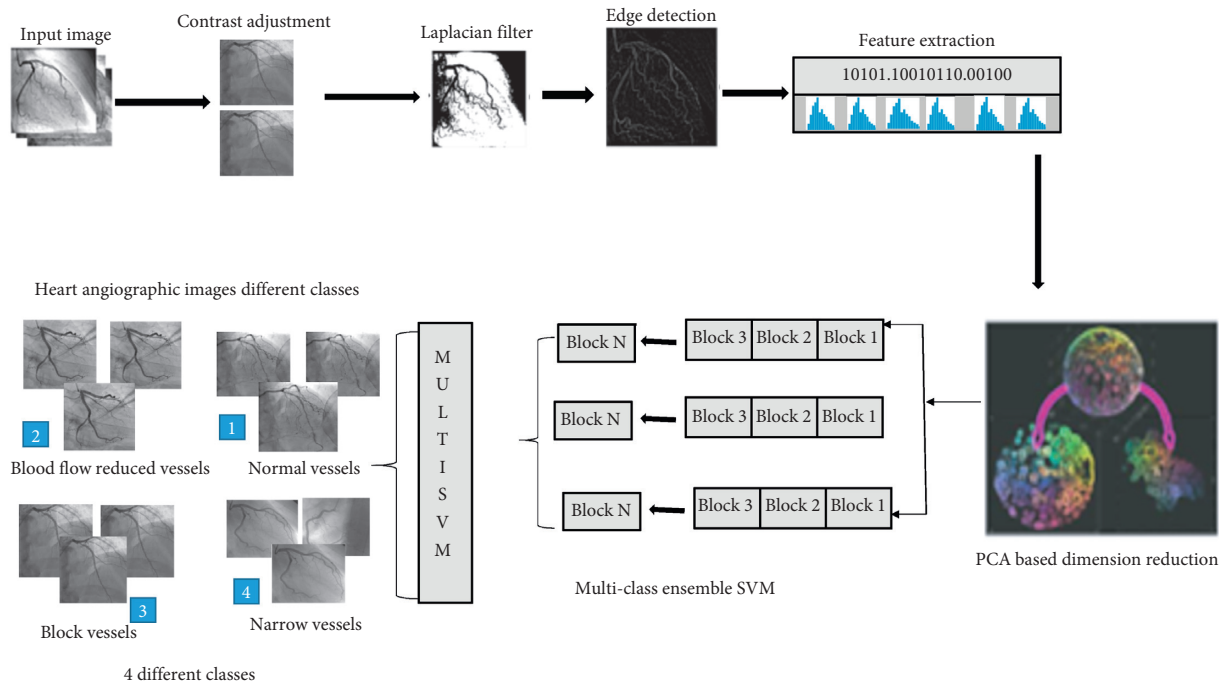


FIGURE 4: The proposed framework for classification and segmentation of heart images.

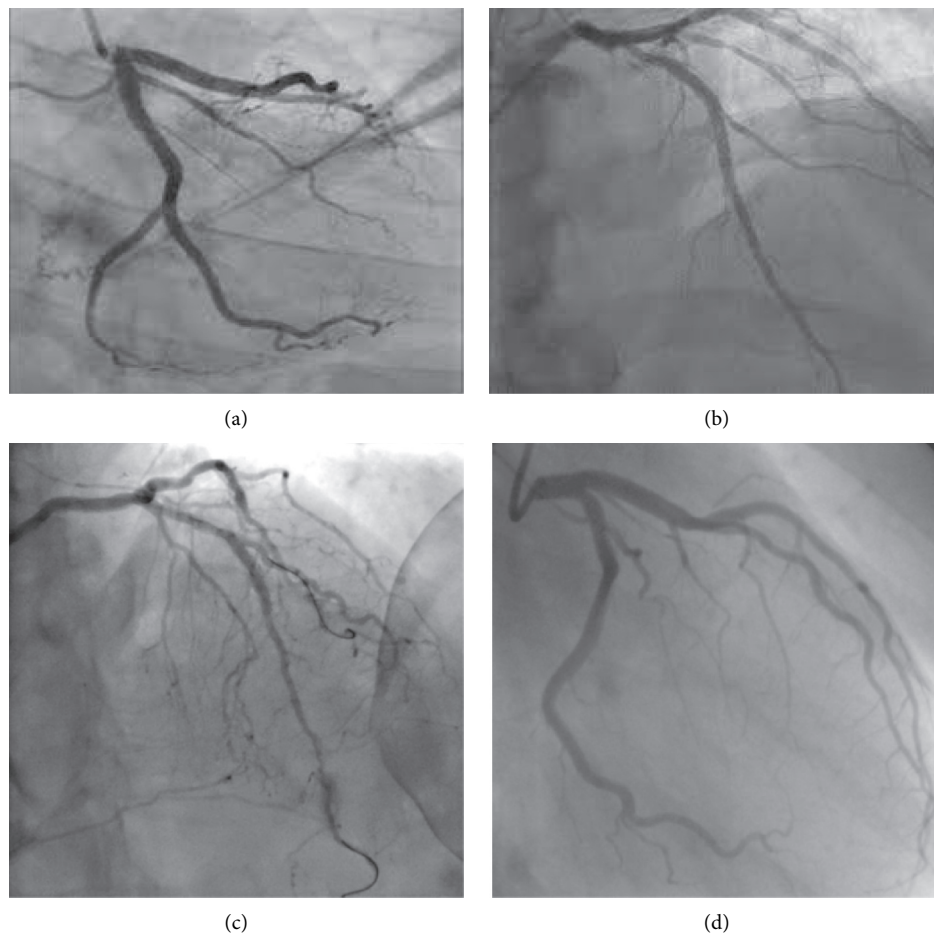
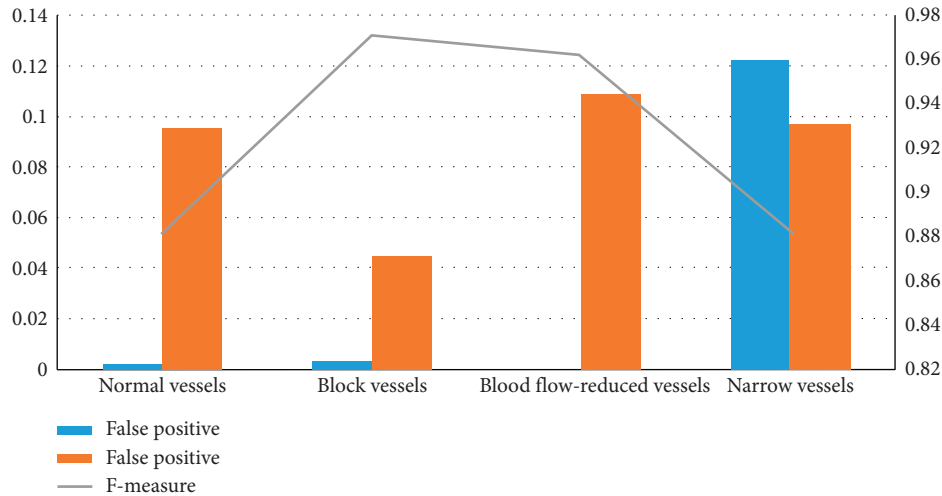


FIGURE 5: Results after classification. (a) Block vessels. (b) Normal vessels. (c) Blood flow-reduced vessels. (d) Narrow vessels.

TABLE 3: Classes with number of images.

S. No.	Human heart interior groups	Number of data sets
I	Block	Hundred
II	Narrow	Hundred
III	Reduction of blood flow	Hundred
IV	Normal	Hundred
V	Overall	Four-hundred

FIGURE 6: Results of false positive, false negative, and F -measure from the data set.

accompanying with the pulverized fact to evaluation the suitability of the predicted heart angiography classification technique. The essentials in between examinations are clarified in the succeeding scraps.

The data set after findings containing four-hundred images were collected from two hospitals of Peshawar in which one is the center of heart diseases cardiac research center of Hayatabad and other one is the teaching hospital of Khyber. All data set samples were in Joint Photographic Experts Group with dimension according to the coordinate system X -axis 960 while Y -axis is 1080 per pixel. Due to the requirement of MATLAB, the data set was adjusted to 128 by 128. Details of the data with repository are given in Table 3.

The convinced evaluation was accomplished to check the classification consequences of the proposed method. In the proposed method, six cardiologists were involved during independent evaluation to point out the parts of human heart in images; these image data sets of different patients are taken through a device called angiogram, and the process is called angiography. Under the supervision of a cardiologist, respective categories, i.e., normal, blood flow-reduced, narrow, and block vessels were set physically. The region of interest (segmentation) by physically of human heart angiography data set as booked minced truth for valuation with heart vessels segmentation by suggested framework. There are three metrics used to find out the false-positive ratio, false-negative ratio, and F -measure as follows:

$$\text{False-positive ratio} = \text{FP}/\text{TN} + \text{FP}.$$

$$\text{False-negative ratio} = \text{FN}/\text{TP} + \text{FN}.$$

$$F\text{-measure} = \text{FPR} + \text{FNR}/\text{FPR} + \text{FNR} * 2.$$

Here, TP is true positive, FP is false positive, and FN is false negative. True positive specifies the interest pixel, true negative specifies noninterested value, false positive identifies the heart vessel's nonconcerned section, and false negative identifies the section where, in the suggested procedure, we are not intended in the heart vessel pixel. Wrongly accepted as noninteresting cell pixels show that the F -measurement of the modalities for four-hundred blood vessels in the heart vessels in all kinds of images in both parts of perception and in the dissecting of vessels in machine learning sections is more than 85 percent. This particular consequence of dissection is more important, widespread, careful, and far closer to the manual splitting off of heart angiography, i.e., real evidence. The false positive, false negative, and F -measure results are shown in Figure 6 and Table 4 [18].

In pattern recognition, information retrieval, and classification (machine learning) precision (also called positive predictive value) is a fraction of the relevant instances among the retrieved instances, while retrieval (also known as sensitivity) is a fraction of the retrieved relevant instances among all relevant instances. Statistical approaches for the study of data from radio-immunoassay are provided. The viability of using the logic transformation has been empirically checked, and the benefits arising from the study of the data in this manner are considered. The statistical transformations take into account the heteroscedasticity of the error inherent in a heart angiography normal and the abnormal extant in the confidence intervals about the estimates of the values of individual samples [14].

TABLE 4: Overall accuracy of the proposed system.

Heart angio image types	False positive	False positive	F-measure
Normal vessels	0.002	0.095	0.882
Block vessels	0.003	0.045	0.971
Blood flow-reduced vessels	0.000	0.109	0.963
Narrow vessels	0.122	0.097	0.882

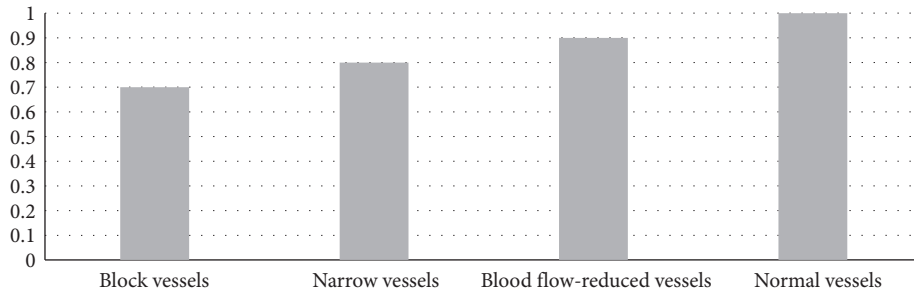


FIGURE 7: Specificity.

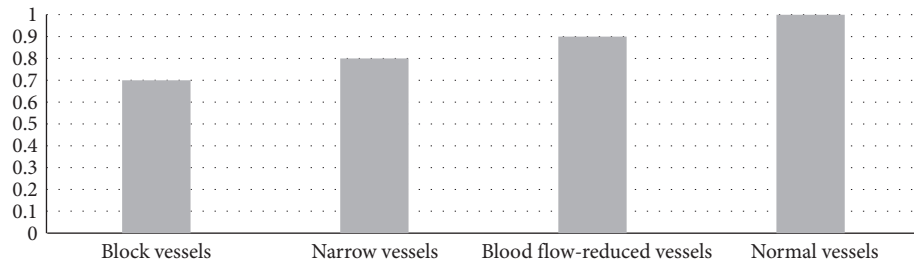


FIGURE 8: Precision.

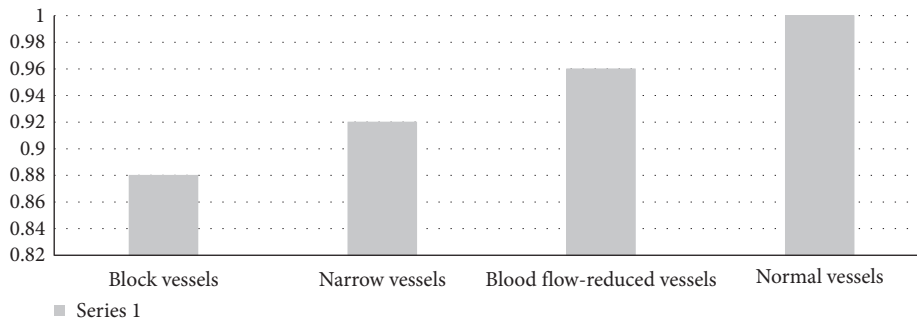


FIGURE 9: Accuracy.

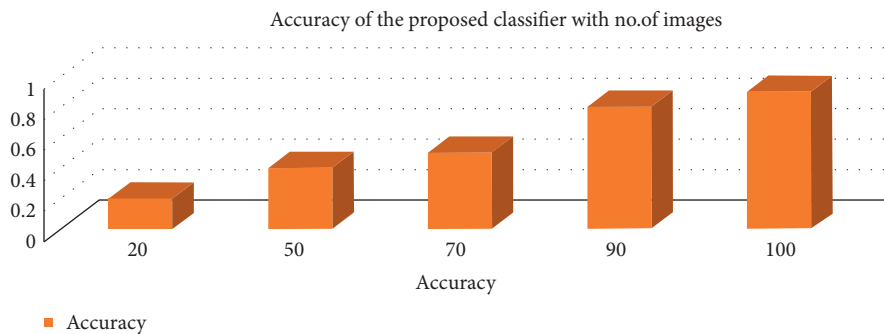


FIGURE 10: Performance of the entire framework.

The specificity = $TN/(TN + FP)$, precision = $TP/(TP + FP)$, and the results are shown in Figures 7 and 8, respectively.

The accuracy of each class is shown in Figure 9 and is given as follows: accuracy = number of correct prediction/total number of prediction.

The accuracy of the overall frame framework is shown in Figure 10 and is given as follows: accuracy = one prediction/total number of prediction.

5. Conclusion

The suggested method of segmentation is to concatenate the identification of the edge and the Laplacian filter. In the proposed method, the combined edge detection system and the Laplacian filter provide full information from cardiac angiographic images. Edge detection plays an important role in the tree, such as the structure of cardiac angiographic images, and the Laplacian filter has two alpha and beta parameters that extract information from cardiac angiographic images below 0.4 on both coordinates. In this paper, we first segment the input image and extract function that gives the best accuracy to the multi-SVM Classifier. The method, however, suffers from the complete classification of narrow boats, but the overall result of the proposed modality is good enough. Efficiency is improved due to the execution of each stage of the framework precisely and mostly in the resizing of images to save processing time. The proposed method provides the doctor/cardiologist and patient with a simple and reliable, time-saving diagnostic approach, as well as a health-related approach [19–21].

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] J. Prinyakupt and C. Pluempitiwiriyawej, "Segmentation of white blood cells and comparison of cell morphology by linear and naïve Bayes classifiers," *BioMedical Engineering OnLine*, vol. 14, no. 1, p. 63, 2015.
- [2] J. Wen, W. Shuai, T. Ding, Y. Feng, J. Zhang, and S. Wang, "Reproductive risk factors for angiographic obstructive coronary artery disease among postmenopausal women," *Menopause*, vol. 27, no. 12, pp. 1403–1410, 2020.
- [3] G. Brown, J. J. Albers, L. D. Fisher et al., "Regression of coronary artery disease as a result of intensive lipid-lowering therapy in men with high levels of apolipoprotein B," *New England Journal of Medicine*, vol. 323, no. 19, pp. 1289–1298, 1990.
- [4] M. Faisal, I. Ali, M. Sajjad Khan, S. Min Kim, and J. Kim, "Establishment of trust in Internet of things by integrating trusted platform module: to counter cyber security challenges," *Complexity*, vol. 2020, Article ID 6612919, 9 pages, 2020.
- [5] J. Li, M. Nazir Jan, and M. Faisal, "Big data, scientific programming, and its role in Internet of industrial things: a decision support system," *Scientific Programming*, vol. 2020, Article ID 8850096, 7 pages, 2020.
- [6] G. Singh, S. J. Al'Aref, M. Van Assen et al., "Machine learning in cardiac CT: basic concepts and contemporary data," *Journal of Cardiovascular Computed Tomography*, vol. 12, no. 3, pp. 192–201, 2018.
- [7] E. Asher, A. Abu-Much, and N. L. Bragazzi, "CHADS2 and CHA2DS2-VASc scores as predictors of platelet reactivity in acute coronary syndrome," *Journal of Cardiology*, vol. 4, 2020.
- [8] A. R. van Rosendael, A. M. Bax, J. M. Smit et al., "Clinical risk factors and atherosclerotic plaque extent to define risk for major events in patients without obstructive coronary artery disease: the long-term coronary computed tomography angiography CONFIRM registry," *European Heart Journal Cardiovascular Imaging*, vol. 21, no. 5, pp. 479–488, 2020.
- [9] M. Zimmerman and R. J. Povinelli, "On improving the classification of myocardial ischemia using Holter ECG data," in *Computers in Cardiology*, pp. 377–380, IEEE, New York, NY, USA, 2004.
- [10] M. Faisal, I. Ali, M. S. Khan, J. Kim, and S. M. Kim, "Cyber Security and Key Management Issues for Internet of Things: Techniques, requirements, and challenges," *Complexity*, vol. 2020, Article ID 6619498, 9 pages, 2020.
- [11] M.-Y. Hung, N. G. Kounis, M.-Y. Lu, and P. Hu, "Myocardial ischemic syndromes, heart failure syndromes, electrocardiographic abnormalities, arrhythmic syndromes and angiographic diagnosis of coronary artery spasm: literature review," *International Journal of Medical Sciences*, vol. 17, no. 8, p. 1071, 2020.
- [12] E. Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10618–10626, 2009.
- [13] K. M. Johnson, H. E. Johnson, Y. Zhao, D. A. Dowe, and L. H. Staib, "Scoring of coronary artery disease characteristics on coronary CT angiograms by using machine learning," *Radiology*, vol. 292, no. 2, pp. 354–362, 2019.
- [14] M. R. Sanjay, P. Madhu, M. Jawaid, P. Sentharamaikkannan, S. Senthil, and S. Pradeep, "Characterization and properties of natural fiber polymer composites: a comprehensive review," *Journal of Cleaner Production*, vol. 172, pp. 566–581, 2018.
- [15] R. P. Durbin, "Letter: acid secretion by gastric mucous membrane," *The American Journal of Physiology*, vol. 229, p. 1726, 1975.
- [16] A. Ukil and S. Bandyopadhyay, "Automated cardiac health screening using smartphone and wearable sensors through anomaly analytics," in *Mobile Solutions and Their Usefulness in Everyday Life*, pp. 145–172, Springer, Berlin, Germany, 2019.
- [17] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical Image Analysis*, vol. 52, pp. 128–143, 2019.
- [18] A. P. Periasamy, R. Ravindranath, S. M. Senthil Kumar, W.-P. Wu, T.-R. Jian, and H.-T. Chang, "Facet- and structure-dependent catalytic activity of cuprous oxide/polypyrrole particles towards the efficient reduction of carbon dioxide to methanol," *Nanoscale*, vol. 10, no. 25, pp. 11869–11880, 2018.
- [19] R. Singh, A. K. Upadhyay, P. Chandra, and D. P. Singh, "Sodium chloride incites reactive oxygen species in green algae *Chlorococcum humicola* and *Chlorella vulgaris*: implication on lipid synthesis, mineral nutrients and antioxidant system," *Bioresource Technology*, vol. 270, pp. 489–497, 2018.

- [20] M. Sajjad, S. Khan, Z. Jan et al., "Leukocytes classification and segmentation in microscopic blood smear: a resource-aware healthcare service in smart cities," *IEEE Access*, vol. 5, pp. 3475–3489, 2017.
- [21] X. Liao, M. Faisal, Q. QingChang, and A. Ali, "Evaluating the role of big data in IIOT-industrial Internet of things for executing ranks using the analytic network process approach," *Scientific Programming*, vol. 2020, Article ID 8859454, 7 pages, 2020.
- [22] G. S. Handelman, H. K. Kok, R. V. Chandra et al., "Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods," *American Journal of Roentgenology*, vol. 212, no. 1, pp. 38–43, 2019.

Research Article

Augmentation in Healthcare: Augmented Biosignal Using Deep Learning and Tensor Representation

Marwa Ibrahim,¹ Mohammad Wedyan ,² Ryan Alturki ,³ Muazzam A. Khan,⁴
and Adel Al-Jumaily ^{5,6}

¹Faculty of Engineering and Information Technology, University of Technology Sydney, NSW 2000, Sydney, Australia

²Faculty of Artificial Intelligence, Al-Balqa Applied University, Al-Salt 19117, Jordan

³Department of Information Sciences, College of Computer and Information Systems, Umm Al-Qura University, P.O. Box 715, Makkah, Saudi Arabia

⁴Department of Computer Science, Quaid-i-Azam University, Islamabad 44000, Pakistan

⁵School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, Australia

⁶College of Computing, Fahad Bin Sultan University, Tabuk, Saudi Arabia

Correspondence should be addressed to Adel Al-Jumaily; adel.al-jumaily@ieee.org

Received 26 October 2020; Revised 22 November 2020; Accepted 12 January 2021; Published 27 January 2021

Academic Editor: Iván García-Magariño

Copyright © 2021 Marwa Ibrahim 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.

In healthcare applications, deep learning is a highly valuable tool. It extracts features from raw data to save time and effort for health practitioners. A deep learning model is capable of learning and extracting the features from raw data by itself without any external intervention. On the other hand, shallow learning feature extraction techniques depend on user experience in selecting a powerful feature extraction algorithm. In this article, we proposed a multistage model that is based on the spectrogram of biosignal. The proposed model provides an appropriate representation of the input raw biosignal that boosts the accuracy of training and testing dataset. In the next stage, smaller datasets are augmented as larger data sets to enhance the accuracy of the classification for biosignal datasets. After that, the augmented dataset is represented in the TensorFlow that provides more services and functionalities, which give more flexibility. The proposed model was compared with different approaches. The results show that the proposed approach is better in terms of testing and training accuracy.

1. Introduction

In healthcare systems, data are not publicly available, and these data are limited in nature too. For example, in the current pandemic, the COVID-19, no data are publicly available and some institutes have very limited data [1, 2]. As a result, machine learning and big data analytics cannot be performed on such limited data. One possible solution is to augment limited data and increase the data for testing and training of various machine learning algorithms. The main purpose of Data Augmentation (DA) is to increase the data size [2]. Also, DA is a technique that strongly invades the field of data mining and processing for regression and classification purposes, particularly in healthcare applications. The expression DA denotes the techniques used to

generate virtual samples. The created latent samples are introduced to the original data to produce a high-dimensional one. The newly generated augmented data are used in training the suggested model. DA algorithms become numerous. The manipulation between DA algorithms is to achieve high accuracy results and at the same time, implementing modest and rapid algorithm is a typical matter of talent [2]. The appropriately selected DA technique drives the accuracy values to a dramatic level. Researchers developed an approach to combine, search, and select the best augmentation scheme between deterministic, marginal, and conditional different augmentations methodologies. This developed approach was applied to three different classes of systems and achieved good results among these systems [1–3].

Augmentation may be applied in two domains; the first domain is the data domain, while the second one is the feature domain [4]. Many studies demonstrated the art of DA by generating numerous training samples [5]. Other studies focused on the advantage of DA and how it might act as an organizer to prevent associated overfitting during training of neural networks [6] and develop the execution to avoid problems that may be correlated with the classes that are not represented equally [7]. Many researchers prospected in the field of DA to achieve high accuracy values and enhance the classifier performance. A punch of distorted and warped samples of characters are generated by the DA technique [8]. This was not the only example of creating deformed samples of characters as it was reported in [9], where the malformed samples are generated in a random manner. The latter mentioned methodology was extended to be applied on backpropagation neural networks and reduced the error rate to 0.4% on the MNIST database [10]. After that, in [4], the researchers followed two augmentations techniques. The first applied augmentation technique was data wrapping or DA on the input of MNIST dataset before being introduced to neural networks, then the output features from the neural networks were augmented in the feature domain. They used SVM, ELM, and backpropagation neural networks as classifiers, where the accuracy percentage ranged between 97.75% and 100% for training samples.

In the same context, DA was hired by generating virtual samples [11]. Generally speaking, the virtual samples can be generated by following two techniques. The first methodology depends on generating virtual samples from important information. For example, in the field of image processing and recognition, we can generate virtual samples from the same image by producing a 3D view, which in turn helps in creating virtual samples for the same image from a different angle [12].

Consequently, the proposed model has the ability to enhance learning performance, especially when dealing with a few samples. A lot of these mentioned sample generation techniques have shown considerable potential to improve classification and prediction performance. In spite of that, none of the previous studies are built on the overlapping found in the features step. Accordingly, this article presents a new model based on generating a virtual sample that also considers solving overlaps between each of the features in the corresponding classes. Moreover, this model is distinct by its ability to create and treat with a massive amount of new virtual samples that are hundreds of thousands of samples rather than tens or hundreds of virtual samples.

In this article, previous works are presented in Section 2. In Section 3, we illustrated the proposed model that consists of data acquisition, random virtual generation equations, and experiment. Then, in Section 4, we presented and explained the results. Finally, in Section 5, the conclusion and future work are discussed.

2. Previous Work

DA was recalled and implemented in many studies. For example, in [13], the researchers established a relation between the iterative computational time for the expectation-

maximization procedure and extension of the space parameter with augmenting the data. The recognized relation was an expansion to the applied space parameter applied along with DA, where the iterative computation time was expected to be shorter. Earlier, scientists calculated the posterior probability for the augmented data when the normal likelihood could not be reached [14]. The DA was used in different fields as, for example, in [15] the concept of replicating the data in the field of chemistry. Image recognition was one of the fields where DA took innovative steps, as in [16]. The scholars applied manual augmentation techniques in conjunction with deep neural network that led to an enhanced achievement. Moreover, the experimenters implemented DA algorithm for hand-drawn dataset and a fine-tuned deep neural network to extract useful features from the introduced dataset [17]. Recently, the authors of [18] applied the DA Markov chain Monte Carlo (MCMC). Furthermore, the scholars in [19] applied augmentation in both data domain and feature domain along with using the neural network for an acoustic signal, while in [20], the authors applied augmentation to the speech signal to prove that the gap between real room impulse response and simulated one was reduced to its minimum value. In addition, the authors of [21] combined the deep belief networks and DA algorithm, adding gamma variables to the original signal. In the field of image processing, the scholars augmented the input image by generating a 3D copy to be processed in the neural network [22]. Finally, researchers in [23] applied the augmentation and balancing to the electroencephalography (EEG) signal.

On the other hand, the tensor representation was used in different research fields to allow for a better representation of the dataset. In 20th century, scientists paid attention to the value of tensors and their applications [24]. In the field of continuum mechanics, tensor fundamentals and enforcement were discussed [25]. A study addressed the tensor decomposition technique [26] and treated it as a generalization for matrix decay. In [27], the concept of deep tensor neural network (DTNN) was initially introduced where one of the layers was substituted by a double projection layer. Those two inserted are totally nonlinear. Therefore, any input vocabulary speech was mapped to the newly introduced in conjunction with a tensor layer. The model was capable of anticipating the next layer in the deep neural network design. The proposed model resulted in reducing error by 3% relatively. The researchers in [26] and in [28] developed a model that was able to estimate an approximation for tensor rank 1 by disintegrating tensor and estimating Canonical Polyadic Decomposition (CPD) by using a sparse matrix of the banded type. In 2014, the inequality of the M tensors was discussed in [29], where the upper and lower values for the eigenvalues were obtained. In [30], the authors demonstrated the modeling of earthquake waveform to estimate the moment tensor solution. The estimation of tensor parameters was deeply analyzed in [31]. Furthermore, tensor decomposition techniques were presented in [32] to give the opportunity for a more latent dataset than that based on the matrix domain. Recently, in [33], the authors applied the tensor decomposition on the genetic expression to a group of

latent components used to find a relation between any biological development and genetic variation.

In [34] to increase the accuracy of the soft sensor under the small sample issue, they proposed a new locally linear embedding based virtual sample generation approach. In the proposed approach, the first step is producing features from the original data space by using locally linear embedding. The next step is generating effective virtual samples in the sparse region of the original data by using a method of random interpolation and a backpropagation neural network. To test the performance of the proposed approach, a couple of studies were conducted: the first study is a process of high-density polyethylene and the second study is developing soft sensors for a production system of purified terephthalic acid. The outcomes showed that the precision with virtual samples improved for the soft sensor. Moreover, the proposed method achieved more accuracy than other approaches in virtual sample generation.

Finally, in [35], the study simulated the process of fishermen rectifying nets; this method was named Kriging-VSG and it was put forward to produce feasible virtual samples in data-sparse zones. This method was based on a distance-based criterion by imposing each dimension to recognize important samples with huge data gaps. Similar to the procedure of fishermen rectifying nets, a specific dimension was fixed at various quantiles. The numerical simulations and a real-world application from a cascade reaction process for high-density polyethylene were achieved to check the performance of the proposed method. The performance was superior to other methods.

3. The Proposed Model

This section shows detailed steps to explain the model we developed in this work, from data acquisition to the generation of the virtual samples required and from constructing the model to the classification tool.

3.1. Data Acquisition. Our proposed model was examined on different datasets. The two datasets are classified into finger movements and the UCI machine learning respiratory. The first dataset was recorded by implementing two surface channels by using FlexComp device. The sensors were of type T9503M and were positioned on the patient forearm, as shown in Figure 1.

Nine participants were asked to perform ten finger movements. Each finger movement consumed five seconds, then there was a rest for another five seconds, and then the participant was requested to execute the next finger movement and so on till finishing the ten finger movement classes, which are shown in Figure 2. The mentioned data collection process were repeated for six times.

The second dataset was for amputee patients. The nine participants missed their left hand. The goal of collecting these data was to classify between six different gestures to understand and analyze the controlled upper limbs prostheses. The six gestures were flexion, index flexion, fine pinch, tripod grip, hook grip, and spherical grip. It was a very

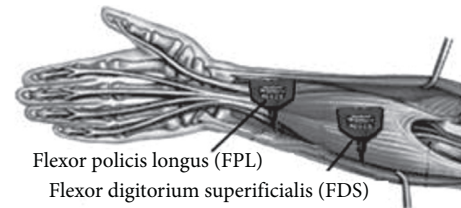


FIGURE 1: Posture of the electrodes.

challenging task to record the surface signal from amputee participants with three different force levels. The skin was cleaned with alcohol and prepared using the abrasive method. The allocated electrodes were Ag/AgCl electrodes.

The surface signal was recorded from 8 channels at three levels of forces for nine amputee participants. The first dataset was amplified by 1000; the first and second datasets were sampled at 2000 Hz. Figure 3 shows the allocation of the electrodes and the collection of the surface signal from amputee participants.

For the above-mentioned datasets, threefold cross-validation was applied, where 2/3 of dataset was assigned to the training set, whereas 1/3 was allocated to the testing set. The data were filtered to secure the precision and removal of noise. The training and testing accuracies were estimated on average basis where the accuracy was calculated per subject and the overall accuracy was the summation of each accuracy per subject divided by the number of subjects.

Other datasets were imported from UCI machine learning respiratory, which was considered as a strong archive that was cited more than 1000 times by machine learning community researchers. The performance of the proposed model was observed on additional five datasets that were archived at the UCI website. Those multiclass datasets were Iris, Breast Cancer, Seeds, Sonar, Mines vs. Rocks, and Indian Liver Patient. The Iris dataset is one of the most popular datasets that has been implemented in the pattern recognition field. The Iris dataset had three classes: one class was linearly independent of the other two classes and could be easily separated, whereas the other two classes were not separable. The three targeted classes for Iris dataset were Iris setosa, Iris virginica, and Iris versicolor. The Breast Cancer dataset was collected from the doctor clinic and classified into six classes, of which two were benign and the other four classes were dedicated for malignant type. The dataset was collected for three different sorts of wheat. The three different classes for the seeds dataset were Kama, Rosa, and Canadian, which were recorded via X-ray plates. The Sonar, Mines vs. Rocks dataset was to discriminate between metal and rock. The last recalled dataset was Indian Liver Patient. The dataset was collected from 441 male Indian participants and 142 female participants to discriminate if this participant could be classified as a liver patient or not.

3.2. Random Virtual Generation Equations. Let us assume that we have dataset $e = (x, f(x))$, where e represents the original training samples, $x \in R_n$, and $f(x) = \{-1, 1\}$. Assume that we have previous information k , and we need to

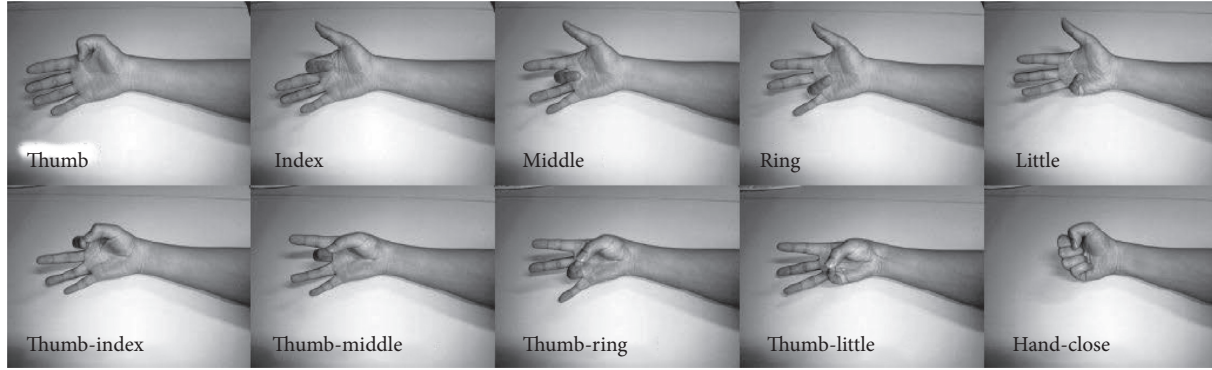


FIGURE 2: Ten different finger gestures (classes).



FIGURE 3: Electrodes allocation for amputee participants.

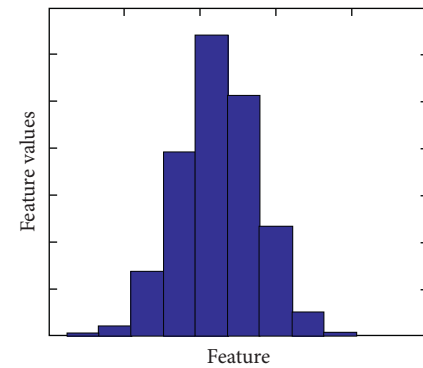


FIGURE 4: The normal distribution for one feature after DA [1, 2].

map our training set to the new domain. Therefore, if we have a convert that equals $T = y_T$, the dataset e will be transformed via this conversion equation to generate virtual samples $(Tx, y_T(f(x)))$. The generation of mathematical transformation T and y_T depends primarily on the previous information, which may result in either simple or complex transformation formula.

However, the second algorithm depends on adding noise to the original signal [36]. Most of the techniques that were used to create virtual samples suffer from the lack of combining reasonableness and adaptability simultaneously. Accordingly, we followed an algorithm to generate virtual Gaussian samples [2]. This method started to calculate the mean and standard deviation for Gaussian distribution, as shown in Figure 4. Then, the virtual samples could be generated following this technique, and finally, the virtual generated samples were added to the original ones [37]. So, $\{x_1, \dots, x_n, x_{n+1}, \dots, x_k\}$ represents the original dataset, which belongs to R . The first n samples of the dataset are continuous, whereas the $k - n$ samples are discrete. The m random variables are generated by Gaussian algorithm $N = (\mu, \eta^2)$ for the first n continuous samples knowing that μ represents the mean and η represents the standard error. However, for the samples $k - n$ that are assigned to be discrete ones, the values will not be transformed and in order to keep the consistency between the discrete and continuous part, we may generate random variables for the discrete part by using $N = (\mu, \eta^2)$ with η^2 equalling zero. Figure 1 shows

the normal distribution for augmented data for one feature only. This technique was utilized in our proposed model, where the main motivation was to secure a normal distribution for the stochastic electromyography signal.

The tensor can be defined as a multidimensional array with respect to a basis; however, for a vector, it can be represented as a single-dimensional array with respect to the same basis. In brief, tensors can be evaluated as a multidimensional vector. Tensors can be deemed as a mathematical method to represent values in a multidimension matrix. Tensors are considered the comprehensive version of matrix, vector, and scalar. Therefore, matrix, vector, and scalar can be measured as subcomponents of tensor. The generation of tensor can be done by following transformation laws. Tensors are characterized as having various coordinate systems. Therefore, the coordinate systems with their transformation laws will be analyzed in the next section.

Assume that we have x^i , where $i = 1, 2, \dots, N$. So, by substituting the different values of i , we can get N values of x in a N -dimensional space x^1, x^2, \dots, x^N . Moreover, the set of \bar{x}^i can be expressed as $\bar{x}^1, \bar{x}^2, \dots, \bar{x}^N$ for N -dimensional coordinates. In the same context, keeping the same transformation laws for x' to \bar{x}' leads to the following transformation equation:

$$x^i = x^i(\bar{x}^1, \bar{x}^2, \dots, \bar{x}^N), i = 1, 2, \dots, N. \quad (1)$$

The above equation creates an independent relation between the two different coordinates x^i and \bar{x}^i for $i = 1, 2, \dots, N$. As long as the relation is kept independent, it can be recalled for transformation.

The Jacobian first-order partial transformation will be estimated as follows:

$$J\left(\frac{x}{\bar{x}}\right) = J\left(\frac{x^1, x^2, \dots, x^N}{\bar{x}^1, \bar{x}^2, \dots, \bar{x}^N}\right) = \begin{vmatrix} \frac{\partial x^1}{\partial \bar{x}^1} & \frac{\partial x^1}{\partial \bar{x}^2} & \dots & \frac{\partial x^1}{\partial \bar{x}^N} \\ \vdots & \vdots & \dots & \vdots \\ \frac{\partial x^N}{\partial \bar{x}^1} & \frac{\partial x^N}{\partial \bar{x}^2} & \dots & \frac{\partial x^N}{\partial \bar{x}^N} \end{vmatrix}. \quad (2)$$

With an inverse transformation,

$$\bar{x}^i = \bar{x}^i(x^1, x^2, \dots, x^N), \quad i = 1, 2, \dots, N. \quad (3)$$

In brief, both equations (1) and (2) can be expressed in the notation formula as follows:

$$x^i = x^i(\bar{x}), \quad i = 1, 2, \dots, N, \quad (4)$$

$$\bar{x}^i = \bar{x}^i(x), \quad i = 1, 2, \dots, N. \quad (5)$$

\bar{x} can be concluded from x and $\bar{\bar{x}}$ can be deduced from \bar{x} by recalling transformation. Assume that $\bar{x} = y$ and $\bar{\bar{x}} = z$. The transformations are represented by T_1 , T_2 , and T_3 , where

$$\begin{aligned} T_1: y^i &= y^i(x^1, x^2, \dots, x^N), \quad i = 1, 2, \dots, N \text{ or } T_1 x = y, \\ T_2: z^i &= z^i(y^1, y^2, \dots, y^N), \quad i = 1, 2, \dots, N \text{ or } T_2 y = z. \end{aligned} \quad (6)$$

T_3 can be deduced by the product of both T_1 and T_2 :

$$\begin{aligned} T_3: z^i &= z^i(y^1(x^1, x^2, \dots, x^N), \dots, y^N(x^1, x^2, \dots, x^N)), \quad i = 1, 2, \dots, N \\ \text{or } T_3 x &= T_2 T_1 x = z \text{ by considering } T_3 = T_2 T_1, \end{aligned} \quad (7)$$

where T_1 , T_2 , and T_3 represent the first, second, and third coordinates transformations, respectively.

4. Experiment

The study implemented two layers of autoencoder: the first layer was 1200 nodes, whereas the second was 900 nodes. The encoder transfer function was purely linear. The suggested model is shown in Figure 5.

We claimed that the suggested paradigm was able to achieve high accuracy values for both the training and the testing sets with a powerful signal representation. In a preparatory step of the model, the input raw biosignal was performed by algorithm. The implemented window size was 200 milliseconds, while window increment was 50 milliseconds. The recommended number of sampling points to calculate the discrete Fourier transform was 1024. The advantage of proceeding lies in providing an appreciated representation for the input raw biosignal, which, in turn, boosted the accuracy values for both training and testing set. The output of representation was fed to the DA stage, where the above-mentioned Gaussian augmentation was used with reiterating represented data 1000 times. Reiterating data 1000 times was followed based on different trials, where 1000 showed the most compromise between simulation time and performance. The DA enriched the data and granted affluence to the data that improved training and testing accuracies in return. As a final stage in representing data, the tensor representation was hired to give us the opportunity to demonstrate the data into a

developed perspective. Then, the data was presented to two layers of autoencoder to learn features from high-quality represented data. The first layer of autoencoder was 1200 nodes, whereas the second one was 900 nodes.

The weight regularization coefficient was set to 0.001, as its default value, for both layers of autoencoder. The coefficient that controlled the weights of the sparsity regularization was set to 4 for both layers. The sparsity proportion factor determined the activation response rate of the autoencoder neuron. The value of sparsity proportion varied from 0 to 1. A lower value promoted and inspired for a higher sparsity. The sparsity proportion was set to 0.05. Eventually, the encoder transfer function was set to purely linear. We executed different transfer functions like logistic and positive saturating linear transfer functions aside from the pure linear one, which led to promoted results. The output features were employed to proceed with the classifier phase. The paper used three different main classifiers, namely, ELM, SVM, and SL. In terms of ELM, five activation functions were used and picked out the activation function that generated the most precise results. The five executed activation functions, namely, Sine, Triangular basis, and Radial basis functions. As for the SVM classifier, the study proceeded with six different functions for SVM. The executed SVM functions were linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian SVM, and the function that performed the best outcome was selected. The accuracies of the three classifiers were presented to the classifier fusion layer to select the best local classifier per class.

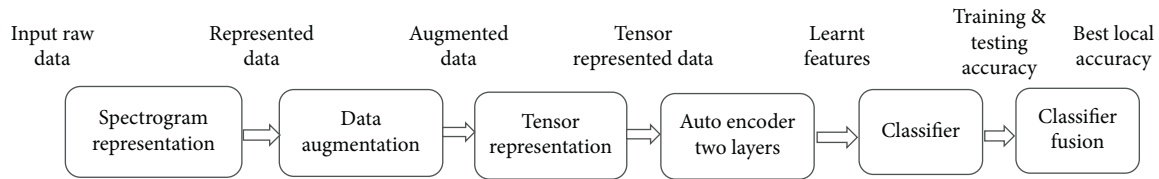


FIGURE 5: Proposed model diagram.

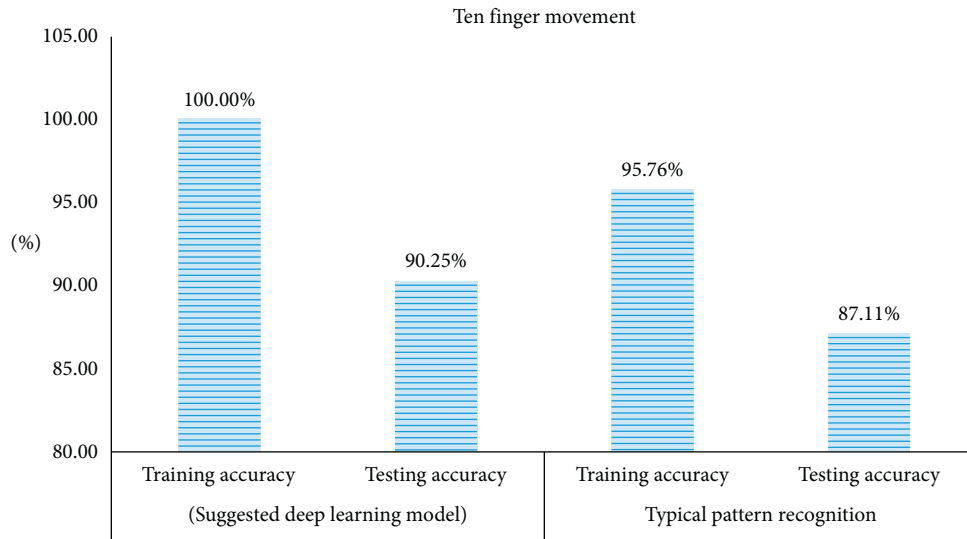


FIGURE 6: Comparison between suggested deep learning model in ten finger movements.

5. Results

The implementation of the classifier layer endorsed the accuracies for both training and testing set as the training set. The classification accuracy of ten finger movements dataset accounted for 100% for the training pack and 90.25% for testing one. As for the high-force six finger movements dataset, the training collection accuracy amounted to 99.74%, whereas accuracy for the testing group achieved 91.85%. Then, the executed data representation techniques mentioned above and the deep neural network were replaced by a typical pattern recognition model where the features were extracted and reduced to a lower number of features by using linear discriminant analysis and that used both the ELM and the SVM as classifiers. However, the performance of the ELM as a classifier was much better than that achieved by the SVM in terms of both the simulation time and accuracy. Based on the used pattern recognition model, the training accuracy for ten finger movements was 95.76%, whereas testing accuracy was 87.11%, as illustrated in Figure 6. In terms of the six finger movements, both the training and the testing accuracies were lower than those values achieved by our proposed model. The training accuracy was 98.57%, whereas the testing one was 89.64%, as illustrated in Figure 7.

This study concluded that our proposed model was explicitly better than the typical pattern recognition model.

Furthermore, the suggested system did not require any feature as it was trained to learn features by itself and independent of the input data. Accordingly, we examined the planned scheme on popular datasets to provide the model with reliability and trustworthiness. The implementation of Iris data resulted in a training accuracy of 100 % and testing accuracy of 98.5%. For Breast Cancer tissue dataset, the training accuracy was 98.58% and testing accuracy was 91.7%. However, using Sonar dataset, the accuracy for training was 85.69% and that for testing was 74.4%. Moreover, executing liver dataset led to 96.47% as training accuracy and 85.1% for testing one. With regard to the data, the training accuracy accounted for 94.57%, whereas for testing one, it amounted to 83.6%. The UCI machine learning respiratory datasets were executed without recalling any classifier fusion layer and were classified by using classifier only. The training simulation time was more than 600 seconds. However, the time consumed for examining the testing set on the trained network was not more than 1.5 seconds. Table 1 shows both training and testing accuracies for all of the above-mentioned datasets. Figure 6 shows a comparison between the testing and the training accuracies for the suggested model and those resulted from implementing a typical pattern recognition technique for classifying the ten finger movements. However, Figure 7 shows the same comparison for the six finger movements. The recommended model did not only show better

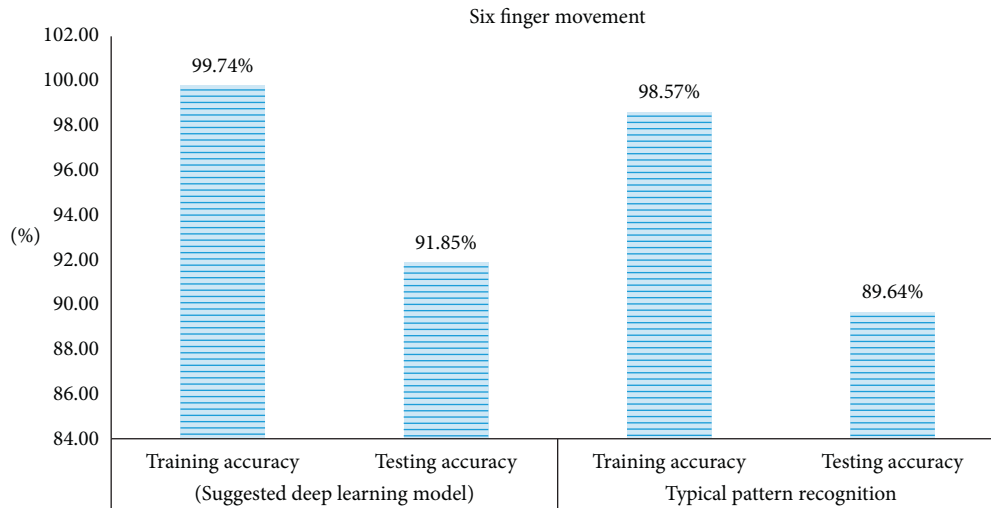


FIGURE 7: Comparison between suggested deep learning model in six finger movements.

TABLE 1: Training and testing accuracies for all implemented datasets.

Dataset	Training accuracy (%)	Testing accuracy (%)
Ten finger movements (suggested deep learning model)	100	90.25
Six finger movements (suggested deep learning model)	99.74	91.85
Ten finger movements (typical pattern recognition)	95.76	87.11
Six finger movements (typical pattern recognition)	98.57	89.64
Iris (suggested deep learning model)	100	98.5
Breast tissue (suggested deep learning model)	98.58	91.7
Sonar (suggested deep learning model)	85.69	74.4
Seeds (suggested deep learning model)	94.57	83.6
Liver (suggested deep learning model)	96.47	85.1

performance on the level of training and testing accuracies but also saved the effort and time that might be wasted in selecting the best features that match the application.

6. Conclusion and Future Work

We suggested a deep learning model where the data were represented, augmented, and then transferred into the tensor domain. Two layers of autoencoder were implemented by adjusting its parameters to have the best results. The SVM, ELM, and SL were applied as classifiers. Also, the best local classifier was applied to select the highest accuracy per class. The proposed model was applied to different datasets to provide it with fidelity and reliability. Ten and six finger movements were used for the advised system and for traditional pattern recognition. The planned diaphragm resulted in higher accuracies than the traditional pattern recognition system with the advantage of the classifier fusion technique. Moreover, the pattern recognition consumed effort and time to extract the best features set that led to better accuracies, whereas the suggested model did not require any features or human interventions as it was capable of learning features by itself regardless of the introduced dataset. The recommended model consumed about 600 seconds to train the network with no more than 1.5 seconds to test the trained network. The planned model

was applied to other popular datasets and brought about accepted accuracy values. The main advantage behind examining data by the model was that we voided the feature extraction engineering handcrafted techniques and fed the model by the data that were capable of learning features by itself and independently of the data type that was introduced, which saved time and effort. Eventually, as a future enhancement, the simulation time may be reduced by implementing different neural networks that may be able to learn features in a superior manner without consuming a long simulation time.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] M. Wedyan, *Augmented reality and novel virtual sample generation algorithm based autism diagnosis system*, Ph.D. thesis, FEIT, UTS, Ultimo, Australia, 2020.

