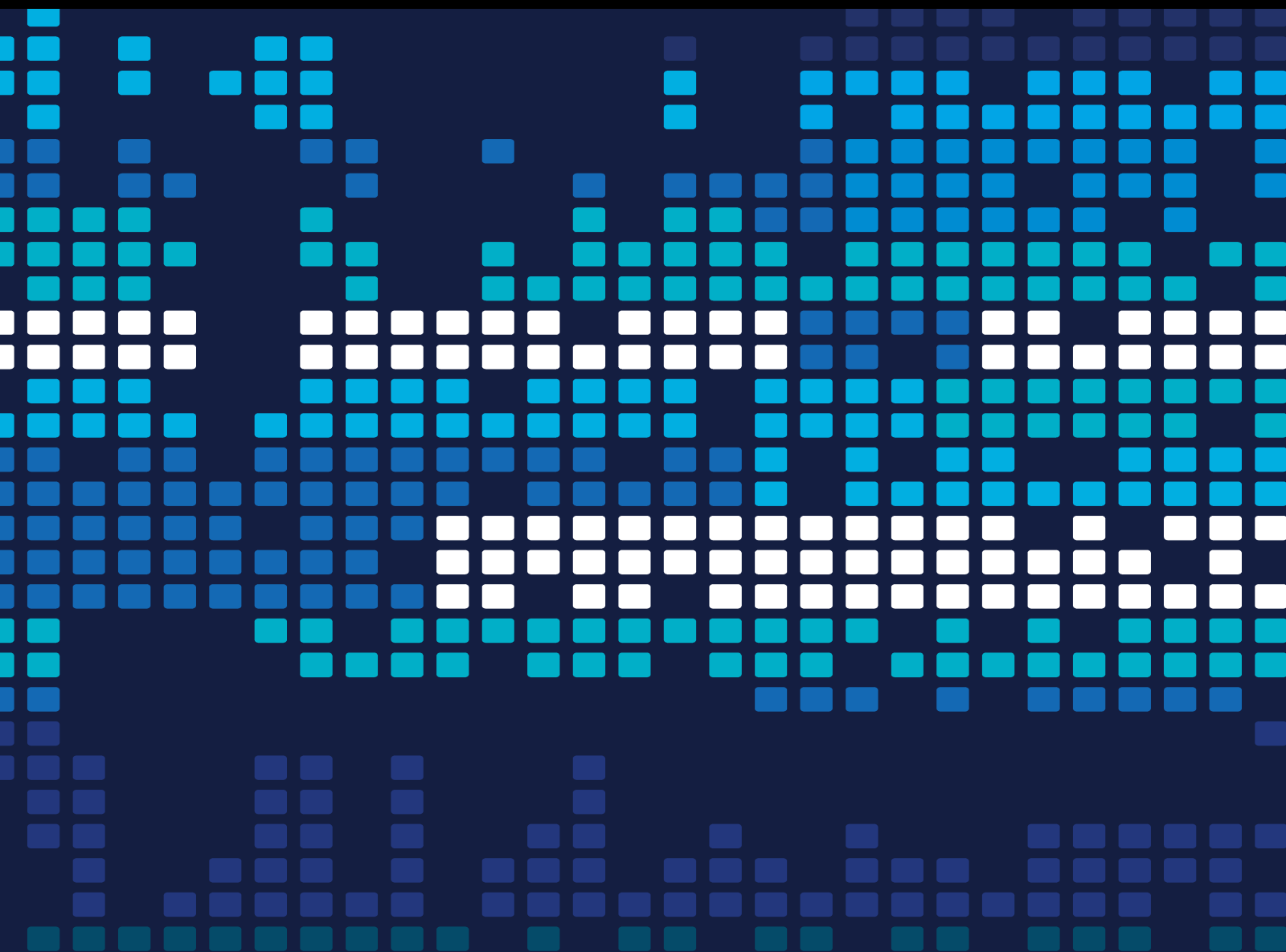


# Healthcare Big Data Management and Analytics in Scientific Programming

Lead Guest Editor: Shah Nazir

Guest Editors: Iván García-Magariño, Rodziah Binti Atan, and Shaukat Ali





---

# **Healthcare Big Data Management and Analytics in Scientific Programming**

Scientific Programming

---

## **Healthcare Big Data Management and Analytics in Scientific Programming**

Lead Guest Editor: Shah Nazir

Guest Editors: Iván García-Magariño, Rodziah Binti Atan, and Shaukat Ali



---

Copyright © 2021 Hindawi Limited. All rights reserved.

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

# Chief Editor

Emiliano Tramontana, Italy


---

## Editorial Board


Manuel E. Acacio Sanchez, Spain  
Marco Aldinucci, Italy  
Sikandar Ali, China  
Davide Ancona, Italy  
Daniela Briola, Italy  
Mu-Chen Chen, Taiwan  
Ferruccio Damiani, Italy  
Sergio Di Martino, Italy  
Bai Yuan Ding, China  
Basilio B. Fragueta, Spain  
Jianping Gou, China  
Ligang He, United Kingdom  
Jiwei Huang, China  
Chin-Yu Huang, Taiwan  
Shujuan Jiang, China  
José E. Labra, Spain  
Maurizio Leotta, Italy  
Zhihan Liu, China  
Piotr Luszczek, USA  
Tomàs Margalef, Spain  
Cristian Mateos, Argentina  
Roberto Natella, Italy  
Shah Nazir, Pakistan  
Francisco Ortin, Spain  
Can Özturan, Turkey  
Zhaoqing Pan, China  
Antonio J. Peña, Spain  
Danilo Pianini, Italy  
Jiangbo Qian, China  
Fabrizio Riguzzi, Italy  
Michele Risi, Italy  
Sebastiano Fabio Schifano, Italy  
Autilia Vitiello, Italy  
Pengwei Wang, China  
Jan Weglarz, Poland  
hong wenxing, China  
Qianchuan Zhao, China

# Contents





## **Healthcare Big Data Management and Analytics in Scientific Programming**

Shah Nazir , Iván García-Magariño , Rodziah Binti Atan, and Shaukat Ali  
Editorial (2 pages), Article ID 9780175, Volume 2021 (2021)














## **EEG Based Aptitude Detection System for Stress Regulation in Health Care Workers**

Tehseen Khan , Huma Javed, Mohammad Amin, Omar Usman, Syed Ishtiaq Hussain, Amjad Mehmood, and Carsten Maple  
Research Article (11 pages), Article ID 4620487, Volume 2021 (2021)



## **Optimal Policy Learning for Disease Prevention Using Reinforcement Learning**

Zahid Alam Khan, Zhengyong Feng, M. Irfan Uddin , Noor Mast, Syed Atif Ali Shah , Muhammad Intiaz, Mahmoud Ahmad Al-Khasawneh , and Marwan Mahmoud   
Research Article (13 pages), Article ID 7627290, Volume 2020 (2020)



## **A New Approach for Enhancing the Services of the 5G Mobile Network and IOT-Related Communication Devices Using Wavelet-OFDM and Its Applications in Healthcare**

Mordecai F. Raji , JianPing Li , Amin Ul Haq , Victor Ejianya , Jalaluddin Khan , Asif Khan , Mudassir Khalil , Amjad Ali , Ghufuran A. Khan , Mohammad Shahid , Bilal Ahamad , Amit Yadav , and Imran Memon   
Research Article (13 pages), Article ID 3204695, Volume 2020 (2020)


## **Deep Learning Algorithm for Brain-Computer Interface**

Asif Mansoor, Muhammad Waleed Usman, Noreen Jamil , and M. Asif Naeem   
Research Article (12 pages), Article ID 5762149, Volume 2020 (2020)


## **Towards a Complete Set of Gym Exercises Detection Using Smartphone Sensors**

Usman Ali Khan, Iftikhar Ahmed Khan , Ahmad Din, Waqas Jadoon, Rab Nawaz Jadoon, Muhammad Amir Khan , Fiaz Gul Khan, and Abdul Nasir Khan  
Research Article (12 pages), Article ID 6471438, Volume 2020 (2020)



## **QuPiD Attack: Machine Learning-Based Privacy Quantification Mechanism for PIR Protocols in Health-Related Web Search**

Rafiullah Khan, Arshad Ahmad , Alhuseen Omar Alsayed, Muhammad Binsawad, Muhammad Arshad Islam, and Mohib Ullah  
Research Article (11 pages), Article ID 8868686, Volume 2020 (2020)

## **A Systematic Review of Healthcare Big Data**

Rakesh Raja , Indrajit Mukherjee, and Bikash Kanti Sarkar  
Review Article (15 pages), Article ID 5471849, Volume 2020 (2020)




## **Data Analytics in Mental Healthcare**

Ayesha Kamran Ul haq, Amira Khattak, Noreen Jamil , M. Asif Naeem , and Farhaan Mirza  
Review Article (9 pages), Article ID 2024160, Volume 2020 (2020)



**Application of Big Data Fusion Based on Cloud Storage in Green Transportation: An Application of Healthcare**

Li Qin Hu , Amit Yadav , Asif Khan, Hong Liu, and Amin Ul Haq  
Research Article (8 pages), Article ID 1593946, Volume 2020 (2020)

**Biomedical Relation Extraction Using Distant Supervision**

Nada Boudjellal , Huaping Zhang , Asif Khan , and Arshad Ahmad   
Review Article (9 pages), Article ID 8893749, Volume 2020 (2020)

**Towards Energy-Efficient Framework for IoT Big Data Healthcare Solutions**

Chong Feng , Muhammad Adnan, Arshad Ahmad , Ayaz Ullah, and Habib Ullah Khan  
Research Article (9 pages), Article ID 7063681, Volume 2020 (2020)

## Editorial

# Healthcare Big Data Management and Analytics in Scientific Programming

**Shah Nazir** <sup>1</sup>, **Iván García-Magariño** <sup>2</sup>, **Rodziah Binti Atan**,<sup>3</sup> and **Shaukat Ali**<sup>4</sup>

<sup>1</sup>Department of Computer Science, University of Swabi, Swabi, Pakistan

<sup>2</sup>Department of Software Engineering and Artificial Intelligence, Complutense University of Madrid, 28040 Madrid, Spain

<sup>3</sup>Faculty of Computer Science and Information Technology, University Putra Malaysia, Selangor, Malaysia

<sup>4</sup>Department of Computer Science, Islamia College, Peshawar, Khyber Pakhtunkhwa, Pakistan

Correspondence should be addressed to Shah Nazir; shahnazir@uoswabi.edu.pk

Received 28 July 2021; Accepted 28 July 2021; Published 15 October 2021

Copyright © 2021 Shah Nazir 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 systems of healthcare are being transmuted by scientific improvements in the information of medical systems, electronic records in medical, smart, and wearable devices, and handheld devices. This growth in medical big data, together with the expansion of computational approaches in the area of healthcare, has aided researchers and practitioners to excerpt and visualize medical big data in a novel scale of research. The role of scientific programming in providing solutions to existing and forthcoming issues exists in the organization of large-scale data in healthcare, such as by assisting in the handing out of vast data volumes, modelling of composite systems, and sourcing derivations from healthcare data and simulations. Visualization is a significant tool in producing diagrams, images, or animations to transfer healthcare messages and improve understandings. Programming tools, including Apache Hadoop, Informatica PowerCenter, and Tableau, analyze data extremely efficiently and enable the visualization of meaningful insights extracted from big data.

Research is needed that explores the integration of big data and healthcare from a scientific programming perspective. Research that considers technological and computational barriers to big data management has clear applications in the area of big data in healthcare. Diverse approaches have been in practice by researchers and practitioners. These areas include decision support systems for big data in healthcare, programming for medical big data visualization and its representation, applications of machine and/or deep learning algorithms and applications in big data

analytics for healthcare, data warehouse and knowledge representation in healthcare technologies, data mining in Internet of Things (IoT) and healthcare systems, and probabilistic computing approaches to the management of big data in healthcare.

The papers included in this Special Issue cover the details of the scientific aspects which are mainly involved in the field of healthcare big data management, its organization, and the role of programming to deal with a particular situation of big data. Raji et al. used the wavelet transform performance against future wireless application system requirements and offered guidelines and methods for wavelet applications in 5G waveform design. The detailed effect of healthcare was targeted. With the help of images as test data, a detailed performance comparison of the Fourier transform and various wavelet transforms have been done with the help of modulation and demodulation complexity, energy efficiency, latency, reliability, spectral efficiency, robustness to time and frequency selective channels, and effect of transmission and reception considered as the key performance indicators. After this, the guidelines are presented for the wavelet transform. Mansoor et al. present an overview of the data acquisition, feature extraction, and classification algorithm approaches adopted by researchers and practitioners in the previous years. Some classification algorithms for EEG-based BCI systems are adaptive classifiers, tensor classifiers, transfer learning approach, and deep learning, as well as some miscellaneous approaches. From the experimental results of the method, it was concluded that using the



adaptive classifiers, precise results were obtained compared to the static classification approach. Deep learning algorithms were adopted for developing and achieving the specified objectives and implementation.

R. Khan et al. investigated the protection level offered by UUP. The query profile distance attack as machine learning-based attack was presented for evaluating the effectiveness of UUP in privacy protection. The approach demonstrates the distance between upcoming query and the user profile. Ten classification algorithms were used for the experimental setup. These algorithms include the tree-based, lazy learner, rule-based, metaheuristic, and Bayesian families for the sake of comparison. Two subsets of an American online dataset including noisy and clean datasets were used for experimental process. Results of the experiments showed that the suggested approach links more than 70% queries to the right users with 72% precision for the clean dataset, while for the noisy dataset, the proposed approach associates more than 40% queries to the right user with a precision of 70%.

Raja et al. presented a systematic literature review of big data related to healthcare. The study evaluated 34 journal articles for the years 2015 to 2019 accordingly to the defined inclusion and exclusion criteria. The study specified the research in the area of big data with its applications and challenges in healthcare implementation. Kamran Ul haq et al. analyzed the studies based on the approach of big data in mental illness and treatment. Diverse types of mental illness such as bipolar disorder, personality disorder, and depression are discussed. The user behavior based on mental illness for drug addiction and suicide are highlighted. An overview of the approaches and tools for predicting the mental condition of a patient based on machine learning and artificial intelligence is given. Hu et al. presented an approach of information research for fusing a large volume of heterogeneous data produced by a charging pile resultant to the new energy electric vehicle in the network of vehicles and present the concept of cloud computing as a module of storage for facilitating the storage and associated expansion of big data. Khan et al. identified 26 exercises of gym from the literature. Out of these, 14 were unique and 12 were common in the existing literature. The study also finds the suitable smartphone attachment position and the number of sensors for predicting exercise with high accuracy.

N. Boudjellal et al. presented an approach of distant supervision for relation extraction, providing a generic architecture of this task based on the existing approaches. The study reviewed the approaches used in the literature targeting the current areas of research with details of knowledge bases used in the process along with the corpora which can be supportive for trainee practitioners looking for knowledge in the field. Feng et al. proposed a framework of fog-based IoT healthcare for reducing the consumption of energy of fog nodes. The results of the study show that the performance of the suggested framework is effective in terms of the network delay and usage of energy. Discussions and suggestions are made for the services of big data for fog devices and analytics of healthcare big data are presented.

## Conflicts of Interest

The editors declare no conflicts of interest.

*Shah Nazir*  
*Iván García-Magariño*  
*Rodziah Binti Atan*  
*Shaukat Ali*

## Research Article

# EEG Based Aptitude Detection System for Stress Regulation in Health Care Workers

**Tehseen Khan** <sup>1</sup>, **Huma Javed**<sup>2</sup>, **Mohammad Amin**<sup>3</sup>, **Omar Usman**<sup>3</sup>,  
**Syed Ishtiaq Hussain**<sup>4</sup>, **Amjad Mehmood**<sup>5,6</sup> and **Carsten Maple**<sup>6</sup>

<sup>1</sup>Department of Computer Science, FAST-NUCES Peshawar, Peshawar, Pakistan

<sup>2</sup>University of Peshawar, Peshawar, Pakistan

<sup>3</sup>FAST-NUCES Peshawar, Peshawar, Pakistan

<sup>4</sup>KP-HED, Peshawar, Pakistan

<sup>5</sup>Kohat University, Kohat, Pakistan

<sup>6</sup>Secure Cyber Systems Research Group, WMG, University of Warwick, Coventry, UK

Correspondence should be addressed to Tehseen Khan; [to\\_tsn@yahoo.com](mailto:to_tsn@yahoo.com)

Received 10 February 2020; Revised 8 April 2020; Accepted 6 June 2021; Published 11 October 2021

Academic Editor: Sebastiano Fabio Schifano

Copyright © 2021 Tehseen 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.

Stress is a complex multifaceted concept that is the result of adverse or demanding circumstances. Workers, especially health care workers, suffer significantly from distress, burnout, and other physical illnesses such as hypertension and diabetes caused by stress. Numerous stress detection systems are realized but they only help in detecting the stress in early stages, and, for regularizing it, these systems employ other means. These systems lack any inherent feature for regularization of stress. In contributing toward this aim, a novel system “EEG-Based Aptitude Detection System” is proposed. This system will help in considering working aptitude of employees working in work places with an intention to help them in assigning proper job roles based on their working aptitude. Selection of right job role for workers not only helps in uplifting productivity but also helps in regulating stress level of employees caused by improper job role assignments and reduces fatigue. Being able to select right job role for workers will help them in providing productive working environment. This paper presents detail layered architecture, implementation details, and outcomes of the proposed novel system. Integration of this system in work places will help supervisors in utilizing the human resource more suitably and will help in regulating stress related issues with improvement in overall performance of entire office. In this work, different implementation architectures based on KNN, SVM, DT, NB, CNN, and LSTM are tested, where LSTM has provided better results and achieved accuracy up to 94% in correctly classifying an EEG signal. The rest of the details can be seen in Sections 3 and 5.

## 1. Introduction

The exponential increase in usage of ubiquitous computing systems and urban living has led to the development of a unique class of complex monitoring and control subsystems. These systems are used in diverse areas and their components can affect sectors such as transport, health, energy, home/buildings, and the environment. Such systems are identified by the general nomenclature of *smart city*. Their functioning involves the usage of a vast number of hardware sensors and they are typically realized using Wireless Sensor Networks, IoT devices, and smart phones but are certainly not restricted to these architectures only. Research in this

field is further classified into areas such as health monitoring [1–4], traffic management [5], intelligent agriculture [6], smart power grids [7], environment monitoring [8–10], human psychology [11–14], smart water grids, smart homes [15], and smart offices [16]. Each of these applications involves the formulation of architecture that integrates sensing, storage, communication, processing, and human computer interfacing. In this context, this manuscript describes an architecture that incorporates aspects of both smart offices and human psychology, with the objective that this architecture can be used to create a conducive and interconnected office environment, where employee productivity can be realized to its full potential.

The term *smart office* generally applies to an environment that benefits both the employees and the organization where it is deployed. Employee's work experience, on the one hand, and office productivity, on the other hand, are improved. Major research problems in the field belong to the domains of communications, human computer interfacing, efficient information processing, office management, adaptation services, and assistance [17, 18].

More recently, aspects such as stress detection and emotion detection, have been included into the context of smart offices [19–27]. Identification of stress levels of employees allows organizations to regulate them at early stages before it affects their performance and becomes cause of deterioration in health. Globally, all workers and especially health care workers are more prone to risks caused by stress. Because of their frequent exposure to risk factors such as high work demands, low work control, and high emotional involvement [1], high exposure to these risk factors enhances stress and mental health complaints and is the main cause behind degrading work performance [2]. These complaints also have other undesirable aspects like quality interaction with patients and colleagues [3]. It is known that moderate and high psychological distress increases the odds for workplace failure and decreases the odds for workplace success [4].

Similarly, emotion detection systems map employees' emotion states with routine activities in an office environment. This manner of parametrizing aspects of human psychology allows managers to forecast and assign appropriate job roles for employees [28–30]. Role assignment can be further optimized, eventually improving productivity. This research work proposes the usage of an additional parameter in the form of *aptitude* for job role assignment.

Aptitude is a qualitative inborn ability to perform a particular task more efficiently than an average person and is also inversely linked to stress [31, 32]. It is usually quantified by means of testing mechanisms. A number of tests are standardized and used globally for various types of professional aptitude assessment. Examples are the GRE (Graduate Record Examinations), GMAT (Graduate Management Admission Test), SAT (Scholastic Aptitude Test), and other similar tests.

This research work performs this quantification by means of Electroencephalography EEG signals and proposes an EEG-based Aptitude Detection architecture. To the best of our knowledge, the inclusion of aptitude as a parameter in smart-office environments is novel. As a proof of concept, analytical skills as a binary ability are considered in this system. Analytical skills, along with IQ level, and dexterities are a number of facets that collectively define aptitude. These additional facets will be addressed in the future. In its current scope, the implementation pipeline includes convolutional neural network (CNN), decision tree (DT), K-nearest neighbors (KNN), Naïve Bayes (NB), support vector machine (SVM), and long-short-term memory (LSTM). This research work reports highest accuracy of 94% using LSTM. For deep networks, it also proposes different topologies and filters for the EEG signals. Lastly, an aptitude-based EEG data set is also a novel contribution of this manuscript.

In the remainder of the manuscript, literature review is given in Section 2, the proposed system and architecture are given in Section 4, and, finally, the outcomes and results are discussed in Section 5, followed by conclusion in Section 6.

## 2. Literature Review

EEG signals are a measure of difference in electrical brain activity attributed to neurons. The signals appear as wave patterns that can be captured using EEG devices [14, 27, 33]. The wavebands represent brain activity due to different types of stimulants such as sensory usage, memory recall, focus and attention, problem solving, relaxation, drowsiness, deep sleep, and others (see Table 1). Some stimulants may result in disappearance of one waveband but an increase in bandwidth of another [41]. EEG signals and EEG devices form the core operations in a number of medical applications, including detection of dementia and epilepsy, sleep disorders, stress or workload measurement [39, 41, 42], and emotion recognition [20–24, 33]. The latter application of emotions detection and recognition has formed one of the core components in smart offices and brain computer interaction (BCI). Different researchers have made contributions in emotion detection and recognition utilizing physiological signals as input (see Table 2).

The highest classification accuracy reported is 99.5%, which is achieved using Electrodermal Activity (EDA) and Heart Rate (HR) signals by means of a fuzzy logic classifier [39]. Here, a single emotion trait is captured. For two emotion traits (arousal and valance), the maximum reported accuracy is 96.6% using multimodal signals using standard statistical features [22]. Here, an ANN is used as a classifier. With four emotion traits (joy, anger, sadness, and pleasure), the maximum reported accuracy is 95% using standard statistical and entropy-based features [20]. Here, Linear Discriminant Analysis is used as a classifier using EMG, ECG, and RSP signals. The highest accuracy using Support Vector Machines is reported in [24] as 92%, followed by 91% in [27]. The former used multimodal physiological signals, while the latter used only EEG. In all cases, a number of factors, including number of modality signals, feature set, and classification techniques, contribute toward an increase in accuracy.

The authors in [32] have proposed aptitude modeling; they have provided multimodal system based on signals such as heart rate, skin temperature, breathing, and Galvanic Skin Response. They have managed to achieve an accuracy up to 96% using a multimodal approach with F1-score of 0.91. The proposed system in this paper is based on encephalographic signal. To the best of the authors' knowledge, no such work has been done before. Before going in to the implementation details of the proposed system, a brief introduction of tools utilized in implementation is covered in Section 3.

## 3. Tools and Methods

A collaborative setup is established using Python and data science/numerical libraries, and the details related to these libraries are provided below. It is worth mentioning that, in implementation of complete system famous NumPy package and respective support has played a very vital role.

TABLE 1: Frequency wavebands of EEG signals considered in this study.

Type	Freq. range (Hz)	Stimulants
Delta	0.5–4.0	Deep sleep and unconsciousness
Theta	4.0–8.0	Drowsiness, fatigue, and day dreaming
Alpha	8.0–13.0	Relaxation and meditation
Beta	13.0–30.0	Focus, attention, and problem solving
Gamma	30–50	Memory and senses

TABLE 2: Literature review.

Ref	Modality signal	Features	Classification	Emotions	Accuracy (%)
[11]	EEG	Energy, entropy	SVM, KNN	Arousal, valence	86
[13]	EEG	Min, Max peak, power	LSTM	Arousal, valence, and liking	87
[14]	EEG	Min, Max peak, power	ANN	Stress, normal	60
[20]	EMG, ECG, RSP	Statistical, energy, entropy	LDA	Joy, anger, sadness, and pleasure	95
[24]	BVP, EMG, EDA, RSP	Statistical features	SVM, Fisher LDA	Amusement, contentment, disgust, fear, sadness, and neutral	92
[26]	EMG, EDA, ECG	No specific feature	No specific classifier	Arousal, valence	NA
[27]	EEG	Statistical features	SVM, ANN	Positive, negative, and neutral	91
[33]	EEG, EMG, Temp, GSR, RSP	Different features	MESAE	Arousal, valence	77
[34]	EEG	No specific features	LDA	Arousal, valence	87
[35]	EEG	DE, PSD	SVM	Negative, positive, and neutral	91.5
[36]	EEG	Spatial, spectral, temporal	CNN	Depression	86
[37]	EDA, HR, EMG	No specific features	HMM	Arousal, valence	81
[38]	EEG	Average PSD, mean, variance, Shannon's entropy, zero crossing	LSSVM	Joy, peace, anger, and depression	65
[39]	EDA, HR	No specific features	Fuzzy logic	Stress	99.5
[40]	EEG	No specific features	Correlation analysis	Neutral, anger, sadness, anxiety, disgust, and surprise	90

Nomenclature for signal modalities: RSP denotes relative spectral power, EEG denotes electroencephalogram, ECG denotes electrocardiogram, GSR denotes galvanic skin response, EDA denotes electrodermal activity, BVP denotes blood volume pulse, HR/HP denotes heart rate/pulse, and Temp denotes temperature. Nomenclature for classifiers: LDA denotes latent discriminant analysis, KNN denotes K-nearest neighbors, ANN denotes artificial neural network, SVM denotes support vector machine, HMM denotes hidden Markov model, LSTM denotes long-short-term memory, DFA denotes deterministic finite automata, MESAE denotes multiple fusion layer based-ensemble classifier of stacked autoencoder, and MEMD denotes multienncoder to multidecoder.

3.1. *Tools.* Different libraries and packages are given below, which are utilized in actual implementation of the proposed system as well as in its validation and testing.

- (1) *NumPy* and *Pandas*. In Python for array processing, mathematical computation, and data sciences, special packages NumPy and Pandas are utilized.
- (2) *h5py*. This library uses traditional batch processing and makes Python compatible with a huge amount of numerical data in HDF5 format. This library takes burden of the system in training our models, especially in case of physiological signals.
- (3) *Matplotlib*. This library is utilized for generating and plotting different graphs and charts for the visualization of results generated by employing the proposed system.
- (4) *Sklearn*. This tool serves as a main constituting part behind generation of confusion matrix and other related metrics. These metrics actually help us in figuring out the actual results generated by our models. These metrics also help in verifying

the authenticity and validation of generated results.

- (5) *Think Gear*. It is a library provided by Neurosky to connect and communicate data between Bluetooth-enabled system and MindWave Mobile EEG device. This library employs COMM port for establishing connection and communication.
- (6) *Neurosky MindWave Mobile EEG Headset*. This headset is used to capture electroencephalogram signal produced by brain as a result of brain activity. It is a single-channel device and it is able to provide 12-bit raw EEG signals with sampling rate  $F_s$  of 512 Hz and band range of 3–100 Hz.

Section 4 covers the details more comprehensively.

3.2. *Data Set.* The proposed system is a novel indigenous system; therefore, no data set is present. So, the first task that needs to be accomplished is to collect and organize the data set with proper labels for the proposed system. Collected data set contains data related to two classes: “with analytical

skills” and “without analytical skills.” For collecting this data, proper experimental setup is created where participants have given analytical reasoning test to solve. While they are solving the test, our system collected the data; this data is then assigned proper labels, which is then utilized in training and validation of our models. See Figure 1 for depicting overall structure of the data flow used in collecting data set, while the rest of the details are provided briefly in Section 4. For availability of data, see Section 7.

#### 4. Proposed System

The proposed architecture is comprised of four different layers illustrated in Figure 2. The first layer is the *sensor and communication layer* which is responsible for capturing various EEG power spectrums. Traditional head gear comprised of multiple electrodes and channels is unfeasible in real-world scenarios due to its preparation and positioning time. Additionally, it is uncomfortable to wear for long periods of time and requires supervised use from trained personnel. Comfortable and cheaper commodity hardware has started to be popular in the last decade. Examples are the Neurosky MindWave Mobile EEG headset, which is popularly used in the entertainment and gaming sector, as well as development of motor development skills in children. This headset is a single-channel device and is able to provide 12-bit raw EEG signals with sampling rate  $F_s$  of 512 Hz and band range of 3–100 Hz. In this case, this device is a server system configured to use the Think-Gear library configured to work with Python. The received raw signals using EEG headset constitute input to this layer and are stored in CSV format.

The second *preprocessing layer* is responsible for cleaning the acquired signal using a number of DSP filters before acquisition of the relevant features. This is essential because when the EEG device captures neuron activity, it also captures crossover noise and other electrical activities within proximity of the electrode point (which may include muscle activity). Since the EEG is a composite signal, its constituent alpha, beta, gamma, delta, and theta wave patterns can be acquired by application of the fast Fourier transform, followed by a bandpass filter of relevant frequency range. An illustration of the complex EEG signal after Fourier transform is given in Figure 3, where imaginary and real parts of signal are superposed on its own amplitude for comparison purpose. Figure 3 also shows the different frequency bands given in Table 1. Here, the amplitude of frequency bands decreases exponentially as their frequency range decreases. This makes lower frequency ranges prone to noise. Effect of crossover noise is mitigated by application of a mean filter. Small and local noise sources (attributed to eye blinking, heart pumping, etc.) are removed using an Independent Component Analysis (ICA) filter, resulting in same amplitude scales for all frequency bands (see Figure 4). It takes raw signal stored in CSV file (generated by sensor and communication layer) as input and after performing necessary processing it stores the output in another CSV file. This newly generated CSV file is then fed as input to the decision layer. The *decision layer* bears three sublayers: *data*

*set*, *modality transform*, and *decision* sublayers. First is the *data set sublayer*. Here, a data set is prepared, comprising the preprocessed signal and its associated labels. This data set is used for training and as input for various machine learning models. The ground truth for the data set is determined experimentally using an alternate work flow and makes use of a MindWave Mobile EEG device (see Figure 1). Experiments are designed, where the state of neuron activity is measured, while the subject is performing analytical tasks. Some example analytical tasks are discussed in [42]. For the work at hand, the authors prepared a test comprising questions of analytical portion of the International GRE. These questions were then given to subject participants to solve in a fixed time interval while being attached to the EEG device. Tests are scored afterwards, and a threshold value is used to determine whether the acquired data belongs to a subject participant *with* or *without* analytical skills. The data set also includes factors such as humidity, mean of ambient noise levels, and room temperature at the time acquisition was being made. The ground truth is collected from male and female candidates aged between 22 and 45 years. The candidates had normal vision and hearing and are free from any kind of neurological disorder. In the *learning* sublayer, the trained model is then used formally for classification using a number of schemes such as DT, KNN, SVM, NB, CNN, and LSTM. For KNN, SVM, and DT, hand-crafted features of frequency domain are used to prepare a feature vector. These include the minimum, maximum, and mean frequencies, as well as their standard deviation. Given that the EEG contains five subbands, this gives a total of 20 hand-crafted features. The authors have built CNN model using Convolutional (ReLU activation), max-pooling, dropout, dense, flattened, and fully connected layers with ReLU and Softmax as activation functions (see Figure 5). The LSTM model includes the average pooling, dense, flattened, dropout, and fully connected layers with sigmoid and hyperbolic tangent ( $\tanh$ ) as activation functions (see Figure 6). These models are implemented using Keras and TensorFlow.

The final layer in the architecture is *Output Layer*, where the decision of the classifier is validated. Interfaces of the architecture support exchanges, transformations, processing, and classification in real time.

#### 5. Results and Discussion

For brevity, the labels *with* and *without* analytical skills are treated as *positive* and *negative* labels, respectively. Using this nomenclature, evaluation can be based on measures of True Positive (TP), i.e., correctly identified positive labels, and True Negative (TN), i.e. correctly identified negative labels. In contrast, we also have False Positive (FP), i.e., positive labels identified as negative labels, and False Negative (FN), i.e., negative labels identified as positive labels. In addition, other metrics such as specificity, recall, precision, and F1-scores can also be formulated. The exact calculation of these measures is given in Table 3. Apart from these metrics, for better understanding of classification probability, receiver operating characteristic (ROC) curve is also being computed. It is a plot of true positive rate (TPR)

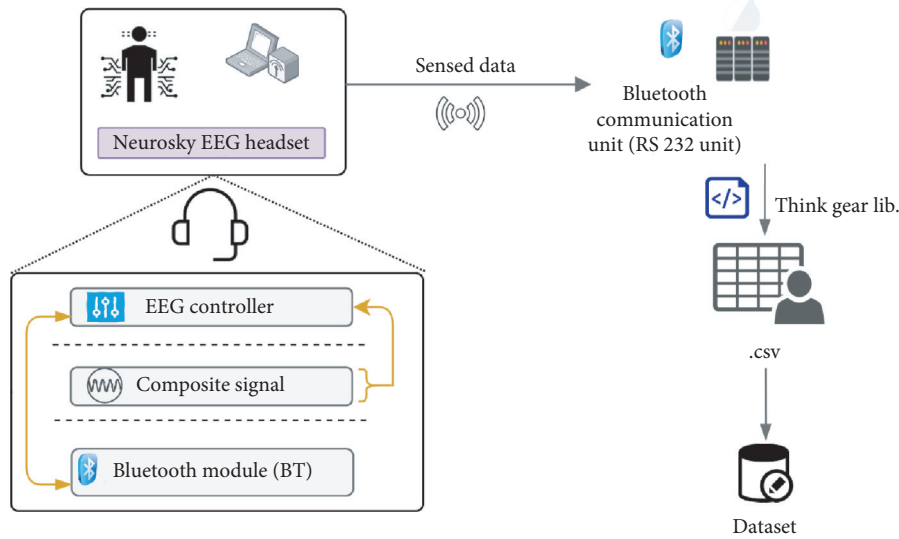


FIGURE 1: Ground truth acquisition work flow.

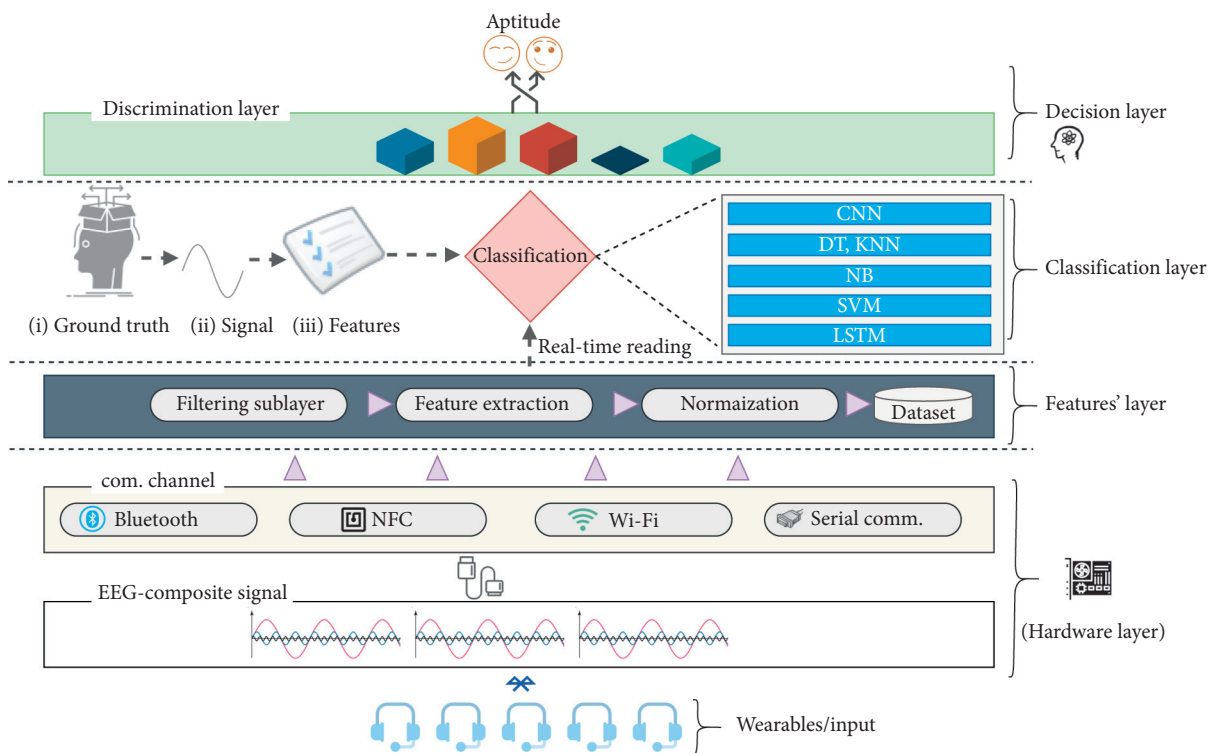


FIGURE 2: Proposed system architecture.

against false positive rate (FPR). The area under ROC curve shows the classification probability of the models. The greater area means better true positive rate and better classification ability of a model. As can be seen in Figure 7, for validation, a tenfold cross-validation technique is utilized. All signals irrespective of class labels are randomly assigned to ten equal-sized chunks. Of these, training is performed using 9 chunks, while the remainder is used for validation. The process is repeated for 250 epochs for each

model. At the end of each epoch, parameters such as accuracy, validation loss, and confusion matrices are extracted. The four labels TP, FP, TN, and FN are then obtained from this confusion matrix. Subsequently, the scores are given in Table 3.

A number of machine learning models were used to perform the classification as depicted in [43, 44], and the maximum, minimum, and average accuracy after 250 epochs are reported in Table 4.

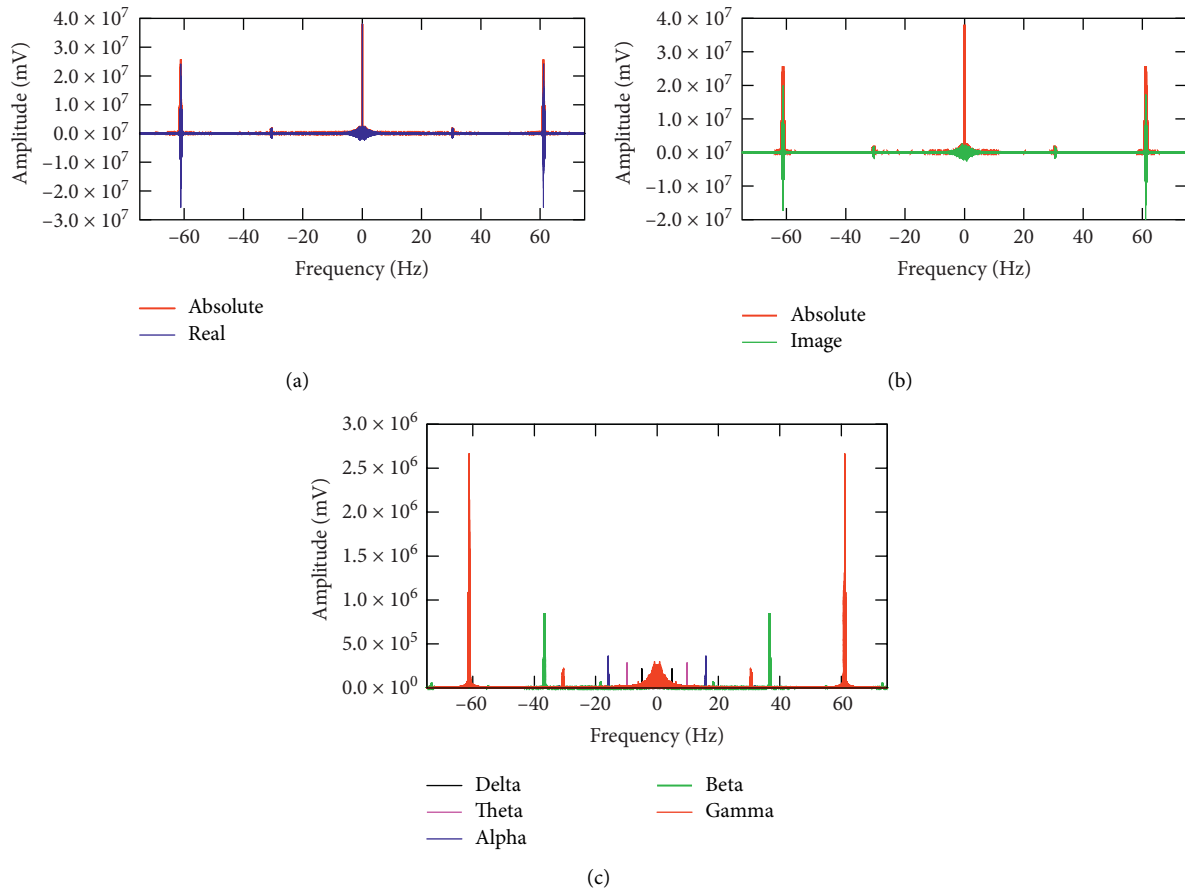


FIGURE 3: (a) Real and (b) imaginary components of transformed EEG signal superposed on magnitude of itself. (c) Illustration of diminishing amplitude for lower-frequency wavebands of the EEG signal.

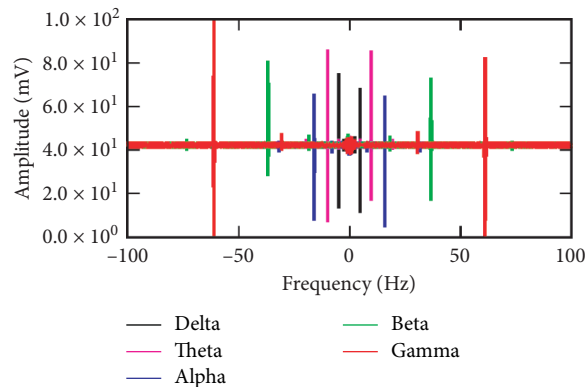


FIGURE 4: ICA-treated wavebands, showing the same amplitude scales as EEG wavebands.

For each of the four labels outlined earlier, the confusion matrix and scores of Table 3 are given in Tables 5 and 6, respectively.

The best results reported are those for LSTM with maximum and average validation accuracy of 100% and 75%, respectively, and with a consistent F1-score,

precision, and specificity of 0.91, 0.99, and 0.99, respectively. SVM provided maximum accuracy of 97%, while its average accuracy was 92%. Its F1-score, precision, and specificity were quite close at 0.93, 0.93, and 0.92, respectively. KNN and DT provide a maximum accuracy of 95%. Their average accuracy is at 89% and 90% (see

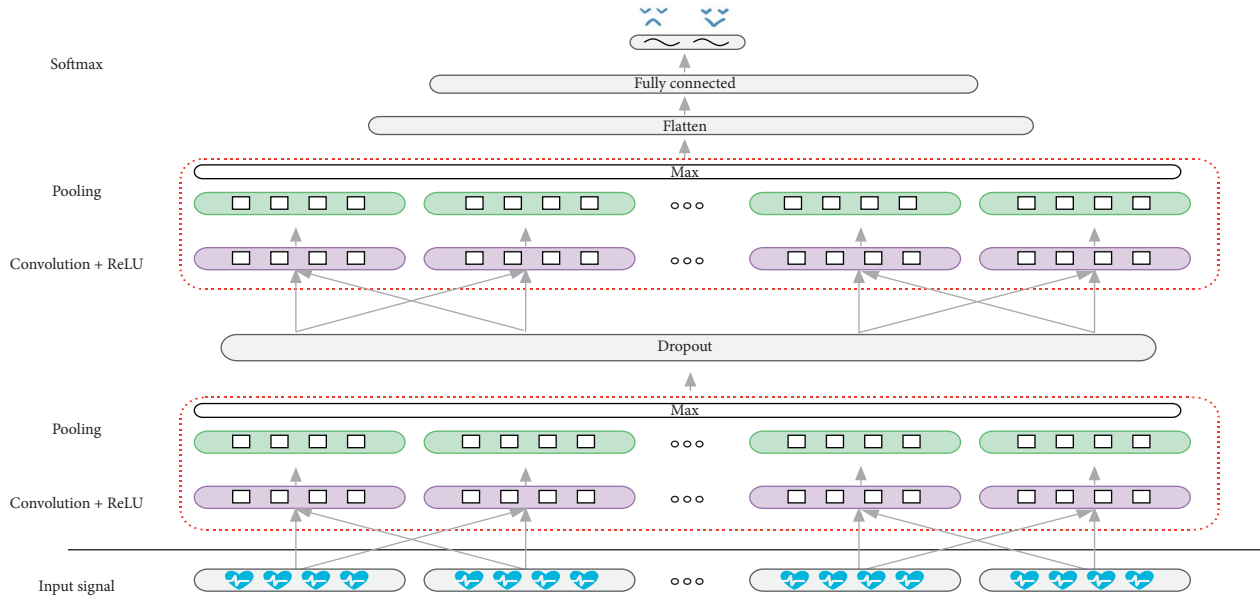


FIGURE 5: CNN layer architecture.

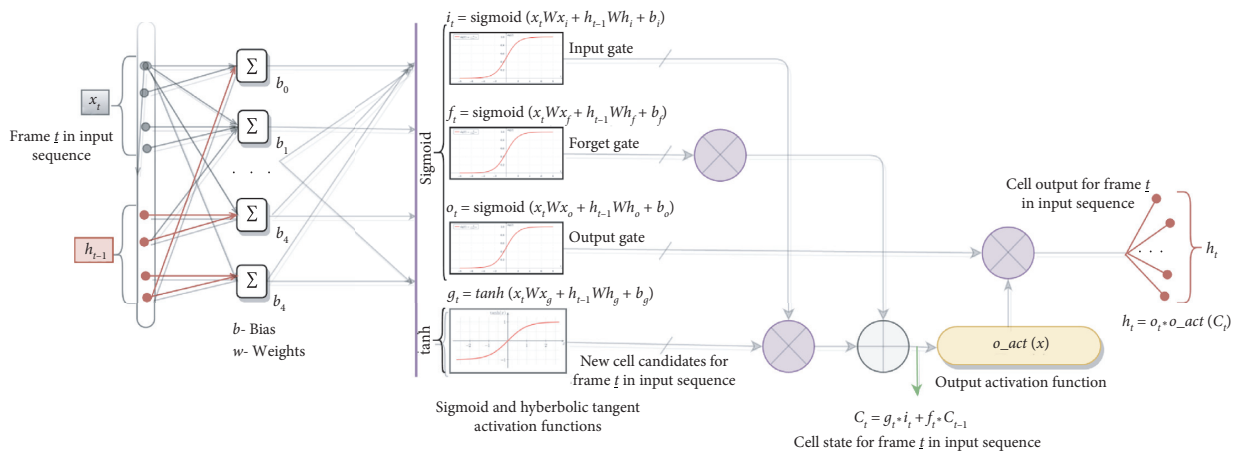


FIGURE 6: LSTM architecture.

TABLE 3: Scoring measures.

Evaluation metric	Evaluation formula
Specificity ( $S$ )	False Positive / (True Negative + False Positive)
Recall ( $R$ )	True Positive / (True Positive + False Negative)
Precision ( $P$ )	True Positive / (True Positive + False Positive)
F1-score	$2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$

Figure 8). However, their F1-score, precision, and specificity are quite less than those of LSTM and SVM (see Figure 9). The architecture of CNN used in this

manuscript gave a maximum accuracy of 99% but a poor average accuracy of 54%. The other scores of CNN were also not consistent. NB scores were not compared to other



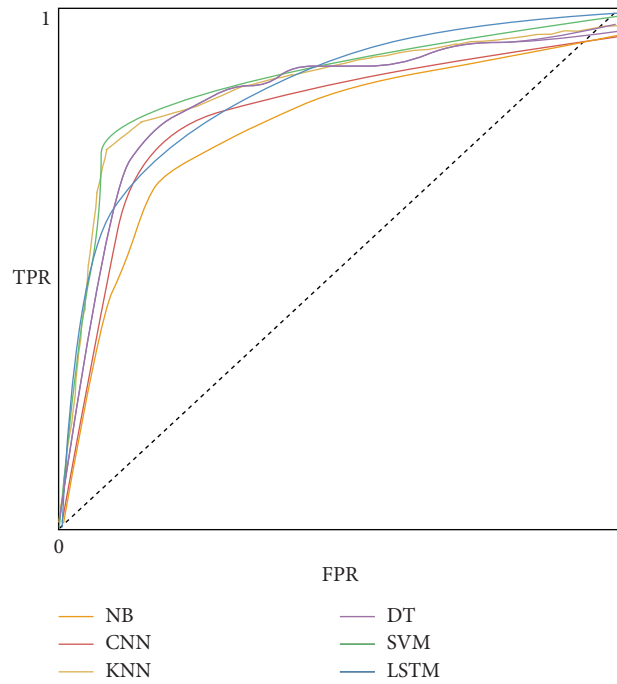


FIGURE 7: Receiver operating characteristic curves for all six models.

TABLE 4: Accuracy for various models.

Model	Maximum	Minimum	Average
Convolutional neural network (training)	81.9	75.2	77.9
Convolutional neural network (validation)	99.2	0	54.2
Decision tree	95.0	86.0	90.5
<i>K</i> -nearest neighbors	95.4	82.6	88.8
Long-short-term memory (training)	94.4	90.4	94.4
Long-short-term memory (validation)	100	0	75.0
Naïve Bayes	76.4	69.2	72.8
Support vector machine	97.6	86.2	92.0

TABLE 5: Confusion matrix.

Model	True Positive	False Negative	False Positive	True Negative
Convolutional neural network	642	144	129	443
Decision tree	706	64	65	525
<i>K</i> -nearest neighbors	696	77	75	512
Long-short-term memory	767	132	5	456
Naive Bayes	513	114	256	477
Support vector machine	723	54	54	529

TABLE 6: Averaged F1-score, precision, recall, and specificity scores for 250 epochs.

Model	F1-score	Precision	Recall	Specificity
Convolutional neural network	0.82	0.83	0.81	0.77
Decision tree	0.91	0.91	0.92	0.89
<i>K</i> -nearest neighbors	0.90	0.90	0.90	0.87
Long-short-term memory	0.91	0.99	0.85	0.99
Naïve Bayes	0.72	0.64	0.80	0.65
Support vector machine	0.93	0.93	0.93	0.92

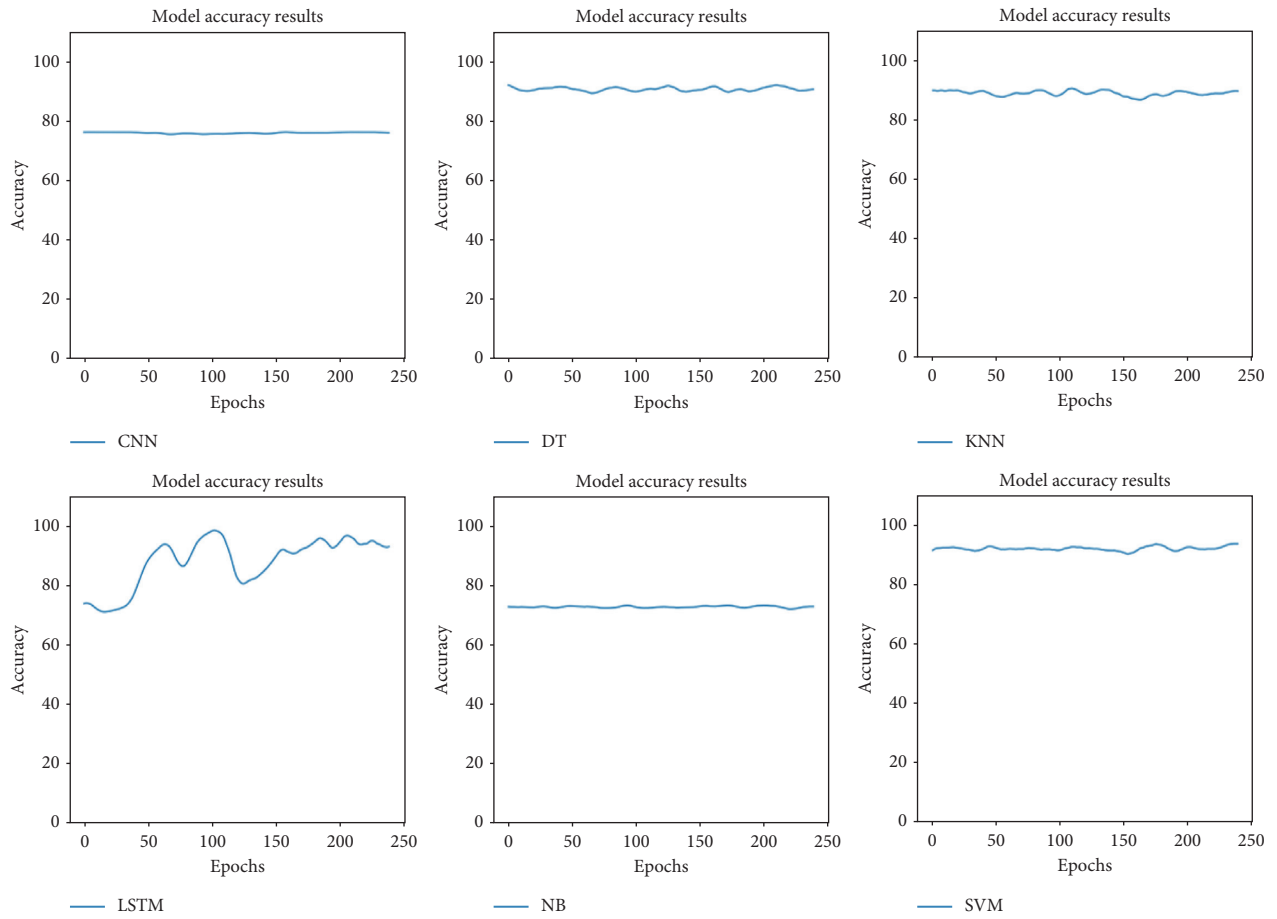


FIGURE 8: Accuracy results of all six models (% age).

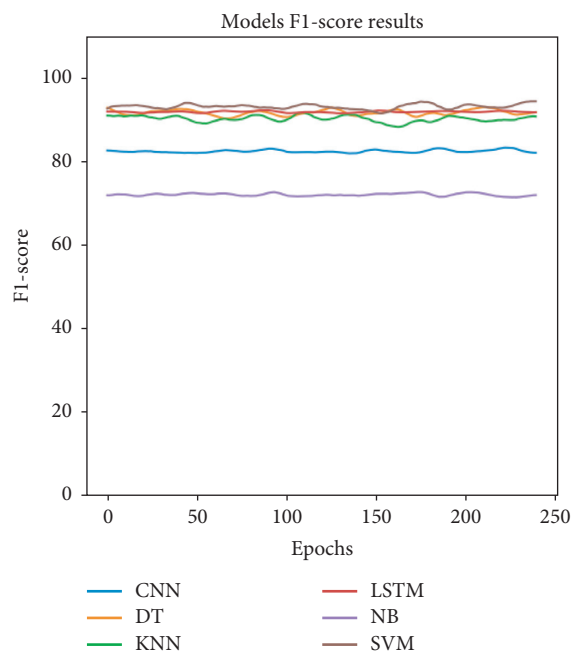


FIGURE 9: F1-scores of all six models (% age).

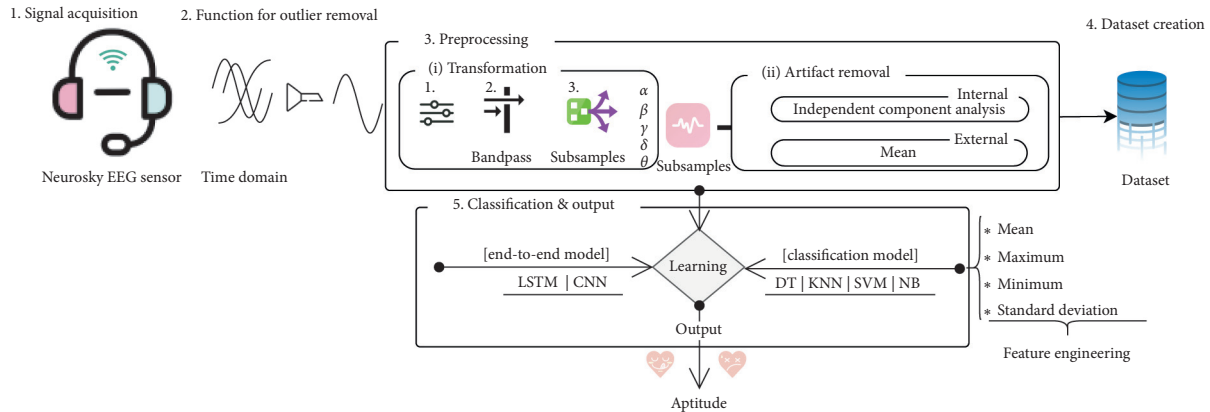


FIGURE 10: Implementation architecture for the proposed system.

models due to high invariance in data. In this study, the implementation pipeline that has been finalized is provided (see Figure 10).

## 6. Conclusion

Aptitude is an innate skill to perform a particular task with ease and perfection. It not only plays a vital in enhancing productivity but also regulates the stress level of employees in work environments. This research work addresses the utilization of aptitude to regulate the stress in a working environment. It is an established fact that if an employee is assigned job roles according to his working aptitude, it helps in reducing stress and fatigue caused by improper job role assignments and overburdening. Keeping this fact in view, an implementation pipeline that makes use of an EEG signal for the detection of aptitude is proposed with detailed implementation. The proposed pipeline is tested with different types of machine learning models. Our findings show good results with LSTM- and SVM-based classifiers, giving achieved accuracy of 94% and 97%, with F1-scores of 0.91 and 0.93, respectively. In this research work, our main focus was on analytical skills of workers. For future work, the binary system can be expanded to include poor, fair, good, better, and outstanding analytical capabilities. Other aptitude facets such as IQ, dexterity, and reasoning can also be work for the future.

## Data Availability

The data set used in this work is propriety data that belongs to institution. Soon after completion of this research work, these data will be made publicly available using GitHub or any other available resource. However, in the meantime, data will be provided upon sending a request to tehsen.khan@nu.edu.pk. Data will only be provided for enhancing the research in this domain only and the requester will clearly mention the purpose of making the request for data. The request should be submitted using an institutional e-mail only.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

- [1] H. Alemdar and C. Ersoy, "Wireless sensor networks for healthcare: a survey," *Computer Networks*, vol. 54, no. 15, pp. 2688–2710, 2010.
- [2] B.-S. Lin, A. M. Wong, and K. C. Tseng, "Community-based ECG monitoring system for patients with cardiovascular diseases," *Journal of Medical Systems*, vol. 40, no. 4, p. 80, 2016.
- [3] T. F. Quatieri, J. R. Williamson, C. J. Smalt et al., "Using EEG to discriminate cognitive workload and performance based on neural activation and connectivity," Technical Report, MIT Lincoln Laboratory, Lexington, MA, US, 2016.
- [4] Y. Li, J. Pan, J. Long et al., "Multimodal BCIS: target detection, multidimensional control, and awareness evaluation in patients with disorder of consciousness," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 332–352, 2015.
- [5] A. Mednis, G. Strazdins, R. Zviedris, G. Kanonirs, and L. Selavo, "Real time pothole detection using android smartphones with accelerometers," in *Proceedings of the 2011 International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS)*, pp. 1–6, IEEE, Barcelona, Spain, June 2011.
- [6] T. Ojha, S. Misra, and N. S. Raghuvanshi, "Wireless sensor networks for agriculture: the state-of-the-art in practice and future challenges," *Computers and Electronics in Agriculture*, vol. 118, pp. 66–84, 2015.
- [7] N. C. Batista, R. Melício, J. C. O. Matias, and J. P. S. Catalão, "Photovoltaic and wind energy systems monitoring and building/home energy management using zigbee devices within a smart grid," *Energy*, vol. 49, pp. 306–315, 2013.
- [8] K. K. Khedo, R. Perseedoss, A. Mungur et al., "A wireless sensor network air pollution monitoring system," *International Journal of Wireless & Mobile Networks*, vol. 2, no. 2, p. 15, 2010.
- [9] N. Maisonneuve, M. Stevens, M. E. Niessen, P. Hanappe, and L. Steels, "Citizen noise pollution monitoring," in *Proceedings of the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government*, pp. 96–103, Puebla, Mexico, May 2009.
- [10] R. Tan, G. Xing, J. Chen, W.-Z. Song, and R. Huang, "Quality-driven volcanic earthquake detection using wireless sensor networks," in *Proceedings of the 2010 31st IEEE Real-Time Systems Symposium*, pp. 271–280, IEEE, San Diego, CA, USA, December 2010.
- [11] Z. Mohammadi, J. Frounchi, and M. Amiri, "Wavelet-based emotion recognition system using EEG signal," *Neural Computing & Applications*, vol. 28, no. 8, pp. 1985–1990, 2017.

- [12] M. Ali, A. H. Mosa, F. A. Machot, and K. Kyamakya, "Emotion recognition involving physiological and speech signals: a comprehensive review," in *Recent Advances in Nonlinear Dynamics and Synchronization*, pp. 287–302, Springer, Berlin, Germany, 2018.
- [13] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *Emotion*, vol. 8, no. 10, pp. 355–358, 2017.
- [14] N. Zhuang, Y. Zeng, L. Tong, C. Zhang, H. Zhang, and B. Yan, "Emotion recognition from EEG signals using multidimensional information in EMD domain," *BioMed Research International*, vol. 2017, Article ID 8317357, 9 pages, 2017.
- [15] H. Zhou and B. Goold, "A domestic adaptable infant monitoring system using wireless sensor networks," in *Proceedings of the IEEE 34th International Performance Computing and Communications Conference (IPCCC)*, pp. 1–2, IEEE, Nanjing, China, December 2015.
- [16] W. Dargie and M. Zimmerling, "Wireless sensor networks in the context of developing countries," in *Proceedings of the IFIP World IT Forum (WITFOR)*, Addis Ababa, Ethiopia, August 2007.
- [17] P. Kumari, L. Mathew, and P. Syal, "Increasing trend of wearables and multimodal interface for human activity monitoring: a review," *Biosensors and Bioelectronics*, vol. 90, pp. 298–307, 2017.
- [18] M. Elgendi, A. Al-Ali, A. Mohamed, and R. Ward, "Improving remote health monitoring: a low-complexity ECG compression approach," *Diagnostics*, vol. 8, no. 1, p. 10, 2018.
- [19] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable eda device," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 410–417, 2009.
- [20] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.
- [21] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 1672–1687, 2004.
- [22] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: first steps towards an automatic system," in *Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems*, pp. 36–48, Springer, Kloster Irsee, Germany, June 2004.
- [23] W. Wan-Hui, Q. Yu-Hui, and L. Guang-Yuan, "Electrocardiography recording, feature extraction and classification for emotion recognition," in *Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering*, pp. 168–172, IEEE, Los Angeles, CA, USA, April 2009.
- [24] C. Maaoui and A. Pruski, "Emotion recognition through physiological signals for human-machine communication," in *Cutting Edge Robotics 2010*IntechOpen, London, UK, 2010.
- [25] J. Kim, "Bimodal emotion recognition using speech and physiological changes," in *Robust Speech Recognition and Understanding*IntechOpen, London, UK, 2007.
- [26] W. Yang, M. Rifqi, C. Marsala, and A. Pinna, "Physiological-based emotion detection and recognition in a video game context," in *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, Rio de Janeiro, Brazil, July 2018.
- [27] R. Mahajan, "Emotion recognition via EEG using neural network classifier," in *Soft Computing: Theories and Applications*, pp. 429–438, Springer, Berlin, Germany, 2018.
- [28] J. Kavanagh, "Determinants of productivity for military personnel. A review of findings on the contribution of experience, training, and aptitude to military performance," Technical Report D, Rand National Defense Research Institution, Santa Monica, CA, USA, 2005.
- [29] S. Cartwright and C. L. Cooper, *Managing Workplace Stress*, Sage, Thousand Oaks, CA, USA, 1997.
- [30] J. E. Hunter, "Cognitive ability, cognitive aptitudes, job knowledge, and job performance," *Journal of Vocational Behavior*, vol. 29, no. 3, pp. 340–362, 1986.
- [31] S. Michie, "Causes and management of stress at work," *Occupational and Environmental Medicine*, vol. 59, no. 1, pp. 67–72, 2002.
- [32] M. Tehseen, H. Javed, A. Mehmood, M. Amin, I. Hussain, and B. Jan, "Multi modal aptitude detection system for smart office," *IEEE Access*, vol. 7, pp. 24 559–24 570, 2019.
- [33] S. Thejaswini, K. M. Ravi Kumar, S. Rupali, and V. Abijith, "EEG based emotion recognition using wavelets and neural networks classifier," *Cognitive Science and Artificial Intelligence*, Springer, Berlin, Germany, pp. 101–112, 2018.
- [34] J. Sorinas, M. D. Grima, J. M. Ferrandez, and E. Fernandez, "Identifying suitable brain regions and trial size segmentation for positive/negative emotion recognition," *International Journal of Neural Systems*, vol. 29, no. 2, Article ID 1850044, 2019.
- [35] H. Huang, Q. Xie, J. Pan et al., "An EEG-based brain computer interface for emotion recognition and its application in patients with disorder of consciousness," *IEEE Transactions on Affective Computing*, vol. 99, 2019.
- [36] X. Li, R. La, Y. Wang et al., "EEG-based mild depression recognition using convolutional neural network," *Medical, & Biological Engineering & Computing*, vol. 57, no. 6, pp. 1341–1352, 2019.
- [37] D. Kulic and E. A. Croft, "Affective state estimation for human-robot interaction," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007.
- [38] J. Cai, W. Chen, and Z. Yin, "Multiple transferable recursive feature elimination technique for emotion recognition based on EEG signals," *Symmetry*, vol. 11, no. 5, p. 683, 2019.
- [39] A. de Santos Sierra, C. Sanchez Avila, J. Guerra Casanova, and G. Bailador del Pozo, "A stress-detection system based on physiological signals and fuzzy logic," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 10, pp. 4857–4865, 2011.
- [40] N. K. Al-Qazzaz, M. K. Sabir, and K. Grammer, "Correlation indices of electroencephalogram-based relative powers during human emotion processing," in *Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology*, pp. 64–70, ACM, Tokyo Japan, March 2019.
- [41] M. W. Miller, J. C. Rietschel, C. G. McDonald, and B. D. Hatfield, "A novel approach to the physiological measurement of mental workload," *International Journal of Psychophysiology*, vol. 80, no. 1, pp. 75–78, 2011.
- [42] G. Funke, B. Knott, V. F. Mancuso et al., "Evaluation of subjective and EEG-based measures of mental workload," in *Proceedings of the International Conference on Human-Computer Interaction*, pp. 412–416, Springer, Las Vegas, NV, USA, July 2013.
- [43] R. Lacuesta, L. Garcia, I. García-Magariño, and J. Lloret, "System to recommend the best place to live based on wellness state of the user employing the heart rate variability," *IEEE Access*, vol. 5, pp. 10 594–10 604, 2017.
- [44] L. García, L. Parra, O. Romero, and J. Lloret, "System for monitoring the wellness state of people in domestic environments employing emoticon-based HCI," *The Journal of Supercomputing*, vol. 75, no. 4, pp. 1869–1893, 2019.

## Research Article

# Optimal Policy Learning for Disease Prevention Using Reinforcement Learning

Zahid Alam Khan,<sup>1</sup> Zhengyong Feng,<sup>1</sup> M. Irfan Uddin ,<sup>2</sup> Noor Mast,<sup>2</sup> Syed Atif Ali Shah ,<sup>3,4</sup> Muhammad Imtiaz,<sup>5</sup> Mahmoud Ahmad Al-Khasawneh ,<sup>4</sup> and Marwan Mahmoud <sup>6</sup>

<sup>1</sup>China West Normal University, Nanchong, China

<sup>2</sup>Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan

<sup>3</sup>Faculty of Engineering and Information Technology, Northern University, Nowshera, Pakistan

<sup>4</sup>Faculty of Computer and Information Technology, Al-Madinah International University, Kuala Lumpur, Malaysia

<sup>5</sup>Faculty of Computer Science, University of Swabi, Swabi, Pakistan

<sup>6</sup>King Abdulaziz University, Jeddah, Saudi Arabia

Correspondence should be addressed to M. Irfan Uddin; irfanuddin@kust.edu.pk

Received 18 February 2020; Accepted 1 October 2020; Published 28 November 2020

Academic Editor: Shaukat Ali

Copyright © 2020 Zahid Alam 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.

Diseases can have a huge impact on the quality of life of the human population. Humans have always been in the quest to find strategies to avoid diseases that are life-threatening or affect the quality of life of humans. Effective use of resources available to human to control different diseases has always been critical. Researchers are recently more interested to find AI-based solutions to control the human population from diseases due to the overwhelming popularity of deep learning. There are many supervised techniques that have always been applied for disease diagnosis. However, the main problem of supervised based solutions is the availability of data, which is not always possible or not always complete. For instance, we do not have enough data that shows the different states of humans and different states of environments, and how all different actions taken by humans or viruses have ultimately resulted in a disease that eventually takes the lives of humans. Therefore, there is a need to find unsupervised based solutions or some techniques that do not have a dependency on the underlying dataset. In this paper, we have explored the reinforcement learning approach. We have tried different reinforcement learning algorithms to research different solutions for the prevention of diseases in the simulation of the human population. We have explored different techniques for controlling the transmission of diseases and its effects on health in the human population simulated in an environment. Our algorithms have found out policies that are best for the human population to protect themselves from the transmission and infection of malaria. The paper concludes that deep learning-based algorithms such as Deep Deterministic Policy Gradient (DDPG) have outperformed traditional algorithms such as Q-Learning or SARSA.

## 1. Introduction

Different types of diseases such as malaria, flu, dengue, and HIV have a huge impact on the quality of life of the human population [1–3]. If we consider malaria only, then according to the World Health Organization's report, approximately 3.2 billion people are infected with malaria. As per their report, in 2016 and 2017, there were 217 and 219

million malaria cases reported, which shows an increase in malaria cases in recent years [4]. Therefore, effective use of resources to get malaria under control has been critical. Insecticide-Treated Nets (ITNs) are the primary method of malaria prevention [5] because there is a type of mosquito called the anopheles mosquito; it bites after 9 p.m. When a mosquito sets on the net, it dies due to the insecticide, which disrupts the reproductive cycle. In addition to ITNs, the

other malaria preventive policies include Indoor Residual Spraying (IRS) [6], larvicide [7] in bodies of water, and malaria vaccination [8–11].

Machine Learning algorithms are applied in different domains and have made tremendous progress [12] where healthcare sector is particularly influenced by machine learning [13–15] in the past few years. These machine learning algorithms are focusing on the diagnosis of diseases [16] or forecasting future results [17], but the treatment of diseases is not explored [18]. It is a very important step to diagnose a disease and is considered as an important step to treat diseases, and machine learning techniques can support healthcare professionals in the treatment to some extent, but it has been a challenging problem to find the best policy to treat patients for medical professionals [19]. Recently, much popularity is gained by reinforcement learning (RL) [20] in video games [21–23], where good and bad actions are learned by the agent through interactions with the environment and the response of the environment. In the context of video games, RL has performed very well, but limited progress has been made in real-world domains like health care. In video games such as AlphaGo and StarCraft, the agent plays a large number of actions in the environment and learns the optimal policy. However, in the context of health care, it is considered unethical to use humans to train RL algorithms and not to mention that this process would be costly and takes years to complete. We are not able to observe everything happening the body of a person. We can measure blood pressure, temperature, and some other measurements at different intervals of time, but these measurements do not represent the complete state of a patient. Similarly, the data collected in health care about patients may exist for one time and may not exist for others. For example, chest X-rays that are used in the treatment of pneumonia [24] are collected before a person is infected and after the person is cured, but the RL model has to know all the estimates of the states the patient goes through. It is very challenging in health care where there are many unknown facts about patients at all time steps.

Reward function is one of the most important functions in RL, and it is challenging in many real-world applications to find a good reward function. In health care, it is even more challenging to search for the reward function that keeps balance between short-term success and overall long-term improvements. For example, in case of sepsis [25], improvements in blood pressure at different durations of time may not cause improvement in the overall success. Similarly, having only a single high reward at the end of an episode (i.e., survived or died) demonstrates that a long route is followed without different intermediary rewards [26, 27]. It is also difficult to know what actions result in reward and what actions result in penalty. All the major breakthroughs are possible by using simulated data in deep RL that is equal to many actual years [28]. When data are generated through simulators, it is not a problem, but in case of health care, it is not possible to generate simulated data for the treatment of different diseases. Generally, the data are very scarce to start with training supervised learning, and the data that exist take efforts to annotate to be used for supervised learning.

Furthermore, hospitals are not willing to share data of patients mainly because of privacy reasons. All these facts further make the application of deep RL to health care challenging.

By nature, the health care data is nonstationary and dynamic [29]. For example, it is possible that patients' symptoms are stored at different intervals of time and maybe different records are stored for different patients. Over time, the objectives of treatments may also change. In literature, different studies [30–32] are focused on reducing the overall mortality. When the condition of a person improves, the focus shifts to a different objective such as the duration of the virus staying in the body. Similarly, viruses or infections may change much more rapidly and may evolve in different dynamics [33–35] that are most probably not observed in the training data used for supervised or semisupervised learning algorithms. Decision-making in medical diagnosis is inherently sequential [36, 37]. It means that a patient visits a health care centre for the treatment of a disease. The doctor, based on previous experiences, decides a treatment to be followed. Later, when the patient returns to the same doctor, the treatment that was previously suggested by the doctor decides the current state of the patient and also helps the doctor in which decision needs to be taken next. In the existing state-of-the-art AI strategies of dealing with disease treatment [38, 39], the sequential nature of the decisions is ignored [40]. These AI systems make decisions on the basis of the present state of the patients. The sequential nature of medical treatment can be effectively modelled as Markov Decision Process (MDP) [41–44] and better solved through RL. The RL algorithms will not only consider the instantaneous outcomes of treatment but also the long-term benefits of the patients [45].

An intervention of actions to avoid malaria are systematically explored in this paper. The paper demonstrates a real-world example of reinforcement learning, where simulated humans are trained to learn an effective technique to avoid malaria. In the literature, AI techniques are used for the prediction, diagnosis, and healthcare planning, but this paper takes a different approach by simulating an environment and using simulated humans to use different reinforcement learning techniques to avoid malaria. A combination of interventions is explored to control the transmission of malaria and learn techniques for malaria avoidance.

The paper is organized as follows: the related works are explained in Section 2. The problem of malaria avoidance and the methodology of reinforcement learning are given in Section 3. Experiments are performed, and their results are analysed in Section 4. Concluding remarks of the paper are given in Section 5.

## 2. Related Work

Recent advancements in machine learning and big data have motivated researchers of different domains to use these algorithms in their problems. Biomedical and health care researchers are getting benefits from these algorithms in early disease recognition, community services, and patients

care. In [46], machine learning and MapReduce algorithms are used to effectively predict different diseases in disease-frequent societies. The paper demonstrated to achieve 94.8% accuracy and convergence speed that is faster than CNN (Convolutional Neural Network) based algorithms. Similarly, deep learning and big data techniques have been used in [47] to predict infectious diseases. The authors have combined Deep Neural Network (DNN) and Long Short-Term Memory (LSTM) and evaluated the performance with Autoregressive Integrated Moving Average (ARIMA) in making the prediction of different diseases one week in the future. Better results have been achieved compared to ARIMA. Automatic diagnosis of malaria enables us to provide reliability in health care services to areas where resources are limited. Machine learning techniques have been tried to investigate the process of automating malaria detection. In [48], malaria classification is performed using CNN. Similarly, in [49], CNN has been used to detect malaria classification and has demonstrated promising accuracy. Deep reinforcement learning (DRL) has recently attained remarkable success, notably in complex games like Atari, Go, and Chess. These achievements are mainly possible because of the powerful function approximation with the help of DNN. DRL has been proved as an effective method in the medical context. Several applications of RL have been found in the context of medicine. For instance, RL methods have been used to develop strategies of treatment for epilepsy [50] and lung cancer [51]. Authors have used the sepsis dataset which is a subset of the MIMIC-III dataset [25]. An action space consisting of vasopressors and IV fluid is selected. Each drug of varying amount is grouped in four bins. Double Deep Q-Network is used for the evaluation. SOFA score which is used for measurements of organ failure is used for the reward function. U-curve is used for evaluation. The mortality rate is used as a function of dosage of policy prescription versus the policy that is actually followed.

In [19], DRL is used to develop a framework that predicts an optimal strategy to deal with Dynamic Treatment Regimes using medical data. The paper has claimed that their RL model is more flexible and adaptive in high dimensional action and state spaces compared to other RL based approaches. The framework models real-world complexity in helping doctors and patients to make a personalized decision in making treatment choices and disease progression. The framework combines supervised learning and DRL using DNN. The dataset is taken from the database of the Centre for International Bone Marrow Transplant Research (CIBMTR) registry. The framework has demonstrated achieving promising accuracy to predict a human doctor's decision and at the same time compute a high reward function. In [52], an RL system is developed that helps diabetes patients to engage in different physical activities. Messages sent to patients were made personalized to patients and the results have demonstrated that participants receiving messages with the RL algorithm increased the number of physical activities and walking speed. A supervised RL with recurrent neural network (SRL-RNN) is combined in a framework to make different treatment recommendations by Wang et al. in [53]. Their results of experiments conducted on MIMIC-3 dataset

have demonstrated that the RL based framework can reduce the estimated mortality and at the same time provide promising accuracy to match doctor's prescriptions. In [54], the authors describe a novel technique that can find the optimal policy that can treat patients with chemo using RL. The authors have used Q-Learning, and, for the action space, a mechanism is used to quantify doses for a given time period that an agent can choose from. The cycle of dose is initiated with a frequency as determined by an expert. At the end of each cycle, transition states are compared. The mean reduction in tumour diameter determines the reward function. Simulated clinical trials are used for the evaluation of the algorithm.

In [55], the authors have taken a different approach that uses the RL techniques to encourage healthy habits instead of looking for direct treatment. In [56], the authors focus on sepsis and RL, but a different approach is taken that uses the RL techniques to control glycemic. In [57], the authors have focused on counterfactual inference and domain adversarial Neural Networks. It is a complicated problem to solve the problem of decision-making under uncertainty. Health care practitioners are facing problems under challenging constraints, with limited tools to make data driven decisions. In [58], the authors have solved the problem of finding an optimal malaria policy as a stochastic multiarmed bandit problem and have developed three agent-based strategies to explore the space of policies. A Gaussian Process regression is applied to the finding of each agent, for compression and for stochastic results from simulating the spread of malaria in a fixed population. The policy generated by the simulation is compared with human experts in the field for direct reference. In [59], the authors have exposed subtleties associated with evaluating RL algorithms in health care. The focus is on the observational setting where RL algorithms have proposed a treatment policy and been evaluated based on historical data. A survey in [60] discusses the different applications of reinforcement learning in health care. The paper provides a systematic understanding of theoretical foundations, methods and techniques, challenges, and new insights into emerging directions. A context aware hierarchical RL scheme [61] has been shown to significantly improve the accuracy of symptom checking over traditional systems while reducing the number of inquiries. Another study that introduces basic concepts of RL and how RL could be effectively used in health care is given in [62].

Policy for malaria control using the reinforcement learning algorithm is explained in [63, 64]. The authors have applied the Genetic Algorithms [65], Bayesian Optimization [66], and Q-Learning with sequence breaking to search for optimal policy for a few years. Their experiments demonstrated the best performance by Q-Learning algorithm. A systematic review of agent-based models for malaria transmission is given in [67]. The paper covers an extensive array of topics covering the spectrum of transmission and intervention of malaria. Machine learning algorithms for the prediction of different diseases are studied in [68]. The authors have used Decision Tree and MapReduce algorithms and have claimed to achieve 94.8% accuracy. Machine learning algorithms have been used to automatically

diagnose malaria in [69]. Deep Convolutional Neural Networks have been used for classification. The authors in [70] have discussed safety applications related to AI in those domains where deep reinforcement learning is applied to the control of automatic mobile robots. An investigation of the risk associated with malaria infection to identify those bottlenecks in different malaria elimination techniques is discussed in [71]. Other relevant studies can be found in [72–74].

### 3. Methodology

Reinforcement learning (RL) [75] is an example of machine learning methods falling between supervised and unsupervised learning, where an agent learns by interacting with the environment. The agent performs certain actions and receives feedback from the environment. This feedback is in the form of negative or positive reward and determines the sequence of good or bad actions to be adapted within a particular situation. As a result, the agent can perform its operation efficiently without any intervention from a human. In other words, RL is a learning method where an agent learns a sequence of actions to eventually increase the reward function. The agent decides which action is the most appropriate and yields a maximum reward. It is possible that an action may not give a positive immediate reward but the long-term reward is also considered. In RL, we have two components, that is, agent and environment as shown in Figure 1. The agent represents the type of RL algorithm, and the environment represents what action returns which reward. The environment is established by sending a state at time  $t$  as  $S_t \in S$ , where  $S$  is the representation of the set of possible states to the agent. The action taken by the agent at time  $t$  is represented by  $A_t \in A(S_t)$ , where  $A(S_t)$  is the representation of the set of actions possible to be taken at state  $S_t$ . The reward to be received by performing that action is represented as  $R_{t+1} \in R$ , where  $R$  is the set of rewards. After one time-step, the next state  $S_{t+1}$  will be sent to the agent by the environment along with reward  $R_{t+1}$ . This reward will eventually help the agent increase its knowledge to be used in evaluating its last action. This process of sending state and receiving reward as an outcome by the agent continues until the environment sends the last or terminal state to the agent.

In addition to the agent and environment, there are four components in a RL environment: (i) policy, (ii) reward, (iii) value function, and (iv) model of the environment.

- (1) *Policy*. A policy defines the behaviour/reaction of an agent at a particular instance of time. Sometimes, a policy can be described as a simple function or as a lookup table, where a policy may involve a lot of computation, for example, the searching process. The policy is considered as a central part of the RL agent because it alone can describe the reaction of the agent. The policy may be stochastic, to determine possibilities for every action. The policy is represented by  $\pi_t$ , where  $\pi_t(a|s)$  demonstrates the probability of  $A_t = a$  if  $S_t = s$

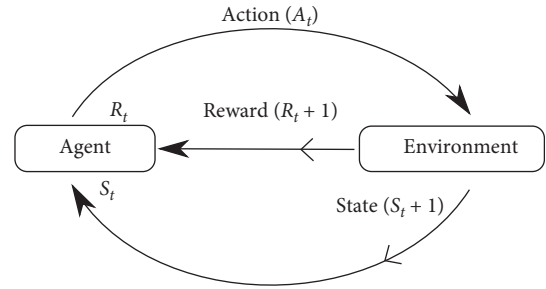


FIGURE 1: A typical reinforcement learning paradigm.

- (2) *Reward*. A reward signal indicates the target of an RL problem. As a result of an action taken by the agent, the environment returns a number, called a reward, at every time step. The objective of the agent is to get most of the total reward over time. Thus, the reward signal identifies that an action is good or bad. The rewards signal determines the action to be taken. If an action returns a low reward, then the policy will be changed to select another action in a similar situation. So generally, a reward signal is the stochastic function of the state and action.
- (3) *Value Function*. A reward signal identifies what is good at the current time, while a value function describes what is good in the long run. In almost all RL algorithms, the most important component to be considered is the method to efficiently estimate the values. More precisely, the current value of the earlier state is adjusted to be closer to the value of the later state. This can be done by moving the earlier state's value a fraction toward the value of the later state. Let  $s$  denote the state before the move, and  $s'$  is the state after the Agent Environment moves; then, the update to the estimated value of  $s$ , denoted as  $V(s)$ , can be written as shown in equation (1), where  $\alpha'$  is a small positive fraction called the step-size parameter, which influences the rate of learning.  $r + \gamma V(s')$  is called Temporal Difference target and is an unbiased estimate for  $V(s')$ . In equation (1),  $r$  represents reward and  $\gamma$  represents the discounting factor. This update rule is an example of a Temporal Difference learning method, called so because its changes are based on a difference,  $V(s') - V(s)$ , that is, difference between estimates at two different times:

$$V(s) \leftarrow V(s) + \alpha[r + \gamma V(s') - V(s)]. \quad (1)$$

- (4) *Model*. A model allows inferences of the actions in an environment. Suppose a state and action are given; then, it is possible that the model determines the resultant next state and reward. The methods that use the models and planning to solve RL Problems are known as model-based methods. Those techniques which are explicitly trail-and-error learner are called model-free methods.



Let us assume that there are finite states and rewards. Let us consider an environment that may respond at time  $t + 1$  to the action taken at time  $t$ . This response actually depends on everything that happened earlier. The complete probability distribution of the dynamics of the system can be defined in equation (2), for all  $r, S$ , and all possible values of the actions in the past represented in the form of action, states, and rewards, that is,  $S_b, A_b$ , and  $R_t$ . However, due to the Markovian property, we can represent the response of the environment at  $t + 1$  that depends only on the state and action at time  $t$ . The dynamics of the environment can be defined as given in equation (3), for all  $r, s', S_b$ , and  $A_t$ . It means that a state or an environment has a Markovian property if and only if equations (2) and (3) are equal. The Markovian property is very important in RL, as decisions and values are a function of the current state. These decisions and values can be effective and carry more information when the state representation carries enough information:

$$\Pr\{R_{t+1} = r, S_{t+1} = S' | S_0, A_0, R_1, \dots, S_{t-1}, A_{t-1}, R_t, S_t, A_t\}, \quad (2)$$

$$p(s', r | s, a) = \Pr\{R_{t+1} = r, S_{t+1} = s' | S_t, A_t\}. \quad (3)$$

The task of RL that satisfied the Markovian property is known by the name Markov Decision Process (MDP). Given a state  $s$  and action  $a$ , the computation of probability of next state  $s'$  along with reward  $r$  is denoted as given in equation (4). The expected value of rewards for the state-action pairs can be computed given in equation (5). The expected rewards for state-action-next-state is given in equation (6):

$$p(s', r | s, a) = \Pr\{S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a\}, \quad (4)$$

$$\begin{aligned} r(s, a) &= E[R_{t+1} | S_t = s, A_t = a] \\ &= \sum_{r \in R} r \sum_{s' \in S} p(s', r | s, a), \end{aligned} \quad (5)$$

$$\begin{aligned} r(s, a, s') &= E[R_{t+1} | S_t = s, A_t = a, S_{t+1} = s'] \\ &= \frac{\sum_{r \in R} r \cdot p(s', r | s, a)}{p(s' | s, a)}. \end{aligned} \quad (6)$$

Value functions, which is a function of states or state-action pairs, are used to estimate the performance of an agent in a given state. This performance is computed in terms of future rewards to be collected. The state value is denoted by  $V_\pi(s)$  given a policy  $\pi$  and state  $s$  and is computed as shown in equation (7), where  $E_\pi[\cdot]$  represents the expectation of variable when an agent follows a policy  $\pi$  at time step  $t$ . Similarly, the action value of a state  $s$  following a policy  $\pi$  represented by  $q_\pi(s, a)$  is given in equation (8), where  $q_\pi$  is the function of action-value when  $\pi$  policy is used:

$$\begin{aligned} V_\pi(s) &= E_\pi[G_t | S_t = s] \\ &= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\right], \end{aligned} \quad (7)$$

$$\begin{aligned} q_\pi(s, a) &= E_\pi[G_t | S_t = s, A_t = a] \\ &= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a\right]. \end{aligned} \quad (8)$$

RL problem is solved by searching for a policy that helps the agent to collect maximum possible rewards over the execution of the simulation. A given policy  $\pi$  is treated as a better policy or equal to another policy  $\pi'$ , if the expectation of the  $\pi$  is greater or equal to the expectation of  $\pi'$  for all states. In other words,  $\pi \geq \pi'$  if and only if  $V_\pi(s) \geq V_{\pi'}(s) \forall s \in S$ . An optimal policy is the policy that is considered good or equal to all possible policies. Optimal policies are represented by  $\pi^*$ . The same state-value function is shared by optimal policies as  $V^*$  and defined as  $V^*(s) = \max V_\pi(s) \forall s \in S$ . They also share same optimal action-value function, represented by  $q^*$  defined as  $q^*(s, a) = \max q_\pi(s, a) \forall s \in S$  and  $a \in A(s)$ .

The model-based RL means the simulation of the dynamics of a given environment. The model learns the probability of moving from the current state  $s_0$ , taking action  $a$  and ending in next state  $s_1$ . Given the learning of transition probability, the agent can determine the probability to enter a state given the current state and action. However, model-based algorithms are not practical because the state space and action space grow. On the other side, the model-free algorithms depend on trial-and-error to update its knowledge. Therefore, space is not required to store all combination of states and actions. In this paper, we are using model-free algorithms. Classification of RL algorithms are made based on on-policy and off-policy. When the value is based on the current action  $a$  and derived from the current policy, it is known as on-policy. When an action  $a^*$  is obtained from a different policy, then it is known as off-policy.

**3.1. Q-Learning.** A well-known algorithm in RL is Q-Learning developed by Watkins [76]. Its proof of convergence is given by Jaakkola [77]. Q-Learning is a simple technique, and it can compute optimal action value without the involvement of intermediary evaluation of cost and the usage of a model [78]. This algorithm is model-free and is considered as off-policy algorithm, which is derived from Bellman Equation as shown in equation (9), where expectation is given by  $E$  and discounting factor is represented by  $\lambda$ . This update equation is shown in Algorithm 1 on line 10. Learning rate is represented by  $\alpha$ . The next state's  $Q$  value determine the next action  $a$  instead of using the current

```

Input:
States:  $S = 1, \dots, n$ 
Actions:  $A = 1, \dots, n$ 
Rewards:  $R: S \times A \rightarrow R$  Transitions:  $T: S \times A \rightarrow S$ 
 $\alpha \in [0, 1]$  and  $\gamma \in [0, 1]$ 
Randomly Initialize  $Q(s, a) \forall s \in S, a \in A(s)$ 
while For every episode do
  Initialize  $S \in S$ 
  Select  $a$  from  $s$  on the basis of exploration strategy (e.g.  $\epsilon$ -greedy)
  while For every step in the episode do
    //Repeat until  $s$  is terminal
    Compute  $\pi$  on the basis of  $Q$  and strategy of exploration (e.g.  $\pi(s) = \operatorname{argmax}_a Q(s, a)$ )
     $a \leftarrow \pi(s)$ 
     $r \leftarrow R(s, a)$ 
     $s \leftarrow T(s, a)$ 
     $Q(s', a) \leftarrow (1 - \alpha) \cdot Q(s, a) + \alpha [r + \max_{a'} Q(s', a')]$ 
     $s \leftarrow s$ 

```

ALGORITHM 1: Q-Learning.

policy. The overall objective of the algorithm is to maximize the Q-value:

$$Q^\pi(s, a) = E_{s'} [r + \lambda Q^\pi(s', a') | s, a]. \quad (9)$$

**3.2. SARSA.** A similar algorithm to Q-Learning is SARSA [79, 80]. In case of Q-Learning, greedy policy is followed, but in case of SARSA on-policy is followed. SARSA learns Q-value by performing actions using the current policy. Algorithm 2 shows the algorithm of SARSA. Current policy is used to carry out selection of actions.

**3.3. Deep Deterministic Policy Gradient.** An actor-critic architecture is called Deep Deterministic Policy Gradient (DDPG) [81, 82]. The parameter  $x$  is tuned for policy by actor as given in equation (10). Using Temporal Difference error, the policy computed by the action is evaluated by critic as demonstrated in equation (11). The policy decided by the actor is shown by  $v$ . The idea of experience replay and separate target network as utilized by Deep Q Network (DQN) [83] is used by DDPG. Algorithm 3 shows the algorithm of DDPG.

$$\pi_\theta(s, a) = P[a | s, \theta], \quad (10)$$

$$r_{t+1} + \gamma V^v(S_{t+1}) - V^v(S_t). \quad (11)$$

$$A(s) = [a_{\text{ITN}}, a_{\text{IRS}}], \quad \text{where } a_{\text{ITN}} \in [0, 1] \text{ and } a_{\text{IRS}} \in [0, 1]. \quad (12)$$

## 4. Simulation and Discussion

In this section, we present the results of algorithms explained in Section 3 obtained in a simulated human population and see which algorithm performs better to prevent humans from diseases. For the evaluation, we need an environment where we have different states, actions, and agents

(representative of human population) looking for the best policy to avoid diseases such as malaria, flu, and HIV. In this section, results are shown for malaria avoidance only, but similar environment with sufficient information can be used for the avoidance of other types of diseases such as flu, HIV, and dengue. An environment where a human, mosquito, and other factors that can influence the transmission of malaria virus to spread to human is shown in Figure 2. The box on the left contains factors relevant to human and the box on the right contains factors pertaining to mosquitoes. Different factors that can influence the disease are shown inside the arrows linking the boxes for humans and mosquitoes. Environment factors and interventions are shown on the top and bottom of the boxes for human and mosquitoes.

The IBM Africa research team has taken steps to control malaria by developing a world-class environment to distribute bed nets and repellents. Their goal is to develop a custom agent that will help identify the best policies for rewards based on the simulation environment. Our work leverages the environment developed by IBM Africa research for reinforcement learning competition on hexagon-ml ([https://compete.hexagon-ml.com/practice/rl\\_competition/38/](https://compete.hexagon-ml.com/practice/rl_competition/38/)) where an agent learns the best policy for the control of diseases, that is, malaria. The environment provides stochastic transmission models for malaria and different researchers can evaluate the impact of different malaria control interventions. In the environment, an agent may explore optimal policies to control the spread of the malaria virus. A diagram representing the environment developed by Hexagon-ML for finding the best policy for avoiding malaria is given in Figure 3. The environment contains five years. Every year is a state. At every state, we take different actions in the form of ITN and IRS.

States are represented as  $S \in \{1, 2, 3, 4, 5\}$ , where each number shows the number of the year. We are trying to solve the problem of making one-shot policy recommendations for the simulation intervention period of 5 years. The main control methods used in different regions are mass-distribution of long-lasting ITNs, IRS with pyrethroids, and the prompt and

**Input:**States:  $S = 1, \dots, n$ Actions:  $A = 1, \dots, n$ Rewards:  $R: S \times A \rightarrow R$ Transitions:  $T: S \times A \rightarrow S$  $\alpha \in [0, 1]$  and  $\gamma \in [0, 1]$  $\lambda \in [0, 1]$  this shows the trade-off between Temporal Difference and Monte Carlo methods.Randomly Initialize  $Q(s, a) \forall s \in S, a \in A(s)$ **while** For every episode do    Randomly initialize  $s \in S$     Initialize  $e$  with 0    Randomly select  $(s, a) \in S \times A$     **while** For every step in the episode do        //Repeat until  $s$  is terminal         $r \leftarrow R(s, a)$          $s' \leftarrow T(s, a)$         Compute  $\pi$  based on  $Q$  using exploration strategy (e.g.  $\epsilon$ -greedy)         $a' \leftarrow \pi(s')$          $e(s, a) \leftarrow e(s, a) + 1$          $\delta \leftarrow r + \gamma \cdot Q(s', a') - Q(s, a)$         **for**  $(s', a') \in S \times A$  do             $Q(s', a') \leftarrow Q(s', a') + \alpha \cdot \delta \cdot e(s', a')$              $e(s', a') \leftarrow \gamma \cdot \lambda \cdot e(s', a')$          $s \leftarrow s'$          $a \leftarrow a'$ 

ALGORITHM 2: SARSA.

- (1) Randomly initialize critic network  $Q(s, a | \theta^Q)$  with weight  $\theta^Q$
- (2) Randomly initialize actor  $\mu(s | \theta^\mu)$  with weight  $\theta^\mu$
- (3) Initialize target network  $Q'$  with weight  $\theta^{Q'} \leftarrow \theta^Q$
- (4) Initialize target network  $\mu'$  with weight  $\theta^{\mu'} \leftarrow \theta^\mu$
- (5) Initialize replay buffer  $R$
- (6) **while** For every episode do
- (7)     Randomly initialize  $N$  for exploration
- (8)     Get initial observation state  $s_1$
- (9)     **while** For every step in the episode do
- (10)         //Repeat until  $s$  is terminal
- (11)         Select action  $a_t = \mu(s_t | \theta^\mu) + N_t$  as per the current policy and exploration strategy
- (12)         Perform action  $a_t$  and monitor rewards  $r_t$  and new states  $s_{t+1}$
- (13)         Store  $(s_t, a_t, r_t, s_{t+1})$  in  $R$
- (14)         Sample a randomly selected minibatch of  $N$  transition  $(s_i, a_i, r_i, s_{i+1})$  from  $R$
- (15)          $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1} | \theta^{\mu'} | \theta^{Q'}))$
- (16)          $L = 1/N \sum_i (y_i - Q(s_i, a_i | \theta^Q))^2$
- (17)         //Update rule for critic to minimize the loss
- (18)          $\Delta_{\theta^\mu} J \approx 1/N \sum_i \Delta_\alpha Q(s, a | \theta^Q)|_{s=s_i, a=\mu(s_i)} \Delta_{\theta^\mu} \mu(s | \theta^\mu)|_{s_i}$
- (19)         //Update rule for actor policy using the sampled policy gradient
- (20)          $\theta^{Q'} \leftarrow \gamma \theta^{Q'} + (1 - \gamma) \theta^Q$
- (21)         //Update rule for target network
- (22)          $\theta^{\mu'} \leftarrow \gamma \theta^{\mu'} + (1 - \gamma) \theta^\mu$

ALGORITHM 3: Deep Deterministic Policy Gradient.

effective treatment of malaria. Actions, represented by  $A(s)$ , are performed in the form of ITN and IRS, where the values of ITN and IRS are infinite real numbers between 0 and 1.

The agent trained on a reinforcement learning algorithm will explore a policy space made up of the first two

components, that is, ITNs and IRS, which are strategies for direct intervention. The prompt and effective treatment is given by the environment parameters and impacts the rewards. The first component. That is, ITN, is the development of nets, defining the population coverage ( $a_{ITN} \in (0, 1)$ ). The second

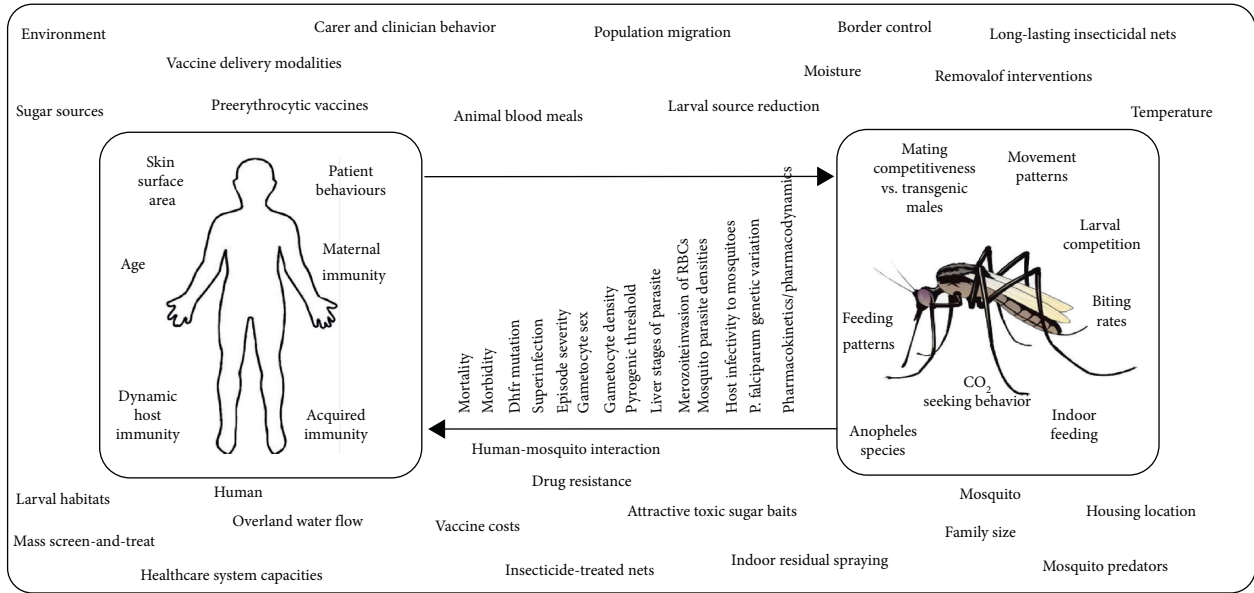


FIGURE 2: Different factors related to humans, mosquitoes, and the environment that influence malaria transmission.

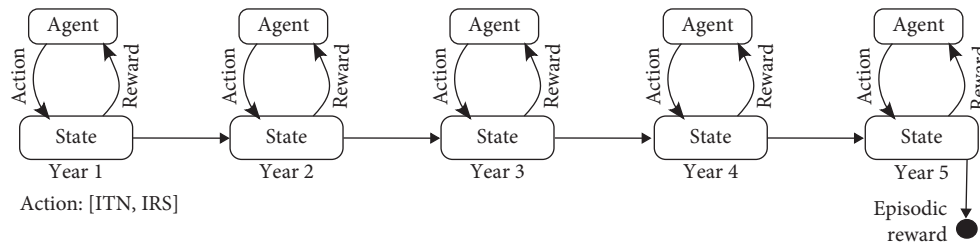


FIGURE 3: A block diagram showing the environment developed by Hexagon-ML for reinforcement learning algorithms to learn malaria intervention.

component is the use of seasonal spraying, and it defines the proportion of population coverage for this intervention ( $a_{IRS} \in (0, 1]$ ). The seasonal spraying is performed through alternating the intervention between April and June every year in different regions. The policy decision is framed in a way of the simulated population to be covered by a particular intervention; the space of policy  $A$  is designed through  $a_i \in A = [a_{ITN}, a_{IRS}]$ .

Health care organizations should be able to explore all possible set of actions for appropriate malaria interventions within the populations. These policies include a mix of actions, like the distribution of ITNs, IRS, larvicide in water, and vaccination for malaria control. The space of possible policies for the control of malaria is not complete and inefficient for health care experts to explore without an adequate decision support system. The environment in simulation handles the distribution of the interventions in the simulated population. The agent is in charge of the complex actions of targeted interventions, which are not reported previously. Although the action space is finite (i.e., finite number of people in the simulation environment) the space size grows exponentially as more interventions are added. The computation time of simulation will also grow

linearly with the number of populations. Therefore, a complex exploration of the entire action space becomes impossible as complexity goes to a real-world equivalent simulation.