- [2] M. Wedyan, A. Crippa, and A. Al-Jumaily, "A novel virtual sample generation method to overcome the small sample size problem in computer aided medical diagnosing," *Algorithms*, vol. 12, no. 8, pp. 160–185, 2019.
- [3] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.
- [4] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?" in *Proceedings of the 2016 international conference on digital image computing: techniques and applications (DICTA)*, November 2016.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in neural information processing systems*, pp. 1097–1105, Lake Tahoe, NV, USA, December 2012.
- [6] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural Computation*, vol. 22, no. 12, pp. 3207–3220, 2010.
- [7] 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, pp. 321–357, 2002.
- [8] H. S. Baird, *Document Image Defect Models, in Structured Document Image Analysis*, pp. 546–556, Springer, Berlin, Germany, 1992.
- [9] L. S. Yaeger, R. F. Lyon, and B. J. Webb, "Effective training of a neural network character classifier for word recognition," in *Proceedings of the Advances in neural information processing systems*, pp. 807–816, Denver, CO, USA, May 1997.
- [10] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," *Icdar*, vol. 3, p. 2003, 2003.
- [11] T. Poggio and T. Vetter, *Recognition and Structure from One 2D Model View: Observations on Prototypes, Object Classes and Symmetries*, MIT Computer Science & Artificial Intelligence Laboratory, Cambridge, MA, USA, 1992.
- [12] R.-w. Xu, L. He, L.-k. Zhang, Z.-y. Tang, and S. Tu, "Research on virtual sample based identification of noise sources in ribbed cylindrical double-shells," *Journal of Vibration and Shock*, vol. 27, no. 5, pp. 32–35, 2008.
- [13] J. S. Liu and Y. N. Wu, "Parameter expansion for data augmentation," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1264–1274, 1999.
- [14] M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 528–540, 1987.
- [15] I. Cortes-Ciriano and A. Bender, "Improved chemical structure-activity modeling through data augmentation," *Journal of Chemical Information and Modeling*, vol. 55, no. 12, pp. 2682–2692, 2015.
- [16] S. Hauberg, O. Freifeld, A. B. L. Larsen, J. Fisher, and L. Hansen, "Dreaming more data: class-dependent distributions over diffeomorphisms for learned data augmentation," in *Proceedings of the Artificial Intelligence and Statistics*, pp. 342–350, Cadiz, Spain, May 2016.
- [17] J. Ahmad, K. Muhammad, and S. W. Baik, "Data augmentation-assisted deep learning of hand-drawn partially colored sketches for visual search," *PloS One*, vol. 12, no. 8, Article ID e0183838, 2017.
- [18] J. Fintzi, X. Cui, J. Wakefield, and V. N. Minin, "Efficient data augmentation for fitting stochastic epidemic models to prevalence data," *Journal of Computational and Graphical Statistics*, vol. 26, no. 4, pp. 918–929, 2017.
- [19] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.
- [20] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, IEEE, New Orleans, LA, USA, March 2017.
- [21] Z. Gan, R. Henao, D. Carlson, and L. Carin, "Learning deep sigmoid belief networks with data augmentation," in *Proceedings of the Artificial Intelligence and Statistics*, pp. 268–276, San Diego, CA, USA, May 2015.
- [22] G. Rogez and C. Schmid, "Mocap-guided data augmentation for 3D pose estimation in the wild," in *Proceedings of the Advances in neural information processing systems*, pp. 3108–3116, Barcelona, Spain, December 2016.
- [23] D. L. Piza, A. Schulze-Bonhage, T. Stieglitz, J. Jacobs, and M. Dümpelmann, "Depuration, augmentation and balancing of training data for supervised learning based detectors of EEG patterns," in *Proceedings of the 2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 497–500, IEEE, Shanghai, China, May 2017.
- [24] A. Boyajian, "The tensor - a new engineering tool," *Electrical Engineering*, vol. 55, no. 8, pp. 856–862, 1936.
- [25] J. Betten, "Applications of tensor functions in continuum damage mechanics," *International Journal of Damage Mechanics*, vol. 1, no. 1, pp. 47–59, 1992.
- [26] A. Bernardi, J. Brachat, P. Comon, and B. Mourrain, "General tensor decomposition, moment matrices and applications," *Journal of Symbolic Computation*, vol. 52, pp. 51–71, 2013.
- [27] D. Yu, L. Deng, and F. Seide, "The deep tensor neural network with applications to large vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 388–396, 2012.
- [28] M. Sørensen and P. Comon, "Tensor decompositions with banded matrix factors," *Linear Algebra and Its Applications*, vol. 438, no. 2, pp. 919–941, 2013.
- [29] J. He and T.-Z. Huang, "Inequalities for M-tensors," *Journal of Inequalities and Applications*, vol. 2014, no. 1, p. 114, 2014.
- [30] K. Biswas and P. Mandal, "Modeling of source parameters and moment tensors of local earthquakes occurring in the eastern Indian shield," *Journal of the Geological Society of India*, vol. 89, no. 6, pp. 619–630, 2017.
- [31] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *Journal of Machine Learning Research*, vol. 15, pp. 2773–2832, 2014.
- [32] A. Cichocki, D. Mandic, L. De Lathauwer et al., "Tensor decompositions for signal processing applications: from two-way to multiway component analysis," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 145–163, 2015.
- [33] V. Hore, A. Viñuela, A. Buil et al., "Tensor decomposition for multiple-tissue gene expression experiments," *Nature Genetics*, vol. 48, no. 9, pp. 1094–1100, 2016.
- [34] Q.-X. Zhu, X.-H. Zhang, and Y.-L. He, "Novel virtual sample generation based on locally linear embedding for optimizing the small sample problem: case of soft sensor applications," *Industrial & Engineering Chemistry Research*, vol. 59, no. 40, pp. 17977–17986, 2020.

- [35] Q.-X. Zhu, Z.-S. Chen, X.-H. Zhang, A. Rajabifard, Y. Xu, and Y.-Q. Chen, "Dealing with small sample size problems in process industry using virtual sample generation: a Kriging-based approach," *Soft Computing*, vol. 24, no. 9, pp. 6889–6902, 2020.
- [36] L. Zhang and G.-H. Chen, "Method for constructing training data set in intrusion detection system," *Jisuanji Gongcheng Yu Yingyong (Computer Engineering and Applications)*, vol. 42, no. 28, pp. 145-146, 2006.
- [37] J. Yang, X. Yu, Z.-Q. Xie, and J.-P. Zhang, "A novel virtual sample generation method based on Gaussian distribution," *Knowledge-Based Systems*, vol. 24, no. 6, pp. 740–748, 2011.

Research Article

Protein-Protein Interaction Analysis through Network Topology (Oral Cancer)

Fazal Wahab Khattak,¹ Yousef Salamah Alhwaiti,² Amjad Ali,¹ Mohammad Faisal ,³ and Muhammad Hameed Siddiqi³

¹Department of Computer and Software Technology, University of Swat, Mingora, KPK, Pakistan

²Department of Computer Science, Jouf University, Sakaka, Aljouf, Saudi Arabia

³Department of CS & IT, University of Malakand, Chakdara, KPK, Pakistan

Correspondence should be addressed to Mohammad Faisal; mfaisal_1981@yahoo.com

Received 4 November 2020; Revised 21 December 2020; Accepted 23 December 2020; Published 16 January 2021

Academic Editor: Shaukat Ali

Copyright © 2021 Fazal Wahab Khattak 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.

Oral cancer is a complex disorder. Its creation and spreading are due to the interaction of several proteins and genes in different biological thoroughfares. To study biological pathways, many high-yield methods have been used. Efforts to merge several data found at separate levels related to biological thoroughfares and interlinkage networks remain elusive. In our research work, we have proposed a technique known as protein-protein interaction network for analysis and exploring the genes involved in oral cancer disorders. The previous studies have not fully analyzed the proteins or genes involved in oral cancer. Our proposed technique is fully interactive and analyzes the data of oral cancer disorder more accurately and efficiently. The methods used here enabled us to observe the wide network consists of one mighty network comprising of 208 nodes 1572 edges which connect these nodes and various detached small networks. In our study, TP53 is a gene that occupied an important position in the network. TP53 has a 113-degree value and 0.03881821 BC value, indicating that TP53 is centrally localized in the network and is a significant bottleneck protein in the oral cancer protein-protein interaction network. These findings suggested that the pathogenesis of oral cancer variation was organized by means of an integrated PPI network, which is centered on TP53. Furthermore, our identification shows that TP53 is the key role-playing protein in the oral cancer network, and its significance in the cellular networks in the body is determined as well. As TP53 (tumor protein 53) is a vital player in the cell division process, the cells may not grow or divide disorderly; it fulfills the function of at least one of the gene groups in oral cancer. However, the latter progression in the area is any measure; the intention of developing these networks is to transfigure sketch of core disease development, prognosis, and treatment.

1. Introduction

Verbal cancer, mouth cancer, and head and neck cancer are all related to each other [1], which are cancers affecting cell development or found in the oral cavity [2]. They may occur as an essential injury beginning in any of the tissues within the oral cavity; its reason may be metastasis process occurring away from oral cavity or is due to change in neighbor physical structures, such as the nasal cavity. Or the oral cancers may originate in any of the mouth tissue and can spread to any cell or tissue: teratoma, adenocarcinoma

inferred from a saliva gland, and tonsillar or other lymphoid tissues cause lymphoma; pigment-producing cells of the oral mucosa are producer of melanoma. Oral cancers are of few kinds, but almost 90% originate within the tissues that join the mouth and lips called squamous cell carcinomas [3].

Oral cancer is the most common cancer in the world with low overall survival rates [4], where oral squamous cell carcinoma (OSCC) is the most common type [5]. The organs that suffer from oral cancer include the buccal mucosa, tongue, and lower lip, and it mostly occurs in the people whose age is more than fifty years [6]. Traditional manual

methods for creating network books, diagrams, or maps have been in use for a long time. The diagrams and other visual data in these valuables were handmade decades ago and have only restricted use. These graphs were based on information of that specific time when created, consisting of unchangeable and interaction less data.

Various biological procedures are enunciated as systems. Networks are demonstrated from the field of atomic sciences such as metabolic networks, protein interaction networks, and gene regulatory networks. Network analysis, sculpting, and visualization are imperative stride about a framework's natural perception of living beings. The visual interpretation of such networks gives the basic idea about the structure of network and also can create an idea about newly generated complex biological data [7].

Genetically, oral cancer has still not been fully analyzed and explored even after several years of research. Many casual or susceptible genes associated with oral cancer have been reported in the research studies. Thus, the focus of our research work is to obtain results that give us useful data related to genes involved in oral cancer and enable us to draw conclusions about gene-gene and related protein-protein interactions.

2. Literature Review

The role of proteins networks in syndromes identification and analyses is very significant. Throughout a period, study in miniature creatures and protein networks plays a vital role in molecular development study enhancement to increase vision in the strength of cells to disruption and also for allotment of new protein occupations. Succeeding those investigations and the present increase of protein interactivity capabilities in mammals, protein networks are progressively helping as toolkit to unknot the molecular foundation of disease. The authors of this paper review favorable uses of protein networks in disease in four main parts:

- (a) Latest disease genes identification
- (b) Study of their network valuables
- (c) Spotting disease-linked subnetworks
- (d) Network-based disease sorting

Proteins networks can also be used in areas like infectious disease, self-medication, and pharmacology [8].

Due to interaction detection methods, millions of interactions between proteins have been discovered. Focus of the paper is based on the graph theory, and the authors derived a new method from graph theory known as spectral method. Spectral method was used to disclose unseen networked structures of complex protein-protein collaboration networks. They come up with the idea that these concealed topological layouts comprise organically germane serviceable clusters. This idea stirs a novel approach to forecast the function of unknown proteins based on the classification of known proteins within network structures. With the use of this procedure, 48 quasi-cliques and six quasi-bipartite were separated from a network consisting of

11855 interactions among 2617 proteins in budding yeast, and 76 uncharacterized proteins were allocated functions [9].

All biological processes were considered important and specifically achieved through protein-protein interactions. They use spectral analysis method as a technique to disclose high-level structures using colossal and multifaceted associations. the structurae of the image radius that the authours obtained by the speculative analysis is used in the nalysis for the topology making is used in the analysis for the topology making for the disclosure of unseen topological structures of a complicated interface network [9].

Proteome functional organization can understand better by using human PPI for the identification of interrelating sets of human proteins scientifically; a protein matrix of 4456 baits and 5632 victims was separated by predetermined yeast two-hybrid (Y2H) interaction coupling. They acknowledged 3186 frequently original relations among 1705 proteins, which produce a huge, highly connected network, following Y2H technique. The Y2H scheme is an influential PPIs identification tool, applicable to high-throughput method in detection of complete proteome of organism interactions. To describe high-level confidence relation among proteins, different techniques were used, for example, topological, GO criteria and scoring system. The network was also used to locate genes without any character assigned to humans' disease proteins players in regulating the cellar trials. Screening human proteins relations systematically can enhance the understanding of protein functionality and cellular procedures [10]. They present their experiments results in the form of different graphs and charts.

There are a huge number of collaboration networks; studying them, finding functional nodes and links, and investigating the inner function of cells are the aims of this research. The authors performed a precise chart theory-based examination of this PPI arranged to build computational models for relating and anticipating the properties of deadly changes and proteins contributing in hereditary contacts, utilitarian bunches, protein complexes, and signaling pathways. Their examination suggests that deadly modifications are not as it were exceedingly related inside the organize structure, but they moreover fulfill an extra property. Their evacuation causes a diversion in organized structure. Creators moreover give proof for the nearness of diverse ways that maintain a strategic distance from practical proteins in PPI systems, whereas such ways do not exist for lethal change [11].

The PPI network includes a minor quantity of exceedingly associated protein hubs (also called centers) and various ineffectively associated hubs. Genome-wide consideration shows that cancellation of a center protein is more likely to be hazardous than cancellation of a nonhub protein; a wonder known as the centrality-lethality runs the show. This run of the show is commonly accumulated to reflect the unprecedented centrality of centers in organizing the courses of action of qualities, which in turn suggests the normal importance of organize plans, a key conviction of frameworks of science. Concurring to creators, for the notoriety of this clarification, the fundamental cause of the

centrality-lethality running the show has never been basically inspected. They propose the concept of basic PPIs to discover out PPIs that are vital for the presence or propagation of a life form. As anticipated, basic PPIs are developmentally more moderated than insignificant PPIs. Considering the part of basic PPIs in deciding quality essentiality, they discover the yeast PPI organize practically more energetic than arbitrary systems, however with less distant sound than the potential ideal. These and other discoveries give unused viewpoints on the organic centrality of arranged structure and vigor [12].

Genomic consideration shows that erasing an exceedingly associated protein hub (center) is more likely to be deadly to a living being than erasing a humble associated hub (nonhub), a strategy recognized as the centrality-lethality running the show. As center points are more imperative than nonhubs in organizing the worldwide arranged structure, the centrality-lethality running the show is broadly accepted to reflect the worth of organized engineering in characterizing organized work, a key idea of frameworks science. Consequently, the centrality-lethality running the show is clarified without the association of organized engineering. Utilizing yeast information, the creators gave down to earth confirmation supporting their speculation [12].

In paper [13], ARIYA et al. suggested a microRNA consisting of multiple genes; actually microRNAs are analyzed, which seemed to be fruitful in high level of tumor. They veiled for the genetic factor; the appearance of these genetic factor remained associated with oral cancer development and evolution. Therefore, recognizing medications that might aim at these genetic factors might benefit plummeting the oral cancer humanity by cultivating development of patients.

In paper [14] by amiri et al., the scheme method has been planned to reconnoiter hereditary difficulty of oral cancer besides recognizing innovative oral cancer associated genes to perceive genomic modifications at molecular smooth, as concluded from variance investigation. Their genetic and communicating examination indicated important enhancement metabolism, motioning trail, and microRNA trails on the road to development. Their integrative method would benefit discovering the genetic variants of oral cancer that can speed up drug detection consequences to mature a healthier sympathetic concerning action approaches for numerous cancer categories.

Protein-protein interaction (PPI) network analysis of cancer has gained focus of medical and biological scientists. Through this approach, examination of the interaction between genes that cause cancer could lead us to improve the diagnosis and treatment of patients [15, 16]. In PPI network analysis, the genes related to the disease are gathered and organized in an integrative structure [15, 17]. The calculation of topological properties of the network, including central parameters such as degree and centrality, provides useful information about molecular mechanism of disease onset and pathology [18]. Introducing selected genes among large number of query genes can lead to specific biomarker panel related to the disease [19].

3. Methodology

The methodology of protein-protein interaction of oral cancer consists of seven steps. The first step is the attainment of the candidate genes of oral cancer, which was done in two ways: Firstly, PolySearch text-mining system is used, which is a web-based tool exploiting various techniques to highlight and align informative text. PolySearch results gave us a large list of genes and proteins. A lot of data are available on cancer. As our specific target is to collect and analyze data on oral cancer, PolySearch tool is used for obtaining the specific data on oral cancer. Secondly, we manually confirmed from reviewing the literature that the obtained genes and proteins were indeed of oral cancer. The second step of the methodology involved the use of STRING database for scanning the protein interactions. Construction of PPI networks, from which an extended network was derived, was used to extract the giant component making up the third step. The fourth step involved the analysis of the PPI network based on its topology. The fifth step was the creation of a backbone network from the giant network based on the highest betweenness centrality (BC) value. The sixth step involved the construction of a subnetwork from the giant network that consisted of all shortest routes among the aspirant genes. The seventh step concluded with TSPO as central protein.

3.1. Dataset Preparation for the Extraction of Genes Associated with Oral Cancer. Candidate genes associated with oral cancer were searched using PolySearch text-mining system, which is a useful tool in producing a list of ideas based on the query of the user. PubMed, OMIM, Drug Bank, and Swiss-Prot are among the few sources of information from which data relevant to the query can be extracted. PolySearch contributes a great deal to extracting information on important biomedical concepts such as genes/proteins, disease, SNPs, drugs, metabolites, pathways, and tissues, which could help in finding the relevant genes of oral cancer [20]. This method is used for finding genes related to oral cancer. Search was performed using the key words “Disease with Gene/Protein Association” and “Oral Cancer.” The literature analysis by PolySearch system returned 274 results. To determine the accuracy, these results were confirmed manually to check whether oral cancer associated genes were viable candidates or not. Discarding the less significant genes resulted in a total of 208 candidate genes for oral cancer by literature survey (Table 1).

3.2. Scanning Protein-Protein Interactions. Oral cancer candidate genes listed in Table 1 were converted to seed proteins. For visualizing the PPIs, STRING database was used. It is an exploratory database for PPIs. Updated version of STRING, 9.1, contains more than 1100 organisms, covering millions of proteins [20].

3.3. Constructing PPI Network from the Extension for Extraction of the Giant Component. A system, which consisted of the seed proteins and has also direct association with their

TABLE 1: List of genes showing association with oral cancer, extracted from the literary database.

SN	Symbol	SN	Symbol	SN	Symbol	SN	Symbol	SN	Symbol	SN	Symbol	SN	Symbol	SN	Symbol	SN	Symbol	SN	Symbol		
1	ABCA1	20	CASP7	39	CP	58	DSC2	77	GPX1	96	IL6	115	MT1A	134	PHLDA1	153	REST	172	SMC2	191	TNFSF10
2	ABCE1	21	CCL20	40	CSNK1E	59	DSG1	78	GRIK2	97	ITGB2	116	MT3	135	PIK3CA	154	RPL30	173	SMURF1	192	TP53
3	ADAM17	22	CCL5	41	CXCL2	60	DSG2	79	GRM5	98	ITGB4	117	MTA1	136	PIM1	155	RPL37A	174	SOD2	193	TP73
4	AKT1	23	CCND1	42	CXCL3	61	DUSP1	80	GSK3A	99	KDR	118	MTHFR	137	POFUT1	156	RPS17	175	SOX2	194	TSP0
5	ALCAM	24	CCND2	43	CXCR1	62	E2F1	81	GSTM1	100	KLF4	119	MTOR	138	POLG2	157	S100A1	176	SRF	195	TWIST1
6	ANG	25	CCNH	44	CXCR2	63	EDNRB	82	GSTO1	101	KLF8	120	MVD	139	POLRMT	158	S100A2	177	STI4	196	TXNRD2
7	ANKRD6	26	CCR4	45	CXCR7	64	EGFR	83	GSTP1	102	KRT19	121	MYCBP	140	PRDM14	159	S100A4	178	STAG2	197	UBB
8	ANXA1	27	CD14	46	CYP11A1	65	EGR1	84	HBA1	103	LAMC2	122	NAT2	141	PREX2	160	S100A7	179	STAT3	198	VEGFA
9	APC2	28	CD207	47	CYP26B1	66	ELAVL3	85	HMG2	104	LBX1	123	NES	142	PRKAB1	161	S100A9	180	TERT	199	WIF1
10	ATP2A2	29	CD44	48	CYP2E1	67	ENG	86	HMG2	105	LCN2	124	NNMT	143	PROM1	162	S100B	181	TFR3	200	WNT2B
11	ATP2A3	30	CD68	49	DCN	68	ERCC2	87	HNRNP35	106	LDHA	125	NRP1	144	PROX1	163	S100P	182	TGFBI	201	WWP1
12	AURKA	31	CDH1	50	DCTN6	69	FADD	88	HOXA10	107	LIN28B	126	OAZ1	145	PTGS2	164	SEMA3C	183	TGFBI	202	XRCC1
13	AXIN2	32	CDK4	51	DDX3X	70	FAM48A	89	HOXA9	108	MAPK14	127	OPRM1	146	PTMA	165	SF3B3	184	TGFBR2	203	XRCC2
14	BECN1	33	CDKN1B	52	DIABLO	71	FGF2	90	HPSE	109	MET	128	ORAOVA1	147	PTPN11	166	SIRT1	185	TIMP3	204	XRCC3
15	BIRC5	34	CDKN1C	53	DICER1	72	FHIT	91	HSPA5	110	MGMT	129	PABPN1	148	PTPRJ	167	SIRT3	186	TIMP4	205	XRCC4
16	BNIP2	35	CDKN2A	54	DNAH8	73	GAD1	92	HSPG2	111	MMP1	130	PARP1	149	PTPRR	168	SLC2A1	187	TLR2	206	YES1
17	BRCA2	36	CHAF1A	55	DNAJA3	74	GDNF	93	ICAM1	112	MMP7	131	PDK1	150	PTPRZ1	169	SLPI	188	TLR3	207	ZBTB7A
18	CA9	37	CHAF1B	56	DNMT3B	75	GHRL	94	ICAM3	113	MOK	132	PDPN	151	RAF1	170	SMAD2	189	TLR4	208	VEGF
19	CALML3	38	COL17A1	57	DPAGT1	76	GJA1	95	IL1F10	114	MSI1	133	PGK1	152	RASGRP3	171	SMAD4	190	TNFRSF11B		

PPI's neighbors, was developed. This system was built using Cytoscape [21], which is an extremely flexible project designed for the purpose of investigations, operations, and visualization of extensive systems. Not only did this study involve expansion on the system incorporating a giant part but also it included two small separate parts. The expectation of this study was that the giant network must consist of hubs with extensive BC value. This was clearly in light of the fact that both separate parts consist of a small number of hubs, and for this purpose the investigation and preparation involved only the giant network with its various parameters. The use of giant system obtained from the extended system was examined and analyzed by this method advantageously.

3.4. Protein Interaction Network Involving Topological Analysis. For accessing the hubs in the system, nodes properties such as betweenness centrality (BC), closeness centrality (CC), and degree (K) were noted and used, especially K and BC values. Quantity of adjacent joints which is based on the determination of the number of connections of one protein to its neighbors determines the most important factor, that is, the degree (K). The number of shortest routes that pass through each node, thus measuring the frequent occurrence of the node which has the greatest restricted routes between other nodes, determines the betweenness centrality. Shortest path or the most limited way is found by measuring the length of every last one of the geodesics to or from the network. The flow in the network is characterized by a high BC node value to great extent. BC may assume an integral part as a global property, since it is a valuable pointer to distinguish bottlenecks in a system. The converse of the normal length of the briefest paths among all the other nodes in the graph, which let us know the topological focus of the network, constitutes a closeness centrality (CC). Estimates of networks based on global topology include average degree, mean most brief path length, and distances utilized in the network.

Average degree ($\langle K \rangle$): mean of total degree values of nodes in a network

Mean shortest path length (mspl): represents connecting average of the steps involved to every pair of nodes using their shortest path

Diameter (D): longest paths among all shortest paths, characteristics of nodes, and measurements used to characterize the network were calculated by Cytoscape software in this paper

3.5. Creating Backbone Network by Searching High BC. This study involved observing the PPI supporting the oral cancer-related gene network.

Consequently, the proteins with high BC ought to have vigorously utilized crossing points as the backbone actually consisted of these proteins and their connections within the network. The critical point with high BC value was set at top 15 genes of the total number of nodes within the network [9, 10]. BC nodes with high values and their interlinkage

were mined utilizing the giant network for creation of a backbone network. Furthermore, BC also served the purpose of measuring the nodes centrality in the network initially since according to the definition, BC values that are high consist of the shortest paths in the network passing through the nodes. These nodes with high BC value are among other nodes of the network function as bottleneck control interaction.

3.6. Construction of a Subnetwork among the Candidate Genes with the Shortest Paths. Some of the candidate genes in the network are not directly connected and that includes the giant network as well. For constructing a subnetwork, genes connected with least number of nodes, in link with oral cancer, are required. For this, each combination of candidate qualities and their most limited way was calculated. These ways were found using Cytoscape software, coming about within the subnetwork comprising all nodes in these ways.

3.7. Central Protein TP53 Validates the Backbone Network. Validation was essential for checking the healthiness of the backbone network. For this, test networks were developed as they were employing a little component of 208 genes as initial seeds. For this, once more Cytoscape computer program was utilized. BC value was determined in these test networks. Segments having nodes with top 15 BC in the test networks agreed with the nodes in the backbone network as per description in the fifth step to determine the accuracy of the backbone. Furthermore, the healthiness of backbone network of oral cancer proteins was determined by calculation of the frequency of the nodes having the largest BC value which gives the accuracy of the backbone nodes.

4. Results

In this research work, we have obtained protein-protein interaction network through applying various steps discussed in the Methodology section. Then, we have obtained the key nodes involved in protein-protein interaction network and perform the analysis for obtaining important protein or genes involved in oral cancer.

4.1. Network with Protein-Protein Interactions. A huge network with several collapsed smaller networks, respectively, found from the seed proteins TSPO and TP53, was also a part of the extended network (Figure 1). Giant network contained of 208 nodes connected with 1572 edges as illustrated in (Figure 2). The backbone network, however, comprised 322 edges linking 29 nodes (Figure 3). Another network consists of 25 nodes linking 87 edges (Figures 4 and 5). The further distinguishing feature is the presence of some very highly connected nodes, though they are very small in number, while others show fewer connections.

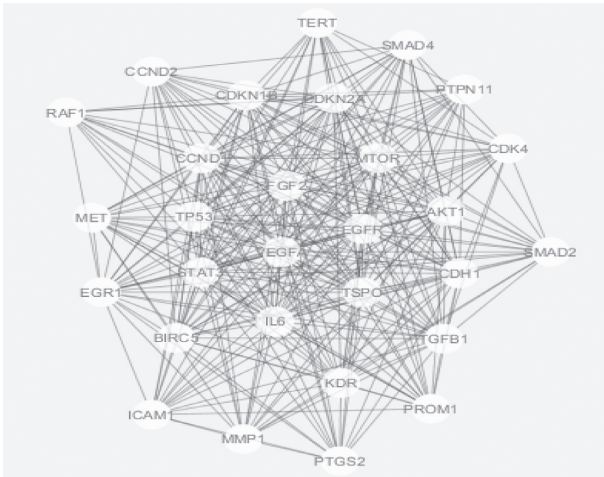


FIGURE 3: Illustration of the backbone obtained from other networks. It consists of 29 nodes with highest betweenness centrality, and nodes' size indicates their BC values.

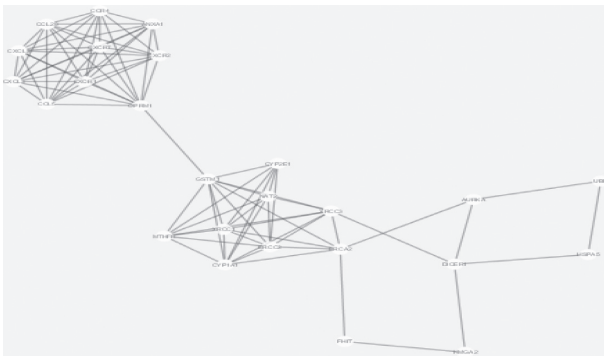


FIGURE 4: Illustration of the backbone obtained from other networks. It consists of 25 nodes and 87 edges.



FIGURE 5: Illustration of the backbone obtained from other networks. It consists of 4 nodes and 5 edges with highest betweenness centrality.

109 nodes had a large degree. 10 nodes had a large degree and high betweenness centrality (Table 2), nodes among the high BC, CC, and degree nodes (Table 3). To visualize the role of each protein in the oral cancer network, different colors and sizes were assigned to highlight them (Figure 2). TP53 is a protein with largest degree, TP53 is a protein with highest BC value, and EGFR is a protein with highest CC value. TP53 occupies central position in the network due to its high degree, BC, and CC value. Signaling pathways in the high betweenness centrality network and their cross-talks

TABLE 2: List of the proteins having high BC and large degree values.

SN	Gene	Degree	BC
1	TSPO	93	0.04809981
2	TP53	113	0.03881821
3	AKT1	90	0.01744922
4	EGFR	79	0.01611723
5	VEGFA	60	0.01249361
6	IL6	60	0.00829916
7	MTOR	49	0.00784171
8	CDH1	66	0.00782314
9	CXCR2	17	0.00712798
10	SMAD4	39	0.00692481

TABLE 3: Nodes with high BC, CC, and degree values.

Gene	Degree	Gene	CC	Gene	BC
TP53	113	EGFR	0.75409836	TSPO	0.04809981
TSPO	93	MAPK14	0.75	TP53	0.03881821
AKT1	90	CDKN1B	0.75	AKT1	0.01744922
EGFR	79	CDK4	0.73684211	EGFR	0.01611723
CDH1	66	TSPO	0.69767442	VEGFA	0.01249361
CCND1	61	STAT3	0.6875	IL6	0.00829916
VEGFA	60	IL6	0.67741935	MTOR	0.00784171
IL6	60	CD68	0.66666667	CDH1	0.00782314
STAT3	56	CHAF1A	0.66666667	CXCR2	0.00712798
FGF2	53	OAZ1	0.66666667	SMAD4	0.00692481
MTOR	49	EGR1	0.64705882	XRCC3	0.00531283
MAPK14	49	FGF2	0.63829787	MET	0.00516178
CDKN2A	47	BIRC5	0.63513514	UBB	0.00501709
TGFB1	47	MMP1	0.63218391	PARP1	0.00451757
KDR	42	RAF1	0.63157895	GJA1	0.00447296

are obtained from the backbone networks. 1st backbone network consists of 29 nodes and 322 edges (Figure 3). Second major backbone network consists of 25 nodes and 87 edges (Figures 4 and 5). Analyzing the values of BC and CC shows that TP53 locates at the center of the backbone network with the highest BC value and the largest degree. TP53 has 29 neighbors: CCND1, SMAD2, TSPO, MTOR, ICAM1, VEGFA, STAT3, EGR1, TERT, MMP1, CCND2, MET, AKT1, KDR, TP53, RAF1, IL6, EGFR, PTGS2, PTPN11, FGF2, CDKN2A, TGFB1, BIRC5, CDKN1B, PROM1, SMAD4, CDH1, and CDK4. Details of other proteins in the backbone network were not included here.

5. Discussion

Genetically oral cancer has not been fully explored; even after several years of research it still remains unexplored. Many casual or susceptible genes associated with oral cancer have been reported. The multiple factors contributing to hepatocarcinogenesis include ecological, transmissible, sustaining, metabolic, and endocrine structures, which are involved specifically or indirectly in the development of oral cancer. The oral cancer pathogenesis indicates the involvement of several genes and can broadly be classified into four majors groups: regulatory genes involved in response to DNA damage, cell cycle control players, those intricate

growth inhibition, apoptosis, and also the genesis involved in cell-cell collaboration and signal transduction. Analyzing the contribution of genes and proteins pertaining to oral cancer pathogenesis in addition to the involvement of other key proteins in the PPI networks by analyzing their topology was the goal of this study. The results obtained gave us useful data related to genes involved in oral cancer and enabled us to draw conclusions about gene-gene and related protein-protein interactions. It was observed that most of the seed proteins (208) connected to giant network in terms of oral cancer and their PPI neighbors. This giant oral cancer network consisted of 1572 edges. Backbone network topology, however, resulted in different small network topologies. As it is not possible to draw conclusion from the giant network, we further split down the backbone network into small subnetworks for obtaining clear results on oral cancer. The splitting of backbone network is performed according to properties of nodes, that is, BC, CC, and degree values. One backbone network consists of 29 nodes and 322 edges. Another subnetwork consists of 27 nodes 87 edges. There is one small subnetwork of 4 nodes and 5 edges. Other subnetworks also exist but they are very small. Literature was analyzed to find further techniques and methods supporting the importance of different genes in oral cancer. TP53 is a gene that occupied important position in the network. TP53 has degree value of 113 and BC value of 0.03881821. Literature was analyzed to find further techniques and methods supporting the importance of TP53 in cancer. In this regard, researchers have clarified TP53 changes in different types of cancers. Moreover, it has been discovered that TP53 reaction pathway is regularly flawed in human diseases and the recurrence of TP53 fluctuates greatly. In another study conducted by Ding and colleagues, polymorphisms were affiliated with this disorder. The study was focused around metainvestigation to better comprehend the relationship between polymorphism of Condon 72 and the tumor protein p53 (TP53) gene, which results in a missense transformation of arginine (R) to proline (p) and causes vulnerability to hepatocellular carcinoma [22, 23]. TSPO is another important gene in oral cancer network having highest BC value of 0.04809981 and degree value of 93. TP53 mutation sites with stalk cell-like gene appearance in addition to association with many cancers were also found in oral cancer with high mortality rate, although the exact sites are unknown as of yet. Moreover, direct sequencing methods are more helpful for reliable results. Form direct sequencing result, it was concluded that TP53 mutations, mainly the blistering mutations R249S and V157F, are linked with meager prognosis for patients with oral cancer. The stem cell-like acquirement of gene manifestation traits might be a contributing factor to the violent behavior of tumors with TP53 mutations having shorter survival rate as compared to wild-type TP53, according to literature survey [23–25].

6. Conclusion

The mainstay network shows clearly all the essentials genes of oral cancer, their associated regulatory pathways,

and the interactions between them. This eventually brought us to the conclusion that TP53 is the protein with the highest degree, TSPO is the protein with largest BC, and EGFR is the protein with highest CC values, but TP53 obtained the important position in the network due to its degree, BC, and CC values, indicating that TP53 is centrally localized in the network and is a significant bottleneck protein in oral cancer protein-protein interaction network. These findings suggested that pathogenesis of oral cancer variation was organized by means of an integrated PPI network, which is centered on TP53. Furthermore, our identification of TP53 as the bottleneck protein in oral cancer network determined its significance in cellular network in the body as well, as P53.

Tumor protein 53 is important in the regulation of cell division; by restricting the cells from growing or dividing uncontrollably, it fulfills the function of at least one of the gene groups in oral cancer.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] R. Lozano, M. Naghavi, K. Foreman et al., “Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010,” *The Lancet*, vol. 380, no. 9859, pp. 2095–2128, 2013.
- [2] J. W. Werning, Ed., *Oral Cancer: Diagnosis, Management, and Rehabilitation*, Thieme, New York, NY, USA, 2011.
- [3] A. C. Chi, T. A. Day, and B. W. Neville, “Oral cavity and oropharyngeal squamous cell carcinoma—an update,” *CA A Cancer Journal for Clinicians*, vol. 65, no. 5, 2015.
- [4] B. Wang, S. Zhang, K. Yue, and X.-D. Wang, “The recurrence and survival of oral squamous cell carcinoma: a report of 275 cases,” *Chinese Journal of Cancer*, vol. 32, no. 11, pp. 614–618, 2013.
- [5] C. Rivera, “Essentials of oral cancer,” *International Journal of Clinical and Experimental Pathology*, vol. 8, no. 9, pp. 11884–11894, 2015.
- [6] S. D’souza and V. Addepalli, “Preventive measures in oral cancer: an overview,” *Biomedicine and Pharmacotherapy*, vol. 107, no. 3, pp. 72–80, 2018.
- [7] A. Kerren and F. Schreiber, “Network visualization for integrative bioinformatics,” in *Approaches in Integrative Bioinformatics*, pp. 173–202, Springer, Berlin, Germany, 2014.
- [8] T. Ideker and R. Sharan, “Protein networks in disease,” *Genome Research*, vol. 18, no. 4, pp. 644–652, 2008.
- [9] D. Bu, Y. Zhao, L. Cai et al., “Topological structure analysis of the protein-protein interaction network in budding yeast,” *Nucleic Acids Research*, vol. 31, no. 9, pp. 2443–2450, 2003.
- [10] U. Stelzl, U. Worm, M. Lalowski et al., “A human protein-protein interaction network: a resource for annotating the proteome,” *Cell*, vol. 122, no. 6, pp. 957–968, 2005.

- [11] N. Pržulj, D. A. Wigle, and I. Jurisica, "Functional topology in a network of protein interactions," *Bioinformatics*, vol. 20, no. 3, pp. 340–348, 2004.
- [12] X. He and J. Zhang, "Why do hubs tend to be essential in protein networks?" *PLoS Genetics*, vol. 2, no. 6, p. e88, 2006.
- [13] S. Ariya, A. James, and B. Joseph, "Computational analysis of oral cancer gene expression profile and identification of MiRNAs and their regulatory hub genes," *Journal of Complementary Medicine Research*, vol. 11, no. 3, pp. 154–159, 2020.
- [14] N. Amiri-Dashatan, M. Koushki, A. Jalilian, N. A. Ahmadi, and M. Rezaei Tavarani, "Integrated bioinformatics analysis of mRNAs and miRNAs identified potential biomarkers of oral squamous cell carcinoma," *Asian Pacific Journal of Cancer Prevention*, vol. 21, no. 6, pp. 1841–1848, 2020.
- [15] H. Zali and M. Rezaei Tavarani, "Meningioma protein-protein interaction network," *Archives of Iranian Medicine*, vol. 17, no. 4, pp. 262–272, 2014.
- [16] N. Safari-Alighiarloo, M. Rezaei-Tavarani, M. Taghizadeh, S. M. Tabatabaei, and S. Namaki, "Network-based analysis of differentially expressed genes in cerebrospinal fluid (CSF) and blood reveals new candidate genes for multiple sclerosis," *PeerJ*, vol. 4, Article ID e2775, 2016.
- [17] H.-A. Abbaszadeh, A. A. Peyvandi, Y. Sadeghi et al., "Er: YAG laser and cyclosporin a effect on cell cycle regulation of human gingival fibroblast cells," *Journal of Lasers in Medical Sciences*, vol. 8, no. 3, pp. 143–149, 2017.
- [18] M. Rezaei-Tavarani, M. Rezaei-Tavarani, V. Mansouri et al., "Introducing crucial protein panel of gastric adenocarcinoma disease," *Gastroenterology and Hepatology from Bed to Bench*, vol. 10, no. 1, pp. 21–28, 2017.
- [19] M. Rezaei Tavarani, F. OkHOVATIAN, M. Zamanian Azodi, and M. Rezaei Tavarani, "Duchenne muscular dystrophy (DMD) protein-protein interaction mapping," *Iranian Journal of Child Neurology*, vol. 11, no. 4, pp. 7–14, 2017.
- [20] J. Li, M. Nazir Jan, and M. Faisal, "Big data, scientific programming, and its role in internet of industrial things: a decision support system," *Scientific Programming*, vol. 2020, Article ID 8850096, 2020.
- [21] X. Liao, M. Faisal, Q. QingChang, and A. Ali, "Evaluating the role of big data in IIOT-industrial internet of things for executing ranks using the analytic network process approach," *Scientific Programming*, vol. 2020, Article ID 8859454, 2020.
- [22] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart, "PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites," *Nucleic Acids Research*, vol. 36, pp. W399–W405, 2008.
- [23] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [24] C. Ding, H. Yu, H. Yu, and H. Qin, "TP53 codon 72 polymorphism with hepatocellular carcinoma: a meta-analysis," *Journal of International Medical Research*, vol. 40, no. 2, pp. 446–454, 2012.
- [25] H. G. Woo, X. W. Wang, A. Budhu et al., "Association of TP53 mutations with stem cell-like gene expression and survival of patients with hepatocellular carcinoma," *Gastroenterology*, vol. 140, no. 3, pp. 1063–1070, 2011.

Research Article

An Online-Offline Certificateless Signature Scheme for Internet of Health Things

Muhammad Asghar Khan ¹, Sajjad Ur Rehman ², M. Irfan Uddin ³, Shibli Nisar,⁴
Fazal Noor,⁵ Ali Alzahrani,⁵ and Insaf Ullah ¹

¹Hamdard Institute of Engineering & Technology, Islamabad 44000, Pakistan

²Department of Electrical Engineering, Namal Institute, Mianwali, Pakistan

³Institute of Computing, Kohat University of Science and Technology, Kohat 26000, Pakistan

⁴Department of Electrical Engineering, National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan

⁵Department of Computer Science and Information Systems, Islamic University of Madinah, Madinah 400411, Saudi Arabia

Correspondence should be addressed to Insaf Ullah; insafktk@gmail.com

Received 17 November 2020; Revised 11 December 2020; Accepted 21 December 2020; Published 31 December 2020

Academic Editor: Shah Nazir

Copyright © 2020 Muhammad Asghar Khan 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 Internet of Health Things (IoHT) is an extended breed of the Internet of Things (IoT), which plays an important role in the remote sharing of data from various physical processes such as patient monitoring, treatment progress, observation, and consultation. The key benefit of the IoHT platform is the ease of time-independent interaction from geographically distant locations by offering preventive or proactive healthcare services at a lower cost. The communication, integration, computation, and interoperability in IoHT are provided by various low-power biomedical sensors equipped with limited computational capabilities. Therefore, conventional cryptographic solutions are not feasible for the majority of IoHT applications. In addition, executing computing-intensive tasks will lead to a slow response time that can deteriorate the performance of IoHT. We strive to resolve such a deficiency, and thus a new scheme has been proposed in this article, called an online-offline signature scheme in certificateless settings. The scheme divides the signing part into two phases, i.e., online and offline. In the absence of a message, the offline phase performs computationally intensive tasks, while lighter computations are executed in the online phase when there is a message. Security analyses and comparisons with the respective existing schemes are carried out to show the feasibility of the proposed scheme. The results obtained authenticate that the proposed scheme offers enhanced security with lower computational and communication costs.

1. Introduction

IoHT is an IoT submarket, capable of grouping all medical devices and applications for gathering, analyzing, and exchanging physiological data of patients over the Internet [1]. Patient data can be collected through biomedical sensors and processed via user terminal devices such as computers, smart phones, smart watches, or even a specific embedded device [2]. Patient data may include breathing rate, blood pressure, chest sound, body temperature, respiratory rate, electrocardiogram (ECG), patient position (accelerometer), etc. [3–7]. In addition to medical applications, IoHT can also be used to

monitor environmental conditions such as patient-care venues, room status, laboratory shift times, treatment times, and staff-to-patient ratios. The user terminal devices are linked to a gateway via short-range wireless technologies such as Bluetooth Low Energy (BLE), Wi-Fi, and Zigbee. The BLE, however, uses strong features such as moderate data rate, low-power consumption, and unlicensed band, making them the most preferable options for connecting wearable sensor nodes. The gateway may be further connected to a (clinical) server or cloud services via fifth-generation (5G) wireless link for high storage and intensive data processing. In a health information system, patient details can be maintained as

electronic health records, which are available to the medical professionals when the patient visits the hospital.

Since a large scale of interactions between biomedical sensors and mobile devices is undertaken on an open wireless channel in IoHT environment, which poses a range of challenges, the most significant of which is the security and privacy of health-related information of patients [8]. To steal or fabricate patient health-related information, an intruder may capture the communication between the sensors and mobile devices. Likewise, with high probability, the attacker may gain access to the disease or health status of the patient. In addition, most devices involved in the IoHT platform have limited computing capabilities and, consequently, fail to perform conventional cryptographic calculations. For example, heavy computations are needed for most of the public key cryptosystems proposed in the literature; therefore, their implementation has not been considered acceptable for IoHT devices. An online-offline approach can be used to address heavy computation issues. When the IoHT devices have reported a message, the online phase is used to perform light computations only, while the offline computations or heavy computations are performed if no message has been recorded by the IoHT devices. Authentication is a major concern for securing IoHT devices. In general, the digital signature is used for authentication in cryptography. Therefore, the digital signature can be used with the online-offline approach for securing IoHT devices. The offline-computed signature value is generated in the offline phase, while the online phase operates with the same offline signature value.

The two basic methods used to validate the public keys are Identity-Based Cryptography (IBC) and Public Key Infrastructure (PKI) in public key cryptosystems. This includes a Certificate Authority (CA) signature, which provides a unique signature link [9]. The CA specifies the public keys with the certificates as defining a participant. However, shortcomings such as distribution, storage, and manufacturing difficulties are associated with PKI systems. Instead, IBC is suggested to decrease the cost of public-key management [10]. The trusted Private Key Generator (PKG) has first-hand data about the participants' private keys with the expense of private key escrow issues [11, 12]. Therefore, certificateless cryptosystem can be used with the signature scheme to accommodate the key escrow problem.

Some computationally hard problems, such as bilinear pairing, Rivest-Shamir-Adleman (RSA), and elliptic curve cryptosystems, usually measure the efficiency of signature schemes. The RSA cryptosystem [13, 14] uses a large key of 1024 bits [15]. Likewise, due to the massive pairing and map-to-point function computation, bilinear pairing is 14.31 times lower than RSA [16]. Similarly, in order to remove the shortcomings of RSA and bilinear pairing, the elliptic curve was introduced [17]. The security hardness and efficiency of elliptic curve cryptography are based on 160-bit keys compared to bilinear pairing and RSA [18]. Despite this, for resource-hungry devices, the 160-bit key is also undesirable and not affordable. Therefore, a new form, the generalization of the elliptic curve, called the hyperelliptic curve was thus suggested [19]. The hyperelliptic curve offers the same

degree of protection as the elliptic curve, bilinear pairing, and RSA using 80-bit keys, identity, and certificate size [20, 21]. For energy-constrained IoHT devices, the hyperelliptic curve would be a better option. Therefore, the data generated by the anticipated massive number of biomedical sensors and IoT devices would need to be collected, processed, and analyzed efficiently in real-time to ensure safe and timely management of patient health [22].

Considering the above objectives, a new scheme, called the online-offline certificateless signature scheme, has been introduced for IoHT. The scheme uses the concept of the hyperelliptic curve and is characterized by the small key size. In comparison, it is uncompromisingly identical to the solutions introduced by the elliptical curve method with half key size.

The research study conducted has the following excellent characteristics:

- (i) A lightweight security scheme, namely, online-offline certificateless signature, has been proposed for an IoHT platform.
- (ii) The proposed scheme divides the certificateless signature scheme into two phases, i.e., online and offline. Lighter computations are performed when there is a message in the online phase, while the offline phase performs computing-intensive tasks in the absence of a message.
- (iii) The scheme uses the hyperelliptic curve cryptography that tackles the limitations faced by IoHT devices such as limited energy and computing capabilities.
- (iv) The proposed scheme has shown to be immune to numerous attacks through formal security analysis.
- (v) Our approach offers better efficiency in terms of computational cost and communication overhead when compared to the existing equivalent schemes.

1.1. Structure of the Paper. The rest of the article is structured as follows. In Section 2, the relevant work is discussed. Section 3 includes preliminaries. The proposed online-offline certificateless signature system is introduced in Section 4. Security analysis can be found in Section 5. The cost analysis is provided in Section 6 with current solutions. Concluding remarks are available in Section 7.

2. Related Work

In scientific literature, the security and privacy concerns using the online-offline approach have not received ample consideration. Thus, the problems need to be thoroughly investigated. A well-designed security framework would greatly minimize the risk of the data being hacked, regardless of the devilish strategy involved. Some research studies are devoted to addressing IoHT platform data security problems.

The offline-online signature technique was first suggested by Even et al. [23], which is suitable for limited-storage devices. When the message to be signed is known, the execution of

their procedure enables the use of the offline mechanism to do moderate computations. After the message is understood to be authenticated, the second phase is carried out electronically. The protection of their method is dependent on the intractability of the large integer factoring mechanism. Their device is protected by chosen messages from attacks. However, their approach is not so successful in practice.

In 2001, to create an effective online-offline signature scheme, Shamir and Tauman [24] used chameleon hash functions based on an ordinary digital signature. In the proposed scheme, the key scale and signature sizes are reduced according to the original scheme. A new type of hash function, called the trapdoor hash function, has been introduced in their model to increase the system security. If the signer repeatedly uses the same hash value to get two signatures on two distinct messages, the recipient can gain a hash collision and use it to retrieve trapdoor information from the signer, which is the secret key of the signer. However, the proposed scheme uses many chameleon hash values for various messages. The main disclosure issue of chameleon hashing is known as this concern.

Yu and Tate [25] suggested an effective online-offline signature scheme that is known to be secure without a random oracle under the RSA assumption. They did not use the hash function at the trapdoor. Therefore, the second key pair did not need to be handled by their scheme and did not have to include in their signature the random commitment attribute. However, the proposed scheme is not affordable for resource-constrained IoHT devices due to the RSA cryptosystem, which is based on hard problems and incurs the high computational cost. Wu et al. [26], using bilinear pairing, suggested a successful online-offline signature scheme. The security of the model is connected to the theoretical Diffie–Hellman assumption in the random oracle model. Addobe et al. [27] also proposed an offline-online signature scheme called the MHCOOS for M-Health devices based on bilinear pairing. However, bilinear pairing involves high pairing and map-to-point function operations, which is not suitable for resource-constrained IoHT devices.

All of the above schemes are based on complex cryptographic techniques, i.e., elliptic curve and bilinear pairing, and thus suffer from high costs of computation and communication overhead. These schemes are thus not compatible with IoHT systems equipped with minimal computing capability. To create a viable IoHT cryptographic solution that needs less computation, there is a critical need to use the state-of-the-art online-offline certificateless signature technique. Our proposed scheme is based on hyperelliptic curve cryptography, which is an advanced version of the elliptic curve. It provides the same degree of protection with the smaller key size as compared to an elliptical curve, bilinear pairing, and modular exponential.

3. Preliminaries

3.1. Hyperelliptic Curve Discrete Logarithm Problem (HCDLP). Suppose a given instance of hyperelliptic curve $\delta = \varepsilon$. Then, the HCDLP is to determine ε from the given instance.

3.2. Threat Model. The security models of the proposed scheme include message c , unforgeability against the adversaries called Type 1 adversary (A_1), and Type 2 adversary (A_2), respectively. A_1 is a malicious adversary who has the ability to replace the user's public key besides the system master keys, while A_2 means an honest-but-curious KGC who knows the system master keys but is not allowed to replace the user's public key. The specific security models under different adversaries are as same as [28] such that unforgeability regarding EUF-CMA- A_1 and unforgeability regarding EUF-CMA- A_2 .

4. Proposed Online-Offline Certificateless Signature Scheme

4.1. Network Model. An initiative to incorporate the proposed scheme must be preceded by careful consideration of the following assumptions:

- (1) Patient data input can be obtained by sensors and analyzed by user terminal devices, such as laptops, tablets, smart watches, or even a particular embedded system
- (2) Each of the medical sensors and the user terminal are connected through BLE
- (3) The user terminal can be further linked with the cloud server using 5G, equipped with cloud computing services
- (4) The medical server presumes the role of administrators
- (5) The medical server is linked with the local computer in which electronic health records (HER) can be viewed by the medical personnel
- (6) The HER is stored securely in the database server for future consultations

IoHT can be implemented in various settings, depending on the requirements as shown in Figure 1. The required gadgets are usually included in the medical sensors according to the patient's illness. Using short-range radio transceivers (i.e., BLE), the sensors can be connected with the gateway router. On a frequency band of 2.4 GHz, the BLE works. There are valid reasons for selecting this level of technology. They function, for example, in the unlicensed spectrum and provide fair data rates and consume very low power [29]. The aggregated data from the patient monitoring sensors may be too big to be handled by the local server. It demands a high ability for storage and computing. Fortunately, with its architecture, the emerging fifth-generation (5G) mobile networking introduces multiaccess edge computing (MEC) facility. MEC performs high storage and intensive processing facilities when integrated into an IoHT setting.

4.2. Construction of the Proposed Scheme. This section covers the construction of the proposed scheme. Notations used in the proposed scheme are illustrated in Table 1. The proposed

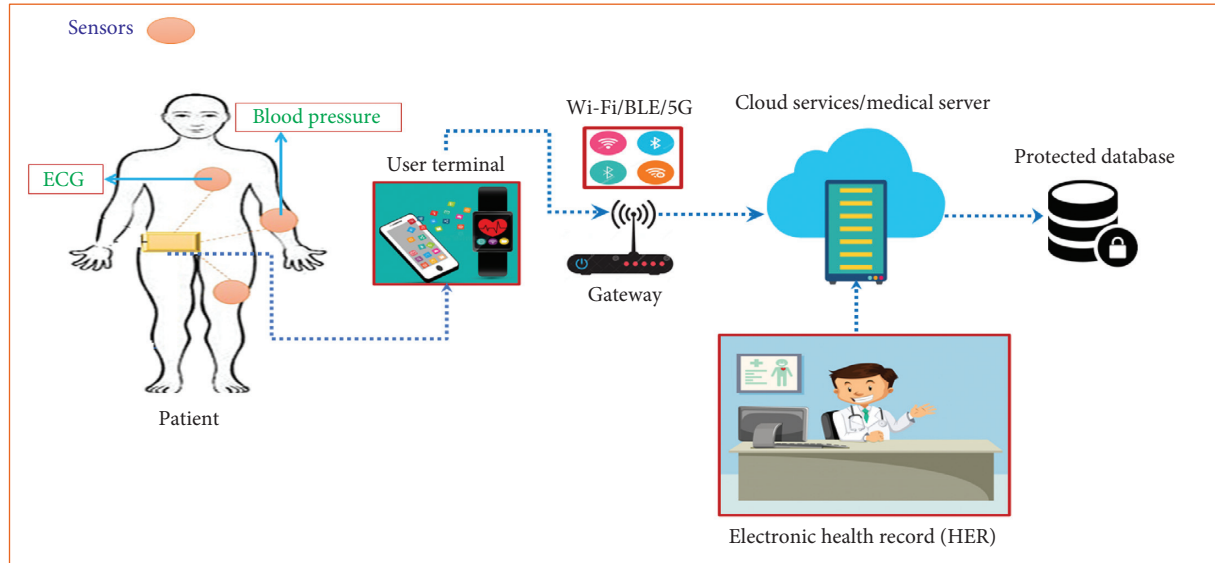


FIGURE 1: Sample network model of IoHT system.

TABLE 1: Notations used.

Notation	Description
η	It represents a security parameter
$\mathcal{H}c$	It represents a hyperelliptic curve
$f(n)$	It represents a finite field of n
n	It represents a large prime number belonging to hyperelliptic curve where the size of $n \geq 2^{80}$
\mathcal{D}	Divisor on the hyperelliptic curve ($\mathcal{H}c$)
\mathcal{Q}	Master private key of the system
\mathcal{K}	Master public key of the system
ψ	It represents a global parameter set that can be available publicly in a network
id_s, id_r	Identity of sender and receiver
Γ_s, Γ_r	They represent partial private key pair for sender and receiver
$\mathcal{N}_s, \mathcal{N}_r$	They represent private key pair for sender and receiver
$\mathcal{Z}_s, \mathcal{Z}_r$	They represent public key pair for sender and receiver
\mathcal{S}	Its represents signature
ϕ	It represents signature pair
h_x, h_y, h_z	Three irreversible and collision resistance hash functions
\perp	It represents null

scheme can be made from the following computational constructions [28]:

Setup: the following computations can be used for this phase:

- (i) The security parameter η can choose by KGC
- (ii) It selects a hyperelliptic curve ($\mathcal{H}c$) with field $f(n)$, where the size of $n \geq 2^{80}$
- (iii) Select a \mathcal{D} divisor from hyperelliptic curve ($\mathcal{H}c$)
- (iv) Then, choose three irreversible and collision resistance hash functions h_x, h_y, h_z
- (v) KGC picks $\mathcal{Q} \in \{1, 2, \dots, n-1\}$ as a master key and then computes the public key as $\mathcal{K} = \mathcal{Q} \cdot \mathcal{D}$
- (vi) KGC produces $\psi = \{\mathcal{K}, h?, h?, h?, \mathcal{D}, \mathcal{H}c, (?), ?\} \geq 2^{80}$ as global parameter set and publishes it publicly

Secret value setting: the participating entity with identity id_i picks $l_i \in \{1, 2, \dots, n-1\}$ as a secret value and computes $\mathcal{V}_i = l_i \cdot \mathcal{D}$ as a public key

Partial private key setting: for a participating entity with identity id_i , the KGC picks $\vartheta_i \in \{1, 2, \dots, n-1\}$, computes $\mu_i = \vartheta_i \cdot \mathcal{D}$, calculates $w_{\vartheta_i} = \vartheta_i + \mathcal{Q}h_x(id_i, \mathcal{V}_i, \mu_i)$, and sends $\Gamma_i = (w_{\vartheta_i}, ?)$ to entity with id_i via secure network

Private key setting: the participating entity, with identity id_i , sets $\mathcal{N}_i = (\Gamma_i, l_i)$ of its private key.

Public key setting: the participating entity, with identity id_i , sets $\mathcal{Z}_i = (\mathcal{V}_i, \mu_i)$ of its public key.

Certificateless online/offline signature: the sender computations can be divided into the following two substeps, e.g., Online and Offline.

Offline phase: this part will be run over the server that is equipped with high resources and the construction step is carried out as follows:

- (i) It picks $e \in \{1, 2, \dots, n-1\}$ and computes $t = d \cdot \mathcal{V}_s$
- (ii) Compute $\mathcal{P} = h(\text{???}, \text{??}, ?, t)$ and $\mathcal{X} = h(\text{???}, \mathcal{V}?, ?, t)$
- (iii) Then, it gives $(d, t, \mathcal{P}, \mathcal{X})$ to the sensor nodes

Online phase: this part will be run on the sensor nodes and the construction step consists as follows:

- (i) Compute $\mathcal{S} = l_s \cdot d - (l_s \cdot \mathcal{X} + \mathcal{P} \cdot w_s)$
- (ii) Set $\phi = (t, \mathcal{S})$ as a signature and send it to the receiver

Certificateless online/offline signature verification: upon reception ϕ , a receiver can verify \mathcal{S} as follows:

- (i) Compute $P = h_y(id_s, \mu_s, m, t)$ and $\chi = h_z(id_s, \mathcal{V}_s, m, t)$
- (ii) Then, it checks if $S \cdot D = t \cdot \chi \mathcal{V}_s - \mathcal{P}(\mu_s + h_x(id_s, \mathcal{V}_s, \mu_s) \mathcal{K})$ holds

4.3. Correctness. The verifier/receptionist can verify the signature if the following computation is successfully processed:

So, if $P = h_y(id_s, \mu_s, m, t)$ and $X = h_z(id_s, \mathcal{V}_s, m, t)$, we acquire

$$\begin{aligned}
 \mathcal{S} \cdot \mathcal{D} &= (l_s \cdot d - (l_s \cdot \mathcal{X} + \mathcal{P} \cdot w_s))D \\
 &= (l_s \cdot d \cdot \mathcal{D} - (l_s \cdot \mathcal{X} + \mathcal{P} \cdot w_s)D) \\
 &= (\mathcal{V}_s \cdot d - (l_s \cdot \mathcal{X} + \mathcal{P} \cdot w_s)D) \\
 &= (t - (l_s \cdot \mathcal{X})\mathcal{D} - (\mathcal{P} \cdot w_s)\mathcal{D}) \\
 &= (t - (l_s \cdot \mathcal{X} \cdot \mathcal{D}) - (\mathcal{P} \cdot w_s)\mathcal{D}) \\
 &= (t - (\mathcal{V}_s \cdot \mathcal{X}) - (\mathcal{P} \cdot (\vartheta_s + \mathcal{Q}h_x(id_s, \mathcal{V}_s, \mu_s))\mathcal{D})) \\
 &= (t - (\mathcal{V}_s \cdot \mathcal{X}) - (\mathcal{P} \cdot (\vartheta_s \cdot \mathcal{D} + \mathcal{Q} \cdot \mathcal{D}h_x(id_s, \mathcal{V}_s, \mu_s)))) \\
 &= (t - (\mathcal{V}_s \cdot \mathcal{X}) - (\mathcal{P} \cdot (\vartheta_s \cdot \mathcal{D} + \mathcal{Q} \cdot \mathcal{D}h_x(id_s, \mathcal{V}_s, \mu_s)))) \\
 &= (t - (\mathcal{V}_s \cdot \mathcal{X}) - (\mathcal{P} \cdot (\mu_s + h_x(id_s, \mathcal{V}_s, \mu_s) \mathcal{K})).
 \end{aligned} \tag{1}$$

This validates the correctness of the proposed scheme.

5. Security Analysis

The purpose of this section is to explain the usefulness of the suggested method in resisting attacks.

Theorem 1. *The proposed scheme resists against an adaptive chosen message attack, if an adversary A_1 would not be able to solve the hyperelliptic curve discrete logarithm problem (HECDLP).*

Proof. Suppose there is a challenger ζ which helps A_1 to extract ℓ from the given instance $f = \ell \cdot \mathcal{D}$ of HECDLP. Further, to figure out HECDLP, ζ can set the master key secret key as $\mathcal{Q} = \ell$ and master public key as $\mathcal{K} = \ell \cdot \mathcal{D}$.

Then, ζ generates ψ as a global parameter set and four empty lists ($L_{h_x}, L_{h_y}, L_{h_z}, L_k$) for holding the value of h_x, h_y, h_z , and keys.

Create (id_i): after reception, Create id_i query, ζ selects $\alpha_i, \beta_i, l_i \in \{1, 2, \dots, n-1\}$ and sets $h_x(id_i, \mathcal{V}_i, \mu_i) = -\beta_i$, $\mathcal{V}_i = l_i \cdot \mathcal{D}$, and $\mu_i = \beta_i \cdot \mathcal{K} - \alpha_i \cdot \mathcal{D}$. Then, ζ answers in the following two steps:

- (i) If $id_i \neq id_s$, with the identity id_i , ζ outputs will be $(\Gamma_i = \nu_i, \mu_i), \mathcal{N}_i = (\perp, l_i)$, and $\mathcal{Z}_i = (\mathcal{V}_i, \mu_i)$, respectively.
- (ii) If $id_i \neq id_s$, with the identity id_i , ζ outputs will be $(\Gamma_i = \nu_i, \mu_i), \mathcal{N}_i = (\Gamma_i, l_i)$, and $\mathcal{Z}_i = (\mathcal{V}_i, \mu_i)$, respectively.

Thus, ζ included $(id_i, \mathcal{V}_i, \mu_i, \beta_i)$ into L_{h_x} and $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ into L_k .

Hash queries (h_x, h_y, h_z): after reception, Hash queries (h_x, h_y, h_z), ζ searches for the values $\Omega_i, \mathcal{P}_i, \mathcal{X}_i$ in lists $L_{h_x}, L_{h_y}, L_{h_z}$; if it finds in these lists then returns to A_1 ; otherwise, the values $\Omega_i, \mathcal{P}_i, \mathcal{X}_i$ for each Hash query will select by ζ in a random manner and send it to the A_1 .

Secret value setting queries: after reception, this query, then, (ζ) answers in the following two steps:

- (i) If $id_i = id_s$, ζ aborts the process.
- (ii) If $id_i \neq id_s$, ζ will look for $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ in L_k ; if such a tuple is found, then it results in l_i ; otherwise, ζ calls Create id_i query and gets $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ and then sends l_i to A_1 .

Partial private key setting queries: after reception, this query, then, (ζ) answers in the following two steps:

- (i) If $id_i = id_s$, ζ aborts the process.
- (ii) If $id_i \neq id_s$, ζ will look for $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ in L_k ; if such a tuple is found, then it sends Γ_i to A_1 .

Public key setting queries: after reception, this query, then, (ζ) answers in the following two steps:

- (i) If $id_i = id_s$, ζ aborts the process.
- (ii) If $id_i \neq id_s$, ζ will look for $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ in L_k ; if such a tuple is found, then it results in $\mathcal{Z}_i = (\mathcal{V}_i, \mu_i)$; otherwise, ζ calls Create id_i query and gets $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ and then sends $\mathcal{Z}_i = (\mathcal{V}_i, \mu_i)$ to A_1 .

Public key replacement queries: after reception, this query, then, (ζ) will look for $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ in L_k and replace \mathcal{Z}_i by \mathcal{Z}_i^* and include $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i^*)$ into L_k . So, ζ sets $w_i = \perp$ and $\mathcal{N}_i = \perp$.

Certificateless online/offline signature queries: after reception, this query, then, (ζ) checks. If $id_i = id_s$, then it aborts the process; otherwise, it will perform the following steps:

- (i) ζ first gets access to L_{h_y}, L_{h_z} , and L_k .

Offline phase:

- (ii) It picks $d_i \in \{1, 2, \dots, n-1\}$ and computes $d_i = d_i \cdot V_s$.

Online phase:

- (iii) Compute $\mathcal{S}_i = l_i \cdot d_i - (l_i \cdot \mathcal{X}_i + \mathcal{P}_i \cdot w_i)$ and it results as a signature $\Phi = t_i, S_i$.

Certificateless online/offline signature verification query: after reception, this query, then, (ζ) checks. If $id_i = id_s$, then it aborts the process; otherwise, it will perform the certificateless online/offline signature verification algorithm for the verifications of signature.

Forgery: at the end, A_1 results a lawful signature ($\Phi = t_i, S_i$). If $id_i = id_s$, ζ aborts the process; otherwise, ζ checks for a list L_{h_x} , and according to forking lemma [], it generates another signature $\Phi^* = (\mathcal{S}_i^*, t_i)$. So, we have $\mathcal{S} \cdot \mathcal{D} = t_s - X \cdot \mathcal{V}_s - \mathcal{P}_s \cdot (\mu_s + \Omega_s \mathcal{K})$ and $\mathcal{S}^* \cdot \mathcal{D} = t_s - X \cdot \mathcal{V}_s - \mathcal{P}_s^* \cdot (\mu_s + \Omega_s \mathcal{K})$. We suppose that $\mu_s = \beta_s \cdot \mathcal{K} + \alpha_s \cdot \mathcal{D}$ and $\mathcal{K} = \ell \cdot \mathcal{D}$. So, when the subtractions between these two equations are performed, then we can get the following computations:

$$\begin{aligned}
 \mathcal{S}_i^* - \mathcal{S} \cdot \mathcal{D} &= (t_s - X \cdot \mathcal{V}_s - \mathcal{P}_s^* \cdot (\mu_s + \Omega_s \mathcal{K})) - (t_s - X \cdot \mathcal{V}_s - \mathcal{P}_s \cdot (\mu_s + \Omega_s \mathcal{K})), \\
 \mathcal{S}_i^* \cdot \mathcal{D} - \mathcal{S} \cdot \mathcal{D} &= t_s - X \mathcal{V}_s - \mathcal{P}_s^* \cdot (\mu_s + \Omega_s \mathcal{K}) - t_s - X \cdot \mathcal{V}_s + \mathcal{P}_s \cdot (\mu_s + \Omega_s \mathcal{K}), \\
 \mathcal{S}_i^* \cdot \mathcal{D} - \mathcal{S} \cdot \mathcal{D} &= \mathcal{P}_s \cdot (\mu_s + \Omega_s \mathcal{K}) - \mathcal{P}_s^* \cdot (\mu_s + \Omega_s \mathcal{K}), \\
 (\mathcal{S}_i^* - \mathcal{S}) \cdot \mathcal{D} - (\mathcal{P}_s - \mathcal{P}_s^*) \alpha_s \cdot \mathcal{D} &= (\mathcal{P}_s - \mathcal{P}_s^*) (\beta_s + \Omega_s) \ell \cdot \mathcal{D}, \\
 ((\mathcal{S}_i^* - \mathcal{S}) - (\mathcal{P}_s - \mathcal{P}_s^*) \alpha_s) \cdot \mathcal{D} &= (\mathcal{P}_s - \mathcal{P}_s^*) (\beta_s + \Omega_s) \ell \cdot \mathcal{D}, \\
 ((\mathcal{S}_i^* - \mathcal{S}) - (\mathcal{P}_s - \mathcal{P}_s^*) \alpha_s) &= (\mathcal{P}_s - \mathcal{P}_s^*) (\beta_s + \Omega_s) \ell, \\
 ((\mathcal{S}_i^* - \mathcal{S}) - (\mathcal{P}_s - \mathcal{P}_s^*) \alpha_s) / (\mathcal{P}_s - \mathcal{P}_s^*) (\beta_s + \Omega_s) &= \ell.
 \end{aligned} \tag{2}$$

So, A_1 can solve HECDLP as $\ell = ((\mathcal{S}_i^* - \mathcal{S}) - (\mathcal{P}_s - \mathcal{P}_s^*) \alpha_s) / (\mathcal{P}_s - \mathcal{P}_s^*) (\beta_s + \Omega_s)$, with the help of challenger ζ . \square

5.1. Probability Analysis. Here, we define the following probability events:

- (i) The winning probability of Create query must be greater than $(1 - Q_{h_x} Q_{\text{create}}/n)$
- (ii) The succeeded probability of h_y must be greater than $(1 - Q_{h_y}/n)$
- (iii) The succeeded probability of h_z must be greater than $(1 - Q_{h_z}/n)$
- (iv) The succeeded probability of certificateless online/offline signature queries must be greater than (Q_s/n)
- (v) $id_i = id_s$ satisfies with probability $(1/Q_{\text{create}})$

Note that Q_{create} , Q_{h_x} , Q_{h_y} , Q_{h_z} , and Q_s represent Create queries and Hash queries to h_x , h_y , h_z , and certificateless online/offline signature queries, respectively.

So, overall advantage of A_1 is towards its success as $\xi^* \geq (1 - Q_{h_x} Q_{\text{create}}/n)(1 - Q_{h_y}/n)(1 - Q_{h_z}/n) (1/Q_{\text{create}})(Q_s/n)$.

Theorem 2. *By using the random oracle model, the proposed scheme resists against an adaptive chosen message attack, if an adversary A_2 would not be able to solve the hyperelliptic curve discrete logarithm problem (HECDLP).*

Proof. Suppose there is a challenger ζ which helps A_1 to extract ℓ from the given instance $f = \ell \cdot \mathcal{D}$ of HECDLP. Further, to figure out HECDLP, ζ picks b and sets master

public key as $\mathcal{K} = b \cdot \mathcal{D}$. Then, ζ generates ψ as a global parameter set, and similar to Theorem 1, it picks four empty lists ($L_{h_x}, L_{h_y}, L_{h_z}, L_k$) for holding the value of h_x, h_y, h_z , and keys.

Create (id_i): after reception, Create id_i query, ζ answers in the following steps:

- (i) If $id_i = id_s$, ζ selects $\alpha_i, \Omega_i \in \{1, 2, \dots, n-1\}$ and sets $h_x(id_i, \mathcal{V}_i, \mu_i) = \Omega_i$, $\mathcal{V}_i = \ell \cdot \mathcal{D}$, $w_i = \alpha_i + b\Omega_i$, and $\mu_i = \alpha_i \cdot \mathcal{D}$. So, it produces $(\Gamma_i = w_i, u_i)$, $\mathcal{N}_i = (\Gamma_i, \perp)$, and $\mathcal{Z}_i = (\mathcal{V}_i, \mu_i)$, respectively.
- (ii) If $id_i \neq id_s$, ζ selects $\alpha_i, l_i, \Omega_i \in \{1, 2, \dots, n-1\}$ and sets $h_x(id_i, \mathcal{V}_i, \mu_i) = \Omega_i$, $\mathcal{V}_i = l_i \cdot \mathcal{D}$, $w_i = \alpha_i b \Omega_i$, and $\mu_i = \alpha_i \cdot \mathcal{D}$.

Thus, ζ included $(id_i, \mathcal{V}_i, \mu_i, \Omega_i)$ into L_{h_x} and $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ into L_k .

Hash queries (h_x, h_y, h_z): these are the same as performed in Theorem 1.

Secret value setting queries: after reception, this query, then, (ζ) answers in the following two steps.

- (i) If $id_i = id_s$, ζ aborts the process.
- (ii) If $id_i \neq id_s$, ζ will look for $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ in L_k ; if such a tuple is found, then it results in l_i ; otherwise, ζ calls Create id_i query and gets $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ and then sends l_i to A_2 .

Partial private key setting queries: after reception, this query, then, (ζ) answers in the following two steps:

- (i) If $id_i = id_s$, ζ aborts the process.
- (ii) If $id_i \neq id_s$, ζ will look for $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{Z}_i)$ in L_k ; if such a tuple is found, then it sends Γ_i to A_2 .

Public key setting queries: after reception, this query, then, (ζ) answers in the following two steps:

- (ii) If $\mathbf{id}_i = \mathbf{id}_s$, ζ aborts the process.
- (iii) If $?? \neq ??$, ζ will look for $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{X}_i)$ in L_k ; if such a tuple is found, then it results in $\mathcal{X}_i = (\mathcal{V}_i, \mu_i)$; otherwise, ζ calls Create id_i query and gets $(id_i, \Gamma_i, \mathcal{N}_i, \mathcal{X}_i)$ and then sends $\mathcal{X}_i = (\mathcal{V}_i, \mu_i)$ to A_2 .

Certificateless online/offline signature queries: after reception, this query, then, (ζ) checks. If $\mathbf{id}_i = \mathbf{id}_s$, then it aborts the process; otherwise, it will perform the following steps:

- (i) ζ first gets access to L_{h_y}, L_{h_z} , and L_k .
Offline phase:
 - (i) It picks $d_i \in \{1, 2, \dots, n-1\}$ and computes $t_i = d_i \cdot \mathcal{V}_s$.
 Online phase:
 - (ii) Compute $\mathcal{S}_i = l_i \cdot d_i - (l_i \cdot X_i + \mathcal{P}_i \cdot w_i)$ and it results as a signature $\mathcal{S} = (\mathcal{t}_\gamma, \mathcal{S}_\gamma)$.

Certificateless online/offline signature verification query: after reception, this query, then, (ζ) checks. If $id_i = id_s$, then it aborts the process; otherwise, it will perform the certificateless online/offline signature verification algorithm for the verifications of signature.

Forgery: at the end, A_1 results in a lawful signature $\phi = (\mathcal{t}_\gamma, \mathcal{S}_i)$. If $id_i = id_s$, ζ aborts the process; otherwise, ζ checks for a list L_{h_x} , and according to forking lemma [], it generates another signature $\Phi^* = (\mathcal{S}_i^*, t_i)$. So, we have $\mathcal{S} \cdot \mathcal{D} = t_s - X \cdot \mathcal{V}_s - \mathcal{P}_s \cdot (\mu_s + \Omega_s \mathcal{K})$ and $\mathcal{S}_i^* \cdot \mathcal{D} = t_s - X \cdot \mathcal{V}_s - \mathcal{P}_s^* \cdot (\mu_s + \Omega_s \mathcal{K})$. We suppose that $\mu_s = \beta_s \cdot \mathcal{K} + \alpha_s \cdot \mathcal{D}$ and $\mathcal{K} = \ell \cdot \mathcal{D}$. So, when the subtractions between these two equations are performed, then we can get the following computations:

$$\begin{aligned}
 \mathcal{S}_i^* \cdot \mathcal{D} - \mathcal{S} \cdot \mathcal{D} &= (t_s - X \cdot \mathcal{V}_s - \mathcal{P}_s^* \cdot (\mu_s + \Omega_s \mathcal{K})) - (t_s - X \cdot \mathcal{V}_s - \mathcal{P}_s \cdot (\mu_s + \Omega_s \mathcal{K})), \\
 \mathcal{S}_i^* \cdot \mathcal{D} - \mathcal{S} \cdot \mathcal{D} &= t_s - X \mathcal{V}_s - \mathcal{P}_s^* \cdot (\mu_s + \Omega_s \mathcal{K}) - t_s + X \cdot \mathcal{V}_s + \mathcal{P}_s \cdot (\mu_s + \Omega_s \mathcal{K}), \\
 \mathcal{S}_i^* \cdot \mathcal{D} - \mathcal{S} \cdot \mathcal{D} &= \mathcal{P}_s \cdot (\mu_s + \Omega_s \mathcal{K}) - \mathcal{P}_s^* \cdot (\mu_s + \Omega_s \mathcal{K}), \\
 (\mathcal{S}_i^* - \mathcal{S}) \cdot \mathcal{D} - (\mathcal{P}_s - \mathcal{P}_s^*) \alpha_s \cdot \mathcal{D} &= (\mathcal{P}_s - \mathcal{P}_s^*) (\beta_s + \Omega_s) \ell \cdot \mathcal{D}, \\
 ((\mathcal{S}_i^* - \mathcal{S}) - (\mathcal{P}_s - \mathcal{P}_s^*) \alpha_s) \cdot \mathcal{D} &= (\mathcal{P}_s - \mathcal{P}_s^*) (\beta_s + \Omega_s) \ell \cdot \mathcal{D}, \\
 ((\mathcal{S}_i^* - \mathcal{S}) - (\mathcal{P}_s - \mathcal{P}_s^*) \alpha_s) &= (\mathcal{P}_s - \mathcal{P}_s^*) (\beta_s + \Omega_s) \ell, \\
 ((\mathcal{S}_i^* - \mathcal{S}) - (\mathcal{P}_s - \mathcal{P}_s^*) \alpha_s) / (\mathcal{P}_s - \mathcal{P}_s^*) &= (\beta_s + \Omega_s) \ell.
 \end{aligned} \tag{3}$$

So, $\ell = (\mathcal{S}_i^* - \mathcal{S}) / (\mathcal{P}_s^* - \mathcal{P}_s)$ as the solution of HECDLP.

The probability analysis is same as Theorem 1 and as follows:

The utilized advantages of A_2 towards its success are as follows:

$$\xi^* \geq (1 - Q_{h_x} Q_{\text{create}}/n)(1 - Q_{h_y}/n)(1 - Q_{h_z}/n)(1/Q_{\text{create}})(Q_s/n). \quad \square$$

6. Cost Analysis

This section contrasts the efficiency of the proposed scheme with the existing equivalents suggested by the schemes of Yu and Tate [25], scheme 1, Yu and Tate [25], scheme 2, Wu et al. [26], and Addobeia et al. [27].

6.1. Computational Cost. Table 2 displays the key results derived from the analysis. Elliptic curve scalar multiplication and bilinear pairings are used in the existing schemes, all of which are more expensive alternatives. Therefore, we add the multiplication of the hyperelliptic divider. Observations have shown that the time it takes for a single scalar multiplication to be processed differs considerably: elliptic curve

point multiplication (ECPM), 0.97 milliseconds; bilinear pairing (P), 14.90 ms; pairing-based point multiplications (BPM), 4.31 ms; and modular exponentiation (E), 1.25 ms [16]. The Multiprecision Integer and Rational Arithmetic C Library (MIRACL) [30] is used to calculate the performance of the proposed system. It checks roughly 1000 times the runtime of specific cryptographic operations. A workstation with the following requirements is used for evaluating simulation results: Intel Core i7-4510U Processor @ 2.0 GHz, 8 GB RAM, and Windows 7 Home Standard 64-bit Operating System [29]. The hyperelliptic curve divisor multiplication (HM) is believed to be 0.48 milliseconds in length due to a smaller key size of 80 bits [31–34]. It is apparent from the results in Tables 2 and 3 that our solution is much more effective in terms of the computational cost as shown in Figure 2.

6.2. Communication Cost. This subsection is aimed at discussing the comparison results from the perspective of communication costs. The proposed approach is compared with the existing schemes presented by Yu and Tate [25] scheme 1, Yu and Tate [25] scheme 2, Wu et al. [26], and

TABLE 2: Computational cost.

Schemes	Signing	Verifying	Total
Yu and Tate [25] scheme 1	$1E + 3BPM$	$3E + 4BPM$	$4E + 7BPM$
Yu and Tate [25] scheme 2	$2E + 3BPM$	$3E + 3BPM$	$5E + 6BPM$
Wu et al. [26]	3BPM	$2P + 2BPM$	$2P + 5BPM$
Addobea et al. [27]	3 BPM	$3P + 4BPM$	$3P + 7BPM$
Proposed	4HM	3HM	7HM

TABLE 3: Computational cost in milliseconds.

Schemes	Signing	Verifying	Total (ms)
Yu and Tate [25] scheme 1	14.18	20.99	35.17
Yu and Tate [25] scheme 2	15.43	16.68	32.11
Wu et al. [26]	12.99	38.42	51.41
Addobea et al. [27]	12.99	61.94	74.93
Proposed	1.92	1.44	3.36

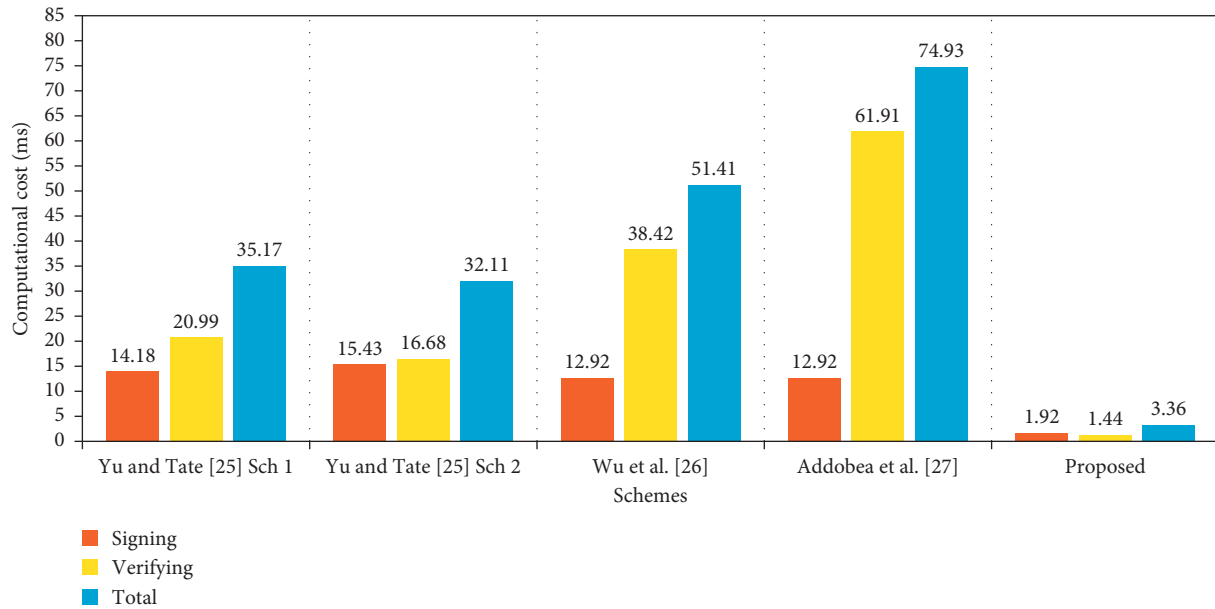


FIGURE 2: Computational cost (in ms).

TABLE 4: Communication cost in bits.

Schemes	Communication cost	Communication cost in bits
Yu and Tate [25] scheme 1	$3 G + m $	4096
Yu and Tate [25] scheme 2	$3 G + m $	4096
Wu et al. [26]	$3 G + m $	4096
Addobea et al. [27]	$3 G + m $	4096
Proposed	$2 n + m $	1184

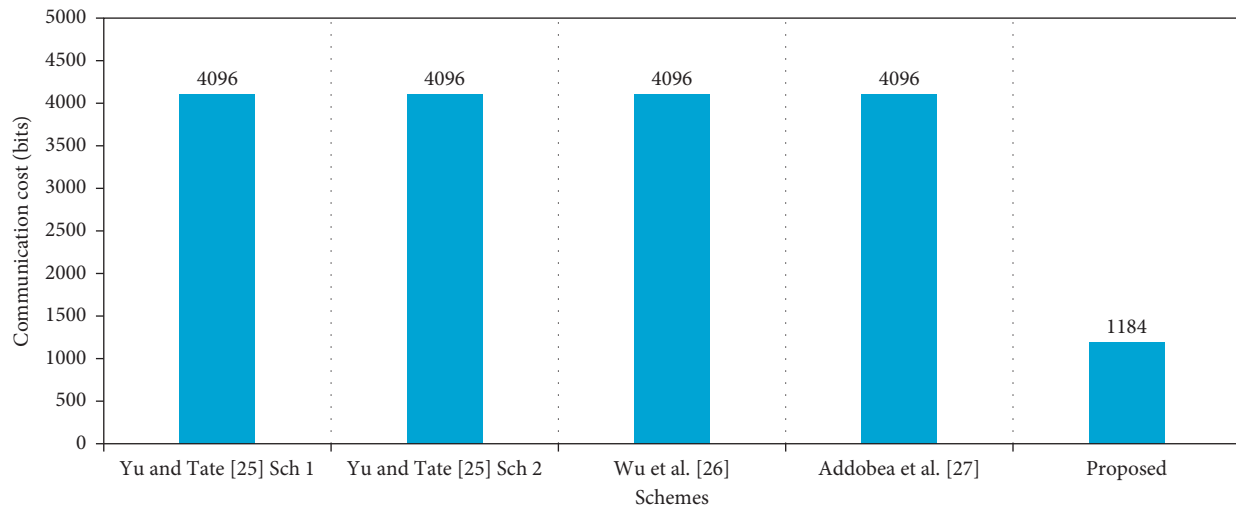


FIGURE 3: Communication cost (in bits).

Addobea et al. [27]. In comparative analysis, the variables, i.e. $|G| = 1024$ bits, $|m| = 1024$ bits, and $|n| = 80$ bits, along with the respective values, are depicted in Table 4 and illustrated in Figure 3.

7. Conclusion

The Internet of Health Things (IoHT) plays an important role as an extension of the Internet of Things (IoT) in the remote data-sharing of multiple physical processes, such as patient monitoring, treatment progression, observation, and consultation. In IoHT, multiple sensors, actuators, and controllers allow communication, computation, and interoperability, thus providing seamless connectivity with efficient resource utilization. However, for the majority of IoHT implementations, conventional cryptographic methods are not feasible due to the energy constraints of low-power embedded devices. Therefore, we suggested a lightweight security scheme in this article, using the idea of the hyperelliptic curve (HEC), called an online-offline certificateless signature scheme. In the limited key size, the HEC solution is powerful and is also acceptable for IoHT environments. The formal security analysis shows the intensity of the proposed approach in avoiding multiple attacks. In addition, after a comparative comparison with the main existing schemes, the proposed scheme proved to be efficient in terms of both computational and communication costs.

An extension of the proposed scheme is required that offers encryption and digital signature in one go. We also plan to improve the security by adding some other aspects of formal analysis, such as the real-or-random (ROR) for the solutions against different attacks. All these aspects are in the development phase and will be taken into account in our future work.

Data Availability

All data generated or analyzed during this study are included in this published article.

Conflicts of Interest

The authors declare no conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

- [1] J. J. P. C. Rodrigues, D. B. De Rezende Segundo, H. A. Junqueira et al., "Enabling technologies for the internet of health things," *IEEE Access*, vol. 6, pp. 13129–13141, 2018.
- [2] S. M. Riazul Islam, D. Daehan Kwak, M. Humaun Kabir, M. Hossain, and K.-S. Kyung-Sup Kwak, "The internet of Things for health care: a comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.
- [3] L. Catarinucci, D. De Donno, L. Mainetti et al., "An IoT-aware architecture for smart healthcare systems," *IEEE Internet of Things Journal*, vol. 2, no. 6, pp. 515–526, 2015.
- [4] Y. Yin, Y. Zeng, X. Chen, and Y. Fan, "The internet of things in healthcare: an overview," *Journal of Industrial Information Integration*, vol. 1, pp. 3–13, 2016.
- [5] M. W. Woo, J. W. Lee, and K. H. Park, "A reliable IoT system for personal healthcare devices," *Future Generation Computer Systems*, vol. 78, pp. 626–640, 2018.
- [6] B. Farahani, F. Firouzi, V. Chang, M. Badaroglu, N. Constant, and K. Mankodiya, "Towards fog-driven IoT eHealth: promises and challenges of IoT in medicine and healthcare," *Future Generation Computer Systems*, vol. 78, pp. 659–676, 2018.
- [7] F. Firouzi, A. M. Rahmani, K. Mankodiya et al., "Internet-of-things and big data for smarter healthcare: from device to architecture, applications and analytics," *Future Generation Computer Systems*, vol. 78, pp. 583–586, 2018.
- [8] X. Lin, R. Lu, X. Shen, Y. Nemoto, and N. Kato, "Sage: a strong privacy preserving scheme against global eavesdropping for ehealth systems," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 4, pp. 365–378, 2009.
- [9] S. Ullah, L. Marcenaro, and B. Rinner, "Secure smart cameras by aggregate-signcryption with decryption fairness for multi-receiver IoT applications," *Sensors*, vol. 19, no. 2, p. 327, 2019.
- [10] A. Shamir, "Identity-based cryptosystems and signature schemes," *Proceedings of the of the CRYPTO 1984*, pp. 19–23, 1984.

- [11] P. Kumar, S. Kumari, V. Sharma, A. K. Sangaiah, J. Wei, and X. Li, "A certificateless aggregate signature scheme for healthcare wireless sensor network," *Sustainable Computing: Informatics and Systems*, vol. 18, pp. 80–89, 2018.
- [12] P. Kumar, S. Kumari, V. Sharma, X. Li, A. K. Sangaiah, and S. H. Islam, "Secure cls and cl-as schemes designed for vanets," *The Journal of Supercomputing*, pp. 1–23, 2019.
- [13] M. Suárez-Albela, P. Fraga-Lamas, and T. Fernández-Caramés, "A practical evaluation on RSA and ECC-based cipher suites for IoT high-security energy-efficient fog and mist computing devices," *Sensors*, vol. 18, no. 11, p. 3868, 2018.
- [14] M. Yu, J. Zhang, J. Wang et al., "Internet of Things security and privacy-preserving method through nodes differentiation, concrete cluster centers, multi-signature, and block-chain," *International Journal of Distributed Sensor Networks*, vol. 14, no. 12, Article ID 155014771881584, 2018.
- [15] A. Braeken, "PUF based authentication protocol for IoT," *Symmetry*, vol. 10, no. 8, p. 352, 2018.
- [16] C. Zhou, Z. Zhao, W. Zhou, and Y. Mei, "Certificateless key-insulated generalized signcryption scheme without bilinear pairings," *Security and Communication Networks*, vol. 2017, Article ID 8405879, 17 pages, 2017.
- [17] S. Kumari, M. Karupiah, A. K. Das, X. Li, F. Wu, and N. Kumar, "A secure authentication scheme based on elliptic curve cryptography for IoT and cloud servers," *The Journal of Supercomputing*, vol. 74, no. 12, pp. 6428–6453, 2017.
- [18] A. A. Omala, A. S. Mbandu, K. D. Mutiria, C. Jin, and F. Li, "Provably secure heterogeneous access control scheme for wireless body area network," *Journal of Medical Systems*, vol. 42, no. 6, 2018.
- [19] C. Tamizhselvan and V. Vijayalakshmi, "An energy efficient secure distributed naming service for IoT," *International Journal of Advanced Studies of Scientific Research*, vol. 3, no. 8, 2019.
- [20] V. S. Naresh, R. Sivaranjani, and N. V.E.S. Murthy, "Provable secure lightweight hyper elliptic curve-based communication system for wireless sensor networks," *International Journal of Communication Systems*, vol. 31, no. 15, p. e3763, 2018.
- [21] A. U. Rahman, I. Ullah, M. Naeem et al., "A lightweight multi-message and multi-receiver heterogeneous hybrid signcryption scheme based on hyper elliptic curve," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 5, p. 5, 2018.
- [22] V. D. Ta, C.-M. Liu, and G. W. Nkabinde, "Big data stream computing in healthcare real-time analytics," in *IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Chengdu, China, July 2016.
- [23] S. Even, O. Goldreich, and S. Micali, "On-line/off-line digital signatures," *Advances in Cryptology—CRYPTO' 89 Proceedings*, pp. 263–275, 1990.
- [24] A. Shamir and Y. Tauman, "Improved online/offline signature schemes," *Advances in Cryptology-CRYPTO 2001*, vol. 2139, pp. 355–367, 2001.
- [25] P. Yu and S. R. Tate, "Online/offline signature schemes for devices with limited computing capabilities," in *The Cryptographers' Track at the RSA Conference 2008 (CT-RSA 2008)*, San Francisco, CA, USA, April 2008.
- [26] T. Wu, Y. Chen, and K. Lin, "ID-based online/offline signature from pairings," in *Proceedings of the International Computer Symposium (ICS2010)*, Tainan City, Taiwan, December 2010.
- [27] A. A. Addobe, J. Hou, and Q. Li, "MHCOOS: An Offline-Online Certificateless Signature Scheme for M-Health Devices," *Security and Communication Networks*, vol. 2020, Article ID 7085623, 12 pages, 2020.
- [28] S. K. H. Islam and G. P. Biswas, "Provably secure and pairing-free certificateless digital signature scheme using elliptic curve cryptography," *International Journal of Computer Mathematics*, vol. 90, no. 11, pp. 2244–2258, 2013.
- [29] M. A. Khan, I. M. Qureshi, and F. Khanzada, "A hybrid communication scheme for efficient and low-cost deployment of future flying ad-hoc network (FANET)," *Drones*, vol. 3, p. 16, 2019.
- [30] Shamus Software Ltd. <http://github.com/miracl/MIRACL>.
- [31] M. A. Khan, I. Ullah, S. Nisar et al., "An efficient and provably secure certificateless key-encapsulated signcryption scheme for flying ad-hoc network," *IEEE Access*, vol. 8, pp. 36807–36828, 2020.
- [32] M. A. Khan, I. M. Qureshi, I. Ullah, S. Khan, F. Khanzada, and F. Noor, "An efficient and provably secure certificateless blind signature scheme for flying ad-hoc network based on multi-access edge computing," *Electronics*, vol. 9, p. 30, 2020.
- [33] M. A. Khan, I. Ullah, S. Nisar et al., "Multiaccess edge computing empowered flying ad hoc networks with secure deployment using identity-based generalized signcryption," *Mobile Information Systems*, vol. 2020, Article ID 8861947, 15 pages, 2020.
- [34] I. Ullah, A. Alomari, N. Ul Amin, M. A. Khan, and H. Khattak, "An energy efficient and formally secured certificate-based signcryption for wireless body area networks with the internet of things," *Electronics*, vol. 8, no. 10, p. 1171, 2019.

Research Article

Performance Assessment of Classification Algorithms on Early Detection of Liver Syndrome

Rashid Naseem ¹, Bilal Khan,² Muhammad Arif Shah,¹ Karzan Wakil,³ Atif Khan ⁴,
Wael Alosaimi,⁵ M. Irfan Uddin ⁶ and Badar Alouffi⁷

¹Department of IT and Computer Science, Pak-Austria Fachhochschule Institute of Applied Sciences and Technology, Haripur, Pakistan

²Department of Computer Science, City University of Science and Information Technology, Peshawar, Pakistan

³Research Center, Sulaimani Polytechnic University, Sulaimani 46001 Kurdistan Region, Sulaymaniyah, Iraq

⁴Department of Computer Science, Islamia College Peshawar, Peshawar, KP, Pakistan

⁵Department of Information Technology, College of Computers and Information Technology, Taif University, P.O.Box 11099, Taif 21944, Saudi Arabia

⁶Institute of Computing, Kohat University of Science and Technology, Kohat 26000, Pakistan

⁷Department of Computer Science, College of Computers and Information Technology, Taif University, P.O.Box 1109, Taif 21944, Saudi Arabia

Correspondence should be addressed to Atif Khan; atifkhan@icp.edu.pk

Received 3 October 2020; Revised 18 November 2020; Accepted 25 November 2020; Published 12 December 2020

Academic Editor: Shah Nazir

Copyright © 2020 Rashid Naseem 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.

In the recent era, a liver syndrome that causes any damage in life capacity is exceptionally normal everywhere throughout the world. It has been found that liver disease is exposed more in young people as a comparison with other aged people. At the point when liver capacity ends up, life endures just up to 1 or 2 days scarcely, and it is very hard to predict such illness in the early stage. Researchers are trying to project a model for early prediction of liver disease utilizing various machine learning approaches. However, this study compares ten classifiers including A1DE, NB, MLP, SVM, KNN, CHIRP, CDT, Forest-PA, J48, and RF to find the optimal solution for early and accurate prediction of liver disease. The datasets utilized in this study are taken from the UCI ML repository and the GitHub repository. The outcomes are assessed via RMSE, RRSE, recall, specificity, precision, G-measure, F-measure, MCC, and accuracy. The exploratory outcomes show a better consequence of RF utilizing the UCI dataset. Assessing RF using RMSE and RRSE, the outcomes are 0.4328 and 87.6766, while the accuracy of RF is 72.1739% that is also better than other employed classifiers. However, utilizing the GitHub dataset, SVM beats other employed techniques in terms of increasing accuracy up to 71.3551%. Moreover, the comprehensive outcomes of this exploration can be utilized as a reference point for further research studies that slight assertion concerning the enhancement in extrapolation through any new technique, model, or framework can be benchmarked and confirmed.

1. Introduction

The liver is well-thought-out to be one of the central organs in any living body with fundamental functions such as processing leftover products, generating enzymes, and eliminating exhausted tissues or cells [1]. We can stay alive merely a couple of days if our liver shuts down. Fortunately, the liver can continue its role even when up to 75% of it is contaminated or removed. This is due to its astonishing

capability to produce new liver tissues from fine fettle liver cells that quiet exist [2]. It shows a significant role in several bodily functions such as protein creation and blood clotting to glucose (sugar), cholesterol, and iron metabolism. It has a range of functions, comprising eliminating toxins from the body, and is crucial for survival [3, 4]. The harm of these functions can reason to momentous destruction to the body. Once the liver is diseased with a virus, injured by chemicals, or under attack from its immune system, the elementary

hazard is similar; that is, the liver will become so spoiled that it can no longer retain an individual alive [3, 5]. According to World Health Organization (WHO) and World Gastroenterology Organization (WGO), 35 million individuals pass away due to chronic diseases, and liver failure is one of the apprehensive diseases stated [6, 7]. It is further stated that more than 50 million grown-ups will be affected with chronic liver disease (CLD), and it requests for instantaneous responsiveness for actions in a conference held in Paris that deliberated the shocking drifts of liver disease worldwide [1, 8]. Moreover, agreeing to the current figures, 25 million US residents are pretentious by the liver or biliary ailment, and out of these, 50% populace have no symptoms. In the United Kingdom, nearly 25% of death due to liver disease is from extreme alcohol drinking [9].

2. Foremost Reasons for Liver Disease

As soon as the liver becomes diseased, it can ground severe destruction to our health. There can be numerous equipment and health conditions that can naively reason for liver damage [10–12].

2.1. Alcohol. Dense alcohol drinking is the utmost collective reason for liver damage. Once individuals drink alcohol, the liver becomes distracted from its other functions and provides attention mostly on converting alcohol into a smaller amount of toxic form.

2.2. Obesity. People who are fat have the leftover quantity of body obese which inclines to accrue nearby the liver causing fatty liver disease (FLD).

2.3. Diabetes. Devising diabetes upturns the hazard of liver disease by 50 percent. Increased level of compelling insulin results in FLD.

3. Common Liver Disorder

3.1. Hepatitis. It is an ailment produced by a virus feast due to manure pollution or direct interaction with the septic bloody fluids [5].

3.2. Cirrhosis. It is the utmost severe liver disease that happens when normal liver cells are swapped by mutilation tissue as the CLD [4, 13].

3.3. Liver Cancer. The danger of consuming liver cancer is higher for individuals who have cirrhosis and another type of hepatitis [12].

In the current era, we have been confronted with a cumulative amount of records kept in several societies such as hospitals, universities, and banks that inspire us to discover an approach to mine information from this huge number of records and to proficiently use them, especially in the healthcare organizations. In the recent era, researchers are focusing on using data from healthcare organisations for

early and accurate prediction of syndromes. Nowadays, data mining (DM) and machine learning (ML) become elementary in healthcare due to its approaches, e.g., classification, clustering, and association rule mining, for determining repeated patterns pragmatic for disease extrapolation on medical data [6, 14].

In the early past, researchers have used different ML techniques for the early and accurate prediction of liver as well as some other diseases. Hassoon et al. [15] used genetic algorithm (GA) for the early prediction of liver syndromes. They have evaluated their model based on accuracy rate, specificity, sensitivity, precision, F1, and false-positive rate. The outcomes are compared with Boosted C5.0, and the results show the best performance of GA with a higher accuracy of 92.23%. Research in [16] focused on liver syndrome by taking ten significant features and using Decision Tree (DT) approaches, Naïve Bayes (NB), and NBTree (NBT) techniques to classify the syndrome's indications. Lastly, they perceived that the NBT technique is most precise than NB for emancipating rules. In [14], for forecasting liver syndrome, they used NB and support vector machine (SVM) for classification, and as a final point, they originate that SVM has better concert and accuracy in liver syndrome classification. A new approach of classification that will relieve suitable and interpretable rules is recursive-rule extraction (Re-RX) which is utilized in [17] to extract more and effective rules for the liver syndrome analysis.

In [18], for discovering the actual rules on liver syndrome analysis, C4.5 procedure smears as one of the well-known DT procedures in classification. Like, in [18, 19], C4.5 technique is used, and the researchers strained to utilize the technique for identifying liver syndrome. We can comprehend that C4.5 has a virtuous response on various types of disease analysis such as diabetes [20] and breast cancer [21]. Likewise, in [6], there is an assessment among C5.0 and CHAID techniques on the liver syndrome, and lastly, they found out that boosted C5.0 has a better response on discovering effectual rules. Boosting is a technique used in the C5.0 technique to increase this version over C4.5. Similarly, it increases the accuracy rate and the runtime of the algorithm [6].

However, the persistence of this study is the performance analysis of various ML classification algorithms on the liver disease dataset taken from UCI ML repository and GitHub repository. The classification algorithms include average one dependency estimator (AIDE), multilayer perceptron (MLP), NB, K-nearest neighbour (KNN), SVM, composite hypercube on iterated random projection (CHIRP), credal decision tree (CDT), forest by penalizing attributes (Forest-PA), decision tree (J48), and random forest (RF). To evaluate the performance analysis of these classifiers, different performance assessment measures are utilized which embrace root relative squared error (RRSE), root mean squared error (RMSE), specificity, precision, recall, F-measure, G-measure, Matthew's correlation coefficient (MCC), and accuracy.

The rest of the paper is prepared as follows: Section 2 contains the methodology of this research that comprises

further subsections of dataset description, performance assessment measures, and review of employed techniques. Section 3 grants the experimental results and discussion, and Section 4 and Section 5, respectively, present the threats to validity and the overall conclusion of this research.

4. Methodology

This research aims to present the performance analysis of ML classification algorithms for liver disease prophecy on two different datasets occupied from GitHub and UCI ML repositories. The complete research is prepared via the procedure shown in Figure 1. After the selection of datasets, a preprocessing step is applied on each dataset for two main purposes: replacing the missing values and changing the class attribute from numerical to categorical due to some of the techniques that do not work on numerical class attributes. After all, when ML techniques are applied to each dataset, the outcomes are assessed using different assessment measures to show the better performance of an individual technique. For this, nine assessment measures, namely, RMSE [22–24], RRSE [25], specificity [26–28], precision [29–31], recall [27, 29, 32], F-measure [29, 30, 33], G-measure [22, 34], MCC [29, 35, 36], and accuracy [3, 37, 38], are utilized to assess the performance of ML classification algorithm going on liver datasets.

4.1. Datasets Description. Each dataset is consisting of some attributes along with a known output class. Respectively, datasets contain numerical data, while the total number of attributes and instances is different. There are two liver datasets utilized in this study. One is taken from the UCI ML repository (<https://archive.ics.uci.edu/ml/datasets/liver+disorders>), and the second is from the GitHub repository (<https://github.com/SanikaVT/Liver-disease-prediction>). Table 1 presents the details of the attributes of the dataset taken from the UCI ML repository, whereas Table 2 presents the same for the dataset taken from the GitHub repository. The first dataset (taken from the UCI ML repository) comprises seven features in which the first five features are all blood examinations which are believed to be thoughtful to liver diseases that might arise from extreme alcohol feeding. There are a total of 345 records in this dataset amid these 345: 145 are liver patients, and the rest of 200 are nonliver patient's records. In the second dataset (taken from GitHub repository), eight features are all blood tests, which is supposed to be thoughtful to liver disorder. This dataset contains a total of 583 records. Among these records, 416 are the liver patients, while the rest 167 are nonliver patient's records. Figure 2 shows the percentage of liver patients and nonliver patients in both datasets. In each dataset, the last attribute is known as a selector containing the value 1 or either 2. Value 1 represents that the person is a positive liver patient, whereas 2 shows the nonliver patients' records. Figure 2 shows the number of liver patients and nonliver patients for each dataset.

4.2. Performance Measurement Parameters. Performance assessment of every model utilized is a significant part of any research study. A model may produce satisfactory results when it is assessed using standard assessment measures. However, in this study, two types of assessment measures are used in which some are utilized for evaluating error rate that includes RMSE [25, 39] and RRSE [25], while others are employed for the assessment of accuracy that comprises specificity [5, 40], precision [32, 41], recall [31, 42], F-measure [29, 36], G-measure [22, 34], MCC [29, 35, 36], and accuracy [3, 37, 38]. Table 3 shows the equation for calculating each assessment measure with equations, where $|y_i - y|$ is the absolute error, n is the number of errors, T_j is the goal value for record ji , P_{ij} is the prediction rate by the particular model I for data j (out of n records), TP presents the true-positive classification, FN shows the false-negative classification, TN grants the true-negative classification, and FP is the rate of false-positive classifications.

5. Summarization of Employed Techniques

This subsection comprises a brief review of techniques employed in this research and contrasted with RF.

5.1. Average One Dependency Estimator. A1DE is a probabilistic technique used for mostly classification problems. It succeeds in extreme precise classification by averaging inclusive of a minor space of different NB-like models that have punier independence suppositions than NB. A1DE was designed to address the attribute-independence issues of a popular NB technique. It was designed to address the attribute-independence issues of the prevalent NB classifier [43].

5.2. Naïve Bayes. NB is known as the kinfolk of modest probabilistic classifiers grounded on Bayes hypothesis with individuality suppositions amid the predictors [44, 45]. NB model is precise simple to construct and can be executed for any dataset containing a large amount of data. The posterior probability $P(c/x)$ is taken as of $P(c)$, $P(x)$, and $P(x/c)$. The consequence of the rate of a predictor (x) on assumed class (c) is autonomous of the rate of other predictors.

5.3. Multilayer Perceptron. MLPs are deliberated as the utmost momentous classes of the neural network comprising an input layer, at least one hidden layer, and an output layer [46, 47]. The techniques behind the neural network are that when data are accessible as the input layer, the network neurons start calculation in the sequential layer till an output value is gained at each of the output neurons. A threshold node is moreover added to the input layer which identifies the weight function [48].

5.4. Support Vector Machine. It is a managed learning technique that has several uses in the ground of classification, biophotonics, and pattern recognition [22]. Firstly, it

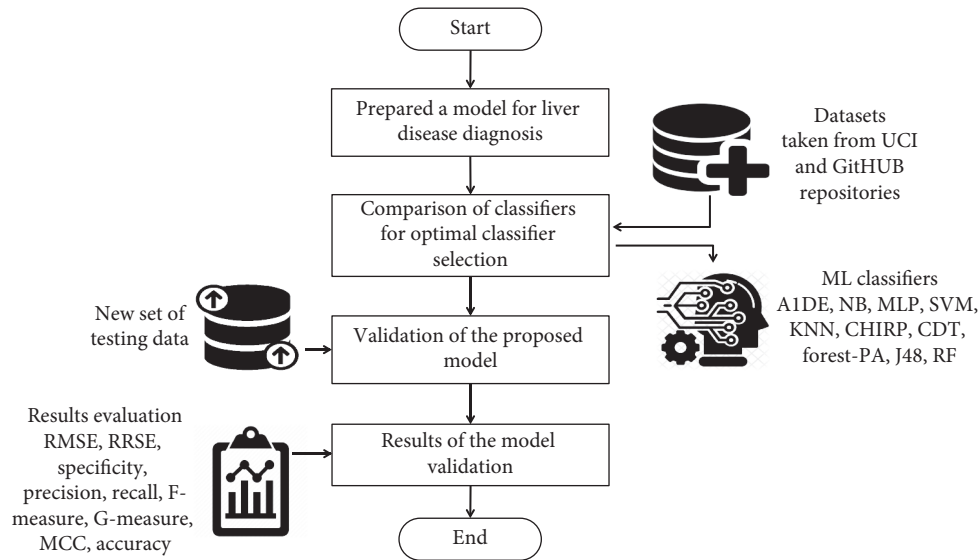


FIGURE 1: Methodology workflow diagram.