The agent learns different rewards during the learning process. The idea of learning is to collect as much reward as possible during the process of execution of the experiment. These rewards are infinite and usually represented by  $R_\pi \in (-\infty, +\infty)$ , where the policy is represented by  $\pi$ . Every policy is associated with a reward represented by  $R_\theta(ai)$  and is a stochastic parameterization of the simulation shown as  $\theta$  which produces random distribution of parameters for the simulated environment.

The environment is executed for 100 episodes, and rewards are collected. An episode consists of five consecutive years. The rewards collected by different algorithms are demonstrated in Figure 4. The random selection algorithm when there is no learning for 100 episodes is given in Figure 4(a). In random policy learning, every time one episode is finished, the environment is initiated with different random states and different policy is tried at random to go from one state to another to collect rewards. In this algorithm, no learning is involved, and this experiment is performed only

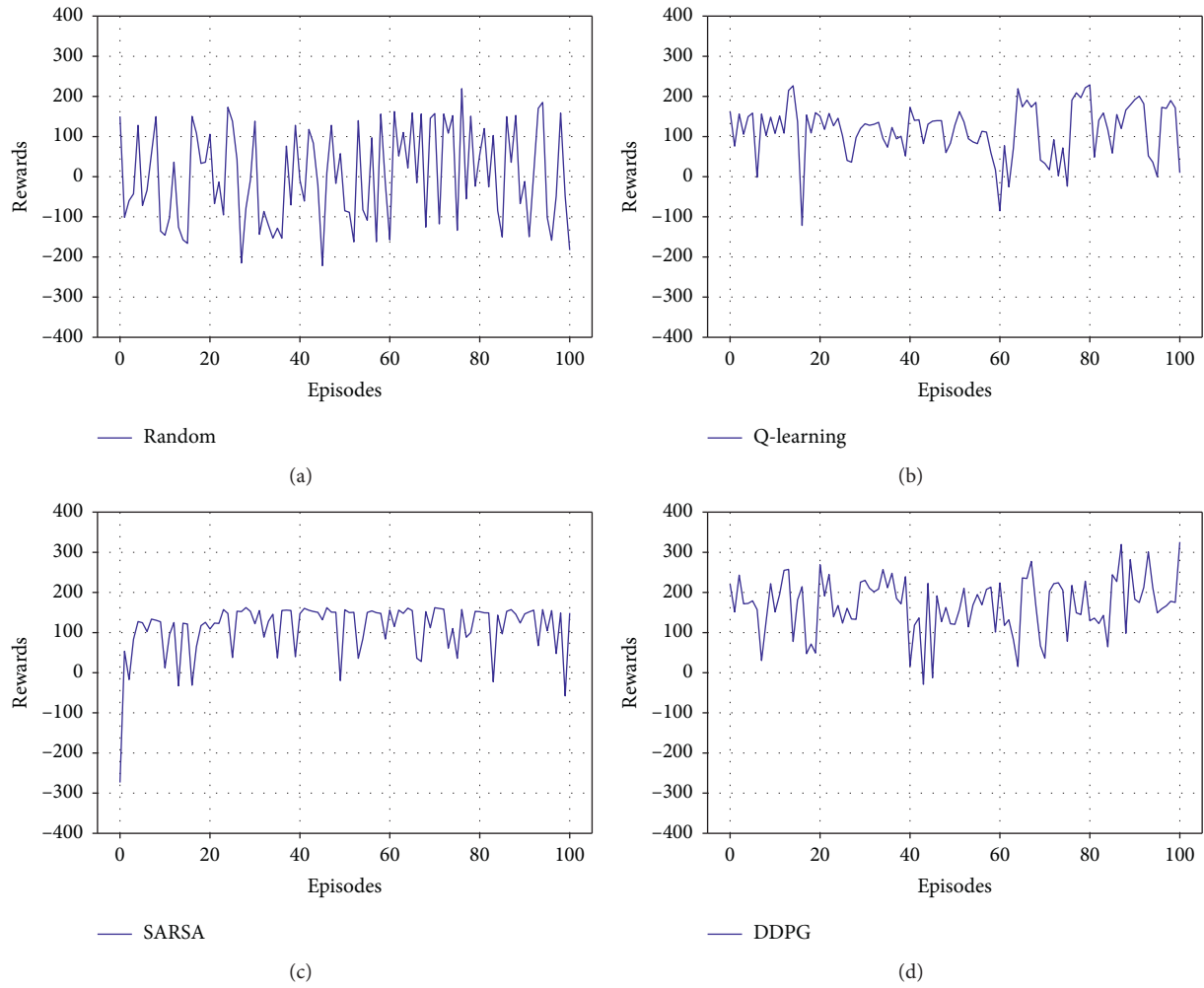


FIGURE 4: Reward collection by agent trained with different reinforcement learning algorithms in 100 episodes. (a) Reward collection when the agent randomly chooses action. (b) Reward collection when the agent is trained with Q-Learning. (c) Reward collection when the agent is trained with SARSA. (d) Reward collection when the agent is trained with DDPG.

to show a base line for comparison with other algorithms. The Q-learning algorithm is shown in Figure 4(b). Compared to random search algorithm, this algorithm has shown improvements as the agent is learning through Q-learning mechanism to collect rewards in the learning process. SARSA algorithm is used, and the result of reward collection is shown in Figure 4(c). The SARSA trained agents are used to look to policy to avoid malaria in a simulated human environment and has shown improvements over simple Q-learning algorithm. An even more sophisticated algorithm known as DDPG is used in the environment to collect rewards, and results are demonstrated in Figure 4(d). This algorithm shows improvements compared to all other three

algorithms and demonstrated that deep learning methods can potentially collect better results in reinforcement learning algorithms.

We have combined the results of the algorithms trained in this paper in Figure 5. In random searching process, there is no learning, and therefore reward is not maximized. But in other algorithms such as Q-learning, SARSA, and DDPG, there is learning involved, and therefore reward is maximized. The overall rewards collected by different algorithms are combined in one figure (Figure 5(b)). The maximum rewards are collected by DDPG because a complex algorithm is used for collection of rewards. This comparison of three algorithms is shown in Table 1. This comparison

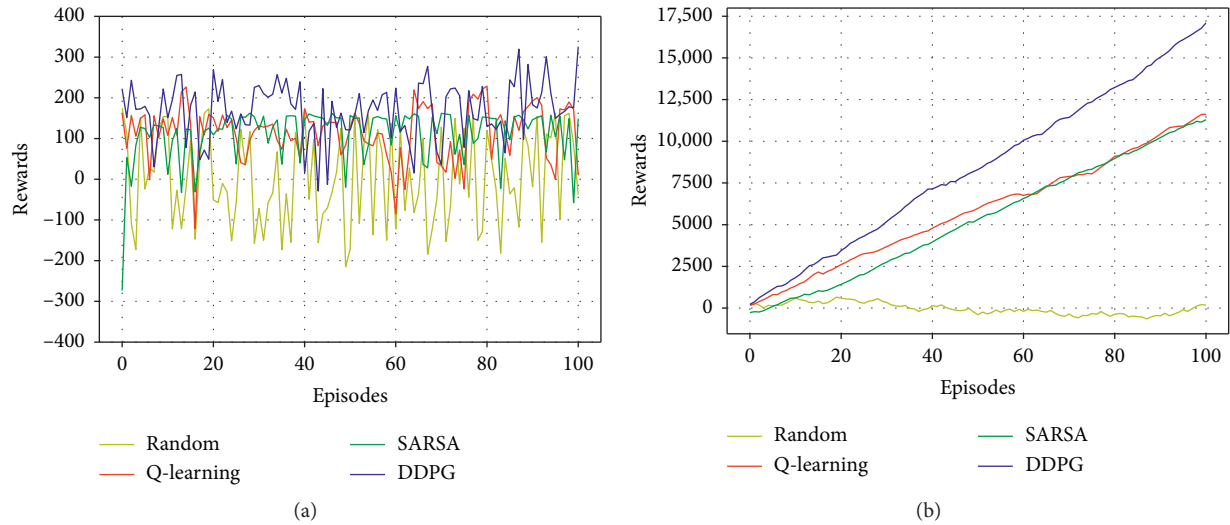


FIGURE 5: Comparison of reward collection by agent trained with different reinforcement learning algorithms, that is, Q-Learning, SARSA, and DDPG in 100 episodes. (a) Reward collection when the agent is trained with different reinforcement learning algorithms, that is, Q-Learning, SARSA, and DDPG. (b) Sum of rewards over time when the agent is trained with different reinforcement learning algorithms Q-Learning, SARSA, and DDPG.

TABLE 1: The comparison of three reinforcement learning algorithms explained in the paper in terms of best rewards and best policy when the agent is executed for 100 episodes.

Algorithm	Best reward	Optimal policy				
		Year 1	Year 2	Year 3	Year 4	Year 5
Random	174.16	[0.2, 0.7]	[0.6, 0.9]	[0.1, 0.8]	[0.4, 0.6]	[0.3, 0.1]
Q-Learning	228.77	[0.3, 0.1]	[0.3, 0.2]	[0.5, 0.2]	[0.9, 0.5]	[0.5, 0.1]
SARSA	161.74	[0.3, 0.1]	[0.3, 0.1]	[0.3, 0.1]	[0.3, 0.1]	[0.3, 0.1]
DDPG	325.55	[1.0, 0.8]	[0.1, 0.0]	[0.1, 0.8]	[0.6, 1.0]	[0.6, 1.0]

demonstrates the best policy obtained by operating in the environment to avoid malaria and the related reward collected by performing the best policy. This table demonstrates that DDPG has outperformed traditional learning algorithms.

## 5. Conclusion

Since the development of human civilizations, humans have always been in the quest to improve the quality of life from different perspectives. We are looking for the most comfortable accommodation, fast and secure transport, clean and healthy food, comfortable clothes, and many other things. But because of the environmental changes and different actions taken by humans, there are possibilities of different viruses entering the body of humans and affecting the quality of life of humans. For instance, malaria, flu, HIV, and dengue are some diseases that not only affect a single individual but also can affect the whole population, as the virus spreads from one person to another person. Humans over time have learned different methods to treat these diseases. There are doctors, who prescribe medicine to treat diseases, and hence diseases are in control. But the problem is that the decision of a doctor requires a huge

amount of knowledge and experience, to effectively cure a disease. We think it is possible that the human effort is minimized, and some AI-based solutions are explored. Different AI-based solutions have also been explored by researchers, in the form of supervised learning such as ANN, KNN, and SVM. However, the problem with these supervised learning is that the model is trained on the existing data to make similar decisions when a similar data is presented as testing. There is a huge gap to further generalize the solution. Therefore, unsupervised learning algorithms and reinforcement learning are becoming popular. In this paper, we have explored reinforcement learning-based algorithms, where an agent interacts with the environment to get feedback and improves its state of knowledge. We have experimented with three different algorithms in reinforcement learning. These algorithms are Q-Learning, SARSA, and DDPG. All these algorithms perform better than random search, as there is learning involved. Q-Learning and SARSA are based on traditional methods of reinforcement learning. However, because of the popularity of deep learning, researchers are interested in introducing deep learning in reinforcement learning. DDPG is a deep learning-based algorithm. Our experiments have demonstrated that deep learning-based

algorithms are the most suitable algorithm for such type of complex environment, where human, their actions, environments, and their feedback play a very important role.

## Data Availability

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

## Conflicts of Interest

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

## Acknowledgments

This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under Grant no. DF-458-156-1441. The authors, therefore, gratefully acknowledge DSR technical and financial support.

## References

- [1] A. Bowling, "The effects of illness on quality of life: findings from a survey of households in great britain," *Journal of Epidemiology and Community Health*, vol. 50, pp. 149–155, 1996.
- [2] C. L. Lam and I. J. Lauder, "The impact of chronic diseases on the health-related quality of life (HRQOL) of Chinese patients in primary care," *Family Practice*, vol. 17, pp. 159–166, 2000.
- [3] R. Somrongthong, D. Hongthong, S. Wongchalee, and N. Wongtongkam, "The influence of chronic illness and lifestyle behaviors on quality of life among older thais," *BioMed Research International*, vol. 2016, pp. 1–7, 2016.
- [4] B. Torto, "Innovative approaches to exploit host plant metabolites in malaria control," *Pest Management Science*, vol. 75, no. 9, pp. 2341–2345, 2019.
- [5] F. Binka and P. Akweongo, "Prevention of malaria using ITNs: potential for achieving the millennium development goals," *Current Molecular Medicine*, vol. 6, pp. 261–267, 2006.
- [6] B. B. Tukei, A. Beke, and H. Lamadrid-Figueroa, "Assessing the effect of indoor residual spraying (IRS) on malaria morbidity in northern Uganda: a before and after study," *Malaria Journal*, vol. 16, no. 1, 2017.
- [7] Y. A. Derua, E. J. Kweka, W. N. Kisinza, A. K. Githeko, and F. W. Mosha, "Bacterial larvicides used for malaria vector control in Sub-Saharan Africa: review of their effectiveness and operational feasibility," *Parasites & Vectors*, vol. 12, no. 1, 2019.
- [8] T. L. I. Diseases, "Malaria vaccination: a major milestone," *The Lancet Infectious Diseases*, vol. 19, p. 559, 2019.
- [9] S. J. Draper, B. K. Sack, C. R. King et al., "Malaria vaccines: Recent advances and new horizons," *Cell Host & Microbe*, vol. 24, pp. 43–56, 2018.
- [10] M. Fatima, A. Baig, and I. Uddin, "Reliable and energy efficient MAC mechanism for patient monitoring in hospitals," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, 2018.
- [11] I. Uddin, A. Baig, and A. Ali, "A controlled environment model for dealing with smart phone addiction," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 9, 2018.
- [12] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *NPJ Computational Materials*, vol. 5, no. 1, 2019.
- [13] K. Y. Ngiam and I. W. Khor, "Big data and machine learning algorithms for health-care delivery," *The Lancet Oncology*, vol. 20, no. 5, p. e262, 2019.
- [14] E. Loh, "Medicine and the rise of the robots: a qualitative review of recent advances of artificial intelligence in health," *BMJ Leader*, vol. 2, no. 2, pp. 59–63, 2018.
- [15] F. Jiang, Y. Jiang, H. Zhi et al., "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.
- [16] O. Frunza, D. Inkpen, and T. Tran, "A machine learning approach for identifying disease-treatment relations in short texts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, pp. 801–814, 2011.
- [17] X. Liu, L. Faes, A. U. Kale et al., "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The Lancet Digital Health*, vol. 1, no. 6, pp. e271–e297, 2019.
- [18] S. Syed, M. Al-Boni, M. N. Khan et al., "Assessment of machine learning detection of environmental enteropathy and celiac disease in children," *JAMA Network Open*, vol. 2, no. 6, Article ID e195822, 2019.
- [19] Y. Liu, B. Logan, N. Liu, Z. Xu, J. Tang, and Y. Wang, "Deep reinforcement learning for dynamic treatment regimes on medical registry data," in *Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI)*, Park City, UT, USA, August 2017.
- [20] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, *An Introduction to Deep Reinforcement Learning*, <http://arxiv.org/abs/1811.12560>, 2018.
- [21] Y. Zheng, *Reinforcement Learning and Video Games*, MSc thesis, University of Sheffield, Sheffield, UK, 2019.
- [22] I. Szita, *Reinforcement Learning in Games*, in M. Wiering, M. van Otterlo (eds) *Reinforcement Learning, Adaptation, Learning, and Optimization*, vol. 12, Springer, Berlin, Germany, 2012, [https://doi.org/10.1007/978-3-642-27645-3\\_17](https://doi.org/10.1007/978-3-642-27645-3_17).
- [23] R. R. Torrado, P. Bontrager, J. Togelius, J. Liu, and D. Perez-Liebana, "Deep reinforcement learning for general video game AI," in *Proceedings of the 14th IEEE Conference on Computational Intelligence and Games, CIG 2018*, Maastricht, Netherlands, August 2018.
- [24] B. A. Kwambana-Adams, E. K. Mulholland, E. K. Mulholland, and C. Satzke, "State-of-the-art in the pneumococcal field: proceedings of the 11th International Symposium on pneumococci and pneumococcal diseases (ISPPD-11)," *Pneumonia*, vol. 12, no. 1, 2020.
- [25] A. Raghu, M. Komorowski, and S. Singh, "Model-based reinforcement learning for sepsis treatment," 2018, <http://arxiv.org/abs/1811.09602>.
- [26] C. P. Janssen and W. D. Gray, "When, what, and how much to reward in reinforcement learning-based models of cognition," *Cognitive Science*, vol. 36, no. 2, pp. 333–358, 2012.
- [27] I. Uddin, "High-level simulation of concurrency operations in microthreaded many-core architectures," *GSTF Journal on Computing*, vol. 4, no. 3, p. 21, 2015.
- [28] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: a brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

- [29] M. Hengge and S. Leonard, *Factor Models for Non-Stationary Series: Estimates of Monthly U.S. GDP*, IHEID Working Papers 13-2017, Economics Section, The Graduate Institute of International Studies, 2017.
- [30] H. Burnett, A. Earley, A. A. Voors et al., “Thirty years of evidence on the efficacy of drug treatments for chronic heart failure with reduced ejection fraction,” *Circulation: Heart Failure*, vol. 10, 2017.
- [31] R. P. Steeds and K. S. Channer, “Drug treatment in heart failure,” *BMJ*, vol. 316, no. 7131, pp. 567-568, Feb. 1998.
- [32] I. Uddin, “One-IPC high-level simulation of microthreaded many-core architectures,” *International Journal of High Performance Computing Applications*, vol. 31, no. 2, pp. 152-162, 2015.
- [33] S. D. W. Frost, B. R. Magalis, and S. L. Kosakovsky Pond, “Neutral theory and rapidly evolving viral pathogens,” *Molecular Biology and Evolution*, vol. 35, no. 6, pp. 1348-1354, 2018.
- [34] R. G. Webster and E. A. Govorkova, “Continuing challenges in influenza,” *Annals of the New York Academy of Sciences*, vol. 1323, no. 1, pp. 115-139, 2014.
- [35] S. Duffy, “Why are RNA virus mutation rates so damn high?” *PLoS Biology*, vol. 16, no. 8, Article ID e3000003, 2018.
- [36] D. J. Hockstra and S. D. Miller, “Sequential games and medical diagnosis,” *Computers and Biomedical Research*, vol. 9, no. 3, pp. 205-215, 1976.
- [37] D. Hausmann, C. Zulian, E. Battagay, and L. Zimmerli, “Tracing the decision-making process of physicians with a decision process matrix,” *BMC Medical Informatics and Decision Making*, vol. 16, no. 1, 2016.
- [38] M. Uddin, Y. Wang, and M. Woodbury-Smith, “Artificial intelligence for precision medicine in neurodevelopmental disorders,” *NPJ Digital Medicine*, vol. 2, no. 1, 2019.
- [39] A. S. Ahuja, “The impact of artificial intelligence in medicine on the future role of the physician,” *PeerJ*, vol. 7, Article ID e7702, 2019.
- [40] D. Zois, “Sequential decision-making in healthcare IOT: real-time health monitoring, treatments and interventions,” in *Proceedings of the 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, pp. 24-29, Reston, VA, USA, December 2016.
- [41] O. Alagoz, H. Hsu, A. Schaefer, and M. Roberts, “Markov decision processes: a tool for sequential decision making under uncertainty,” *Medical decision making*, *An International Journal of the Society for Medical Decision Making*, vol. 30, pp. 474-483, 2010.
- [42] C. C. Bennett and K. Hauser, “Artificial intelligence framework for simulating clinical decision-making: a markov decision process approach,” *Artificial Intelligence in Medicine*, vol. 57, no. 1, pp. 9-19, 2013.
- [43] S. A. A. Shah, I. Uddin, F. Aziz, S. Ahmad, M. A. Al-Khasawneh, and M. Sharaf, “An enhanced deep neural network for predicting workplace absenteeism,” *Complexity*, vol. 2020, Article ID 5843932, 12 pages, 2020.
- [44] M. I. Uddin, N. Zada, F. Aziz et al., “Prediction of future terrorist activities using deep neural networks,” *Complexity*, vol. 2020, Article ID 1373087, 16 pages, 2020.
- [45] S. Parisi, D. Tateo, M. Hensel, C. D’Eramo, J. Peters, and J. Pajarinen, “Long-term visitation value for deep exploration in sparse reward reinforcement learning,” 2020, <http://arxiv.org/abs/2001.00119>.
- [46] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities,” *IEEE Access*, vol. 5, pp. 8869-8879, 2017.
- [47] 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.
- [48] Y. Dong, Z. Jiang, H. Shen, and W. D. Pan, “Classification accuracies of malaria infected cells using deep convolutional neural networks based on decompressed images,” in *Proceedings of the 2017 Southeastern Conference*, pp. 1-6, Charlotte, NC, USA, 2017.
- [49] S. Rajaraman, S. K. Antani, M. Poostchi et al., “Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images,” *PeerJ*, vol. 6, Article ID e4568, 2018.
- [50] J. Pineau, A. Guez, R. Vincent, G. Panuccio, and M. Avoli, “Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach,” *International Journal of Neural Systems*, vol. 19, no. 4, pp. 227-240, 2009.
- [51] Y. Zhao, D. Zeng, M. A. Socinski, and M. R. Kosorok, “Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer,” *Biometrics*, vol. 67, pp. 1422-1433, 2011.
- [52] E. Yom-Tov, G. Feraru, M. Kozdoba, S. Mannor, M. Tennenholtz, and I. Hochberg, “Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system,” *Journal of Medical Internet Research*, vol. 19, no. 10, p. e338, 2017.
- [53] L. Wang, W. Zhang, X. He, and H. Zha, “Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD’18*, pp. 2447-2456, New York, NY, USA, 2018.
- [54] G. Yauney and P. Shah, “Classification accuracies of malaria infected cells using deep convolutional neural networks based on decompressed images reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection,” in *Proceedings of the 3rd Machine Learning for Healthcare Conference*, pp. 161-226, Palo Alto, CA, USA, August 2018.
- [55] I. Hochberg, G. Feraru, M. Kozdoba, S. Mannor, M. Tennenholtz, and E. Yom-Tov, “A reinforcement learning system to encourage physical activity in diabetes patients,” 2016, <http://arxiv.org/abs/1605.04070>.
- [56] W.-H. Weng, M. Gao, Z. He, S. Yan, and P. Szolovits, *Representation and Reinforcement Learning for Personalized Glycemic Control in Septic Patients*, 2017.
- [57] O. Atan, W. R. Zame, and M. van der Schaar, “Learning optimal policies from observational data,” 2018, <http://arxiv.org/abs/1802.08679>.
- [58] O. Bent, S. Remy, S. Roberts, and A. Walcott-Bryant, *Novel Exploration Techniques (Nets) for Malaria Policy Interventions*, 2017, <https://arxiv.org/abs/1712.00428>.
- [59] O. Gottesman, F. Johansson, J. Meier et al., *Evaluating Reinforcement Learning Algorithms in Observational Health Settings*, 2018, <https://arxiv.org/abs/1805.12298>.
- [60] C. Yu, J. Liu, and S. Nemati, *Reinforcement Learning in Healthcare: A Survey*, 2019, <https://arxiv.org/abs/1908.08796>.
- [61] H.-C. Kao, K.-F. Tang, and E. Y. Chang, “Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LO, USA, February 2018.
- [62] A. Jonsson, “Deep reinforcement learning in medicine,” *Kidney Diseases*, vol. 5, pp. 18-22, 2018.



- [63] V. B. Nguyen, B. M. Karim, B. L. Vu, J. Schlötterer, and M. Granitzer, *Policy Learning for Malaria Control*, 2019, <https://arxiv.org/abs/1910.08926>.
- [64] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.
- [65] K. F. Man, K. S. Tang, and S. Kwong, "Genetic algorithms: concepts and applications in engineering design," *IEEE Transactions on Industrial Electronics*, vol. 43, no. 5, pp. 519–534, Oct 1996.
- [66] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 2951–2959, New York, NY, USA, 2012.
- [67] N. R. Smith, J. M. Trauer, M. Gambhir et al., "Agent-based models of malaria transmission: a systematic review," *Malaria Journal*, vol. 17, 2018.
- [68] S. Vinitha, S. Sweetlin, H. M. Vinusha, and S. Sajini, "Disease prediction using machine learning over big data," *SSRN Electronic Journal*, 2018.
- [69] Y. Dong, Z. Jiang, H. Shen, and W. D. Pan, "Classification accuracies of malaria infected cells using deep convolutional neural networks based on decompressed images," in *Proceedings of the SoutheastCon 2017*, pp. 1–6, Charlotte, NC, USA, 2017.
- [70] T. Namba and Y. Yamada, "Risks of deep reinforcement learning applied to fall prevention assist by autonomous mobile robots in the hospital," *Big Data and Cognitive Computing*, vol. 2, no. 2, p. 13, June 2018.
- [71] G. Tiburce, S. Laurentine, H. N. Ngum, I. C. Etso, and C. N.-D. Hugues, "Investigating risk factors associated with the persistence of malaria in the obang valley, north west region, Cameroon," *Journal of Public Health and Epidemiology*, vol. 10, no. 10, pp. 380–386, 2018.
- [72] J.-e. Liu and F.-P. An, "Image classification algorithm based on deep learning-kernel function," *Scientific Programming*, vol. 2020, pp. 1–14, Article ID 7607612, 2020.
- [73] E. Torti, M. Musci, F. Guareschi, F. Leporati, and M. Piastra, "Deep recurrent neural networks for edge monitoring of personal risk and warning situations," *Scientific Programming*, vol. 2019, pp. 1–10, Article ID 9135196, 2019.
- [74] B. Ramzan, I. S. Bajwa, N. Jamil et al., "An intelligent data analysis for recommendation systems using machine learning," *Scientific Programming*, vol. 2019, pp. 1–20, Article ID 5941096, 2019.
- [75] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, A Bradford Book, Cambridge, MA, USA, 2018.
- [76] C. Watkins, *Learning from Delayed Rewards*, Ph.D. thesis, University of Cambridge, Cambridge, UK, 1989.
- [77] T. Jaakkola, M. I. Jordan, and S. P. Singh, "Convergence of stochastic iterative dynamic programming algorithms," in *Proceedings of the 6th International Conference on Neural Information Processing Systems (NIPS'93)*, pp. 703–710, San Francisco, CA, USA, 1993.
- [78] C. H. C. Ribeiro, *A Tutorial on Reinforcement Learning Techniques*, University of Michigan, Ann Arbor, MI, USA, 1999.
- [79] D. Zhao, H. Wang, K. Shao, and Y. Zhu, "Deep reinforcement learning with experience replay based on SARSA," in *Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–6, Athens, Greece, 2016.
- [80] Z.-x. Xu, L. Cao, C. Xiliang, C.-x. Li, Y.-l. Zhang, and J. Lai, "Deep reinforcement learning with sarsa and Q-learning: a hybrid approach," *IEICE Transactions on Information and Systems*, vol. E101, pp. 2315–2322, 2018.
- [81] G. Yang, F. Zhang, C. Gong, and S. Zhang, "Application of a deep deterministic policy gradient algorithm for energy-aimed timetable rescheduling problem," *Energies*, vol. 12, no. 18, p. 3461, 2019.
- [82] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8577–8588, 2019.
- [83] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," 2020, <http://arxiv.org/abs/1901.00137>.

## Research Article

# A New Approach for Enhancing the Services of the 5G Mobile Network and IOT-Related Communication Devices Using Wavelet-OFDM and Its Applications in Healthcare

Mordecai F. Raji <sup>1</sup>, JianPing Li <sup>1</sup>, Amin Ul Haq <sup>1</sup>, Victor Ejianya <sup>2</sup>,  
Jalaluddin Khan <sup>1</sup>, Asif Khan <sup>1</sup>, Mudassir Khalil <sup>1</sup>, Amjad Ali <sup>3</sup>, Ghufraan A. Khan <sup>4</sup>,  
Mohammad Shahid <sup>5</sup>, Bilal Ahamad <sup>6</sup>, Amit Yadav <sup>7</sup>, and Imran Memon <sup>8</sup>

<sup>1</sup>School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

<sup>2</sup>School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, China

<sup>3</sup>Department of Computer Science and Software Technology, University of Swat, Mingora, Pakistan

<sup>4</sup>School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China

<sup>5</sup>Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

<sup>6</sup>Department of Computer Science, Shaqra University, Shaqra, Saudi Arabia

<sup>7</sup>Department of Information Management, Chengdu Neusoft University, Chengdu, Sichuan, China

<sup>8</sup>Department of Computer Science, Bahria University, Islamabad, Sindh, Pakistan

Correspondence should be addressed to Mordecai F. Raji; [mraji@qq.com](mailto:mraji@qq.com), JianPing Li; [jpli2222@uestc.edu.cn](mailto:jpli2222@uestc.edu.cn), Amin Ul Haq; [khan.amin50@yahoo.com](mailto:khan.amin50@yahoo.com), and Asif Khan; [asifkhan@uestc.edu.cn](mailto:asifkhan@uestc.edu.cn)

Received 30 November 2019; Revised 14 January 2020; Accepted 6 February 2020; Published 10 October 2020

Academic Editor: Shaukat Ali

Copyright © 2020 Mordecai F. Raji 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 heart of the current wireless communication systems (including 5G) is the Fourier transform-based orthogonal frequency division multiplex (OFDM). Over time, a lot of research has proposed the wavelet transform-based OFDM as a better replacement of Fourier in the physical layer solutions because of its performance and ability to support network-intensive applications such as the Internet of Things (IoT). In this paper, we weigh the wavelet transform performances against the future wireless application system requirements and propose guidelines and approaches for wavelet applications in 5G waveform design. This is followed by a detailed impact on healthcare. Using an image as the test data, a comprehensive performance comparison between Fourier transform and various wavelet transforms has been done considering the following 5G key performance indicators (KPIs): energy efficiency, modulation and demodulation complexity, reliability, latency, spectral efficiency, effect of transmission/reception under asynchronous transmission, and robustness to time-/frequency-selective channels. Finally, the guidelines for wavelet transform use are presented. The guidelines are sufficient to serve as approaches for tradeoffs and also as the guide for further developments.

## 1. Introduction

The number of devices connected to the internet is on the increase. The advent of the IoT will make it an explosive one [1, 2]. Almost every 10 years, a new generation of wireless communication standard is developed to meet the exponentially growing demand for fast and reliable connections.

Wireless service consumers have experienced the growth and maturity of 2G, 3G, and 4G mobile network systems. Now, the International Telecommunication Union (ITU) has defined the expectations for 5G New Radio (NR). The 5G requirements are classified into enhanced mobile broadband (eMBB), ultrareliable low-latency communication (URLLC), and massive machine-type communication (mMTC). All

these classifications are subsets of IOT's requirements. Their requirements are not limited to multi-gigabit-per-second (Gbps) data rates, low latency, high spectral efficiency, high mobility, and high connection density [3]. We can, therefore, conclude that 5G will have to cope with a high degree of heterogeneity in terms of services and requirements. Surprisingly, the increasing number of certain new applications as well as newer, yet to be conceived applications, will require even greater data rates among other requirements than what 5G NR can offer [4], and some of those applications might not even fall completely within a single defined use-case [5]. A few of them are virtual/augmented reality (VR/AR), wireless cognition, and wireless backhaul. Furthermore, because of 5G's diversity, there possibly exist other undiscovered areas of applications such as in healthcare, of which IoT is obviously the enabler. Advances in a research area will impact the other; it is like a chain reaction; for example, 5G (as the underlying technology) will unlock IoT capabilities that were previously unattainable on 4G networks. It will introduce major innovations, such as higher connection speed, greater capacity, and lower latency. Over time, these benefits will lead to technological milestones that will have a significant impact on healthcare and other areas. Therefore, for an efficient IoT system, there is a need for a diverse network such as 5G. The advent of 5G has led to further exploration of new methodology, hardware, waveform design, underutilized millimeter-wave (mm-wave) frequency, etc. The most prominent of these explorations but most challenging would be to design waveform for efficient signaling in 5G. The summary of the proposed waveforms is given in [6].

OFDM remains the key ingredient in waveform design for many multicarrier wireless communication schemes [7]. Proposed waveforms for 5G are mostly OFDM-inspired. This reason is partly due to OFDM's use in 4G LTE and other standards, familiarity among the wireless society, and its maturity as a technology [8, 9]. However, the gradual migration from cell-centric to user-centric processing is rendering OFDM unfeasible due to its intrinsic drawbacks. Therefore, OFDM limitations such as the reduction in transmission throughput due to the use of cyclic prefix (CP), high peak-to-average power ratio (PAPR), sensitivity to carrier offset, and out-of-band emissions (OOB) are some of the problems solved in the proposed waveforms.

Attempts at combating some of the OFDM's limitations require an understanding of the root causes. For example, OOB emission (or spectral leakage) is introduced by band egress noise to neighboring bands and ingress noise from neighboring bands. This is because the spectral localization of the subcarriers is weak, resulting in spectral leakage. Filter bank multicarrier (FBMC) attempts to solve the aforementioned problem by direct suppression of the sidelobes using special filters. FBMC is spectrum-efficient, does not require redundant CP, and is robust to narrow-band jammers. However, practical applications indicate that FBMC is vulnerable to multipath distortion due to a lack of CP. As a result, in practical cases whereby channel

state information is not perfect, OFDM will perform better. Attempt to improve FBMC and adopt it in MIMO channels has been reported in [10] with a focus on channel uncertainty.

There are other evident disadvantages of FBMC such as the introduction of overhead in overlapping symbols in the filter bank in the time domain and loss in bandwidth efficiency when transmitting short data packets. As a result, FBMC was extended to the generalized frequency division multiplex (GFDM) in [11]. GFDM is based on controlling the OOB of the transmitted signal by an adjustable pulse shaping filter applied to the individual carriers [12]. GFDM can be applied successfully in non-accurate synchronization of users without problems. However, the GFDM scheme is not perfect; adjacent synchronization is affected by some level of interference. Attempts at solving this problem have been exploited in [12, 13], but not without a significant increase in the transmitter's complexity.

Research progress led to circular FBMC (C-FBMC), a concept developed from GFDM and FBMC. C-FBMC is less complex, is easily extensible to the multiple-input multiple-output (MIMO) antenna, and preserves the orthogonality of the subcarrier's symbols. Most of the other proposed waveforms are based on the schemes described above. They either apply filtering or windowing operation (or both) in either the time or frequency domain. The aim of this section is not to review the literature but provide the basis for the waveform design. Therefore, to keep the content of this article concise, no further review will be pursued here. A detailed review is given in [14].

The discrete wavelet transform (DWT) is another potential waveform candidate for OFDM design. Its application in the OFDM is known as orthogonal wavelet division multiplex (OWDM). The DWT-based signal coding was introduced in [15] followed by many research studies. Some of them are documented in [16–22]. However, these research scopes are generally not broad enough to support the research insights into future wireless application requirements. OWDM's fit for 5G is scarcely reported in the literature. The contribution of this paper is addressing this by checking for both merits and demerits of using OWDM and weighing them against what the future wireless applications require. The 3rd Generation Partnership Project (3GPP) group has selected CP-OFDM (a holdover from 4G) as the signaling option for 5G for the 3GPP's Release 15. Therefore, we compared OWDM to CP-OFDM. We made a detailed comparison between Fourier transform-based CP-OFDM and OWDM using MATLAB, detailed its impact on healthcare, and proposed a guideline on how to approach wavelet application in 5G and beyond.

The rest of this paper is organized as follows: 5G design criteria and its impacts on healthcare are discussed in Section 2. Sections 3 and 4 are about the OFDM and OWDM system model. Section 5 contains simulations and results. Section 6 evaluates the results based on the requirements, while Section 7 draws conclusions and reviews based on the results.

## 2. 5G Design Criteria and Its Impacts on Healthcare

In order to give detailed simulation and comparison, we have taken into consideration all the necessary 5G KPIs [23] in gaging transmitter/receiver systems' efficiency.

*2.1. High Energy Efficiency.* In Section 1, we defined PAPR as the peak-to-average power ratio of the transmission and reception energy. Energy efficiency is mostly defined by the PAPR and degree of computational complexity. At the physical layer, waveforms exhibit peaks and lows. The implication is that, at transmission (or reception), if the difference (ratio) between the peaks and lows is large, higher transient energy will result in the power amplifiers which in turn will lead to higher energy consumption. Therefore, low PAPR is necessary for power-efficient transmissions at uplink (UL), downlink (DL), and side link (SL). Furthermore, computational complexity should be minimal, especially in power-stringent (battery-operated) devices such as the mobile phone or an IOT field sensor. In an IoT system, a further attempt at increasing the energy efficiency at the network layer has been reported in a recent study in [24]. It presents an information-centric networking (ICN) caching strategy that fits well in the energy-efficient IoT environment by utilizing Packet Update Caching (PUC). Considering all these advancements in research, it is safe to say the longevity of battery life and performance will be improved to ensure continuous and uninterrupted remote monitoring. Actually, in 5G, low-power sensors will be designed to operate on the same battery for the full duration of medical operation. This duration can be as much as 10 years [25].

*2.2. Low Device Complexity.* This section could also be termed as transceiver baseband complexity. It is the number of operations required to be performed to transmit and, more importantly, receive a signal successfully with the minimum possible processing overhead. At a very high frequency (e.g., millimeter-wave) and large bandwidths, severe RF impairments may result. With the presence of impairments, the waveform design standard should maintain the least possible computational complexity and processing overhead that could result from filtering, reduction of intersymbol interference (ISI) and intercarrier interference (ICI), windowing, interference cancellation, etc. The direct implication of this is battery longevity. This is similar to what is discussed above.

*2.3. High Reliability.* Reliability is evaluated by bit error rate (BER); it is the capability of a network to carry out a preferred operation with very low error rates. Biomedical sensors with IoT capabilities generate a huge amount of data, therefore error-sensitive. More so, considerable amounts of errors might lead to an increase in latency. Considering the signaling traffic from a massive number of these sensors would even lead to more increase in latency as each sensor is trying to retransmit. The low BER scheme will be supported

by the 5G network. One of the main advantages of the proposed wavelet-OFDM scheme in this article is low BER compared to OFDM.

*2.4. Low Latency.* Ultralow latency defines the network which is optimized to process huge amounts of data packet with a very low tolerance for delay. 5G maximum allowed latency for the ultrareliable low-latency communication (URLLC) applications is less than 1 ms. This is due to services such as machine-to-machine communication requiring fast response time to allow efficient sporadic transmission of small packets. So, it requires a transmission mode with low air-interface latency enabled by very short frames. The efficient transmission of short frames is very much essential for medical IoT. For example, in telesurgery, the maximum acceptable latency (end-to-end) is 200 ms [26]. Comparing this to the 5G network typical latency of less than 1 ms, "telesurgeons" can be guaranteed low latency communication and optimum stability in receiving haptic feedback and improved wireless data rates for better visualization and greater precision. In the future, this will serve as an enabling technology for new telesurgery applications alongside other real-time applications with stricter latency requirements. This can also serve as an enabling technology for emerging e-health fields requiring some forms of wireless transfer of big data and machine learning for early detection of certain diseases. These diseases are not limited to heart disease [27], Parkinson [28], and breast cancer [29]. For further studies about latency reduction in 5G, refer to [30].

*2.5. High Spectral Efficiency or High Bandwidth.* Bandwidth refers to the transmission capacity of a network per given time. Spectral efficiency refers to the efficient use of the available bandwidth. In the 3G and 4G networks, biomedical sensors can only send a limited amount of data due to restricted bandwidth [31]. This limitation is mitigated in 5G by exploring the available (untapped) spectrum at higher frequencies (as high as 10 GHz). Also, at such high frequencies, a higher transmission rate on the order of Gbps is achieved. With this, seamless remote monitoring can be achieved; physicians can view ultra-high-definition contents (videos and pictures) and be able to make better-informed decisions. Furthermore, online consultations can be carried out whereby patients, ordinary citizens, different civic associations, experts, and executive bodies can contribute and exchange medical information.

Having discussed the benefits of 5G's high spectral efficiency, it is necessary to note that several factors could also contribute to the degradation of its spectral efficiency (Sections 5.3 and 5.4). This is why the nature of the waveform is more important. This is one of the problems addressed in the waveform scheme proposed in this article.

*2.6. Massive Asynchronous Transmission.* Massive asynchronous transmission or asynchronous coexistence of a huge number of nodes will be achieved with 5G. This will be achieved with D2D [32]. In D2D communications, each

terminal can communicate with each other directly without routing through gateways and base stations. Therefore, a highly dense network problem can be solved through D2D communications. D2D communications will enable asynchronous coexistence between a large number of medical sensors, wearables, and devices, and monitoring equipment can communicate with minimal interference.

In the 4G network, all communications are routed through gateways and base stations. This routing is inefficient, especially when devices are near each other. Achieving D2D communication will require waveforms with less strict synchronization requirements [33]. The effect of poor synchronization is phase noise. Phase noise including carrier frequency offset (CFO) is both caused by differences in the transmitter and receiver oscillator. The mathematical description of the CFO is the multiplication of a signal in the time domain by a time-varying complex exponential function. CFO will cause a received signal to be shifted in the frequency. Therefore, sampling of the received signal will be done at an offset point, which is not the peak point. The result is a raised ICI [8].

*2.7. MIMO Compatibility.* Massive MIMO is a key technology in delivering mobile 5G. This is essentially grouping together large-scale antennas at the transmitter and receiver to increase throughput and improve spectrum efficiency at the access nodes. Other benefits of massive MIMO are lower latency, simplification of the media access control (MAC) layer, and robustness against (intentional) jamming. Furthermore, beamforming is necessary for overcoming high propagation losses especially at very high frequencies. One of the direct impacts and implications of this is large file transfers. Hospitals will possess the ability to transfer large image files such as medical images. Overall, hospitals will be able to process more patients, given the same amount of time.

One of the various checks of our proposed waveform is its suitability and ease of integrating MIMO with it. It should be less complex and practicable.

*2.8. Robustness to Frequency-Selective and Time-Selective Channels.* Very harsh propagation conditions can cause poor performance in a wireless communication system and result in a loss of signal power without loss of noise power, consequently resulting in a poor signal-to-noise ratio (SNR). OFDM is generally used to combat selective fading because of its robustness to frequency-selective channels. Among other methods are MIMO, rake receivers, space-time codes, forward error correction, interleaving, etc. The robustness of the 5G waveform design should include adapting to this impairment as well as time-selective fading occurring in high-speed scenarios. An example of such a case is the vehicle to anything (V2X). In this case, we can consider the (efficient) transmission of the medical ultrasound video stream from a fast-moving ambulance to the host hospital.

### 3. OFDM System Model

OFDM is the most popular multicarrier modulation scheme that is currently being employed in many standards such as the downlink of 4G LTE and the IEEE 802.11 family [34]. In an OFDM system, at the transmitter, data to be transmitted are mapped to a constellation, split into parallel, and modulated using the inverse fast Fourier transform (IFFT). The modulation process is shown in Figure 1. Guard band and cyclic prefix (CP) are inserted to prevent a delayed version of symbol overlapping with the adjacent symbol and mitigate the delay spread, respectively. The orthogonal signals are then mixed. The key process here is modulation, where signals are mapped from the frequency domain to the time domain and multiplexed. The modulation process is mathematically the summation of  $N$  tones described in [35] and mathematically expressed in the following equation:

$$x_n(t) = \sum_{k=0}^{N-1} s_k[n] e^{j((2k\pi)/T)t}. \quad (1)$$

$x_n[t]$  is the summation of the number of complex-valued sinusoids  $N$  with period  $T$  in the continuous-time domain  $t$ .  $n$  denotes the discrete-time index. The data symbol is  $s_k[n]$ , with frequency component  $k=0, 1, \dots, N-1$ . The multiplexed data symbol  $y_n(t)$  on the baseband is passed through a channel with transfer function  $H(f)$ , which is expressed mathematically in the following equation:

$$y_n(t) = \sum_{k=0}^{N-1} H\left(\frac{k}{t}\right) s_k[n] e^{j(2k\pi/T)t}. \quad (2)$$

The sinusoids  $N$  are located at frequencies  $f = 0, 1/T, 2/T, \dots, (N-1)/T$  for orthogonality. The orthogonality between two adjacent sinusoids is defined in the following equation:

$$\sum_{t=N/2}^{N/2} e^{-j2\pi kt/N} e^{-j2\pi pt/N} = 0, \quad \forall p \neq k. \quad (3)$$

After modulation, the CP denoted as  $p$  is added by copying the last part of the modulated IFFT signal and appending it to the beginning as a guard interval to prevent ISI. In the 4G LTE system, the CP is hard-coded into the waveform, while for 5G NR Release 15, the CP is determined by the maximum delay present in individual channels.

At the receiver, the transmitting process is reversed to decode the received data. For demodulation, the fast Fourier transform (FFT) is employed. OFDM, as opposed to a single-carrier system, has the ability to cope with frequency-selective fading because data are divided and transmitted in parallel streams on a modulated set of subcarriers. This approach results in the efficient use of bandwidth.

### 4. OWDM System Model

Some of the wavelet transform applications are in source and channel coding, signal denoising, and data compression [36, 37]. DWT and inverse discrete wavelet transform (IDWT) are used in OWDM [38, 39]. This replacement is

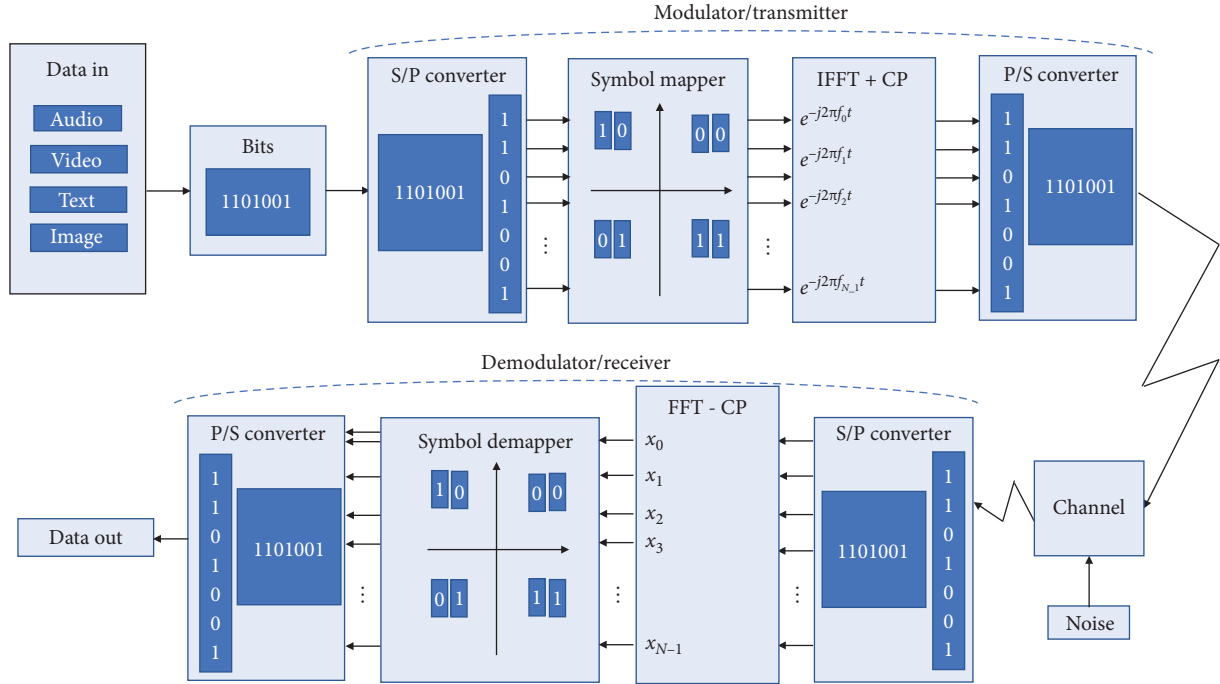


FIGURE 1: Fourier-OFDM modulation and demodulation system process model.

due to its properties such as orthonormality [40–42] and the ability to decompose signals effectively in the time-frequency domains by scaling and shifting. In DWT, the scaling ( $j$ ) and shifting ( $k$ ) results are generated by the mother wavelet denoted by  $\Psi$  as a function of time ( $t$ ) and mathematically expressed in the following equation:

$$\Psi_{j,k}(t) = 2^{-j/2} \Psi(2^{-j}t - k). \quad (4)$$

The discrete wavelet and the inverse discrete wavelet representation of a signal are given by equations (5) and (6), respectively [43]:

$$X_{\text{DWT}}^j_k = \int_{-\infty}^{\infty} x(t) 2^{j/2} \Psi(2^j t - k) dt, \quad (5)$$

$$X_{\text{IDWT}}(t) = \sum_{j=-\infty}^{\infty} \cdot \sum_{k=-\infty}^{\infty} X_k^j 2^{j/2} \Psi(2^j t - k) dt. \quad (6)$$

$x$  is the input signal in the time domain. Therefore, the OWDM symbol can be expressed as the weighted sum of wavelet and scale carriers and is mathematically expressed in the following equation:

$$S(t) = \sum_{j \leq J} \cdot \sum_k w_{j,k}(t) \cdot \Psi_{j,k}(t) + \sum_k a_{j,k} \cdot \Phi_{j,k}(t). \quad (7)$$

$w_{j,k}$  is the sequence of the wavelet, and  $a_{j,k}$  are the approximation coefficients.

The OWDM transmission and reception concept is the same with the OFDM; however, the modulation process is different. It makes use of IDWT for modulation and DWT

for demodulation. This is highlighted in Figure 2. Starting with the modulation process, the data bits are mapped into OFDM symbols of parallel data stream corresponding to the number of subcarriers. This signal is converted to serial  $X(n)$  and fed into IDWT. Here, the signal is vector-transposed ( $V_t$ ), upsampled ( $u_s$ ), and convoluted with a low-pass filter (LPF) and high-pass filter (HPF) [44] using approximated coefficients ( $A_c$ ) and detailed coefficients ( $D_c$ ), respectively. The decomposed outputs from the LPF and HPF are  $L(n)$  and  $H(n)$ , respectively. Different wavelet families have different filter lengths, so the length of the zero pads is adjusted accordingly to maintain orthonormality and orthogonality in  $L(n)$  and  $H(n)$ .  $L(n)$  and  $H(n)$  are mixed and then transmitted.

At the receiver, the modulation process is simply reversed to decode the received data. The received data are decomposed back into  $L(n)$  and  $H(n)$  using the LPF and HPF, respectively.  $H(n)$  contains noise, so it is discarded, while  $L(n)$  is processed for data recovery. In this paper, for our analysis, we considered Haar, Daubechies (DB2), biorthogonal (Bior5.5), and Symlet (Sym4) wavelet transforms. Our consideration is based on the results in [45–47]. The results show that these transforms are the most suitable for waveform design because of their resiliency to channel noise.

**4.1. Haar Wavelet.** The Haar wavelet, discovered by a Hungarian mathematician [48], is a sequence of rescaled “square-shaped” functions which together form a wavelet family or basis. Both wavelet function and scaling function are square-shaped as shown in Figures 3 and 4. The simplest example of an orthogonal wavelet is the Haar function

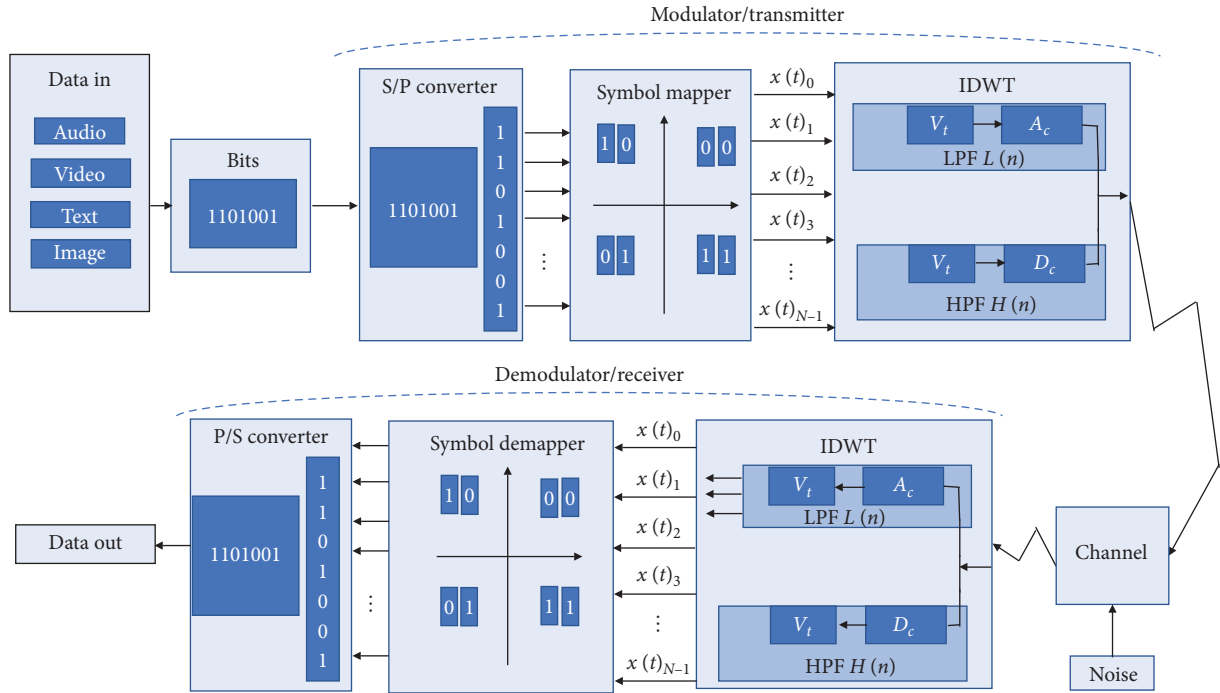


FIGURE 2: Wavelet-OVDM modulation and demodulation system process model.

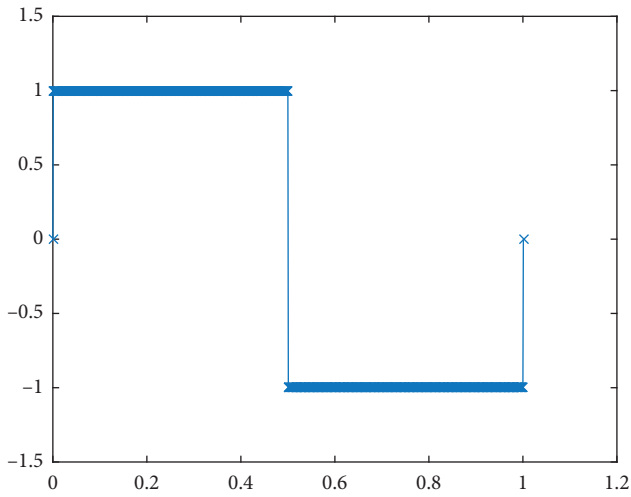


FIGURE 3: Haar wavelet function.

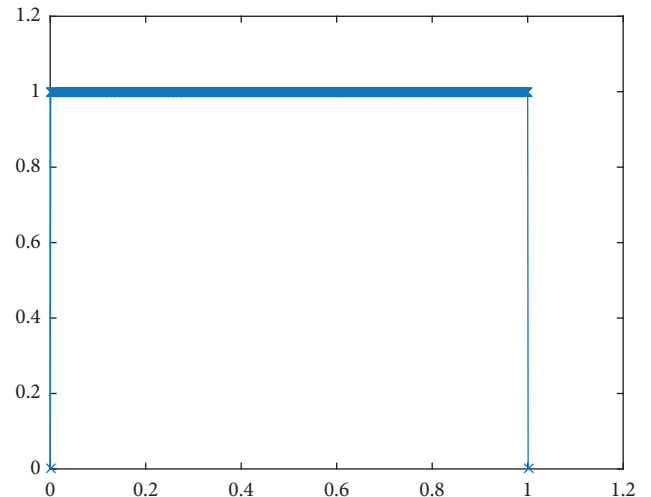


FIGURE 4: Approximation of the Haar scaling function.

denoted by  $\Psi H$  and defined by [49] and mathematically expressed in equations (8) and (9):

$$\Psi(t) = \begin{cases} 1, & \text{for } 0 \leq x < \frac{1}{2}, \\ -1, & \text{for } \frac{1}{2} \leq x < 1, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

Its scaling function  $\varphi(t)$  can be described as

$$\Phi(t) = \begin{cases} 1, & 0 \leq t < 1, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

**4.2. Daubechies Wavelet.** Daubechies wavelet is orthogonal, and it is characterized by a maximal number of vanishing moments for some given support. Daubechies wavelets are usually characterized by  $dbN$ ,  $N$  referring to the order number. Figures 5 and 6 demonstrate the Db4 wavelet and scaling functions. In this paper, the Db4 wavelet is used.

**4.3. Symlet Wavelet.** The Symlet wavelet transforms are in the  $N$  order (Sym4); in this paper, we used the Sym4 wavelet. The Symlets are orthogonal, nearly symmetrical, and bi-orthogonal wavelets. They are a modification of the Db family to improve symmetry [50]. Figures 7 and 8 represent the Sym4 wavelet and scaling functions.

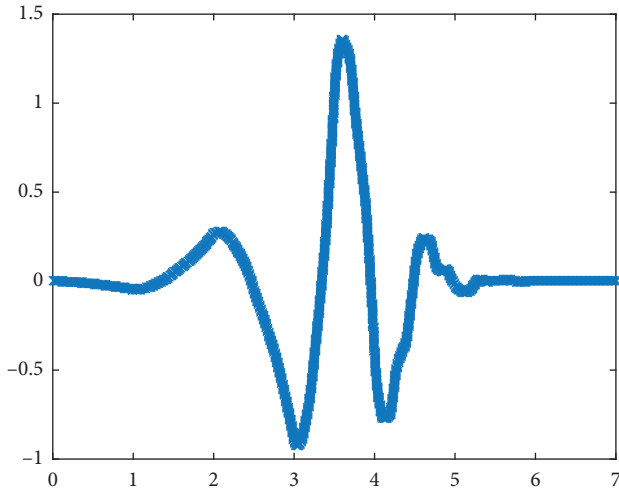


FIGURE 5: Approximation of the Db4 wavelet function.

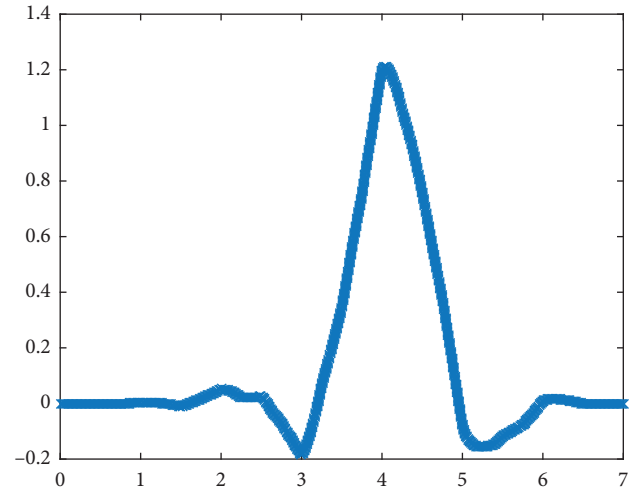


FIGURE 8: Approximation of the Sym4 scaling function.

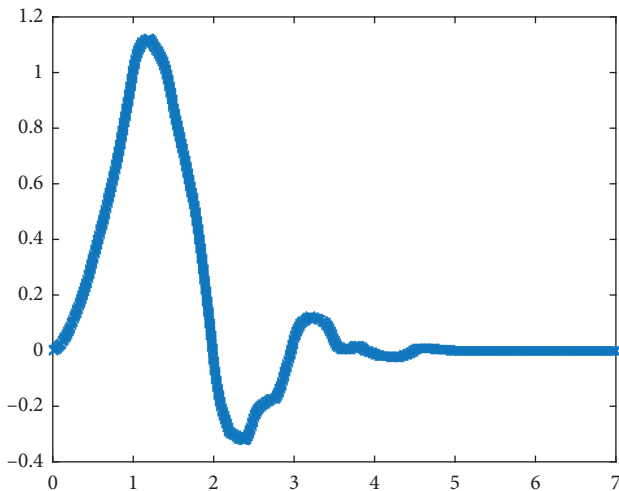


FIGURE 6: Approximation of the Db4 scaling function.

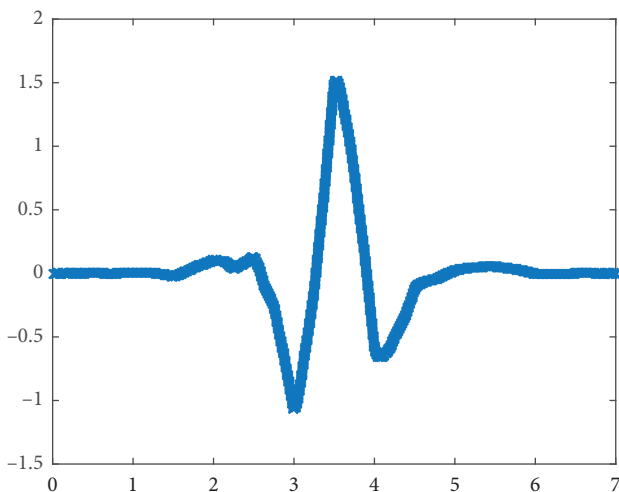


FIGURE 7: Approximation of the Sym4 wavelet function.

**4.4. Biorthogonal Wavelet.** This is a wavelet function where the associated wavelet transform is invertible but not necessarily orthogonal. In many cases, it is required to choose a wavelet and scaling function such that they meet the condition of mirrored impulse response filters, without the loss of orthogonality [51]. In the biorthogonal case, there are two scaling functions for decomposition and reconstruction which must satisfy the following bi-orthogonality condition and are expressed mathematically in equation (10) and graphically demonstrated in Figures 9 and 10.

$$\sum_{n \in \mathbb{Z}} a_n \tilde{a}_{n+2m} = 2 \cdot \delta_{m,0}. \quad (10)$$

## 5. Simulation Results' Analysis

Here, a grayscale image having resolution  $800 \times 800$  pixels of 8 bits depth is fed into the MATLAB simulator as the test data. In total, the image has 5,120,000 bits, which are enough for our simulation. The simulation parameters for the OFDM and OWDM are shown together in Table 1. The major impacting factor is the communication channel. For a more practical approach, we considered the additive white Gaussian noise (AWGN) channel.

**5.1. BER Performance Analysis.** Here, we present the BER versus SNR plot for each candidate and also some images to show their qualities at a varied SNR. An AWGN channel is assumed. From the graphical result, the wavelet transforms show a lower BER than the Fourier transforms and wider BER performance with increasing SNR in Figure 11. The lower BER result is due to the wavelet-based waveform having properties such as orthonormality and the ability to encode signals both with time and frequency localization simultaneously as opposed to Fourier transform which can



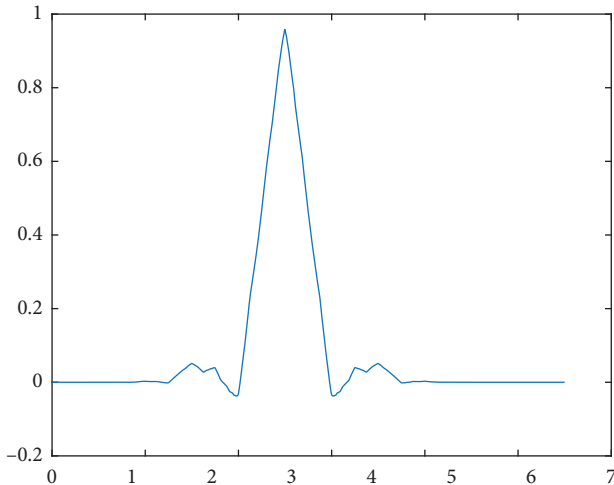


FIGURE 9: Approximation of the biorthogonal wavelet function.

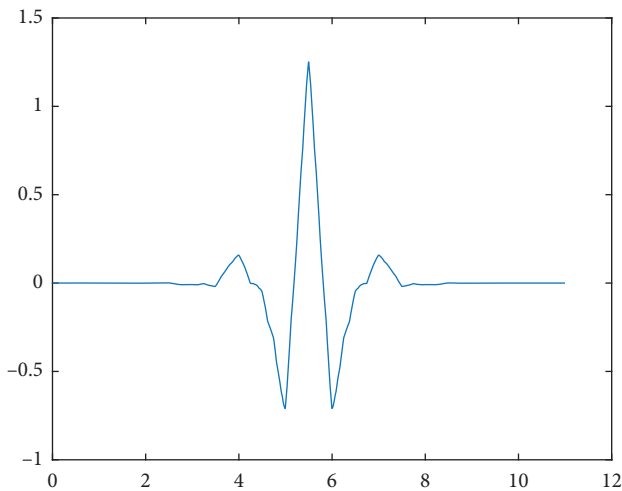


FIGURE 10: Approximation of the biorthogonal scaling function.

TABLE 1: Simulation parameters for OFDM and OWDM.

Modulation	QPSK
FFT size	1024
Number of subcarriers	128
CP length	16
SNR range	1 db–10 db
Channel	AWGN, Rayleigh
Maximum Doppler frequency	0 Hz
Sampling period	1e3
Path delays	0, 1e-3, 3.5e-5, 12e-5
Path gains	0, -1, -1, -3
Carrier frequency	7 GHz
CFO	100 Hz
	Biorthogonal 5.5
Wavelet transforms	Haar Symlet4 Daubechies4

only analyze signals with time localization only. This wavelet transform property enables efficient encoding (and decoding) of signals, hence the robustness to the channel error.

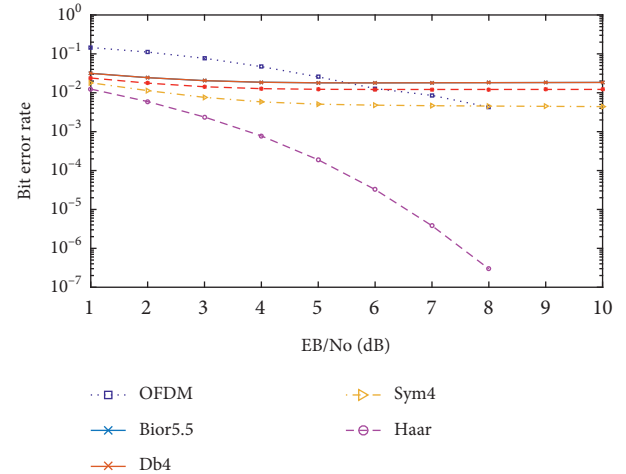


FIGURE 11: BER performance comparison in the AWGN channel.

**5.2. PSD and PAPR Performance.** The PAPR is a KPI of a wireless communication system as it goes a long way in shaping the design of the power amplifier's level of linearity and therefore its cost. PAPR is evaluated by computing the complementary cumulative distribution function (CCDF) of the waveform with respect to the number of subcarriers and constellation order. Transmit PAPR performance is given in Table 2, and the plot of the received PAPR against a varied SNR is shown in Figure 12 for the Fourier-OFDM and the wavelet modulations.

At transmission and at reception with a varied SNR of 1 dB to 10 dB, all the wavelet variants have better and distinct PAPR performance than the OFDM with the Haar wavelet having the best performance. Although the wavelet transforms have lower PAPR, they use more transmit power than Fourier. In Figure 12, a plot of the received PAPR at various SNRs is shown, including the power spectral density performance (PSD images) of various transforms at 128 subcarriers each.

In a nonlinear amplifier, OFDM is susceptible to spectral regrowth compared to other wavelets due to its high PAPR. The theory behind this is the broadening of a bandwidth of a modulated signal with large envelope fluctuations due to nonlinearities that generate mixing products between the individual frequency components of the spectrum. The end result is adjacent channel interference. Figures 13–17 are the PSD plots of various transforms. OFDM has a higher sidelobe (noise) than the rest of the wavelet transforms.

Table 3 is the bandwidth result from transmitting the test data at 7 GHz as specified at the beginning of this section. The very high use of bandwidth by Fourier (low spectral efficiency) is as a result of the need for cyclic prefixing.

**5.3. Effect of the Carrier Frequency Offset.** In this section, we assume that both users are perfectly synchronized in the time domain but with an offset between their respective carrier frequencies. In Figure 18, we present the BER curve at CFO = 0.1 kHz. The simulation result shows that OFDM suffers the most effect from CFO, while Haar incurs the least effect. Also, it is observed that the performance disparity between the transforms increases roughly with the SNR.

TABLE 2: Transmit PAPR of various transforms and the transmit power.

Transform	Transmit PAPR (dB)	Transmit power (dB)
Fourier	18.95	-21.10
Haar	$1.93e-15$	-6.04
Sym4	5.35	-6.04
Db4	7.18	-6.04
Bior5.5	6.87	-3.67

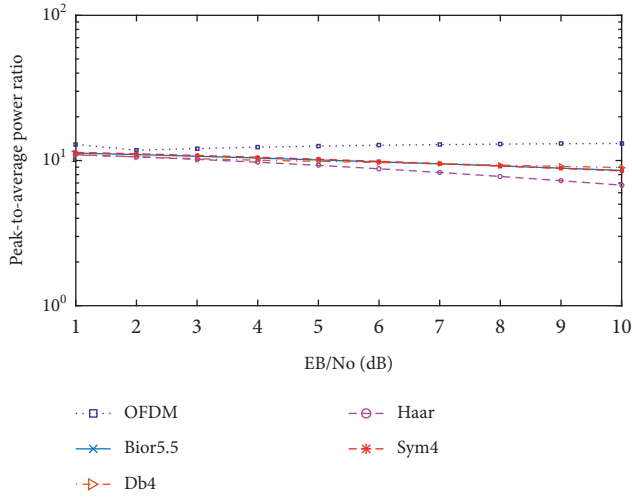


FIGURE 12: Received PAPR performance of various transforms.

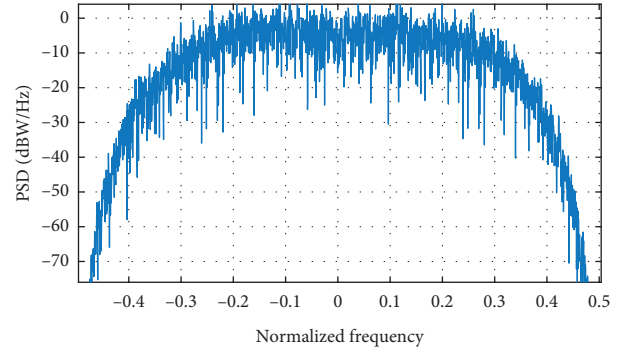


FIGURE 15: Sym4 PSD plot.

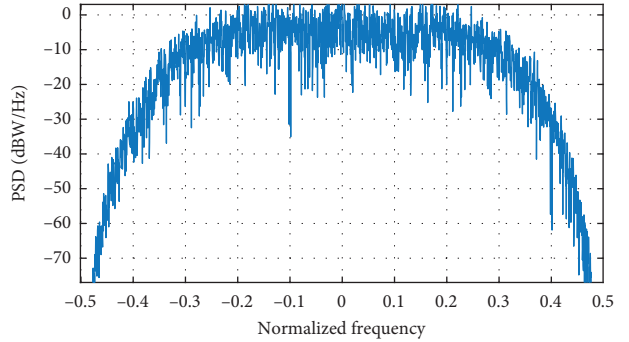


FIGURE 16: Db4 PSD plot.

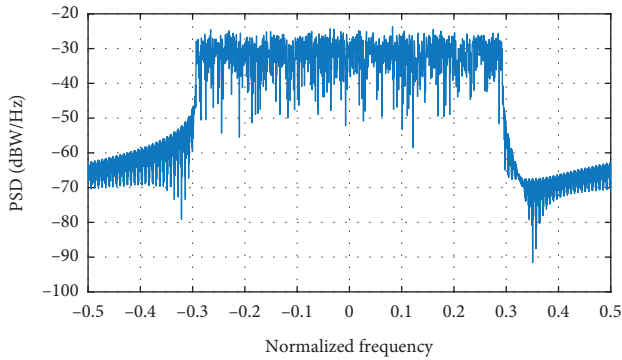


FIGURE 13: OFDM PSD plot.

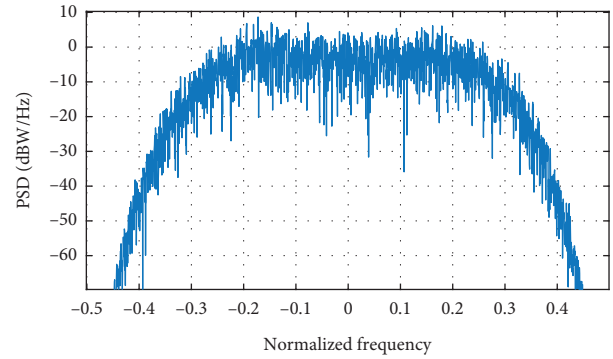


FIGURE 17: Bior5.5 PSD plot.

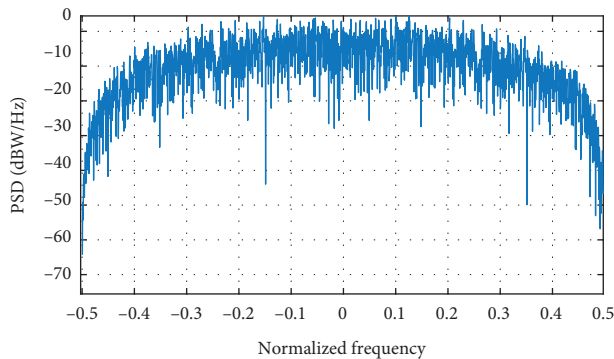


FIGURE 14: Haar PSD plot.

TABLE 3: Occupied bandwidth of various transforms at 7 GHz.

Transform	Bandwidth (Hz)
Fourier	418.615
Bior5.5	286.212
Haar	363.605
Sym4	376.248
Db4	374.307

5.4. *Effect of Multipath Fading.* The test data as described at the beginning of Section 5 are transmitted and received using various transforms to measure the impact of multipath fading on them. The simulation parameters are also exactly as given in the aforementioned section, i.e., four taps with their respective channel gains in a Rayleigh multipath fading channel. In order to observe the maximum BER per SNR, channel state

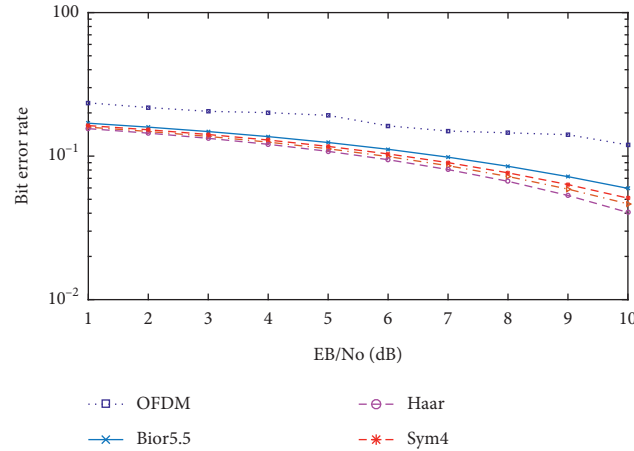


FIGURE 18: BER versus Eb/No at CFO = 0.1 kHz.

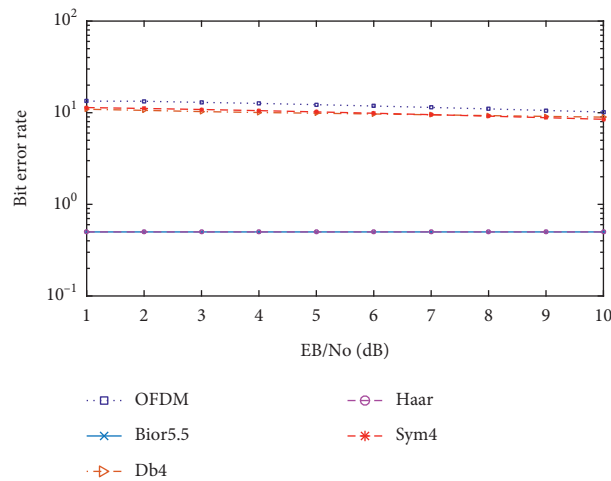


FIGURE 19: OFDM and OWDM BER vs. SNR in the frequency-selective Rayleigh fading channel.

TABLE 4: Overhead of various transforms.

Transform	Time (ms)		Peak Mem (MB)
	Modulation + demodulation		Modulation + demodulation
Fourier	212		90
Haar	1980		320
Bior5.5	2215		308
Sym4	2209		312
Db4	2239		312

information (CSI) and equalization were not considered. It is observed that the OWDM transforms are considerably robust to the multipath fading effect. In Figure 19, OFDM and OWDM's BER vs. SNR in the frequency-selective Rayleigh fading channel are graphically demonstrated.

**5.5. Overhead Cost and Complexity of Implementation.** The time and memory requirements for both at the transmitter and at the receiver are simulated and shown in Table 4. The simulation of the overhead cost is based on the transmission and reception of the image with the size as

described at the beginning of Section 4. For the purpose of accuracy, the time and memory computation costs are limited to IFFT/FFT and IDWT/DWT only.

In Table 4, it is observed that although OWDMs might seem like an attractive choice, they generally require more computational resources than the OFDM.

## 6. OFDM and OWDM Evaluation Metrics

Using the simulation results above, we assess OFDM and OWDM for a number of KPIs, taking the 5G set standards into consideration. These comparisons are outlined in Table 5.

TABLE 5: Comparative analysis of OFDM and OWDM.

5G performance metrics	Fourier	Haar	Db2 Sym4 Bior5.5
High energy efficiency	PAPR is high. Computational complexity is low.	PAPR is low, and computational complexity is high. Haar has the lowest processing time, but memory requirement is higher than other wavelet pairs.	
Low device complexity	Needs one block per process; IFFT and FFT.	Require a larger device processing capacity. Wavelet transforms generally need two blocks (i.e., their filter banks) for each process (i.e., IDWT and DWT) and more memory than Fourier. However, advances in the clock speeds and memory sizes of communication devices have increased by 4 to 6 orders of magnitude in the past 40 years [52]. Therefore, future wise, complexity might not be a challenge.	
High reliability	Has poorer BER performance compared to OWDM.	Have better BER performance under a wide varied SNR.	
Low latency	Requires the use of CP to reduce ISI which resultantly reduces throughput.	Do not require CP.	
High spectral efficiency	It has a considerable spectral efficiency.	They have better spectral efficiency compared to Fourier due to Fourier's requirement of higher bandwidth (CP).	
Massive asynchronous transmission	Spectral localization of the subcarriers is weak and highly susceptible to the CFO effect.	They are more localized in time and space. Consequently, they are more immune to the CFO effect.	
MIMO compatibility	OFDM is quite compatible with MIMO because of its ease of implementation.	The use of OWDM with MIMO is feasible and has been documented in the literature [36].	
Frequency-time channel selectivity	OFDM is robust to frequency-selective channels and can be made robust to time channel selectivity by a proper choice of subcarrier spacing.	OWDM is more robust to frequency-selective channels, especially at low SNRs.	
Flexibility and scalability	By proper choice of design parameters, OFDM is a flexible waveform that can support diverse services, with several evolved versions.	OWDM is also a flexible waveform. More so, there are different wavelet transforms with different scalable properties. Therefore, its flexibility [53] could be explored and tailored to the requirements of the 5G NR classifications.	

TABLE 6: Proposed assessment of OFDM and OWDM for 5G NR.

Performance indicators	OFDM assessment	OWDM assessment	Downlink requirement	Uplink requirement	Side link requirement	V2X requirement	Backhaul requirement
Spectral efficiency	High	Very high	Very high	Very high	High	Very high	Very high
MIMO compatibility	High	High	Very high	Very high	High	Very high	Very high
Time localization	High	Very high	High	High	High	Very high	Very high
Transceiver baseband complexity	Low	High	Very high	High	Very high	High	High
Robustness to frequency-selective channels	High	High	High	High	High	High	High
Robustness to time-selective channels	Medium	High	High	High	High	Very high	Low
Robustness to phase noise	Medium	High	High	High	High	High	High
Robustness to sync errors	High	Very high	Low	High	High	Medium	Low
PAPR	High (can be reduced)	Low	Low	High	High	Medium	Low
Frequency localization	Low (can be reduced)	High	Medium	Medium	Medium	Medium	Low

The proposed application guidelines are given in Table 6 based on the 5G NR design criteria and waveform assessment discussed in [54].

## 7. Conclusion

The research presented in this paper provides a deep insight into the consideration of OWDM for 5G NR physical layer design and beyond. KPI assessment of OFDM and OWDM and their respective performance simulation results have been carried out. Our results show that OWDM outperforms OFDM in many of the KPIs, and it is, therefore, a more efficient IoT enabler for healthcare and other vital areas. However, our results also show that the main drawback of OWDM is high memory and time cost. In view of this, careful tradeoff considerations are required. Furthermore, we also include uplink, downlink, side link, V2X, and backhaul links as further assessments. These various link assessments are necessary for OWDM waveform numerology establishments for 5G's diverse applications. Finally, we observed that, in the establishment of OWDM waveform numerologies for 5G, CP will not be required. This means the establishment process can be less complicated than that of OFDM.

## Data Availability

No data were used to support this study.

## Conflicts of Interest

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

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 61370073), the National High Technology Research and Development Program of China (Grant no. 2007AA01Z423), and the Project of Science and Technology Department of Sichuan Province.

## References

- [1] A. Ijaz, L. Zhang, M. Grau et al., "Enabling Massive IoT in 5G and beyond systems: PHY radio frame design considerations," *IEEE Access*, vol. 4, pp. 3322–3339, 2016.
- [2] S. B. Baker, W. Xiang, and I. Atkinson, "Internet of things for smart healthcare: technologies, challenges, and opportunities," *IEEE Access*, vol. 5, pp. 26521–26544, 2017.
- [3] S. K. Goudos, M. Deruyck, D. Plets et al., "A novel design approach for 5G massive MIMO and NB-IOT green networks using a hybrid Jaya-differential evolution algorithm," *IEEE Access*, vol. 7, pp. 105687–105700, 2019.
- [4] T. S. Rappaport, Y. Xing, O. Kanhere et al., "Wireless communications and applications above 100 GHz: opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019.
- [5] A. B. Kihero, M. S. J. Solaija, and H. Arslan, "Inter-numerology interference for beyond 5G," *IEEE Access*, vol. 7, pp. 146512–146523, 2019.
- [6] Y. Medjahdi, S. Traverso, R. Gerzaguet et al., "On the road to 5G: comparative study of physical layer in MTC context," *IEEE Access*, vol. 5, pp. 26556–26581, 2017.
- [7] C. An and H. Ryu, "CPW-OFDM (cyclic postfix windowing OFDM) for the B5G (Beyond 5th Generation) waveform," in *Proceedings of the IEEE 10th Latin-American Conference on Communications (LATINCOM)*, pp. 1–4, Guadalajara, Mexico, November 2018.
- [8] F. Kalbat, A. Al-Dweik, B. Sharif, and G. K. Karagiannidis, "Performance analysis of precoded wireless OFDM with carrier frequency offset," *IEEE Systems Journal*, vol. 14, 2020.
- [9] O. Daoud, "Power reallocation and complexity enhancement for beyond 4G systems," *China Communications*, vol. 16, no. 6, pp. 114–128, 2019.
- [10] M. Payaró, A. Pascual-Iserte, and M. Najar, "Performance comparison between FBMC and OFDM in MIMO systems under channel uncertainty," in *Proceedings of the European Wireless Conference (EW)*, pp. 1023–1030, Lucca, Italy, April 2010.
- [11] G. Fettweis, M. Krondorf, and S. Bittner, "GFDM-generalized frequency division multiplexing," in *Proceedings of the IEEE 69th Vehicular Technology Conference (VTC)*, pp. 1–4, Barcelona, Spain, April 2009.
- [12] N. Michailow, I. Gaspar, S. Krone, M. Lentmaier, and G. Fettweis, "Generalized frequency division multiplexing: analysis of an alternative multi-carrier technique for next generation cellular systems," in *Proceedings of the 2012 International Symposium on Wireless Communication Systems (ISWCS)*, pp. 171–175, Paris, France, August 2012.
- [13] M. Matthe, L. L. Mendes, and G. Fettweis, "Space-time coding for generalized frequency division multiplexing," in *Proceedings of the 20th European Wireless Conference*, pp. 1–5, Barcelona, Spain, May 2014.
- [14] B. Farhang-Boroujeny and H. Moradi, "OFDM inspired waveforms for 5G," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2474–2492, 2016.
- [15] M. A. Tzannes and M. C. Tzannes, "Bit-by-bit channel coding using wavelets," in *Proceedings of the [Conference Record] GLOBECOM'92-Communications for Global Users*, vol. 2, pp. 684–688, Orlando, FL, USA, December 1992.
- [16] M. Chafii, Y. J. Harbi, and A. G. Burr, "Wavelet-OFDM vs. OFDM: performance comparison," in *Proceedings of the 23rd International Conference on Telecommunications (ICT)*, pp. 1–5, Thessaloniki, Greece, May 2016.
- [17] S. A. Dawood, F. Malek, M. S. Anuar, and S. Q. Hadi, "Discrete multiwavelet critical-sampling transform-based OFDM system over Rayleigh fading channels," *Mathematical Problems in Engineering*, vol. 2015, Article ID 676217, 10 pages, 2015.
- [18] K. Lavish, V. Sharma, and J. S. Malhotra, "MIMO-WiMAX system incorporated with diverse transformation for 5G applications," *Frontiers of Optoelectronics*, vol. 12, pp. 1–15, 2019.
- [19] R. Ayaswarya and N. Amutha Prabha, "Fractional wavelet transform based OFDM system with cancellation of ICI," *Journal of Ambient Intelligent Humanized Computing*, 2019.
- [20] S. Tripathi, A. Rastogi, K. Sachdeva, M. K. Sharma, and P. Sharma, "PAPR reduction in OFDM system using DWT with nonlinear high-power amplifier," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, no. 5, pp. 2278–3075, 2013.
- [21] N. Ali, M. I. Youssef, and I. F. Tarrad, "ICI reduction by parallel concatenated encoder using wavelet transforms," *Advances in Intelligent Systems and Computing*, vol. 933, 2020.

- [22] J. Lee and H. Ryu, "Design and comparison of discrete wavelet transform based OFDM (DWT-OFDM) system," in *Proceedings of the Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 881–885, Prague, Czech Republic, July 2018.
- [23] A. F. Demir, M. Elkourdi, M. Ibrahim, and H. Arslan, "Waveform design for 5G and beyond," *5G Networks: Fundamental Requirements, Enabling Technologies, and Operations Management*, pp. 51–76, 2018.
- [24] I. U. Din, S. Hassan, A. Almogren, F. Ayub, and M. Guizani, "PUC: packet update caching for energy efficient IoT-based information-centric networking," *Future Generation Computer Systems*, vol. 111, pp. 634–643, 2020.
- [25] H. T. Mouftah, M. Erol-Kantarci, and M. H. Rehmani, *Transportation and Power Grid in Smart Cities: Communication Networks and Services*, Wiley, Hoboken, NJ, USA, 2018.
- [26] Q. Han, S. Liang, and H. Zhang, "Mobile cloud sensing big data and 5G networks make an intelligent and smart world," *IEEE Networks*, vol. 29, no. 2, pp. 40–45, 2015.
- [27] U. H. Amin, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, Article ID 3860146, 21 pages, 2018.
- [28] A. U. Haq, J. P. Li, M. H. Memon et al., "Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings," *IEEE Access*, vol. 7, pp. 37718–37734, 2019.
- [29] U. H. Amin, "A novel integrated diagnosis method for breast cancer detection," *Journal of Intelligent and Fuzzy Systems*, vol. 38, pp. 1–16, 2020.
- [30] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile Internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, 2016.
- [31] O. O. Fagbohun, "Comparative studies on 3G,4G and 5G wireless technology," *IOSR Journal of Electronics and Communication Engineering*, vol. 9, no. 2, pp. 133–139, 2014.
- [32] V. Kumar, S. Yadav, D. N. Sandeep, S. Dhok, R. K. Barik, and H. Dubey, "5G cellular: concept research work and enabling technologies," in *Advances in Data and Information Sciences*, pp. 327–338, Springer, Singapore, 2019.
- [33] A. Ahad, M. Tahir, and K. A. Yau, "5G-Based smart healthcare network: architecture, taxonomy, challenges and future research directions," *IEEE Access*, vol. 7, pp. 100747–100762, 2019.
- [34] T. Hwang, C. Yang, G. Wu, S. Li, and G. Ye Li, "OFDM and its wireless applications: a survey," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 1673–1694, 2009.
- [35] M. Veena and S. Shanmukha, "Performance analysis of DWT based OFDM over FFT based OFDM and implementing on FPGA," *International Journal of VLSI Design & Communication Systems*, vol. 2, pp. 119–130, 2011.
- [36] G. K. Rao and A. S. S. Rao, "Performance analysis of adaptive MIMO based OFDM using FFT and DWT," *International Journal of Advanced Research in Science and Technology*, vol. 4, pp. 262–266, 2015.
- [37] M. Chafii, J. Palicot, and R. Gribonval, "Wavelet modulation: an alternative modulation with low energy consumption," *Comptes Rendus Physique*, vol. 18, no. 2, pp. 156–167, 2017.
- [38] S. Linfoot, M. Ibrahim, and M. Al-akaidi, "Orthogonal wavelet division multiplex: an alternative to OFDM," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 278–284, 2007.
- [39] A. R. Lindsey, "Wavelet packet modulation for orthogonally multiplexed communication," *IEEE Transactions on Signal Processing*, vol. 45, no. 5, pp. 1336–1339, 1997.
- [40] K. Kaur, "Discrete wavelet transform based OFDM system using convolutional encoding," M.S. thesis, Thapar University, Patiala, India, 2014.
- [41] S. Sheela, T. P. Surekha, and R. Arjun, "Analysis of BER in OFDM using wavelet and FFT based method," in *Proceedings of the International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCT-CEEC)*, pp. 473–476, Mysore, India, September 2017.
- [42] M. GovindaRaju and B. V. Uma, "Design and simulation of wavelet OFDM with wavelet denoising on AWGN channel," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 8, pp. 3015–3018, 2013.
- [43] K. Anitha, K. Dharmistan, and N. J. R. Muniraj, "Modified lifting based DWT/IDWT architecture for OFDM on virtex-5 FPGA," *Global Journal of Research in Engineering*, vol. 12, 2012.
- [44] K. Abdullah and Z. M. Hussain, "Simulation of models and BER performances of DWT-OFDM versus FFT-OFDM," *Discrete Wavelet Transforms*, IntechOpen, London, UK, 2011.
- [45] R. Kanti and M. RaiDr., "Comparative analysis of different wavelets in OWDM with OFDM for DVB-T," *International Journal of Advancements in Research & Technology*, vol. 2, no. 3, 2013.
- [46] R. Bodhe, S. Joshi, and S. Narkhede, "Performance comparison of FFT and DWT based OFDM and selection of mother wavelet for OFDM," *International Journal of Computer Science and Information Technologies*, vol. 3, no. 3, pp. 3993–3997, 2012.
- [47] G. Gowri, G. Uma Maheswari, E. Vishnupriya, S. Prabha, D. Meenakshi, and N. R. Raajan, "Performance analysis of DWTOFDM and FFT-OFDM systems," *International Journal of Engineering and Technology*, vol. 5, 2013.
- [48] V. MB and M. N. S. Swamy, "Low power pipelined DWT-IDWT architecture for OFDM system on FPGA," *Procedia Engineering*, vol. 30, pp. 466–474, 2012.
- [49] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, Cambridge, MA, USA, 1999.
- [50] M. S. Chavan, N. Mastorakis, and M. Gaikwad, "in Proceedings of the Joint WSEAS International Conferences on Recent Researches in Communications," *Automation, Signal Processing, Nanotechnology, Astronomy and Nuclear Physics*, pp. 37–41, Cambridge, UK, 2011.
- [51] G. Pajares and J. Manuel de la Cruz, *A Wavelet-Based Image Fusion Tutorial Pattern Recognition*, pp. 1855–1872, Elsevier, Amsterdam, Netherlands, 2004.
- [52] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3029–3056, 2015.
- [53] E. F. James and P. Beatrice, "An overview on wavelets in source coding, communications, and networks," *EURASIP Journal on Image and Video Processing*, vol. 2007, no. 1, 2007.
- [54] A. A. Zaidi, R. Baldemair, H. Tullberg et al., "Waveform and numerology to support 5G services and requirements," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 90–98, 2016.

## Research Article

# Deep Learning Algorithm for Brain-Computer Interface

Asif Mansoor,<sup>1</sup> Muhammad Waleed Usman,<sup>2</sup> Noreen Jamil ,<sup>2</sup> and M. Asif Naeem <sup>2</sup>

<sup>1</sup>National University of Sciences and Technology, Islamabad, Pakistan

<sup>2</sup>National University of Computer and Emerging Sciences, Islamabad, Pakistan

Correspondence should be addressed to Noreen Jamil; [noreen.jamil@nu.edu.pk](mailto:noreen.jamil@nu.edu.pk)

Received 23 January 2020; Accepted 10 July 2020; Published 25 August 2020

Academic Editor: Iván García-Magariño

Copyright © 2020 Asif Mansoor 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.

Electroencephalography-(EEG-) based control is a noninvasive technique which employs brain signals to control electrical devices/circuits. Currently, the brain-computer interface (BCI) systems provide two types of signals, raw signals and logic state signals. The latter signals are used to turn on/off the devices. In this paper, the capabilities of BCI systems are explored, and a survey is conducted how to extend and enhance the reliability and accuracy of the BCI systems. A structured overview was provided which consists of the data acquisition, feature extraction, and classification algorithm methods used by different researchers in the past few years. Some classification algorithms for EEG-based BCI systems are adaptive classifiers, tensor classifiers, transfer learning approach, and deep learning, as well as some miscellaneous techniques. Based on our assessment, we generally concluded that, through adaptive classifiers, accurate results are acquired as compared to the static classification techniques. Deep learning techniques were developed to achieve the desired objectives and their real-time implementation as compared to other algorithms.

## 1. Introduction

**1.1. Background.** Brain-computer interface (or BCI) is basically setting up a connection between the human brain and the computer device to control or to perform certain activity using brain signals. These brain signals are translated as an action for a device. The interface thus provides a one-to-one communication pathway between the brain and the target.

The technology has advanced from mechanical devices and touch systems, and now, world is approaching towards use of neural waves as the input. Even though it is not widely applied for now, it has a promising future. Especially for the physically impaired people who face difficulties in performing physical activities and lose their brain signal to move their muscles, it is the only way to function.

A BCI system includes a device with electrodes that act as sensors and measure brain signals, an amplifier to raise the weak neural signals, and a computer which decodes the signals into controlling signals to operate devices. Mostly, the BCI device is a headset which is portable and wearable.

The BCI device has two functions. Firstly, it records the data reviewed at its electrodes, and secondly, it interprets or decodes neural signals.

Nervous system resembles an electrical system which passes nerve impulses as a message. This means neurons (brain cells) communicate by transmitting and receiving very small electrical waves, merely in range of microvolts. Now, to sense and record these signals, we require precise and advanced sensors.

Electrodes are set directly on the scalp or embedded in the brain which requires surgical procedure. The non-surgical method of electrode placement though does not damage the brain, it yields poor-quality brain signals. Those that are recorded directly from the scalp yield better results but at the risk of surgery that may induce damage in the brain. The risk of damaging brain tissues exceeds the quality obtained through the surgical method. BCI is therefore a better pathway for neurorehabilitation for paralyzed people. Apart from these, other techniques include functional MRI (fMRI) and magnetoencephalography (MEG). fMRI maps brain activity with an MRI scanner, while MEG is a brain imaging process that identifies brain activity. Electric currents flowing through the brain produce magnetic field, and these are sensed by highly sensitive magnetometers. Both fMRI and MEG techniques use large and expensive machines. Another noninvasive

methodology is near-infrared spectroscopy (NIRS). In this process, neural signals are recorded by passing NI light through the head. The quality of the brain activity measurement is not adequate for the brain computer interface.

In case of healthy people, the brain transmits signals from the central nervous system to the muscles, and thus, they can move the muscles of the body. However, in case of people suffering from stroke or neuromuscular illness, the transmission of signals between the brain and the rest of body muscles is distorted. The patient's body becomes paralyzed or loses the capability to control muscle movement, like cerebral palsy. It is observed that a patient may not be able to move a muscle, but a brain can transmit the neural signal. This means that the neural signal is transmitted from the CNS but not received by target muscles. A BCI can be designed to utilize those commands to control devices or computer programs.

Each part of the body is controlled by a particular part of the brain as shown in the figure. Using BCI techniques, it is observed which part of the brain is active and transmitting the signal. Through this, the BCI system can predict the muscle locomotion from the brain activity [1].

BCI systems can be advanced, and multiple new applications can be developed using a fact that a variety of other brain activities can also be recognized. For instance, while one performs a numeric calculation, the frontal lobe is activated, and when one comprehends a language, Wernicke's area is activated.

Currently, numerous groups are contributing to the evolution of BCIs so as to develop numerous applications, specific for each category of the consumer. Each day, scientists and engineers are improving algorithms, BCI sensor devices, and techniques for quality attainment of data and improved accuracy of systems.

The problem is which method is optimal to analyze these complex time-varying neural responses and map them accordingly to the output response desired. These signals are merely in the range of microvolts. So, these electrical signals are passed through several processes to remove noise and to gather useful signals. Next, algorithms and classification techniques are applied to the data obtained [2].

*1.2. Preliminaries.* To attain a better understanding of BCI systems and the processes that undergo within them, an explanation of the terminologies and the said processes is presented as follows.

*1.2.1. Brain Waves.* Brain waves are oscillating voltages bearing amplitudes from microvolts to some millivolts; there are 5 widely known brain waves bearing different frequency ranges exhibiting states of the brain as shown in Table 1 [3].

*1.2.2. Brain Activity Recording Methods for the BCI.* The neural activity of the brain can be analyzed and understood based on the recording methods used. Recording methods of the BCI can be categorized as follows:

TABLE 1: Brain waves and associated frequencies.

Frequency band	Frequency	Brain states
Gamma	>35 Hz	Concentration
Alpha	12–35 Hz	Anxiety, relaxed, external attention
Beta	8–12 Hz	Very relaxed, passive attention
Theta	4–8 Hz	Deeply relaxed, inward focus
Delta	0.5–4 Hz	Sleep

(1) *Invasive Recording Techniques.* Invasive recording methods are those in which the electrodes are inserted deep in the brain using surgical methods, and the quality of the signal generated is better as compared to its noninvasive counterpart; however, issues arise from long-term stability, and protection is required to hinder them from creating infections. One such example is electrocorticography (ECoG), which measures the brain activity from the neural cortex.

(2) *Noninvasive Recording Techniques.* Noninvasive techniques do not require any surgical treatment and thus safe from causing any sort of infections; though their signal quality is low, it is still a popular means of brain signal acquisition.

These techniques include electroencephalography (EEG) in which the electrical activity is recorded from the scalp of the brain and magnetoencephalography (MEG) in which magnetic properties exhibited due to the difference in oxygenated and deoxygenated hemoglobin are recorded.

For our project, we will opt for an EEG-based signal recording technique and explain its characteristics in the following [1].

*1.3. Electroencephalography (EEG).* Introduced by Hans Berger in 1929, EEG is a measurement of voltage levels that underlines the activity of the brain in response to an event or a stimulus. EEG method comprises electrodes placed on the scalp of the brain at different locations as specified in Figure 1 with temporary glue. The electric signals are generated due to the ionic content present in the brain consisting of Na<sup>+</sup>, Ca<sup>++</sup>, K<sup>+</sup>, and Cl<sup>-</sup> ions; the transportation of these ions invokes the electric potential used in EEG.

The EEG signals are of low quality because of different layers of tissues between the EEG cap and the signal source as shown in Figure 2. The potential created is in a range of tens of microvolts, and these electrodes need to have powerful amplifiers in order to acquire meaningful signals.

*1.4. Need of BCI.* Brain-computer interface-based technology is a developing field, and it has been under focus by many industries to innovate and make everyday life tasks easier. One of the questions which arises in the mind is why we need BCI systems? BCI system is a complex technology, no doubt, however, leading to a simpler life.

Following are the main reasons why we need to focus on this technology:

- (i) Control of devices can be made easy through just our thoughts



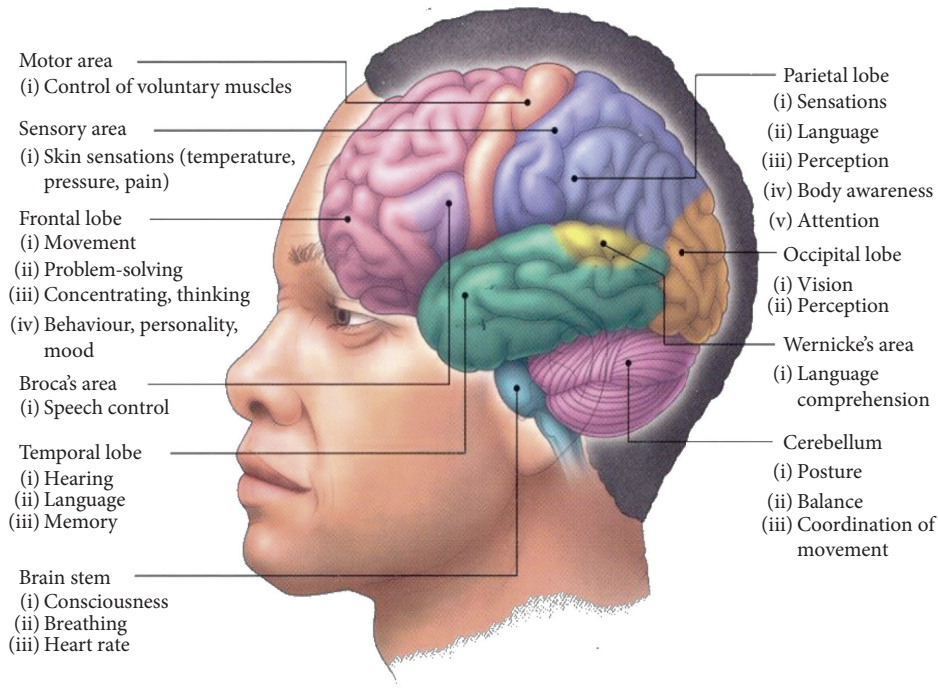
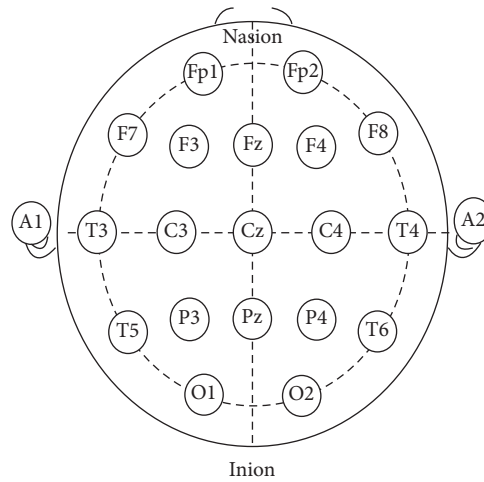


FIGURE 1: Brain anatomy.