TABLE 1: List of dataset attributes and descriptions taken from the UCI ML repository.

S. no.	Attribute	Value type	Normal value range	Description
1	MCV	Integer	75–95	Mean corpuscular volume
2	Alkphos	Integer	63–2110	Alkaline phosphatase
3	SGPT	Integer	10–2000	Alanine aminotransferase
4	SGOT	Integer	10–4929	Aspartate aminotransferase
5	GammaGT	Integer	12–64	Gamma-glutamyl transpeptidase
6	Drinks	Real	-	Number of half-pint equivalents of alcoholic beverages % drunk per day
7	Selector	Selector {1, 2}	-	Field used to split data into two sets

TABLE 2: List of dataset attributes and descriptions taken from the GitHub repository.

S. no.	Attribute	Value type	Normal value range	Description
1	Age	Integer	4–90 years	Age of the patient
2	Gander	Text	Male/female	Gander of the patient
3	TB	Integer	0.4–75	Total bilirubin
4	DB	Integer	0.1–19.7	Direct bilirubin
5	Alkphos	Integer	63–2110	Alkaline phosphatase
6	SGPT	Integer	10–2000	Alanine aminotransferase
7	SGOT	Integer	10–4929	Aspartate aminotransferase
8	TP	Integer	2.7–9.6	Total proteins
9	ALB	Integer	0.9–5.5	Albumin
10	A/G ratio	Integer	0.3–2.8	Albumin and globulin ratio
11	Selector	Integer	1–2	Field used to split data into two sets

was developed for binary classification; however, it can also be used for multiple classes [41]. In binary classification, SVM classifies data by finding the best hyperplane that separates all data points in one class from those in the other class. In that case, if data are linearly inseparable, a mathematical function is utilized to transmute the records to an advanced dimensional space such that it possibly will grow into linear divisible in the new space [27].

5.5. K-Nearest Neighbour. KNN is a supervised learning technique where the preparation of features attributes to

forecast the class of new test data. KNN classifies the first-hand data grounded on leased space from the new records to the k -nearest neighbors [9]. The nearest distance can be found using different distance functions like Manhattan distance (MD), Euclidean distance (ED), and Minkowski distance (MkD) [49].

5.6. Composite Hypercube on Iterated Random Projection. It is a reiterative module of three levels: anticipating, binning, and covering, which projected to a defrayal with the thorn in your side of computational unpredictability,

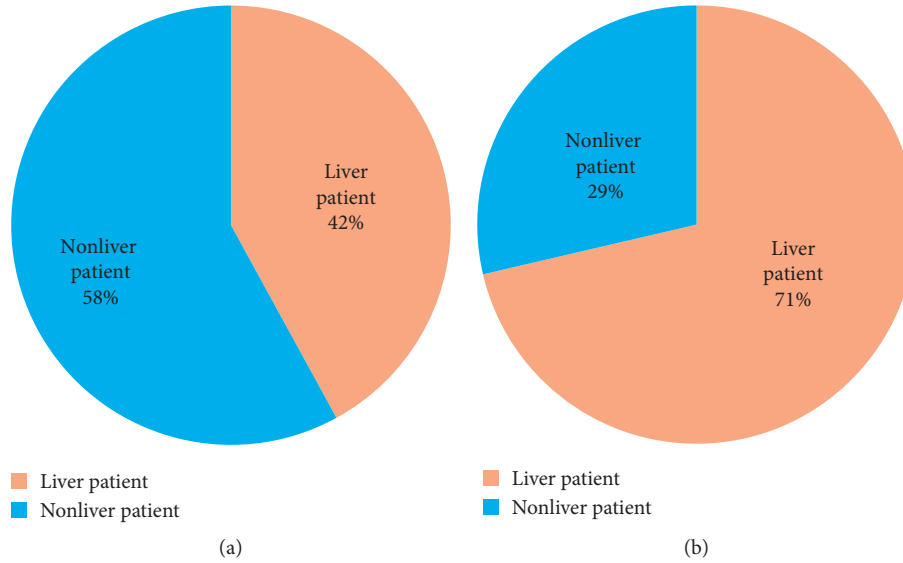


FIGURE 2: Number of liver patient and nonliver patient for each dataset: (a) UCI ML repository; (b) GitHub repository.

TABLE 3: Performance assessment measures to evaluate the experimental results.

S. no.	Measure	Description and equation
1	RMSE	$RMSE = \sqrt{1/2 \sum_{j=1}^n (y_i - 1)^2}$
2	RRSE	$RRSE = \sqrt{\sum_{j=1}^n (P_{ij} - T_j)^2 / \sum_{j=1}^n (T_j - T)^2}$
3	Specificity	$Specificity = TN / (FP + TN)$
4	Precision	$Precision = TP / (TP + FP)$
5	Recall	$Recall = TP / (TP + FN)$
6	F-measure	$FM = 2 * precision * recall / (precision + recall)$
7	G-measure	$GM = 2 * recall * specificity / (recall + specificity)$
8	MCC	$MCC = (TN * TP) - (FN * FP) / \sqrt{(FP + TP)(FN + TP)(TN + FP)(TN + FN)}$
9	Accuracy	$Accuracy = TP + TN / (TP + TN + FP + FN)$

dimensionality, and nonlinear recognisability [50]. CHIRP is not the cascading of diverse techniques, also not the enhancement or modification of attractive techniques; it utilizes new packaging techniques. The exactness of this technique usually utilized unbiased datasets and leaves behind the accuracy of contestants. The CHIRP uses computationally convincing ways to deal with accumulating 2D predictions and sets of quadrangular regions on those predictions that include valuations from a separable crowd of data. CHIRP categorizes these crowds of forecasts and segments them into a final incline for the accumulation of new data estimation [51].

5.7. Credal Decision Tree. CDT is a technique to design classifiers grounded on inexact possibilities and improbability measures [52]. Throughout the creation procedure of a CDT, toward sidestep producing an also problematical decision tree, a new standard remained presented: stop once the overall improbability rises because of the splitting of the decision tree. The function utilized in the overall indecision dimension can be fleetingly articulated as in [53, 54].

5.8. Forest by Penalizing Attributes. Forest-PA uses bootstrap samples and penalized attributes. It purposes to construct a group of extremely precise decision trees by manipulating the strong point of entirely nonclass features presented in a dataset, not like certain current techniques that utilized a subgroup of the nonclass features. Next to a similar time to support robust assortment, Forest-PA enforces disadvantages (detrimental weights) en route for individual's features that contributed to happening on the newest tree to produce the consequent trees. Forest-PA moreover consumes a contrivance toward step-by-step rise loads from the features that have not been verified in the consequent tree(s) [55].

5.9. Decision Tree (J48). This is the basic C4.5 Decision Tree (DT) used for classification problems [37]. It is the deviation of information gain (IG), usually utilized to stun the result of biasness. An attribute using a maximum gain ratio is nominated in direction to shape a tree as a dividing attribute. Gain ratio- (GR-) based DT performs well as compare to IG, in terms of accuracy [4].

5.10. Random Forest. RF produces a set of techniques that involve constructing an ensemble or so-termed as a forest of decision trees from a randomized variation in tree induction techniques [1]. RF works through forming a mass of decision trees at the preparation period and outputting the group in the approach of the group output by a single tree. It is deliberated as one of the utmost techniques which is extremely proficient for both classification and regression problems [56].

6. Experimental Results

This section comprises the experimental analysis of liver syndrome prophecy utilizing ten ML classifiers. For training and testing, 10-fold cross-validation is used which is a standard methodology for assessments [41]. The ML classifiers are evaluated on the dataset available online on the UCI ML repository and GitHub repository. The overall experimental analysis shows the error rates (achieved via RMSE and RRSE) as well as accuracy (succeeded through specificity, recall, precision, G-measure, F-measure, MCC, and accuracy). The experimental analysis is subdivided into two sections that are scenario 1 and scenario 2. Scenario 1 represents the outcomes of algorithms employed on dataset taken from the UCI ML repository, while scenario 2 represents the same on dataset taken from the GitHub repository.

6.1. Experimental Results: Scenario 1 (UCI Dataset). Here, firstly, we discuss the experiments carried out to find the minimum error rate assessed by RMSE and RRSE achieved via each classifier. These results are given in Table 4 where the second column shows the list of employed classifiers while the third column and fourth column, respectively, represent the results of RMSE and RRSE. This table shows that RF outperforms other classifiers in terms of reducing error rates; the results are 0.4328 for RMSE and 87.6766 for RRSE. In the rest of the classifiers, MLP produces better results in reducing both RMSE and RRSE, and the results achieved are 0.4532 and 91.6375, respectively.

Table 5 shows the detail of correctly classified instances (CCIs) and incorrectly classified instances (ICIs) amid an overall of 345 instances. The greater CCI rates show the best performance of an individual classifier. Table 6 represents the standings of confusion matrix (CM), while Table 7 represents the CM for all the assessments calculated throughout experimentations. There are binary classes in which predicting is promising, i.e., class 1 and class 2. Class 1 is also known as positive, while class 2 is known as negative. If we predict the existence of a disease, in the case, class 1 proceeds that the individual ensures the disease, while class 2 proceeds that the individual does not ought to the disease. Here, TP is the situation where the persistent as positive (they ought to the disease), and FP is likewise the condition of positive, but they ought no to the disease, which is known as type 1 error. FN illustrates the negative conditions, but they in fact ought to the disease which is called type 2 error. TN demonstrates a negative situation, which indicates that

TABLE 4: RMSE and RRSE outcomes assessments.

S. no.	Classifier	RMSE	RRSE
1	A1DE	0.4995	101.1922
2	NB	0.5083	102.9673
3	MLP	0.4523	91.6375
4	SVM	0.6461	130.8811
5	KNN	0.6072	123.0036
6	CHIRP	0.5357	108.5209
7	CDT	0.5005	101.3988
8	Forest-PA	0.4563	92.4357
9	J48	0.5025	101.8061
10	RF	0.4328	87.6766

TABLE 5: Results of CCI and ICI achieved via each classifier.

S. no.	Technique	CCI	ICI
1	A1DE	194 (56.2%)	151 (43.8%)
2	NB	191 (55.4%)	154 (44.6%)
3	MLP	247 (71.6%)	98 (28.4%)
4	SVM	201 (58.3%)	144 (41.7%)
5	KNN	217 (62.9%)	128 (37.1%)
6	CHIRP	246 (71.3%)	99 (28.7%)
7	CDT	219 (63.5%)	126 (36.5%)
8	Forest-PA	241 (69.9%)	104 (30.1%)
9	J48	237 (68.7%)	108 (31.3%)
10	RF	249 (72.2%)	96 (27.8%)

TABLE 6: Terms of confusion matrix.

	Positive or class 1 (1)	Negative or class 2 (0)
Positive or class 1 (1)	True positive	False positive
Negative or class 2 (0)	False negative	True negative

TABLE 7: Confusion matrix for all classifiers.

S. no.	Technique	TP	FP	FN	TN
1	A1DE	33	112	39	161
2	NB	111	34	120	80
3	MLP	83	62	36	164
4	SVM	1	144	0	200
5	KNN	82	63	65	135
6	CHIRP	82	63	36	164
7	CDT	60	85	41	159
8	Forest-PA	74	71	33	167
9	J48	77	68	40	160
10	RF	90	55	41	159

they ought not to the disease. The values of CM are employed in finding complete accuracy outcomes. In our case, these are specificity, recall, precision, G-measure, F-measure, MCC, and accuracy according to equations (see Table 3).

Table 8 signifies the assessed outcomes of specificity, precision, recall, F-measure, G-measure, MCC, and accuracy concerning each classifier. The values of each of these measures are calculated with help of CM (see Table 7). The best performance of each classifier assessed via every

TABLE 8: Outcomes assessed via specificity, recall, precision, G-measure, F-measure, MCC, and accuracy.

S. no.	Technique	Specificity	Precision	Recall	F-measure	G-measure	MCC	Accuracy
1	A1DE	0.5897	0.2276	0.4583	0.3041	0.5158	0.0396	56.2319
2	NB	0.7018	0.7655	0.4805	0.5904	0.5704	0.1737	55.3623
3	MLP	0.7257	0.5724	0.6975	0.6288	0.7113	0.4075	71.5942
4	SVM	0.5814	0.0069	1	0.0137	0.7353	0.0633	58.2609
5	KNN	0.6818	0.5655	0.5578	0.5616	0.6136	0.2401	62.8986
6	CHIRP	0.7225	0.5655	0.6949	0.6236	0.7084	0.4011	71.3043
7	CDT	0.6516	0.4138	0.5941	0.4878	0.6215	0.2265	63.4783
8	Forest-PA	0.7017	0.5103	0.6916	0.5873	0.6966	0.3685	69.8551
9	J48	0.7018	0.531	0.6581	0.5878	0.6792	0.3452	68.6957
10	RF	0.743	0.6207	0.687	0.6522	0.7139	0.4228	72.1739

evaluation metric is mentioned in bold. This analysis shows that, by evaluating each classifier through specificity, F-measure, MCC, and accuracy, RF outperforms other classifiers and achieved better results. The details of according to these measures are presented in Figure 3 while Figure 4 presents the accuracy details. In the case of precision, NB results are better than the rest of the classifiers while on recall and G-measure, SVM outperforms other classifiers employed. Figure 5 shows the percentage difference in terms of accuracy between RF and other employed classifiers. The difference is calculated via the following equation:

$$\text{percentage difference} = \frac{|v_i - v_j|}{(v_i + v_j/2)} * 100, \quad (1)$$

where v_i and v_j are the values in which the difference is to be calculated.

Figure 4 illustrates that there is very little difference between RF and MLP and RF and CHIRP, which is 0.81% and 1.21%, respectively.

6.2. Experimental Results: Scenario 2 (GitHub Dataset). Here, first, we discuss the experiment carried out to find the minimum error rate assessed by RMSE and RRSE achieved via an individual classifier. The outcomes are shown in Table 9 where the second column represents the list of employed classifiers while the third column and fourth column, respectively, represent the results of RMSE and RRSE. This table shows that RF outperforms other classifiers in terms of reducing error rates, and the results are 0.4225 for RMSE and 93.4416 for RRSE. Despite the classifiers, MLP outperforms other classifiers in terms of reducing the error rate. The results achieved via MLP are 0.4276 and 94.5776 in that order for RMSE and RRSE.

Table 10 presents the details of CCI and ICI among a total of 583 instances. The larger ICI rate shows the best performance of that classifier. Table 11 represents the CM for all the assessments assessed throughout the experiments.

Table 12 signifies the outcome assessed via specificity, precision, recall, F-measure, G-measure, MCC, and accuracy. These outcomes show the best performance of three different classifiers for different assessment measures. According to these analyses, A1DE beats other classifiers in

terms of better results of specificity and G-measure that are 0.4680 and 0.5934 accordingly. NB outperforms other techniques in terms of good results for recall and MCC that are 0.9540 and 0.3469, respectively. However, SVM outperforms other classifiers by increasing the rate of precision, F-measure, and accuracy. The results achieved are 1 for precision, 0.8328 for F-measure, and 71.3551% accuracy. These outcomes are illustrated in Figure 6, while Figure 7 represents the accuracy details of each classifier which shows the best performance of SVM. The accuracy difference between SVM and other classifiers is presented in Figure 8.

7. Results Discussion

This research focuses on the performance analysis of ten various and well-known ML classification algorithms on two different liver disease datasets taken from the UCI ML repository and GitHub repository. On both datasets, results, after the evaluation is different due to each dataset, contain different amounts of instances, attributes, dataset according to attributes, and, the most important, different amount (percentage) of affected and nonaffected patient records. Table 13 shows the better performance of optimal classifiers on both datasets concerning each assessment measure. These analyses illustrate that, in terms of reducing the error rate on both datasets, RF outperforms other classifiers. Moreover, RF also outclasses additional employed techniques in reports of increasing accuracy on the dataset in use from the UCI ML repository. This is because RF is an excessive classifier with high-dimensional data; meanwhile, we are at work with subsets of data. To succeed in the prediction using the trained RF, classifier desires to permit the test features through the information of each randomly generated tree [7, 57]. RFs agonize fewer overfitting to a specific dataset than simple trees. RFs were constructed via merging the forecasts of numerous trees that are trained in separation, which provide valuable internal assessments of strength, error, correlation, and variable prominence [29, 58]. However, on the UCI dataset, SVM produces better results for recall and G-measure assessment measures. On the contrary, on the dataset taken from the GitHub repository, SVM performs better in terms of increasing accuracy as well as precision and F-measure. The SVM is the progressive tool with thoroughgoing classification algorithms surrounded in statistical learning theory [14]. It utilizes a nonlinear

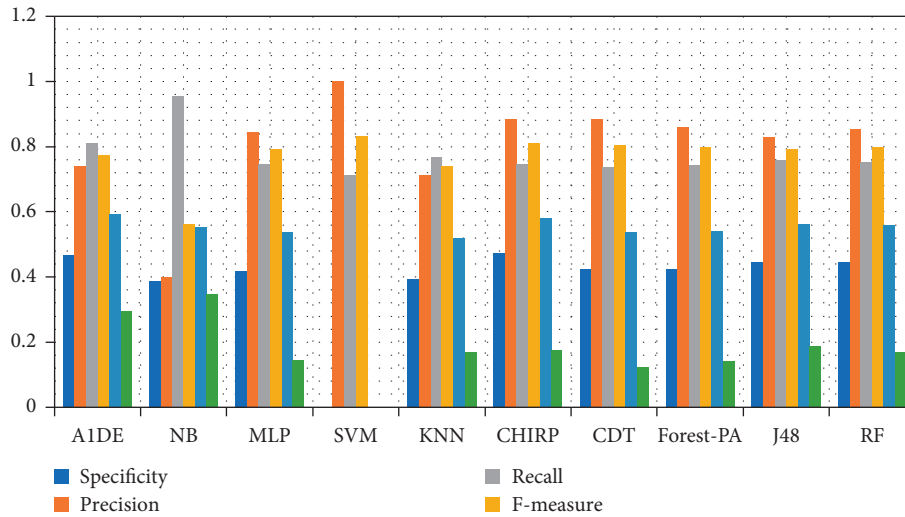


FIGURE 3: Specificity, precision, recall, F-measure, G-measure, and MCC analysis representation.

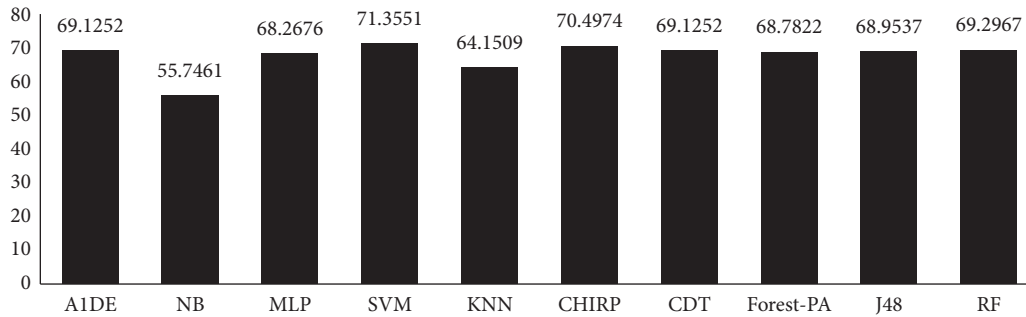


FIGURE 4: Accuracy achieved via each classifier.

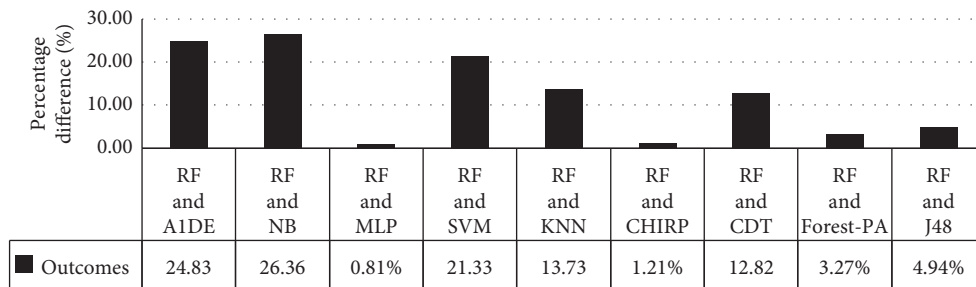


FIGURE 5: Accuracy percentage difference between RF and other employed classifiers.

TABLE 9: RMSE and RRSE outcomes assessments.

S. no.	Technique	RMSE	RRSE
1	A1DE	0.4479	99.074
2	NB	0.6541	144.6684
3	MLP	0.4276	94.5776
4	SVM	0.5352	118.3788
5	KNN	0.5976	132.1834
6	CHIRP	0.5432	120.1379
7	CDT	0.4492	99.3545
8	Forest-PA	0.4379	96.8574
9	J48	0.4797	106.1058
10	RF	0.4225	93.4416

TABLE 10: Results of CCI and ICI achieved via each classifier.

S. no.	Technique	CCI	ICI
1	A1DE	403 (69.1%)	180 (30.9%)
2	NB	325 (55.7%)	258 (44.3%)
3	MLP	398 (68.3%)	185 (31.7%)
4	SVM	416 (71.4%)	167 (28.6%)
5	KNN	374 (64.2%)	209 (35.8%)
6	CHIRP	411 (70.5%)	172 (29.5%)
7	CDT	403 (69.1%)	180 (30.9%)
8	Forest-PA	401 (68.8%)	182 (31.2%)
9	J48	402 (69%)	181 (31%)
10	RF	404 (69.3%)	179 (30.7%)

TABLE 11: Confusion matrix for all classifiers.

S. no.	Technique	TP	FP	FN	TN
1	A1DE	308	108	72	95
2	NB	166	250	8	159
3	MLP	351	65	120	47
4	SVM	416	0	167	0
5	KNN	297	119	90	77
6	CHIRP	368	48	124	43
7	CDT	367	49	131	36
8	Forest-PA	358	58	124	43
9	J48	345	71	110	57
10	RF	355	61	118	49

TABLE 12: Outcomes assessed via specificity, recall, precision, G-measure, F-measure, MCC, and accuracy.

S. no.	Technique	Specificity	Precision	Recall	F-measure	G-measure	MCC	Accuracy
1	A1DE	0.4680	0.7404	0.8105	0.7739	0.5934	0.2935	69.1252
2	NB	0.3886	0.399	0.9540	0.5627	0.5524	0.3469	55.7461
3	MLP	0.4196	0.8438	0.7452	0.7914	0.5369	0.1437	68.2676
4	SVM	#DIV/0!	1	0.7136	0.8328	#DIV/0!	#DIV/0!	71.3551
5	KNN	0.3929	0.7139	0.7674	0.7397	0.5197	0.1675	64.1509
6	CHIRP	0.4725	0.8846	0.748	0.8106	0.5792	0.177	70.4974
7	CDT	0.4235	0.8822	0.7369	0.8031	0.5379	0.1253	69.1252
8	Forest-PA	0.4257	0.8606	0.7427	0.7973	0.5412	0.141	68.7822
9	J48	0.4453	0.8293	0.7582	0.7922	0.5611	0.1864	68.9537
10	RF	0.4454	0.8534	0.7505	0.7987	0.5591	0.1696	69.2967

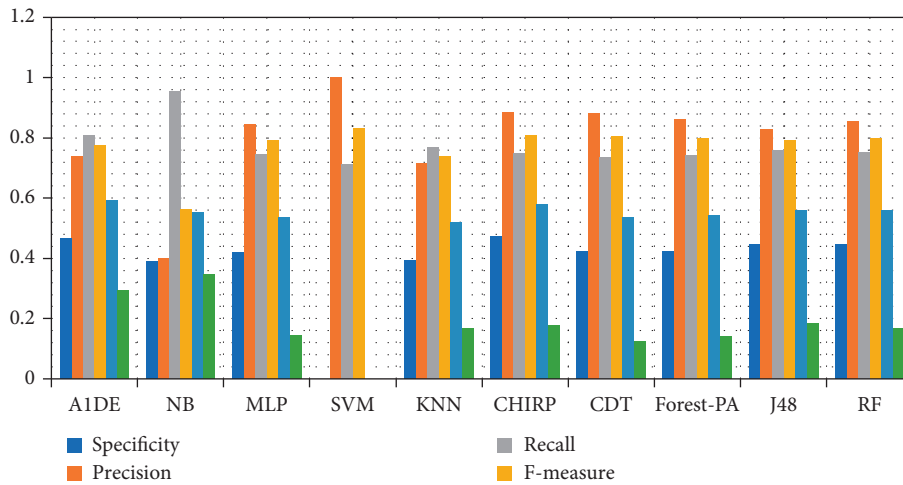


FIGURE 6: Specificity, recall, precision, MCC, F-measure, and G-measure analysis representation.

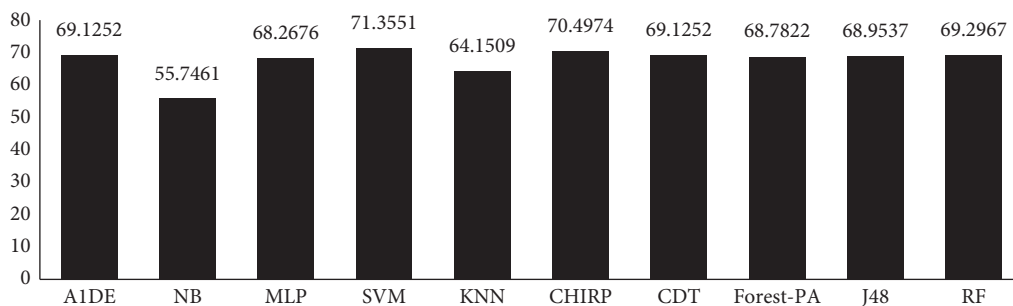


FIGURE 7: Accuracy representation.

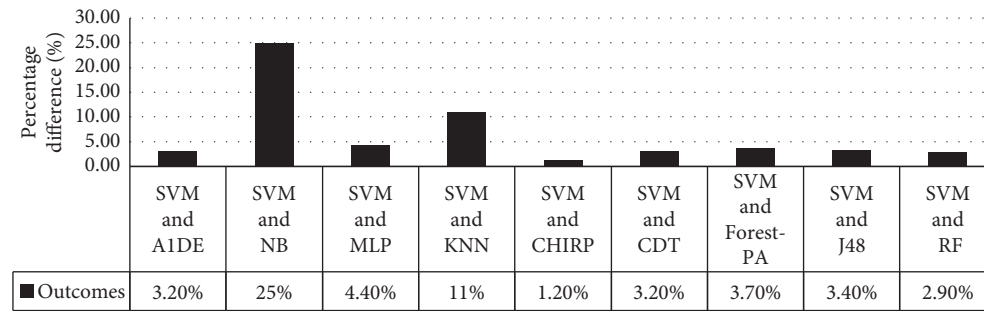


FIGURE 8: Accuracy percentage difference between SVM and other employed classifiers.

TABLE 13: Performance of optimal classifiers on datasets according to each assessment measures.

S. no.	Assessment measures	Dataset from UCI ML repository	Dataset from GitHub repository
1	RMSE	RF	RF
2	RRSE	RF	RF
3	Specificity	RF	A1DE
4	Precision	NB	SVM
5	Recall	SVM	NB
6	F-measure	RF	SVM
7	G-measure	SVM	A1DE
8	MCC	RF	NB
9	Accuracy	RF	SVM

mapping to recondition the exclusive training data keen on a higher dimension [59]. Conversely, on the same dataset, A1DE also performs better in terms of increasing the rate of specificity and G-measure while NB does the same for recall and MCC.

7.1. Model Preparation. A model for liver syndrome prophecy is proposed, evaluated, and validated to test and compare results of ten various ML classification algorithms including A1DE, NB, MLP, SVM, KNN, CHIRP, CDT, Forest-PA, J48, and RF, and as the results revealed that RF is best suitable classifier in the environment related to prediction of liver syndromes in reports of both increasing accuracy and reducing error rate on the dataset occupied from UCI ML repository. However, on the dataset taken from GitHub repository, SVM is the optimal solution for increasing accuracy although RF is the best solution to reduce the error rates.

7.2. Objectives Accomplished

7.2.1. Objective 1. It was to propose a model for liver syndrome prophecy that will help to increase the accuracy and reduce error rate in early prophecy.

7.2.2. Objective 2. It was to compare the results of classification algorithms to achieve most optimal solution for early and accurate prediction of liver syndromes.

7.3. Threats to Validity. This section contains the effects that might anguish the cogency of this research work.

Internal Validity. The exploration of this research is grounded proceeding diverse and very familiar evaluation standards that are used in the past in various studies. Amid these standards, several techniques are used to assess the error rate while certain techniques were used to assess the accuracy. So, the treat can be that renewal of new evaluation standards as a replacement for utilized standards can decrease the accuracy. Furthermore, the techniques used in this exploration can be supplanted using several newest techniques or can be cascaded with each other that can harvest enhanced outcomes as compared to the employed techniques.

External Validity. This study piloted investigations on two datasets occupied from UCI ML and GitHub repositories. The threat to validity might rise due to the condition of relating the projected techniques in other existent data composed from the various medical organizations or replacing these datasets with some other datasets, which may distress the outcomes while growing the error rates. Likewise, the projected technique possibly will not be capable toward harvesting improved forecast in outcomes via certain additional datasets. Hence, this research concentrated on datasets available on UCI ML repository and GitHub repository to measure the performance of the employed techniques.

Construct Validity. In this research, diverse ML techniques remain benchmarked through each other, going on liver dataset occupied from UCI ML and GitHub repositories using several valuation measures. The assortment of techniques utilized in this study is on the center of their progressive characteristic above the other techniques that ought to be exploited by the canvassers in the last decades. However, it can be a threat if we put on some other new

techniques, and the outcomes can be improved probably than the projected techniques. In addition, the increase or decrease in training or testing samples from the dataset has a significant impact on the error rate. Likewise, choosing a different number of folds during K-fold validation has a dramatic effect on the error rate. The newest evaluation standards can also produce improved outcomes that can beat current accomplished outcomes.

8. Conclusions

Liver diseases are rising on daily basis, and it is hard to foresee these ailments in the early premise. Researchers have utilized a large number of ML techniques to foresee such ailments in the initial stage, but still there is need to improve accuracy as well as reduce error rates in the projected models. However, in this study, ten different ML classifiers including A1DE, NB, MLP, SVM, KNN, CHIRP, CDT, Forest-PA, J48 and RF are benchmarked on two different liver disease datasets taken from UCI ML repository and GitHub repository. For the assessments of these classifiers, nine standard assessment standards are utilized which are RMSE, RRSE, specificity, recall, precision, G-measure, F-measure, MCC, and accuracy. The overall experiments in use on UCI ML repository dataset show the best performance of RF. RMSE and RRSE results of RF are 0.4328 and 87.6766 correspondingly, while accuracy is 72.1739%. Moreover, RF also performs better in terms of reducing error rate on the dataset from GitHub repository, and the achieved results are 0.4225 and 93.4416, respectively, for RMSE and RRSE. However, in terms of increasing accuracy on the GitHub repository dataset, SVM achieved a higher accuracy of 71.3551%.

8.1. The Major Contributions of This Research

We associate the results of ten ML classifiers including A1DE, NB, MLP, SVM, KNN, CHIRP, CDT, Forest-PA, J48, and RF.

We acquit a series of experiments on liver disease datasets accessible on UCI ML and GitHub repositories.

To deliver vision into the experimental outcomes, evaluation is conceded out via RRSE, RMSE, specificity, recall, precision, G-measure, F-measure, MCC, and accuracy.

8.2. Significance Statement. In this study, we employed ten ML classifiers on two different liver disease datasets that are occupied from the UCI ML repository including 345 cases and GitHub repository enclosing 583 cases. The results of stated techniques have been compared with characterising the utmost accurate technique that conveys around categorizing the affected and nonaffected patients with less error rate and high accuracy. This study recommended the RF and SVM are the best techniques that can be employed by physicians so as to exterminate treatment and diagnostic errors.

Data Availability

The data utilized for finding the outcomes of this research have been taken from UCI ML and GitHub repositories available at <https://archive.ics.uci.edu/ml/datasets/liver+disorders> and <https://github.com/SanikaVT/Liver-disease-prediction>, respectively.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by Taif University Researchers Supporting Project number (TURSP-2020/254), Taif University, Taif, Saudi Arabia.

References

- [1] A. N. Arbain and B. Y. P. Balakrishnan, "A comparison of data mining algorithms for liver disease prediction on imbalanced data," *International Journal of Data Science and Analytics*, vol. 1, 2019.
- [2] S. Dhamodharan, "Liver disease prediction using bayesian classification," in *Proceedings of the 4th National Conference on Emerging Computing Technologies*, pp. 1–3, Maharashtra, India, May 2014.
- [3] N. Nahar and F. Ara, "Liver disease prediction by using different decision tree techniques," *International Journal of Data Mining & Knowledge Management Process*, vol. 8, no. 2, pp. 01–09, 2018.
- [4] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "A systematic machine learning based approach for the diagnosis of non-alcoholic fatty liver disease risk and progression," *Scientific Reports*, vol. 8, no. 1, pp. 1–12, 2018.
- [5] M. Abdar, N. Y. Yen, and J. C.-S. Hung, "Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees," *Journal of Medical and Biological Engineering*, vol. 38, no. 6, pp. 953–965, 2018.
- [6] M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, "Performance analysis of classification algorithms on early detection of liver disease," *Expert Systems with Applications*, vol. 67, pp. 239–251, 2017.
- [7] L. Lau, Y. Kankanige, B. Rubinstein et al., "Machine-learning algorithms predict graft failure after liver transplantation," *Transplantation*, vol. 101, no. 4, pp. e125–e132, 2017.
- [8] B. Khan, R. Naseem, M. Ali, M. Arshad, and N. Jan, "Machine learning approaches for liver disease diagnosing," *International Journal of Data Science and Advanced Analytics*, vol. 1, pp. 27–31.
- [9] U. R. Acharya, H. Fujita, S. Bhat et al., "Decision support system for fatty liver disease using GIST descriptors extracted from ultrasound images," *Information Fusion*, vol. 29, pp. 32–39, 2016.
- [10] D. Grissa, D. Nytoft Rasmussen, A. Krag, S. Brunak, and L. Juhl Jensen, "Alcoholic liver disease: a registry view on comorbidities and disease prediction," *PLoS Computational Biology*, vol. 16, no. 9, Article ID e1008244, 2020.
- [11] S. Hashem, M. ElHefnawi, S. Habashy et al., "Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease," *Computer*

- Methods and Programs in Biomedicine*, vol. 196, p. 105551, 2020.
- [12] B. Losic, A. J. Craig, C. Villacorta-Martin et al., "Intratympanic heterogeneity and clonal evolution in liver cancer," *Nature Communications*, vol. 11, no. 1, pp. 1–15, 2020.
- [13] J. D. Yang, F. Ahmed, K. C. Mara et al., "Diabetes is associated with increased risk of hepatocellular carcinoma in patients with cirrhosis from nonalcoholic fatty liver disease," *Hepatology*, vol. 71, no. 3, pp. 907–916, 2020.
- [14] M. S. D. Dr and S. Vijayarani1, "Liver disease prediction using SVM and Naïve Bayes algorithms," *International Journal of Engineering Sciences & Research*, vol. 4, no. 4, pp. 816–820, 2015.
- [15] M. Hassoon, M. S. Kouhi, M. Zomorodi-Moghadam, and M. Abdar, "Rule optimization of boosted C5.0 classification using genetic algorithm for liver disease prediction," in *Proceedings of the International Conference on Computer and Applications (ICCA)*, pp. 299–305, Doha, UA, September 2017.
- [16] S. N. N. Alfisahrin and T. Mantoro, "Data mining techniques for optimization of liver disease classification," in *Proceedings of the 2013 International Conference on Advanced Computer Science Applications and Technologies*, pp. 379–384, Kuching, Malaysia, December 2013.
- [17] Y. Hayashi and K. Fukunaga, "Accuracy of rule extraction using a recursive-rule extraction algorithm with continuous attributes combined with a sampling selection technique for the diagnosis of liver disease," *Informatics in Medicine Unlocked*, vol. 5, pp. 26–38, 2016.
- [18] S. Sankaranarayanan and T. P. Perumal, "A predictive approach for diabetes mellitus disease through data mining technologies," in *Proceedings of the 2014 World Congress on Computing and Communication Technologies*, vol. 1, pp. 231–233, Trichirappalli, India, March 2014.
- [19] X. Zhou, Y. Zhang, M. Shi, H. Shi, and Z. Zheng, "Early detection of liver disease using data visualisation and classification method," *Biomedical Signal Processing and Control*, vol. 11, no. 1, pp. 27–35, 2014.
- [20] K. S. Purushottam, K. Saxena, and R. Sharma, "Diabetes mellitus prediction system evaluation using C4.5 rules and partial tree," in *Proceedings of the 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, pp. 1–6, UP, India, September 2015.
- [21] B. Padmapriya and T. Velmurugan, "A survey on breast cancer analysis using data mining techniques," in *Proceedings of the 2014 IEEE International Conference on Computational Intelligence and Computing Research*, Coimbatore, India, December 2014.
- [22] A. Alsaeedi and M. Z. Khan, *Software defect prediction using supervised machine learning and ensemble techniques: a comparative study*, *Journal of Software Engineering and Applications*, vol. 12, no. 5, pp. 85–100, 2019.
- [23] K. Balasaravanan and M. Prakash, "Detection of dengue disease using artificial neural network based classification technique," *International Journal of Engineering & Technology*, vol. 7, no. 1.3, pp. 13–15, 2018.
- [24] 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.
- [25] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An empirical evaluation of machine learning techniques for chronic kidney disease prophecy," *IEEE Access*, vol. 8, pp. 55012–55022, 2020.
- [26] A. Amini, O. Varsaneux, H. Kelly et al., "Diagnostic accuracy of tests to detect hepatitis B surface antigen: a systematic review of the literature and meta-analysis," *BMC Infectious Diseases*, vol. 17, no. 1, 2017.
- [27] C. Davi, A. Pastor, T. Oliveira et al., "Severe dengue prognosis using human genome data and machine learning," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2861–2868, 2019.
- [28] H.-H. Rau, C.-Y. Hsu, Y.-A. Lin et al., "Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network," *Computer Methods and Programs in Biomedicine*, vol. 125, pp. 58–65, 2016.
- [29] A. Iqbal, S. Aftab, U. Ali et al., "Performance analysis of machine learning techniques on software defect prediction using NASA datasets," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 300–308, 2019.
- [30] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of roman-Urdu opinions using Naïve bayesian, decision tree and KNN classification techniques," *Journal of King Saud University-Computer and Information Sciences*, vol. 28, no. 3, pp. 330–344, 2016.
- [31] H. Jin, S. Kim, and J. Kim, "Decision factors on effective liver patient data prediction," *International Journal of Bio-Science and Bio-Technology*, vol. 6, no. 4, pp. 167–178, 2014.
- [32] T. Menzies, A. Dekhtyar, J. Distefano, and J. Greenwald, "Problems with precision: a response to 'comments on 'data mining static code attributes to learn defect predictors,'" *IEEE Transactions on Software Engineering*, vol. 33, no. 9, pp. 637–640, 2007.
- [33] J. Li, P. He, J. Zhu, and M. R. Lyu, "Software defect prediction via convolutional neural network," in *Proceedings of the 2017 IEEE International Conference on Software Quality, Reliability and Security (QRS)*, pp. 318–328, Prague, Czech Republic, Europe, July 2017.
- [34] J. Chen, Y. Yang, K. Hu, Q. Xuan, Y. Liu, and C. Yang, "Multiview transfer learning for software defect prediction," *IEEE Access*, vol. 7, pp. 8901–8916, 2019.
- [35] Q. Song, Y. Guo, and M. Shepperd, "A comprehensive investigation of the role of imbalanced learning for software defect prediction," *IEEE Transactions on Software Engineering*, vol. 45, no. 12, pp. 1253–1269, 2019.
- [36] H. Tong, B. Liu, and S. Wang, "Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning," *Information and Software Technology*, vol. 96, pp. 94–111, 2018.
- [37] M. M. Saritas, "Performance analysis of ANN and naive Bayes classification algorithm for data classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, 2019.
- [38] C. Wu, S.-C. Kao, C.-H. Shih, and M.-H. Kan, "Open data mining for Taiwan's dengue epidemic," *Acta Tropica*, vol. 183, pp. 1–7, 2018.
- [39] P. Guo, T. Liu, Q. Zhang et al., "Developing a dengue forecast model using machine learning: a case study in China," *PLoS Neglected Tropical Diseases*, vol. 11, no. 10, Article ID e0005973, 2017.
- [40] T. C.-F. Yip, A. J. Ma, V. W.-S. Wong et al., "Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general

- population,” *Alimentary Pharmacology & Therapeutics*, vol. 46, no. 4, pp. 447–456, 2017.
- [41] S. Khan, R. Ullah, A. Khan, N. Wahab, M. Bilal, and M. Ahmed, “Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM),” *Biomedical Optics Express*, vol. 7, no. 6, p. 2249, 2016.
- [42] D. A. E. H. Omran, A. H. Awad, M. A. E. R. Mabrouk, A. F. Soliman, and A. O. A. Aziz, “Application of data mining techniques to explore predictors of HCC in Egyptian patients with HCV-related chronic liver disease,” *Asian Pacific Journal of Cancer Prevention*, vol. 16, no. 1, pp. 381–385, 2015.
- [43] S. Picek, A. Heuser, and S. Guillely, A. Heuser and S. Guillely, “Template attack versus Bayes classifier Picek,” *Journal of Cryptographic Engineering*, vol. 7, no. 4, pp. 343–351, 2017.
- [44] A. Naik and L. Samant, “Correlation review of classification algorithm using data mining tool: WEKA, rapidminer, tanagra, orange and knime,” *Procedia Computer Science*, vol. 85, pp. 662–668, 2016.
- [45] T. R. Baitharu and S. K. Pani, “Analysis of data mining techniques for healthcare decision support system using liver disorder dataset,” *Procedia Computer Science*, vol. 85, pp. 862–870, 2016.
- [46] K. A. Otunaiya and G. Muhammad, “Performance of data-mining techniques in the prediction of chronic kidney disease,” *Computer Science and Information Technology*, vol. 7, no. 2, pp. 48–53, 2019.
- [47] S. Chatterjee, N. Dey, F. Shi, A. S. Ashour, S. J. Fong, and S. Sen, “Clinical application of modified bag-of-features coupled with hybrid neural-based classifier in dengue fever classification using gene expression data,” *Medical & Biological Engineering & Computing*, vol. 56, no. 4, pp. 709–720, 2018.
- [48] A. B. Nassif, D. Ho, and L. F. Capretz, “Towards an early software estimation using log-linear regression and a multi-layer perceptron model,” *Journal of Systems and Software*, vol. 86, no. 1, pp. 144–160, 2013.
- [49] E. K. Hashi, M. S. U. Zaman, and M. R. Hasan, “An expert clinical decision support system to predict disease using classification techniques,” in *Proceedings of the 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 396–400, Kolatoli, Bangladesh, September 2017.
- [50] L. Wilkinson, A. Anand, D. N. Tuan, and C. H. I. R. P., “Chirp,” in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*, pp. 6–14, San Diego California USA, August 2011.
- [51] K. L. Bouman, M. D. Johnson, D. Zoran, V. L. Fish, S. S. Doelman, and W. T. Freeman, “Computational imaging for VLBI image reconstruction,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2016, 922 pages, 2016.
- [52] C. J. Mantas and J. Abellán, “Credal decision trees to classify noisy data sets,” in *Proceedings of the 9th International Conference on Hybrid Artificial Intelligence Systems*, vol. 8480, pp. 683–688, Salamanca, Spain, June 2014.
- [53] Q. He, Z. Xu, S. Li et al., “Novel entropy and rotation forest-based credal decision tree classifier for landslide susceptibility modeling,” *Entropy*, vol. 21, no. 2, p. 106, 2019.
- [54] J. Abellán and A. R. Masegosa, “An ensemble method using credal decision trees,” *European Journal of Operational Research*, vol. 205, no. 1, pp. 218–226, 2010.
- [55] M. N. Adnan, M. Z. Islam, and P. T. U. S. C. R., “Forest PA: constructing a decision forest by penalizing attributes used in previous trees,” *Expert Systems with Applications*, vol. 89, p. 389, 2017.
- [56] A. Gulia, R. Vohra, and P. Rani, “Liver patient classification using intelligent techniques,” vol. 5, no. 4, pp. 5110–5115, 2014.
- [57] T. K. Ho, “Random decision forests,” in *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, pp. 278–282, Barcelona, Spain, May 1995.
- [58] G. Louppe, *Understanding Random Forests: From Theory to Practice*, <https://arxiv.org/abs/1407.7502>, 2014.
- [59] J. Nayak, B. Naik, and H. S. Behera, “A Comprehensive Survey on Support Vector Machine in Data, Mining Tasks: Applications & Challenges,” *International Journal of Database Theory and Application*, vol. 8, no. 1, 2015.

Research Article

IoMT-Based Automated Detection and Classification of Leukemia Using Deep Learning

Nighat Bibi ¹, Misba Sikandar ¹, Ikram Ud Din ¹, Ahmad Almogren ²,
and Sikandar Ali ³

¹Department of Information Technology, The University of Haripur, Haripur 22620, Pakistan

²Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11633, Saudi Arabia

³Beijing Key Laboratory of Petroleum Data Mining, China University of Petroleum-Beijing, Beijing 102249, China

Correspondence should be addressed to Sikandar Ali; sikandar@cup.edu.cn

Received 22 October 2020; Revised 2 November 2020; Accepted 21 November 2020; Published 4 December 2020

Academic Editor: Shah Nazir

Copyright © 2020 Nighat Bibi 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.

For the last few years, computer-aided diagnosis (CAD) has been increasing rapidly. Numerous machine learning algorithms have been developed to identify different diseases, e.g., leukemia. Leukemia is a white blood cells- (WBC-) related illness affecting the bone marrow and/or blood. A quick, safe, and accurate early-stage diagnosis of leukemia plays a key role in curing and saving patients' lives. Based on developments, leukemia consists of two primary forms, i.e., acute and chronic leukemia. Each form can be subcategorized as myeloid and lymphoid. There are, therefore, four leukemia subtypes. Various approaches have been developed to identify leukemia with respect to its subtypes. However, in terms of effectiveness, learning process, and performance, these methods require improvements. This study provides an Internet of Medical Things- (IoMT-) based framework to enhance and provide a quick and safe identification of leukemia. In the proposed IoMT system, with the help of cloud computing, clinical gadgets are linked to network resources. The system allows real-time coordination for testing, diagnosis, and treatment of leukemia among patients and healthcare professionals, which may save both time and efforts of patients and clinicians. Moreover, the presented framework is also helpful for resolving the problems of patients with critical condition in pandemics such as COVID-19. The methods used for the identification of leukemia subtypes in the suggested framework are Dense Convolutional Neural Network (DenseNet-121) and Residual Convolutional Neural Network (ResNet-34). Two publicly available datasets for leukemia, i.e., ALL-IDB and ASH image bank, are used in this study. The results demonstrated that the suggested models supersede the other well-known machine learning algorithms used for healthy-versus-leukemia-subtypes identification.

1. Introduction

Internet of Things (IoT) is deployed in several areas, such as vehicular communications [1, 2], smart cities [3], cloud computing [4, 5], smart ecosystems [6, 7], smart campus [8], mobile communications [9], smart agriculture [10], and Healthcare or Internet of Medical Things (IoMT) [11]. However, a large section of the research community is attracted by IoMT or simply Healthcare IoT. IoMT [12, 13] is indeed a set of WiFi smart medical gadgets and smart applications connected to IT health systems through computer networks [14–16]. Smart medical devices are equipped with sensors or other computing resources and are exclusively intended for healthcare in the body at home,

clinic, hospital, and community [17, 18]. These smart devices are linked to the cloud platforms to analyze collected data for further processing [19]. The IoMT technology includes virtual care for patients with long-term illnesses, portable mHealth devices for patients, monitoring of patients' medication, tracking the location of hospitalized patients [20, 21], and the ability to provide information to caregivers [22, 23]. The IoMT technology saves time and efforts of patients and doctors [11]. Connecting patients to their doctors and enabling the transfer of medical data over a secure network reduce the burden on health systems [24]. A rapid increase in the development and use of IoMT opens the door to deploying such frameworks that can fastly, securely, and accurately examine the patients' health and

diagnose and cure different diseases remotely [25]. IoMT-based frameworks are abundant, particularly for those diseases which are more crucial with respect to patients' life, such as leukemia.

1.1. Leukemia. Leukemia is a disease related to white blood cells (WBC). Platelets, red blood cells (RBC), and WBC are various components of blood. Platelets help to clot and control bleeding. RBC known as erythrocytes are responsible for the transfer of oxygen through lungs to the body tissues. While WBC, also known as leukocytes, are responsible for fighting against diseases and infections. Leukemia refers to the production of large numbers of immature WBC. It is a type of cancer that affects the bone marrow and blood while destroying the immune system of a human body. Two main categories of leukemia based on progress are acute and chronic leukemia. Infected WBC grow rapidly in acute leukemia and do not perform in a normal way, whereas in chronic leukemia, WBC can act normally and grow less quickly. However, this can be severe since it may not be distinguished easily from the normal WBC. In addition, two types of acute and chronic leukemia are lymphoid and myeloid leukemia based on the size and shape of WBC, where both of them can be further divided into two subtypes each, i.e., acute lymphocytic leukemia (ALL), chronic lymphocytic leukemia (CLL), acute myeloid leukemia (AML), and chronic myeloid leukemia (CML).

1.1.1. Acute Lymphocytic Leukemia. ALL is mostly seen in children, which is a WBC cancer caused by the constant multiplication and overproduction of immature WBC in the bone marrow. The symptoms of ALL are quite similar to flu and other common diseases, such as exhaustion, weakness, and pain in joints and bones, making it very difficult to diagnose this disease. Three types of ALL are classified as L1, L2, and L3 [26].

1.1.2. Acute Myeloid Leukemia. The most common type of acute leukemia is AML, which happens when the bone marrow starts producing blasts and immature WBC. It may also create RBC and platelets that are abnormal. The common symptoms of early-stage AML may be similar to those of influenza or other common illnesses. Based on the types of blood cell affected, signs and symptoms may vary. The signs of AML are fever, bone pain, tiredness and fatigue, shortness of breath, pale skin, frequent infections, easy bruising, and abnormal bleeding, such as frequent nosebleeds and gum bleeding. AML has eight different subtypes that differentiate it from the other types of leukemia [27].

1.1.3. Chronic Lymphoblastic Leukemia. CLL is a hematological sickness that gets worse slowly. It is commonly observed in adults and is very uncommon in children. The

symptoms of CLL include loss in weight, fever, night sweats, and periodic infections.

1.1.4. Chronic Myeloid Leukemia. CML, also known as chronic myelogenous leukemia, is a form of slow growing leukemia, but it can develop into acute leukemia, which is fast growing and difficult to treat. This may be viewed in three stages, i.e., chronic, accelerated, and blast stages. In the chronic stage, the leukemia is inside the strongest situation and grows slowly. Within the second stage, the blood cells are immature, usually referred to as extended stage. The third stage is the blast stage, which is also known as the acute stage or blast transformation stage. The pictorial representation of blood structure and leukemia types is shown in Figure 1.

It is necessary for hematologists to identify the existence of leukemia along with its particular form in order to avoid medical risks and to determine the correct leukemia treatment. A crucial and time-consuming step is the identification of leukemia through optical blood smears examination monitored by a specialist. To solve such problems, many CAD methodologies for quantitative analysis of the peripheral blood samples have been developed using machine learning and deep learning methods. However, these approaches have some drawbacks and need improvements in terms of accuracy, learning process, and efficiency.

Thus, to tackle these drawbacks and by keeping in view the vitality of healthcare, an IoMT-based framework for the automatic identification of leukemia subtypes is proposed in this study. In the proposed framework, IoT-enabled microscope uploads the blood smear images to the leukemia cloud. The leukemia with respect to its type(s) is diagnosed by using the deep learning models, i.e., ResNet-34 [28] or DenseNet-121 [29]. Deep learning is a branch of machine learning used to solve a variety of problems and describes the abstract concepts through different layers of data processing to discover better learning algorithms and representations that are less dependent on feature engineering [30]. The high prediction power of deep learning algorithms and surprising ability of features extraction extend their use to a wide range of research areas. Therefore, in this study, deep transfer learning-based methods are proposed to identify microscopic blood images as healthy, ALL, AML, CLL, and CML. The detail of the models is given in Section 3. The diagnostic results of leukemia with respect to its types predicted by the suggested models can be shown on the clinician's computer and accordingly, medical care may be offered to leukemia patients. The proposed framework is demonstrated in Figure 2.

The suggested framework is also helpful for patients and doctors in pandemics such as coronavirus disease 2019 (COVID-19). Due to the spread of COVID-19, most of the countries imposed sudden lockdown in major cities; as a result, almost ten billion citizens were quarantined. During this pandemic, people put their focus on accumulating more necessary care. However, some patients with chronic diseases leave their homes for medical help opening a gap in

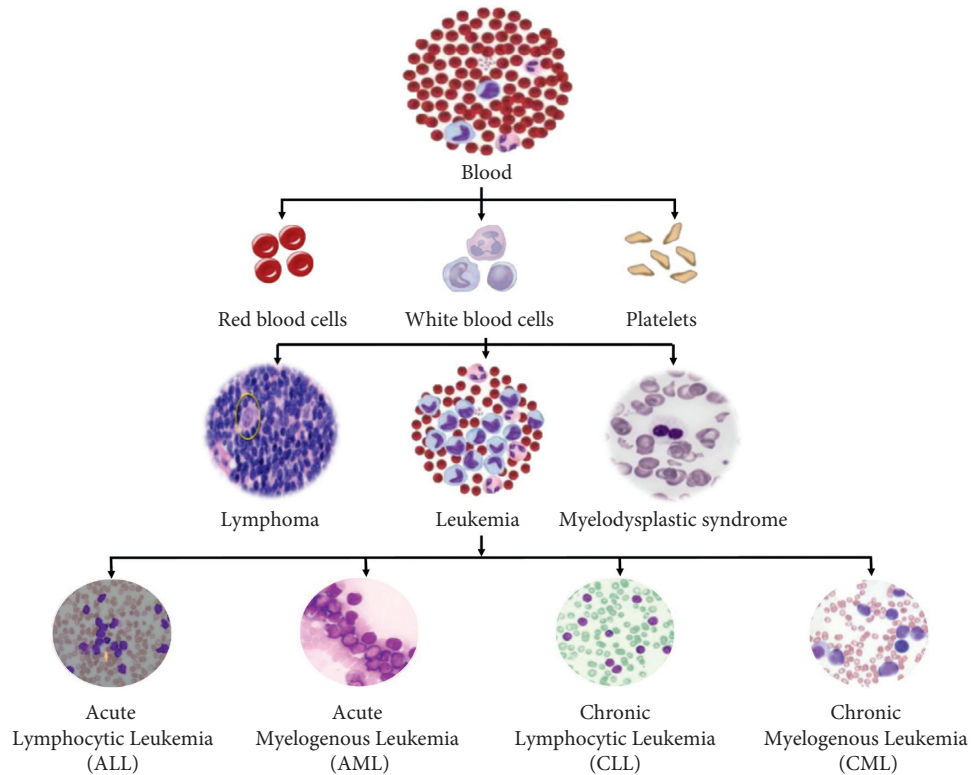


FIGURE 1: Blood and leukemia types.

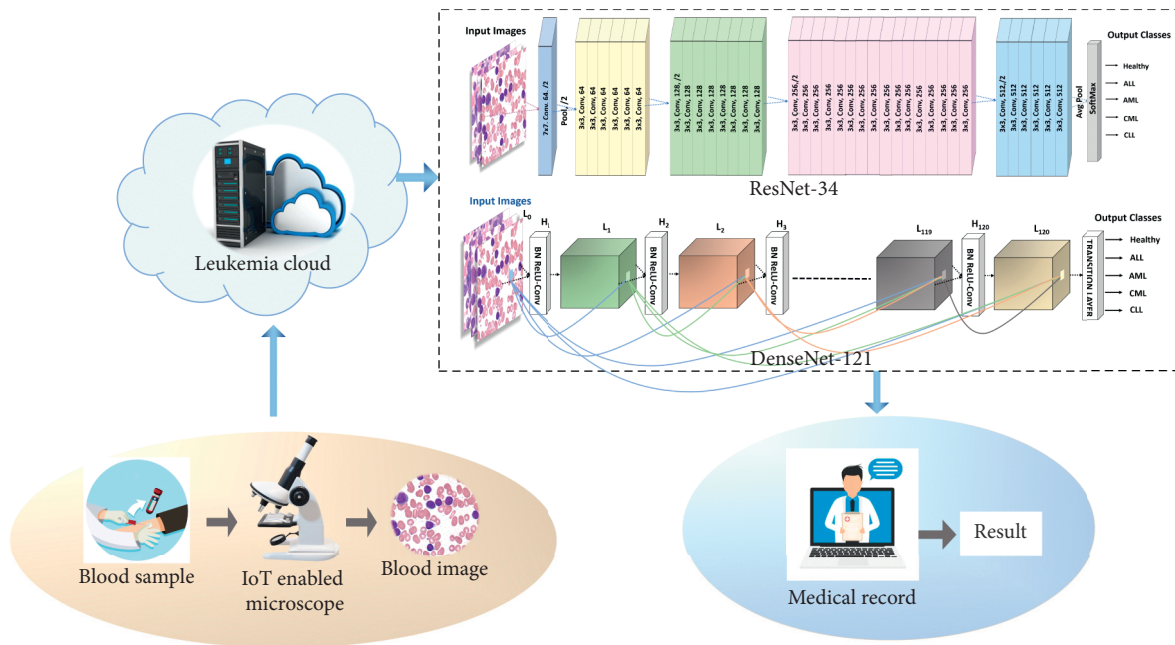


FIGURE 2: Proposed framework for automated leukemia diagnosis.

quarantine measures, which is a threat to disease control. Hence, if the proposed framework is applied to the current crisis, it would provide a medical platform that may help patients to receive adequate medical care at homes. The proposed approach is fast and accurate in addition to

reducing the need for an expert oncologist to diagnose leukemia or any of its subtypes.

We delineate our experiments in the following sections. The related work is presented in Section 2. Section 3 elaborates the data exploration, augmentation, and the proposed

models. Section 4 describes the experimental results, while Section 5 concludes the conducted study and presents future research directions.

2. Literature Survey

For the computer-aided diagnosis (CAD) of leukemia, various studies have been conducted by the IoMT research community. These studies present different machine and deep learning algorithms for the identification of leukemia. In [31], the classification and detection of WBC cancer and some of its subtypes are done by using Random Forest with 94.3% accuracy. In [32], the proposed model detected ALL using KNN and Naive Bayes Classifier with 92.8% accuracy. The classifier is tested on 60 sample images. In [33], a novel Principal Component Analysis (PCA) based on ABC-BPNN scheme is suggested for the classification of leukemia cells and attained 98.72% average accuracy with reduction in the computation time. In [34], the ALL is identified. Firstly, leukemia images are segmented by BSA-based clustering. Secondly, Jaya technique is applied in integration with some standard classification methods such as Naive Bayes, K-nearest neighbor, linear discriminant analysis, support vector machine (SVM), ensemble random under-sampling boost, and decision tree. However, Jaya with SVM and Jaya with decision tree gave better accuracy.

The Linear Discriminant Analysis- (LDA-) based PCA model for diagnosing leukemia by utilizing Discrete Orthogonal Stockwell Transform (DOST) method to extract the features of microscopic images is presented in [35]. In [36], for feature extraction, three pre-trained CNN architectures are applied. However, for the leukemia classification on the hybrid database, SVM without segmentation is used. In [37], for the identification of acute leukemia from microscopic images initially to extract features, the images were segmented by using unsharp masking sub-imaging bounding box and $L \times a \times b$ color Fuzzy C-Means Clustering after applying the Nearest Neighbor. Then, to identify ALL, the extracted features were passed to SVM, which attained 95% accuracy. In [38], a robust feature extraction and selection technique for the identification of lymphocytes versus ALL is proposed. The classification is done by KNN using Euclidean Distance with 92.5% accuracy. In [39], the classification is done using CNN, where only ALL vs healthy samples are classified with 88.25% accuracy and leukemia subtypes are classified with 81.74% accuracy.

In [35, 40–43], ALL-IDB dataset is used to detect ALL. In [40], K-medoids is presented with 98.60% accuracy. In [35], DOST, PCA, and LDA are proposed with 99.66% accuracy. In [41], Generative Adversarial Optimization- (GAO-) based method is investigated with 93.84% accuracy. In [42], Genetic Algorithm (GA) and Artificial Neural Network (ANN) are presented with 97.07% accuracy. In [43], Chronological Sine Cosine Algorithm (SCA) is tested, which achieved 98.70% accuracy. In [44], ASH image bank and ALL-IDB1 datasets are utilized for classifying lymphoblast cells. Convnet is investigated and achieved 81.74% accuracy. In [36], heterogeneous database ALL-IDB1 and ALL-IDB2 datasets

are used. The diagnosis of leukemia (pathological or non-pathological) is done using pre-trained CNN with SVM, which attained 99% accuracy. In [45], the ALL-IDB1 dataset is employed where the diagnosis of leukemia (normal vs abnormal) is performed by CNN and achieved 96.60% accuracy.

In [46], the ALL-IDB1 and ALL-IDB2 datasets are employed for the ALL detection. However, the scheme used is SVM, which attained 89.81% accuracy. In [47], the ALL-IDB2 dataset is used for the identification of ALL where the method investigated is customized KNN with 96.25% accuracy. In [48], by utilizing ALL-IDB1 dataset, ALL is classified on the basis of cell energy features by employing SVM, which attained 94.00% accuracy. In [49], by utilizing ASH image bank dataset, firstly FAB ALL subtypes are identified by applying GA with multilayer perceptron kernel (MLP) function and acquired 97.1% accuracy. Secondly, FAB AML subtypes are identified by employing genetic phenomena with Gaussian radial basis kernel function and achieved 98.5% accuracy. Finally, healthy, ALL, and AML are identified by using a GA with Gaussian Radial Basis kernel and achieved 99.50% accuracy.

It is exhibited from the literature that almost all previous approaches identified leukemia with respect to healthy, AML or ALL types. However, these approaches did not address the problem of identifying leukemia with respect to all its subtypes, i.e., ALL, AML, CLL, and CML. Therefore, in this study, the deep CNN-based approaches are presented to classify leukemia in terms of all its types.

The proposed work is an advancement of the study conducted in [39], where the authors have done the classification of ALL vs. healthy using a simple CNN architecture. However, our proposed work focuses on the classification of four subtypes of leukemia, i.e., ALL, AML, CLL, and CML, using advanced CNN architectures—Residual Network-34 (ResNet-34) [28] and Dense Convolutional Network-121 (DenseNet-121) [29]. Hence, it is elaborated in the result section that the proposed deep CNNs, i.e., DenseNet-121 and ResNet-34, outperform the existing schemes with 99.91% and 99.96% accuracy, respectively. Furthermore, for investigating the proposed methods, ALL-IDB and ASH image bank datasets are utilized.

3. Experimental Setup

This section delineates the experimental setup for conducting the study. Firstly, the dataset retrieval is described followed by data augmentation, while in the end, the deep learning models used to classify the subtypes of leukemia are explained. The proposed methodology is presented in Figure 3.

3.1. Leukemia Database Description. The dataset is collected from two different sources: ASH image bank [50] and ALL-IDB [51, 52]. The ASH image bank is freely accessible on the Internet and contains a complete bank of images on a variety

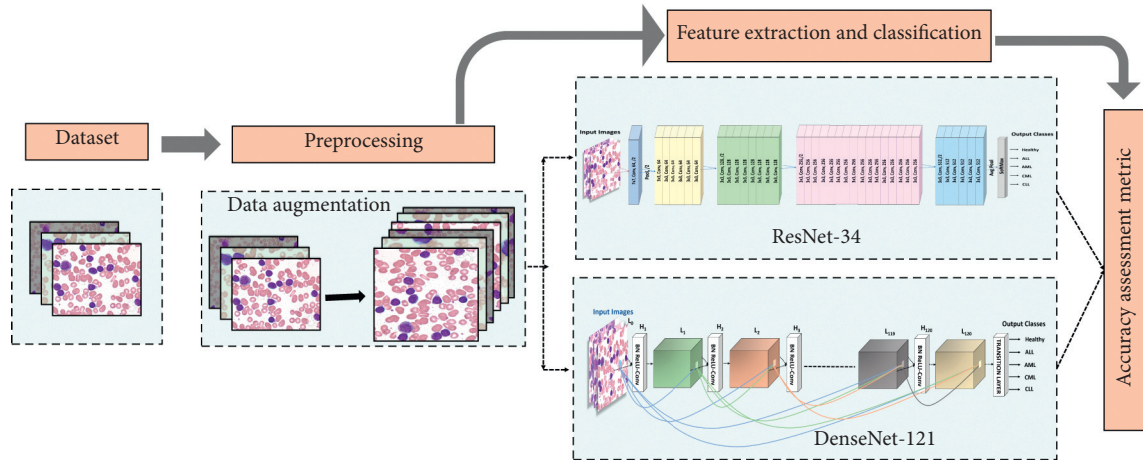


FIGURE 3: Proposed methodology.

of hematological topics. In this article, all accessible annotated cell images with leukemia of blood are selected including any one of the four subtypes.

The ALL-IDB dataset provides annotated microscopic images of blood cells that were developed for segmentation, evaluation, and classification. ALL-IDB contains only healthy and ALL types of leukemia samples. The remaining mentioned leukemia subtypes are not available in the ALL-IDB dataset. The ALL-IDB is considered more reliable because experienced oncologists provided the ALL classification for each image in the dataset. Some samples of images from datasets are shown in Figure 4.

3.2. Reducing Overfitting. Data augmentation methods are commonly used for maximizing the size of datasets and reducing overfitting, especially in the deep learning models. Various image transformation schemes, such as rotation, flipping, and shifting, have been used to obtain distinct images from the original image. They would have more generalization capabilities if machine learning and deep learning models were trained with the original images along with their sub-versions. In different image segmentation studies with CNNs, numerous methods of information replication have reduced the model error rate by giving better speculation. Here, a better assortment of microscopic blood cell images is provided by the retrieved datasets, but both datasets contain samples of limited numbers of leukemia subtypes. As the sample datasets are very limited for deep learning techniques, it can lead to overfitting. Thus, image transformation or data augmentation techniques are used to maximize the dataset; i.e., rotation, height shift, width shift, horizontal flip, zoom, and shearing, are applied. The sample size after applying the methods of image transformation is increased in both datasets. Exact number of samples per leukemia subtype before and after augmentation is shown in Table 1. After applying the data augmentation, the sample size reached 3277 and 2359 in the ASH image bank and ALL-IDB datasets, respectively.

3.3. Deep Learning Models. After performing augmentation on the dataset, it is passed on to CNNs—the deep learning models. Instead of using traditional machine learning techniques which are provided with hand-crafted features to learn and require time and effort, in this study, CNNs are utilized. CNNs have the ability to learn automatically from raw data. A typical CNN begins with convolutional and pooling layers and ends with a fully connected layer. New models can be created from a CNN by assembling some convolutional, pooling, and dense layers. By arranging some layers of CNNs architectures, new light CNN models have been developed such as VggNet [53] and AlexNet [54]. However, advanced CNN models, such as ResNet [28] and DenseNet [29], are deeper and more complex having the ability to learn better, as elaborated in the result section. Therefore, in the presented work, ResNet-34 and DenseNet-121 models are used in a supervised learning mode to identify and categorize leukemia with respect to its types. The details of each model are described in the following subsections.

3.3.1. ResNet-34. ResNet-34 is a pre-trained 34-layer model. The performance of deep neural networks depends on architecture and dataset. A deep network of CNNs and large dataset produce better performance. However, the performance deteriorates after a certain depth when the network gets deeper. The reason of this problem is the vanishing gradient. The ResNet solves this problem as gradients flow from starting layers to the final ones by skipping some layers. Mathematically, the layers of the ResNet model can be calculated according to

$$Y = f(x) + id(x) = f(x) + x. \quad (1)$$

By skipping the connections between layers, the gradient can easily flow and the training of the layers becomes faster. ResNet-34 consists of a total of 34 layers wherein one is convolutional and pooling layer in addition to four other layers with the same pattern. Each layer is convolved with 3×3 convolution with a feature map of sizes 64, 128, 256,

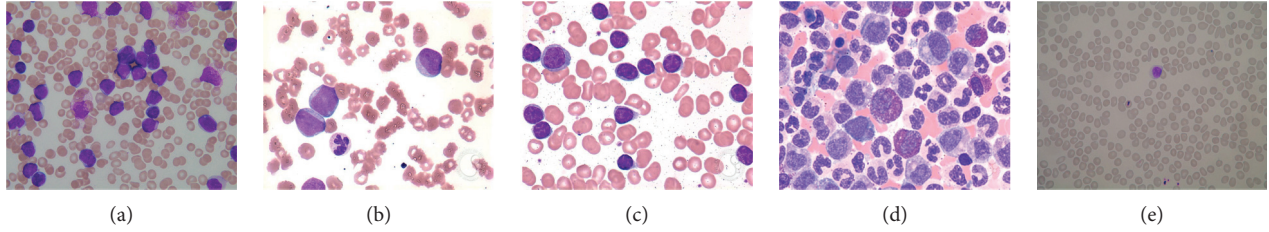


FIGURE 4: Leukemia subtype images. (a) Acute lymphocytic leukemia (ALL). (b) Acute myelogenous leukemia (AML). (c) Chronic lymphocytic leukemia (CLL). (d) Chronic myelogenous leukemia (CML). (e) Healthy.

TABLE 1: Distribution of leukemia subtypes before and after augmentation.

Leukemia type	Dataset	Before augmentation	After augmentation
ALL	ALL-IDB	181	1079
AML	ASH image bank	55	1194
CLL	ASH image bank	38	840
CML	ASH image bank	57	1243
Healthy	ALL-IDB	187	1280

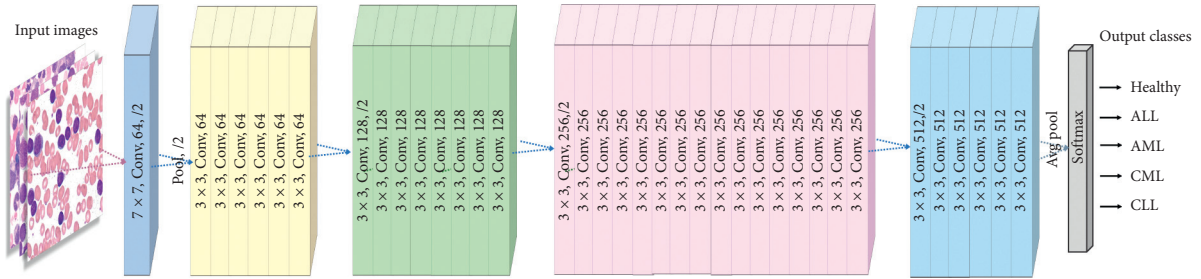


FIGURE 5: General architecture of ResNet-34 (adapted from [28]).

and 512, respectively [28]. The general architecture of ResNet-34 is presented in Figure 5.

3.3.2. *DenseNet-121*. Dense convolutional networks or DenseNet achieved the best classification results on CIFAR-10 and ImageNet datasets [29]. Dense connections are used in the DenseNet architecture, such as ResNet architecture. DenseNet-121 consists of 121 layers. In DenseNet architecture, each layer is connected to all subsequent layers. Thus, each layer receives important features learned by any preceding layers of the network that makes training of the network more efficient [55]. The DenseNet architecture uses fewer parameters than ResNet for the training of the network. Small datasets make the model overfit, while the dense connection solves that overfitting problem [29]. A significant part of DenseNet is a dense block, which is used for enhancing the information flow between layers. It consists of BN, ReLU, and 3×3 conv. The particular formula for the dense block is provided in

$$L_l = H_l([L_0, L_1, \dots, L_{l-1}]), \quad (2)$$

where $(0, 1, \dots, l-1)$ represents the layers of DenseNet-121 and $[L_0, L_1, \dots, L_{l-1}]$ is the concatenation of feature map obtained from the layers of DenseNet-121. The composite function of BN, ReLU, and 3×3 conv. operations on the l^{th}

layer is presented by $H_l(\cdot)$ [29]. The DenseNet architecture is presented in Figure 6.

Both pre-trained models are implemented with *Python* using open source fastai [56] and the deep learning library, which makes the implementation of the models simpler. All the experiments are performed on Google Colab [57].

4. Experimental Results

In order to evaluate the proposed models, the performance measures used are precision, recall, F1 score, and accuracy. The mathematical description of each of these parameters is given in Table 2. Here, TP is a true positive rate and refers to positive class determined as positive; FP is false positive rate and refers to negative class determined as positive; TN is true negative rate and refers to negative class determined as negative; and FN is false negative rate and refers to positive class determined as negative [58]. However, Recall is the accuracy of prediction for the known leukemia subtype class. Accuracy is the prediction for the known leukemia and non-leukemia subtype classes [59]. Precision is the ratio of correct positive predicted leukemia subtype classes to the predicted positive leukemia subtype classes, while F1 score is the harmonic mean of precision and recall [23].

To show the efficiency of the utilized methods, various parameters such as precision, recall, F1 score, and accuracy

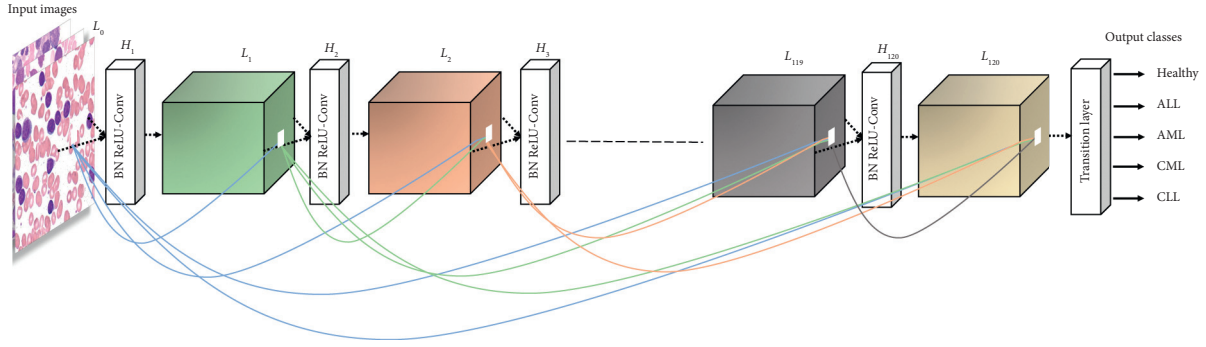


FIGURE 6: General architecture of DenseNet-121 (adapted from [29]).

TABLE 2: Performance measures mathematical description.

Measure	Derivations
Accuracy	$ACC = (TP + TN)/(P + N)$
Precision	$PPV = TP/(TP + FP)$
Recall	$TPR = TP/(TP + FN)$
F1 score	$F1 = 2TP/(2TP + FP + FN)$

	ALL	AML	CLL	CML	Healthy
ALL	225	0	0	0	0
AML	0	227	0	1	0
CLL	0	2	173	0	0
CML	0	1	1	247	0
Healthy	0	0	0	0	250

FIGURE 7: Confusion matrix of ResNet-34 for leukemia subtype classification.

are used. Figures 7 and 8 show the confusion matrix for the classification of subtypes of leukemia using ResNet-34 and DenseNet-121, respectively. It is evident from Figures 7 and 8 that the proposed models predicate very well. Furthermore, Tables 3 and 4 present the accuracy, precision, recall, and F1 score for each type of leukemia on the benchmark dataset, ALL-IDB and ASH image bank.

It is exhibited from Tables 3 and 4 that the ResNet-34 and DenseNet-121 prediction accuracy for ALL and healthy cases is 100%, while precision, recall, and F1 score are also 100%, i.e., 1.0. The prediction accuracy of ResNet-34 for AML is 99.65%, precision is 1.0%, recall is 0.99%, and F1 score is also 0.99%. In case of CLL, the prediction accuracy of ResNet-34 is 99.73%. However, precision, recall, and F1 score are 0.99%. For CML, the prediction accuracy of

ResNet-34 is 99.73%, precision is 0.99%, recall is 1.0%, and F1 score is 0.99%, whereas in case of DenseNet-121, the prediction accuracy for AML is 99.91%. However, precision, recall, and F1 score are 1.0%. In case of CLL, the prediction accuracy of DenseNet-121 is 99.91%, precision is 1.0%, recall is 0.99%, and F1 score is 1.0%. For CML, the DenseNet-121 prediction accuracy, precision, recall, and F1 score are 100%.

Moreover, the training and validation loss for ResNet-34 and DenseNet-121 are shown in Figures 9 and 10, respectively. It is depicted from Figures 9 and 10 that the training and validation loss of ResNet-34 and DenseNet-121 are nearly 0, which means that the accuracy is also near 100%. However, in case of DenseNet-121, the training and validation loss are nearer 0 as compared to ResNet-34. Hence, it is concluded from the results that the prediction performance of ResNet-34 and DenseNet-121 is better for identifying the leukemia subtypes. However, DenseNet-121 seems to supersede ResNet-34. The DenseNet-121 training and validation loss are near 0 when compared to ResNet-34, while for some subtypes identification, i.e., AML, CLL, and CML, DenseNet-121 outperforms ResNet-34. The performance comparison of ResNet-34 and DenseNet-121 with respect to leukemia subtypes identification is shown in Figure 11.

To show their efficiency, a comparison of the proposed models is done with previous approaches, i.e., GA with SVM [49] and CNN [39]. GA with SVM performs the identification of ALL, AML, and healthy samples, while CNN performs the leukemia subtype identification. Nevertheless, it is elucidated in Figure 12 that the utilized models, i.e., ResNet-34 and DenseNet-121, outperform the existing schemes, such as GA with SVM and CNN. It is exhibited from Figure 12 that GA with SVM has 99.50%, CNN has 81.74%, ResNet-34 has 99.56%, and DenseNet-121 has 99.91% accuracy, respectively. Thus, DenseNet-121 supercedes the other approaches. Previously, numerous machine learning techniques were using the same datasets, i.e., ALL-IDB and ASH image bank, for the detection of leukemia and its subtypes. In those studies, the feature extraction and classification methods were used, which require time and effort. However, the proposed models do not need hand-crafted feature set for making predictions and save time and efforts. A thorough comparison of the proposed models with the previous approaches with respect to accuracy is presented in Table 5. It is depicted from Table 5 that the

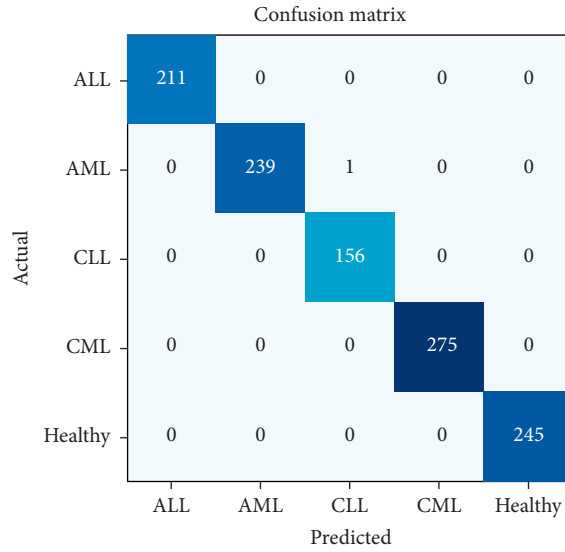


FIGURE 8: Confusion matrix of DenseNet-121 for leukemia subtype classification.

TABLE 3: Performance of the ResNet-34 model for leukemia subtype classification.

Leukemia type	Accuracy	Precision	Recall	F1 score
ALL	100	1.0	1.0	1.0
AML	99.65	1.0	0.99	0.99
CLL	99.73	0.99	0.99	0.99
CML	99.73	0.99	1.0	0.99
Healthy	100	1.0	1.0	1.0

TABLE 4: Performance of the DenseNet-121 model for leukemia subtype classification.

Leukemia type	Accuracy	Precision	Recall	F1 score
ALL	100	1.0	1.0	1.0
AML	99.91	1.0	1.0	1.0
CLL	99.91	1.0	0.99	1.0
CML	100	1.0	1.0	1.0
Healthy	100	1.0	1.0	1.0

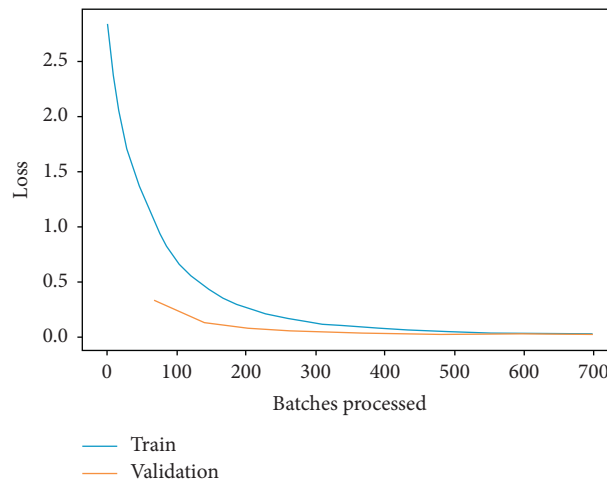


FIGURE 9: Training and validation loss of ResNet-34.

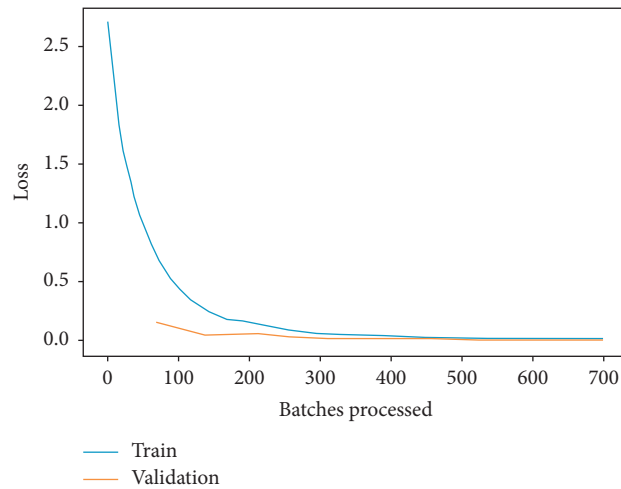


FIGURE 10: Training and validation loss of DenseNet-121.

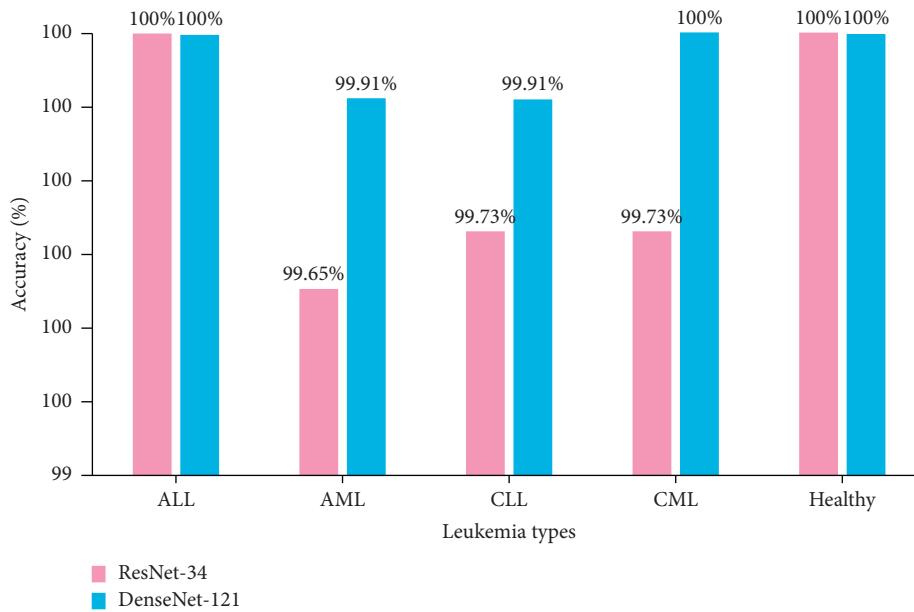


FIGURE 11: Accuracy comparison of the ResNet-34 and DenseNet-121 for leukemia subtype classification.

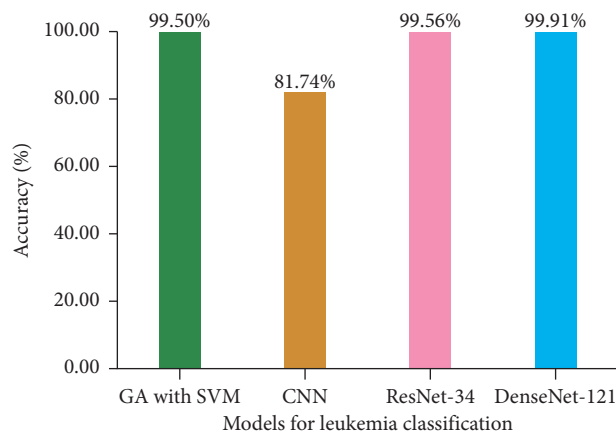


FIGURE 12: Comparison of the studies on automated detection of subtypes of leukemia.

TABLE 5: A comparison of the proposed models with the previous approaches for automated detection of leukemia and its subtypes using the same datasets with respect to average accuracy.

Reference	Dataset	Classification	Classifier	Accuracy (%)
Ahmed et al. [39]	ALL-IDB	Leukemia vs healthy	CNN	88.25
			Naive Bayes	69.69
			Decision tree	62.94
			KNN	58.57
			SVM	50.09
	ALL-IDB, ASH image bank	Leukemia subtypes classification	CNN	81.74
			Naive Bayes	52.68
			Decision tree	45.92
			KNN	43.51
			SVM	20.84
Shafique et al. [26]	ALL-IDB	Acute lymphoblastic leukemia detection	AlexNet	99.50
		Subtypes of acute lymphoblastic leukemia	AlexNet	96.06
Jothi et al. [34]	ALL-IDB	Acute lymphoblastic leukemia detection	Jaya, SVM	99.00
Acharya et al. [40]	ALL-IDB	White blood cells	K-medoids algorithm	98.60
Mishra et al. [35]	ALL-IDB1	Acute lymphoblastic leukemia detection	DOST, PCA, LDA	99.66
Tuba et al. [41]	ALL-IDB2	Acute lymphoblastic leukemia detection	GAO-based methods	93.84
Al-jaboriy et al. [42]	ALL-IDB1	Acute lymphoblastic leukemia detection	GA and ANN	97.07
Jha et al. [43]	ALL-IDB2	Acute lymphoblastic leukemia detection	SCA-based deep CNN	98.70
Pansombut et al. [44]	ASH image bank, ALL-IDB1	Lymphoblast cells	CNN-based convnet	81.74
Vogado et al. [36]	Heterogeneous database ALL-IDB1, ALL-IDB2	Diagnose leukemia (pathological or not)	Pre-trained CNN with SVM	99
Thanh et al. [45]	ALL-IDB1	Diagnose leukemia (normal vs abnormal)	CNN	96.60
Moshavash et al. [46]	ALL-IDB1, ALL-IDB2, Dr. Juan Bruno Zayas Alfonso Hospital, Santiago de Cuba	Acute lymphoblastic leukemia detection	Two ensemble classifiers with SVM	89.81
Umamaheswari et al. [47]	ALL-IDB2	Acute lymphoblastic leukemia	Customized KNN	96.25
Agaian et al. [48]	ALL-IDB1	Acute lymphoblastic leukemia	Cell energy feature with SVM	94.00
Rawat et al. [49]	ASH image bank	ALL		97.10
		AML	GA with SVM	98.50
		Healthy, AML, ALL		99.50
Proposed work	ALL-IDB, ASH image bank	Healthy, ALL, AML, CLL and CML	ResNet-34 DenseNet-121	99.56 99.91

proposed models outperform the previous approaches with average accuracy of 99.56% and 99.91% for ResNet-34 and DenseNet-121, respectively.

5. Conclusion and Research Directions

In this study, an IoMT-based framework is proposed for the leukemia subtypes detection. In the proposed framework, an IoT-enabled microscope uploads the blood smear images to the leukemia cloud. The leukemia is diagnosed by using the ResNet-34 or DenseNet-121 models. It is observed that the diagnosing power of ResNet-34 and DenseNet-121 supersedes all the previous approaches. By using data augmentation techniques, ResNet-34 and DenseNet-121 both process large numbers of image patterns. After diagnosis, the result is sent to the doctor's computer where s/he provides medical care on

the basis of test report through the IoMT framework. Furthermore, the proposed framework facilitates the patients in pandemics such as COVID-19.

In the future, the dataset can be extended by adding new samples of blood images and utilizing new augmentation techniques to achieve better performances. Furthermore, the proposed IoMT-based framework can be equipped with the functionality of diagnosing the subcategories of each leukemia type. Moreover, the proposed models can also be used to find other abnormalities in the blood.

Data Availability

The following two datasets have been used in the research: ASH image bank and ALL-IDB publicly available datasets, which have been cited in the paper at proper positions.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this work.

Acknowledgments

This work was supported in part by King Saud University, Saudi Arabia, through research supporting project number RSP-2020/184, and in part by Fundamental Research Funds for Central Universities (No. 2462020YJRC001).

References

- [1] F. M. Malik, H. A. Khattak, A. Almogren, O. Bouachir, I. U. Din, and A. Altameem, "Performance evaluation of data dissemination protocols for connected autonomous vehicles," *IEEE Access*, vol. 8, pp. 126 896–126 906, 2020.
- [2] K. A. Awan, I. U. Din, A. Almogren, M. Guizani, and S. Khan, "Stabtrust—a stable and centralized trust-based clustering mechanism for iot enabled vehicular ad-hoc networks," *Ieee Access*, vol. 8, pp. 21 159–21 177, 2020.
- [3] H. A. Khattak, K. Tehreem, A. Almogren, Z. Ameer, I. U. Din, and M. Adnan, "Dynamic pricing in industrial internet of things: blockchain application for energy management in smart cities," *Journal of Information Security and Applications*, vol. 55, 2020.
- [4] K. Haseeb, I. U. Din, A. Almogren, Z. Jan, N. Abbas, and M. Adnan, "Ddr-esc: a distributed and data reliability model for mobile edge-based sensor-cloud," *IEEE Access*, vol. 8, pp. 185 752–185 760, 2020.
- [5] K. Haseeb, A. Almogren, I. U. Din, N. Islam, and A. Altameem, "Sasc: secure and authentication-based sensor cloud architecture for intelligent internet of things," *Sensors*, vol. 20, no. 9, p. 2468, 2020.
- [6] W. Ali, I. U. Din, A. Almogren, M. Guizani, and M. Zuair, "A lightweight privacy-aware iot-based metering scheme for smart industrial ecosystems," *IEEE Transactions on Industrial Informatics*, vol. 20, 2020.
- [7] W. Ali, I. U. Din, A. Almogren, and N. Kumar, "Alpha: an anonymous orthogonal code-based privacy preserving scheme for industrial cyber physical systems," *IEEE Transactions on Industrial Informatics*, vol. 20, 2020.
- [8] Z. Ali, M. A. Shah, A. Almogren, I. U. Din, C. Maple, and H. A. Khattak, "Named data networking for efficient iot-based disaster management in a smart campus," *Sustainability*, vol. 12, no. 8, 2020.
- [9] K. Haseeb, I. U. Din, A. Almogren, N. Islam, and A. Altameem, "Rts: a robust and trusted scheme for iot-based mobile wireless mesh networks," *IEEE Access*, vol. 8, 2020.
- [10] K. Haseeb, I. U. Din, A. Almogren, and N. Islam, "An energy efficient and secure iot-based wsn framework: an application to smart agriculture," *Sensors*, vol. 20, no. 7, p. 2081, 2020.
- [11] I. U. Din, A. Almogren, M. Guizani, and M. Zuair, "A decade of internet of things: analysis in the light of healthcare applications," *IEEE Access*, vol. 7, pp. 89 967–89 979, 2019.
- [12] H. A. Al Hamid, S. M. M. Rahman, M. S. Hossain, A. Almogren, and A. Alamri, "A security model for preserving the privacy of medical big data in a healthcare cloud using a fog computing facility with pairing-based cryptography," *IEEE Access*, vol. 5, pp. 22 313–22 328, 2017.
- [13] M. G. R. Alam, M. M. Hassan, M. Z. Uddin, A. Almogren, and G. Fortino, "Autonomic computation offloading in mobile edge for iot applications," *Future Generation Computer Systems*, vol. 90, pp. 149–157, 2019.
- [14] I. U. Din, M. Guizani, S. Hassan et al., "The internet of things: a review of enabled technologies and future challenges," *Ieee Access*, vol. 7, pp. 7606–7640, 2018.
- [15] I. U. Din, M. Guizani, J. J. P. C. Rodrigues, S. Hassan, and V. V. Korotaev, "Machine learning in the internet of things: designed techniques for smart cities," *Future Generation Computer Systems*, vol. 100, pp. 826–843, 2019.
- [16] E. K. Wang, C.-M. Chen, M. M. Hassan, and A. Almogren, "A deep learning based medical image segmentation technique in internet-of-medical-things domain," *Future Generation Computer Systems*, vol. 108, pp. 135–144, 2020.
- [17] S. R. Khan, M. Sikandar, A. Almogren, I. U. Din, A. Guerrieri, and G. Fortino, "Iomt-based computational approach for detecting brain tumor," *Future Generation Computer Systems*, vol. 109, pp. 360–367, 2020.
- [18] N. Islam, Y. Faheem, I. U. Din, M. Talha, M. Guizani, and M. Khalil, "A blockchain-based fog computing framework for activity recognition as an application to e-healthcare services," *Future Generation Computer Systems*, vol. 100, pp. 569–578, 2019.
- [19] K. Janjua, M. A. Shah, A. Almogren, H. A. Khattak, C. Maple, and I. U. Din, "Proactive forensics in iot: privacy-aware log-preservation architecture in fog-enabled-cloud using holochain and containerization technologies," *Electronics*, vol. 9, no. 7, p. 1172, 2020.
- [20] S. U. Khan, N. Islam, Z. Jan, I. U. Din, A. Khan, and Y. Faheem, "An e-health care services framework for the detection and classification of breast cancer in breast cytology images as an iomt application," *Future Generation Computer Systems*, vol. 98, pp. 286–296, 2019.
- [21] K. Haseeb, N. Islam, A. Almogren, and I. U. Din, "Intrusion prevention framework for secure routing in wsn-based mobile internet of things," *Ieee Access*, vol. 7, pp. 185 496–185 505, 2019.
- [22] S. Khan, N. Islam, Z. Jan, I. U. Din, and J. J. P. C. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognition Letters*, vol. 125, pp. 1–6, 2019.
- [23] M. Sikandar, W. Anwar, A. Almogren, I. U. Din, and N. Guizani, "Iomt-based association rule mining for the prediction of human protein complexes," *IEEE Access*, vol. 8, pp. 6226–6237, 2020.
- [24] K. A. Awan, I. U. Din, A. Almogren, H. Almajed, I. Mohiuddin, and M. Guizani, "Neurotrust-artificial neural network-based intelligent trust management mechanism for large-scale internet of medical things," *IEEE Internet of Things Journal*, vol. 8, 2020.
- [25] A. Almogren, I. Mohiuddin, I. U. Din, and H. Al Majed, "Ftm-iomt: Fuzzy-based trust management for preventing sybil attacks in internet of medical things," *IEEE Internet of Things Journal*, vol. 8, 2020.
- [26] S. Shafique and S. Tehsin, "Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks," *Technology in Cancer Research & Treatment*, vol. 17, 2018.
- [27] <https://www.cancercenter.com/cancer-types/leukemia/types/acute-myeloid-leukemia>.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016.
- [29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2017.

- [30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [31] H. Mohamed, R. Omar, N. Saeed et al., "Automated detection of white blood cells cancer diseases," 2018.
- [32] S. Kumar, S. Mishra, and P. Asthana, "Automated detection of acute leukemia using k-mean clustering algorithm," *Advances in Computer and Computational Sciences*, vol. 554, pp. 655–670, 2018.
- [33] R. Sharma and R. Kumar, "A novel approach for the classification of leukemia using artificial bee colony optimization technique and back-propagation neural networks: ICCN 2018," 2019.
- [34] J. Ganesan, H. H. Inbarani, A. T. Azar, and K. R. Devi, "Rough set theory with jaya optimization for acute lymphoblastic leukemia classification," *Neural Computing and Applications*, vol. 55, pp. 1–20, 2018.
- [35] S. Mishra, B. Majhi, and P. K. Sa, "Texture feature based classification on microscopic blood smear for acute lymphoblastic leukemia detection," *Biomedical Signal Processing and Control*, vol. 47, pp. 303–311, 2019.
- [36] L. H. S. Vogado, R. M. S. Veras, F. H. D. Araujo, R. R. V. Silva, and K. R. T. Aires, "Leukemia diagnosis in blood slides using transfer learning in cnns and svm for classification," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 415–422, 2018.
- [37] D. patra and S. satpathi, "Image analysis of blood microscopic images for acute leukemia detection," 2010.
- [38] S. A. kareem and H. ariffin, "A robust feature extraction and selection method for the recognition of lymphocytes versus acute lymphoblastic leukemia," 2012.
- [39] N. Ahmed, A. Yigit, Z. Isik, and A. Alpkocak, "Identification of leukemia subtypes from microscopic images using convolutional neural network," *Diagnostics*, vol. 9, no. 3, p. 104, 2019.
- [40] V. Acharya and P. Kumar, "Detection of acute lymphoblastic leukemia using image segmentation and data mining algorithms," *Medical & Biological Engineering & Computing*, vol. 57, p. 6, 2019.
- [41] M. Tuba and E. Tuba, "Generative adversarial optimization (Goa) for acute lymphocytic leukemia detection," *Studies in Informatics and Control*, vol. 28, pp. 245–254, 10 2019.
- [42] S. S. Al-jaboriy, N. N. A. Sjarif, S. Chuprat, and W. M. Abdullah, "Acute lymphoblastic leukemia segmentation using local pixel information," *Pattern Recognition Letters*, vol. 125, pp. 85–90, 2019.
- [43] K. K. Jha and H. S. Dutta, "Mutual information based hybrid model and deep learning for acute lymphocytic leukemia detection in single cell blood smear images," *Computer Methods and Programs in Biomedicine*, vol. 179, 2019.
- [44] T. Pansombut, S. Wikaisuksakul, K. Khongkraphan, and A. Phon-on, "Convolutional neural networks for recognition of lymphoblast cell images," *Computational Intelligence and Neuroscience*, vol. 2019, 2019.
- [45] T. T. P. Thanh, C. Vununu, C. Vununu, S. Atoev, S.-H. Lee, and K.-R. Kwon, "Leukemia blood cell image classification using convolutional neural network," *International Journal of Computer Theory and Engineering*, vol. 10, no. 2, pp. 54–58, 2018.
- [46] Z. Moshavash, H. Danyali, and M. S. Helfroush, "An automatic and robust decision support system for accurate acute leukemia diagnosis from blood microscopic images," *Journal of Digital Imaging*, vol. 31, no. 5, pp. 702–717, 2018.
- [47] D. Umamaheswari and S. Geetha, "A framework for efficient recognition and classification of acute lymphoblastic leukemia with a novel customized-knn classifier," *Journal of Computing and Information Technology*, vol. 26, no. 2, pp. 131–140, 2018.
- [48] S. Agaian, M. Madhukar, and A. T. Chronopoulos, "A new acute leukaemia-automated classification system," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 303–314, 2018.
- [49] J. Rawat, A. Singh, B. Hs, J. Virmani, and J. S. Devgun, "Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia," *Bio-cybernetics and Biomedical Engineering*, vol. 37, no. 4, pp. 637–654, 2017.
- [50] ASH Image Bank, "The american society of hematology," 2017.
- [51] R. D. Labati, V. Piuri, and F. Scotti, "All-idb: the acute lymphoblastic leukemia image database for image processing," 2011.
- [52] ALL-IDB, "Acute lymphoblastic leukemia image database processing," 2017.
- [53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.
- [54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," 2012.
- [55] L. Li, J. Liang, M. Weng, and H. Zhu, "A multiple-feature reuse network to extract buildings from remote sensing imagery," *Remote Sensing*, vol. 10, p. 1350, 2018.
- [56] Fastai, "The fastai deep learning library," 2018.
- [57] Colab, "Google colab," 2018.
- [58] E. Urtnasan, J.-U. Park, and K.-J. Lee, "Multiclass classification of obstructive sleep apnea/hypopnea based on a convolutional neural network from a single-lead electrocardiogram," 2018.
- [59] Z.-C. Li, Y.-H. Lai, L.-L. Chen, X. Zhou, Z. Dai, and X.-Y. Zou, "Identification of human protein complexes from local subgraphs of protein-protein interaction network based on random forest with topological structure features," *Analytica Chimica Acta*, vol. 718, pp. 32–41, 2012.

Research Article

Big Data-Enabled Analysis of DRGs-Based Payment on Stroke Patients in Jiaozuo, China

Dawei Qiao ¹, Yanru Zhang ¹, Ateeq ur Rehman ², and Mohammad R. Khosravi ³

¹School of Medicine, Henan Polytechnic University, Jiaozuo 454000, China

²Computer Science Department, Abdul Wali Khan University Mardan, Mardan, KPK, Pakistan

³Department of Computer Engineering, Persian Gulf University, Bushehr 75168, Iran

Correspondence should be addressed to Yanru Zhang; zyr@hpu.edu.cn

Received 6 October 2020; Revised 6 November 2020; Accepted 23 November 2020; Published 3 December 2020

Academic Editor: Sara Shahzad

Copyright © 2020 Dawei Qiao 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.

Stroke is the first leading cause of mortality in China with annual 2 million deaths. According to the National Health Commission of the People's Republic of China, the annual in-hospital costs for the stroke patients in China reach ¥20.71 billion. Moreover, multivariate stepwise linear regression is a prevalent big data analysis tool employing the statistical significance to determine the explanatory variables. In light of this fact, this paper aims to analyze the pertinent influence factors of diagnosis related groups- (DRGs-) based stroke patients on the in-hospital costs in Jiaozuo city of Henan province, China, to provide the theoretical guidance for medical payment and medical resource allocation in Jiaozuo city of Henan province, China. All medical data records of 3,590 stroke patients were from the First Affiliated Hospital of Henan Polytechnic University between 1 January 2019 and 31 December 2019, which is a Class A tertiary comprehensive hospital in Jiaozuo city. By using the classical statistical and multivariate linear regression analysis of big data related algorithms, this study is conducted to investigate the influence factors of the stroke patients on in-hospital costs, such as age, gender, length of stay (LoS), and outcomes. The essential findings of this paper are shown as follows: (1) age, LoS, and outcomes have significant effects on the in-hospital costs of stroke patients; (2) gender is not a statistically significant influence factor on the in-hospital costs of the stroke patients; (3) DRGs classification of the stroke patients manifests not only a reduced mean LoS but also a peculiar shape of the distribution of LoS.

1. Introduction

Stroke is an acute cerebrovascular disease, which is characterized by the sudden numbness of some parts of the body, such as face, arm, or leg. This is a group of diseases caused by brain tissue damage, caused by blood flow to the brain due to blocked blood vessels or sudden rupture of blood vessels in the brain [1]. With the development of society and the change of life style, stroke becomes one of the main causes of disability and death in the world. According to the report of World Health Organization (WHO), stroke is the second leading cause of deaths in the world, accounting for 11.3% of the total deaths, and near 5.8 million people died of stroke [2, 3]. Global Burden of Stroke (GBS) reported that there are three characteristics for mortality rate caused by stroke in the low-income, lower-middle-income, upper-middle-

income, and high-income countries [4]. The mortality rate of stroke in low-income countries is about 43 per 100,000 population. In lower-middle-income and high-income countries, the mortality rate of stroke is about 62 per 100,000 population, while the number of deaths in upper-middle-income countries reaches up to 117 per 100,000 population. This is more than twice as high as in low-income countries.

Recently, with the number of stroke patients increasing, in-hospital costs spent on stroke are increasing year by year. In western counties, the expenditure on stroke accounts for 2%–4% of total healthcare expenditure [5]. In 2008, the total expense of stroke patients in the United States reached up to \$40.9 billion, accounting for 2.6% of the annual healthcare expenditure. The total annual expenditure of stroke in the United Kingdom was as much as £25.6 billion in 2019 [6]. Excepting institutionalization cost, the overall expenditure

of the stroke for European Union (EU) was €45 billion, accounting for 28% of the total expenditure of cardiovascular diseases [7]. The Aggregate National Healthcare expenditure of Brazil for the ischemic stroke was about £326.9 million from 2006 to 2007 [8]. This cost accounted for a large proportion of the overall healthcare expenditure since only 18% of Brazilians had private health insurance. In 2012, the entire economic expenditure of stroke was \$5 billion in Australia, and the loss of healthy life and the total burden of disease cost in 2012 was \$49.4 billion [9]. In Korea, the total economic expenditure of the stroke was \$3.53 billion, of which \$1.74 billion and \$1.79 billion were direct costs and indirect costs, respectively [10]. The authors in [11] presented a detailed analysis of financial data on the direct in-hospital costs of the stroke treatment in Lebanon.

In China, stroke is the first leading cause of mortality and disability of adults. In recent years, the incidence, morbidity, mortality, and disability-adjusted life years (DALY) of the stroke in China are on the rise, and the disease burden caused by stroke is quite serious [12–17]. The Global Burden of Disease (GBD) showed that the number of the stroke patients in China was 4.03 million in 2016, and the average growth rate was 8.3% in 1997–2016. With the ageing of the society, the acceleration of urbanization process, and the popular unhealthy life style of residents, the incidence and mortality of the stroke are increasing dramatically in China [18]. The incidences of DALY caused by stroke, respectively, accounted for 4.6% and 9.71% of all diseases in the world and in China [19], where the disease burden of stroke was more than twice the global average. In 2014, the total cost of the stroke outpatient and inpatient service in China was ¥20.71 billion, and the average annual growth rate of medical cost of the stroke was 24.96% [20]. Due to the imbalance of the economic level and specialized stroke resources among the areas of China, the prevention and management of the stroke are facing great challenges in Henan province, China. Henan is located in the center of China, whose resident population in 2018 was 96.05 million, accounting for 6.9% of the total population of China [21]. According to the Chinese Stroke Association (CSA), the incidence of the stroke of Henan province in 2012 was 3.22%, which was the highest among all provinces of China [22]. The number of stroke patients was about 1.5 million, and the direct and indirect cost of stroke were ¥10 billion. Jiaozuo is located in the northwest of Henan, whose population was 3.55 million in 2019. The differences between urban and rural areas and the regional imbalance in stroke prevention and treatment in Henan are more significant, which are prominent problems that need to be solved at present. Therefore, it is of profound importance to investigate the costs of stroke in Jiaozuo. Table 1 presents the incidence and mortality caused by stroke between 25 and 74 years old from 1987 to 1993. Table 1 reveals that Henan has higher incident and mortality than the average of China.