(a)



(b)

FIGURE 2: EEG: (a) subject wearing a 32-electrode EEG cap; (b) standardized electrode placements.

- (ii) Making a decision and then performing a task takes time, while operating a device using thoughts or technically our brain waves is easier
- (iii) Re-establishing communication path from the brain to artificial limbs and assisting those affected by brain-related diseases

1.4.1. *Individuals in Need of a BCI to Re-Establish Motor Control and Communication.* A wide range of neurological diseases such as motor neuron diseases and spinal cord injury may lead to severe paralysis of the motor muscles, restricting the patient to control artificial devices through

only a few muscles and thereby termed as “locked-in,” while people who have completely lost their motor control are termed as “completely locked-in.”

Evident that the normal communication channel from the brain to the limbs is lost, BCI is used to re-establish the communication through an alternative route.

Even being applicable to a healthy person, BCI systems can be used to employ numerous tasks from the users using the signals generated from the brains to control applications as presented in the following [4]:

(1) *Noninvasive Brain-Computer Interface Research at the Wadsworth Center.* The research conducted at the

Wadsworth Center was to study different approaches employed in the BCI to control a computer screen cursor to analyze their advantages and disadvantages; one approach was sensory-based rhythm control in which the selected features in the frequency domain were based on the potentials created by motor imagery and linear regression was employed so that they can be converted as control signals to move the cursor.

The other procedure was the P300-based cursor control in which the user focuses attention on the desired symbol and is provided with a  $6 * 6$  matrix to produce time-varying stimuli and linear regression is utilized to allow these signals as a control input to move the cursor.

The research suggested that the BCI is an application-oriented approach and depends entirely on user training; the EEG features dictate the BCI system for speed, accuracy, bit rate, and usefulness.

Sensorimotor Rhythms (SMR) is an approach employing better results for control tasks such as controlling a screen cursor, while the P300-BCI system was slower as compared to the SMR-BCI.

(2) *The Berlin Brain-Computer Interface: Machine Learning-Based Detection of User-Specific Brain States.* The researchers for the Berlin brain-computer interface employed sensory motor rhythms, i.e., thinking of moving the left hand or right hand and used machine learning-based detection of the user specific brain states. While testing their trained model, they achieved an information transfer rate above 35 bits per minute (bpm), and overall spelling speed was 4.5 letters per minute including correcting the mistakes, using 128-channel EEG and using feedback control for untrained users in order to properly train the machine learning algorithms, thereby reducing the training user time used in the voluntary control approach [2].

1.5. *Structure of the BCI.* The steps of the brain computer interface system include the following:

- (1) Brain activity measurement/recording methods of the BCI
- (2) Preprocessing techniques
- (3) Feature extraction
- (4) Machine learning implementation/classification
- (5) Translation to control signal

1.5.1. *Preprocessing.* In BCI, preprocessing techniques consist of data acquisition of brain signals to check the signal quality without losing important information; the recorded signals are cleaned and conduct noise removal to acquire relevant information encoded in the signal. As mentioned above, the EEG signals are of poor quality; even the commercial 50 Hz frequency, due to nearby appliances, can corrupt the EEG signals, and the users are also advised not to think anything else apart from the stimuli as presented. In preprocessing, using Fourier transform or Fourier series, the signals are taken into the frequency domain and studied what

frequency content is present in the signal. The undesired 60 Hz frequency signal and undesired signal produced by performing actions other than the said stimuli are then filtered out using a notch filter as mentioned in Figure 3.

1.5.2. *Feature Extraction.* Feature extraction plays a vital role in brain-computer interface applications; the raw EEG signals are nonstationary signals that are corrupted by noise or due to artifacts present in the environment where they are being recorded, but still meaningful information can be extracted from them. The data dimensionality is also reduced to process it better, and machine learning models are applied. This method is essential to increase the classification accuracy of the BCI system.

EEG signal is a time-domain nonstationary signal, and the relevant information such as signal energy is analyzed as a function of time or frequency later; relevant statistic measures are adapted to properly explain the characteristics of the signal.

Some of the commonly used feature extraction techniques are listed as follows.

Short-time Fourier transform is a frequency-domain feature extraction technique in which the EEG signal is convolved with a window function  $w$  to extract the relevant frequency features of the brain which are broken down as sinusoids at different frequency ranges.

Mathematically, it is represented as

$$Xstft[m, n] = \sum_{k=0}^{L-1} x[k]w[n-k]e^{-j2\pi mk/L}, \quad (1)$$

where  $x[k]$  is the input EEG and  $w[n-k]$  is the window multiplied to extract the frequency features as shown in the figure [4].

Discrete-time and continuous-time wavelet transform is a time frequency-based feature extraction technique that allows better temporal and spatial resolution in which the EEG signals are produced in the form of wavelets at different frequency ranges of interest as shown in Figure 4.

The technique of wavelet transforms as adapted in the literature is a filtering continuous application of high-pass and low-pass filters to extract the wavelets of the signal which, when added together, constitute the original signal and downsampled by a factor of 2 as shown in Figure 5.

$g$  and  $h$  are consecutive high-pass and low-pass filters which produce the relevant detailed and approximated coefficients of the EEG signals [5].

1.5.3. *Neural Networks.* Before starting to explain what deep learning is, it is first beneficial to explain the role of deep learning and its fundamental blocks.

Deep learning is a classification tool used in a variety of daily applications which is composed of speech recognition and computer vision to natural language processing in the context of the BCI; the input features which are different brain frequency bands are classified according to what activity the user is performing at the moment.

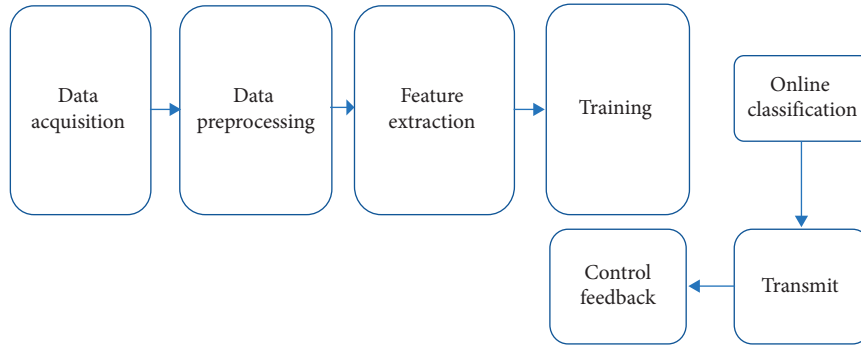


FIGURE 3: BCI block diagram.

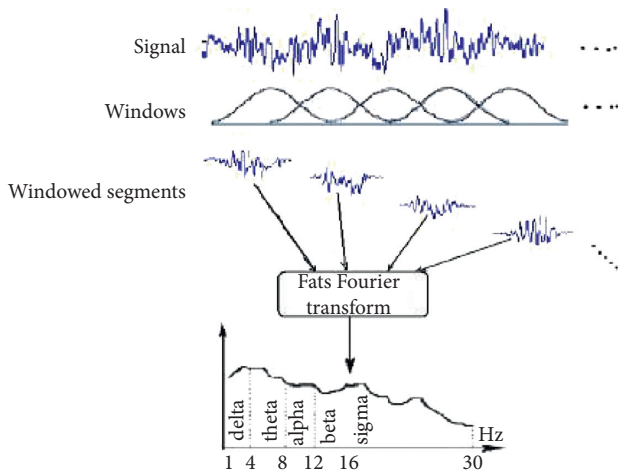


FIGURE 4: Feature extraction using short-time Fourier transform.

*Neural Network.* A neural network is a model similar to that of a neuron in our brain that has input nodes and output nodes; the mathematical model for a neural network is given by the following equation:

$$v = wx + b, \tag{2}$$

where  $v$  is the weighted sum of the inputs and the bias term which will be fed at the output node,  $b$  is a bias term which is mostly set to 1, and  $w$  is the random weights assigned that are multiplied with the input in order to reach closer to the desire output.

The neural network is shown in Figure 6.

These calculations are often preceded in the form of matrices; the input, the weight terms, the output, and the bias are as follows:

$$v = (w_1 \times x_1) + (w_2 \times x_2) + (w_3 \times x_3) + b. \tag{3}$$

Finally, the output node is passed to an activation function and provides the final output; the activation function calculates the characteristic of the node. The activation function acts to map the corresponding inputs to the right output  $y$  present at the output node:

$$y = \varphi(v). \tag{4}$$

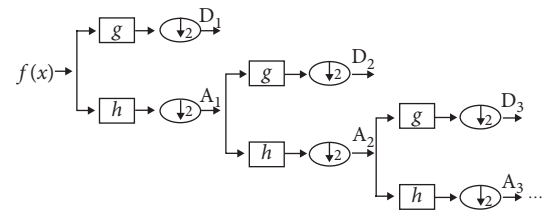


FIGURE 5: Wavelet transform.

The neural network does not get its right output at the first attempt. It needs to be trained a lot, and so a training rule is assigned to neural networks to get the right output. Many training rules are adapted, but one of the most commonly used is the delta rule, and the rule is expressed using the following equation:

$$w_{ij} \leftarrow w_{ij} + \alpha e_i x_j, \tag{5}$$

where  $x_j$  represents the number of inputs,  $e_i$  is the error generated at the output node, and  $\alpha$  is the learning rule between  $(0 < \alpha < 1)$ .

The training rule is summarized as follows:

- (1) Assign adequate values to the weights.
- (2) Obtain the input from the training data and feed it into the neural network which will give an output  $d$ ; subtract the output  $d$  to obtain the correct output at the output node.

$$e_i = d_i - y_i. \tag{6}$$

- (3) Calculate the weight updates:

$$\Delta w_{ij} = \alpha e_i x_j. \tag{7}$$

- (4) Adjust the weights accordingly until the correct output or that has small tolerance is obtained:

$$w_{ij} \leftarrow w_{ij} + \Delta w_{ij}. \tag{8}$$

The above explanation was presented for a single-layer neural network; the architecture of neural networks is

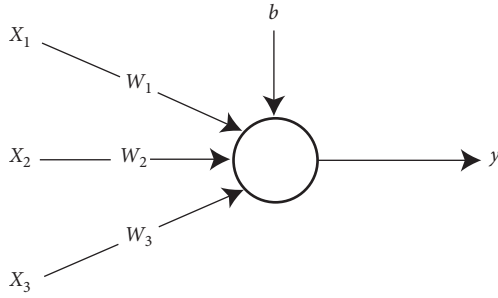


FIGURE 6: Neural network.

becoming better with the cost of greater memory, but with higher classification accuracy, we use deep neural networks which are the same as the single-layer neural network but with hidden layers added in between the input and output nodes, as shown in Figure 7.

The concepts are similar to those of a single neural network but with the adjustments of added hidden layers and a different training rule because the delta rule has a drawback of not propagating the output to the hidden layers, thereby the weights are not adjusted.

To explain how the deep neural network works, the above explained single neural network is set as basis.

In Figure 8, given a multiple-layered neural network, the weighted sum obtained at the first hidden layer is presented as

$$\begin{bmatrix} v_1^{(1)} \\ v_2^{(1)} \end{bmatrix} = \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (9)$$

$$\triangleq W_1 x.$$

The outputs are calculated via the sigmoid activation function:

$$\begin{bmatrix} y_1^{(1)} \\ y_2^{(1)} \end{bmatrix} = \begin{bmatrix} \varphi(v_1^{(1)}) \\ \varphi(v_2^{(1)}) \end{bmatrix}. \quad (10)$$

The process is repeated, and the outputs obtained are treated as the inputs to the other nodes, and we get the outputs as

$$\begin{bmatrix} v_1^{(1)} \\ v_2^{(1)} \end{bmatrix} = \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{bmatrix} \begin{bmatrix} y_1^{(1)} \\ y_2^{(1)} \end{bmatrix} \quad (11)$$

$$\triangleq W_1 y^{(1)}.$$

And lastly, the weighted sum is being inserted into the activation function, and we return our final output:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \varphi(v_1) \\ \varphi(v_2) \end{bmatrix}. \quad (12)$$

Deep learning training rule is given in the following.

Backpropagation algorithm is commonly used as the training instruction for the deep neural networks; the procedure is summarized as follows:

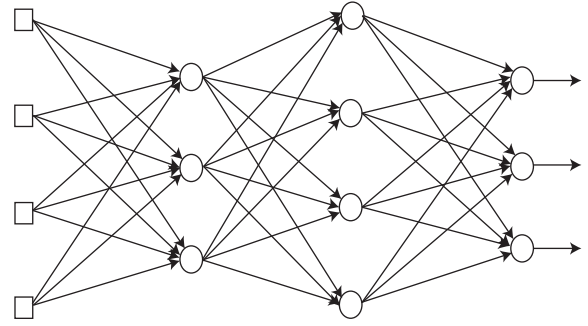


FIGURE 7: Structure of the deep neural network.

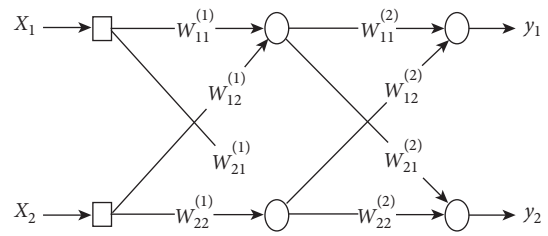


FIGURE 8: Multilayered network.

- (1) Assign adequate values to the weights.
- (2) Take the input from the training data and feed it into the neural network which will give an output  $d$ . Subtract the output  $d$  to obtain the correct output at the output node and the delta ( $\delta$ ) of the output nodes:

$$\begin{aligned} e &= d - y, \\ \delta &= \varphi'(v)e. \end{aligned} \quad (13)$$

- (3) Propagate the delta back towards the hidden nodes, and determine respective delta  $\delta$  of nodes:

$$\begin{aligned} e^k &= W^T \delta, \\ \delta^{(k)} &= \varphi'(v^{(k)})e^{(k)}. \end{aligned} \quad (14)$$

- (4) Repeat until it reaches the input nodes.
- (5) Modify the weights according to the rule:

$$\begin{aligned} \Delta w_{ij} &= \alpha \delta_i x_j, \\ w_{ij} &\leftarrow w_{ij} + \Delta w_{ij}. \end{aligned} \quad (15)$$

- (6) These steps are repeated until the neural network is utterly trained as shown in Figure 6.

Now, the alpha-beta ranges are extracted, and consecutive energies are calculated. The features are fed into the deep learning neural network, and using the backpropagation learning rule, the model is trained, yielding an accuracy of 97.83%, as shown in Figure 9 [5].

## 2. Critical Review of the Related Literature

A brain-computer interface involves various stages, and development in each stage leads to an improved and efficient system. Here, the literature review of major steps including data acquisition, feature extraction, classification algorithm, and applications is presented.

- (1) Improvement of EEG signal acquisition: An electrical aspect for state of the art of front end. Computational intelligence and neuroscience: a research paper published by Ali Bulent Usakli, the NCO Academy, Turkey, presented some applicable concerns for acquiring quality EEG signals which are proven helpful for users and design engineers. One of the most important considerations is selecting suitable electrodes and headset. In the EEG-based BCI, electrodes, signal processing components including mental and environmental conditions, filtration of noise, amplification, signal translation, and data storage affect the recording process. The data acquisition is an important step in any machine learning procedure. Brain signals need to be cleaned and preprocessed so that a good result can be obtained [1].
- (2) P300 wave detection using Emotiv EPOC+ headset: effects of matrix size, flash duration, and colors: Saleh Ibrahim Alzahrani conducted research on P300 wave detection using the Emotive EPOC+ headset to study the effects of the size of the matrix, flash duration, and colors. In this study, P300 signals were obtained from five subjects with Emotive EPOC+ using all 14 channels. For this research, EEG signals obtained were communicated to software OpenViBE through a USB dongle. A sample was taken every 8 seconds at a rate of 128 samples per second. The configured sampling rate provides ample samples for the four frequency bands, containing significant ERP data. During process of signal recording, the subjects were shown a matrix of  $6 \times 6$  or  $3 \times 3$  on the computer screen. They were instructed to stay calm, avoid any needless movement, and set all on letter which they desire to spell. It is reported in the study that one of the advantages of using the Emotiv EPOC+ headset is that it takes merely two to three minutes for preparation as compared to other headsets that take more than ten minutes. The quality of sensors is verified through Emotiv Xavier SDK which provides feedback report of each sensor. To decrease the contact impedance, saline liquid was applied to the sensor surface. Primary objective was to assess the potential of Emotiv EPOC+ to perceive P300 signals. Finding the electrodes proficient at providing target signals helps minimize the number of channels which makes signal processing a lot easier. The results of this experiment provide evidence of the capability of Emotiv EPOC+ to detect the P300 signals from two channels, O1 and O2 [2].
- (3) Automatic seizure detection in SEEG using high frequency activities in wavelet domain. Medical engineering and physics: in this research paper, the researcher has found the method for detection of seizures using the high-frequency analysis in the wavelet domain. This method is used highly in the high-frequency domain. Because of seizure detection, the method is usually done using high frequency in the range of 80–250 Hz. Also, it can handle fast ripples in the range of 250 to 500 Hz. The biggest advantage in the seizure detection is that it can detect the seizure offset. The methodology consists of the Continuous Wavelet Transform (CWT), which was computed by convolving the SEEG signal which has to make the feature extraction, and it also includes the complex conjugate of the wavelet basis function (Ayoubian, 2013).
- (4) Classification of epilepsy EEG signals using DWT-based envelope analysis and neural network ensemble: envelope detection is a very efficient method for detecting the impact of the signals which are based on the biological change or diagnosis. In this paper, the researchers used the Hilbert transform which has a good impact on the resultant signals so that the signals are then widespread using the DWH technique which has a unique behavior regarding the biological change; the researchers detected the changes using this method [5].
- (5) Feature selection for motor imagery EEG classification based on firefly algorithm and learning automata: in this research paper, the researchers implemented spectral linear discriminant analysis for the classification of motor imagery signals. Feature extraction method used was basically common spatial patterns; the advantage of using this feature extraction method is basically of two-class discrimination problems. This maximizes the variance of one class and decreases the variance of the other class, which is the advantage, but the disadvantage is because of the multiclass overlap structure in this method, it is not used for multiclass prediction [6].
- (6) Unsupervised adaptation of electroencephalogram signal processing based on fuzzy C-means algorithm. International Journal of Adaptive Control and Signal Processing: this research paper presented the techniques of brain mapping with emphasis on multichannel EEG and functional brain imaging techniques. During training and testing, the concentration of the subject on the target object is one of the concerns which signifies the capacity to operate a device. They have used different algorithms such as LDA and fuzzy C-means. Fuzzy C-means is an adaptive classifier, and this is probably

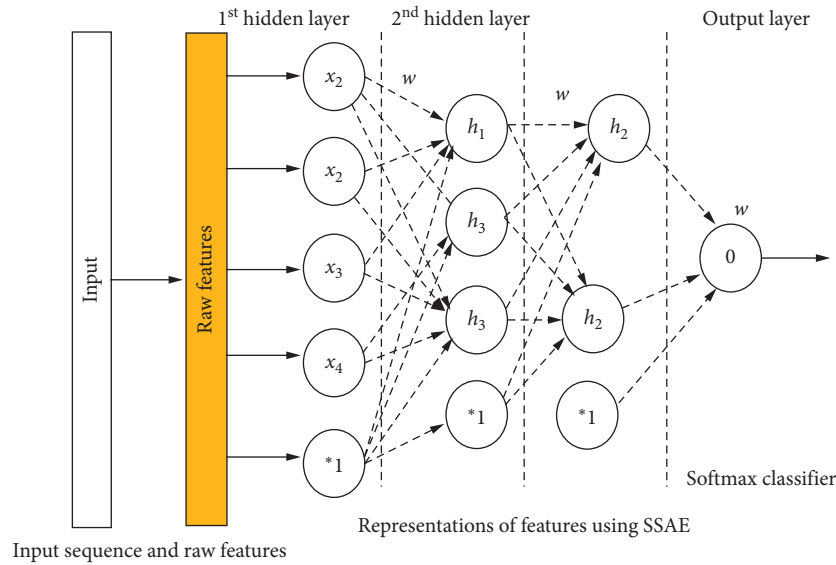


FIGURE 9: General diagram of neural networks in the BCI.

used where the device behavior is not synchronized with the classification model [7].

- (7) A review of classification algorithms for EEG-based brain-computer interfaces: a 10-year update: this paper is the latest review of the BCI classification techniques; there are some algorithms discussed in this paper taken from different papers, and their accuracies, optimization, the method of feature extraction used were compared, and every algorithm has its own advantages and disadvantages. In this paper, they have used the event-related potentials. Feature extraction method they have used is based on spatial filtering which is the most optimized filter; it can further be optimized by using the calibration [8].
- (8) Sequential non-stationary dynamic classification with sparse feedback: in this research paper, basically, they have discussed the technique for the classification of the nonstationary and nonlinear signals. As we know that the BCI signals are nonstationary in nature, sparse feedback can be used for the stability of the brain signals and classifying them using the RBF classifiers which are radial basis functions. The signals are acquired in a nonlinear manner, and we can apply linear models to them, but again, multiclass prediction is not able to be performed. This is because of the sparse feedback matrices involved [9].
- (9) Motor imagery and direct brain-computer communication: Gert Pfurtscheller and Neuper researched about the technique for the motor imaginary signals by the imagination of the left hand, right hand, and foot movements. In the neurofeedback method, real-time prediction of brain signals is difficult to achieve. We have nonlinear signals at the input of the neurofeedback

method so that we can use the Hidden Markov Method (HMM) to make predictions in real time but the accuracy is a tradeoff [10].

- (10) Toward unsupervised adaptation of LDA for brain-computer interfaces. IEEE Transactions on Biomedical Engineering; the firefly algorithm (FA) is an efficient algorithm for selecting the most appropriate subset of features and helps improving accuracy of classification. When the problem of entrapping of FA in the local optimum arises, a procedure for combining the firefly algorithm and learning automata (LA) is proposed which optimizes feature selection for motor imagery EEG. For the expected outcome of the high-dimensional feature set, a process of combining the common spatial pattern (CSP) and local characteristic-scale decomposition (LCD) algorithms is used. It is further classified by the use of the spectral regression discriminant analysis (SRDA) classifier [11].

### 3. Comparison of Classification Algorithms

Table 2 shows the comparison of classification algorithms.

### 4. Discussion

There is a large range of classifiers developed by scientists and engineers around the world. These classification algorithms can be divided into four groups.

**4.1. Adaptive Classifiers.** Adaptive classifiers are listed as those classifiers in which parameters are progressively recalculated and also updated with procurement of new EEG data signals which are nonstationary, and adaptive classifiers are capable to follow the change in the feature distribution.

TABLE 2: Comparison of classification algorithms.

Title	Algorithm	Input features	Efficiency	Advantages	Drawbacks/tradeoff
Vidaurre et al. [12]. Toward unsupervised adaptation of LDA for brain-computer interfaces. IEEE transactions on biomedical engineering, 587–597.	Linear discriminant analysis uses hyperplanes for different classes, assuming normal distribution, with equal covariance matrix for both classes; to solve an NC class problem, several hyperplanes are used.	Separating hyperplane is obtained by seeking the projection that maximizes the distance between two classes' means and minimizes the interclass variance.	Suitable for online BCI and provides generally good result, and fluctuations in the training data set do not affect much.	Very low computational requirement so suitable for online BCI system.	Linearity that can provide poor results on complex nonlinear EEG data (not immune to noise).
Li et al. [5]. Classification of epilepsy EEG signals using DWT-based envelope analysis and neural network ensemble. Biomedical signal processing and control, 357–365.	Support vector machine uses a support vector hyperplane to identify classes.	Selected hyperplane is the one that maximizes margins, i.e., the distance from nearest training points.	Enables classification using linear decision boundaries, (linear SVM) has been applied, generalization capabilities, to be insensitive to overtraining and to the curse of dimensionality.	Maximizing margins and regularization are known to increase the accuracy.	These advantages are gained at the expense of a low speed of execution.
Lotte et al. [4]. A review of classification algorithms for EEG-based brain-computer interfaces: a 10-year update. Journal of neural engineering, 031005.	Neural network consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function which utilizes a supervised learning technique called backpropagation for training.	Neurons of the output layer determine the class of the input feature vector.	Applied to almost all BCI applications.	Because universal approximators are composed of enough neurons and hidden layers, they can approximate and classify any continuous signal.	Sensitive to overtraining, especially with such noisy and nonstationary data as EEG; therefore, careful architecture selection and regularization is required.
Usakli [1]. Improvement of EEG signal acquisition: An electrical aspect for state of the art of front end. Computational intelligence and neuroscience.	Consists of the noninvasive technique for recording brain signals which is based on the electromagnetic resonance signals compared to that of the EEG scalp signals.	Brain signals as an electrical pulse coming from the brain.	Suitable for all BCI systems. They are based on the noninvasive technique in real-time monitoring of signals.	Proven helpful for users and design engineers. One of the most important considerations is selecting suitable electrodes and headset.	Costly design because of the gold electrodes. They are costly, and everyone cannot afford that system.
Alzahrani [2]. P300 wave detection using Emotive Epoc+ headset: Effects of matrix size, flash duration,	Emotive Epoc+ 14 channel sensor was used which has 14 channels for EEG and a neutral channel as well.	Input features consist of P300 steady-state evoked potentials.	Suitable for the brain signals which are collected by the Emotive Epoc+ sensor. The signals are carried out using Emotive Epoc.	Advantage is basically being optimized, and the device is very cheap in price with affordable accuracy.	For more number of class predictions, the accuracy becomes low, and the output is affected.

TABLE 2: Continued.

Title	Algorithm	Input features	Efficiency	Advantages	Drawbacks/tradeoff
Ayoubian, L. a. (2013). Automatic seizure detection in SEEG using high frequency activities in wavelet domain. Medical engineering and physics, 35, 319–328.	Based on the continuous wavelet transform, the brain signals were computed by convolving the SEEG signal.	Brain signals were collected from the Stellate Harmonic system for EEG monitoring purpose. These signals were passed through a band-pass filter.	Suitable for the detection of the seizure. A seizure onset is added to the signals and then compared with the normal brain signal.	For automatic seizure detection, it is very useful, and it can be used for the patients who are not able to calculate when they have seizure.	The disadvantage is basically for some high-frequency seizures. The high-frequency seizures are not detected easily because of the band-pass filter.
Liu et al. [6]. Feature selection for motor imagery EEG classification based on firefly algorithm and learning automata. Sensors.	Spectral regression discriminant analysis (SRDA) is widely used in the feature classification; in this paper, they have implemented this algorithm.	Separating hyperplane is obtained by seeking the projection that maximizes distance between two classes' means and minimizes the interclass variance.	Suitable for online BCI and provides generally good result, and fluctuations in the training data set do not affect much.	Very low computational requirement so suitable for online BCI system, simple to use, and generally provides good results.	Linearity that can provide poor results on complex nonlinear EEG data (not immune to noise).
Lowne et al. [9]. Sequential non-stationary dynamic classification with sparse feedback. Pattern recognition, 897--905.	Spectral regression discriminant analysis (SRDA) is widely used in the feature classification; in this paper, they have implemented this algorithm.	Separating hyperplane is obtained by seeking the projection that maximizes distance between two classes' means and minimizes the interclass variance.	Suitable for online BCI and provides generally good result, and fluctuations in the training data set do not affect much.	Very low computational requirement so suitable for online BCI system, simple to use, and generally provides good results.	Linearity that can provide poor results on complex nonlinear EEG data (not immune to noise).
Liu et al. [6]. Unsupervised adaptation of electroencephalogram signal processing based on fuzzy C-means algorithm. International journal of adaptive control and signal processing.	Common spectral patterns were used for the feature extraction and the linear discriminant analysis, and fuzzy C-means was used for the feature classification.	The maximum distance was calculated for each fuzzy C-means, and then the mean was calculated; after that, the features were classified.	They are suitable for nonlinear EEG signals having different amplitudes for different people.	Very low computational requirement so suitable for online BCI system, simple to use, and generally provides good results.	Fuzzy behavior can be seen in the output when the frequency changes at the input.
Pfurtscheller and Neuper [10]. Motor imagery and direct brain-computer communication. Proceedings of the IEEE, 1123–1134.	Hidden Markov model was used for the classification of EEG signals as they are nonlinear in nature, so we can tune the Markov model accordingly.	The input signals were obtained from the two channels, and these signals were transformed into the HMM network.	For two channels, EEG signals, this is good, and it has a fast classification.	The method used in this paper uses low computational power, and the model functions are optimized.	Output accuracy depends on the linear behavior of the signals. When the frequency fluctuates the output, control will also change.

As mentioned in [10], a model for a motor-imagery-based self-paced BCI structure for operating a robot was proposed. They used a basic synchronous BCI, devised earlier for recording data for offline training classification before conducting the online self-paced procedure. They extracted logarithmic band power as features, and features were extracted from EEG signals. Feature selection

was manual so as to gain quality frequency bands. To extract the features, chosen frequency bands were digitally bandpass-filtered, squared, and averaged over 1 second sliding window, and natural log was applied. Utilizing the features and associated labels, two linear discriminant analysis (LDA) classifiers were trained, with one to recognize left imagery from right and the other to



isolate right imagery movement development from others.

Features related to subject's control brain signals are extracted and constantly classified by the offline LDA algorithm. These features are then used to control the robot. They used parameters of the LDA algorithm in place of accommodating threshold and dwell features. Subject's control intention and timing is the basis for adaptation for online training. The methodology proposed entailed information about the user's control needed to train and adapt which could show promise of improving the accuracy and performance. The paper used Kalman filters for online parameters to be classified using LDA. Event label assignment was introduced with a slower learning process [12].

One of the advantages of implementing adaptive classifiers is that they can employ both supervised and unsupervised. This means that even if there is no information of true labels of data being received, they can be executed. From multiple research studies, it is inferred that unsupervised learning has yielded better results than static classifiers.

Now, most of the real-world applications of BCI do not present class labels, and for this purpose, unsupervised adaption classifiers require more development [7].

*4.2. Matrix and Tensor Classifiers.* The approach which the researchers have used in this paper holds fewer stages for classification as compared to the classical machine learning algorithms, and they are simpler as well. Compared to other standard classifiers, Riemannian classifier does not need any parameter-tuning techniques such as cross validation to properly train and verify the accuracy of the produced model, which makes it far more robust and accurate all due to its logarithmic tendencies. Likewise, the inborn Riemannian separation for the SPD matrix is invariant both to inversion of the matrix and to any direct invertible change of the information, for example, any outside interference added to the EEG sources does not change the separations among the witnessed covariance matrices. These properties partially clarify why Riemannian classification techniques give a decent speculation ability, which empowered analysts to set up adjustment-free versatile ERP-BCIs utilizing basic subject-to-subject and session-to-session exchange learning methodologies [9].

It is shown that several approaches were implemented and gave higher performance than CSP + LDA procedures on motor imagery EEG data. The biggest advantage is quality performance. However, this is a gain at tradeoff between performance and greater number of weights because of elevated expansion in input feature dimensionality which makes suitable regularization a much-needed step [12].

EEG data can be represented in the form of tensors and are treated as analysis tools for EEG data tools for EEG data analysis which includes feature extraction, data clustering, and data classification in the BCI. EEG data are represented in more than one dimension; this includes time, frequency space, and trails and hence, these are presented by the tensor order. EEG data that have time frequency and space can be

represented by 3-dimensional tensor. Tensors have been used frequently for motor imagery-based analysis even with a large amount of data containing different categories which can be represented by the tensor and its order [10].

*4.3. Transfer Learning and Deep Learning.* Transfer learning is a crucial tool when it comes to session-to-session and subject-to-subject decoding performance. If transfer learning is improved enough, BCI system can be operated without calibration, and this will revolutionize the BCI systems forever.

It is observed that calibration sessions are strenuous and mentally exhausting for subjects. It is explained that accepting the input from the earliest starting point of their BCI system is promising for started subjects.

Deep learning is categorized as a ML algorithm, where features and the classifier are collectively learned straight from EEG data. There exist multiple deep learning algorithms. One of the most explored and commonly used is deep neural networks (DNNs). DNN is also performed online for slow cortical potentials (SCP) and motion-onset visual evoked potential (MVEP) which are not commonly used. The very first research conducted and paper was published on the P300-based BCI by Ciotte et al. Two convolutional layers were constructed followed by completely connected layer. One convolutional layer has a purpose to learn spatial filters and the second one was to learn temporal filters. The model was proven better than the P300 experiment, but the SVM model had more accuracy [2].

Deep extreme learning machine is used for Slow Cortical Potentials (SCP) classifications. This technique consists of multiple layers, and the last one was Kernel ELM. However, in this project, number of units, network structure, hyperparameter, and input features were not reasoned. This did not prove to be better than multiplayer ELM or standard ELM [4].

*4.4. Miscellaneous Classifiers.* In order to classify more than two mental tasks, two main approaches can be used to obtain a multiclass classification function. The first approach consists in directly estimating the class using multiclass techniques such as decision trees, multilayer perceptron, naive Bayes classifiers, or  $k$ -nearest neighbors. The second approach consists of decomposing the problem into several binary classification problems.

Multiclass and multilabel approaches therefore aim to recognize more than two commands. It is therefore necessary to choose carefully the mapping between mental commands and corresponding labels. However, the errors may be possible during the classification. In particular, the set of estimated labels, sometimes, may not correspond to any class, and several classes may be at equal distances, thus causing class confusion [13].

## 5. Methodology

Here are some methods which are discussed in the research papers for the past few years in brain-computer interface systems, as shown in Table 3.

TABLE 3: Summary of various methodologies in BCI systems.

Classification methods	Input EEG pattern	Features	References
Adaptive classifiers	Motor imaginary-based	Frequency band power, EEG time	[10, 12, 14]
Matrix and tensor classifiers	Steady-state visual evoked potentials, P300	Frequency band power, raw data	[2, 5, 7, 15]
Transfer learning and deep learning	Motor imaginary-based, P300, SSVEP	EEG time points, frequency band power, raw EEG data	[2, 4, 6, 7]
Miscellaneous classifiers	Motor imaginary-based, P300	Not specified	[1, 4, 15]

## 6. Conclusion

During this course of work, a question arises whether it is possible to create a brain computer interface which is affordable, with high accuracy and optimization. So, after reviewing different papers, the conclusion is that if we need an optimized model with high accuracy for the noninvasive technique of brain signals, the artificial neural network has a high accuracy and is optimal. However, there are some tradeoffs as well that are model compatibility with the brain signals. From 10 years of BCI review, we have obtained that the ANN has a high response and accuracy; after all, it optimizes the system as well. However, further research studies have been done to make it more accurate because this has to be used in health care.

Due to the fast processing that ANN allows, a form of guidance could be provided during training that enables a user to improve the classifier's performance and let it reflect his/her intents more often. This guidance was very useful for one of the three subjects.

Also, the statistical test was examined on whether it performs in a way that can be expected. Therefore, after the offline classification, it yielded an accuracy above 90 percent.

The control provided by the developed system has been sufficient to conclude that it provides enough control that a user can command an arbitrary computerized device. Also, it showed to be easily trainable.

In the future, the proposed model can provide support on multiplatforms. This can be achieved by developing applications which can help humanity and make everyday tasks easier. Furthermore, the system can be controlled with a smartphone that can override EEG headset commands. This will act as fail-safe if the BCI system experiences any malfunctioning. On the basis of this development, better EEG can be designed with higher efficiency and which is less dependent on offline classification.

## Data Availability

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

## Conflicts of Interest

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

## References

- [1] A. B. Usakli, "Improvement of EEG signal acquisition: an electrical aspect for state of the art of front end," *Computational Intelligence and Neuroscience*, vol. 2010, Article ID 630649, p. 12, 2010.
- [2] S. I. Alzahrani, "P300 Wave Detection Using Emotiv EPOC+ Headset: Effects of Matrix Size, Flash Duration, and Colors," Colorado State University, Fort Collins, CO, USA, Doctoral dissertation, 2016.
- [3] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in brain-computer interfaces: a review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1753–1762, 2016.
- [4] F. Lotte, L. Bougrain, A. Cichocki et al., "A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, Article ID 031005, 2018.
- [5] M. Li, W. Chen, and T. Zhang, "Classification of epilepsy EEG signals using DWT-based envelope analysis and neural network ensemble," *Biomedical Signal Processing and Control*, vol. 31, pp. 357–365, 2017.
- [6] A. Liu, K. Chen, Q. Liu, Q. Ai, Y. Xie, and A. Chen, "Feature selection for motor imagery EEG classification based on firefly algorithm and learning automata," *Sensors*, vol. 17, no. 11, p. 2576, 2017.
- [7] G. Liu, D. Zhang, J. Meng, G. Huang, and X. Zhu, "Unsupervised adaptation of electroencephalogram signal processing based on fuzzy C-means algorithm," *International Journal of Adaptive Control and Signal Processing*, vol. 26, no. 6, pp. 482–495, 2012.
- [8] A. Andreev, A. Barachant, F. Lotte, and M. Congedo, *Recreational Applications of OpenViBE: Brain Invaders and Use-The-Force*, Wiley Online Library, Hoboken, NY, USA, 2016.
- [9] D. R. Lowne, S. J. Roberts, and R. Garnett, "Sequential non-stationary dynamic classification with sparse feedback," *Pattern Recognition*, vol. 43, no. 3, pp. 897–905, 2010.
- [10] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.
- [11] A. Schlögl, C. Vidaurre, and K. R. Müller, *Adaptive Methods in BCI Research-an Introductory Tutorial*. In *Brain-Computer Interfaces*, Springer, Berlin, Heidelberg, Germany, 2009.
- [12] C. Vidaurre, M. Kawanabe, P. von Büna, B. Blankertz, and K. R. Müller, "Toward unsupervised adaptation of LDA for brain-computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 587–597, 2010.
- [13] A. B. R. Suleiman and T. A. H. Fatehi, *Features Extraction Techniques of EEG Signal for BCI Applications*, University of Mosul, Mosul, Iraq, 2007.
- [14] I. A. Fouad and F. E. Z. M. Labib, "Using emotiv EPOC neuroheadset to acquire data in brain-computer interface," *International Journal*, vol. 3, no. 11, pp. 1012–1017, 2015.
- [15] G. A.-R. Dornhege, *Toward Brain-Computer Interfacing*, MIT Press, Cambridge, MA, USA, 2007.

## Research Article

# Towards a Complete Set of Gym Exercises Detection Using Smartphone Sensors

Usman Ali Khan,<sup>1</sup> Iftikhar Ahmed Khan ,<sup>1</sup> Ahmad Din,<sup>1</sup> Waqas Jadoon,<sup>1</sup> Rab Nawaz Jadoon,<sup>1</sup> Muhammad Amir Khan ,<sup>2</sup> Fiaz Gul Khan,<sup>1</sup> and Abdul Nasir Khan<sup>1</sup>

<sup>1</sup>Department of Computer Science, COMSATS University Islamabad, Abbottabad Campus, TOBE CAMP, Abbottabad 22060, Pakistan

<sup>2</sup>Department of Electrical and Computer Engineering, COMSATS University Islamabad, Abbottabad Campus 22060, Pakistan

Correspondence should be addressed to Iftikhar Ahmed Khan; [iftikharahmed@cuiatd.edu.pk](mailto:iftikharahmed@cuiatd.edu.pk)

Received 23 December 2019; Accepted 20 February 2020; Published 22 July 2020

Guest Editor: Iván García-Magariño

Copyright © 2020 Usman Ali 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.

Smartphones with gym exercises predictors can act as trainers for the gym-goers. However, various available solutions do not have the complete set of most practiced exercises. Therefore, in this research, a complete set of most practiced 26 exercises was identified from the literature. Among the exercises, 14 were unique and 12 were common to the existing literature. Furthermore, finding suitable smartphone attachment position(s) and the number of sensors to predict exercises with the highest possible accuracy were also the objectives of the research. Besides, this study considered the most number of participants (20) as compared to the existing literature (maximum 10). The results indicate three key lessons: (a) the most suitable classifier to predict a class (exercise) from the sensor-based data was found to be KNN (K-nearest neighbors); (b) the sensors placed at the three positions (arm, belly, and leg) could be more accurate than other positions for the gym exercises; and (c) accelerometer and gyroscope when combined can provide accurate classification up to 99.72% (using KNN as classifier at all 3 positions).

## 1. Introduction

The advancement in the technology troubled humans by making their lives busy. This is affecting their health negatively [1]. However, the technology also helps humans to improve their health, education, business, and social relationships [2]. The beneficial impact of technology is tremendous, especially in the health sector. Multiple hardware and software [3, 4] are used to improve overall human health. Among various sources of maintaining health, gyms are the major source of physical fitness.

People join gyms to achieve goals like bodybuilding, physical fitness, or losing weight. In the modern world, technology has replaced the traditional concepts of guidance and training to stay healthy and fit. The tools like smartphones and devices like wearable gadgets are among the many resources that are helping to stay healthy and fit [5–8]. There is also some researches like [9–11] that support the

notion that technology can help to achieve fitness objectives. Besides, there are various smartphone applications like [12–14] that can track different physical activities, e.g., walking, running, sitting, and standing with the corresponding calorie burn out. The sensors (accelerometer, gyroscope, etc.) are used to track the activities.

Many wearable devices and smartphone applications track physical activities and calorie burnout like [3, 4]. However, none of the studies provides the information appropriate to measure major gym activities. For example, research studies such as [9, 11, 15] targeted a group of upper body muscles along with some warm-up exercises only. This research is a similar attempt yet is different in many aspects. First, in this research, 14 exercises of different muscle groups (abdominal, upper body, and lower body) are added to move towards a complete solution.

Second, in most of the existing research, the position of the sensors or the devices was only at the arm. Seeger et al.

[16] used three sensors at the following positions: a single accelerometer at the wrist, a hand glove, and a sensor at torso position. We hypothesized that the use of accelerometer and gyroscope at three body positions (arm, leg, and belly) could enhance accuracy because of the dependence of various gym exercises on either of the positions individually or in combination.

The third aim is to determine the number of sensors required to detect an exercise accurately. At the hypothesized three positions, five sensors are used: at the arm and the leg, the accelerometer and gyroscope together, while on the belly, only the accelerometer. The single sensor at the belly will only be used to determine the laying ( $x$ -axis), standing ( $y$ -axis), or in-between position that is often used in the gym exercises (e.g., angle leg press). The contribution of these sensors towards the accuracy has been analyzed in this research as well.

The classification algorithms used to detect exercises in the related work such as [7, 9, 11, 15–17] are linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), K-nearest neighbor (KNN), Naïve Bayes (NB), support vector machine (SVM), and dynamic time wrapping (DTW) algorithms. The result of the accuracies achieved by the studies was promising. However, their used datasets were quite sparse (collected from 8 to 10 persons) for 43 unique exercises. The exercises were also related to each other or were exercises from the same muscle groups having the same activity motion/patterns. The fourth aim of this study is to increase the number of participants in the real-world settings to bring more rigor to the findings. The increase in the number of participants and thus dataset could also affect the choice of the exercise detection algorithm. This forms the fifth aim, the selection of the most appropriate algorithm(s) to detect gym exercises.

The rest of the paper is organized as follows. Section 2 discusses the relevant literature in the context of the aims of this study. Section 3 is about the materials and methods. In Section 4, experimental setup is elaborated. In Section 5, the analysis and results are discussed. Section 6 concludes the paper as well as identifies some limitations. The section also embarks upon the possible future work.

## 2. Literature Review

In this section, related work is discussed in the context of the aims of this study. Therefore, this section is divided into the following three subsections: (1) exercise detection, (2) positioning and the number of sensors, and (3) exercise detection algorithms. The participant's selection is described in the Materials and Methods section.

*2.1. Exercise Selection.* The first activity recognition study based on wearable sensors device [18] was published in 2000. In this study, they attached two accelerometers inside the trousers' pocket to recognize daily life activities. The study [19] examines the use of a single smartphone accelerometer in activity recognition. The reported results showed accuracies between 80% and 97% depending on the set of

activities used and the processing techniques. Muehlbauer et al. [7] used the arm position to attach an Arm-Hostler with a fixed sensor to recognize a set of ten upper body gym exercises. They reported 93.6% accuracy in more than 90% of the cases they studied. MyHealthAssistant [16] classified the gym exercises using three accelerometers (on the hand glove, wrist, and torso). They trained a Bayesian classifier on the mean and variance features collected via an accelerometer. They collected the data of 11 exercises and achieved 92% accuracy. Chang et al. [8] used 2 accelerometers (on the hand and waist position) and examined a Hidden Markov Model (HMM) and a Bayes Classifier to identify exercises. They achieved 90% accuracy for the set of nine exercises and around 5% of the overall miss-count rate.

The activity recognition data collected from the literature corresponding from the years 2006–2018 found only 6 of 25 research studies related to gym exercises while the remaining 19 of 25 research papers were about daily life physical activities, emotional recognition, and elderly fall detection [20]. All the 25 papers were used to extract the information like the type of sensors used, features used to recognize activities, and the classification algorithms used.

*2.2. Position and Number of Sensors.* In most of the literature, only a single sensor for activity recognition is utilized. However, some studies used more than one sensor as well. For example, the authors in [21] put both an accelerometer and a gyroscope together and stated that the gyroscope adds nothing to the recognition results. However, some contradictory results are reported by the authors in [22]. The study [10] reported a 3.1 to 13.4 percent increase in recognition accuracy for 08 of 09 activities when an accelerometer is combined with a gyroscope while using the KNN classification algorithm. The average accuracy reported was 83.7% with an accelerometer and 90.2% with both accelerometer and gyroscope with an increase of 6.5% in average accuracy. The study also revealed that the sensor combination provides better results as compared to accelerometer alone. However, the paper does not report individual accuracies, thus resulting in an ambiguity whether the gyroscope or the accelerometer played a major role in the accuracies.

Table 1 provides the details of the number of sensors and device positions as per the literature while Table 2 describes the use of the combination of sensors (sensor fusion) as well as their accuracies.

From the analysis of Table 2, it can be argued that the combination of accelerometer and gyroscope provides the strongest accuracy results. Moreover, in most cases, a gyroscope does improve the recognition accuracy from 3.1% to 13.4% when used in combination with an accelerometer [10]. The magnetometer's role in activity recognition was poor.

*2.3. Exercise Detection Algorithms.* The related literature has also used different classification algorithms. For example, the authors of [21] used KNN combined with support vector machine (SVM) and the authors of [22] used KNN combined with decision tree and Naïve Bayes, while the authors

TABLE 1: Details of sensors/positions used.

S. no.	(Number)/position of smartphones/devices	Names of sensors used	Accuracy acquired (%)	Reference
1	(01)/arm	Accelerometer, gyroscope, magnetometer, electromyography EMG	75.70	[11]
2	(01)/arm	Accelerometer	89.00	[15]
3	(01)/arm	Accelerometer and gyroscope	93.00	[9]
4	(03)/01 wrist, 01 hand glove, and 01 torso	03 accelerometers only	92.00	[16]
5	(01)/arm	Accelerometer and gyroscope	93.60	[7]
6	(01)/arm	Accelerometer	90.00	[17]

TABLE 2: Detail of sensors.

Sensor name	Times used	Reason	Accuracies min., max., and avg.	References
Accelerometer	12	The accelerometer is the most powerful sensor in smartphones. It can be used for activity recognition by inferring the user's movements, such as walking, standing, running, sitting, and gym activities.	Min. accuracy = 82.2% Max. accuracy = 97.3% Avg. accuracy = 87.7%	[15, 23–33]
Accelerometer and gyroscope	08	Accelerometer and a gyroscope, to be used in recognizing physical activity and providing the strongest result.	Min. accuracy = 67.8% Max. accuracy = 97% Avg. accuracy = 88.3%	[10, 11, 31, 34–38]
Accelerometer, gyroscope, and magnetometer	05	Adding a magnetometer with an accelerometer and gyroscope. The results are not encouraging because magnetometer causes overfitting in training classifiers due to its dependence on directions.	Min. accuracy = 71.6% Max. accuracy = 96% Avg. accuracy = 82.6%	[10, 11, 34, 37, 39]

of [40] used J48 Decision Tree combined with Naive Bayes for exercise recognition. They reported an average accuracy of 95%, 90.2%, and 88%, respectively. Table 3 shows the top three classifiers Naïve Bayes (NB), decision trees (J48), and K-nearest neighbor (KNN) being abundantly used for the activity recognition purpose.

The accuracy of the results depends on the suitable selection of the classification algorithm as well as on the selection of the suitable parameters for them. Table 4 shows the top three most used features in the literature, that is, mean, standard deviation, and minimum and maximum as classification algorithm parameters.

### 3. Materials and Methods

In this section, the materials and methods used in the study are discussed. Section 3.1 discusses the selection of gym exercises for the current study. Section 3.2 is about the development of the application used for the data collection process. Section 4 is about experiment and data collection methods used in the study.

*3.1. Selection of the Exercises.* The process of the selection was started by collecting a list of all the gym exercises from the two sources [41, 42]. The sources listed a total of 74 gym exercises which will be called set TE (Total Exercises). To

verify the repeatability of the exercises in gyms, one of the authors visited four most known and commonly used gyms of the city to meet with the gym-goers. They were interviewed about the most common exercises the gym-goers trained on. The results were a subset of 54 most used gym exercises. The set will be called the set SE (Subexercises).

The set SE was compared with a set of the common exercises mentioned in the literature which resulted in (Common Exercises) set CE having 35 exercises. The exercises in CE were further categorized into exercises group along with information like exercise positions and equipment used to do the exercise.

Further analysis of the set CE revealed that five exercises were repeating in different muscle groups with different names. One of the exercises occurred three times and the rest of the four exercises twice in each muscle group. Removing the repetitions from the set CE resulted in a set of 29 exercises.

From the 29 exercises, 3 exercises were related to the head and are considered as warm-up exercises in the literature [43]. These 3 exercises are also removed from the list of 29 exercises reducing the final exercise set (Total Final Exercises) TFE to 26.

The exercises mentioned in the above paragraph were extracted from research papers like [7, 9, 11, 15–17]. Among these references, the study in [17] was related to only gym warm-up exercises and thus was not included in the exercise

TABLE 3: Details of classifier algorithms.

Algorithm	Times used	References
Naïve Bayes	5	[4, 5, 12, 14, 15]
Decision trees	5	[4, 9, 11, 12, 14]
K-nearest neighbor	3	[3, 4, 14]
Linear discriminant analysis	2	[21, 40]
Hidden Markov model (HMM)	2	[10, 15]
Quadratic discriminant analysis QDA	1	[40]
Logistic regression	2	[2, 9]
Support vector machine	1	[16]
Dynamic time warping (DTW)	1	[13]

selection. The remaining five papers were used to form 5 exercise sets (EP1–EP5). Here, E stands for exercise and P for the paper. Thus, EP1 represents exercise set extracted from paper 1, that is, reference [7] and so on. The union of the exercise sets EP1–EP5 was taken resulting in set (Total Exercises from Papers) TEP containing 43 exercises considered in the literature. The set TFE was subtracted from the set TEP to provide 14 unique exercises and 12 exercises that are considered in the literature (Table 5).

**3.2. Application Development.** To accomplish the objectives of this research, the first requirement was to develop an application to collect data from the participants. For the purpose, an android based smartphone application was developed. The users could add, view, edit, and delete personal profiles. The user interface of the developed application is depicted in Figure 1, whereas the flow of the user’s interaction with the application is elaborated in Figure 2. The application is also provided as a supplementary file with the paper for the researchers who want to replicate the research.

Figure 2 shows the overall process followed in the developed application for the data collection. The start screen provides options for the new users to register themselves, while already registered users can go to the registered users’ screen. After the selection of the new registration option, the user could move to the signup screen option. There they either can enter their profile information such as height and weight to register themselves with the new user profile or could go back to the main screen without registration. After clicking already registered users option, users could move to already registered users profile list to select their profile by name. The selected profile screen with information appears about the users from where they can start recording the exercise data and will also start doing the exercise. They can also view their stored records or could go back to the main screen. The users could exit the application from the main screen.

## 4. Methods

The developed application was installed on 3 smartphones and was positioned as shown in Figure 3. An LG Model F180 was attached to the leg while another similar model was

attached to the arm. This model supports both the accelerometer and gyroscope sensors providing the values of acceleration and rotation. For the belly position, we required only one sensor to determine the state (sitting, laying) of the participant. For the purpose Q-Mobile model, i7 was attached at the belly position having the support of only the accelerometer sensor.

The research also aimed to increase the number of participants and to collect varying data. Therefore, 20 participants with two sets of a total of 10 repetitions for each participant were used for the purpose. The 10 repetitions are used in the related literature before such as [11]. The data were collected against a selected set of 26 exercises. The smartphones were attached at three different body positions (arm, belly, and leg). All the gym-goers taking part in data collection were asked to behave normally as their usual exercising day. The sensors X, Y, and Z values were being recorded and stored in a file by the application while performing exercises. All the activities were carried out indoors in a gym.

**4.1. Experimental Setup.** The experimental setup section is divided into a further four subsections. Section 4.1.1 is about ethical compliance as per involvement and data collection of the participant. Section 4.1.2 explains the demographics of the participants. Section 4.1.3 is about the data collection process. The preparation of the data for the analysis is discussed in Section 4.1.4.

**4.1.1. Ethical Compliance.** The departmental ethics committee, called Project Research and Evaluation Committee (PREC), approved the study design and the procedure as defined in the above section. Informed consent for the study was obtained from the participants of this study.

**4.1.2. Participants.** For the selection of the participants, the busiest gym in the center of the city was selected. The gym-goers used to visit the gym regularly were approached and the aims and objectives of the data collection were explained to them. The 20 participants all males volunteered to participate in the data collection process. The participants were between the age brackets of 20 and 35. Their mean age was 25.85 years with SD of 4.13. Their heights ranged from 162 to 181 cm with a mean of 171.1 cm and SD of 5.34. Their weights ranged from 62 to 80 kg with a mean of 68.1 kg and SD of 5.56, respectively. The gym experience of the participant was between 2 and 19 months with a mean of 9.35 months and SD of 4.90. All the exercises were completed with free weights (participants choose weights themselves).

**4.1.3. Data Collection.** The data were recorded from 5 sensors (two sensors of the smartphone attached at the leg, two attached at the arm, and one attached at belly). All the sensors recorded X, Y, and Z values while the participant was doing the exercise. A triaxial accelerometer estimates the acceleration along X, Y, and Z axis and gyroscope (Pitch, Yaw, and Roll) helps the accelerometer to predict the

TABLE 4: Details of features.

Features	No. of papers in which used	References
Mean	15	[2-4, 6-9, 11, 14-16, 19, 21, 22, 40]
Standard deviation	11	[3, 4, 8, 9, 14, 15, 18, 19, 21, 22, 40]
Minimum/maximum	09	[3, 4, 13, 15, 16, 18, 32, 33, 40]

TABLE 5: The set of final 26 exercises considered in this paper.

S. no.	Repeating R/unique U	Exercise group	Exercise name	Exercise position	Equipment used
1	Unique exercises group	Shoulder	Face pull	Standing	Cable/rope
2		Shoulder	Cable front raise	Standing	Cable/rope
3		Biceps	Scott curl	Sitting bend	Dumbbell
4		Biceps	Smith machine drag curl	Sitting	Barbell
5		Triceps	Triceps with bar	Standing	Cable/rope
6		Triceps	Decline close grip bench press	Lying bench decline	Barbell
7		Chest	Standing cable cross	Standing	Cable/rope
8		Back	Wide grip pull up	Standing hanging	Fix rods
9		Back	T bar rows	Sitting hanging	Barbell
10		Back	Chin ups	Standing hanging	Fix rods
11	Common exercises group	Abs	Adjustable sit-up bench	Sitting and lying	Bench
12		Abs	Abs wheel	Lying	Abs wheel
13		Abs	Roman chairs	Lying	Roman chair
14		Abs	Flutter kick	Lying	Flat surface
15		Legs	Leg press	Lying	Leg press machine
16		Legs	Romanian deadlift	Standing	Barbell
17		Legs	Barbell squat	Sit-stand	Barbell
18		Shoulder	Incline press wide grip	Lying	Barbell
19		Shoulder	Standing barbell press	Standing	Barbell
20		Biceps	Inclined dumbbell curl	Sitting incline	Dumbbell
21		Biceps	Barbell preacher curl	Sitting	Barbell
22		Triceps	Triceps press with cable	Standing	Cable/rope
23		Chest	Machine bench press	Sitting	Bench machine
24		Chest	Dips	Standing hanging	Fix rods
25		Chest	Pec deck machine (butterfly)	Sitting	Pec deck machine
26		Shoulder	Seated barbell shoulder press	Sitting	Barbell

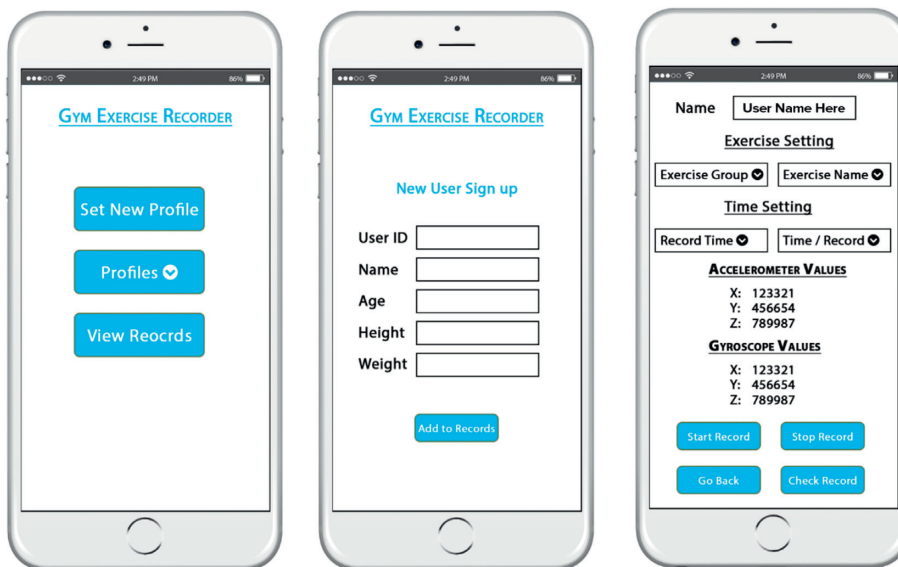


FIGURE 1: Application user interface.

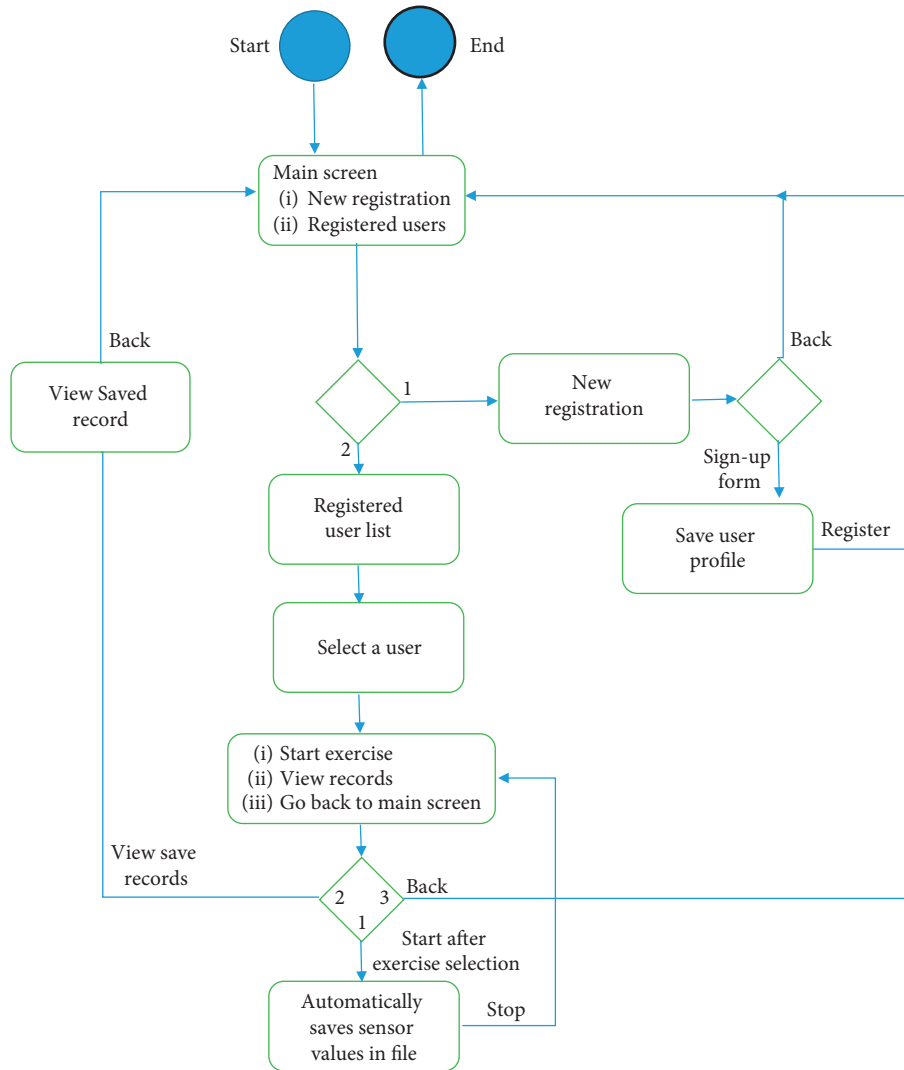


FIGURE 2: Application activity flow diagram.

orientation of the sensor. Three smartphones were synchronized to get the time from the server. The time was recorded up to millisecond along with X, Y, and Z values. This resulted in the 15 X, Y, and Z values along with a timestamp, the category of the exercise, and the exercise name. The dataset available from the literature [17, 44, 45] was not used because of the nonavailability of the data of 14 unique exercises. We also decided to collect data for the exercises whose data was available because of the probable setup differences between the existing studies and this study. This may have help in countering the bias and variations.

**4.1.4. Data Preparation.** The recognition process includes a collection of exercises data using multiple sensors. The data is preprocessed and segmented and the features are extracted and classified as the last step [11, 46]. The same process is followed in this research as well.

Three different files containing exercise data from each smartphone were combined carefully to match the participant's assigned ids and time stamps. In the second step, the

recorded data from CSV files were preprocessed to remove the extra noise. For example, at the start and the end of an exercise, the participant's movements were very random as well as jerky and were not aligned with the required exercise. Therefore, to remove this noise we removed the data from the first 3 seconds and the last 3 seconds of the recorded data of each exercise. For each exercise, there were 2 sets, each set of 10 repetitions and with an average participant time consumption for an exercise of 38 seconds. After pre-processing, we considered the data of 32 seconds only. The application was programmed to record 4 samples in a minute.

Various previous studies such as [11] used a 4-second window to extract required features and 1 minute of the slide to vary the data. We adopted the same strategy. The features extracted were based on the most used features (mean, standard deviation, and minimum and maximum) for the similar nature of the data as presented in Table 3. For each of the X, Y, and Z values, these four features were extracted forming a total of 60 features (36 accelerometer features and 24 gyroscope features).





FIGURE 3: Smartphone positioned on participant body.

TABLE 6: Table of classification results.

S number	Sensor name	Smartphone positions	Naïve Bayes classification accuracy in percentage (%)	K-NN classification accuracy in percentage (%)	Decision tree (J-48) classification accuracy in percentage (%)	Sensors features (mean, std. deviation, min, and max) A (x, y, z) G (x, y, z)
Single sensor used						
1	Accelerometer	Arm	52.61	95.87	87.73	12
2		Belly	40.99	95.16	86.29	12
3		Leg	47.59	96.39	91.58	12
4	Gyroscope	Arm	31.93	91.83	80.23	12
5		Leg	23.27	87.39	75.00	12
Two sensors used						
6	Accelerometers = 2	Arm and belly	69.92	98.40	92.96	24
7		Arm and leg	74.18	99.17	95.31	24
8		Belly and leg	63.10	98.65	93.85	24
9	Gyroscopes = 2	Arm and leg	42.52	96.29	86.40	24
10		Arm				
11	Accelerometer + gyroscope	62.49%	98.04	90.63	24	
		Leg	47.72	97.76	91.40	24
Three sensors used						
12	Accelerometers = 2, gyroscope = 1	Arm and belly	77.42	99.41	93.94	36
13	Accelerometer = 2, gyroscope = 1	Belly and leg	77.59	99.02	94.21	36
14	Accelerometers = 3	Arm, belly, and leg	79.59	99.51	89.0	36
Four sensors used						
15	Accelerometers = 2, gyroscopes = 2	Arm and leg	79.32	99.63	96.05	48
Five sensors used						
16	Accelerometers = 3, gyroscopes = 2	Arm, belly, and leg	80.72	99.72	96.29	60

TABLE 7: Table for best accuracy with the number of sensors used.

Sensors used	Smartphone position	Classifier used	Maximum accuracy achieved (%)	Variance
Single sensor used	Leg	KNN	96.39	—
Two sensors used	Arm and leg	KNN	99.27	2.88%
Three sensors used	Arm, belly, and leg	KNN	99.51	0.23%
Four sensors used	Arm and leg	KNN	99.63	0.12%
Five sensors used	Arm, belly, and leg	KNN	99.72	0.14%

TABLE 8: Comparison table for KNN classification results at three body positions (exercisewise) of the only accelerometer and both accelerometer and gyroscope sensors.

Exercise group	Exercise name	Three accelerometers' accuracy (%)	Three accelerometers + two gyroscopes' accuracy (%)	Difference (%)	Avg. time (in sec) consumed at exercise, 2 sets each of 10 reps
Unique exercises group	Face pull	100	100	0.0	26.8
	Cable front raise	100	100	0.0	38.5
	Scott curl	98.4	99.2	+0.6	33.2
	Smith machine drag curl	95.6	100	+4.4	30.9
	Triceps with bar	98.1	99.0	+0.9	26.8
	Decline close grip bench press	100	100	0.0	26.3
	Standing cable cross	100	100	0.0	28.7
	Wide grip pull up	99.5	99.7	+0.2	23.8
	T bar rows	100	100	0.0	24.2
	Chin ups	100	100	0.0	23.7
	Adjustable sit-up bench	100	100	0.0	37.9
	Abs wheel	100	100	0.0	46.5
	Roman chairs	100	100	0.0	37.8
	Flutter kick	100	100	0.0	28.7
	Seated barbell shoulder press	100	97.4	-2.6	27.1
	Incline press wide grip	99.1	100	+0.9	26.6
	Standing barbell press	97.0	99.1	+2.9	27.4
	Inclined dumbbell curl	100	100	0.0	37.2
	Barbell preacher curl	99.3	99.3	0.0	34.4
Common exercises group	Triceps press with cable	98.9	99.0	+0.1	29.2
	Machine bench press	100	100	0.0	25.2
	Leg press	100	100	0.0	41.3
	Romanian deadlift	100	100	0.0	37.2
	Barbell squat	100	100	0.0	30.2
	Dips	100	100	0.0	25.3
Both groups	Butterfly	99.5	100	+0.5	41.0
	All exercises average accuracy	99.51	99.72	+0.21	31.4

To analyze the preprocessed data, we used WEKA (Waikato Environment for Knowledge Analysis) [47]. The preprocessed data (extracted features) were converted to ARFF (Attribute-Relation File Format). The listed attributes were named as per the following strategy. In the name of the

attribute, the first position character 'a' stands for an arm, 'b' stands for the belly, and 'l' stands for the leg. The second position character 'a' or 'g' stands for accelerometer or gyroscope. The third position character 'x', 'y', or 'z' stands for axis values X, Y, and Z. The selected classifiers NB, KNN,

Confusion matrix

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	<- Classified as	
432	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	a = Seated barbell shoulder press
0	441	0	2	0	0	0	0	1	1	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	b = Face Pull
0	2	401	6	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	c = Incline press wide grip
0	1	2	570	4	0	0	0	3	6	5	0	0	1	0	0	0	0	0	0	2	1	0	0	1	0	0	d = Cable front raise
0	5	0	0	430	0	0	0	1	2	4	0	0	1	0	0	1	0	0	0	0	1	0	0	0	2	0	e = Standing barbell press
0	0	0	0	0	593	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	f = Inclined dumbbell curl
0	0	0	0	0	0	524	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	g = Barbell preacher curl
0	1	0	0	0	0	1	488	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	h = Scott curl
0	1	0	8	4	0	0	0	464	5	5	0	0	0	0	0	2	1	2	0	1	1	0	0	2	3	0	i = Smith machine drag curl
0	0	0	1	1	0	0	0	3	444	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	j = Triceps press with cable
0	0	0	2	0	0	0	0	1	14	452	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	k = Triceps with bar
0	0	0	0	0	0	0	0	0	0	0	419	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	l = Decline close grip bench press
0	0	0	0	0	0	0	0	0	0	0	0	430	0	0	1	0	0	0	0	2	0	0	0	0	0	0	m = Machine bench press
0	4	0	0	3	0	0	0	2	4	3	0	0	420	4	0	1	0	0	0	0	0	0	0	1	0	0	n = Standing cable cross
0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	379	0	1	0	0	0	0	0	0	1	3	0	o = Dips
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	618	0	0	0	1	0	0	0	0	0	0	p = Pec deck machine (butterfly)
0	1	2	2	6	0	0	0	1	1	2	0	0	2	0	2	0	373	0	18	1	0	0	0	0	0	0	q = Wide grip pull up
0	0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	0	390	1	0	0	4	0	0	3	0	0	r = T bar rows
0	0	0	1	7	0	1	0	1	0	1	0	1	2	0	0	13	0	345	0	0	0	0	0	1	0	0	s = Chin ups
1	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	587	0	0	0	0	0	0	t = Adjustable sit up bench
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	1	737	0	0	0	0	0	0	u = Abs wheel
0	0	0	0	1	0	0	0	3	2	3	0	0	0	0	0	0	1	0	0	0	605	0	0	0	0	0	v = Roman chairs
0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	482	4	0	0	w = Flutter kick
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	643	0	0	x = Leg press
0	0	0	1	0	0	0	0	3	2	4	0	0	0	0	0	0	2	1	0	4	1	0	0	601	5	0	y = Romanian deadlift
0	2	0	0	1	0	0	0	0	0	0	0	2	2	1	0	0	0	0	0	2	2	0	0	10	481	0	z = Barbell squat

FIGURE 4: Confusion matrix of KNN classification using all smartphones and sensors.

and J48 as per Table 3 were utilized with default configuration settings. In the test options, the percentage split with 80% training and 20% testing option was selected as used also by [38, 48] to evaluate the performance and accuracy of the classifiers.

### 5. Analysis and Results

The existing research mostly used three classifiers, namely, NB, KNN, and J48 (cf. Table 3) and hence they were also utilized in this research. All of the above-mentioned algorithms can create multiform class boundaries and, therefore, are suitable for the data collected via sensors and devices [10]. Furthermore, for practical applications, these methods are fast and are easily implementable [10].

We examined the values of both the sensors (accelerometer and gyroscope) with the above-mentioned classifiers at three different body positions (arm, belly, and leg). The analysis was done in five ways: firstly, the analysis of the exercises considering the data from three sensors of the same nature, that is, accelerometer (rows 1, 2, and 3 of Table 6) attached at three positions (arm, leg, and belly). As there were only two gyro sensors at arm and leg positions, data from the positions are analyzed and presented as per rows 4 and 5 of Table 6. The same process is continued for the combination of three, four, and five sensors as illustrated in Table 6.

In Table 6, column “sensor name” represents the name of the sensor from which the data is acquired. The number in front of the sensor name represents the count of sensors used to acquire and analyze the data. For example, S numbers 6, 7,

and 8 in Table 6 display “accelerometer = 2” which is an indication that two accelerometers attached at the body positions (displayed in the next column) were used to analyze the data. The results of the input data from the chosen three classifiers are presented in the classifier names columns in the form of accuracy. The last column represents the number of features used in the analysis. A single sensor used at body position will have 12 features, two sensors will have 24 features, and so on.

The results revealed that the best accuracy of 99.72% was achieved with the KNN classifier using five sensors at three attachment positions (arm, belly, and leg). However, as can be seen from the summary as per Table 7, this is not a big variation from the accuracy of the KNN using two sensors at two attachment positions. A minimum of two sensors used at the arm and leg position provided an accuracy of 99.27% which is equatable to five sensor positions.

For each (exercise) activity, the accuracies achieved using the KNN classifier with both the accelerometer and gyroscope are a little better than using only the accelerometer. The accuracy results and their difference are shown in Table 8.

The classification confusion matrix in Figure 4 shows that the highest accuracy is achieved using data from all the sensors of the smartphones and with a KNN classifier. Examining the confusion matrix, the results show that most of the classes (exercises) are accurately being predicted. However, a couple of classes (exercises) were not differentiable because of the similarity in the exercise position and nature. For example, the Triceps group (triceps press with cable and triceps press with bar) are having similar motion

patterns. However, we still can differentiate between them based on the execution time differences as per Table 8. The table shows that triceps press with bar takes 26.8 seconds for 20 repetitions while triceps with cable take 29.2 seconds for the same 20 repetitions thus having a difference of 2.4 seconds.

## 6. Conclusion and Future Work

The goal of this study was to predict gym exercises with the help of smartphone sensors in real-world settings. To achieve the goal, exercises from the literature were extracted for which prediction research work was conducted and was intersected with a set of the most used exercises in the gym. The result was 14 unique exercises for this study. Besides, 12 common exercises were also considered for comparison purposes. Furthermore, finding the sensors suitable attachment positions, as well as the number of sensors to utilize in predicting the exercise accurately, was also one of the goals of this research. Also, we conducted the exercises with the greatest number of participants (20) as compared to the existing literature (avg. max. 10). The results indicated three key lessons derived from this study while examining the goals. (a) The most suitable classifier to predict a class from the sensor-based data was found to be KNN. (b) The sensors placed at three positions (arm, belly, and leg) could provide better accuracy than other positions when the gym exercises are under the question and (c) smartphone sensors accelerometer and gyroscope in combination can provide accurate classification (using KNN as classifier at all 3 positions) in most of the activities averaging up to 99.72% accuracy. Their combination can increase accuracy by up to 0.21%.

The research can be implemented in the form of a smartphone application that can be turned on by the users while doing exercises in the gym. In the future, this application can be embedded with a calorie burn out tracker that should be able to guide gym-goers to do which exercise and for how much time? The output could be in the form of sound notifications as well as sound messages that could advise to change or stop the exercise.

The research has some limitations as well. In this research, only 14 unique exercises are considered taking the considered exercises in the literature to 55 exercises. In this context, of the total of 74 exercises as per sources [41, 42], nineteen (19) gym exercises still remain to be predicted though not most often used. The future research work can consider these exercises as well. In addition, in this research, no female participants were involved thus having a probability of nonapplicability of this research for the female participants. The future research could also hire female participants to increase further accuracy.

## Data Availability

The data are available within the supplementary information file. However, any query about the research conducted in this paper is highly appreciated and can be asked from the

principal authors (Usman Ali Khan and Dr. Iftikhar Ahmed Khan).

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Supplementary Materials

An ARFF (Attribute Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software. A complete description including information on how to read this file is available at the following URL. [https://docs.rapidminer.com/latest/studio/operators/data\\_access/files/read/read\\_arff.html](https://docs.rapidminer.com/latest/studio/operators/data_access/files/read/read_arff.html). (*Supplementary Materials*)

## References

- [1] V. Burke, L. J. Beilin, K. Durkin, W. G. K. Stritzke, S. Houghton, and C. A. Cameron, "Television, computer use, physical activity, diet and fatness in Australian adolescents," *International Journal of Pediatric Obesity*, vol. 1, no. 4, pp. 248–255, 2006.
- [2] S. Stieglitz and T. Brockmann, "The impact of smartphones on e-participation," in *Proceedings of Annual Hawaii International Conference on System Sciences*, pp. 1734–1742, Maui, HI, USA, January 2013.
- [3] C. Smith, *Nike Fitbit Flex*, Wareable.com, Malaysia, 2020, <https://www.wareable.com/fitness-trackers/not-so-happy-birthday-nike-fuelband-2351>.
- [4] J. Park and E. Friedman, *FITBIT*, Fitbit.com, USA, 2020, <https://www.fitbit.com/home>.
- [5] O. Banos, M. Damas, H. Pomares, A. Prieto, and I. Rojas, "Daily living activity recognition based on statistical feature quality group selection," *Expert Systems with Applications*, vol. 39, no. 9, pp. 8013–8021, 2012.
- [6] M. Zhang and A. A. Sawchuk, "Human daily activity recognition with sparse representation using wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 553–560, 2013.
- [7] M. Muehlbauer, G. Bahle, and P. Lukowicz, "What can an arm holster worn smartphone do for activity recognition?" in *Proceedings of the 2016 ACM International Symposium Wearable Comput. ISWC*, Heidelberg, Germany, September 2011.
- [8] K.-h. Chang, M. Y. Chen, and J. Canny, "Tracking free-weight exercises," *UbiComp 2007: Ubiquitous Computing*, vol. 4717, pp. 19–37, 2007.
- [9] D. Morris, T. Saponas, A. Guillory, and I. Kelner, "RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises," in *Proceedings of the 32nd Annual Conference on Computer Security*, pp. 3225–3234, New Orleans, LA, USA, October 2014.
- [10] M. Shoaib, H. Scholten, and P. Havinga, "Towards physical activity recognition using smartphone sensors," in *IEEE 10th International Conference on Ubiquitous Intelligence and Computing*, pp. 80–87, Vietri sul Mare, Italy, December 2013.
- [11] H. Koskimäki, "MyoGym-introducing an open gym data set for activity recognition collected using myo armband," *UbiComp/Iswc*, vol. '17, pp. 537–546, 2017.

- [12] D. M. Bravata, C. Smith-Spangler, V. Sundaram et al., "Using pedometers to increase physical activity and improve health," *JAMA*, vol. 298, no. 19, pp. 2296–2304, 2007.
- [13] C. B. Chan, D. A. J. Ryan, and C. Tudor-Locke, "Health benefits of a pedometer-based physical activity intervention in sedentary workers," *Preventive Medicine*, vol. 39, no. 6, pp. 1215–1222, 2004.
- [14] D. Merom, C. Rissel, P. Phongsavan et al., "Promoting walking with pedometers in the Community: The step-by-step trial," *American Journal of Preventive Medicine*, vol. 32, no. 4, pp. 290–297, 2007.
- [15] I. Pernek, K. A. Hummel, and P. Kokol, "Exercise repetition detection for resistance training based on smartphones," *Personal and Ubiquitous Computing*, vol. 17, no. 4, pp. 771–782, 2013.
- [16] C. Seeger, A. Buchmann, and K. Van Laerhoven, "myHealthAssistant: a phone-based body sensor network that captures the wearer's exercises throughout the day," in *Proceedings of the 6th ICST Conference on Body Area Networks*, Beijing, China, November, 2011.
- [17] O. Baños, M. Damas, H. Pomares, I. Rojas, M. Tóth, and O. Amft, "A benchmark dataset to evaluate sensor displacement in activity recognition," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, p. 1026, Pittsburgh PA, USA, September 2012.
- [18] K. Van Laerhoven and O. Cakmakci, "What shall we teach our pants?" in *Proceedings of the 2nd IEEE International Symposium on Wearable Computer*, pp. 77–83, Cambridge, MA, USA, September 2000.
- [19] O. D. Incel, M. Kose, and C. Ersoy, "A review and taxonomy of activity recognition on mobile phones," *Bionanoscience*, vol. 3, no. 2, pp. 145–171, 2013.
- [20] M. Habib, M. Mohktar, S. Kamaruzzaman, K. Lim, T. Pin, and F. Ibrahim, "Smartphone-based solutions for fall detection and prevention: challenges and open issues," *Sensors*, vol. 14, no. 4, pp. 7181–7208, 2014.
- [21] A. Anjum and M. U. Ilyas, "Activity recognition using smartphone sensors," in *Proceedings of the 2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, pp. 914–919, Las Vegas, NV, USA, January 2013.
- [22] W. Wu, S. Dasgupta, E. E. Ramirez, C. Peterson, and G. J. Norman, "Classification accuracies of physical activities using smartphone motion sensors," *Journal of Medical Internet Research*, vol. 14, no. 5, pp. 1–9, 2012.
- [23] I. E. Smith and W. G. Griswold, "Ubiquitous computing," *UbiComp*, vol. 4206, 2006.
- [24] M. B. Berchtold, D. Gordon, H. R. Schmidtke, M. Beigl, and ActiServ, "Activity recognition service for mobile phones," in *Proceedings International Symposium on Wearable Computers, ISWC*, Seoul, South Korea, October 2010.
- [25] G. Bieber, P. Koldrack, C. Sablowski, C. Peter, and B. Urban, "Mobile physical activity recognition of stand-up and sit-down transitions for user behavior analysis," in *Proceedings of the 3rd International Conference on Pervasive Technologies PETRA*, Samos, Greece, January 2010.
- [26] A. Henpraserttae, S. Thiemjarus, and S. Marukatat, "Accurate activity recognition using a mobile phone regardless of device orientation and location," in *Proceedings of the 2011 International Conference on Body Sensor Networks (BSN)*, Zakopane, Poland, June 2011.
- [27] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.
- [28] Y.-S. Lee and S.-B. Cho, "Activity recognition using hierarchical hidden Markov models on a smartphone with 3D accelerometer," *Lecture Notes in Computer Science*, pp. 460–467, 2011.
- [29] A. F. Olsen and J. Torresen, "Smartphone accelerometer data used for detecting human emotions," in *Proceedings of the 3rd International Conference on Systems and Informatics*, pp. 410–415, Shanghai, China, November 2016.
- [30] Z. Wang, D. Wu, R. Gravina, G. Fortino, Y. Jiang, and K. Tang, "Kernel fusion based extreme learning machine for cross-location activity recognition," *Information Fusion*, vol. 37, pp. 1–9, 2017.
- [31] B. J. Mortazavi, M. Pourhomayoun, G. Alsheikh, N. Alshurafa, S. I. Lee, and M. Sarrafzadeh, "Determining the single best axis for exercise repetition recognition and counting on smartwatches," in *Proceedings of the 2015 IEEE 11th International Conference on Wireless and Mobile Implantable Body Sensor Networks (BSN)*, Jeju Island, Korea, July 2014.
- [32] M. Kose, O. Incel, and C. Ersoy, "Online human activity recognition on smart phones," in *Proceedings of the 2nd International Work Mobile Sensor From Smartphones Wearables to Big Data*, Beijing, China, October 2012.
- [33] F. Li, K. Shirahama, M. Nisar, L. Köping, and M. Grzegorzek, "Comparison of feature learning methods for human activity recognition using wearable sensors," *Sensors*, vol. 18, no. 3, p. 679, 2018.
- [34] Y. E. Ustev, O. Durmaz Incel, and C. Ersoy, "User, device, and orientation independent human activity recognition on mobile phones," in *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct*, pp. 1427–1436, Zurich, Switzerland, September, 2013.
- [35] T. Saponas, J. Lester, J. Froehlich, J. Fogarty, and J. Landay, *iLearn on the iPhone: Real-Time Human Activity Classification on Commodity Mobile Phones*, pp. 8–4, University of Washington, Seattle, WA, USA, 2008.
- [36] M. Nilsson and H. Wilén, *Push-up Tracking through Smartphone Sensors*, MS Thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2016.
- [37] K. Liu, Y. Wang, R. Chen, T. Chu, and J. Bi, "A survey of human activity recognition using smartphones," *Journal of Residuals Science & Technology*, vol. 13, no. 8, pp. 1–10, 2016.
- [38] C. A. Ronao and S.-B. Cho, "Recognizing human activities from smartphone sensors using hierarchical continuous hidden Markov models," *International Journal of Distributed Sensor Networks*, vol. 13, no. 1, p. 155014771668368, 2017.
- [39] L. Liu, Y. Peng, M. Liu, and Z. Huang, "Sensor-based human activity recognition system with a multilayered model using time series shapelets," *Knowledge-Based Systems*, vol. 90, pp. 138–152, 2015.
- [40] H. Martín, A. M. Bernardos, J. Iglesias, and J. R. Casar, "Activity logging using lightweight classification techniques in mobile devices," *Personal and Ubiquitous Computing*, vol. 17, no. 4, pp. 675–695, 2013.
- [41] Bodycraft Exercise Guide, 2020, <https://www.bodycraft.com/pdfs/exercise/ExerciseBook.pdf>.
- [42] Bodybuilding.com, 2020, The Personal Training System, URL: <https://www.bodybuilding.com/fun/guide.pdf>.
- [43] S. Woods, T. Bridge, D. Nelson, K. Risse, and D. M. Pincivero, "The effects of rest interval length on ratings of perceived exertion during dynamic knee extension exercise," *The Journal of Strength and Conditioning Research*, vol. 18, no. 3, pp. 540–545, 2004.

- [44] D. Micucci, M. Mobilio, and P. Napolitano, “UniMiB SHAR: a new dataset for human activity recognition using acceleration data from smartphones,” *Applied Sciences*, vol. 10, p. 1101, 2016.
- [45] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “A public domain dataset for human activity recognition using smartphones, eur. Symp. Artif. Neural networks,” *Computational Intelligence*, pp. 24–26, 2013.
- [46] A. Bulling, U. Blanke, and B. Schiele, “A tutorial on human activity recognition using body-worn inertial sensors,” *ACM Computing Surveys*, vol. 46, no. 3, pp. 1–33, 2014.
- [47] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
- [48] B.-J. Ho, R. Liu, H.-Y. Tseng, and M. Srivastava, “MyoBuddy: detecting barbell weight using electromyogram sensors,” *Proceedings of the 1<sup>st</sup> Work. Digit. Biomarkers*, pp. 27–32, New York, NY, USA, 2017.

## Research Article

# QuPiD Attack: Machine Learning-Based Privacy Quantification Mechanism for PIR Protocols in Health-Related Web Search

Rafiullah Khan,<sup>1,2</sup> Arshad Ahmad ,<sup>3</sup> Alhuseen Omar Alsayed,<sup>4</sup> Muhammad Binsawad,<sup>5</sup> Muhammad Arshad Islam,<sup>6</sup> and Mohib Ullah<sup>1,2</sup>

<sup>1</sup>*Institute of Computer Science and Information Technology, The University of Agriculture, Peshawar, Pakistan*

<sup>2</sup>*Capital University of Science and Technology, Islamabad, Pakistan*

<sup>3</sup>*Department of Computer Science, University of Swabi, Anbar, Pakistan*

<sup>4</sup>*Deanship of Scientific Research, King Abdulaziz University Jeddah, Jeddah, Saudi Arabia*

<sup>5</sup>*Faculty of Computer Information Systems, King Abdulaziz University Jeddah, Jeddah, Saudi Arabia*

<sup>6</sup>*National University of Computer and Emerging Sciences, Islamabad, Pakistan*

Correspondence should be addressed to Arshad Ahmad; [yaarshad@gmail.com](mailto:yaarshad@gmail.com)

Received 16 March 2020; Accepted 22 April 2020; Published 14 July 2020

Academic Editor: Rodziah Binti Atan

Copyright © 2020 Rafiullah 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.

With the advancement in ICT, web search engines have become a preferred source to find health-related information published over the Internet. Google alone receives more than one billion health-related queries on a daily basis. However, in order to provide the results most relevant to the user, WSEs maintain the users' profiles. These profiles may contain private and sensitive information such as the user's health condition, disease status, and others. Health-related queries contain privacy-sensitive information that may infringe user's privacy, as the identity of a user is exposed and may be misused by the WSE and third parties. This raises serious concerns since the identity of a user is exposed and may be misused by third parties. One well-known solution to preserve privacy involves issuing the queries via peer-to-peer private information retrieval protocol, such as useless user profile (UUP), thereby hiding the user's identity from the WSE. This paper investigates the level of protection offered by UUP. For this purpose, we present QuPiD (query profile distance) attack: a machine learning-based attack that evaluates the effectiveness of UUP in privacy protection. QuPiD attack determines the distance between the user's profile (web search history) and upcoming query using our proposed novel feature vector. The experiments were conducted using ten classification algorithms belonging to the tree-based, rule-based, lazy learner, metaheuristic, and Bayesian families for the sake of comparison. Furthermore, two subsets of an America Online dataset (noisy and clean datasets) were used for experimentation. The results show that the proposed QuPiD attack associates more than 70% queries to the correct user with a precision of over 72% for the clean dataset, while for the noisy dataset, the proposed QuPiD attack associates more than 40% queries to the correct user with 70% precision.

## 1. Introduction

Currently, web search engines (WSEs) have become the preferred way to find health care-related content on the World Wide Web. A recent survey reports that more than 80% of patients use WSE to seek health-related information before consulting the physician [1], while according to the report published by Pew Research Center, 35% of American adults consulted WSE to diagnose medical conditions [2]. However, while using the web search services, the user usually posts their physical condition and health information as a query [3]. Web search engines claim that they

collect and maintain user queries as user profile for various activities such as result ranking [4], market research [3], personalization [5], targeted advertisements [6], and others. On the brighter side, maintaining users profile can actually improve the quality of results and user experience, while on the darker side, this indiscriminate collection of users' queries may cause critical privacy breaches as users' queries may contain sensitive and personal information [7]. This issue of users' privacy breach received significant attention in 2005 when the US Department of Justice compelled Google to submit records of users' queries [8]. Later, America Online (AOL) released (pseudonymized) 20

million queries of more than 650,000 users submitted in three months of time [9], from which the identities of some users had been inferred through personal information enclosed in their queries [10].

Patient's health information is considered to be a sensitive issue since ancient times, and it is also reflected in the Hippocratic Oath [11] that physician will keep the patient's information secret [12]. However, in online and public health facility services, user privacy is just becoming behavior tracking [12]. Consider a scenario when a user posts a series of private queries related to his/her health condition such as "HIV" or "diabetes." WSE may sell this information to the advertisement agencies or other companies for business purposes, which ultimately breaches the user's privacy [3]. Such kind of privacy disclosure happened in 2006 when the New York Times managed to deduce and infer personal information from the search history from the pseudonymized log published by AOL. One of them was a 62-year-old widow who conducted hundreds of searches related to her health condition such as "hand tremors," dry mouth," and "nicotine effects on the body" which were linked back to her [13].

To address this issue of privacy infringement, several methods have been proposed. These methods include user profile obfuscation [14], query scrambling [15], anonymizing networks [16], and private information retrieval (PIR) protocols [17–20]. In a user profile obfuscation, a user profile is contaminated with fake queries to mislead the WSE. In the query scrambling technique, the user query is replaced by a set of blurred and benign synonyms and later posted to WSE. Techniques based on anonymizing network forward the user query through a series of routers to make it difficult for WSE to trace the origin of the query. These methods hide the IP address while the user is still traceable through cookies and device fingerprints [21]. In PIR protocols, a group of users submits queries on behalf of each other to hide their identity.

Despite the fact that the aforementioned methods improve the user privacy, yet some previous studies [22–25] using a machine learning algorithm and user profile (i.e., user history or logged user queries) show that an adversary is able to break profile obfuscation and anonymizing network methods. However, it is not clear if an adversary is able to break PIR protocols using machine learning techniques. Therefore, in this research, we propose a machine learning-based attack in order to evaluate the effectiveness of popular PIR protocol, i.e., useless user profile (UUP) [17, 18].

A higher-level goal of this work is to analyze the effectiveness of PIR protocol in preserving users' privacy against an adverse WSE (from here on, we will call the PIR protocols as UUP, for simplicity of presentation without loss of generality). In UUP, a group of users exchanges their queries with each other in such a way that the identity of the query originator node remains hidden from other group mates. In the next step, all group members submit the received queries to the WSE and results are broadcasted in the group. On the WSE side, the user's query is received in plain text but with a different identity, and thus WSE cannot identify the originator of the queries. We set out to investigate whether it is possible (and to what extent) for an

adverse WSE—equipped with users' web search profile (histories)—to link the queries coming out of UUP exit user to the original users and thus undermine the privacy provided by UUP.

To better understand the limits of UUP on user's privacy, we present in this paper a study of UUP focusing on active users. This study is conducted with QuPiD attack, a machine learning-based attack that determines the distance between the user's profile and query. We conducted our experiments with randomly selected active 100 users from publicly available AOL dataset and treated them as users of UUP. The AOL dataset is composed of over 20 million queries submitted during the period of March 1, 2006, to May 31, 2006, by 6.5 million users. The data of the first two months are used as training data while the last month data are used as testing data. We measured the efficiency of attack using some known machine learning matrices: precision, recall, F-measure, and true-positive rate. The results showed that our proposed QuPiD attack associates more than 70% queries to the correct user with more than 72% precision. Based on the results, we can conclude that most of the users are vulnerable to privacy infringement despite using UUP. The contributions of this work are as follows:

- (1) Proposed QuPiD attack: a machine learning-based attack for privacy evaluation of PIR protocols
- (2) A proposed new vector for query classification
- (3) Recommendation of a suitable machine learning algorithm for query classification

The remainder of the paper is organized as follows: In Section 2, we describe the proposed QuPiD attack. Experimentation setup, preprocessing of the dataset, feature vector construction, and classification algorithms are discussed in Section 3. Section 4 presents the experimental results. Section 5 presents the conclusions and outlines directions for future work.

## 2. Adverse Model and QuPiD Attack

Users are more concerned about the privacy risks of querying WSEs. In this work, we investigated the robustness of popular PIR protocol, i.e., UUP. As mentioned earlier, WSE receives a user's query with a different identity due to the shuffling process. Therefore, the entries of queries will never appear with their true originator in the weblog. However, the weakness of this protocol is the timing of query submission by all group members. After the query shuffling step, every group member submits the received query to WSE almost at the same time. Due to which their entries appeared close to each other in the weblog. Figure 1 illustrates an example of query entries in the weblog. In Figure 1, exhibit 1 shows the users' queries before the shuffling process while exhibit 2 shows the queries after the shuffling process. After shuffling, the queries are submitted to WSE (Figure 1, exhibit 3).

In the proposed adverse model, WSE is assumed to be an entity whose goal is to work against the privacy-preserving solution and identify the user of interest (UoI) queries for profiling purposes. It is assumed that WSE is equipped with



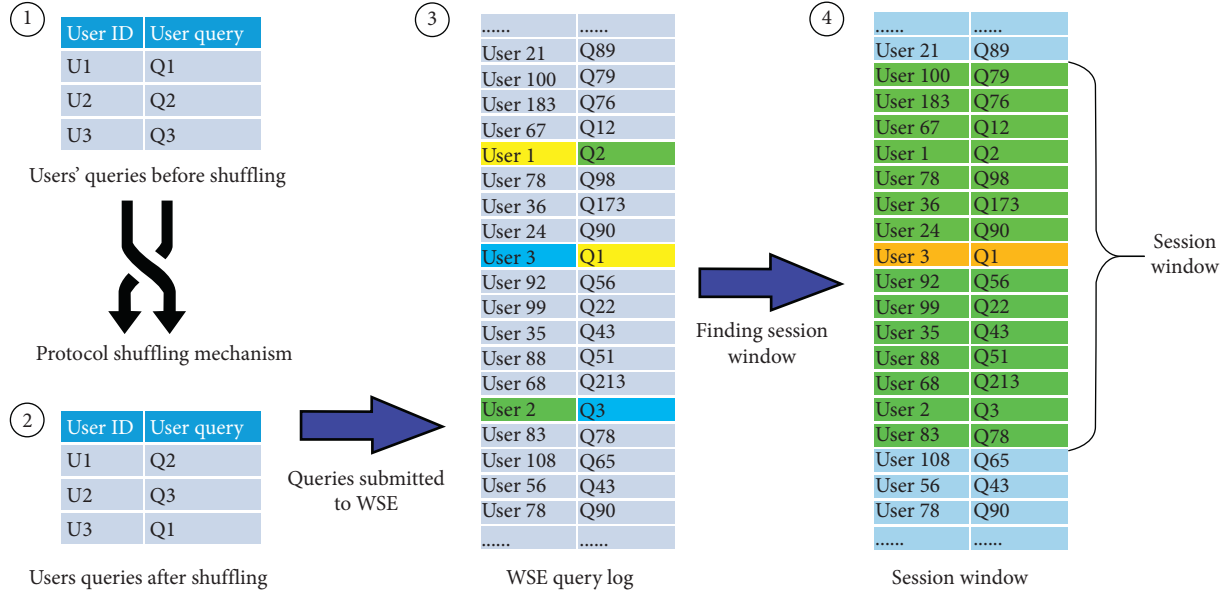


FIGURE 1: Query entry in the weblog and session window.

the user's search history (i.e., user profile) PU. The user profile contains queries submitted by the user in the past without using any UUP protocol shown in equation (1) (where  $P_{q,i}$  shows the queries in the UoI profile).

$$PU = [P_{q,1}, P_{q,2}, P_{q,3}, \dots, P_{q,n}]. \quad (1)$$

The user profile PU is used as training data for building the classification model. As the dataset used for experimentation is spread across three months' duration, the first two months' data are used as a training set, while the UUP protocol is simulated with the third month data to create an anonymized log (as shown in Figure 1, exhibit 3). The anonymized log is used as the test set. For testing, all session windows of the UoI are drawn out from the query logs. Here, the session window is a block of records (query entries in the log) in an anonymized log that contains the entry of UoI, but with another user [26, 27]. In other words, the session window is composed of the selected number of queries' entries in the WSE query log, which appeared immediately before and after the query of UoI. As shown in Figure 1 (exhibit 4), our UoI is "User 3" and the session window size is 15 records (7 records before UoI and 7 after UoI). For this research, we have used the window size of 251 records. Each session window ( $S_{win}$ ) is composed of 125 queries appearing before and 125 queries appearing after the query of UoI (as per the recommendation of [27]). A generic session window  $S_{win}$  is shown in equation (2) (where  $q_i$  represents a query in the session window). The collection of all session windows  $GS_{win}$  is shown in equation (3).

$$S_{win} = [q1, q2, q3, \dots, q125, qUoI, q126, \dots, q251], \quad (2)$$

$$GS_{win} = [S_{win}1, S_{win}2, S_{win}3, \dots, S_{win}n]. \quad (3)$$

As shown in the query log, the target user who uses any PIR protocol will remain hidden since his/her query is exchanged with a query of another user in the group.

Therefore, a session window is used to reduce the testing data. Both PU (training set) and  $GS_{win}$  (testing set) are used as input to the algorithm of the adverse model. The working of the adverse model is presented in Algorithm 1 and depicted in Figure 2. The working of the algorithm is as follows:

$$PU_v = [P_{q,1v}, P_{q,2v}, P_{q,3v}, \dots, P_{q,nv}], \quad (4)$$

$$S_{winv} = [q1v, q2v, q3v, \dots, q125v, qUoIv, q126v, \dots, q251v]. \quad (5)$$

For experimentation purposes, two subsets of 100 users were created from the AOL dataset constituting a three-month web query log of AOL users. Each subset was divided into two portions, i.e., training and testing data. Training data are composed of the first two months of the log, while the testing data are composed of the last month of the log. The details of the user selection criteria and dataset formation are discussed in Section 4.

### 3. Methodology

**3.1. AOL Dataset.** We used the real-world web search query log released by AOL in 2006 for the evaluation of our proposed adverse model. The AOL dataset consists of over 20 million queries submitted during the period of March 1, 2006, to May 31, 2006, by 6.5 million users. Although the AOL dataset is old and has a lot of deficiencies as compared to the current situation, we are forced to use this dataset due to a lack of availability of the benchmark dataset. The attributes of the query log are user ID, query, date and time of the query, the rank of the content clicked, and the clicked URL. For experimentation purposes, the data of the first two months were used as user profile (PU) or training data while the third month's data were the new queries to be classified

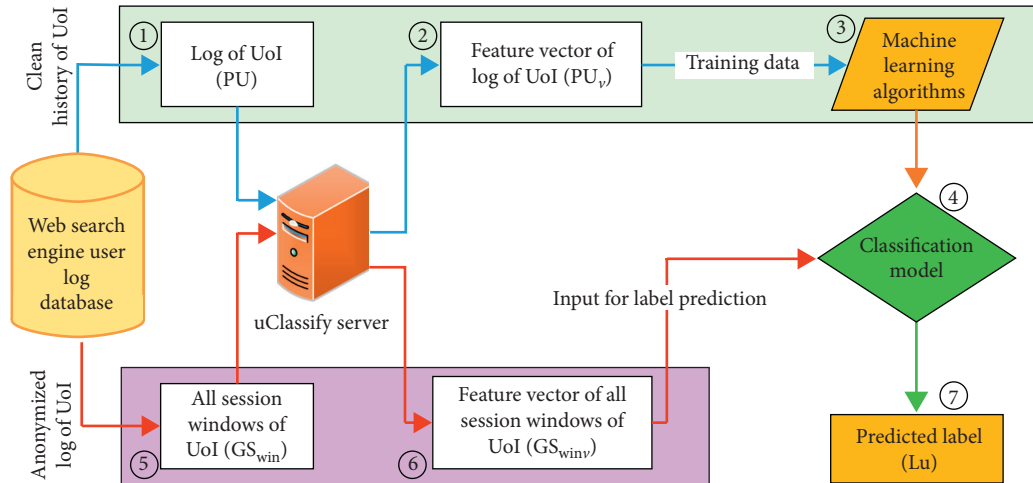


FIGURE 2: Operation of the adverse model.

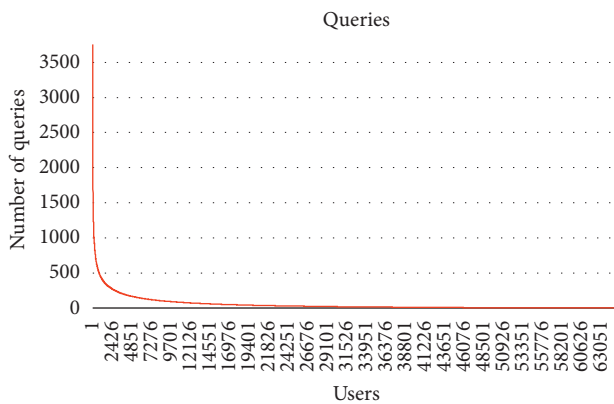


FIGURE 3: Distribution of the number of queries issued per user in the selected dataset.

(i.e., testing data). The distribution of the number of queries issued per user in the selected dataset is shown in Figure 3. For experimentation, we chose 100 users with high query frequency instead of concentrating on all users. The user selection criteria are discussed in Section 3.3, while the summary of the dataset is provided in Table 1.

**3.2. Feature Vector Extraction.** The dataset is composed of five attributes: user ID, query, date and time of the query, the rank of the content clicked, and the clicked URL. Since our adverse model works with the user ID, submitted query, and query score in ten major topics, we neglect the remaining features. To obtain query scores in ten major classes, we used uClassify service that provides classifiers for topics, age, gender, sentiments, language detection, and many others. In this paper, the topic classifier is employed that provides the numeric value of 10 categories against each query. The topic classifier uses a subset of topics from the Open Directory Project (ODP) directory in which topics are placed in a hierarchy. The classes are Arts, Business, Computers, Games, Health, Home, Recreation, Science, Society, and Sports. The classifier provides the percentage of each query in each

category. For example, for query “olive oil,” the score for each topic is shown in Table 2.