How to solve the problem of the rapid growth of costs for the stroke has always been a worldwide concern, which has sparked a great deal of research interest. The pioneering work was conducted by Fetter et al. in Yale University, from which diagnosis related groups (DRGs) were originally

proposed [24]. DRGs were first used to monitor the quality of medical services in medical institutions [25]. Then, the second generation DRGs have been realized in 1983, which were introduced as the basis of the hospital paying of healthcare systems. It was shown that, due to the application of DRGs in the United States from 1983 to 1990, the proportion of the total medical expenditure of gross domestic product (GDP) decreased from 16–18% to 7–8% [26]. Since then, DRGs have been adopted by the most developed countries as the prospective payment system (PPS) to control in-hospital cost [27–32]. It is manifested that DRGs-based PPS can not only improve the utilization efficiency of medical resources, but also reduce the length of stay (LoS) in hospital and the stroke burden.

In China, it can be divided into three stages for the DRGs [33]: (1) the first stage is exploring; (2) the second stage is piloting; and (3) the last one is completing. The first stage started from 1980s to 2001. In this stage, DRGs were introduced and explored by Beijing and Tianjin to provide a strong financial incentive for the healthcare of public hospitals [34, 35]. The second stage was motivated by new rural cooperative medical system (NCMS) for rural Chinese residents [36] ‘Notice on Piloting the Simplified DRG-PPS’ of China’s Ministry of Health (CMH) and advanced Beijing-DRGs (BJ-DRGs), where the DRGs-PPS/genuine DRGs have been piloted in some public hospitals of China [37, 38]. In the third stage, CMH has released two issues of ‘2008 Quality Supervision and Management Manual of Simplified DRGs-PPS’ of Central People’s Government of the People’s Republic of China (CPGPRC) and ‘Five Key Reforms on Medicine and Health’ to assist the promotion of DRGs across the entire China [39, 40]. In [41], DRGs-based payment has been conducted on a trial basis in the selected hospitals in Shanghai. Therefore, unlike the DRGs from other developed countries, the DRGs currently being adopted by China can be regarded as a transitional version of other countries’ DRGs [33].

Henan has the largest population of China, especially the rural population, which is urgent to keep the balance of medical resources between rural and urban. In 2018, Health Commission of Henan Province (HCHP) has issued the ‘Notification on the Key Points of Provincial Medical and Political Work’, which stated that quality control of the first page of medical records should be strengthened [42]. Jiaozuo city is in the northwest of Henan province. Its total area is 4,071 square kilometres, and the city has a permanent population of 3.5971 million. In addition, the urbanization ratio has reached 60.94%, and Jiaozuo has just issued to promptly explore DRGs payment systems in 2019 [43].

In addition, with the development of information and computer technologies, massive growth of data is a great challenge for further applications. To solve this problem, big data mining and analyzing of collected data to extract the valuable information have become main tasks. For this aspect, multivariate linear regression analysis has been identified as an effective evaluation way of big data by employing the statistical significance to determine the explanatory variables. Big data has been extensively adopted in many fields, such as information communication, finance,

TABLE 1: Summary of average age-standardized incident and mortality of acute stroke in 25–74-year-old population in 1987–1993^a [23].

Monitored area	Average monitored population		Incident (/0.1 million)		Mortality (/0.1 million)		Fatality rate (%)	
	Male	Female	Male	Female	Male	Female	Male	Female
Heilongjiang	293, 929	284, 955	646	368	129	89	20	24
Jilin ^b	193, 436	179, 135	508	256	104	68	20	26
Guangdong ^b	37, 780	32, 880	330	167	94	44	28	26
Liaoning	248, 140	243, 237	276	137	113	68	41	49
Beijing	234, 776	241, 248	247	196	91	72	33	36
Henan ^b	65, 005	63, 403	254	191	140	105	55	54
Hebei	60, 673	61, 413	236	166	101	81	43	48
Inner Mongolia ^b	88, 629	86, 754	217	169	77	58	25	34
Shandong ^b	58, 922	52, 374	210	134	65	59	31	44
Fujian ^b	29, 708	28, 905	174	71	112	43	64	60
Xinjiang ^b	17, 898	16, 437	174	198	41	51	23	26
Shanghai	124, 014	133, 591	150	117	72	53	48	45
Szechwan	68, 089	68, 234	133	80	72	46	54	57
Jiangxi ^b	58, 618	54, 853	102	74	46	32	45	43
Jiangsu	112, 749	114, 785	95	55	50	33	52	60
Anhui	38, 107	36, 843	63	45	43	30	68	66
Total	1,730, 473	1, 699, 047	270	161	89	61	33	38

^aThe age-standardized rate is the standardized rate of the world population. ^bThe marked cooperative province in 1987–1989 (Shandong, Fujian, Jiangxi, Henan, and Guangdong) or 1987–1991 (Inner Mongolia).

security, energy, and electricity [19–23]. With successful applications of big data in the above fields, it has provided many conditions and experience for its application in the healthcare field.

Motivated by the above discussion, this paper aims to investigate the effects of China Healthcare Security-DRGs (CHS-DRGs)-based hospital payment on the stroke patients in Jiaozuo area of Henan province. In the following, we simply use “DRGs” to refer to “CHS-DRGs”. Some influence factors of in-hospital cost of the stroke patients are analyzed via multivariate linear regression analysis, which is one of big data algorithms [11], such as age, gender, LoS, and outcomes. In addition, contrary to the existing works, we also study the distributions of length of stay (LoS) for DRGs-based stroke patients since they reveal the burden not only for stroke patients, but also for medical resources for hospitals. This also presents an incentive for hospital to reassign the medical resources for different stroke patients. The synthetization indicates that DRGs can save the medical costs, improve medical service quality, and reduce LoS. Finally, we carry out the analysis of the impact of outcomes on the in-hospital cost with boxplots.

2. Materials and Methods

2.1. Description of Data. This study is conducted based on the data collections of the First Affiliated Hospital of Henan Polytechnic University of Jiaozuo; 3,590 discharged stroke cases from 1 January 2019 to 31 December 2019 were selected. In order to avoid the extreme casemix, four cases are ignored: (1) LoS is smaller than 1; (2) LoS is larger than 60; (3) the in-hospital cost is smaller than 7000; and (4) the in-hospital cost is larger than 300,000. We select the data of 2019 since the Department of Human Resources and Social Security of Henan Province (DHRSSHP) has issued the

notification that DRGs have been determined as the payment of Henan public hospitals of DHRSSHP [44]. The extracted data includes admission number, age, gender, occupation, dates and modes of admission and discharge, admission condition, expense, and diagnostics of cerebral infarction (ICD-10 codes: I60–I63). In addition, patients transferred from other hospitals and dying before discharge are excluded in the extracted data.

2.2. Statistical Analysis of Data. The objective of this paper is to analyze the effects of pertinent influence factors of DRGs-related stroke patients on the in-hospital cost in Jiaozuo public hospitals of Henan province. As in China Healthcare Security-DRGs (CHS-DRGs) [45], the data of the stroke patients can be divided into four stroke-related groups based on the major diagnosis, LoS, admission mode, type of stroke, and outcomes. According to the above discussion, the number of effective data is 3,590. The characteristics of stroke-related groups in Jiaozuo hospital are illustrated in Table 2. In this study, readmission and priority patients are omitted since the number of patients of this stroke kind is only two. For the purpose of convenience, LQ and UQ represent lower quartile and upper quartile, respectively. The average payment includes cure fee, bunk fee, western medicine fee, Chinese medicine fee, examination fee, emission fee, surgery fee, lab fee, inspection fee, sanitary materials fee, and other fees.

Groups I60–I62 are of the type hemorrhage, while group I63 is of the type ischemic. As shown in Table 2, I61 has the most LQ of LoS, UQ of LoS, and median of LoS, which are, respectively, 12 days, 34 days, and 22 days. Then, LQ of LoS, UQ of LoS, and median of LoS in I62 are 11 days, 29 days, and 21 days, respectively. I63 has the least LQ of LoS, UQ of LoS, and median LoS of 8 days, 17 days, and 14 days.

TABLE 2: Characteristics of stroke-related groups in Jiaozuo hospital (2019)

Code	LQ LoS	UQ LoS	Type of the stroke	Admission mode (%)	Median LoS	Average payment
I60	9 days	24 days	Hemorrhage	Outpatient service (23.33), emergency treatment (76.67)	17 days	¥110,022.70
I61	12 days	34 days	Hemorrhage	Outpatient service (36.20), emergency treatment (63.80)	22 days	¥40,481.09
I62	11 days	29 days	Hemorrhage	Outpatient service (30.56), emergency treatment (69.44)	21 days	¥45,414.65
I63	8 days	17 days	Ischemic	Outpatient service (62.27), emergency treatment (37.73)	14 days	¥18,644.21

Although I60 has smaller LQ of LoS (9 days), UQ of LoS (24 days), and median LoS (17 days) than those of I61 and I62, it has the most expensive group due to larger treatment fee, western medicine fee, and sanitary materials fee. The expenses of the other three groups are ¥40,481.09, ¥45,414.56, and ¥18,644.21, respectively. Compared with I60, I61, and I62, I63 has a large proportion of outpatient service. The outpatient service and emergency treatment account for 62.27% and 37.73%, respectively.

From Table 2, it is clearly shown that all groups have large proportions of cure and improvement. Compared with other groups, I62 has the largest death rate, about 5.71%, which is about twice larger than that of I61 and about six times that of I63. Finally, as can be seen, the ischemic group has the least in-hospital cost compared to other groups due to the least LoS and largest improvement.

2.3. Measurement of LoS. To analyze the effects of DRGs-related stroke on the LoS and in-hospital cost, we investigate the distribution of the LoS of in-hospital stroke patients, which can be divided into one most expensive group (I60), two more expensive groups (I61 and I62), and one least expenditure group (I63). In light of this fact, the distributions of LoS of four groups are presented in the form of histograms to distinguish the empirical distributions of LoS. To this end, the in-hospital patients can be divided into hemorrhage stroke and ischemic stroke. The distributions of LoS of the above two types are illustrated by the histograms, which can distinguish the empirical distribution of LoS of the two types. The effects of age, gender, and LoS on the cure, improvement, and death for the two groups are analyzed. It is noted that the other terms are removed since they include many kinds of cases.

2.4. Measurement of Outcomes. For the purpose of evaluating the performance of the hospital treating of the stroke patients, four outcomes are considered, namely, cure, improvement, unhealed, and death. The above cases are computed on the basis of discharge mode. Moreover, we investigate the distributions of cure, improvement, unhealed, death, and others according to the type of the stroke (hemorrhage and ischemic), as shown in Table 3. This comparison is adopted for the following two reasons: (1) stroke is classified into hemorrhagic and ischemic, which can keep representative datasets; (2) some sampling errors can be controlled by avoiding the cases of small number of stroke patients. Therefore, unhealed of these two groups is removed due to very small number of the stroke patients. The variation, the central tendency, and the outliers for the outcomes of the stroke patients are evaluated by utilizing boxplots. On

this basis, the variation is characterized by the first quartile and the third quartile, while the central tendency is measured by the median. The outliers capture the data points outside the boxplot whiskers.

3. Results

In this section, we will explore the influence factors of hospitalization expenses of the stroke patients. First, considering age, gender, LoS, and outcome as dependent variables, we study the effects of DRGs of stroke patients on the hospitalization expense via multivariate linear regression analysis. Then, the distributions of LoS of DRGs-related stroke patients for I60, I61, I62, and I63 are discussed. Finally, we present the effects of outcomes of the stroke patients on the hospitalization expenses.

3.1. Stroke-Related DRGs. In order to obtain more insights, by utilizing multivariate linear regression analysis [46], we carry out the analysis of influence effects on the total in-hospital costs of the DRGs stroke patients as shown in Table 4. In this table, we take total in-hospital costs as the independent variable and the patient's age, gender, LoS, and outcomes as the dependent variables. Using multivariate linear regression analysis, we can conclude that although the unstandardized coefficient of gender is not statistically significant, the multiple linear regression model has a good agreement with $R = 0.539$, $R^2 = 0.291$, and $\Delta R = 0.291$ due to a P value less than 0.05. In addition, the unstandardized coefficients of age, LoS, and outcomes are statistically significant.

3.2. Analysis of the Distribution of LoS. As in Figures 1–4, the mean LoS of DRGs-related stroke inpatients for I60, I61, I62, and I63, based on the stroke patient data from the First Affiliated Hospital of Henan Polytechnic University, is 17, 22, 21, and 14 days, respectively. In Figures 1–4, we show the histograms of density of LoS for different groups according to DRGs. From Figures 1 and 3, we can find that there are some vacancies due to the small number of the stroke patients. The histograms in Figures 1–3 almost have the same peak of LoS, which is 13.30%, 13.30%, 12.58%, respectively, while the density of Figure 4 is about 19.95%. This means that ischemic stroke patients have higher peak of LoS than the other three groups. Finally, we can also conclude that DRGs classification of the stroke patients, using LoS as a grouping variable, manifests not only a reduced mean LoS but also a peculiar shape of the distribution of LoS. It happens that most stroke patients stay in hospital for all the groups between 10 and 20 days.

TABLE 3: Number and proportion of the outcomes of the stroke patients (2019)

Code	Cure (%)	Improvement (%)	Unhealed (%)	Death (%)	Others (%)	Number of the stroke patients (%)
G_1	124 (32.98)	161 (42.82)	2 (0.53)	10 (21.01)	79 (21.67)	376 (10.47)
G_1	869 (27.04)	2,221 (69.10)	8 (0.25)	29 (0.90)	87 (2.71)	3,214 (89.53)

TABLE 4: Analysis of regression model of the stroke patients and impact factors^c.

Model	Unstandardized coefficients		Standardized coefficients	T	P	Collinearity statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-5,524.606	2,984.944		-1.851	0.064		
Age	-176.765	40.343	-0.062	-4.382	0.000	0.989	1.012
Gender	-961.111	1,079.232	-0.013	-0.891	0.373	0.990	1.010
LoS	1,448.955	40.034	0.513	36.193	0.000	0.984	1.016
Outcomes	9,159.222	571.752	0.227	16.020	0.000	0.983	1.017

^c $R = 0.539$, $R^2 = 0.291$, adjusted $R^2 = 0.290$, F variation = 367.897, $\Delta R = 0.291$.

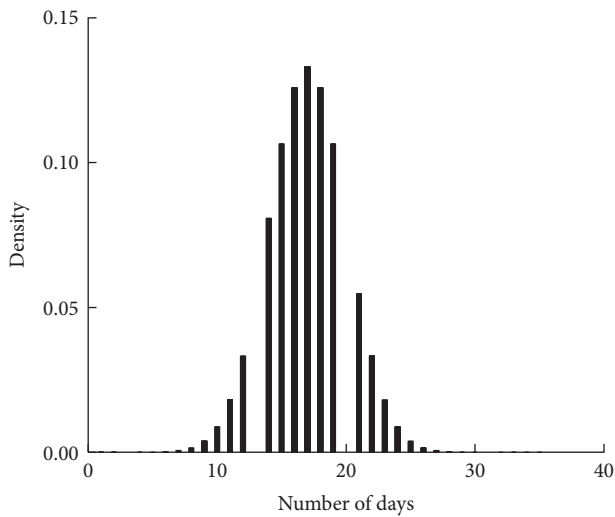


FIGURE 1: Effects of DRGs-related stroke on the LoS for I60.

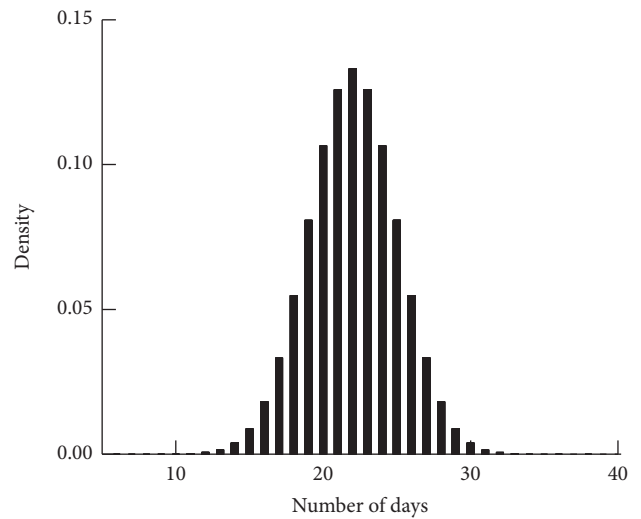


FIGURE 2: Effects of DRGs-related stroke on the LoS for I61.

3.3. Analysis of Outcomes on Hospitalization Expenses.

In this subsection, the boxplots are provided to demonstrate the impacts of outcomes on the in-hospital cost. For analytical accuracy, we remove other terms since they include different influence factors. In addition, the values of hospitalization expenses larger than 200,000 and smaller than 7,000 are removed.

Figures 5–8 show boxplots of the hospitalization expenses versus different outcomes, namely, cure, improvement, unhealed, and death. As shown in Figures 5–8, the means of cure and improvement are ¥39,565.73 and ¥56,059.41 compared with ¥25,513.28 and ¥42,597.81 of unhealed and death. The interquartile intervals of cure, improvement, unhealed, and death are [¥11,942.11; ¥54,103.61], [¥16,686.12; ¥76,446.23], [¥16,686.12; ¥25,964.6], and [¥16,124.41; ¥47,837.57], respectively. This means that interquartile intervals of unhealed are the smallest. The differences of others are not very large and concentrate between ¥10,000 and ¥60,000.

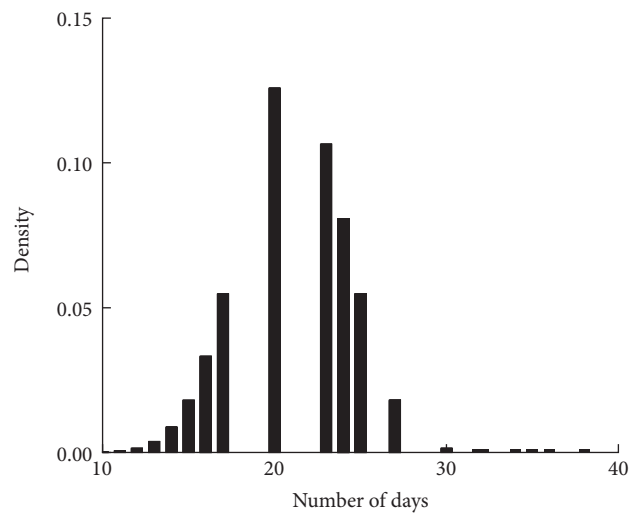


FIGURE 3: Effects of DRGs-related stroke on the LoS for I62.

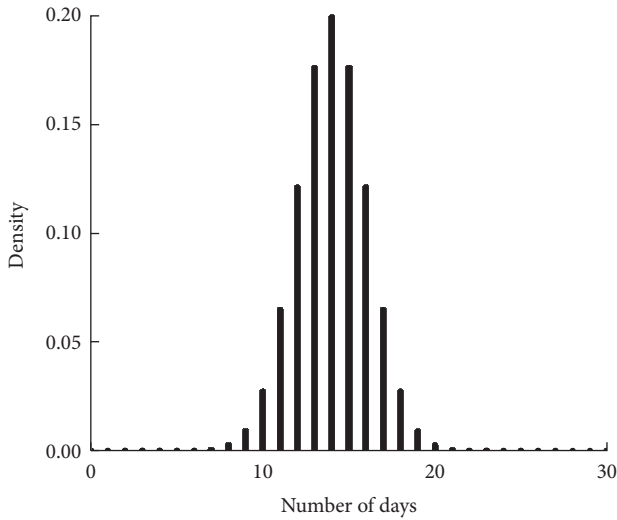


FIGURE 4: Effects of DRGs-related stroke on the LoS for I63.

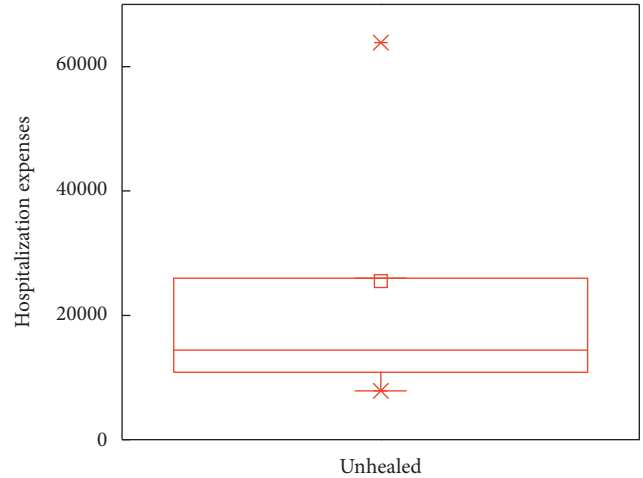


FIGURE 7: Effects of DRGs-related stroke on the in-hospital cost for unhealed.

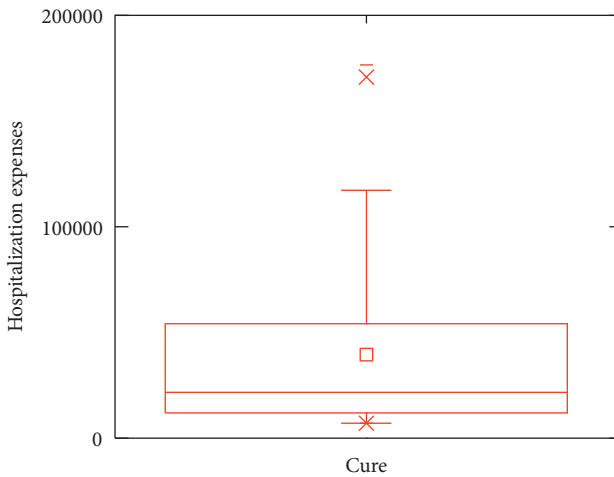


FIGURE 5: Effects of DRGs-related stroke on the in-hospital cost for cure.

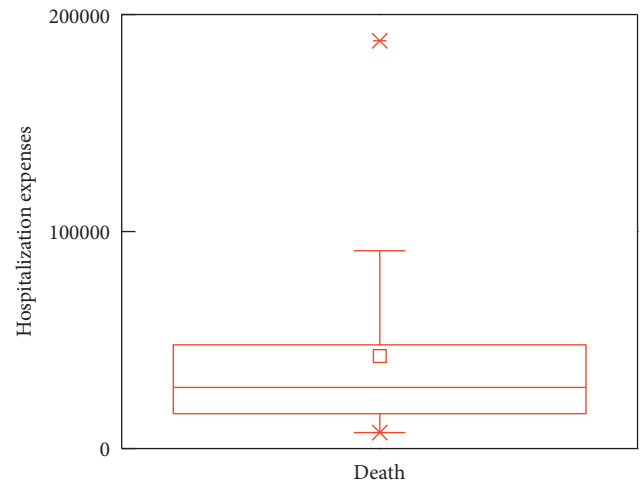


FIGURE 8: Effects of DRGs-related stroke on the in-hospital cost for death.

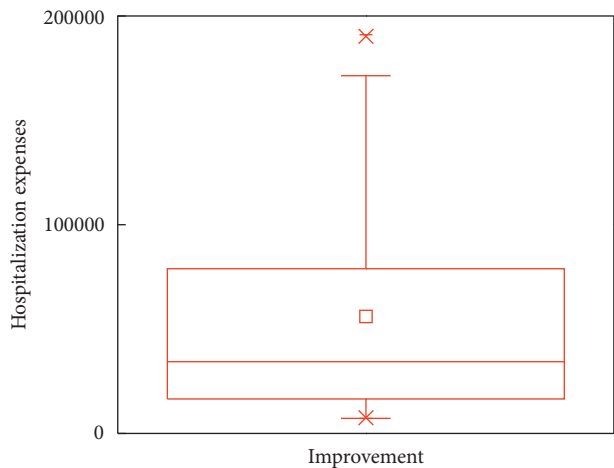


FIGURE 6: Effects of DRGs-related stroke on the in-hospital cost for improvement.

4. Discussion

Recently, the rapid increase of in-hospital cost has become a common problem in our country. Stroke has become the first leading cause of death among urban and rural residents in China. It not only leads to the loss of labor ability, but also dramatically increases the cost of diagnostic hospitalization. Its burden of disease puts enormous economic pressure on families and society. In addition, payment systems have significant effects on treatment behaviors for stroke patients and hospitals. DRGs are effective ways to improve quality of service and reduce unnecessary medical resources, which has been identified as an important direction for China's medical reform, and DRGs have been adopted as the main payment system [46]. Therefore, it is of great significance to analyze the influencing factors of stroke hospitalization expenses for social and economic benefits.

Based on the multivariate linear regression analysis, we show that stroke is a common cardiovascular disease in

Jiaozuo area. Among all the influence factors, gender is the least one, which has no effect on the hospitalization expense. In the future hospital payment system, gender should not be included in group pricing of the payment standard. Moreover, age, LoS, and outcomes are the three significant influence factors that should be included in the payment standard. Finally, we can conclude that the DRGs yield perfect coefficient of variation (CV) and reduction in variation (RIV) results for the medical expense control, which is consistent with the existing literature [47]. Motivated by this, DRGs are the main directions of Jiaozuo's medical reform.

In general, LoS can be classified into value-adding patient days and non-value-adding patient days [48]. The value-adding patient days refer to the days that are meaningful to the diagnosis and prognosis of patients and that patients or payers of medical expenses are willing to pay their expense. Non-value-adding patient days refer to those days that are not necessary, just increasing the cost of hospitalization, and are meaningless to patient's diagnosis. Thus, LoS is an important influence factor of hospitalization expenses of the stroke patients. Note that throughout this paper we use LoS to refer to value-adding patient days. By optimizing medical source allocation and improving medical service, LoS can be shortened and then hospitalization expense can be reduced. LoS of most stroke patients is between 10 and 30 days, and the peaks of density of LoS for hemorrhage and ischemic stroke are lower than 14% and 20%.

Among influence factors, outcomes have significant impacts on the hospitalization expenses, among which improvement and unhealed are the most important ones. This is because improvement has large interquartile intervals and unhealed has the largest mean. Therefore, it is of great significance to analyze the impact of outcomes on hospitalization expenses in Jiaozuo area of Henan province.

5. Conclusion and Future Work

The economic conditions and level of development of China do not meet the regulatory requirements for full implementation of DRGs. The improved DRGs schemes have been explored in some cities of China, and they will be gradually piloted across most provinces of China. Jiaozuo of Henan province, as a pilot province of DRGs, has made a lot of important explorations. Under the conditions of the existing resources, however, beginning to screen for common and frequent diseases and investigating the related diseases in each disease group and the key factors for these diseases affecting the medical expenses, according to the standard-setting process simplified based on standardization of clinical path and the accurate cost accounting, will greatly improve the execution efficiency of related diagnosis of Jiaozuo in Henan province.

Utilizing big data related algorithm, this work analyzes the influence factors of DRGs-based stroke patients on in-hospital costs via DRGs; however, how to design a DRGs model for the payment of Jiaozuo hospital will be our further work. In addition, aiming to further enhance the accuracy of the analysis, some advanced machine learning algorithms

should be involved, such as decision tree, neural network, and support vector machine. Our analysis methods can be extended to other chronic diseases (hypertension, coronary heart disease, cancer, and diabetes), which are set aside as our future work.

Data Availability

The data used to support this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

D. Qiao and Y. Zhang jointly designed the study. D. Qiao collected and analyzed the data. Y. Zhang, A. U. Rehman, and M. R. Khosravi reviewed and edited the manuscript.

References

- [1] Institute for Health Metrics and Evaluation (IHME), "Global trends in disability," Technical Reports, Seattle, WA, USA, 2017.
- [2] World Health Organization (WHO), "The top 10 causes of death," Technical Reports, Seattle, WA, USA, 2018.
- [3] W. Felix, K. Denise, and J. Schmitt, "Impact of complex quality-interventions on patient outcome: a systematic overview of systematic reviews," *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, vol. 56, pp. 1–11, 2019.
- [4] K. Mira and L. Andreas, "Global burden of stroke," *Seminars in Neurology*, vol. 38, no. 2, pp. 208–211, 2018.
- [5] X. Zhang, "Research on hospitalization costs and case mix of stroke patients of Lanzhou Tertiary Hospital," Master's thesis, Lanzhou University, Lanzhou, China, 2018.
- [6] P. Anita, B. Vladislav, Q. Zahidul, D. King, M. Knapp, and R. Wittenberg, "Estimated societal costs of stroke in the UK based on a discrete event simulation," *Age and Ageing*, vol. 49, pp. 270–276, 2020.
- [7] E. Wilkins, L. Wilson, K. Wickramasinghe et al., *European Cardiovascular Disease Statistics 2017*, EuropeanHeart Network, Brussels, Belgium, 2017, <http://www.ehnheart.org/images/CVD-tatistics-eportAugust.pdf>.
- [8] C. C. Michael, V. Raul, S. S. Gisele et al., "Acute treatment costs of stroke in Brazil," *Neuroepidemiology*, vol. 32, no. 2, pp. 142–149, 2009.
- [9] Stroke Foundation, *The Economic Impact of Stroke in Australia*, National Stroke Foundation, Centennial, CO, USA, 2013, <https://strokefoundation.org.au/Media-Releases/2015/05/28/The-economic-impact-of-stroke-in-Australia>.
- [10] H. J. Kim, Y. A. Kim, H. Y. Seo, E. J. Kim, S.-J. Yoon, and I.-H. Oh, "The economic burden of stroke in 2010 in Korea," *Journal of the Korean Medical Association*, vol. 55, no. 12, pp. 1226–1236, 2010.
- [11] A. Rachel, A. Halim, S. Pascale, N. Jomaa, R. G. Rizk, and H. H. Hosseini, "Direct medical cost of hospitalization for acute stroke in Lebanon: a prospective incidence-based multicenter cost-of-illness study," *Inquiry: A Journal of Health Care Organization, Provision, and Financing*, vol. 55, pp. 1–11, 2018.

- [12] L. F. Valery, V. K. Rita, P. Priya et al., "Update on the global burden of ischemic and haemorrhagic stroke in 1990-2013: the GBD 2013 study," *Neuroepidemiology*, vol. 45, no. 3, pp. 161-176, 2015.
- [13] L. F. Valery, A. M. George, N. Bo, J. L. M. Christopher, and A. R. Gregory, "Atlas of the global burden of stroke (1990-2013): the GBD 2013 study," *Neuroepidemiology*, vol. 45, no. 3, pp. 230-236, 2015.
- [14] V. K. Rita, E. M. Andrew, L. F. Valery et al., "Stroke prevalence, mortality and disability-adjusted life years in adults aged 20-64 years in 1990-2013: data from the global burden of disease 2013 study," *Neuroepidemiology*, vol. 45, no. 3, pp. 190-202, 2015.
- [15] V. K. Rita, D. Gabrielle, L. F. Valery et al., "Stroke prevalence, mortality and disability-adjusted life years in children and youth aged 0-19 years: data from the global and regional burden of stroke 2013," *Neuroepidemiology*, vol. 45, no. 3, pp. 177-189, 2015.
- [16] N. Mohsen, A. A. Amanuel, A. Cristiana et al., "Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the global burden of disease study 2016," *Lancet*, vol. 300, pp. 1151-2110, Article ID 10100, 2015.
- [17] J. Li, J. Liu, Y. Ma, P. Peng, X. He, and W. Guo, "Corrigendum to "imbalanced regional development of acute ischemic stroke care in emergency departments in China," *Emergency Medicine International*, vol. 2019, Article ID 9752671, 2 pages, 2019.
- [18] H. Sun, D. Huang, H. Wang et al., "Association between serum copeptin and stroke in rural areas of northern China: a matched case-control study," *Disease Markers*, vol. 2018, Article ID 9316162, 8 pages, 2018.
- [19] Global Burden of Disease Collaborative Network, *Global Burden of Disease Study 2013 (GBD 2013) Risk Factor Results 1990-2013*, Institute for Health Metrics and Evaluation (IHME), Seattle, WA, USA, 2015.
- [20] L. Wang, *Report on Stroke Prevention and Treatment in China*, Peoples Medical Publishing House, Beijing, China, 1st edition, 2018.
- [21] National Bureau of Statistics (NBS), National Bureau of Statistics, Beijing, China, 2018, <http://www.stats.gov.cn/>.
- [22] Chinese Stroke Association (CSA), Chinese Stroke Association, Beijing, China, 2019, <http://www.chinastroke.net/default.aspx>.
- [23] Z. Wu, C. Yao, and D. Zhao, "Epidemiological study on stroke incidence and mortality in China," *Chinese Journal of Epidemiology*, vol. 24, no. 3, pp. 236-239, 2003.
- [24] B. F. Robert, S. Youngsoo, L. F. Jean, F. A. Richard, and D. T. John, "Case mix definition by diagnosis-related groups," *Medical Care*, vol. 18, no. 2, pp. 1-53, 1980.
- [25] N. Goldfield, "The evolution of diagnosis-related groups (DRGs): from its beginnings in case-mix and resource user theory, to its implementation for payment and now for its current utilization for quality within and outside the hospital," *Qual Manag Health Care*, vol. 19, no. 16, pp. 3-16, 2010.
- [26] X. Gao and Q. Zeng, "Research status and problems of case combinations," *Chinese Journal of Hospital Statistics*, vol. 11, no. 3, pp. 53-55, 2004.
- [27] B. Reinhard, G. Alexander, A. Ain et al., *Diagnosis Related Groups in Europe: Moving towards Transparency, Efficiency and Quality in Hospitals*, BMJ Publishing Group Ltd, London, UK, 2013.
- [28] S. K. David, Q. Wilm, and B. Reinhard, "DRG-based hospital payment systems and technological innovation in 12 European countries," *Value in Health*, vol. 14, no. 18, pp. 1166-1172, 2011.
- [29] M. Inke and W. Friedrich, "Hospital payment systems based on diagnosis-related groups: experiences in low-and middle-income countries," *Bull World Health Organization*, vol. 91, no. 10, pp. 745-756, 2013.
- [30] M. Hennamari, K. Ilmo, and H. Unto, "DRG-related prices applied in a public health care system-can Finland learn from Norway and Sweden?" *Health Policy*, vol. 59, no. 1, pp. 37-51, 2002.
- [31] O. Zeynep, "Implementation of DRG payment in France: issues and recent developments," *Health Policy*, vol. 117, no. 2, pp. 146-150, 2014.
- [32] K. H. Uwe and S. K. David, "Policy trends and reforms in the German DRG-based hospital payment system," *Health Policy*, vol. 119, no. 3, pp. 252-257, 2015.
- [33] R. Liu, J. Shi, B. Yang et al., "Charting a path forward: policy analysis of China's evolved DRG-based hospital payment system," *International Health*, vol. 9, no. 5, pp. 317-324, 2015.
- [34] W. Y. Jian, M. Lu, X. M. Zhang, J. Yang, and M. Hu, "The grouping process and method of diagnosis related groups, Beijing version (BJ-DRGs)," *Chinese Journal of Hospital Administration*, vol. 27, no. 11, pp. 829-831, 2011.
- [35] B. Zhu, "Study on the hospitalization expenses and case mix for diabetes inpatients in Tianjin," *Chinese Journal of Hospital Administration*, Tianjin Medical University, Tianjin, China, 2012.
- [36] GMW, "Henan Piloted the case-based reimbursement," 2006, <http://www.gmw.cn/01gmr/b/2006-08/13/content%20463754.htm>.
- [37] W. Jian, M. Hu, W. Jian, and X. Zhang, "Evaluation on the comprehensiveness of diagnosis and treatment capacity based on diagnosis related groups," *Chinese Hospital Management*, vol. 30, no. 8, pp. 17-19, 2010.
- [38] S. Xiao, W. Luo, and L. Yao, "Analysis on the game of stakeholders of single diseases in cities," *Health Economics Research*, vol. 7, pp. 38-40, 2010.
- [39] Central Peoples Government of the Peoples Republic of China (CPGPRC), "Notification issued by the ministry of health on the second batch of quality management index on simplified-drugs," 2016, <http://www.gov.cn/gzdt/2010-12/13/content1764678.htm>.
- [40] Central Peoples Government of the Peoples Republic of China (CPGPRC), "Notification issued by the general office of the state council on the work arrangement of five key reform on medicine and health," 2011, <http://www.gov.cn/zwgk/2011-02/17/content1805068.htm>.
- [41] Z. Wang, R. Liu, P. Li, and C. Jiang, "Exploring the transition to DRGs in developing countries: a case study in Shanghai, China," *Journal of Medical Sciences*, vol. 30, no. 2, pp. 250-255, 2014.
- [42] Health Commission of Henan Province (HCHP), "Notification on the key point points of provincial medical and political work," 2018, <http://www.henanyz.com/index.php?op=1&id=18012210205400701>.
- [43] Jiaozuo City Peoples Government (JCPG), "Jiaozuo City medical insurance focus on six work," 2019, <http://www.henan.gov.cn/2019/04-30/789521.html>.
- [44] Department of Human Resources and Social Security of Henan Province (DHRSSHP), "Notification on the issuance of national pilot technical specifications and diagnosis related groups (DRGs)," 2019, <http://www.henan.gov.cn/2019/04-30/789521.html>.

- [45] National Healthcare Security Administration (NHSA), “Notification on the issuance of national pilot technical specifications and diagnosis related groups (DRGs),” 2019, <http://www.nhsa.gov.cn/art/2019/10/24/art371878.html>.
- [46] R. Sreeja, P. Varghese, G. M. Varun, J. Sunil, and P. G. Vinod, “Brain-controlled adaptive lower limb exoskeleton for rehabilitation of post-stroke paralyzed,” *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2531–2538, 2020.
- [47] Z. Wang, R. Liu, P. Li, C. Jiang, and M. Hao, “How to make diagnosis related groups payment more feasible in developing countries—a case study in Shanghai, China,” *Iranian Journal of Public Health*, vol. 43, no. 5, pp. 572–578, 2014.
- [48] A. V. Straten, J. H. P. V. D. Meulen, V. D. Bos, and M. Limburg, “Length of hospital stay and discharge delays in stroke patients,” *Stroke*, vol. 28, no. 1, pp. 137–140, 1997.

Research Article

Wireless Sensor Network Applications in Healthcare and Precision Agriculture

Naila Nawaz Malik,¹ Wael Alosaimi,² M. Irfan Uddin ,¹ Bader Alouffi,³ and Hashem Alyami³

¹*Institute of Computing, Kohat University of Science and Technology, Kohat 26000, KPK, Pakistan*

²*Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia*

³*Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia*

Correspondence should be addressed to M. Irfan Uddin; irfanuddin@kust.edu.pk

Received 5 September 2020; Revised 5 November 2020; Accepted 13 November 2020; Published 30 November 2020

Academic Editor: Shah Nazir

Copyright © 2020 Naila Nawaz Malik 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.

A wireless sensor network is a large sensor hub with a confined power supply that performs limited calculations. Due to the degree of restricted correspondence and the large size of the sensor hub, packets sent through the sensor network are based primarily on multihop data transmission. Current wireless sensor networks are widely used in a range of applications, such as precision agriculture, healthcare, and smart cities. The network covers a wide domain and addresses multiple aspects in agriculture, such as soil moisture, temperature, and humidity. Therefore, issues of precision agriculture at the output of the network are analyzed using a star and mesh topology with TCP as the transmission protocol. The system is equipped with two sensors: Arduino DFRobot for soil moisture and DHT11 for relative temperature and humidity. The experiments are performed using the NS₂ simulator, which provides an improved interface to analyze the results. The results showed that the proposed mechanism has good performance and output.

1. Introduction

Movement types (like satellite trajectories), sensor systems, network preparedness, pervasive figuring, and the associated selection support improve the outline and important association limits [1]. Wireless sensor networks in Figure 1 are a collection of specific transducers with an exchange base to monitor and record the conditions at nodes, which routinely include the parameters of temperature, stickiness, weight, wind headings, and pace. This information helps motivate and drive prerequisites, key body limits, sound power voltage, undermining level, vibration power, and compound obsessions.

A wireless sensor network (WSN) is also called a wireless sensor or wireless sensor advance network (WSAN) and is a collection of free sensors that screen physical or steady conditions, such as temperature, sound, and weight. This

information is transmitted through the network to a focal area. Current networks are bi-directional and enable sensor control. Advanced WSNs have been used for military applications as battle zone affirmation. Such networks are also used in various mechanical and customer applications for event monitoring, network status assessment, machine accomplishment sensing, etc.

Sensors are used to gather information about common physical quantities; however, actuators are used to react to and control the reported circumstances. The sensors collect information that describes the object or environment and are used to provide information on people, areas, objects, and their states. Association securing shows the conditions of spaces that have social events at the time of collection, including for agribusiness. Creating space addresses some necessities, including (1) air collected, harvest, and soil data, (2) monitoring of coursed zones, (3) various yields on an

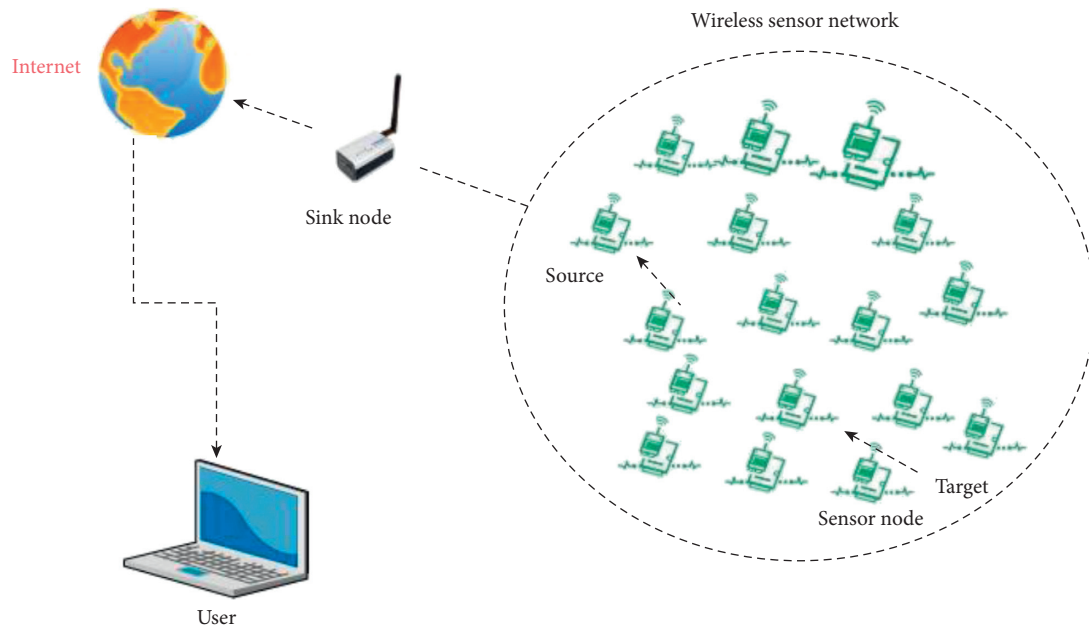


FIGURE 1: Wireless sensor network (a sample architecture).

area bundle, (4) distinctive compost and water that are central to different uneven zones, (5) diverse necessities of harvests for various air and soil conditions, and (6) proactive strategies instead of responsive blueprints. A simple structure of sensor is given in Figure 2.

There are several applications of sensor networks. These include frustration encouraging operations, drop sensor focus from a plane over a quickly spreading fire where every point measures the temperature to derive a “temperature map.” In addition, these sensors can generate biodiversity maps or watch wildlife. Intelligent networks (or expansions) can help reduce centrality waste based on the formed stickiness, ventilation, and cooling control. This allows estimating room inhabitation, temperature, and wind streams. Such networks can screen mechanical issues after earthquakes and machine acknowledgment for preventive upkeep. Embedded sensing/control distinguishes changes in systems, such as tire weight checking or accurately determining the status of compost/pesticides/watering systems. In the health industry, such applications include long-term perception of continually handicapped patients or the elderly.

Sensors have been used for a long time in various networks. The rule indoor controller was established in 1883, and many consider this as the basic, present-day, assembled sensor. Infrared sensors have been around since the late 1940s and have recently become standard devices. Progression markers have been utilized for many years. The sink node assembles data in wireless sensor systems, which means that data gathering may skip samples or perform multi-bounce where all sensors store data that is sent to the base

station, which is called the sink focus. Every middle point in the sensor network contains three subsystems. The first senses the environment, the second readies the subsystem to perform neighborhood estimation on the perceived data, and the third handles message exchange.

A static sink focus is typically used for data gathering for a wireless sensor system using multicorochet sending, indicating that it is more critical to be utilized near base points on the base station and transferring data from other sites. An adaptable sink focus is generally used to collect all the data from sensor focus points and send it to the base station. A WSN routinely incorporates expansive totals of focuses performing nearby to wirelessly shape networks. Each middle point is self-overseeing and has a short total range, indicating they are valuable and viable over a large zone. The typical sections of a network are as follows:

- (i) Sensors: reaped or set away from power hotspots for party and transmitting data about the environment
- (ii) Access network: sink focus gathering data from a collection of sensors and engaging correspondence with a control focus or outer segments
- (iii) Middleware: software to gather and process the information
- (iv) Application platform: a progression stage for the exceptional use of a WSN for a specific application

Many studies proposed very efficient techniques to get significant results; many of them discussed data gathering, energy consumption, monitoring mechanism, and network

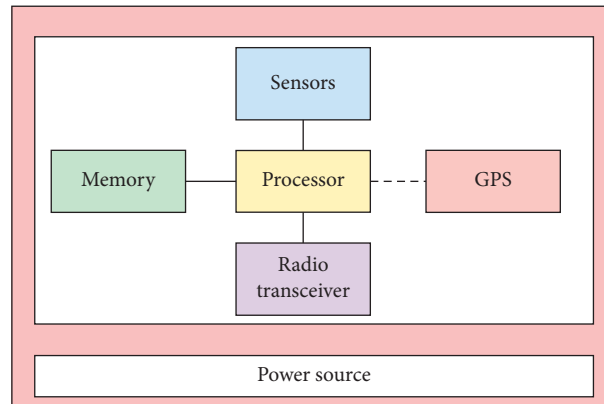


FIGURE 2: Basic components of sensor.

techniques such as Zigbee, Bluetooth GPRS, and others. After literature review, some research questions are raised:

- (i) How much fast data is delivered?
- (ii) What data rate is needed? How much throughput of the network could be good for data transmission?
- (iii) What is the data delay among the nodes?
- (iv) What is the ratio of packet lost?

To find the answers of the above-mentioned questions, we proposed a mechanism to get reliable results. The results show the actual performance of network output. In our mechanism, we take 2 sensors and deploy them via mesh topology by using NS₂ simulator.

2. Literature Review

Akyildiz et al. [2] found that farming is a rich application space in which the limits of utilizing WSN and WSN are high. The WSN persistently works in surprising conditions and its wireless directivity is slip-up slanted due to the obstructions brought by environmental conditions. The transmission operations in WSN are unequivocally required in some key applications. Kang et al. [3] showed that WSNs have been considered for their potential application in various fields like prosperity care, disaster organizations, global positioning systems, furnished powers, environmental sciences, and cultivation. In sensor applications, the openness of benefits is adaptable and available to achieve a sensible network lifetime. Chen and Varshne [4] found that, in wireless sensor networks, it is an issue to delineate sensible guiding customs to execute different applications in different environment. In this partnership, Aces have proposed specific controlling customs to overhaul execution requests for unmistakable applications through the network layer of conventional wireless sensors.

Tanwar et al. [5] indicated that WSNs are self-sorting and include different types of sensor focuses for wireless operability. Creating, battle zone and military perception, living space monitoring, security learning, achievement

monitoring, observations of standard catastrophes like timberland fires, mechanical technique control, and speedy transportation are some of the applications of WSNs. Wireless sensor focuses rely on inherent and compelled essential resources and are placed in wireless environments to collect information. Reducing the power consumption of a sensor focus point helps extend its lifetime.

Mahfooz et al. [6] illustrated that WSNs include the correspondence of small perceiving segments that work together to gather information, arrange, and surrender over wireless channels regarding physical phenomena. The self-overseeing and criticalness of these important networks can be prominent means to monitor underground mining, untamed wildlife, and contrasting physical establishments for cases, stages, pipelines, and networks. Developing countries have a multifaceted test in utilizing and tracking important resources. While the explanations for inefficient utilization of preferences are insufficient and without clear solutions, focus has shifted to using small-scale electronic contraptions to handle these issues, which require reporting properties of particular physical phenomena. The interface with the physical world creates adaptability for operations and wireless control.

Abouzar et al. [7] also proposed a mechanism to solve the inference problem in message delivering by using Receive Signal Strength Indication-based self-localization by deploying Bayesian algorithm for information aggregation. Praynlin and Ida Jenson et al. [8] selected two datasets, namely, SoDa (Solar Radiation Data) and NCEP (National Center for Environmental Predictions), to forecast average daily solar irradiance using two models (RBFN and BPN). Two-year data was used for training of the network and one-year data was used for testing. The RMSE values were observed to be less for BPN, precisely 3.12 MJ/m² and 3.212 MJ/m² for the two datasets.

Srivastava et al. [9] performed prediction of average daily solar intensity 6 days ahead using Model Averaged Neural Network (MANN). Nine parameters including time, average temperature, minimum temperature, maximum temperature, rain, wind, dew point, atmospheric pressure, and azimuth were given as inputs to neural networks. Model was

compared with FFNN, RBFN, and BPN. Average of monthly RMSE for 12 months was observed to be less for MANN with a value of 204.52 W/m^2 .

3. Materials and Methods

NS₂ is a discrete-event network simulator that was developed in 1989 as an improvement to an earlier network simulator. This simulator is a combination of C++ and OTcl, making it an object-oriented scripting language. Support for wireless networks was added in 1997, which was designed to simulate wireless LAN protocols, though this was later expanded to mobile ad hoc networks. A project at the Naval Research Laboratory produced an extension to the NS₂ for sensor webs. This extension adds a phenomenon channel module to model physical phenomena such as sensor nodes and the environment. Although NS₂ has been used to evaluate WSNs, the accuracy of the results is questionable as the MAC protocols, packet format, and energy models are different from those of typical sensor web platforms. NS₂ began as a variant of the REAL network simulator in 1989 and has evolved substantially over the past few years.

3.1. NS₂ Scenario Generator (NSG). The NS₂ scenario generator (NSG) is a tool that can run on any platform and generate TCL scripts for wired and wireless scenarios. The main features of the NSG are as follows:

- (1) Create wired and wireless nodes via drag and drop
- (2) Create simplex and duplex links for wired networks
- (3) Create grid, random, and chain topologies
- (4) Create TCP and UDP agents while supporting TCP
- (5) Tahoe, TCP Reno, TCP New-Reno, and TCP Vegas
- (6) Support ad hoc routing protocols such as DSDV
- (7) AODV, DSR, and TORA
- (8) Support FTP and CBR applications
- (9) Support node mobility
- (10) Set the packet size, start time of simulation, and end time
- (11) Time of simulation, transmission range, and interference
- (12) Range of wireless networks
- (13) Set other network parameters, such as bandwidth for wireless scenarios

The NS₂ was developed by UC Berkeley and is an open-source software simulation platform for network technologies. The program is a discrete event simulator with a virtual clock where all the simulations are driven by discrete events. The abundant modules it contains, which include nearly all aspects of network technologies, allow researchers to easily develop a network technology. NS₂ has become one

of the first selected software applications to implement network simulations in academia.

4. Results

Two sensors (Arduino DFRobot for soil moisture and DHT11 for relative temperature and humidity) are used in the experiments for different networks. The results are analyzed using the NS₂ simulator. The DSR wireless routing protocol is used in the simulation experiments. To communicate between sensors in the TCt3g5fr7u6t5P wireless communication protocol, ten, twenty, fifty, and one hundred nodes are used to make the network and validate the data rate, throughput, packet loss, data delay, and data delivery ratio for the sensors.

4.1. First Scenario. The Arduino DFRobot soil moisture sensor was used to create a network of ten nodes for the simulations as shown in Figure 3, which gave the following results. We transmitted 2685 packets over the network and received 2622 packets. There were 272 routing packets. A total of nineteen packets were analyzed and data lost were 3236 bytes in 10 nodes network. Throughput and delay were 213.02 and 792.21, respectively. As shown in Figure 4, there was an apparent loss of packets. The data loss ratio was low and the throughput was low.

4.2. Second Scenario. The number of nodes was increased to twenty and the same procedure was applied in the NS₂ simulator. For the twenty-node network with a star topology, it was seen that the packet loss ratio increased, and the throughput decreased relative to the ten-node network. 2351 packets were sent and 2293 received where 290 were routing packets. The number of dropped data packets and dropped data was 23 and 1895 bytes, respectively, with a packet loss ratio of 0.07%. Throughput and delay were 186.31 and 727.84 approximately. In Figures 5 and 6 a better view of the results could be seen.

4.3. Third Scenario. As shown in Figure 7, the third simulation considered fifty nodes with the reasonable results. In this scenario, we transmitted 3047 packets and received 2985 with the loss of 62 packets, where the packet loss ratio was 0.12; we got an average delay of 672.62 ms and throughput of 242.94 kbps. Dropped data bytes were 5536 recorded. 420 packets were considered as routing packets. We present the above-mentioned results in Figure 8 for better analysis and understanding.