In some cases, uClassify was unable to find the score of the dominant topic of the submitted query. For example, uClassify is unable to find the dominant class for the query “glenliviet 18.” Therefore, in that case, uClassify just divided an equal score in each class, i.e., 10% for each class. We refer to this kind of query as a “confused query” (shown in Table 2). In the dataset of selected 100 users, uClassify marked 28% of the queries as confusing queries. Therefore, we conducted our experiments using two datasets. One dataset was comprised of both confused and unconfused queries, while the other dataset was comprised of only unconfused queries to find the impact of confusing queries over the results of a classifier. From this point onwards, the dataset with confused queries will be referred to as a noisy dataset while the dataset with only unconfused queries will be referred to as the clean dataset. The details of both datasets are given in Table 3.

**3.3. User Selection and Subset Construction.** Instead of conducting experiments using all users, we focused on a few users who were considered to be active. Active users are those users who submitted more than 300 queries for at least 61 days during the entire period. From the analysis of the dataset, we found only 21,407 (3.29%) users to be active users. From those active users, we randomly selected 100 users as UoI. The cumulative distribution of queries in both noisy and clean dataset is shown in Figure 4. To see the effects of the size of the training data, we divide both noisy and clean datasets into five groups based on the average of query frequency. The selected 100 users are divided into 5 groups in both datasets. The average number of total, training, and testing instances in all groups for both datasets is given in Table 4.

**3.4. Anonymized Log Creation.** As mentioned earlier, the AOL data spans across three months. For experimentation purposes, we have considered the first two months’ data as

**Input:** User Profile (PU); all session windows belong to the user ( $GS_{win}$ ).

**Output:** Expected User Label (Lu)

```

(1) procedure QUERY ASSOCIATION (PU,  $GS_{win}$ )
(2)   for  $P_{qi} \in PU$  do
(3)      $PU_v \leftarrow$  get feature vector for ( $P_{qi}$ )
(4)      $P_{Model} \leftarrow$  Classification Algorithm ( $PU_v$ )
(5)     for  $S_{win}^i \in GS_{win}$  do
(6)       for  $q_k \in S_{win}^j$  do
(7)          $q_{k^v} \leftarrow$  get feature Vector for ( $q_k$ )
(8)          $Lu \leftarrow P_{Model}(q_{k^v})$ 
(9)   return Lu

```

- (1) Firstly, the user profile (PU) feature vector is acquired for training purposes. The user profile with the feature vector ( $PU_v$ ) is shown in equation (4). The feature vector is acquired from the uClassify (<http://www.uclassify.com>) service, a machine learning web service that provides numerous different classifiers for text classification. We have selected the “Topics” classifier that gives the score of each phrase or query in 10 major classes including Sports, Society, Science, Recreation, Home, Health, Games, Computers, Business, and Arts.
- (2) In the second step, a classification model  $P_{Model}$  is built using  $PU_v$  and supervised machine learning algorithms. To test the response of the data with different classification techniques, 10 classification algorithms are selected from tree-based, rule-based, lazy learner, metaheuristic, and Bayesian families.
- (3) After the classification model ( $P_{Model}$ ), the third step is to acquire the feature vector  $S_{win^v}$  shown in equation (5) for the queries of session window  $S_{win}$  from uClassify for testing data.
- (4) In the last step, each query of  $S_{win^v}$  is provided to the classification model for the expected label Lu. The label Lu shows whether the incoming query belongs to UoI or not.

ALGORITHM 1: Associating incoming query to the user using the prior profile.

TABLE 1: AOL dataset properties.

Total queries	36,389,567
Total users	657,426
Unique queries	10,154,742
Attributes	5 (AnonID, query, query time, item rank, click URL)
Time duration	01 March, 2006–31 May, 2006

TABLE 2: The score of queries from uClassify.

Query	Arts	Business	Computers	Games	Health	Home	Recreation	Science	Society	Sports
Olive oil	0.0386	0.0974	0.0280	0.0396	0.0569	0.4659	0.0652	0.1028	0.0874	0.0182
Glenlivet 18	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

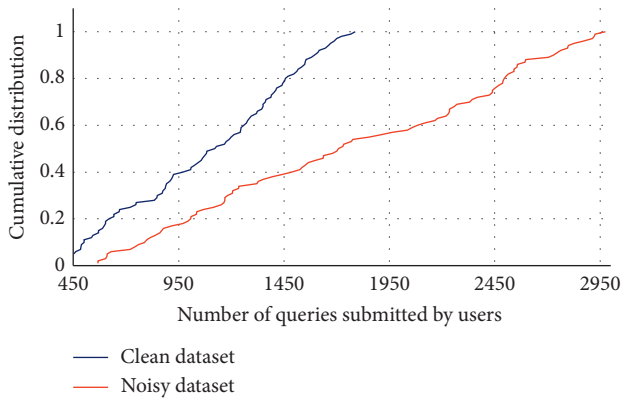


FIGURE 4: Distribution of the number of queries submitted by users in the clean and noisy datasets.

the clean history of UoI available to the search engine and last month’s data as new queries to be classified. The selected PIR protocol, i.e., (UUP) is simulated with the third month’s query log to create the anonymized log of UoI. The parameters considered for simulations are group size and the number of queries submitted by the respective users. According to the literature, UUP is tested with a group size of 3, 4, 5, and 10 users [17, 18]. Another study indicated that a bigger group size offers more privacy [27]. We, therefore, considered a group size of 20 users. The number of queries submitted by the target user is dependent on the actual query frequency of the selected user in the third month queries log.

3.5. *Classification Algorithms.* In several previous studies, Peddinti et al. [23, 24] and Petit [21] used Random Forest,

AD Tree, Zero R, Regression, and SVM algorithms for the classification of the data queries. In both studies, the classification model was biclass, i.e., the query is machine or user generated. Moreover, the model was built based on two attributes like query and assigned label. In our work, however, the classification model is multiclass, i.e., in the testing data, the model will decide which query belongs to which user and the model is based on twelve attributes (discussed in Section 3.2). We selected ten off-the-shelf (with default settings) different families' classification algorithms. We chose J48 [28] and Logistic Model Tree (LMT) [29] from the tree-based family, Decision Table [30], JRip [31], and OneR [32] from rule-based family, IBK [33] and KStar [34] from lazy learner family, Bagging [35] and LogitBoost [36] from metaheuristic family, and Bayes Net [37] from Bayesian family. Rep Tree [38] and Regression are used as base classifiers for Bagging and LogitBoost algorithms.

**3.6. Performance Evaluation Metrics.** Three metrics, precision, recall, and F-measure, are usually used to evaluate the performance of a classifier. Precision represents how many of the identified samples are correct and recall describes how many of the total samples are correctly identified. Both precision and recall are mathematically represented in the following equations:

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}, \quad (6)$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \quad (7)$$

where true positive represents the actual positives that are correctly identified cases by the classifier and false positive is the proportion of all negatives that still yield positive test outcomes, while false negative represents the proportion of positive which yields negative test outcomes with the test. The trade-off between precision and recall is represented by a unified metric called F-measure. The value of F-measure is in the range from 0 to 1, where 0 shows none of the samples is classified correctly, while 1 shows perfect classification. Mathematically, F-measure is represented as

$$F - \text{measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (8)$$

## 4. Results and Discussion

The primary aim of this study is to propose and evaluate a privacy quantification model for PIR protocols. Experiments are performed with two datasets: noisy and clean (Section 3.3), each set composed of 100 users having variable query frequencies distributed over five groups. For each UoI, we measured precision, recall, and true-positive percentage of correctly classified queries from an anonymized log.

Tables 5 and 6 illustrate the true-positive percentage of the queries of *UoI* in both datasets. According to Table 5, all

algorithms correctly identified more than 89% queries of 2 users in the noisy dataset except OneR and LogitBoost. OneR correctly identified 80% to 90% queries of 4 users. Overall, IBK correctly identified more than 50% queries of 36 users followed by Bagging and KStar with 30 and 28 users, respectively, in the noisy dataset. Similarly, in the clean dataset, LMT and IBK were able to correctly identify more than 89% queries of 14 users followed by J48 and Bagging with 12 users each. Overall, IBK correctly identified more than 50% queries of all 100 users followed by KStar and Bagging with 96 and 92 users in the clean dataset. The detailed performance of all algorithms (in terms of true-positive rate) of the clean dataset is given in Table 6. In both datasets, the performance of lazy learner family algorithms (i.e., IBK and KStar) is better when compared to other selected algorithms.

As mentioned earlier, both datasets are further divided into 5 groups of 20 users (Table 4) in order to observe the impact of the size of training on the accuracy of results. Table 7 shows the comparison of the performance of all algorithms with a variation of the training dataset size in the noisy dataset. The performance of each algorithm is measured in precision and recall. IBK and KStar associated more than 40% queries to the correct user with the precision of above 60% in all cases, while Bagging, J48, Decision Table, and Bayes Net associated more than 25% queries to the correct user with the precision of above 60% in all cases. From the perspective of the size of the training dataset, it is slightly difficult to draw a conclusion about its effect on accuracy. Almost every algorithm shows irregular behavior with a variation in the training dataset size. For the first three groups, the performance of IBK, J48, KStar, and LMT is observed more accurately. However, unexpectedly, the rate of recall drops for the last two groups. The results of precision and recall of noisy data are plotted in Figure 5.

In the clean dataset, however, a clear pattern of improvement in the recall is visible. According to Table 8, the performance of all algorithms is improving as the size of the training dataset increases. IBK and KStar associated more than 62% queries to the correct users with the precision of above 70% in all cases, while Bagging, J48, Decision Table, and LMT associated more than 51.68% to 82.84% queries to the correct user with the precision of above 60% in all cases. Among other algorithms, Bayes Net was able to associate more than 70% of the queries in some cases. Although the increase in recall with the increase in training data is not linear, an improvement pattern is clearly visible in the clean dataset. The results of precision and recall of clean data are plotted in Figure 6.

Overall, IBK and Bagging associated 45.1% and 43% queries to the correct user with above 70% precision for the noisy dataset, while J48, KStar, and LMT associated 42.2%, 41.7%, and 40.6% queries to the correct user with the precision of 70.9%, 73.5%, and 70.2%. Similarly, in the clean dataset, IBK and Bagging associated 79.5% and 75.7% queries to the correct user with 79.6% and 75.9% precision, while J48, KStar, and LMT associated 73.9%, 74.4%, and 72% queries to the correct user with the precision of 73.9%,

TABLE 3: Properties of noisy and clean datasets.

Properties	Noisy dataset	Clean dataset
Training instances	116101	71817
Testing instances	59809	36998
Total instances	175911	108815
Max queries by the single user	2975	1788
Min queries by the single user	567	365
Distinct queries	69164	49662

TABLE 4: Average dataset instances (queries).

Dataset	Group	Total data	Training data	Testing data
Noisy	Group 1	777.55	513.183	264.37
	Group 2	1215.15	801.99	413.15
	Group 3	1752.45	1156.62	595.833
	Group 4	2332.1	1539.18	792.91
	Group 5	2718.3	1794.08	924.22
Clean	Group 1	509.55	336.30	173.25
	Group 2	820.95	541.83	279.12
	Group 3	1132	747.12	384.88
	Group 4	1367	902.22	464.78
	Group 5	1611.25	1063.43	547.83

TABLE 5: Percentage of users in a group based on true-positive values of the noisy dataset.

True-positive percentage bands	Tree-based		Rule-based			Lazy learner		Metaheuristic		Bayesian
	J48	LMT	DT	JRip	OneR	IBK	KStar	Bagging	LogitBoost	Bayes Net
100%-90%	2	2	2	2	0	2	2	2	0	2
90%-80%	2	2	2	2	4	2	0	2	2	2
80%-70%	4	2	4	2	0	4	4	4	2	4
70%-60%	4	8	4	2	6	4	6	4	0	2
60%-50%	14	2	4	6	0	24	16	18	2	14
50%-40%	26	28	24	4	10	18	22	20	6	20
Below 40%	48	56	60	82	80	46	50	50	88	56

TABLE 6: Percentage of users in a group based on true-positive values of the clean dataset.

True-positive percentage bands	Tree-based		Rule-based			Lazy learner		Metaheuristic		Bayesian
	J48	LMT	DT	JRip	OneR	IBK	KStar	Bagging	LogitBoost	Bayes Net
100%-90%	12	14	10	10	4	14	8	12	0	4
90%-80%	18	12	14	8	4	32	26	24	4	12
80%-70%	26	22	22	6	8	24	26	22	2	18
70%-60%	20	26	16	8	6	22	20	18	0	18
60%-50%	12	16	10	10	18	8	16	16	6	14
50%-40%	12	10	20	16	18	0	4	8	10	18
Below 40%	0	0	8	42	42	0	0	0	78	16

76.1%, and 72.6%. The top three algorithms in terms of F-measure (trade-off between precision and recall) for the noisy dataset are IBK, Bagging, and J48 with the score of 0.514, 0.487, and 0.477, respectively, while for the clean dataset, the top three algorithms are IBK, Bagging, and KStar

with the score of 0.793, 0.753, and 0.745, respectively. Hence, IBK is determined to be a more appropriate algorithm for the feature vector “categories.” The results of the average F-measure of the noisy and the clean dataset are plotted in Figure 7.

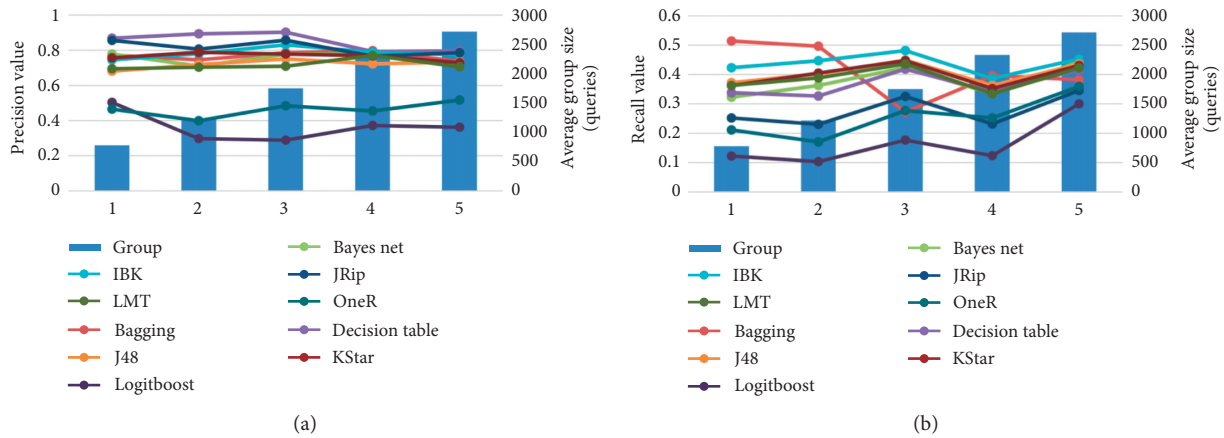


FIGURE 5: Noisy dataset's groupwise precision and recall in different groups. (a) Noisy dataset precision. (b) Noisy dataset recall.

TABLE 7: Precision and recall of noisy dataset in different groups.

Group			Group 1	Group 2	Group 3	Group 4	Group 5
Tree-based	J48	Precision	0.68	0.71	0.75	0.72	0.72
		Recall	0.37	0.40	0.44	0.36	0.43
	LMT	Precision	0.69	0.70	0.70	0.75	0.72
		Recall	0.36	0.38	0.43	0.33	0.42
Rule-based	Decision Table	Precision	0.86	0.89	0.90	0.79	0.79
		Recall	0.33	0.32	0.41	0.34	0.41
	JRip	Precision	0.85	0.80	0.85	0.77	0.78
		Recall	0.25	0.23	0.32	0.23	0.34
OneR	Precision	0.46	0.39	0.48	0.46	0.51	
	Recall	0.21	0.17	0.27	0.25	0.35	
Lazy learner	IBK	Precision	0.74	0.78	0.83	0.78	0.77
		Recall	0.42	0.44	0.48	0.38	0.45
	KStar	Precision	0.75	0.78	0.77	0.76	0.72
		Recall	0.36	0.40	0.44	0.35	0.72
Metaheuristic	Bagging	Precision	0.77	0.74	0.78	0.79	0.73
		Recall	0.37	0.41	0.45	0.36	0.44
	LogitBoost	Precision	0.50	0.29	0.28	0.37	0.36
		Recall	0.12	0.10	0.17	0.12	0.30
Bayesian	Bayes Net	Precision	0.77	0.71	0.77	0.78	0.69
		Recall	0.32	0.36	0.42	0.33	0.44

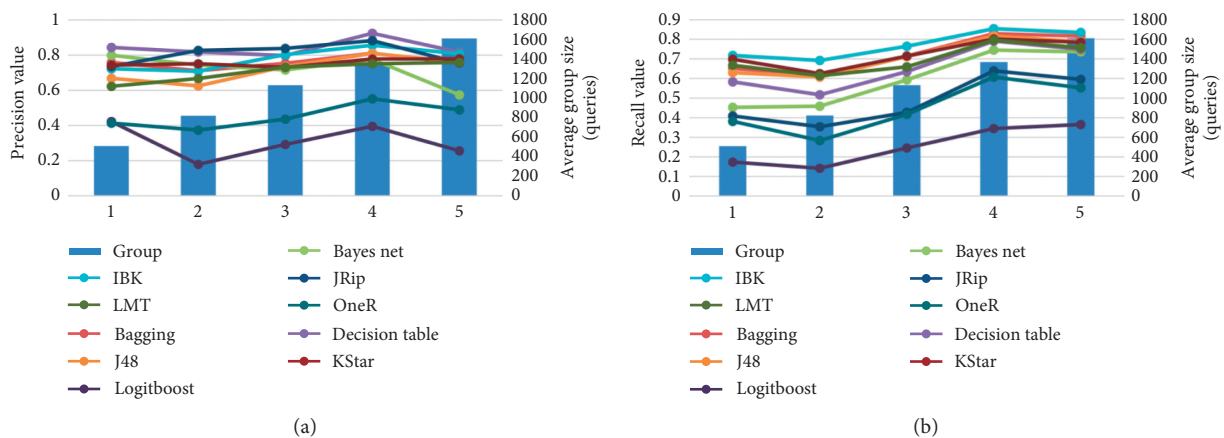


FIGURE 6: Clean dataset's groupwise precision and recall in different groups. (a) Clean dataset precision. (b) Clean dataset recall.

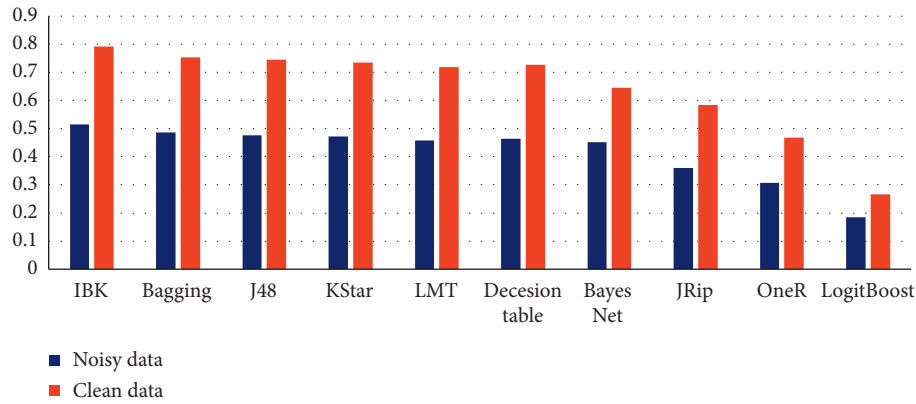


FIGURE 7: Average F-measure of all selected classification algorithms for noisy and clean datasets.

TABLE 8: Precision and recall of clean dataset in different groups.

Group			Group 1	Group 2	Group 3	Group 4	Group 5
Tree-based	J48	Precision	0.66	0.62	0.73	0.80	0.76
		Recall	0.62	0.60	0.71	0.81	0.78
	LMT	Precision	0.62	0.66	0.73	0.75	0.75
		Recall	0.66	0.61	0.65	0.79	0.75
Rule-based	Decision Table	Precision	0.84	0.81	0.79	0.92	0.81
		Recall	0.58	0.51	0.63	0.79	0.74
	JRip	Precision	0.73	0.82	0.83	0.88	0.75
		Recall	0.40	0.35	0.42	0.63	0.59
	OneR	Precision	0.41	0.37	0.43	0.55	0.48
		Recall	0.38	0.28	0.41	0.60	0.55
Lazy learner	IBK	Precision	0.72	0.70	0.80	0.85	0.80
		Recall	0.71	0.69	0.76	0.85	0.83
	KStar	Precision	0.74	0.75	0.73	0.77	0.77
		Recall	0.69	0.62	0.71	0.80	0.78
Metaheuristic	Bagging	Precision	0.75	0.71	0.75	0.81	0.75
		Recall	0.65	0.61	0.71	0.82	0.81
	LogitBoost	Precision	0.42	0.17	0.29	0.39	0.20
		Recall	0.19	0.14	0.23	0.34	0.38
Bayesian	Bayes Net	Precision	0.79	0.74	0.71	0.77	0.57
		Recall	0.45	0.45	0.59	0.74	0.73

## 5. Conclusions

Health information has been regarded as sensitive private information since ancient times. However, WSE collects this information for selling and targeted advertisements, which can infringe user's privacy. This paper presents QuPiD attack: a machine learning-based attack that quantifies the level of protection provided by popular PIR protocol UUP. The QuPiD attack uses a classification algorithm and the history of the user to classify an incoming query. We used two subsets (noisy and clean datasets) of real-world web data to test the proposed model. We showed that our proposed attack succeeds in correctly associating incoming queries to their real originator at a high ratio. For the selection of the best classification algorithm, we conducted our experiments with ten classification algorithms from different families. J48 and LMT from the tree-based family, Decision Table, JRip, and OneR from rule-based family, IBK and KStar from lazy

learner family, Bagging and LogitBoost from metaheuristic family, and Bayes Net from Bayesian family were selected. The results showed that IBK is the most appropriate algorithm if the "categories" feature vector is used.

During the analysis of the noisy dataset, almost every algorithm showed irregular behavior with the variation in the training dataset size. However, analyzing the clean dataset, we found that when increasing the size of the training data while building the classification model, the testing data in terms of recall are improving. We, therefore, conclude that noise is one of the factors responsible for unsteady behavior. Our analysis shows that PIR protocols are vulnerable to machine learning attacks, even with the first-degree classification tags of queries. This situation is alarming for currently available PIR protocols. Any web search engine or even web service armed with a profile of the user can expose a targeted user. In the future, we are interested to assess the proposed attack from different

perspectives, such as the impact of group size, the number of queries in a session, user profile size, and others. Moreover, we are excited to explore the unsteady behavior of classification algorithms.

### Data Availability

The data used to support the findings of the study are available at <http://www.radiounderground.net/aol-data/>.

### Conflicts of Interest

The authors declare that they have no conflicts of interest.

### Acknowledgments

This research was funded by the Deanship of Scientific Research, King Abdulaziz University (KAU), Jeddah, Saudi Arabia.

### References

- [1] M. W. Ng, R. Smith, N. Wickramesinghe, P. J. Smart, and N. Lawrentschuk, "Health on the net: do website searches return reliable health information on hemorrhoids and their treatment?" *International Surgery*, vol. 102, no. 5-6, pp. 216–221, 2017.
- [2] S. Fox and M. Duggan, "Health online 2013," *Health*, pp. 1–55, Pew Research Center, Washington, DC, USA, 2013.
- [3] R. Khan, M. A. Islam, M. Ullah, M. Aleem, and M. A. Iqbal, "Privacy exposure measure: a privacy-preserving technique for health-related web search," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 6, pp. 1196–1204, 2019.
- [4] P. Thomas, B. Billerbeck, N. Craswell, and R. W. White, "Investigating searchers' mental models to inform search explanations," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 1, pp. 1–25, 2020.
- [5] H. Yoganarasimhan, "Search personalization using machine learning," *Management Science*, vol. 66, no. 3, pp. 1045–1070, 2020.
- [6] F. Long, K. Jerath, and M. Sarvary, "Leveraging information from sponsored advertising at online retail marketplaces," *Kenan Institute of Private Enterprise Research Paper*, no. 20-03, <https://ssrn.com/abstract=3516104>, 2019.
- [7] S. B. Mokhtar, A. Boutet, P. Felber, M. Pasin, R. Pires, and V. Schiavoni, "X-search: revisiting private web search using intel sgx," in *Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference*, pp. 198–208, Las Vegas, NV, USA, December 2017.
- [8] K. Hafner and M. Richtel, *Google Resists US Subpoena of Search Data*, New York Times, New York, NY, USA, 2006.
- [9] G. Pass, A. Chowdhury, and C. Torgeson, "A picture of search. Infoscale' 06, Hong Kong," in *Proceedings of the 1st International Conference on Scalable Information Systems*, ACM, New York, NY, USA, 2006.
- [10] I. Lundberg, A. Narayanan, K. Levy, and M. J. Salganik, "Privacy, ethics, and data access: a case study of the fragile families challenge," 2018, <https://arxiv.org/abs/1809.00103>.
- [11] L. Edelstein, "The hippocratic oath: text, translation and interpretation," in *Ancient Medicine: Selected Papers of Ludwig Edelstein*, pp. 3–63, Johns Hopkins Press, Baltimore, MD, USA, 1943.
- [12] T. Libert, "Privacy implications of health information seeking on the web," *Communications of the ACM*, vol. 58, no. 3, pp. 68–77, 2015.
- [13] M. Barbaro, T. Zeller, and S. Hansell, *A Face is Exposed for AOL Searcher No. 4417749*, New York Times, New York, NY, USA, 2006.
- [14] V. Toubiana, L. Subramanian, and H. Nissenbaum, "Trackmenot: enhancing the privacy of web search," 2011, <https://arxiv.org/abs/1109.4677>.
- [15] A. Arampatzis, G. Drosatos, and P. S. Efraimidis, "Versatile query scrambling for private web search," *Information Retrieval Journal*, vol. 18, no. 4, pp. 331–358, 2015.
- [16] R. Dingleline, N. Mathewson, and P. Syverson, *Tor: the Second-Generation Onion Router*, Naval Research Lab, Washington, DC, USA, 2004.
- [17] C. Romero-Tris, J. Castellà-Roca, and A. Viejo, "Distributed system for private web search with untrusted partners," *Computer Networks*, vol. 67, pp. 26–42, 2014.
- [18] C. Romero-Tris, A. Viejo, and J. Castellà-Roca, "Multi-party methods for privacy-preserving web search: survey and contributions," in *Advanced Research in Data Privacy*, pp. 367–387, Springer, Berlin, Germany, 2015.
- [19] K. Stokes and M. Bras-Amorós, "Optimal configurations for peer-to-peer user-private information retrieval," *Computers & Mathematics with Applications*, vol. 59, no. 4, pp. 1568–1577, 2010.
- [20] M. Ullah, M. A. Islam, R. Khan, M. Aleem, and M. A. Iqbal, "ObSecure logging (OSLo): a framework to protect and evaluate the web search privacy in health care domain," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 6, pp. 1181–1190, 2019.
- [21] A. Petit, *Introducing Privacy in Current Web Search Engines*, Université de Lyon, Lyon, France, 2017.
- [22] A. Gervais, R. Shokri, A. Singla, S. Capkun, and V. Lenders, "Quantifying web-search privacy," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 966–977, Scottsdale, AZ, USA, November 2014.
- [23] S. T. Peddinti and N. Saxena, "On the privacy of web search based on query obfuscation: a case study of TrackMeNot," in *Proceedings of the International Symposium on Privacy Enhancing Technologies Symposium*, pp. 19–37, Berlin, Germany, July 2010.
- [24] S. T. Peddinti and N. Saxena, "On the effectiveness of anonymizing networks for web search privacy," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, pp. 483–489, Hong Kong, China, March 2011.
- [25] A. Petit, T. Cerqueus, A. Boutet et al., "SimAttack: private web search under fire," *Journal of Internet Services and Applications*, vol. 7, no. 2, 2016.
- [26] R. Khan and M. A. Islam, "Quantification of PIR protocols privacy," in *Proceedings of the 2017 International Conference on Communication, Computing and Digital Systems (C-CODE)*, pp. 90–95, Islamabad, Pakistan, March 2017.
- [27] R. Khan, M. Ullah, and M. A. Islam, "Revealing pir protocols protected users," in *Proceedings of the 2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, pp. 535–541, Dublin, Ireland, August 2016.
- [28] J. R. Quinlan, *C4. 5: Programs for Machine Learning*, Elsevier, Amsterdam, Netherlands, 2014.
- [29] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine Learning*, vol. 59, no. 1-2, pp. 161–205, 2005.
- [30] R. Kohavi, "The power of decision tables," in *Proceedings of the European Conference on Machine Learning*, pp. 174–189, Heraclion, Greece, April 1995.
- [31] W. W. Cohen, "Fast effective rule induction," in *Machine Learning Proceedings 1995*, pp. 115–123, Elsevier, Amsterdam, Netherlands, 1995.
- [32] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, no. 1, pp. 63–90, 1993.
- [33] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991.



- [34] J. G. Cleary and L. E. Trigg, "K\*: an instance-based learner using an entropic distance measure," in *Machine Learning Proceedings 1995*, pp. 108–114, Elsevier, Amsterdam, Netherlands, 1995.
- [35] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [36] E. Frank, M. Hall, G. Holmes et al., "Weka-a machine learning workbench for data mining," in *Data Mining and Knowledge Discovery Handbook*, pp. 1269–1277, Springer, Berlin, Germany, 2009.
- [37] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2/3, pp. 131–163, 1997.
- [38] N. Midha and V. Singh, "Classification of E-commerce products using RepTree and K-means hybrid approach," in *Big Data Analytics*, pp. 265–273, Springer, Berlin, Germany, 2018.

## Review Article

# A Systematic Review of Healthcare Big Data

**Rakesh Raja** , **Indrajit Mukherjee**, and **Bikash Kanti Sarkar**

*Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, India*

Correspondence should be addressed to Rakesh Raja; [rajarakeshchauhan@gmail.com](mailto:rajarakeshchauhan@gmail.com)

Received 24 December 2019; Revised 14 March 2020; Accepted 20 June 2020; Published 13 July 2020

Academic Editor: Shaukat Ali

Copyright © 2020 Rakesh Raja 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.

Over the past decade, data recorded (due to digitization) in healthcare sectors have continued to increase, intriguing the thought about big data in healthcare. There already exists plenty of information, ready for analysis. Researchers are always putting their best effort to find valuable insight from the healthcare big data for quality medical services. This article provides a systematic review study on healthcare big data based on the systematic literature review (SLR) protocol. In particular, the present study highlights some valuable research aspects on healthcare big data, evaluating 34 journal articles (between 2015 and 2019) according to the defined inclusion-exclusion criteria. More specifically, the present study focuses to determine the extent of healthcare big data analytics together with its applications and challenges in healthcare adoption. Besides, the article discusses big data produced by these healthcare systems, big data characteristics, and various issues in dealing with big data, as well as how big data analytics contributes to achieve a meaningful insight on these data set. In short, the article summarizes the existing literature based on healthcare big data, and it also helps the researchers with a foundation for future study in healthcare contexts.

## 1. Introduction

The era of big data has opened the door in the healthcare industry as a response to the digitization of healthcare data. Over the past decade, the exponential growth in data [1] has introduced a new domain called big data within the field of information technology (IT) and data science. The term big data is commonly used to describe a large amount of data which are too big and not easy to handle using traditional techniques of the database management system. The idea of big data is not very new, but the manner in which it is characterized is continuously changing. In 1997, Michael Cox and David Ellsworth introduced the term “big data” for the first time in the world during a paper conferred at an IEEE conference to explain the visual representation of data and the difficulties it exhibits to computer systems [2]. The data that go beyond the processing capacity of traditional database management systems are termed as big data. These data are so large that they do not fit the structure of typical database management systems.

The notion of big data given by Doug Laney was characterized by volume, velocity, and variety known as 3Vs [3]. Generally, big data can be defined as a collection of very

large amount of data with a wide range of types, making it very hard to process using conventional database management systems. As per the author in [4], big data is a data set with large volume, high speed, and high diversity that requires a new style of processing to facilitate decision-making and exploring knowledge and optimization of techniques. Typically, a massive volume of data may be referred to as big data when capturing, analysing, and visualizing of data with current technologies are overwhelming. Big data plays an important role in the current digital era due to the significant advancement of healthcare technologies [5]. As the sources of big data concerned in healthcare industries and various sectors are well known for their volume and diversity, hence, the healthcare domain gained its effect through the impact of big data. The healthcare industries have generated enormous amount of healthcare data over the past couple of years. These healthcare data are similar to the big data in terms of their characteristics, therefore named as healthcare big data. Healthcare data generally incorporate electronic medical records (EMRs) such as patient’s medical history, physician notes, clinical reports, biometric data, and other medical data related to health. All these data together result in healthcare big data. The evolution of healthcare big data is

advance and cost-effective for both public and private healthcare. The success of healthcare applications with regard to big data entirely relies upon the underlying architecture and use of suitable tools as proven in pioneering research efforts. It also gives an idea of the analytics of big data in healthcare systems. More specifically, big data analytical tools and techniques have the potential to improve the quality of medical services and reduce the medical cost of patients by exploring the association and understanding the nature of healthcare data. In 2016, Kohli et al. discuss how electronic health records (EHR) facilitate integration of patient health history for planning safe and proper treatment [6]. More about big data and healthcare big data definition are presented in Table 1.

## 2. Systematic Literature Review (SLR) Method

The purpose of the research process for conducting a systematic literature review (SLR) (based on the relevant articles and studies published in academic journals) focuses on the following objectives:

- Analysing different perspectives about the concept of big data in healthcare
- Exploring the origins of healthcare big data
- Identifying tools and techniques for healthcare big data analytics
- Highlighting the potential advantages and applications of big data in healthcare
- Drawing attention to overcome the big data challenges in healthcare

By discussing these goals in depth, the systematic review aims to assist in understanding the overall context of big data and its applications in the healthcare sector.

*2.1. Research Questions.* The following are the key research questions that are to be addressed for conducting the SLR of the proposed study:

- RQ1. What are the characteristics of big data in the healthcare domain?
- RQ2. What are the challenges and opportunities of healthcare big data?
- RQ3. What are the features of big data analytics in healthcare?
- RQ4. What techniques are used for big data analytics in healthcare?
- RQ5. What are the applications of big data analytics in healthcare?
- RQ6. What research has been pursued in healthcare big data since 2015?

*2.2. SLR Protocol.* Based on the SLR protocol designed in [12], this literature review follows the below mentioned guidelines.

*2.2.1. Search Strategy.* The two main electronic research databases: ScienceDirect and IEEE Xplore, were used to search for the collection of relevant articles related to the proposed research. However, some good and relevant works published by Springer publ. are also included in the present study.

*2.2.2. Search String.* The keywords defined by the authors for search process were “Big data,” “Healthcare,” and “Big data analytics” in context to the research domain. To conduct an SLR, the search process was carried out to identify the relevant articles for addressing the research questions based on predefined keywords using Boolean operators.

*2.2.3. Selection Criteria.* The authors agreed to select articles based on the following inclusion-exclusion criteria:

*(1) Inclusion Criteria*

- The articles relevant to healthcare big data and big data analytics
- The articles published during year 2015 to 2019
- The articles from journals publications only
- The articles written in the English language

*(2) Exclusion Criteria*

- The articles not in the range of 2015 to 2019
- The articles other than journal publications

*2.2.4. Study Selection Process.* The methodology for the literature review process was performed in different stages. The details of the study selection process of SLR are shown in Figure 1. Initially, all the articles relevant to big data, healthcare big data, and big data analytics were selected in the preliminary stage of screening as per the searching keywords. Based on inclusion-exclusion criteria, these articles were screened in the first stage, and irrelevant articles which were not published between 2015 and 2019 were excluded. During the second stage of screening, the selected articles were further screened on the basis of title, abstract, and keywords. The articles which were not associated with the proposed study were excluded. Finally, in the last stage of screening, these articles were further screened on the basis of abstract using the Boolean AND operator applied to all the three authors’ defined searching keywords. As a result, 34 articles relevant to the research domain were selected from 8355 articles, for further study by the authors.

*2.2.5. Quality Assessment.* During the review, quality assessment plays a significant role in the SLR protocol. The quality assessment of articles was done by all authors after the analysis and evaluation of abstracts of selected articles. These articles were selected with respect to each defined key research question based on inclusion-exclusion criteria.

TABLE 1: Big data and healthcare big data definition.

Sources	Definition
[7]	Healthcare big data can be defined as digitalized version of health information which is so vast and complex that they are not easy to manage using traditional software and/or hardware, nor can they be easily handled using conventional data management tools and methods
[8]	Big data means enormous amount of digital data that organizations and governments collect about individuals and their general surrounding environments where the generated data are about 2500 petabytes or even more
[9]	Big data in healthcare refers to the data sets with $\log(n \times p) \geq 7$ , and that they have high variety and high-speed characteristics
[10]	Big data can be defined as wealth of information described by massive volume, high velocity, and wide variety in order to have specific technology and analytical techniques to transform it into worth
[11]	Healthcare data generally incorporates electronic health records (EHRs) such as patient’s medical history, physician notes, clinical reports, biometric data, and other medical data related to health, as well as social media posts such as blog posts, tweets, Facebook postnotifications, and publications in medical journals

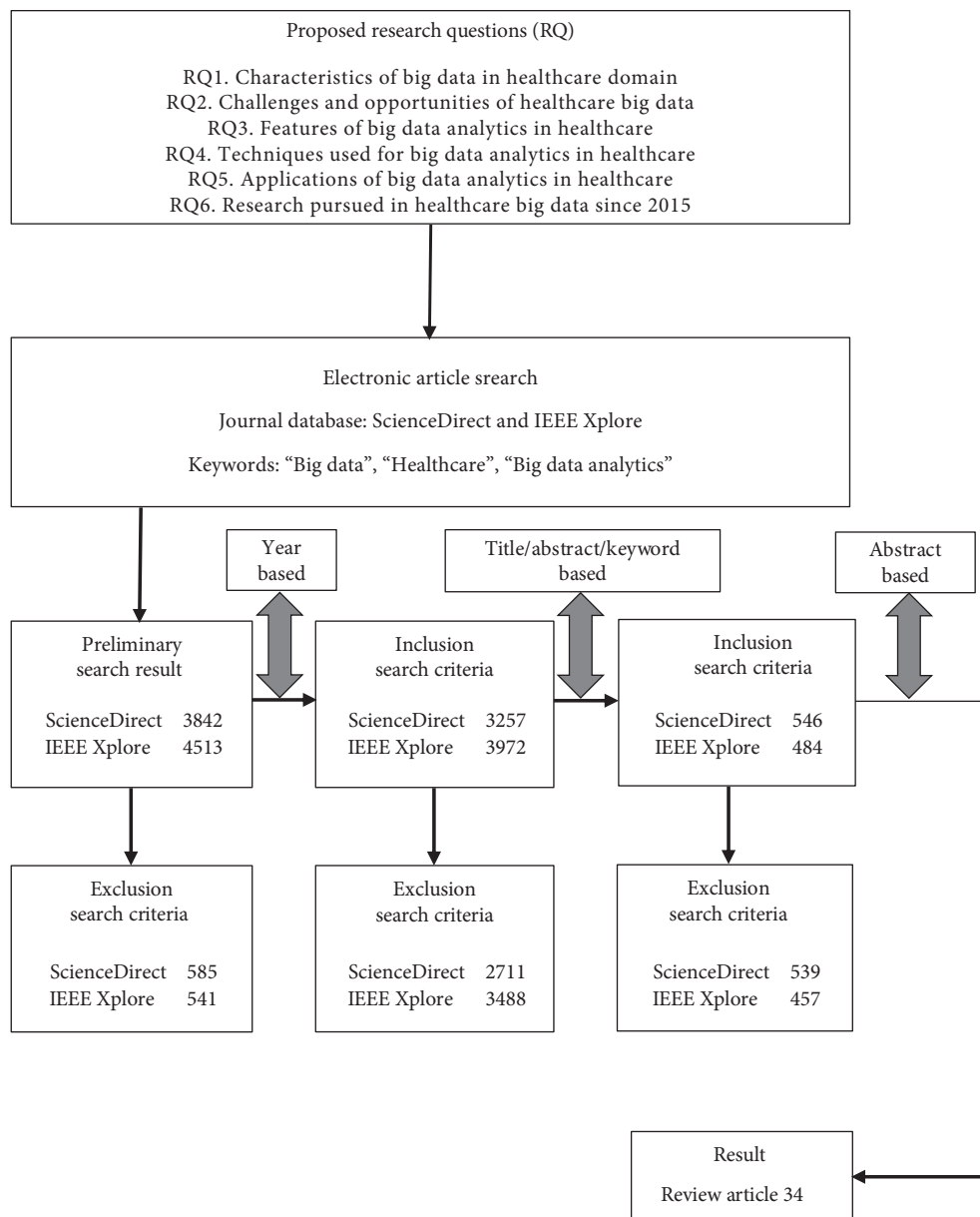


FIGURE 1: SLR process.

**2.2.6. Results and Discussion.** During the SLR process of the proposed research article, a collection of review articles related to defined research questions based on authors' defined search string (keywords) were identified by performing a search operation on the two most common electronic databases: ScienceDirect and IEEE Xplore. Around 7699 articles were filtered for the years 2015–2019 from the preliminary stage. Based on the title, abstract, and keywords, a total of 1030 articles were selected in the next stage. All of these articles were finally screened on the basis of the abstract using the Boolean AND operator applied to all three searching strings (keywords). As a result, 34 articles with respect to each defined research question were selected for further study by the authors according to the inclusion-exclusion criteria.

Table 2 shows the three screening stages of articles. Based on the main research objectives, the contents from these articles were extracted, and the proposed research article was organized into different sections: comprehensive overview of big data in the healthcare domain, sources of healthcare big data, challenges of big data in healthcare, big data analytics in healthcare, and application and potential benefits of big data in healthcare.

**2.3. Trend of Big Data Research in Healthcare Domain.** With the rapid growth of data, big data has given researchers an exposure to utilize it in more noticeable manner for decision-making in several healthcare applications. The trend of big data in the field of healthcare domain for the year 2015–2019 is described in Figure 2 with respect to Tables 2 and 3 of the revised version of the article. Figure 2 shows the increasing tendency of doing innovative research studies (published in reputed journals) in the area of healthcare big data.

### 3. Big Data: A Comprehensive Overview

**3.1. Big Data in General.** Big data refers to a collection of extensive and complicated data sets that are hard to handle using conventional database systems. As per the *zdnet.com*, big data pertains to the tools and techniques that allow an organization to generate, exploit, and maintain vast amounts of data with storage facilities. Each one of us is continuously producing enormous amount of data. And, big data is being generated by every computerized system as well as social networking sites. It is transmitted by the digital system, sensor devices, cameras, handheld devices, smartphones, and their applications [13]. Big data arrives at an unprecedented rate, large data size, and greater diversity from various sources. To extract significant worth from such large amount of data, we need high computational power, analytical capabilities, and expertise. This explosion of data attempts to change the opinion of people to think about everything in terms of big data. In recent times, transactional data, web-based data, sensor data, and electronic medical data keep developing with rapid speed. These data can be classified into web-based data, sensor-based data, demographic data,

transactional data, and machine-generated data [14] (as stated below):

Web-based data are acquired from social networking sites such as Facebook, Twitter, and Blogs

Machine-generated data are extracted from sensor-based devices and other gadgets

Transactional data are retrieved from biometrics, vital sign, radiology, and other medical images

Human-generated data comprise E-mails, doctor's prescriptions, and digitalized version of medical reports

This remarkable development in data growth has led to this new concept known as big data. In article [15], it is stated that big data is a complex set of data that has a significant impact on the ability of conventional data warehouses to store, maintain, perform, and analyse data. A formal definition of big data has been provided in [10]. It is stated there as follows: big data is a wealth of information described by huge quantity, high velocity, and wide variety in order to have specific technology and analytical techniques to transform it into worth. Looking at it another way, the McKinsey Global Institute defines big data as data sets whose size exceeds the capability of conventional database systems to collect, store, maintain, and analyse data. According to the authors in [16], big data is the assemblage of data collected from different sources such as corporate databases, websites, maps, movies, and public databases.

**3.2. Characteristics of Big Data.** The common characteristics of big data are illustrated in the following:

**Volume:** this implies data size usually measured in terabytes (TB =  $10^{12}$  bytes), petabytes (PB =  $10^{15}$  bytes), and zettabytes (ZB =  $10^{21}$  bytes), and so forth

**Velocity:** this indicates the rate of generation of data

**Variety:** this refers to the nature of data which big data can include such as structured, semistructured, and unstructured data

**Veracity:** this refers to the trustworthiness of the data

**Value:** the term itself is related to the worth of data being extracted

Apart from the abovementioned features of big data, several researchers and scientists have introduced new features to big data due to various applications available; i.e., the big data definition keeps changing according to the advancement of technology, data storage, and data transmission rate, as well as other system capabilities. The different explanations for the definition of big data are from 3Vs to 4Vs [17, 18], 5Vs [19], and 10Vs [20]. In particular, these dimensions are expanding as time goes by; and we currently have 42 distinct dimensions for big data till 2017 as per [21], and also the dimensions will keep on expanding as the big data evolves further. Figure 3 describes the generic notion of big data.

TABLE 2: Screening stages of research articles.

Electronic search		Article selection based on search string			Result
		Big data	Big data, healthcare	BDA	
Journals	ScienceDirect	2852	642	348	3842
	IEEE Xplore	4080	129	304	4513
<i>First screening based on year (2015–2019)</i>					
Journals	ScienceDirect	2424	543	290	3257
	IEEE Xplore	3578	123	271	3972
<i>Second screening based on title/abstract/keywords</i>					
Journals	ScienceDirect	464	34	48	546
	IEEE Xplore	429	32	23	484
<i>Final screening based on abstract</i>					
Journals	ScienceDirect		7		34
	IEEE Xplore		27		

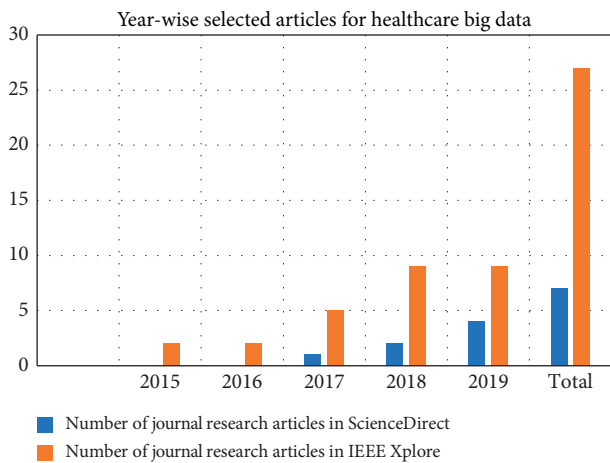


FIGURE 2: Trend of healthcare big data research from 2015 to 2019.

TABLE 3: Number of research articles year-wise.

Year	Number of journal research articles	
	ScienceDirect Articles	IEEE Xplore Articles
2015	0	2
2016	0	2
2017	1	5
2018	2	9
2019	4	9
Total	7	27
Outline	Studies which discuss the significance of BDA and the usefulness of big data in the field of healthcare	

3.3. *Big Data Definitions.* Big data and healthcare big data definitions are given in Table 1.

### 4. Big Data in Healthcare Domain

4.1. *Healthcare Big Data.* A pioneering renovation is taking place in the healthcare industry. The healthcare industry is generating a large volume of healthcare data due to the advancement in technology and digitization of medical records. In recent years, health information technology

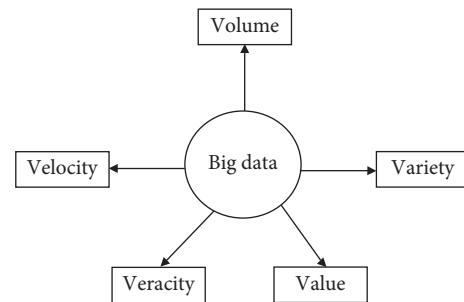


FIGURE 3: Generic concept of big data.

(HIT) has developed the power to generate, store, and transmit data electronically worldwide within seconds and also has the potential to deliver tremendously better productivity and service quality to healthcare. It allows each stakeholder in healthcare sectors to possess his/her own database of patients’ medical records in a digital form. The healthcare sectors have produced huge amounts of healthcare data by keeping records, consent and regulatory requirements, and patient care [11]. All these data together form healthcare big data. To be more specific, healthcare big data can be defined as electronic medical records (EMRs) which incorporates patient’s medical history, physician notes, clinical reports, biometric data, and other medical data related to health, as well as social media posts such as blog posts, tweets, Facebook postnotifications, and publications in medical journals [11]. Importantly, the exponential growth of healthcare data is another major issue in the current healthcare information systems (HISs). This transformation is not only about the large volume of healthcare data; however, we are also experiencing an exponential rise in the velocity at which these data are generated, as well as large diversity of medical data.

The evolution of advancement in technologies like sensor systems, cameras, and smartphones is a significant source of healthcare data. Everyday new sources of data are introduced. This makes it much more difficult to process or analyse big data in healthcare using common database management tools. Typically, when massive volume of healthcare data are captured, stored, and analysed properly in order to gain insight, it will enhance the healthcare service

outcome through smart decisions and also reduce healthcare costs. However, effective data analytical tools and techniques as well as powerful computing systems are required for this purpose. Healthcare big data analytics (BDA) in particular has started to emerge as a promising tool for taking care of issues in numerous healthcare disciplines. In addition, the role of a data analyst is to mine the big data, exploring the association and understanding trends and patterns of healthcare data. This enhances the health and improves the quality of life of an individual, as well as provides appropriate early-stage treatment at low cost.

The amount of data stored in healthcare sectors continued to increase curiosity about healthcare big data. There is an enormous amount of data ready to be analysed. One of the principle motivations behind big data is to focus on healthcare. The basic motive of nations around the world is to improve the healthcare facilities and decrease the medical costs. However, the revolution of massive volume of data in healthcare remains a barrier for achieving this goal. Electronic healthcare data from all around the world were estimated at 500 petabytes in 2012, reaching 25 petabytes by 2020 [22]. Thus, healthcare can be described as a wide variety of services offered by medical professionals to people, families, or societies to encourage, maintain, or restore better health. The quality of the healthcare system is significant because it determines hospital sustainable growth and helps people to maintain the optimal state of health. In certain cases, the quality of healthcare services is too high, and it ends up costly for patients. Consequently, it is essential to address the key healthcare procedures and related quality parameters that act in collaboration to ensure the best possible outcomes for patients and reduce the healthcare costs.

**4.2. Sources of Healthcare Big Data.** This section deals with several important sources of healthcare data. Big data in healthcare can revolutionize the medical field through early-stage disease detection using adequate analytical tools and techniques by incorporating and analysing health-related information in a comprehensive manner. Currently, the evolution of advancement in technologies like sensor systems, cameras, wearable devices, and mobile applications is widely used in the domain of the medical field [23, 24]. As a result, more medical information is being explored in a consistent manner. Data in medicinal services are fragmented and dispersed, originating from disparate sources with multiple formats [25]. The facts confirm that information on health is large and heterogeneous. The reason is on the ground that they originate from various internal and external sources accessible at multiple locations. External sources include web data, social media data, and machine-generated data, and internal sources include transactional data, biometric data, and human-generated data. Various healthcare data and their sources are summarized in Table 4.