4.4. Fourth Scenario. The final scenario used one-hundred-node network with the same process. Sent and received packets were 2657 and 2597, respectively, with 0.11% packet

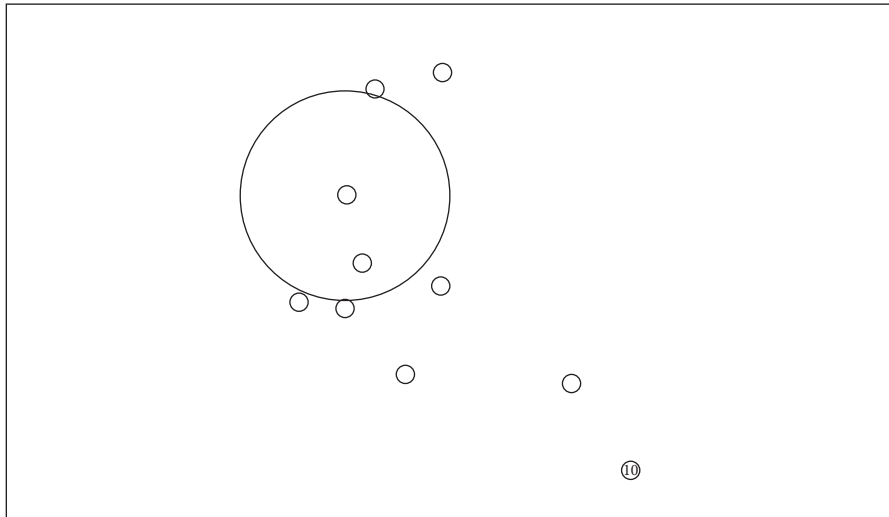


FIGURE 3: Simulation with ten nodes in the wireless sensor network.

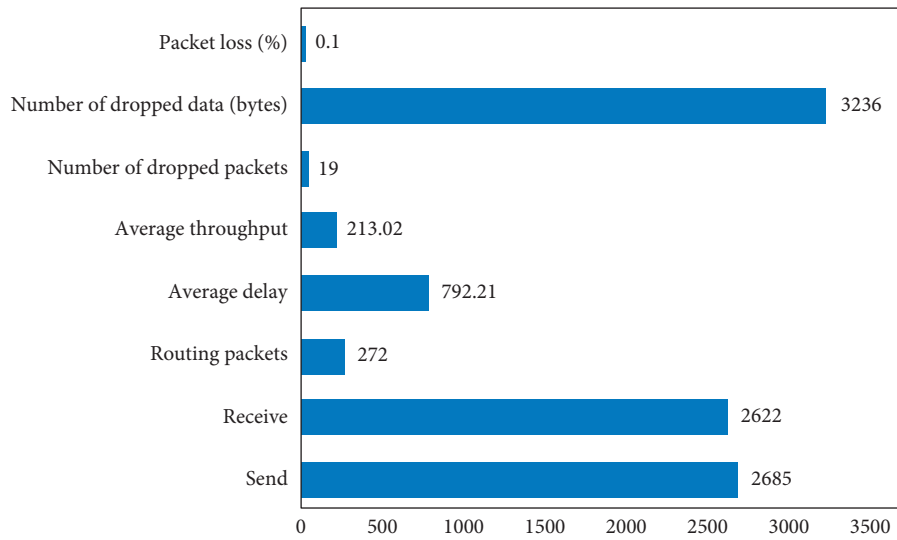


FIGURE 4: Soil moisture measurements from the DFRobot over ten nodes.

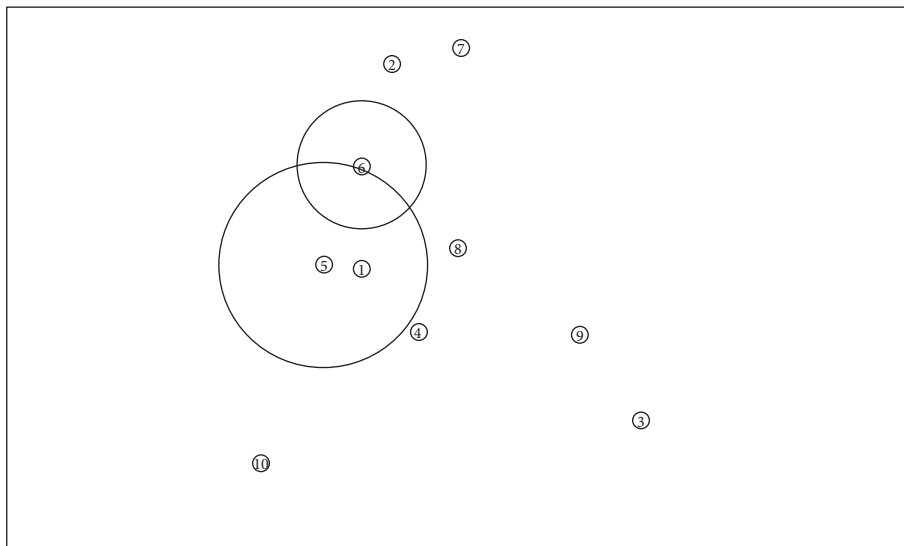


FIGURE 5: Simulation with twenty nodes in the wireless sensor network.

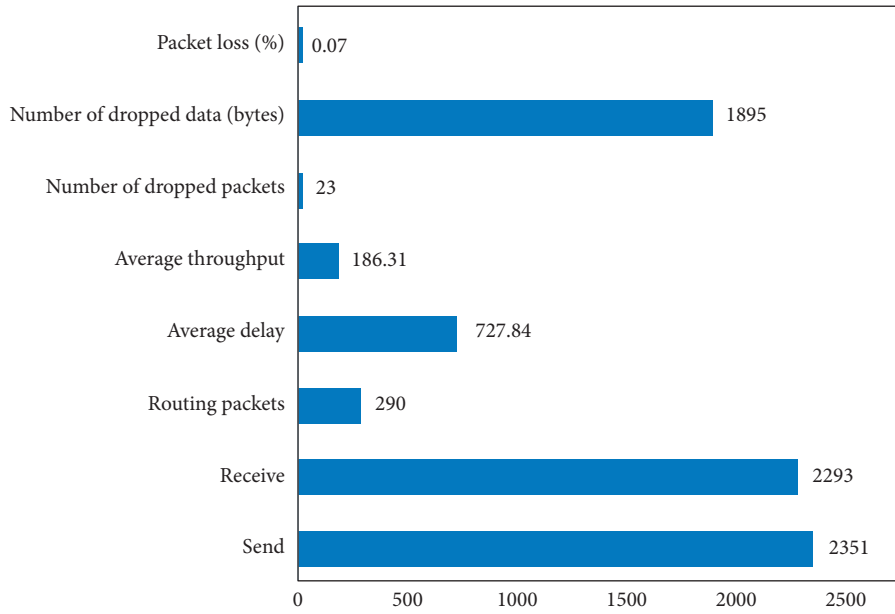


FIGURE 6: Soil moisture measured from the DFRobot for a twenty-node network.

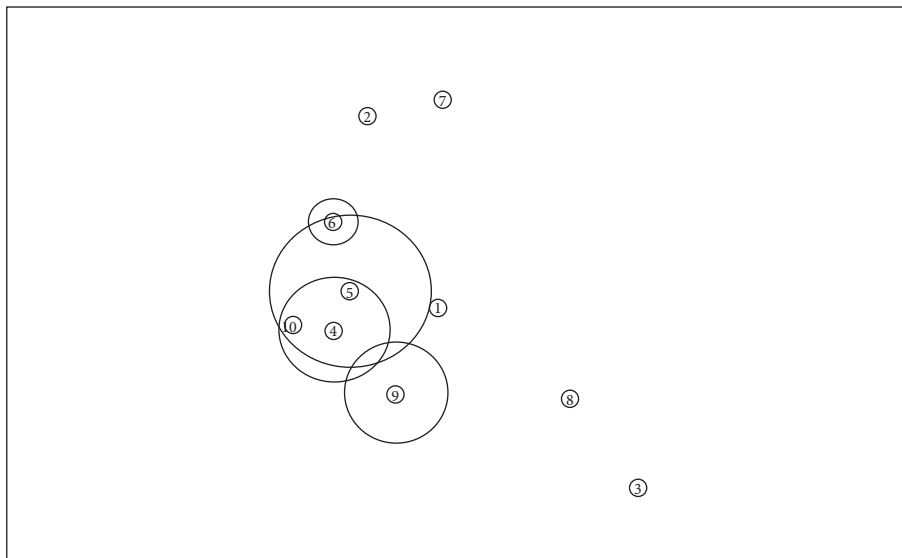


FIGURE 7: Simulations results for a fifty-node wireless network using the DHT11 sensor.

loss. During transmission, 60 packets were dropped and dropped data ratio was 6516 bytes. We noticed 779.40 ms average delay and 211.21 kbps throughput.

The simulated output results for the two sensors differed as the number of nodes increased. When there were more nodes, the packet loss ratio increased, and the throughput decreased. Analyzing the procedures for precision

agriculture indicates there are no problems with the sensors, but there are slight problems in the network side at the back end. In the first simulation, there are only ten nodes. The analyzed output of the simulations indicates that if there is a small number of packets, then the percent packet loss is higher. However, in the fourth scenario, the situation is different, as shown in Figures 9 and 10 for the one-hundred-

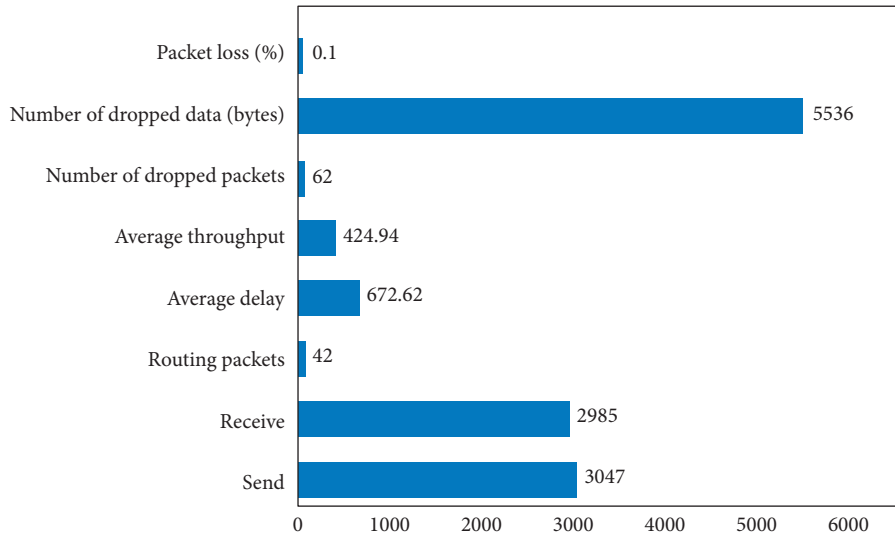


FIGURE 8: Data from the DHT11 sensor with a fifty-node wireless network.

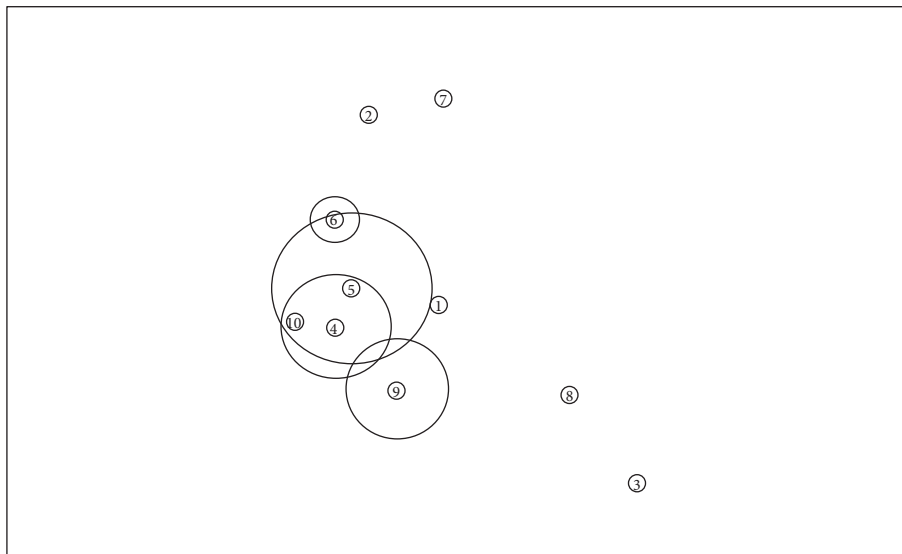


FIGURE 9: Simulation results for one-hundred-node wireless network using the DHT11 sensor.

node network in both sensors' simulations. If a larger number of nodes are required, the network will collapse because the data lost ratio increases. However, it is more important to analyze the data throughput, which is a key aspect of successful communication networks. If the throughput is high, the network is considered to be working successfully. However, the simulations indicate that

throughput decreased with the number of nodes, which is harmful to the network.

Precision agriculture is a vast domain with several sensors used to measure many aspects. However, the simulations only considered two sensors to determine the issues that may occur when establishing a network. In the experiments, all the issues were from the network, such as

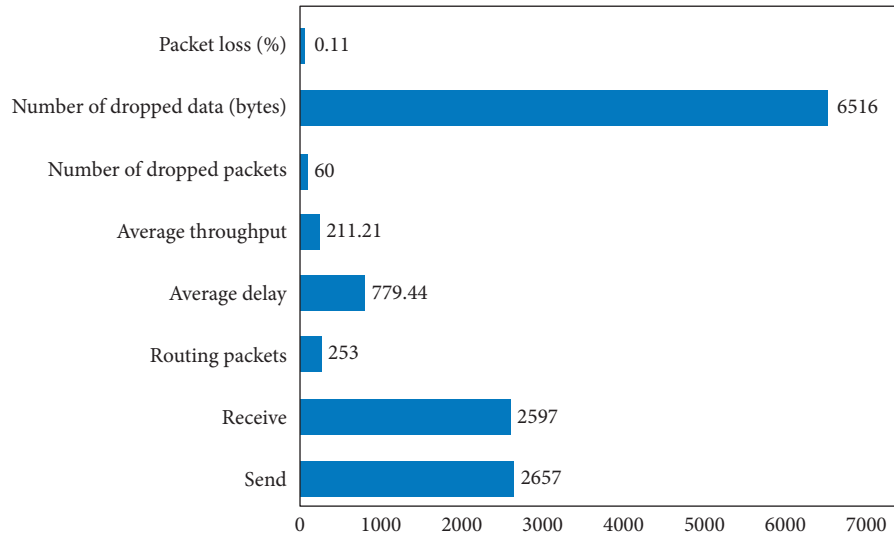


FIGURE 10: DHT11 sensor with one hundred nodes.

packet loss, packet drops, decreased throughput, and decreased data delay.

5. Conclusion

A WSN, also called a wireless sensor or WSAN, includes wrapped sensors that screen physical phenomena or steady conditions to monitor phenomena such as temperature, sound, and weight and pushes their measurements through the network to a focal area. Most existing networks are bidirectional to enable sensor control. The development of WSNs includes two or three of the following: (1) data for air collection, harvesting, and soil, (2) monitoring coursed zones, (3) various yields from an area, (4) different fertilizers and water fundamentals to various parts of an uneven zone, (5) diverse necessities of harvests for various air and soil conditions, and (6) proactive strategies instead of responsive blueprints. Sensors have been around for some time in various networks. The rule indoor controller was made in 1883, and many consider this as the basis for current assembled sensors. Infrared sensors have been around since the late 1940s and have become ubiquitous in recent years. Progression markers have also been utilized for many years. In this study, NS₂ simulator was used for analyzing the results which have better environment for evaluation and experiments. We set up a virtual environment. We used different topologies for different scenarios to get better numbers. The simulation results are from two different sensor types that have differing outputs for a greater number of nodes in the network. It is seen that as the number of nodes increased, the packet loss ratio increased, and the throughput decreased. Analyzing the procedures in precision agriculture indicates there are no problems with the sensors, but there are issues in the network at the back end. In the first simulation, there are only ten nodes. The output of the simulations suggests that a small number of packets have a greater percent loss of packets. However, the results of the fourth scenario are quite different, as shown in the

graphical representation of the one-hundred-node network in both sensor simulations. Only network issues in precision agriculture are discussed here. Packet loss in all scenarios is low but loss of data is not better for communication. Delay must be decreased in the future, but the throughput should be increased. There are many opportunities for future research to analyze the algorithms to overcome these issues to determine sensor functionality concerns.

Data Availability

The datasets used for this study can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was supported by Taif University Researchers Supporting Project (number TURSP-2020/254), Taif University, Taif, Saudi Arabia.

References

- [1] Aqeel-ur-Rehman and Z. A. Shaikh, *Smart Agriculture, Application of Modern High Performance Networks*, Bentham Science Publishers Ltd., Sharjah, UAE, 2009.
- [2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, no. 4, pp. 393–422, 2002.
- [3] J. Kang, Y. Zhang, and B. Nath, "End-to-end channel capacity measurement for congestion control in sensor networks," in *Proceedings of the 2nd International Workshop on Sensor and Actor Network Protocols and Applications (SANPA)*, Boston, MA, USA, 2004.
- [4] D. Chen and P. K. Varshne, "QoS support in wireless sensor networks: a survey," in *in Proceedings of the International*

- Conference on Wireless Networks*, pp. 1–7, vol. 233, no. 1, Las Vegas, NV, USA, June 2004.
- [5] S. Tanwar, N. Kumar, and J. J. Rodrigues, “A systematic review on heterogeneous routing protocols for wireless sensor network,” *Journal of Network and Computer Applications*, vol. 53, pp. 39–56, 2015.
- [6] O. Mahfooz, M. Memon, and J. Poncela, “Review on use of wireless sensor network to overcome agricultural problems of Pakistan,” *Pakistan Journal of Engineering, Technology & Science*, vol. 5, no. 1, 2016.
- [7] P. Abouzar, D. G. Michelson, and Maziyar Hamdi, “RSSI-based distributed self-localization for wireless sensor networks used in precision agriculture,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 6638–6650, 2016.
- [8] E. Praynlin and J. Ida Jenson, “Solar radiation forecasting using artificial neural network,” in *Proceedings of the 2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pp. 1–7, IEEE, Vellore, India, April 2017.
- [9] R. Srivastava, A. N. Tiwari, and V. K. Giri, “Forecasting of solar radiation in India using various ANN models,” in *Proceedings of the 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pp. 1–6, Gorakhpur, India, November 2018.

Research Article

Future Location Prediction for Emergency Vehicles Using Big Data: A Case Study of Healthcare Engineering

Muhammad Daud Kamal ¹, Ali Tahir ¹, Muhammad Babar Kamal ²
and M. Asif Naeem ^{3,4}

¹Institute of Geographical Information Systems, National University of Sciences and Technology, Islamabad, Pakistan

²Department of Computer Science, COMSATS University, Islamabad, Pakistan

³Department of Computer Science, National University of Computer and Emerging Sciences (NUCES), Islamabad, Pakistan

⁴School of Engineering, Computer & Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand

Correspondence should be addressed to Ali Tahir; ali.tahir@igis.nust.edu.pk

Received 7 October 2020; Revised 6 November 2020; Accepted 17 November 2020; Published 28 November 2020

Academic Editor: Shaukat Ali

Copyright © 2020 Muhammad Daud Kamal 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 number of devices equipped with GPS sensors has increased enormously, which generates a massive amount of data. To analyse this huge data for various applications is still challenging. One such application is to predict the future location of an ambulance in the healthcare system based on its previous locations. For example, many smart city applications rely on user movement and location prediction like SnapTrends and Geofeedia. There are many models and algorithms which help predict the future location with high probabilities. However, in terms of efficiency and accuracy, the existing algorithms are still improving. In this study, a novel algorithm, NextSTMove, is proposed according to the available dataset which results in lower latency and higher probability. Apache Spark, a big data platform, was used for reducing the processing time and efficiently managing computing resources. The algorithm achieved 75% to 85% accuracy and in some cases 100% accuracy, where the users do not change their daily routine frequently. After comparing the prediction results of our algorithm, it was experimentally found that it predicts processes up to 300% faster than traditional algorithms. NextSTMove is therefore compared with and without Apache Spark and can help in finding useful knowledge for healthcare medical information systems and other data analytics related solutions especially healthcare engineering.

1. Introduction

Analysing the movement pattern has always been of a keen area of interest, may it be automobile, humans, or any other moving object. These movement patterns can help analysts in making a decision related to the behaviour patterns of an object. For example, the idea of geo-marketing can be evolved if the pattern of the people who are shopping is observed.

Similarly, different location-aware applications can help urban planning by observing traffic patterns. Approximately 3.5 billion mobile phone users are predicted worldwide in 2020 [1]. A mobile user location is better estimated these days by the techniques that are currently being developed

and used by the telecommunication providers. Therefore, mobile user's patterns and activities are sensed by using different mobility data records that are saved by telecommunication companies [2].

The objective of observing mobility data is to see why and when the objects move. To accomplish these objectives, various data sources are used such as Global System for Mobile communications (GSM) or Global Positioning System (GPS) for analysing and later transforming them into meaningful predicting patterns. The process of predicting patterns is known as Knowledge Discovery (KD), i.e., produced from raw data and converted into meaningful knowledge [3]. A hybrid system for location recognition and prediction which addressed key issues of location-based

services, such as location recognition and prediction, was proposed by [4]. The system used a hybrid method combining k -Nearest Neighbour (kNN) and decision tree to effectively recognize the locations not only in the outdoor environment but also in the indoor environment. NextPlace was presented by [5], an approach for spatio-temporal user location prediction based on nonlinear analysis of the time series of start times and duration times of visits to significant locations. This approach allows forecasting not only the next location of a user but also his/her arrival and residence time, i.e., the interval of time spent in that location. With a particular objective to settle on an informed decision concerning which advancements to understand, information was assembled from a few previous literature studies.

The main motivation of this research is the availability of massive data with the industry which was never utilized for business intelligence in the context of Pakistan. The data has been gathered over the years and only used for real-time monitoring of vehicles. There is a huge potential to explore and perform geo visual analytics on available data on big data platforms such as Apache Spark.

The main purpose of this research is to analyse the spatio-temporal mobility patterns of Global Positioning System (GPS) data using new technologies for big data; i.e., Apache Spark is used to reduce the time taken per job for discovering useful information, which can help assist decision-making for real-world scenarios. The future location of vehicles is predicted from a large pool of data with more than 100 million rows of records after developing a novel algorithm, Next Spatio-Temporal Move (NextST-Move), on Apache Spark to optimize the time taken and later the predicted locations are verified against the real data.

The main contribution of our research is our newly proposed NextSTMove algorithm which is more efficient and accurate than existing algorithms. Moreover, we have used the real data of a local tracker company. The results of our algorithm can be very useful for long-term strategic and business advantages in healthcare engineering.

The remainder of this paper is structured as follows.

Section 2 describes the related literature review. Section 3 presents a detailed methodology and proposed algorithm. Results and discussion are presented in Section 4. Finally, conclusions are outlined in Section 5.

2. Related Work

There are many trajectory prediction algorithms that exist in the literature. Over the years, various researchers have proposed novel algorithms which cater to their needs. Broadly, trajectory prediction algorithms are derived from machine learning approaches such as Bayesian networks [6], hidden Markov models [7], decision trees [8], neural networks [9], and state predictor methods [10]. This section describes existing work in this field while commenting on the above-mentioned parameters.

Research in mobility data is not that new. However, in the last few years, it has gained popularity for data mining and artificial intelligence, and health engineering [11–13]. Substantial amounts of information are produced by GPS

and telecommunication technologies advancement. In the survey paper [14], five algorithms are used for four users that had different patterns. Innovations and advancement are giving hints of producing pervasive computing for mobility data which helps predict the accuracy. The trajectories that are stored for the semantics of mobility data are aiding in finding useful information about the movements of the objects [15].

Likewise, paper [16] presents visual techniques to generate trajectories (spatio-temporal sequences) using GPS data to assist in efficient trajectory projection of emergency vehicles in highly urbanized cities. Furthermore, papers [17, 18] use visual analysis to implement intelligent transportation enabling efficient utilization of new knowledge and complex data.

A spatial-temporal prediction method was proposed by [19] which is called Spatial-Temporal Recurrent Neural Networks (STRNN). The experimental results on real datasets showed that STRNN outperformed the state-of-the-art methods and can well model the spatial and temporal contexts. In [20, 21], authors discussed that with the growing data volume arises a need for processing spatio-temporal queries efficiently. For this, they used parallel processing in Secondo for geospatial big data analysis, while in [22] the context of time and space in a massive geospatial big data database is analysed using High-Performance Computing (HPC). A classification was presented by [23] for approaching decision trees to predict the next place of mobile users. The authors implemented an optimizer to find the best parameter combination for each user since users had widely varying behaviour. Finally, the performance of the approach was demonstrated by the results of the experiments on the real-life dataset of 80 mobile users provided by Nokia. The existing solutions for geolocation prediction (GP) and divided geolocation prediction into two primary parts were reviewed by [24]. The initial step proposed to manufacture a geolocation expectation show is Mining Popular Geolocation Region (MPGR), and the second is Mining Personal Trajectory (MPT). The results described the basic concepts of GP, the characteristics of MPGR, and MPT. They also discussed the limitations, openings, and future geolocation prediction analytical trends for mobility big data. Similarly, paper [25] proposed a methodology for the prediction of a user's outdoor location derived from contextual data (current location, day of the week, time, and speed), which were collected with a GPS device and with a smartphone. This methodology was based on spatial clustering of data and on-time segmentation to find points of interest that the user visits every day and every hour.

An investigation in 2013 by [26] worked on the perspectives identified with data accumulation and taking care of trajectories that are feeding to the databases with proper data. The trajectories recreation for producing meaningful trajectories includes procedures for gathering movement data and cleansing the data gathered, compression of data, and map coordinates to deliver noise-free trajectories. For the production of semantically compliant trajectories, raw spatial data from the common repository need to be recovered using different remaking tasks along with semantic

trajectories. Moreover, paper [27] defined the concept behind the management of trajectory and their representations. The focus of the research was analysis on an extensive scale for phenomena related to mobility with more focus on the semantic behaviour of the data. The main goal of analysing the behaviour is indicating which behaviour defines which moving object.

An unprecedented amount of geospatial data gathered from moving objects defies human capability to analyse it. A study by [28] found new methods for processing and mining moving objects. For modelling and representing trajectories, paper [29] discusses the problem in the context of database systems. Moving objects databases represent a set of moving objects using abstract data types and maintain complete histories of movement.

An open-source software, Secondo, has a framework for big trajectory data whose data model is not fixed. This Database Management Systems (DBMS) prototype can be used for different data models. "WhereNext" are previously visited trajectory patterns that were extracted [30] that use previously extracted trajectory patterns. In most of the studies, few aspects of trajectory prediction are discussed. For example, some studies focus on indoor and outdoor navigation. Similarly, other studies highlight public and private datasets. In many studies, the authors have validated the accuracy of these algorithms on given datasets. In our research, we proposed a novel algorithm, NextSTMov, using Apache Spark to minimize the query and processing time for GPS big data. The reason is because Apache Spark is becoming de facto for processing big data in the computing world. We used it to predict future locations of vehicle GPS data.

Bayes-based predictors were used to add to the performance of their prediction for leveraging big data [31]. They studied a large Call Detail Record (CDR) dataset. At first, they explored the dataset and found that they can use call activity to generate prior probabilities for use in a Bayes predictor. With this reasoning, they developed an enhanced Bayes predictor that uses a distance threshold and the users' regular location to improve the generation of prior probabilities. Experimental results show that the enhancements they proposed increase accuracy of the Bayes based predictor by 17 percentage points. In the end, they concluded that it is feasible to leverage big cellular data to enhance location predictors without relying on external data.

Apache Spark developed in the year 2009 at Berkeley's lab [32] is said to have achieved the lowest latency rate in comparison with Secondo and Parallel Secondo. It is freely available for several operating systems such as Windows, Linux, and Mac Operating systems. Apache Spark is a Unified Analytics Engine for big data processing and management that supports streaming data, batched data, SQL, Graph, and machine learning processes. Apache Spark for point cloud spatial data management was used to achieve a lower latency rate [33]. They found, in comparison to the traditional methods for point cloud management, a file system storage, a single processing server, and a distributed approach based on Apache Spark were able to achieve a more agile speed and higher robustness

and fault tolerance support. A comparison was made between the processing time taken by Relational Database Management System (PostgreSQL) and the time taken using Apache Spark. They achieved up to 300% of reduced latency rate, which shows Apache Spark is faster compared to these DBMS. As the number of nodes in the cluster was increased, the processing capabilities of the system increased. Increasing the number of points did not affect the query execution time of Apache Spark much, whereas queries run over PostgreSQL slow almost immediately. A new platform for geospatial big data was developed by [34] inside Apache Spark a GeoSpark SQL framework that was able to carry out geospatial SQL queries over an Apache Spark system.

The research showed that Apache Spark has a better performance than traditional Relational Database Management Systems (RDBMS) for a huge number of geospatial type queries. The methods for inserting the point data into the Apache Spark data structure are represented in [33]. The data were sliced into rectangular areas and each area was ingested in a separate document. Rectangular areas were numbered by a Geohash system and were stored in MongoDB. These structures allowed executions of operations by MapReduce for point cloud data, sometimes MongoDB or from an external framework like Apache Hadoop [35]. Similarly, paper [36] compared quadtree and R-tree on Spark for finding the difference in the query efficiency.

The purpose of this research is to analyse the spatial-temporal mobility patterns of GPS data using Apache Spark to reduce the time taken per job for discovering useful information, which can support the decision-making process for real-world problems.

3. Materials and Methods

3.1. Overview. In general, Apache Spark software is used for clustering of systems for very fast query response. It provides an executable environment for all the Spark applications in the Kernel of Spark core. The actual advantage of Apache Spark is that, compared with other technologies like Hadoop and MapReduce which only use disk for memory, Apache Spark uses memories and can also make use of the disk for the processes. Apache Spark is versatile, unlike the Hadoop ecosystem, as it does not have its own distributed file system but can make use of Hadoop Distributed File System (HDFS).

Apache Spark is a standalone software that does not use any resource manager. However, if we use it for more than one node and environment setup, we can use Yet Another Resource Negotiator (YARN) or Multiple Equivalent Simultaneous Offers (MESOs) for resource management, along with a distributed file system such as HDFS or Amazon Simple Storage Service (S3).

3.2. Spark SQL. Spark has a built-in library for processing structured data. This can be used for complicated SQL database queries and algorithm-based analytics. Spark SQL

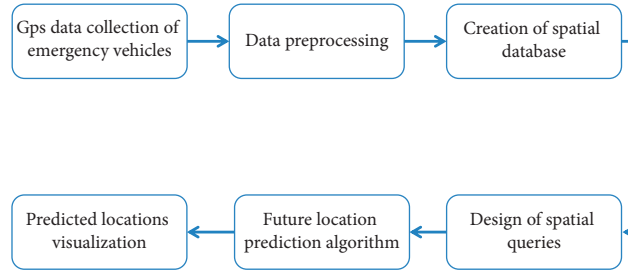


FIGURE 1: Detailed flowchart of methodology.

TABLE 1: Raw data of a vehicle.

Veh. no.	Place of visit	Latitude	Longitude	Date	Day	Time
A	Subway Blue Area	33.71146667	73.0577	6/1/2016	Wednesday	00:00
A	Subway Blue Area	33.71146667	73.0577	6/1/2016	Wednesday	00:14
A	Jinnah super franchise	33.70936667	73.05383333	6/1/2016	Wednesday	00:21
A	Pak printing press	33.68015	73.07563333	6/1/2016	Wednesday	00:26
A	Shell quick fill F/Station	33.64775	73.0999	6/1/2016	Wednesday	00:31
A	Total Parco petrol diesel gas station	33.61301667	73.12583333	6/1/2016	Wednesday	00:36
A	TCS office	33.60621667	73.11566667	6/1/2016	Wednesday	00:41
A	Islamabad/Pindi airport	33.60576667	73.09916667	6/1/2016	Wednesday	00:46
A	Islamabad/Pindi airport	33.60641667	73.09966667	6/1/2016	Wednesday	00:51
A	Islamabad/Pindi airport	33.60628333	73.0989	6/1/2016	Wednesday	00:55

```

while  $i \leftarrow$  unique locations do
  total count  $\leftarrow$  0
  while  $j \leftarrow$  all locations do
    IF ( $j == i$ )
      total count  $\leftarrow$  total count + 1
      current probability  $\leftarrow$  total count divide by count of all the locations and multiply by hundred
    END IF
  end
  IF (current probability > first probability)
    third probability  $\leftarrow$  second probability
    second probability  $\leftarrow$  first probability
    first probability  $\leftarrow$  current probability
    first location  $\leftarrow i$ 
  ELSEIF (current probability > second probability)
    third probability  $\leftarrow$  second probability
    second probability  $\leftarrow$  current probability
    second location  $\leftarrow i$ 
  ELSEIF (current probability > third probability)
    third probability  $\leftarrow$  current probability
    third location  $\leftarrow i$ 
  END IF
  IF (third probability = sys.maxsize)
    third location  $\leftarrow i$ 
  END IF

```

ALGORITHM 1: NextSTMove: algorithm for predicting the top three locations.

supports HIVE, SQL-like HiveQL query, Java Database Connectivity (JDBC), and Open Database Connectivity (ODBC). This can also enable some degree of connections with existing databases, warehouses, and business intelligence environments.

3.3. *Deployment of Apache Spark.* NextSTMove for predicting the future location of vehicles using Python programming is designed and implemented. PySpark utility was installed for Windows 7 using PIP. PySpark was locally installed in the system.

```

IF (third probability – sys.maxsize)
dataframe ← Select locationname,
Latitude, Longitude from TwelveMonthGPSData where locationName is equal to third location
thirdlatitude ← dataframe.first().Latitude
third longitude ← dataframe.first().Longitude

print ( “Probability for Third location is ‘%s’ is %s and Latitude %s Longitude %s”%
( third location, third probability, third latitude, third longitude ) )

(third location, third probability, third latitude, third longitude )
ENDIF
ELSE IF (second probability – sys.maxsize)
dataframe ← Select location name,
Latitude, Longitude from TwelveMonthGPSData where locationName
← second location;
second latitude ← dataframe.first().Latitude;
second longitude ← dataframe.first().Longitude;

print ( “Probability for second location is ‘%s’ is %s and Latitude: %s Longitude:
%s”% ( second location, second probability, second latitude, second longitude ) )

END IF( first probability – sys.maxsize)
dataframe ← Select location name,
Latitude, Longitude from TwelveMonthGPSData where locationName
← first location;
first latitude ← dataframe.first().Latitude;
first longitude ← dataframe.first().Longitude

print ( “Probability for first location is ‘%s’ is %s and Latitude: %s Longitude:
%s”% ( first location, first probability, first latitude, first longitude ) )

ELSE
print (“No Records for the above query!”)
ENDIF
    
```

ALGORITHM 2: NextSTMove: location extraction from user’s data.

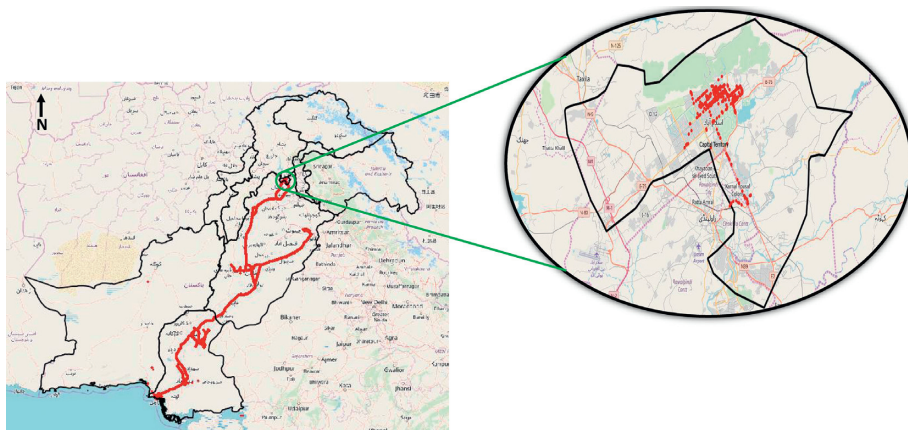


FIGURE 2: One month (January) data across Pakistan.

TABLE 2: Queries for predicting the top three locations.

Query no.	Vehicle no.	Day	Time from	Time to
Query 1	User A	Wednesday	12:00:00	01:00:00
Query 2	User B	Monday	09:00:00	10:00:00
Query 3	User B	Tuesday	09:00:00	10:00:00

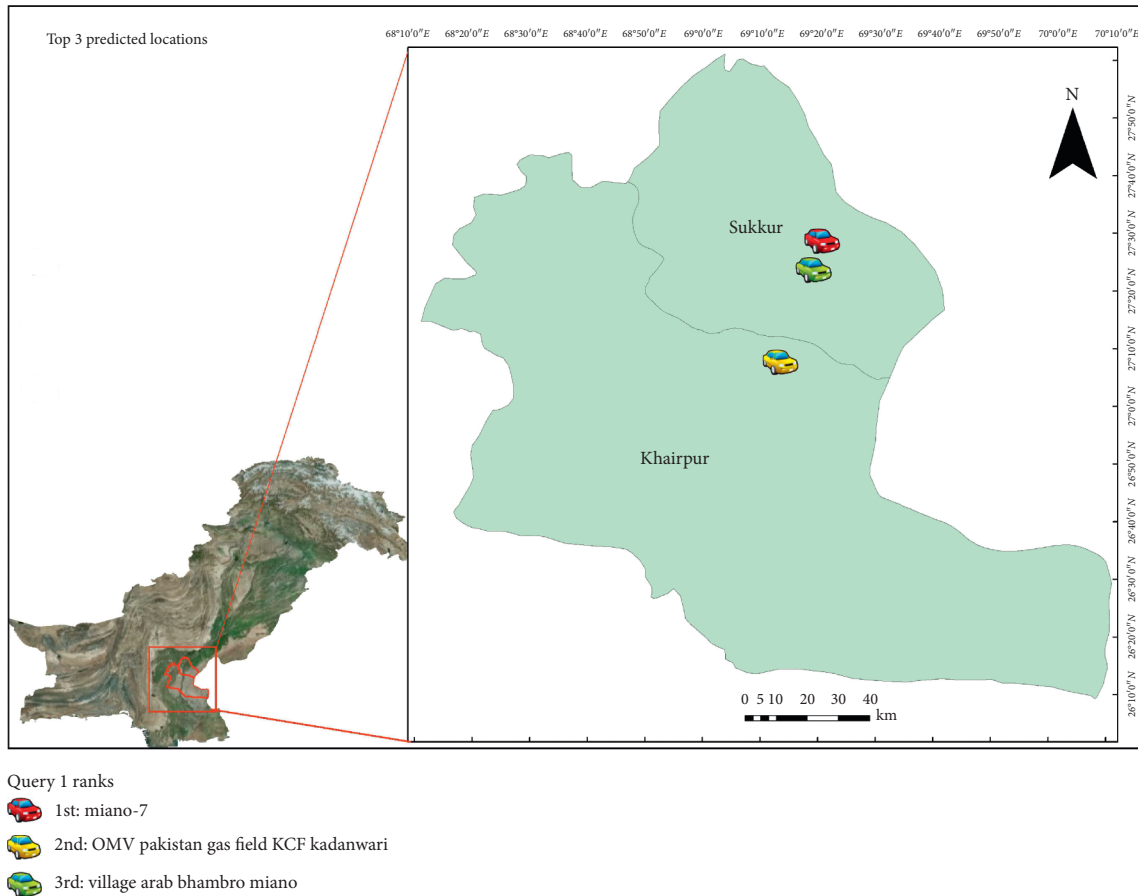


FIGURE 3: Results of query 1.

3.4. Approach. In this section, we describe the approach used for this research. The approach is divided into data collection, data pre-processing, creation of the spatio-temporal database in Apache Spark, creation of locations from user's data, and spatio-temporal queries. A flowchart of the methodology carried out is shown in Figure 1.

3.5. Raw GPS Data Collection. Real-time data from a vehicle GPS tracker company is used for this research.

The GPS data was received in MS-SQL database. The data spanned from 1st January 2016 and ended on 31st December 2016 with a total of 105,096,953 records in all twelve tables for each month of the year 2016. A total of 2261 vehicles contributed to the data. Table 1 shows the data of the anonymized vehicle.

3.6. Data Pre-Processing. Data pre-processing involved cleaning of data, removal of unwanted data fields, and removal of missing fields to avoid null data in columns and adding new columns. Unlike [29] where the authors used an algorithm kNN to identify the latitudes and longitudes that are associated with each other, the data received were already assigned location names to a cluster of latitudes and longitudes which were quite accurate.

The GPS data received had thirty-three columns and most of them were not useful for the algorithm. To extract only the useful five columns from all the twelve tables of each month from the MS-SQL database to a single Comma Separated Values (CSV) file, the following batch command query was used on windows command-line environment:

```

bc "SELECT ReportGroupDate, vehicleRegistrationNo,
locationName, latitude, longitude FROM GPSDataMonth1
union "SELECT ReportGroupDate, vehicleRegistrationNo,
locationName, latitude, longitude FROM GPSDataMonth2
union "SELECT ReportGroupDate, vehicleRegistrationNo,
locationName, latitude, longitude FROM GPSDataMonth3
...
union "SELECT ReportGroupDate, vehicleRegistrationNo,
locationName, latitude, longitude FROM GPSDataMonth12
queryout E:/TwelveMonthsTablesData.csv -t, -c -S. -d
ReceievedGPSDatabase --T

```

The data from each table of each month is now stored in one file and the total size of the data is reduced from 30 GB to 10 GB. There are more than 100 million records used in the Apache Spark database. Data within the CSV is arranged in the following sequence:

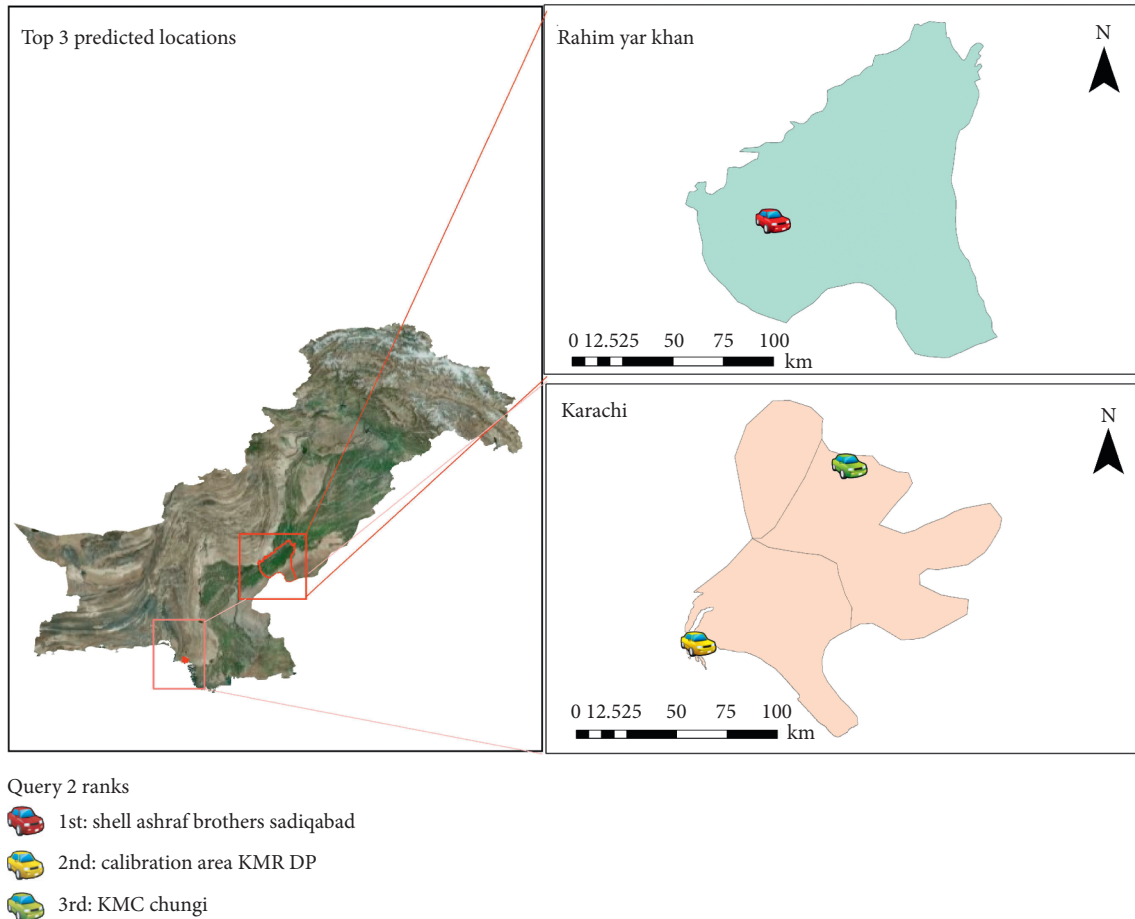


FIGURE 4: Results of query 2.

Date, Time, Day, vehicleRegistrationNo, LocationName, Latitude, Longitude

3.7. Creation of Spatial Database in Spark. A new database was created and then the data was stored in the Spark SQL table. PySpark syntax was used for storing the data in the Spark database. We used Spark Context to define the cores that are to be used by the local system. The Python command we used for this purpose is as follows:

```
sc = SparkContext ("local[*]", "User")
spark = SparkSession.builder
.master("local")
.appName ("Data cleaning")
.getOrCreate ()
```

We used DataFrames in Apache Spark version 2.0 Application Programming Interface (API) for managing our data. The final CSV file was generated for all the twelve months of the year and has been pre-processed and was assigned to a DataFrame using the following lines of code:

```
SparkDataFrame = spark.read.format ("csv").option
("header", "true").option
("mode", "DROPMALFORMED").load ("E:
TwelveMonthsTablesData.csv")
```

```
SparkDataFrame.createOrReplaceTempView
("TwelveMonthGPSData")
```

Here, SparkDataFrame is the DataFrame we are using and Spark's API to read the CSV file after loading it. A spatio-temporal database was created by using the createOrReplaceTempView library of Apache Spark. The 'SparkDataFrame' was stored in our Spark SQL table which is used for NextSTMOver Algorithm.

3.8. Design of Spatial Queries. As the focus of this work is future location prediction using Apache Spark, the design of the query included the spatio-temporal aspect of the data, i.e., where and when. We queried for where a used vehicle will be at a given time and on a given day, for example, the location of a particular user, for example, "User A" on "Monday" between "9 Am" to "9:30 Am."

We asked the user to input a valid vehicle number, the day they want to inquire, and the time between which they want to predict the vehicle location. After the user has input

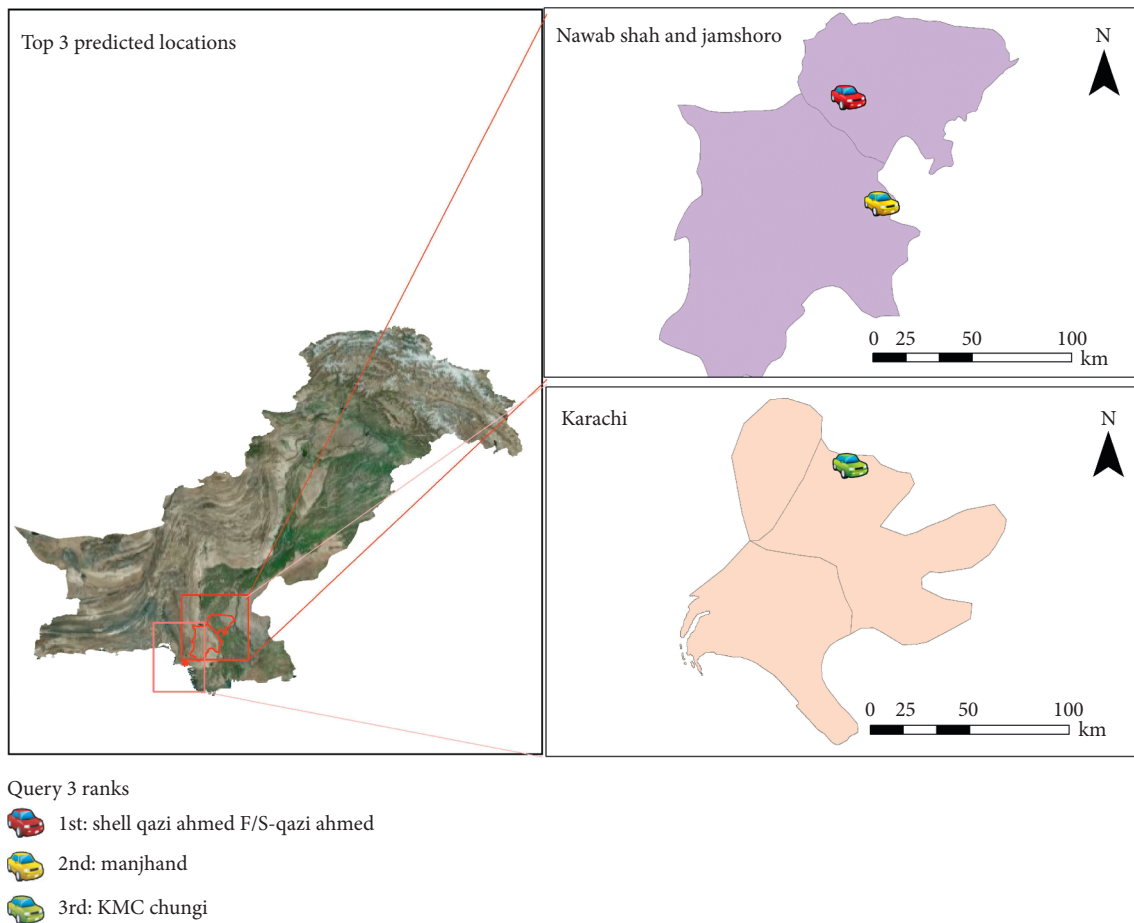


FIGURE 5: Results of query 3.

these parameters, a query gives the necessary results. The Spark SQL query is as follows:

```
SparkDataFrame = spark.sql ("SELECT * from Twelve-MonthGPSData where vehicleRegistrationNo = '\%s' AND Day = '\%s' AND (Time = '\%s' or Time = '\%s')" \% (vehicleRegistrationNo, Day, TimeWindow1, TimeWindow2))
```

3.9. Future Location Prediction Algorithm. As the DataFrame is updated and records all locations for the requested query. Each time a new location comes, it is considered as a key and their repetition is considered as tokens. The keys are checked against the total locations present in the DataFrame. Every time a key is repeated, respective tokens are also incremented. At the end of this nested loop, the count of each key is stored in their respective tokens. The top three keys having the most tokens are considered as the top three probable locations. These top three locations are stored for further processing.

Algorithm 1 explains the steps:

After the successful iterations of the above algorithm, three locations and the probability of their occurrences are output in the algorithm.

3.10. Creation of Location from Users Data. The results from the algorithm were visualized on the web using geospatial visualization libraries of Python. The result was extracted for web maps on run time to avoid any delays in data generation. The findings are discussed in Section 4. Algorithm 2 explains the logical steps involved:

The coordinates of the top three locations along with their names and probabilities are ready to be mapped. Folium library in Python is used to generate a web map.

4. Results and Discussion

Figure 2 illustrates all the vehicles visited in one month. The visualization seems cluttered. Therefore, a further zoomed-in view on the city of a single user for the 1st month can also be seen.

4.1. Top Predicted Locations. The queries were applied to the data to generate the top three probable locations between two-time intervals. Table 2 shows the queries while Figures 3–5 display their results. The predictions of queries 2 and 3 are the same; however, the location predictions are quite apart because user B may be a frequent visitor of

TABLE 3: Queries with and without using Apache Spark.

Query no.	Vehicle no.	Day	Time from	Time to
Query 1	User A	Monday	10:00:00	12:00:00
Query 2	User B	Tuesday	11:00:00	13:00:00
Query 3	User B	Wednesday	12:00:00	14:00:00
Query 4	User D	Thursday	13:00:00	15:00:00
Query 5	User E	Friday	14:00:00	16:00:00

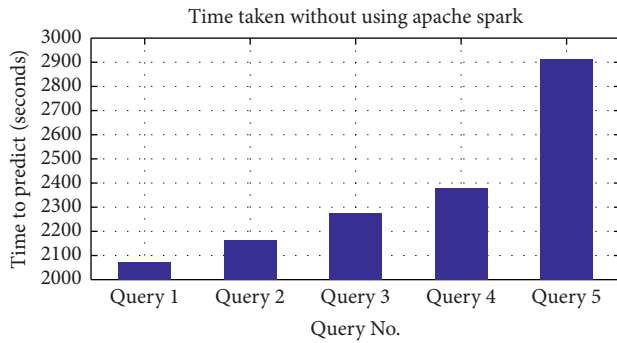


FIGURE 6: Time taken without using Apache Spark.

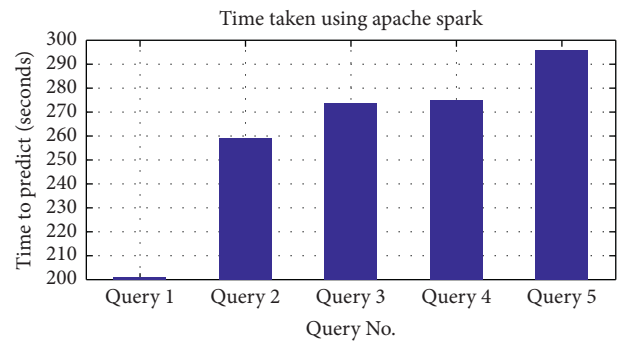


FIGURE 7: Time taken using Apache Spark.

Karachi and interior province to provide ambulance services to the larger city from underdeveloped areas.

4.2. Latency for Apache Spark. A comparison is carried out on algorithms with and without Apache Spark. Initially, the algorithm was designed using a simple Python library. The NextSTMov algorithm was developed using queries and the latency of the process was calculated. The algorithm was then developed on Apache Spark using PySpark and queries applied are shown in Table 3. Up to 200 queries were applied and a sample of five random queries is shown in Table 3.

After this, the algorithm was developed using Apache Spark for the queries shown in Table 3. We achieved a remarkable amount of decrease in the time taken by the queries. The job took more than two thousand seconds without using Apache Spark as illustrated in Figure 6. After using Apache Spark, the queries took less than 300 seconds. Figure 7 illustrates the time taken in detail.

4.3. Accuracy of Predicted Locations. Six months were used to predict the future locations of users and then the data from the next six months were used to find the accuracy of the predicted locations. We compared the real-time locations from the next six months' data with the predicted output for the queries. Table 3 shows the queries whose accuracy percentage is illustrated in Figure 8.

The three bars for each query in Figure 8 show the accuracy of the top three locations that were queried. As shown in Figure 8, query 4 has achieved a 100% accuracy, the reason being that this user has not changed his pattern for that time of the day. Therefore, the algorithm predicted it accurately. Similarly, in query 3 the algorithm achieves an accuracy of 90% for the top 2nd and 3rd predicted location, which depicts that this user was mostly using the same route, or he/she was present at the same location mostly.

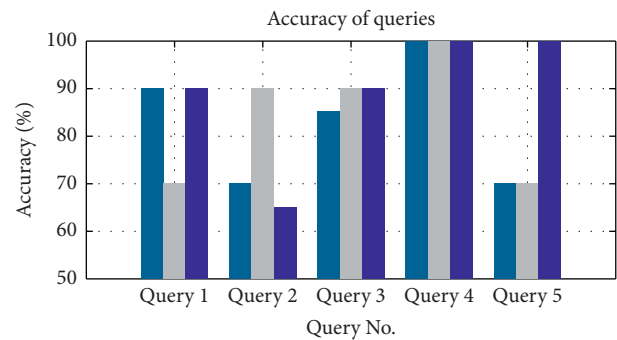


FIGURE 8: Accuracy of queries. Three predicted locations are shown for each query where the first, second, and third bars represent the first, second, and third locations, respectively.

The top 1st location for query 3 has a prediction rate of 85% which means the user showed varied behaviour.

5. Conclusions

Our work presents a novel algorithm, NextSTMov, where vehicle future movements are successfully predicted with and without Apache Spark. The algorithm achieved 75% to 85% accuracy and in typical cases 100% accuracy, where the users follow a repetitive pattern. The main aim of this research was to improve the latency and efficiency as compared to existing algorithms such as NextPlace [5]. Apache Spark, a big data platform, was fully utilized to achieve this. The algorithm reduced the processing time to up to 300%. This processing was done on a total of 2261 users having approximately 100 million data points.

This study is significant in predicting future locations of emergency vehicles. This can facilitate users to perform spatial tasks while improving the analytical knowledge gained from understanding their behaviours. The emergency

vehicle tracker data reveal their spatio-temporal patterns. This research work can also help in solving many geospatial big data applications from both a commercial and security viewpoint.

As part of a future road map, we plan to expand our work by including real-time streaming of big data instead of processing with batched data only. Furthermore, we plan to introduce more nodes to the distributed processing to enhance the efficiency of the system running the queries. More data attributes can be introduced to analyse additional information which can reveal meaningful information and patterns for real-time applications.

Further analysis can be carried out in answering the question of how and why a user visited a particular location. This can help find the semantics of trajectories and carry out their analysis. Similarly, another future area in the algorithm can be predicting the next location of a vehicle using its previous history, i.e., where a user will be next after a specific location, by making a system to predict a route for vehicles that will be congested for a specific time and ask the emergency vehicles if they want to avoid that road. This study opens up further avenues for research. The main concern for using Apache Spark for NextSTMovE is that during the loading of queries the first query takes more time to process as compared to the rest of the queries. Also, Apache Spark gets batched data, while other platforms such as Apache Flink can work with streaming data as well. Therefore, to increase processing capabilities, streaming data processing can be embedded along with it as part of our future work.

Data Availability

Some sample data of a few vehicles might be provided on request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] News, I.67+ Revealing Smartphone Statistics for 2020, 2020.
- [2] F. Giannotti and D. Pedreschi, "Mobility, data mining and privacy: geographic knowledge discovery," Springer, Berlin, Germany, 2008.
- [3] C. Renso, S. Spaccapietra, and E. Zimányi, *Mobility Data*, Cambridge University Press, Cambridge, UK, 2013.
- [4] S.-B. Cho, "Exploiting machine learning techniques for location recognition and prediction with smartphone logs," *Neurocomputing*, vol. 176, pp. 98–106, 2016.
- [5] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, "Nextplace: a spatio-temporal prediction framework for pervasive systems," in *Proceedings of the International Conference on Pervasive Computing*, pp. 152–169, San Francisco, CA, USA, June 2011.
- [6] F. V. Jensen, *An introduction to Bayesian Networks*, Vol. 210, UCL Press, London, UK, 1996.
- [7] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine* 1986, vol. 3, no. 1, pp. 4–16.
- [8] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [9] D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, no. 1, pp. 109–118, 1990.
- [10] J. Garnier, D. J. Osguthorpe, and B. Robson, "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins," *Journal of Molecular Biology*, vol. 120, pp. 97–120, 1978.
- [11] M. C. Chyu, T. Austin, F. Calisir et al., "Healthcare engineering defined: a white paper," *Journal of Healthcare Engineering*, vol. 6, Article ID 724584, 14 pages, 2015.
- [12] A. Sharma and R. Kumar, "Service-level agreement—energy cooperative quickest ambulance routing for critical healthcare services," *Arabian Journal for Science and Engineering* 2019, vol. 44, pp. 3831–3848.
- [13] M. A. Usman, N. Y. Philip, and C. Politis, "5G enabled mobile healthcare for ambulances," in *Proceedings of the 2019 IEEE globecom workshops (GC wkshps)*, December 2019.
- [14] J. Petzold, F. Bagci, W. Trumler, and T. Ungerer, "Next Location Prediction within a Smart Office Building," *Cognitive Science Research Paper-University of Sussex CSR2005*, vol. 577, p. 69.
- [15] M. R. Vieira, P. Bakalov, and V. J. Tsotras, "On-line discovery of flock patterns in spatio-temporal data," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 286–295, ACM, Seattle, WA, USA, November 2009.
- [16] U. Maqsood, A. Tahir, K. Fatima et al., "Interpreting rescue vehicle patterns using geovisual analytics for spatiotemporal resource allocation," *Arabian Journal of Geosciences*, vol. 13, p. 660, 2020.
- [17] G. Andrienko, N. Andrienko, W. Chen, R. Maciejewski, and Y. Zhao, "Visual analytics of mobility and transportation: state of the art and further research directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 2232–2249, 2017.
- [18] N. Andrienko, T. Lammarsch, G. Andrienko et al., "Viewing visual analytics as model building," *Computer Graphics Forum*, vol. 37, no. 6, pp. 275–299, 2018.
- [19] Q. Liu, S. Wu, L. Wang, and T. Tan, "Predicting the next location: a recurrent model with spatial and temporal contexts," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, Phoenix, AZ, USA, February 2016.
- [20] C. Zhou, F. Su, F. Harvey, and J. Xu, *Spatial Data Handling in Big Data Era*, Springer, Beijing, China, 2016.
- [21] M. Ashfaq, A. Tahir, F. M. Orakzai, G. McArdle, and M. Bertolotto, "Using T-Drive and BerlinMod in parallel SECONDO for performance evaluation of geospatial big data processing," in *Spatial data handling in big data era*, pp. 3–19, Springer, Singapore, 2017.
- [22] Z. Li, "Geospatial Big Data Handling with High Performance Computing: Current Approaches and Future Directions," in *High Performance Computing For Geospatial Applications*, pp. 53–76, Springer, Berlin, Germany, 2020.
- [23] L. H. Tran, M. Catasta, L. K. McDowell, and K. Aberer, "Next place prediction using mobile data," in *Proceedings of the Mobile Data Challenge Workshop (MDC 2012)*, Ulm, Germany, June 2012.
- [24] G. Xu, S. Gao, M. Daneshmand, C. Wang, and Y. Liu, "A survey for mobility big data analytics for geolocation prediction," *IEEE Wireless Communications*, vol. 24, pp. 111–119, 2016.

- [25] D. Guessoum, M. Miraoui, and C. Tadj, "Contextual location prediction using spatio-temporal clustering," *International Journal of Pervasive Computing and Communications*, vol. 12, no. 3, pp. 290–309, 2016.
- [26] G. Marketos, M. L. Damiani, N. Pelekis, Y. Theodoridis, and Z. Yan, *Trajectory Collection and Reconstruction*, Cambridge University Press, Cambridge, UK, 2013.
- [27] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot, "A conceptual view on trajectories," *Data & Knowledge Engineering*, vol. 65, no. 1, pp. 126–146, 2008.
- [28] M. Morzy, "Prediction of moving object location based on frequent trajectories," in *Proceedings of the International Symposium on Computer and Information Sciences*, pp. 583–592, Istanbul, Turkey, November 2006.
- [29] R. H. Güting, T. Behr, C. Düntgen et al., "SECONDO: a platform for moving objects database research and for publishing and integrating research implementations," *IEEE Data Engineering Bulletin*, vol. 33, pp. 56–63, 2010.
- [30] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "Wherenext: a location predictor on trajectory pattern mining," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 637–646, ACM, Paris, France, July 2009.
- [31] D. Matekenya, M. Ito, R. Shibasaki, and K. Sezaki, "Enhancing location prediction with big data: evidence from dhaka," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pp. 753–762, ACM, Heidelberg, Germany, September 2016.
- [32] A. Spark, "Mobile subscriptions near the 7-billion mark," *Does Almost Everyone Have a Phone?*, 2019.
- [33] V. Pajić, M. Govedarica, and M. Amović, "Model of point cloud data management system in big data paradigm," *ISPRS International Journal of Geo-Information*, vol. 7, p. 265, 2018.
- [34] Z. Huang, Y. Chen, L. Wan, and X. Peng, "GeoSpark SQL: an effective framework enabling spatial queries on 345 spark," *ISPRS International Journal of Geo-Information*, vol. 6, p. 285, 2017.
- [35] J. Boehm, K. Liu, and C. Alis, "Sideload-ingestion of large point clouds into the apache spark big data engine," in *Proceedings of the XXIII ISPRS Congress*, pp. 343–348, Prague, Czech Republic, July 2016.
- [36] X. Xie, Z. Xiong, X. Hu, G. Zhou, and J. Ni, "On massive spatial data retrieval based on spark," in *Proceedings of the International Conference on Web-Age Information Management*, pp. 200–208, Macau, China, June 2014.

Research Article

Enabling Clustering for Privacy-Aware Data Dissemination Based on Medical Healthcare-IoTs (MH-IoTs) for Wireless Body Area Network

Fasee Ullah,¹ Izhar Ullah,² Atif Khan ,³ M. Irfan Uddin ,⁴ Hashem Alyami,⁵ and Wael Alosaimi⁶

¹Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China

²Institute of Business and Management Sciences, Peshawar, KP, Pakistan

³Department of Computer Science, Islamia College, Peshawar, KP, Pakistan

⁴Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan

⁵Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia

⁶Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia

Correspondence should be addressed to Atif Khan; atifkhan@icp.edu.pk and M. Irfan Uddin; irfanuddin@kust.edu.pk

Received 14 September 2020; Revised 2 November 2020; Accepted 9 November 2020; Published 28 November 2020

Academic Editor: Shah Nazir

Copyright © 2020 Fasee 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.

There is a need to develop an effective data preservation scheme with minimal information loss when the patient's data are shared in public interest for different research activities. Prior studies have devised different approaches for data preservation in healthcare domains; however, there is still room for improvement in the design of an elegant data preservation approach. With that motivation behind, this study has proposed a medical healthcare-IoTs-based infrastructure with restricted access. The infrastructure comprises two algorithms. The first algorithm protects the sensitivity information of a patient with quantifying minimum information loss during the anonymization process. The algorithm has also designed the access policies comprising the public access, doctor access, and the nurse access, to access the sensitivity information of a patient based on the clustering concept. The second suggested algorithm is K -anonymity privacy preservation based on local coding, which is based on cell suppression. This algorithm utilizes a mapping method to classify the data into different regions in such a manner that the data of the same group are placed in the same region. The benefit of using local coding is to restrict third-party users, such as doctors and nurses, when trying to insert incorrect values in order to access real patient data. Efficiency of the proposed algorithm is evaluated against the state-of-the-art algorithm by performing extensive simulations. Simulation results demonstrate benefits of the proposed algorithms in terms of efficient cluster formation in minimum time, minimum information loss, and execution time for data dissemination.

1. Introduction

The wireless body area network (WBAN) for patient health monitoring is a leading technology of the current decade. Different biomedical sensors (BMSs) are used in WBAN to monitor the patient's vital signs. The vital signs of patient are heart beat rate, respiratory rate, blood pressure, and temperature [1]. Moreover, different BMSs are installed on the patient's body and inside the patient's body, and some BMSs

are placed around the patient's body to monitor different physical activities. These BMSs monitor patient's vital signs, and the monitored data are transmitted to the body coordinator (centralized node), which is responsible to immediately transmit all the patient's health information to the physicians in real time, and if an emergency situation is detected, the physician will instantly inform the patient through the computer system by sending suitable messages or alarms. The whole scenario is implemented on the

Medical Healthcare IoT, as shown in Figure 1. The data transmission is categorized into three phases. The first phase collects the sensory data using a centralized node. In the second stage, the centralized node forwards the sensory data to the base station. The base station transmits the received data to the medical staff, in the third phase. Moreover, this whole network is established via IoT network-based concepts [2]. The transmission contains reading of the sensory data and other details such as patient's name, disease, zip code, and age. There is a possibility of the data privacy issue with patient's data for sharing and processing with a doctor or nurse who may see all the information. In addition, the patient data may be stolen between the three phases in the transmission process. Thus, there is a need to design an efficient and secure data privacy technique based on machine learning when data are at risk of being stolen.

In this work, the K -medoid machine learning algorithm is used for clusters formation and local recording algorithm is employed for data privacy. In local recoding, there is a less chance of information loss during data transmission. Prior studies have also presented data privacy algorithms based on clustering such K member and OKA [3]. However, it has been noted that these algorithms take an enormous amount of time in cluster formation and suffer from information loss. In literature, major research has used anonymity, data masking, and data padding for data privacy [4, 5]. These techniques have issues of privacy, standardization, digital forgetting, mobility, object name servers (ONSs), naming, traffic characterization, quality support, data integrity, and authentication [6]. These issues are NP hard problems in the existing privacy algorithms, and most of them have considered issues of incognito, clustering, global recoding, and diversity in privacy. In addition, cluster mechanisms and local coding techniques have a greater impact on the privacy of data in the healthcare sector. The personal information privacy is the most important aspect. In privacy leakage, the invader can use any information for diverse purpose. Thus, this work has focused on the challenging issues of data privacy when data are communicated through IoT devices because any physical entity or object can be compromised. Majority of devices and sensors are operated through batteries in IoT environment with small battery and little processing energy consumption compared to data privacy being the major concern.

In this perspective, this paper proposes a clustering mechanism for privacy-aware data dissemination based on medical healthcare-IoTs (MH-IoTs) for wireless body area network. The proposed mechanism is compared with different machine learning algorithms using standard datasets for patient's data privacy when a medical doctor reviews her health report. Specially, the design of the proposed mechanisms is broadly divided into three folds:

- (i) To propose an algorithm based on the clustering technique (K -medoid), to enumerate the information loss at the anonymization process, with the aim of reducing the information loss, and to preserve the personal identification in healthcare domain.

- (ii) To design a K -anonymity model for privacy preserving using local recoding, which is based on cell suppression. The local recoding algorithm utilizes a mapping method to classify the data into different regions and places the same kind of data into the same region. Thus, the strength of using local recoding is to restrict the third-party users such as doctors and nurses when trying to insert incorrect values in order to access real patient data.
- (iii) Extensive simulations are carried out to assess the performances of the existing machine learning-based algorithms in terms of information loss and execution time for cluster formation and data privacy.

The rest of the work is arranged in following manner: Section 2 presents the existing work. Section 3 introduces the proposed clustering and K -anonymity-based data privacy mechanism (using local recording). Section 4 addresses the empirical assessment of the proposed algorithm against state-of-the-art algorithm. Section 5 concludes the research work.

2. Related Work

Numerous studies have been presented on patient data privacy based on the medical healthcare-IoTs (MH-IoTs) infrastructure. This paper [7] has focused on data privacy issues in social networks. The study also highlighted the privacy issues of the nodes deployed in the networks. In addition, the top-down approach, based on K -anonymization, has been adopted to protect privacy for individuals and organizations [8]. All information is stored in graph nodes, and the similarity index between the two nodes represents the weight of the edges in graph. The study [9] showed that the K -anonymization is a NP hard problem and proposed an algorithm, known as graph-based local recording for data anonymization. They adopted real and synthetic datasets for the simulations for K -anonymity problems [10]. On the contrary, these simulations have opened up security challenges in a distributed environment for IoT to know how to handle the Internet in the future. The existing Internet is connected to different types of nodes, such as sensors, systems, software, and applications. They work together in a single environment, known as IoT. However, the issue of data privacy in a diverse environment is a challenge. For this purpose, this study [11] has developed the cooperative distributed systems (CDSs) by connecting with the contract net protocol (CNP).

Existing studies have described security vulnerabilities of the operation layer in IoT, such as network layers, physical layers, application layers, and processing layers. [12]. IoT is essential in our lives to take effective action to protect the privacy/confidentiality and security of the user. [13]. The study in [14] has tried to establish anonymity in IoT networks by classifying the networks into coverage and the technologies deployed in IoT. Through these classifications, it is easy to get information of sources. However, Lopez et al.

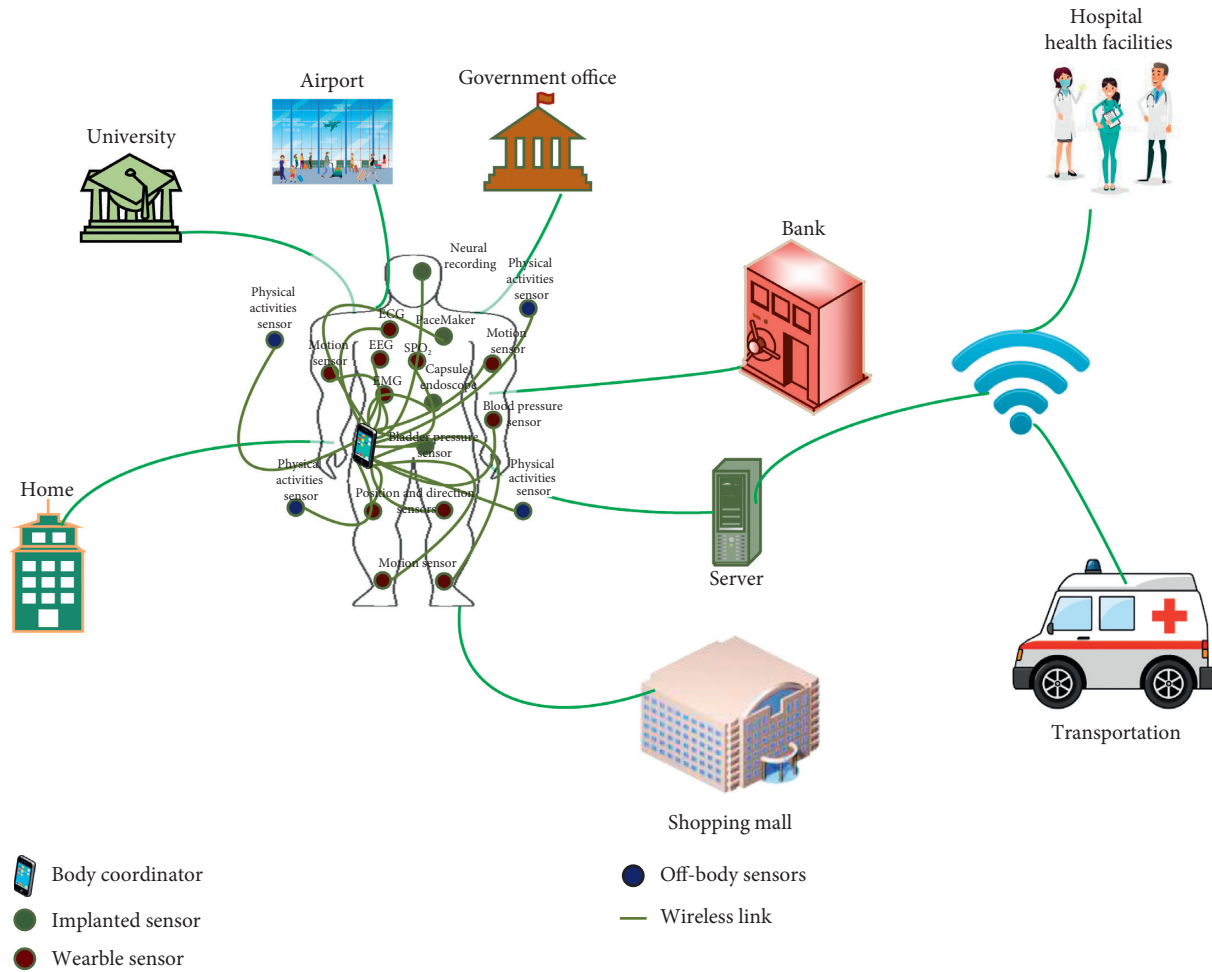


FIGURE 1: Medical healthcare-IoT-based patient's health monitoring and data dissemination.

[15] have observed that the study in [14] has not considered the privacy risk. Alabdulatif et al. [16] have explored the aim of creating a health surveillance system, making health groups smarter with the goal of detecting frequent variations in the patient's signs, elders living in assisted living surroundings, and healthy people living in smart home. The patient's data privacy problem has been handled with cloud computing by sharing the reading of vital signs to medical staff via the protected IoT environment. The narrative-based anonymity model has been developed, known as PPDC, and has been implemented on the client-server model to take care of anonymous records with reduction of privacy issues [17]. Luo et al. [18] have inherited the basic merits in [11] by developing a scheme, known as the secret sharing scheme (SW-SSS) with optimization of the secret share and exactly repairs of the shared data with storage of the patient's data in cloud. This study also claims that the patient data are safe if a server is comprised. The study [19] has evaluated and identified the privacy means, tradeoff between privacy, efficiency, and quality of the model. Zhou et al. [20] investigated the data privacy problem caused when there is increase in operators and by the approval of IoT technologies. Moreover, the framework was settled to inspect and find the consequences of privacy and security in provision of

IoT new technologies [21]. This paper [22] has developed a chaos-based encryption model for a patient's data privacy. The patient data are encrypted in form of image with casualness behaviour which ensures efficiency and the uppermost level of safekeeping from counter attacks. The electrocardio graph (ECG) for IoT-based medical care was developed for validation by removing the noises with privacy protection [23]. To improve the IoT environment for data privacy, the fog computing environment has been embedded for fast response with low latency [24]. However, the implementation of fog computing has improved the connectivity problems, but there is a need to efficiently handle data privacy in fog-based IoT deployment. The patient's revocation has been suggested in [25] with the core concept implementation of block chain technology for healthcare. Moreover, the smart cross-domain data sharing, self-adaptive access control, and smart deduplication supports have been introduced for data privacy, sharing in restricted mode and a user revocation [26]. The authors in [27] suggested EETP-MAC protocol to transmit the patient's data using prioritization by classifying into different perceptions with consideration of reduced energy consumption. The efficient design of MAC superframe structure is presented in [28] for controlling the nonemergency data with

enhanced performances. However, these papers have not considered handling the patients' data from multiple environments. The anonymization problem has validated through mathematical testing for cluster formation and information exchange [29, 30]. Thus, the existing studies on the data privacy of patient's health monitoring-based IoT dissemination have motivated to design and develop efficient mechanisms with required minimum time for data transmission with high data privacy.

3. The Proposed Work

3.1. Overview. There are various BMSs deployed to monitor vital signs of patients, and these BMSs have connected in STAR/MESH topology to the body coordinator. The aim of this proposed work is to protect the privacy of patient's data such as name, disease, age, gender, and zip code from medical doctor, public, and paramedic staff. Hence, this study has proposed two schemes that are cluster-based privacy preserving and local recoding using K -anonymity model for privacy preserving, which are explained below.

3.2. Cluster-Based Technique for Privacy Preserving. The most efficient way of the resource allocation with certain restricted conditions among public, doctor, and medical staff has been acquired via the cluster technique, as depicted in Figure 2. The cluster divides the large spaces into n spaces and allocates the private spaces according to the policies. The advantage of the cluster technique is to efficiently manage the patient's data with access policies, and through this way, privacy preserving is achieved by losing minimum information. Figure 2 shows a clustering technique by classifying health monitoring and data forwarding, main transceiver, and cluster-based restriction to access patient's data. The health monitoring and data forwarding monitor health of m patients, and the monitored data are then forwarded to the body coordinator. Moreover, the body coordinator sends the monitored data to access points wherein the access points forward data to the base station. The most important stage is the cluster-based restriction to access patient's data that have been achieved via the anonymity technique. The anonymity technique is based on the machine learning approach and is known as K -medoid.

The working procedure of K -medoid machine learning algorithm is to identify the nearest objects in the whole data elements and assigns the identified objects to the same cluster of data elements. The identified objects of the same cluster of the neighbours are assigned to groups to the anonymity algorithm. The anonymity algorithm is based on K -medoid which is implemented on the cluster technique and allows to access the patient's data according to the group and identity of personnel. Through this implementation, all the processes are executed in short time.

K -medoid is a partitioning algorithm that splits the n data points of dataset into K -nonoverlapping predefined distinctive groups known as clusters, where each data point goes into a single cluster. It is more robust and less prone to noise and outliers as compared to the K -means clustering

algorithm. Medoids (actual points) are used as cluster centers instead of K -means average points. In addition, it is simple and converges in a certain number of steps. K -medoid identifies each cluster with a single data point inside known as medoid of a cluster. It is also known as partitioning around medoid (PAM). The term medoid is the point inside a cluster whose average similarity with other data points in the cluster is greater. Aim of K -medoid algorithm is to minimize the dissimilarities summation among cluster medoid and data point in a cluster, as depicted in Algorithm 1. K -medoid cost is given as follows:

$$\sum_{P_i \in C_i} \sum_{P_j \in C_i} |P_i - C_i|. \quad (1)$$

3.3. Local Recoding-Based K -Anonymity Model for Privacy Preserving. The proposed K -anonymity model of local recoding for privacy preserving algorithm is based on cell suppression. This algorithm classifies data into different regions and places the classified data to the same region of the same group via a mapping method. Thus, the benefit of using local recoding is to restrict the third party users like doctors and nurses when they try to insert wrong values for accessing the real data of patients.

The proposed algorithm 2 for local recoding is presented. First, this algorithm computes all generalized domain (GD) of each class. Subsequently, the second step indicates to compute again all possible attributes of GD (e.g., G0 and B0). After, the generation of all generalized domains is verified whether GD is a, k anonymous. If it is true, then the generalization domain is selected. However, this process will repeat and verify all GD until an appropriate selection is achieved (e.g., size of quasi). Continuing in the steps, if the size of quasi-identifier (Q) is reached to threshold values, then it selects GD table based on distortion. Otherwise, it will go again to the second step to compute all possible attributes of GD. Finally, the achieved generalization domain data are applied to table by achieving an a, k -anonymity. Figure 3 shows the flowchart of the proposed local recoding algorithm.

4. Performance Evaluation

This section explains the performance simulation of the proposed work and has been compared with the state-of-the-art works. It is divided into three sections. First section describes the simulation environment. Second section describes the performance evaluation metrics which have used in performance measurement of the proposed work and compared with state-of-the-art works. Thirdly, the results of the proposed algorithms have been evaluated and discussed in phase 1, while the phase 2 represents the performance comparison of the proposed algorithms with the state-of-the-art algorithms.

4.1. Simulation Environment. The simulation environment has been setup on Dell machine having a CPU speed of 2.4 GH along with 6 GB RAM. Microsoft windows 10

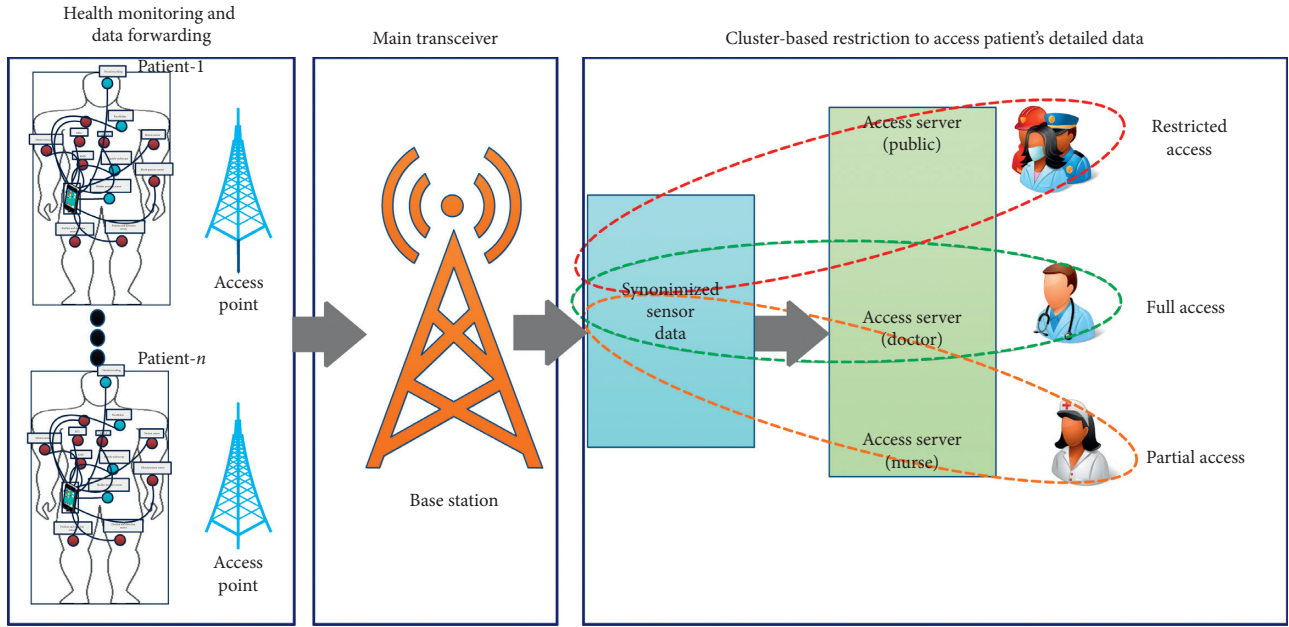


FIGURE 2: Proposed architecture for data privacy preserving based on cluster.

operating system was used, and the simulation has been carried out in Python Anaconda Spyder v2.7. This simulation has used datasets of UC Irvine Machine Learning repository. Moreover, the datasets contain numerical and categorical data elements of 14 attributes with 48842 instances. There are nine attributes chosen for K -anonymity comprising gender, marital status, education, occupation, race, and native country as a quasi-identifier. Age, education, and salary are treated numerical attributes, while the others categorical.

4.2. Simulation Performance Metrics. The selected simulation performance metrics have reduced information loss and increased execution efficiency, as explained in detail.

4.2.1. Reduction in Information Loss. The attributes in table contain numeric and categorical values with minimum and maximum distances of the X generalized equivalence classes. Thus, the information loss (LS) for numerical attributes can be expressed as follows:

$$LS = \frac{\text{Max}_x - \text{Min}_x}{\text{Max} - \text{Min}} \quad (2)$$

The information loss for categorical attributes is represented in tree with n levels, as follows:

$$LS = \frac{S * n}{S}, \quad (3)$$

$$LS = j = \sum_i^n \frac{LSj}{n}, \quad (4)$$

where S is denoted by the size of the attribute. The execution performance is improved.

To compute the system efficiency and analyse the time consumption of the local recoding algorithm in processing patient data via BMSs, the number of suppression and generalization procedures is incorporated to get local recoding and cluster-based anonymity.

4.3. Analysis of Results and Discussion. There are three parameters considered to measure the performance of the proposed work compared with state-of-the-art works that are information loss, execution time, and number of cluster (k).

4.3.1. Information Loss in K -Medoid Based on K . Information loss based on K -medoid algorithm has been tested for number of clusters (K), as shown in Figure 4. For every value, the graph will have different grounds on K values and information loss. For instance, if $K=10$, information loss is 11.24% of the anonymity algorithm. It depicts K values, and information loss is directly related. Decreasing the K will automatically result in decreasing information loss. The mentioned graph is created from six experimental values of K , i.e., 10, 20, 30, 40, 50, and 100 with corresponding information loss. The NCP value is 11.24% at $K=10$, and the value is 15.03% at $K=20$, as highlighted in Table 1, showing various values of K and NCP in information loss.

4.3.2. Execution Time in Anonymity Based on K . Performance of anonymity algorithm has been presented in terms of K values versus time, as shown in Figure 5. Changing the value of K results in varied time. For each value, the performance will be changed. At $K=10$, the anonymity algorithm will spend 1.42 seconds, and 0.93 seconds will be spent for the K value of 20. Therefore, K and time are

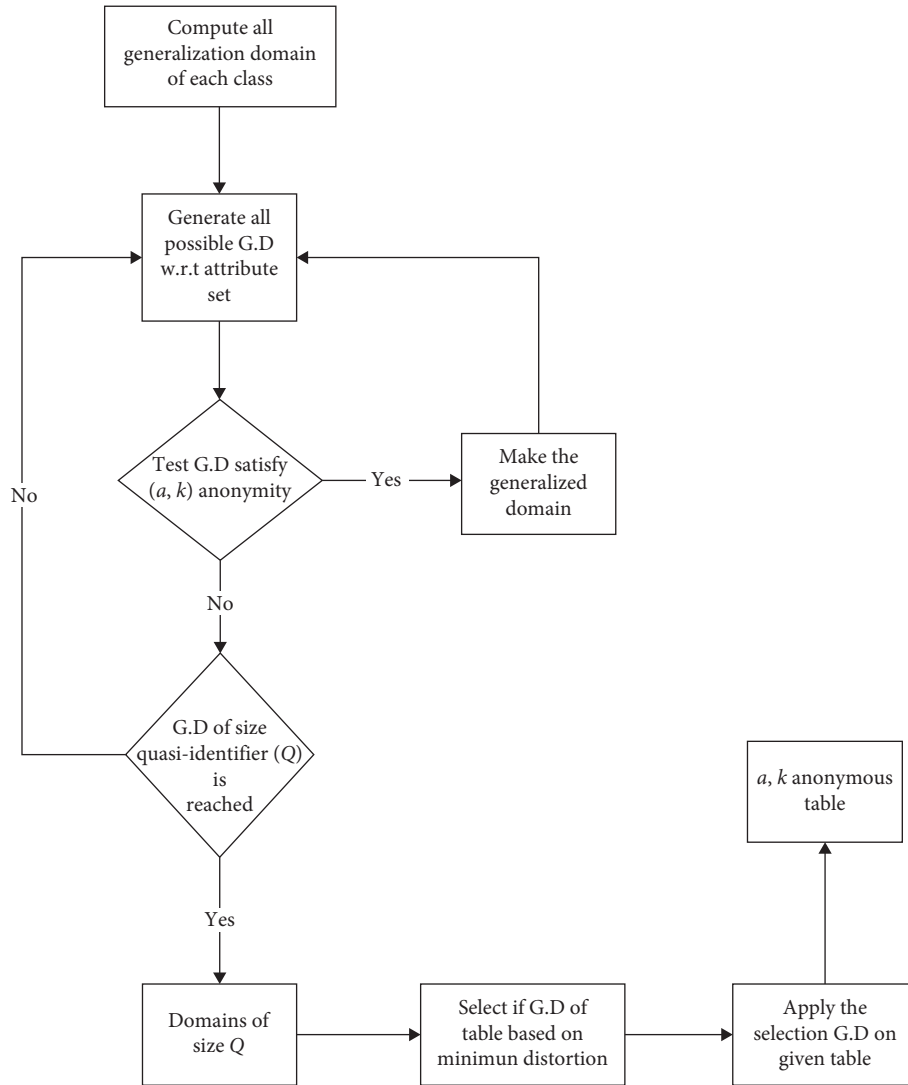


FIGURE 3: Local recoding-based anonymity.

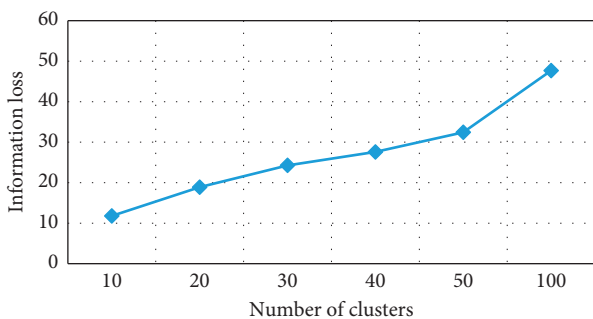


FIGURE 4: Number of clusters vs information loss in K-medoid.

indirectly related to each other. The presented results are taken from six experimental values of K , i.e., 10, 20, 30, 40, and 50. Some minor variation occurs at $K=30$ that spends 0.82 seconds. Table 2 presents statistics analysis of number of clusters versus execution time.

4.3.3. Information Loss and Execution Time of OKA (One-Pass K-Means Algorithm) Based on K . Information loss of the algorithm is 16.78% when the value K is set as 10, and at $K=20$, the resultant NCP is 25.47%, as shown in Figure 6. These NCP values gradually go up as more K values are added. Increasing the value of K will result in increasing information loss and vice versa that unfolds that both K values and information loss are directly related. The presented results are picked up from six experimental values of K , i.e., 10, 20, 30, 40, 50, and 100. In the same way, the OKA algorithm requires 7300.11 sec times for execution of 10 clusters (K), as shown in Figure 7. Here, also changing the value of K takes varied time that shows K and time are indirectly related.

4.3.4. Execution Time and Information Loss of K Member Based on K . The execution time performance has been compared based on values of K , as shown in Figure 8. To

(1) Initialization: select medoids by randomly selecting k points from a set of n data points.
 (2) Link each data point to the nearby medoid using Manhattan distance.
 (3) While the cost reduces for every medoid c , for every data point p not as medoid.

ALGORITHM 1: Pseudocode for K -medoid algorithm.

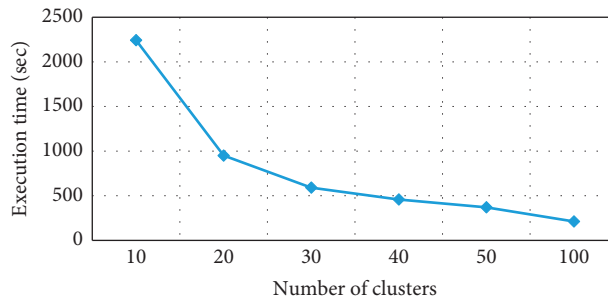


FIGURE 5: Number of clusters vs execution time in K -medoid.

Input: dataset D , quasi-identifier Q , a sensitive attribute S or a sensitive value in S , an integer k , and a fraction α
Output: (α, k) -anonymous view V
 test if D has an (α, k) anonymous table and return FALSE if not $V \leftarrow \Phi$
while $D \neq \Phi$ **do**
 Let D' contain all precisely deassociated trunks
 $D_r D \leftarrow D'$
 $V \leftarrow V \cup D'$
 $q_{\max} \leftarrow [Dr] - [(D_r, S/\alpha)]$
 choose a set of at most q_{\max} tuples in D_r
 $D \leftarrow D - D'$
 $V \leftarrow V \cup D''$
 if $D \neq \Phi$ **then**
 choose one attribute A in Q with the highest entropy generalize D according to attribute A
 end if
end while
return V

ALGORITHM 2: Pseudocode for local recoding.

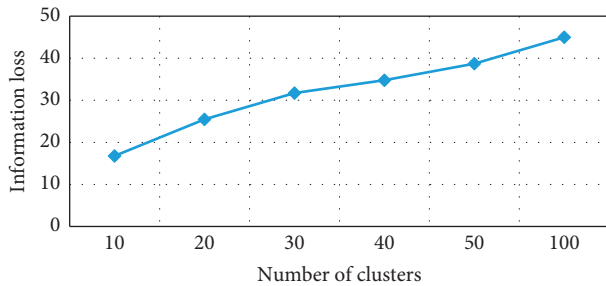


FIGURE 6: Number of clusters vs NCP in OKA.

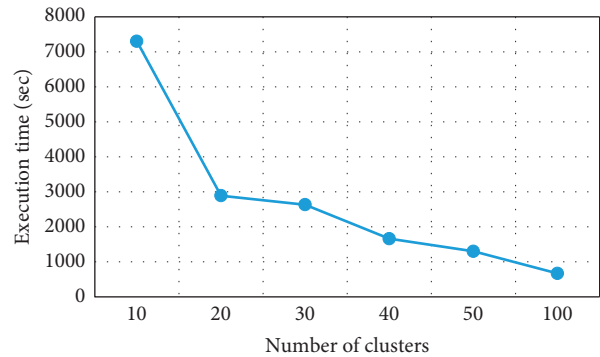


FIGURE 7: Number of clusters vs execution time of OKA.

alter the value of K , change will occur in time although time will not continuously increase but depend on K . Figure 9 shows the information loss of 11.24% at $K = 10$. Both K and information loss are directly related. The results presented in

the graph are generated from six different experimental K values, i.e., 10, 20, 30, 40, 50, and 100. Change in the information loss is also noted. For instance, its value is 11.24% at K value 10 and 15.03% at K value 20.



FIGURE 8: Number of clusters vs execution time in K -member.

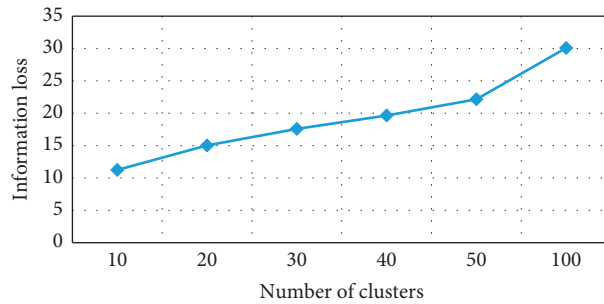


FIGURE 9: Number of clusters vs execution time in K -member.

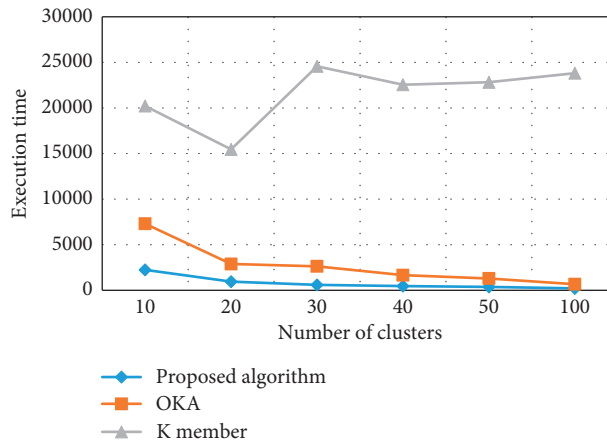


FIGURE 10: Number of different clusters vs execution time.

4.4. Phase 2: Results Comparison of the Proposed Algorithms with State-of-the-Art Algorithms

4.4.1. Execution Time of All Clusters Based on K . The performance of the proposed algorithm for cluster formation and the required execution time is compared with OKA and K member, as presented in Figure 10. When a cluster initial value is set 10, then the execution time for a cluster formation of K is very high as compared to OKA. In the same way, the proposed algorithm outperforms in terms of less time required for cluster formation when a cluster initial value is 10 as compared to OKA and K member. The reason

is that the proposed algorithm selects a node with minimum value or distance as compared to both algorithms. The number of clusters along with NCP values is given in Table 3.

4.4.2. Information Loss in Clusters Based on K . The information loss of the proposed algorithm is compared with two clusters algorithms, that is, OKA algorithm and K member algorithm, as depicted in Figure 11. The information loss ratio is increased as the network extends in terms of further cluster formation. Thus, the performance of the proposed algorithm is comparatively better to OKA and K member.

TABLE 1: Information loss in K -medoid algorithm.

K	NCP
10	11.79
20	18.88
30	24.26
40	27.58
50	32.44
100	47.67

TABLE 2: Execution time in K -medoid.

Number of clusters	Execution time (sec)
10	2242.87
20	950.25
30	591.36
40	458.01
50	370.42
100	211.93

TABLE 3: Information loss in OKA algorithm.

Number of clusters	NCP (%)
10	16.78
20	25.47
30	31.72
40	34.77
50	38.71
100	44.99

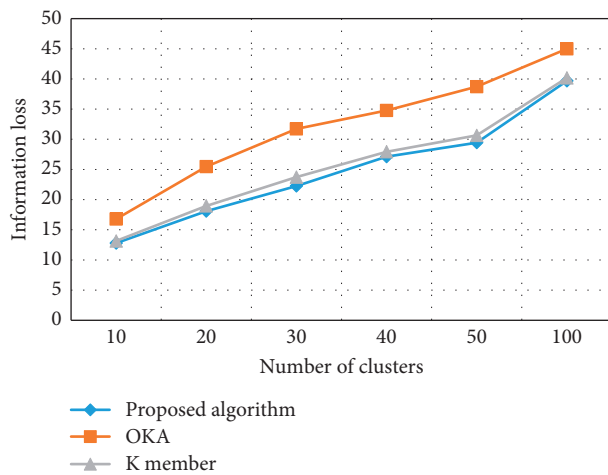


FIGURE 11: Information loss ratio in clusters.

Performance measurement for information loss of OKA is higher than K member because K member thoroughly verifies the existing nodes in clusters caused by more information loss comparatively. Moreover, there is increase in information loss if k unceasingly alters which means the more the number of clusters, the more the information loss. Conversely, decreasing the number of clusters will result in decreasing value of information loss. The proposed algorithm has features of clusters set in the network. These clusters lessen the execution by adopting the substitute

channels for providing the stream of bits from one end to another in the network. Inserting multiple clusters can condense the information loss during transmission due to incessant data monitoring by each cluster.

Time complexity is the central point to present proficient results. OKA takes $O(n^2/K)$, and k takes $O(n^2)$. The entire time complexity of the proposed algorithm is $O(nk + nd)$, where $O(d)$ specifies calculating a distance to one example. $O(nd)$ reveals to compute the distance to every cluster. $O(nk)$ is the entire time to find out the closest cluster in k .

5. Conclusion

The cluster-based resources sharing using machine learning approaches is one of the prominent concepts in WBAN. The first proposed algorithm is employed for preserving personal identification using cluster concepts with polices of the access restrictions. This algorithm has been efficiently performed for preserving personal identification in data sharing and information gathering in clusters. Also, the maximum number of clusters formation has consumed optimal time as compared to the existing machine learning algorithms. The second proposed algorithm has lost minimum information comparatively in OKA and K member algorithms. These performances significantly increase the privacy of the patient's data in a better way.

In the future, the proposed algorithm will improve and compare with a deep learning algorithm for the K -anonymity problem for patient's data privacy.

Data Availability

The data can be obtained from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by Taif University Researchers Supporting Project number (TURSP-2020/254), Taif University, Taif, Saudi Arabia.

References

- [1] F. Ullah, A. H. Abdullah, O. Kaiwartya, J. Lloret, and M. M. Arshad, "EETP-MAC: energy efficient traffic prioritization for medium access control in wireless body area networks," *Telecommunication System*, vol. 75, pp. 181–203, 2020.
- [2] T. Manna and I. S. Misra, "Performance analysis of scheduled access mode of the IEEE 802.15.6 MAC protocol under non-ideal channel conditions," *IEEE Transactions on Mobile Computing*, vol. 19, no. 4, pp. 935–953, 2020.
- [3] W. Xue, K. Yu, X. Hua, Q. Li, W. Qiu, and B. Zhou, "APs' virtual positions-based reference point clustering and physical distance-based weighting for indoor wi-fi positioning," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 3031–3042, 2018.

- [4] P. Zhao, H. Jiang, J. C. S. Lui et al., "P3-LOC: a privacy-preserving paradigm-driven framework for indoor localization," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2856–2869, 2018.
- [5] X. Wang, J.-K. Chou, W. Chen et al., "A utility-aware visual approach for anonymizing multi-attribute tabular data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 351–360, 2018.
- [6] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi, and S. Shamshirband, "A new K-nearest neighbors classifier for big data based on efficient data pruning," *Mathematics*, vol. 8, no. 2, p. 286, 2020.
- [7] L. Atzori, A. Iera, and G. Morabito, "The internet of things: a survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.
- [8] J. Li, J.-J. Yang, Y. Zhao, and B. Liu, "A top-down approach for approximate data anonymisation," *Enterprise Information Systems*, vol. 7, no. 3, pp. 272–302, 2013.
- [9] K. S. Babu, S. K. Jena, J. Hota, and B. Moharana, "Anonymizing social networks: a generalization approach," *Computers & Electrical Engineering*, vol. 39, no. 7, pp. 1947–1961, 2013.
- [10] R. Roman, J. Zhou, and J. Lopez, "On the features and challenges of security and privacy in distributed internet of things," *Computer Networks*, vol. 57, no. 10, pp. 2266–2279, 2013.
- [11] A. Samani, H. H. Ghenniwa, and A. Wahaishi, "Privacy in internet of things: a model and protection framework," *Procedia Computer Science*, vol. 52, pp. 606–613, 2015.
- [12] A. Crabtree and R. Mortier, "Personal data, privacy and the internet of things: the shifting locus of agency and control," *SSRN Electronic Journal*, vol. 2016, 2016.
- [13] A. Wahab, O. Ahmad, M. Muhammad, and M. Ali, "A comprehensive analysis on the security threats and their countermeasures of IoT," *International Journal of Advanced Computer Science & Applications*, vol. 8, no. 7, 2017.
- [14] L. Hellebrandt, O. Hujňák, P. Hanáček, and I. Homoliak, "Survey of privacy enabling strategies in IoT networks," in *Proceedings of the 2017 International Conference On Computer Science And Artificial Intelligence - CSAI 2017*, pp. 216–221, Jakarta, Indonesia, December 2017.
- [15] J. Lopez, R. Rios, F. Bao, and G. Wang, "Evolving privacy: from sensors to the internet of things," *Future Generation Computer Systems*, vol. 75, pp. 46–57, 2017.
- [16] A. Alabdulatif, I. Khalil, A. R. M. Forkan, and M. Atiquzzaman, "Real-time secure health surveillance for smarter health communities," *IEEE Communications Magazine*, vol. 57, no. 1, pp. 122–129, 2019.
- [17] H. Li, F. Guo, W. Zhang, J. Wang, and J. Xing, "Anonymous scheme for privacy-preserving data collection in IoT-based healthcare services systems," *Journal of Medical Systems*, vol. 42, no. 3, p. 56, 2018.
- [18] E. Luo, M. Z. A. Bhuiyan, G. Wang, M. A. Rahman, J. Wu, and M. Atiquzzaman, "PrivacyProtector: privacy-protected patient data collection in IoT-based healthcare systems," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 163–168, 2018.
- [19] S. Sharma, K. Chen, and A. Sheth, "Toward practical privacy-preserving analytics for IoT and cloud-based healthcare systems," *IEEE Internet Computing*, vol. 22, no. 2, pp. 42–51, 2018.
- [20] W. Zhou, Y. Jia, A. Peng, Y. Zhang, and P. Liu, "The effect of IoT new features on security and privacy: new threats, existing solutions, and challenges yet to be solved," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1606–1616, 2019.
- [21] M. N. Alraja, M. M. J. Farooque, and B. Khashab, "The effect of security, privacy, familiarity, and trust on users' attitudes toward the use of the IoT-based healthcare: the mediation role of risk perception," *IEEE Access*, vol. 7, pp. 111341–111354, 2019.
- [22] R. Hamza, Z. Yan, K. Muhammad, P. Bellavista, and F. Titouna, "A privacy-preserving cryptosystem for IoT E-healthcare," *Information Sciences*, vol. 527, pp. 493–510, 2020.
- [23] P. Huang, L. Guo, M. Li, and Y. Fang, "Practical privacy-preserving ECG-based authentication for IoT-based healthcare," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 9200–9210, 2019.
- [24] R. Saha, G. Kumar, M. K. Rai, R. Thomas, and S.-J. Lim, "Privacy Ensured $\{e\}$ -healthcare for fog-enhanced IoT based applications," *IEEE Access*, vol. 7, pp. 44536–44543, 2019.
- [25] J. Xu, K. Xue, S. Li et al., "Healthchain: a blockchain-based privacy preserving scheme for large-scale health data," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8770–8781, 2019.
- [26] Y. Yang, X. Zheng, W. Guo, X. Liu, and V. Chang, "Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system," *Information Sciences*, vol. 479, pp. 567–592, 2019.
- [27] F. Ullah, A. H. Abdullah, O. Kaiwartya, J. Lloret, and M. M. Arshad, "EETP-MAC: energy efficient traffic prioritization for medium access control in wireless body area networks," *Telecommunication System*, vol. 75, pp. 181–203, 2020.
- [28] F. Ullah, A. H. Abdullah, O. Kaiwartya et al., "TraPy-MAC: traffic priority aware medium access control protocol for wireless body area network," *Journal of Medical Systems*, vol. 41, p. 93, 2017.
- [29] X. C. Yin, Z. G. Liu, B. Ndibanje, L. Nkenyereye, and S. M. Riazul Islam, "An IoT-based anonymous function for security and privacy in healthcare sensor networks," *Sensors*, vol. 19, no. 14, p. 3146, 2019.
- [30] X. He, H. Chen, Y. Chen, Y. Dong, P. Wang, and Z. Huang, *Clustering-Based K-Anonymity*, Springer-Verlag, Berlin, Heidelberg, Germany, 2012.