**4.3. The 5Vs of Healthcare Big Data Characteristics.** In this section, the important Vs about healthcare data are briefly stated. The five key characteristics that have been found in

most literature [12, 35] to define healthcare big data are as follows:

*Volume.* Based on the general discussion of big data, healthcare data are a perfect case of big data. The volume refers to the data size that grows exponentially day to day, and by 2020, the volume of big data may reach to 44 zettabytes [36]. Compared to most of the industries, the healthcare sector generates massive amounts of data in the form of electronic medical records (EMRs), biometric data, clinical data, radiology images, genomics, etc. All these data collectively form healthcare big data [37–39]. Obviously, the utilization of several tools such as Hadoop, MapReduce, and MongoDB is getting more popular among healthcare organizations due to their ability to store and measure massive volume of data [40, 41].

*Velocity.* Velocity refers to the speed at which data are generated, as well as data acquired from various healthcare systems [42].

*Variety.* Variety refers to the heterogeneity and diversity of data. The healthcare industry generates and collects data at a staggering rate from different sources such as social networking sites, sensor devices, cameras, and smartphones. However, these healthcare data may be in any one of the forms, structured, unstructured, or semistructured. Example of structured data is clinical data, whereas data such as physician notes, images, social media data, mobile data, and radiograph films are unstructured or semistructured. Figure 4 depicts the types of healthcare data, along with examples.

*Veracity.* The veracity characteristics of healthcare data refer to the trustworthiness of the data, which in this context is equivalent to quality assurance of data. It gives the degree of authenticity about healthcare knowledge.

*Value.* Value is the most important and distinctive characteristics of all the 5Vs of healthcare big data, as it has the ability to transform healthcare data into worth of information. Its concept is exactly in line with that of healthcare data.

## 5. Big Data Challenges in Healthcare

The evolution of big data introduces several challenges, constraints, and problems due to exponential growth of healthcare data. Big data is constantly changing, and this change of data presents a lot of challenges in storing, analysing, and retrieving the massive volume of data. Certainly, the conventional database systems could not be used to store, process, and extract the information due to its massive size and diversity of data.

The main challenges encountered by healthcare BDA are as follows:

- Quality and storage of data
- Data analysis of good quality
- Expertise in data analytics

TABLE 4: Important sources of healthcare data.

Healthcare data	Features	Sources
Clinical data	Data within electronic health records (EHRs) can be either structured (e.g., EMRs and clinical data), unstructured (e.g., clinical trials data), or semistructured (e.g., claims data) forms	[26]
Patient-generated data	Biometric data, social media data, online data (e.g., blogs, Facebook posts, and Twitter)	[27]
Sensor data	Data produced by sensor-based devices (e.g., vital signs, ECG, and handheld devices)	[28–30]
Genomic data	Gene typing (e.g., gene expression and DNA sequence)	[11, 31, 32]
Clinical research data	Health product data (e.g., drug information)	[33]
External data	Insurance data (e.g., financial data) Biometric data (e.g., fingerprints)	[34]

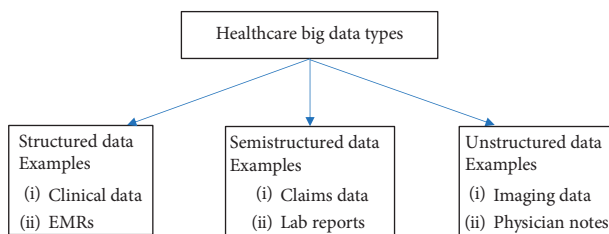


FIGURE 4: Types of healthcare big data.

Data security and confidentiality

Multiple sources of data

Healthcare big data challenges encountered are no different. Big data characteristics are the main issues that need to be addressed. It is vital to move towards big data technology in order to provide better medical facilities. Big data technology, however, introduces a potential risk to certain categories.

5.1. *Issues in Healthcare Big Data.* Big data issues that generally occur in the healthcare organizations are covered by four main categories [35, 43]:

*Data Governance.* Data management and regulation is the governance of data. As the healthcare sector moves towards data analytics, data governance is a major challenge. Healthcare data generated are diversified in nature, requiring standardization and governance.

*Economic Challenges.* The facilities in the medical field between patients and healthcare professionals throughout clinical visits depend on the paid service. Subsequently, advancement in technologies associated with this process places a burden on the medical community and generates an unnecessary impact for the personnel against such unpaid services.

*Big Data Technology Challenges.* Big data in healthcare is enormous and highly fragmented which causes problems in quality of information, as well as technology-wise, big data creates a barrier to accomplish the healthcare vision [44].

*Security and Privacy Issues.* In the era of big data, the privacy of healthcare data must be seriously considered due to the potentially sensitive information about individual healthcare stakeholders. Healthcare data are highly sensitive data which must be secured from unauthorized access so that it cannot be made publicly available, as well as healthcare fraud can also be prevented from attackers. Therefore, data security is one of the most important challenging tasks in the healthcare domain.

While studying and analysing several published research papers with reference to the SLR protocol, this research focuses on how recent developments in ICT (information and communication technologies) together with big data techniques can be effectively incorporated to address these challenges of healthcare big data and make a significant contribution towards healthcare services [45–51]. Based on [17, 52], we the authors classify healthcare BDA into three categories, namely, descriptive analytics, predictive analytics, and prescriptive analytics. Among these different BDA techniques, this literature review reflects that there are various tools, for example, Hadoop [53] and MapReduce [54] that have been developed for healthcare big data management. These are described in Section 6. A few of the well-known BDA techniques used in the areas of healthcare are described in Table 5. The categories provided in Table 5 are drawn from the literature [12, 66–68].

6. Big Data Analytics in Healthcare

Healthcare BDA has a potential to improve the quality of care and reduce the medical cost of patients by finding the associations from massive volume of healthcare data, thereby offering a wider perspective of clinical expertise based on medical evidence and various tests. Advanced analytical tools and techniques used in healthcare systems provide services that satisfy a growing need and enable healthcare agencies to process massive volume of data, analyse it in real time, and extract knowledge from medical records of all patients. In 2017, Palanisamy and Thirunavukarasu have presented various analytical avenues that exist in the patient centric healthcare system from the perspective of various stakeholders [69]. The main goal of the article is to assist researchers and data scientists to make



TABLE 5: BDA techniques.

BDA techniques	Healthcare application areas	Examples	Sources
Machine learning	Early detection of diseases	Predicting epidemics, disease monitoring	[12, 55]
Data mining	Prediction of heart disease at early stages	Health analytics, determination of epidemics	[56–59]
Neural network	Diagnosis of chronic diseases	Predicting of patients' future disease, patient safety	[60–62]
Pattern recognition	Improvement of public health disease surveillance	Empowering public health, health literacy, improving quality of care	[63, 64]
NLP	Improve the quality of care and accuracy of clinical decisions	Cost reduction, high-risk factor identification	[57, 65]

informed healthcare decision and enhance the performance of the healthcare centre, so that people live a healthier lifestyle. In particular, this includes numerous analytical techniques such as machine learning, pattern recognition, statistical analysis, visualization, and data mining to interpret feature relationships and discover knowledge. BDA is based on the concept of data mining that incorporates various analytical techniques to evaluate and explore large volume of data to extract significant and useful information. Researchers may find ample information about BDA and healthcare from the articles [66, 70–72].

*6.1. Types of Healthcare Big Data Analytics.* BDA mainly perform three types of analytics, namely, *descriptive analytics*, *predictive analytics*, and *prescriptive analytics*. The descriptive analytics facilitates to explore insights and allows healthcare practitioners to understand what is happening in a given situation [73, 74]. In the context of healthcare data, the descriptive analytics analyses the data gathered in order to interpret, understand, summarize, and visualize significant health-related information. On the other hand, predictive analytics assist healthcare stakeholders to identify the healthcare services and responding appropriately according to the requirements of patients. It also enables clinicians to be capable of making patient-related decisions on the basis of system predictions [73, 74]. Predictive analytics involves various statistical techniques used to analyse and extract valuable insights from big data [17]. Hadoop/MapReduce is one of the most widely used techniques to develop a predictive model for healthcare systems. Prescriptive analytics is comparatively a modern type of analytics that combines descriptive and predictive analytics [75]. Though predictive analytics recommends what will happen in the future, prescriptive analytics provides the best course of action to be taken by healthcare providers in the future [73, 74]. By incorporating clinical and genomic data, prescriptive analytics continuously repredicts the healthcare services and improves the predictive accuracy in order to provide more appropriate diagnoses and treatments for healthcare providers [76, 77].

The medical industry is flooded with enormous volumes of data that require validation and analysis. BDA has a power and capability to perform essential computing and analytical ability to process large volumes of healthcare data. It facilitates medical professionals, clinical researchers, and healthcare stakeholders to improve their results through the use of their internal and

external sources of big data [78, 79]. As per the healthcare providers, the assessment of patient data, which incorporates patient medical history (EHR), doctors' prescriptions, diagnostic reports, biometric data, clinical tests, and other medical data related to health, assists them to follow the advancement of a recommended course of treatment and interrupt the course so that changes can be made if necessary. Thus, it helps to eliminate unnecessary visits and reduce readmission rates. Furthermore, the drug company and other medical organizations take benefit of analytical advantages in designing marketing strategies. Indeed, pharmaceutical industries can study their current market status by capturing and analysing the healthcare data such as sales record and interpretation of drug information prescribed by healthcare professionals for each patient and disease to develop the strategic goals. Therefore, the health insurance company can develop an appropriate health plan for every patient by analysing their demographic data, clinical trials, and statistical data related to health factors [69].

An enormous amount of data are accumulated in the healthcare sector from patients' medical histories, clinical trials, and diagnostic reports. Like healthcare big data, data analytics can be characterized by volume, velocity, and variety known as 3Vs [3, 17]. BDA is the use of advanced analytical techniques to analyse, extract, and discover meaningful patterns and insight from large data sets [80, 81]. BDA plays a crucial role in enhancing healthcare facilities and increases patients' clinical outcomes. It therefore has the ability to improve the quality of care and life styles and reduce medical costs. Based on the systematic review on the current state of big data research by Wang and Hajli, BDA in the context of healthcare can be characterized as the capability to acquire, store, process, and analyse large amounts of health data in different forms and provide meaningful information to users, which enables them to explore business values and insights in a timely manner [82].

*6.2. Necessity of Healthcare Big Data Analytics.* BDA in healthcare is needed to enhance the healthcare quality by taking the associated healthcare services into account:

*Provision of Personalized Healthcare.* Big data in healthcare can revolutionize the medical field through early-stage disease detection and reduce medical cost for the patients using appropriate analytical tools in a

comprehensive manner. This helps to develop a personalized healthcare system for healthcare stakeholders [83, 84].

*Early Detection of Spread of Diseases.* This concentrates on early prediction of viral (infectious) diseases (*i.e.*, before spreading) on the basis of social network analysis. More and more social media of the patients suffering from a disease in a specific geographical area are monitored to identify the development and spread of viral disease. This assists the healthcare experts to counsel the sufferers to take the necessary preventive action.

*Monitoring the Clinical Performance.* There is a lot of enthusiasm to evaluate clinical performance in order to screen and enhance the quality of healthcare services. The reform of the hospital is of major concern in the strategic plan of the healthcare sectors. This can be achieved by monitoring and setting up the hospital in accordance with medical council's standards.

**6.3. Big Data Analytical Techniques in Healthcare.** In the past, traditional technologies and data warehouses were used by the data analyst to store, process, and manage data. However, the revolution of massive volume of data in healthcare cannot be handled using conventional database systems, tools, and techniques. Nowadays, many advanced technologies with high computing power and storage capacity have been developed in order to address the low performance and difficulty of traditional systems. Accordingly, in [85], "big data technologies can be referred to as advanced technologies that have a high computing power and analytical ability to process large volumes of data collected from various sources to extract insight from it." As per the authors in [86, 87], big data techniques cover a wide range of fields such as machine learning, statistical analysis, and image analysis. A few of the well-known BDA techniques used in the areas of healthcare are shown in Table 5. The categories that are generated in Table 5 are taken from the literature [12, 66–68, 86]. Big data plays a significant role across all domains such as government organizations, trade associations, healthcare industries, education, and research and development. BDA also empowers the secondary use of clinical data in the healthcare sector [88]. Big data acceptance has shown enormous growth from 17 percent in 2015 to 53 percent in 2017 according to Forbes [89].

In the current digital era, healthcare is one of the sectors that generates a large volume of healthcare data, and these healthcare big data can be characterized by its volume, velocity, and variety known as 3Vs [3]. Data mining techniques can be applied on this massive amount of healthcare data so as to identify new interesting patterns and valuable insights for quality medical services. The Hadoop is an open source software framework for BDA in healthcare as well as the most popular implementation of the MapReduce programming model [90]. It allows distributed storage and processing of large variety of healthcare big data whether it is structured, semistructured, or unstructured such as patient's EHR, physician's notes, laboratory data, clinical trials

reports, and insurance data as compared to conventional database systems. Figure 5 shows a general conceptual architecture of big data analytics [7].

#### 6.4. Platform and Tools for Healthcare Big Data Analytics.

There are currently several techniques available for performing BDA. The few tools and techniques that support the Hadoop distributed platform are being discussed below [91, 92]:

*Hadoop Common.* It refers to the set of common utilities that assist other modules of the Hadoop framework. Hadoop Common is a fundamental part of the Apache platform in addition to the HDFS, YARN, and MapReduce. Hadoop Common is generally called as Hadoop Core.

*Apache HDFS.* HDFS refers to the Hadoop Distributed File System that can be used to process unstructured data on commodity hardware predominantly. HDFS is the primary data storage where each file is divided into blocks of fixed size and distributed across numerous servers (nodes). HDFS employs the master/slave architecture using NameNode (master node) and DataNode (slave node) [93, 94].

*Hadoop MapReduce.* MapReduce is a programming framework that enables us to process massive amount of data in parallel in a distributed computing environment. This framework consists of two main functions, namely, Map and Reduce that can effectively manage structured as well as unstructured data [95, 96]. As the name MapReduce indicates, reducer function occurs after the completion of the mapper function.

*Apache Hive.* Hive is a data warehouse framework designed to query and analyse huge amount of data stored in Hadoop HDFS. It is an ETL (extract, transform, and load) tool for the Hadoop ecosystem. Hive is built on top of the Hadoop platform and provides a declarative language similar to SQL known as the Hive query language (HiveQL) that enables SQL programmers to perform data analysis conveniently [97].

*Pig.* Apache Pig is a parallel computing framework that runs on the Apache Hadoop platform. Pig Latin is the language for this platform which is used for analysing large volumes of data due to its distributed architecture. In fact, Pig Latin is like the SQL language and is easy to learn. The main distinction is that Pig Latin can process semistructured and unstructured data [93, 98].

*HBase.* Apache HBase is an open source, multidimensional distributed database system in a Hadoop ecosystem. It runs on the top of the Hadoop Distributed File System (HDFS). HBase can store large volumes of data usually measured in terabytes (TB) to petabytes (PB) and does not support a structured query language like SQL; indeed, HBase employs a NoSQL approach.

*Mahout.* Apache Mahout is an open source distributed framework that supports BDA on the Hadoop platform and is designed for machine learning using the MapReduce program. The Apache Mahout enables us to

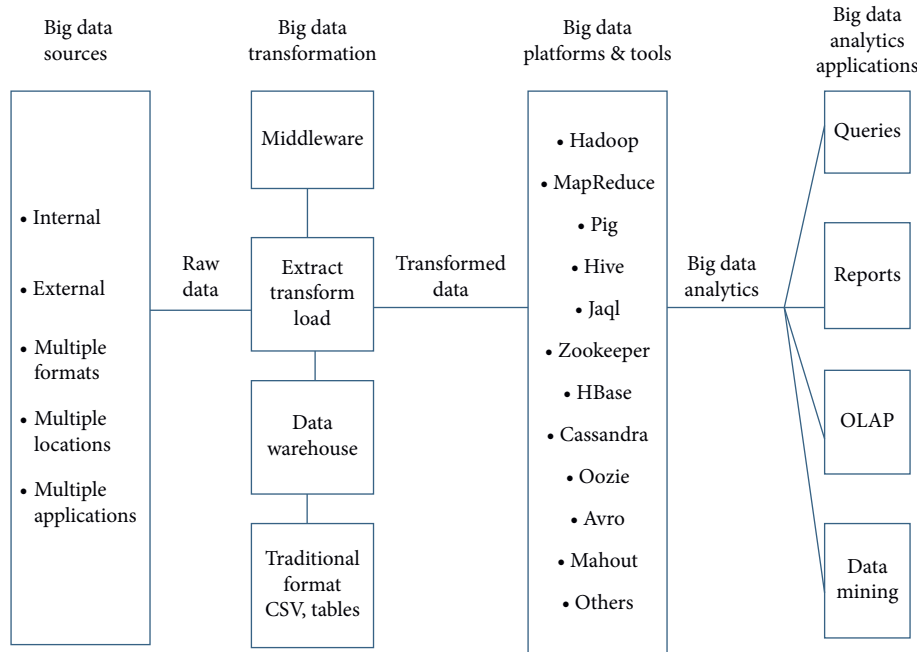


FIGURE 5: A conceptual architecture of big data analytics [7].

develop collaborative filtering, classification, clustering, association mining, and statistical algorithms related to machine learning with the help of data science techniques [93, 99].

For more about the tools and techniques, one may refer to [7, 53].

## 7. Big Data Benefits in Healthcare Sector

Healthcare sectors extending from a single physician's office to a large set of networks of healthcare service providers have a potential to acknowledge significant benefits by digitizing, integrating, and effectively using big data analytical tools and techniques in healthcare.

Based on the recently published studies [65, 66, 100], following are some of the major benefits:

**Clinical operations:** the information on healthcare helps to determine methods of diagnosing and treating patients that are more clinically important and cost-effective

**Patients:** healthcare information can help patients to make the right decision at the right time and improve patients' health while reducing the healthcare cost

**Healthcare providers:** the data acquired from medical organizations assist the stakeholders to develop new healthcare strategies for patients to minimize the unnecessary hospitalizations

**Research and development:** healthcare data support researchers and scientists to enhance healthcare services through more precise and appropriate treatments

**Public health:** healthcare data also assist to assess the health risks as well as analyses trends of diseases to enhance public health surveillance

## 8. Application of Big Data Analytics in Healthcare

The buzzword big data in the digital world is highly in demand in every sector especially in the field of healthcare. This has laid a foundation for various applications in the healthcare sector. Healthcare BDA has a potential to improve the quality of care and reduce the medical cost of patients by discovering the associations from massive volume of healthcare data, thereby offering a wider perspective of clinical expertise based on medical evidences and various tests [101]. Healthcare BDA also helps the clinicians and policy makers to develop public policy and service delivery based on open health prescribed data, disease prevalence data, and economic deprivation data [102]. As per the authors in [100, 101, 103, 104], the major areas for the applications of BDA in healthcare are as follows:

**Healthcare Monitoring.** Healthcare data analytics can be used to continuously monitor the health status of the users (patients) in order to enhance their lifestyle [93].

**Healthcare Risk Prediction.** A deep analysis of healthcare data helps healthcare stakeholders and medical practitioners to develop solutions for risk prediction. It also enables clinicians to be capable of making patient-related decisions on the basis of system predictions [73, 74]. Data analytics in healthcare can also be used to identify and manage high-risk and high-cost patients [105].

**Behavioural Monitoring.** Another prospective implementation of BDA in healthcare is monitoring of patients with abnormal behaviour [66]. In 2005, Nambu et al. proposed the home healthcare system to capture the behavioural data of patients for diagnosing their health conditions [106].

*Fraud Detection and Prevention.* One of the major and important application of data analytics in the healthcare sector is fraud detection and prevention. As per the authors in [107], data mining and machine learning techniques are mainly used for fraud detection in healthcare.

*Clinical Decision Support Systems.* In the medical field, clinical decision support systems are designed to facilitate healthcare professionals in making clinical decisions to diagnose diseases based on patient's health condition [108, 109].

*Personalized Healthcare Recommendation System.* Big data plays a significant role in the healthcare domain to develop a personalized recommendation system to give precise and relevant medical recommendation (advice) to an individual (patient) based on their current health status and medical history [110]. The authors in [111] proposed an intelligence-based health recommendation system using BDA to study and research health records of patients, assess risk and the severity of different diseases, and then provide recommendations based on outcomes of prediction. The authors in [112] suggested a clinical recommendation system that is beneficial for patients to access accurate recommendations based on their own health status.

*Drug Discovery and Clinical Trials.* Healthcare BDA is widely used by the pharmaceutical industry for drug discoveries so that it can help physicians, pharmaceutical developers, and other healthcare professionals for getting the right drug to the right patient at the right time [107, 113, 114].

*Image Informatics and Telediagnosis.* Imaging informatics is the study of methods for generating, managing, and representing imaging information in various biomedical applications. It is concerned with how medical images are exchanged and analysed throughout complex healthcare systems [115, 116]. The authors of the study [117] introduce a novel telemammography system for early detection of breast cancer with the help of image processing and machine learning techniques. Computer-aided diagnosis plays a significant role in medical imaging [118].

*Healthcare Knowledge System.* According to [119], a knowledge management system is developed based on healthcare big data in order to support clinical decision-making and disease diagnosis. The healthcare knowledge system is based on a variety of databases such as electronic health record (EHR), medical imaging data, and unstructured clinical notes and genetic data.

*Public Health Information.* As per [115, 120, 121], BDA in healthcare can also be used to track and monitor public health status for decision-making and policy development.

Based on the studies of different authors, it is revealed that the BDA in healthcare has a potential to improve the quality of healthcare, decreasing the readmission rates and

reducing the medical cost of patients by exploring the association and understanding the nature of healthcare data [7, 93, 122]. Furthermore, image processing, signal processing, and genomics are presently the three main areas for the application of data analytics in the healthcare domain [123].

## 9. Conclusion

This systematic review focuses on the existing literature to study healthcare big data based upon defined keywords and research aspects in the healthcare domain. The proposed research uses an SLR protocol and guidelines to review the systematic study of the past and the cutting-edge articles of the big data in healthcare. The purpose of an SLR protocol is based on the following objectives:

- Analysing different perspectives about the concept of big data in healthcare
- Exploring the origins of healthcare big data
- Identifying tools and techniques for healthcare big data analytics
- Highlighting the potential advantages and applications of big data in healthcare
- Drawing attention to overcome the big data challenges in healthcare

The present study will help the researchers with a useful base for future work to understand the overall context of healthcare big data and its applications. The limitation of the proposed research is that the electronic search process was performed in only two journal databases from 2015 to 2019, and the rest of the databases were skipped while accessing the quality of journal articles which can be addressed in future research.

## Data Availability

Data sharing is not applicable to this article as no data sets were generated or analysed during the current study.

## Conflicts of Interest

The authors declare that no conflicts of interest exist regarding this publication.

## References

- [1] D. P. Augustine, "Leveraging big data analytics and Hadoop in developing India's healthcare services," *International Journal of Computer Applications*, vol. 89, no. 16, pp. 44–50, 2014.
- [2] M. Cox and D. Ellsworth, "Application-controlled demand paging for out-of-core visualization," in *Proceedings. Visualization '97 (Cat. No. 97CB36155)*, pp. 235–244, IEEE, Phoenix, AZ, USA, October, 1997.
- [3] D. Laney, "3D data management: controlling data volume, velocity and variety," *META Group Research Note*, vol. 6, 2001.

- [4] C. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: a survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014.
- [5] X. Wang, Y. Wang, C. Gao, K. Lin, and Y. Li, "Automatic diagnosis with efficient medical case searching based on evolving graphs," *IEEE Access*, vol. 6, pp. 53307–53318, 2018.
- [6] R. Kohli, S. S. L. Tan, and S. S.-L. Tan, "Electronic health records: how can IS researchers contribute to transforming healthcare?" *MIS Quarterly*, vol. 40, no. 3, pp. 553–573, 2016.
- [7] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, p. 3, 2014.
- [8] M. Vivekanand and B. M. Vidyavathi, "Security challenges in big data," *International Journal of Advanced Research in Computer Science*, vol. 6, no. 6, 2015.
- [9] E. Baro, S. Degoul, R. Beuscart, and E. Chazard, "Toward a literature-driven definition of big data in healthcare," *BioMed Research International*, vol. 2015, Article ID 639021, 9 pages, 2015.
- [10] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of big data based on its essential features," *Library Review*, vol. 65, no. 3, pp. 122–135, 2016.
- [11] K. Priyanka and N. Kulennavar, "A survey on big data analytics in health care," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 4, pp. 5865–5868, 2014.
- [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. Sagioglu and D. Sinanc, "Big data: a review," in *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS)*, IEEE, San Diego, CA, USA, pp. 42–47, May 2013.
- [14] I. Lee, "Big data: dimensions, evolution, impacts, and challenges," *Business Horizons*, vol. 60, no. 3, pp. 293–303, 2017.
- [15] S. Tiwari, H. M. Wee, and Y. Daryanto, "Big data analytics in supply chain management between 2010 and 2016: insights to industries," *Computers & Industrial Engineering*, vol. 115, pp. 319–330, 2018.
- [16] G. Silahtaroglu and N. Yılmaztürk, "Data analysis in health and big data: a machine learning medical diagnosis model based on patients' complaints," *Communications in Statistics-Theory and Methods*, pp. 1–10, 2019.
- [17] A. Gandomi and M. Haider, "Beyond the hype: big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.
- [18] C. H. Lee and H.-J. Yoon, "Medical big data: promise and challenges," *Kidney Research and Clinical Practice*, vol. 36, no. 1, pp. 3–11, 2017.
- [19] R. Y. Zhong, S. T. Newman, G. Q. Huang, and S. Lan, "Big data for supply chain management in the service and manufacturing sectors: challenges, opportunities, and future perspectives," *Computers & Industrial Engineering*, vol. 101, pp. 572–591, 2016.
- [20] K. F. Tiampo, S. McGinnis, Y. Kropivnitskaya, J. Qin, and M. A. Bauer, "Big data challenges and hazards modeling," in *Risk Modeling for Hazards and Disasters*, pp. 193–210, Elsevier, 2018.
- [21] L. Wang and C. A. Alexander, "Big data in medical applications and health care," *American Medical Journal*, vol. 6, no. 1, pp. 1–8, 2015.
- [22] J. Sun and C. K. Reddy, "Big data analytics for healthcare," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13*, Chicago, IL, USA, 2013.
- [23] A. T. Lo'ai, R. Mehmood, E. Benkhelifa, and H. Song, "Mobile cloud computing model and big data analysis for healthcare applications," *IEEE Access*, vol. 4, pp. 6171–6180, 2016.
- [24] I. García-Magariño, R. Lacuesta, and J. Lloret, "Agent-based simulation of smart beds with internet-of-things for exploring big data analytics," *IEEE Access*, vol. 6, pp. 366–379, 2017.
- [25] W. B. Rouse and N. Serban, *Understanding and Managing the Complexity of Healthcare*, MIT Press, Cambridge, MA, USA, 2014.
- [26] S. Yang, M. Njoku, and C. F. Mackenzie, "Big data approaches to trauma outcome prediction and autonomous resuscitation," *British Journal of Hospital Medicine*, vol. 75, no. 11, pp. 637–641, 2014.
- [27] N. P. Terry, "Protecting patient privacy in the age of big data," *SSRN Electronic Journal*, vol. 81, p. 385, 2012.
- [28] R. B. Shrestha, "Big data and cloud computing," *Applied Radiology*, vol. 43, no. 3, p. 32, 2014.
- [29] A. Rizwan, A. Zoha, R. Zhang et al., "A review on the role of nano-communication in future healthcare systems: a big data analytics perspective," *IEEE Access*, vol. 6, pp. 41903–41920, 2018.
- [30] L. Carnevale, R. S. Calabrò, A. Celesti et al., "Toward improving robotic-assisted gait training: can big data analysis help us?" *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1419–1426, 2018.
- [31] P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "—Omic and electronic health record big data analytics for precision medicine," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 2, pp. 263–273, 2016.
- [32] F. Celesti, A. Celesti, J. Wan, and M. Villari, "Why deep learning is changing the way to approach NGS data processing: a review," *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 68–76, 2018.
- [33] K. Miller, "Big data analytics in biomedical research," *Biomedical Computation Review*, vol. 2, pp. 14–21, 2012.
- [34] S. C. Helm-Murtagh, "Use of big data by blue cross and blue shield of North Carolina," *North Carolina Medical Journal*, vol. 75, no. 3, pp. 195–197, 2014.
- [35] B. K. Sarkar, "Big data for secure healthcare system: a conceptual design," *Complex & Intelligent Systems*, vol. 3, no. 2, pp. 133–151, 2017.
- [36] M. S. Hajirahimova and A. S. Aliyeva, "About big data measurement methodologies and indicators," *International Journal of Modern Education and Computer Science*, vol. 9, no. 10, p. 1, 2017.
- [37] A. Widmer, R. Schaer, D. Markonis, and H. Müller, "Gesture interaction for content-based medical image retrieval," in *Proceedings of International Conference on Multimedia Retrieval*, pp. 503–506, April 2014, Glasgow, Scotland.
- [38] J. A. Seibert, "Modalities and data acquisition," in *Practical Imaging Informatics*, pp. 49–66, Springer, New York, NY, USA, 2009.
- [39] A. S. Panayides, M. S. Pattichis, S. Leandrou, C. Pitris, A. Constantinidou, and C. S. Pattichis, "Radiogenomics for precision medicine with a big data analytics perspective," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 5, pp. 2063–2079, 2018.

- [40] G. Adrián, G. E. Francisco, M. Marcela, A. Baum, L. Daniel, and G. B. de Quirós Fernán, “Mongoddb: an open source alternative for HL7-CDA clinical documents management,” in *Proceedings of the Open Source International Conference (CISL’13)*, Buenos Aires, Argentina, 2013.
- [41] K. Kaur and R. Rani, “Managing data in healthcare information systems: many models, one solution,” *Computer*, vol. 48, no. 3, pp. 52–59, 2015.
- [42] B. Cyganek, M. Graña, B. Krawczyk et al., “A survey of big data issues in electronic health record analysis,” *Applied Artificial Intelligence*, vol. 30, no. 6, pp. 497–520, 2016.
- [43] D. V. Dimitrov, “Medical internet of things and big data in healthcare,” *Healthcare Informatics Research*, vol. 22, no. 3, pp. 156–163, 2016.
- [44] F. Firouzi, B. Farahani, M. Ibrahim, and K. Chakrabarty, “Keynote paper: from EDA to IoT eHealth: promises, challenges, and solutions,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 12, pp. 2965–2978, 2018.
- [45] S. Sakr and A. Elgammal, “Towards a comprehensive data analytics framework for smart healthcare services,” *Big Data Research*, vol. 4, pp. 44–58, 2016.
- [46] Y. Zhang, M. Qiu, C. W. Tsai, M. M. Hassan, and A. Alamri, “Health-CPS: healthcare cyber-physical system assisted by cloud and big data,” *IEEE Systems Journal*, vol. 11, no. 1, pp. 88–95, 2015.
- [47] S. K. Sharma and X. Wang, “Live data analytics with collaborative edge and cloud processing in wireless IoT networks,” *IEEE Access*, vol. 5, pp. 4621–4635, 2017.
- [48] M. S. Hadi, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, “Patient-centric cellular networks optimization using big data analytics,” *IEEE Access*, vol. 7, pp. 49279–49296, 2019.
- [49] S. El-Sappagh, F. Ali, S. El-Masri, K. Kim, A. Ali, and K. S. Kwak, “Mobile health technologies for diabetes mellitus: current state and future challenges,” *IEEE Access*, vol. 7, pp. 21917–21947, 2018.
- [50] Z. Hong, W. Chen, H. Huang, S. Guo, and Z. Zheng, “Multi-hop cooperative computation offloading for industrial IoT-edge-cloud computing environments,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 12, pp. 2759–2774, 2019.
- [51] D. C. Yacchirema, D. Sarabia-Jácome, C. E. Palau, and M. Esteve, “A smart system for sleep monitoring by integrating IoT with big data analytics,” *IEEE Access*, vol. 6, pp. 35988–36001, 2018.
- [52] M. H. U. Rehman, V. Chang, A. Batool, and T. Y. Wah, “Big data reduction framework for value creation in sustainable enterprises,” *International Journal of Information Management*, vol. 36, no. 6, pp. 917–928, 2016.
- [53] S. Kumar and M. Singh, “Big data analytics for healthcare industry: impact, applications, and tools,” *Big Data Mining and Analytics*, vol. 2, no. 1, pp. 48–57, 2018.
- [54] H. Jiang, Y. Chen, Z. Qiao, T.-H. Weng, and K.-C. Li, “Scaling up MapReduce-based big data processing on multi-GPU systems,” *Cluster Computing*, vol. 18, no. 1, pp. 369–383, 2015.
- [55] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities,” *IEEE Access*, vol. 5, pp. 8869–8879, 2017.
- [56] K. R. Ghani, K. Zheng, J. T. Wei, and C. P. Friedman, “Harnessing big data for health care and research: are urologists ready?” *European Urology*, vol. 66, no. 6, pp. 975–977, 2014.
- [57] J. Roski, G. W. Bo-Linn, and T. A. Andrews, “Creating value in health care through big data: opportunities and policy implications,” *Health Affairs*, vol. 33, no. 7, pp. 1115–1122, 2014.
- [58] T. U. Mane, “Smart heart disease prediction system using improved K-means and ID3 on big data,” in *Proceedings of the 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, IEEE, Pune, India, pp. 239–245, February 2017.
- [59] K. Rahimi, D. Bennett, N. Conrad et al., “Risk prediction in patients with heart failure,” *JACC: Heart Failure*, vol. 2, no. 5, pp. 440–446, 2014.
- [60] D. Al-Jumeily, A. Hussain, C. Mallucci, and C. Oliver, *Applied Computing in Medicine and Health*, Morgan Kaufmann, Burlington, MA, USA, 2015.
- [61] F. J. Martin-Sanchez, V. Aguiar-Pulido, G. H. Lopez-Campos, N. Peek, and L. Sacchi, “Secondary use and analysis of big data collected for patient care,” *Yearbook of Medical Informatics*, vol. 26, no. 01, pp. 28–37, 2017.
- [62] A. Jindal, A. Dua, N. Kumar, A. K. Das, A. V. Vasilakos, and J. J. P. C. Rodrigues, “Providing healthcare-as-a-service using fuzzy rule based big data analytics in cloud computing,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1605–1618, 2018.
- [63] D. D. Luxton, “An introduction to artificial intelligence in behavioral and mental health care,” in *Artificial Intelligence in Behavioral and Mental Health Care*, pp. 1–26, Academic Press, Cambridge, MA, USA, 2016.
- [64] 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.
- [65] Y. Wang, L. Kung, and T. A. Byrd, “Big data analytics: understanding its capabilities and potential benefits for healthcare organizations,” *Technological Forecasting and Social Change*, vol. 126, pp. 3–13, 2018.
- [66] N. Mehta and A. Pandit, “Concurrence of big data analytics and healthcare: a systematic review,” *International Journal of Medical Informatics*, vol. 114, pp. 57–65, 2018.
- [67] C. S. Kruse, R. Goswamy, Y. Raval, and S. Marawi, “Challenges and opportunities of big data in health care: a systematic review,” *JMIR Medical Informatics*, vol. 4, no. 4, p. e38, 2016.
- [68] 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.
- [69] V. Palanisamy and R. Thirunavukarasu, “Implications of big data analytics in developing healthcare frameworks—a review,” *Journal of King Saud University—Computer and Information Sciences*, vol. 31, no. 4, pp. 415–425, 2019.
- [70] G. Harerimana, B. Jang, J. W. Kim, and H. K. Park, “Health big data analytics: a technology survey,” *IEEE Access*, vol. 6, pp. 65661–65678, 2018.
- [71] A. Celesti, O. Amft, and M. Villari, “Guest editorial special section on cloud computing, edge computing, internet of things, and big data analytics applications for healthcare industry 4.0,” *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 454–456, 2019.
- [72] J. Qadir, M. Mujeeb-U-Rahman, M. H. Rehmani et al., “IEEE access special section editorial: health informatics for the developing world,” *IEEE Access*, vol. 5, pp. 27818–27823, 2017.

- [73] G. Phillips-Wren, L. S. Iyer, U. Kulkarni, and T. Ariyachandra, "Business analytics in the context of big data: a roadmap for research," *Communications of the Association for Information Systems*, vol. 37, no. 1, p. 23, 2015.
- [74] H. J. Watson, "Tutorial: big data analytics: concepts, technologies, and applications," *Communications of the Association for Information Systems*, vol. 34, no. 1, p. 65, 2014.
- [75] D. Delen, *Real-World Data Mining: Applied Business Analytics and Decision Making*, FT Press, Upper Saddle River, NJ, USA, 2014.
- [76] M. Riabacke, M. Danielson, and L. Ekenberg, "State-of-the-art prescriptive criteria weight elicitation," *Advances in Decision Sciences*, vol. 2012, Article ID 276584, 24 pages, 2012.
- [77] Z. Pang, H. Yuan, Y.-T. Zhang, and M. Packirisamy, "Guest editorial health engineering driven by the industry 4.0 for aging society," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 6, pp. 1709-1710, 2018.
- [78] D. Rajeshwari, "State of the art of big data analytics: a survey," *International Journal of Computer Applications*, vol. 120, no. 22, 2015.
- [79] S. Shafqat, S. Kishwer, R. U. Rasool, J. Qadir, T. Amjad, and H. F. Ahmad, "Big data analytics enhanced healthcare systems: a review," *The Journal of Supercomputing*, vol. 76, no. 3, pp. 1754-1799, 2018.
- [80] M. Chen, S. Mao, Y. Zhang, and V. C. Leung, *Big Data: Related Technologies, Challenges and Future Prospects*, Springer, Berlin, Germany, 2014.
- [81] Z. Zhou, W. Gaaloul, P. C. K. Hung, L. Shu, and W. Tan, "IEEE access special session editorial: big data services and computational intelligence for industrial systems," *IEEE Access*, vol. 3, pp. 3085-3088, 2015.
- [82] 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.
- [83] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: big data for personalized healthcare," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1209-1215, 2015.
- [84] Y. Zhang, L. Zhang, E. Oki, N. V. Chawla, and A. Kos, "IEEE Access special section editorial: big data analytics for smart and connected health," *IEEE Access*, vol. 4, pp. 9906-9909, 2016.
- [85] J. Gantz and D. Reinsel, "Extracting value from chaos," *IDC Iview*, vol. 1142, pp. 1-12, 2011.
- [86] C. Ngufor and J. Wojtusiak, "Learning from large-scale distributed health data: an approximate logistic regression approach," in *Proceedings of the ICML 13: Role of Machine Learning in Transforming Healthcare*, Atlanta, GA, USA, 2013.
- [87] R. Zhang, G. Simon, and F. Yu, "Advancing Alzheimer's research: a review of big data promises," *International Journal of Medical Informatics*, vol. 106, pp. 48-56, 2017.
- [88] I. Cano, A. Tenyi, E. Vela, F. Miralles, and J. Roca, "Perspectives on big data applications of health information," *Current Opinion in Systems Biology*, vol. 3, pp. 36-42, 2017.
- [89] <https://www.forbes.com/sites/louiscolombus/2017/12/24/53-of-companies-are-adoptingbig-data-analytics/#50bf384239a1>.
- [90] M. Bakratsas, P. Basaras, D. Katsaros, and L. Tassioulas, "Hadoop mapreduce performance on SSDs for analyzing social networks," *Big Data Research*, vol. 11, pp. 1-10, 2018.
- [91] P. Zikopoulos, D. Deroos, K. Parasuraman, T. Deutsch, J. Giles, and D. Corrigan, *Harness the Power of Big Data the IBM Big Data Platform*, McGraw Hill Professional, New York, NY, USA, 2012.
- [92] P. Zikopoulos and C. Eaton, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill Osborne Media, New York, NY, USA, 2011.
- [93] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big data technologies: a survey," *Journal of King Saud University—Computer and Information Sciences*, vol. 30, no. 4, pp. 431-448, 2018.
- [94] V. Rajaraman, "Big data analytics," *Resonance*, vol. 21, no. 8, pp. 695-716, 2016.
- [95] M. Idris, S. Hussain, M. Ali et al., "Context-aware scheduling in MapReduce: a compact review," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 17, pp. 5332-5349, 2015.
- [96] H. Senger, V. Gil-Costa, L. Arantes et al., "BSP cost and scalability analysis for MapReduce operations," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 8, pp. 2503-2527, 2016.
- [97] <https://www.datasciencecentral.com/profiles/blogs/the-hadoop-ecosystem-hdfs-yarn-hivepig-hbase-and-growing>.
- [98] A. K. Bhadani and D. Jothimani, "Big data: challenges, opportunities, and realities," in *Effective Big Data Management and Opportunities for Implementation*, pp. 1-24, IGI Global, Pennsylvania, PA, USA, 2016.
- [99] N. Khan, I. Yaqoob, I. A. T. Hashem et al., "Big data: survey, technologies, opportunities, and challenges," *The Scientific World Journal*, vol. 2014, Article ID 712826, 18 pages, 2014.
- [100] S. Bahri, N. Zoghalmi, M. Abed, and J. M. R. S. Tavares, "Big data for healthcare: a survey," *IEEE Access*, vol. 7, pp. 7397-7408, 2019.
- [101] S. R. Sukumar, R. Natarajan, and R. K. Ferrell, "Quality of big data in health care," *International Journal of Health Care Quality Assurance*, vol. 28, no. 6, pp. 621-634, 2015.
- [102] B. Cleland, J. Wallace, R. Bond et al., "Insights into antidepressant prescribing using open health data," *Big Data Research*, vol. 12, pp. 41-48, 2018.
- [103] N. Agnihotri and A. K. Sharma, "Proposed algorithms for effective real time stream analysis in big data," in *Proceedings of the 2015 Third International Conference on Image Information Processing (ICIIP)*, pp. 348-352, IEEE, Wanknaghat, India, December 2015.
- [104] J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, "Big data analytics to improve cardiovascular care: promise and challenges," *Nature Reviews Cardiology*, vol. 13, no. 6, pp. 350-359, 2016.
- [105] 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.
- [106] M. Nambu, K. Nakajima, M. Noshiro, and T. Tamura, "An algorithm for the automatic detection of health conditions," *IEEE Engineering in Medicine and Biology Magazine*, vol. 24, no. 4, pp. 38-42, 2005.
- [107] R. Platt, R. Carnahan, J. S. Brown et al., "The U.S. Food and drug administration's mini-sentinel program," *Pharmacoepidemiology and Drug Safety*, vol. 21, pp. 1-303, 2012.
- [108] E. S. Berner, *Clinical Decision Support Systems*, vol. 233, Springer Science+ Business Media, LLC, New York, NY, USA, 2007.
- [109] C. S. Mayo, J. M. Moran, W. Bosch et al., "American association of physicists in medicine task group 263: standardizing nomenclatures in radiation oncology," *International Journal of Radiation Oncology\*Biophysics\*Physics*, vol. 100, no. 4, pp. 1057-1066, 2018.

- [110] A. Kos and A. Umek, "Wearable sensor devices for prevention and rehabilitation in healthcare: swimming exercise with real-time therapist feedback," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1331–1341, 2018.
- [111] A. K. Sahoo, S. Mallik, C. Pradhan, B. S. P. Mishra, R. K. Barik, and H. Das, "Intelligence-based health recommendation system using big data analytics," in *Big Data Analytics for Intelligent Healthcare Management*, pp. 227–246, Academic Press, Cambridge, MA, USA, 2019.
- [112] T. R. Hoens, M. Blanton, A. Steele, and N. V. Chawla, "Reliable medical recommendation systems with patient privacy," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4, pp. 1–31, 2013.
- [113] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," *New England Journal of Medicine*, vol. 363, no. 4, pp. 301–304, 2010.
- [114] G. Wang, K. Jung, R. Winnenburg, and N. H. Shah, "A method for systematic discovery of adverse drug events from clinical notes," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1196–1204, 2015.
- [115] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big data application in biomedical research and health care: a literature review," *Biomedical Informatics Insights*, vol. 8, Article ID BILS31559, 2016.
- [116] T. Saheb and L. Izadi, "Paradigm of IoT big data analytics in healthcare industry: a review of scientific literature and mapping of research trends," *Telematics and Informatics*, vol. 41, pp. 70–85, 2019.
- [117] L. Syed, S. Jabeen, and S. Manimala, "Telemammography: a novel approach for early detection of breast cancer through wavelets based image processing and machine learning techniques," in *Advances in Soft Computing and Machine Learning in Image Processing*, pp. 149–183, Springer, Cham, Switzerland, 2018.
- [118] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4-5, pp. 198–211, 2007.
- [119] G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K. M. Abbas, and R. Sundarsekar, "Big data knowledge system in healthcare," in *Internet of Things and Big Data Technologies for Next Generation Healthcare*, pp. 133–157, Springer, Cham, Switzerland, 2017.
- [120] T. Heart, O. Ben-Assuli, and I. Shabtai, "A review of PHR, EMR and EHR integration: a more personalized healthcare and public health policy," *Health Policy and Technology*, vol. 6, no. 1, pp. 20–25, 2017.
- [121] P. Galetsi, K. Katsaliaki, and S. Kumar, "Values, challenges and future directions of big data analytics in healthcare: a systematic review," *Social Science & Medicine*, vol. 241, p. 112533, 2019.
- [122] A. Kankanhalli, J. Hahn, S. Tan, and G. Gao, "Big data and analytics in healthcare: introduction to the special section," *Information Systems Frontiers*, vol. 18, no. 2, pp. 233–235, 2016.
- [123] A. Belle, R. Thiagarajan, S. M. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big data analytics in healthcare," *BioMed Research International*, vol. 2015, Article ID 370194, 16 pages, 2015.



## Review Article

# Data Analytics in Mental Healthcare

**Ayesha Kamran Ul haq,<sup>1</sup> Amira Khattak,<sup>2</sup> Noreen Jamil ,<sup>1</sup> M. Asif Naeem ,<sup>1,3</sup> and Farhaan Mirza<sup>3</sup>**

<sup>1</sup>National University of Computer and Emerging Sciences, Islamabad, Pakistan

<sup>2</sup>Prince Sultan University, Riyadh, Saudi Arabia

<sup>3</sup>Auckland University of Technology, Auckland, New Zealand

Correspondence should be addressed to Noreen Jamil; [noreen.jamil@nu.edu.pk](mailto:noreen.jamil@nu.edu.pk)

Received 20 January 2020; Revised 13 March 2020; Accepted 12 June 2020; Published 4 July 2020

Academic Editor: Shaukat Ali

Copyright © 2020 Ayesha Kamran Ul haq 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.

Worldwide, about 700 million people are estimated to suffer from mental illnesses. In recent years, due to the extensive growth rate in mental disorders, it is essential to better understand the inadequate outcomes from mental health problems. Mental health research is challenging given the perceived limitations of ethical principles such as the protection of autonomy, consent, threat, and damage. In this survey, we aimed to investigate studies where big data approaches were used in mental illness and treatment. Firstly, different types of mental illness, for instance, bipolar disorder, depression, and personality disorders, are discussed. The effects of mental health on user's behavior such as suicide and drug addiction are highlighted. A description of the methodologies and tools is presented to predict the mental condition of the patient under the supervision of artificial intelligence and machine learning.

## 1. Introduction

Recently the term “big data” has become exceedingly popular all over the world.

Over the last few years, big data has started to set foot in healthcare system. In this context, scientists have been working on improving the public health strategies, medical research, and the care provided to patients by analyzing big datasets related to their health.

Data is coming from different sources like providers (pharmacy and patient's history) and nonproviders (cell phone and internet searches). One of the outstanding possibilities available from huge data utilization is evident inside the healthcare industry. Healthcare organizations have a big quantity of information available to them and a big portion of it is unstructured and clinically applicable. The use of big data is expected to grow in the medical field and it will continue to pose lucrative opportunities for solutions that can help in saving lives of patients. Big data needs to be interpreted correctly in order to predict future data so that

final result can be estimated. To solve this problem, researchers are working on AI algorithms that have a high impact on analysis of huge quantities of raw data and extract useful information from it. There are varieties of AI algorithms that are used to predict patient disease by observing past data. A variety of wearable sensors have been developed to deal with both physical and social interactions practically.

Mental health of a person is measured by a high grade of affective disorder which results in major depression and different anxiety disorders. There are many conditions which are recognized as mental disorders including anxiety disorder, depressive disorder, mood disorder, and personality disorder. There are lots of mobile apps, smart devices like smartwatches, and smart bands which increase healthcare facilities in mobile mental healthcare systems. Personalized psychiatry also plays an important role in predicting bipolar disorder and improving diagnosis and optimized treatment. Most of the smart techniques are not pursued due to lack of resources especially in underdeveloping countries. Like, in Pakistan, 0-1% of the government health budget is being

spent on the mental health system. There is a need for an affordable solution to detect depression in Pakistan so that everyone could be able to pay attention to it.

Researchers are working on many machine learning algorithms to analyze raw data to deduce meaningful information. It is now impossible to manage data in healthcare with traditional database management tools as data is in terabytes and petabytes now. In this survey, we analyzed different issues related to mental healthcare by usage of big data. We analyze different mental disorders like bipolar disease, opioid use disorder, personality disorder, different anxiety disorders, and depression. Social media is one of the biggest and most powerful resources for data collection as every 9 out of 10 people use social networking sites nowadays. Twitter is the main focus of interest for most researchers as people write 500,000 tweets on average per minute. Twitter is being used for sentimental analyses and opinion mining in the business field in order to check the popularity of a product by observing customer tweets. We have a lot of structure and unstructured data in order to reach any decision; data must be processed and stored in such a manner that follows the same structure. We analyzed and compared the working of different storage models under different conditions like mongo DB and Hadoop which are two different approaches to store large amounts of data. Hadoop works on cloud computing that helps to accomplish different operations on distributed data in a systematic manner.

In this survey we discuss the mental health problems with big data into further four sections. The second section describes related work regarding mental healthcare and the latest research on it. The third section describes different types of mental illness and their solutions within the data science. The fourth section describes the different illegal issues faced by the mental patients and early detection of these types of activities. The fifth section describes different approaches of data science towards mental healthcare systems such as different training and testing methods of health data for early prediction like supervised and unsupervised learning methods and artificial neural network (ANN).

## 2. Literature Review

There are a lot of mental disorders like bipolar one, depression, and different forms of anxieties. Bauer et al. [1] conducted a paper-based survey in which 1222 patients from 17 countries were participated to detect bipolar disorder in adults. This survey was translated into 12 different languages with some limitation that it did not contain any question about technology usage in older adults. According to Bauer et al. [1], digital treatment is not suitable for the older adults with bipolar disorder.

Researchers are working on the most interesting and unique method of tremendous interest to check the personality of a person just by looking at the way he or she is using the mobile phone. De Montjoye [2] collected dataset from US Research University and created a framework that analyzed phone call and text messages to check the personality of the user. Participants who did 300 calls or text per

year failed to complete personality measures. They choose optimal sample size that is 69 with mean age = 30.4, S. D. = 6.1, and 1 missing value. Similarly, Bleidorn and Hopwood [3] adopted a comprehensive machine learning approach to test the personality of the user using social media and digital records. Main 9 recommendations for how to amalgamate machine learning techniques provided by the researcher enhance the big five of the personality assessments. Focusing on minor details of the user comprehends and validates the result. Digital mental health has been revolutionized and its innovations are growing at a high rate. The National Health Service (NHS) has recognized its importance in mental healthcare and is looking for innovations to provide services at low cost. Hill et al. [4] presented a study of challenges and considerations in innovations in digital mental healthcare. They also suggested collaboration between clinicians, industry workers, and service users so that these challenges can be overcome and successful innovations of e-therapies and digital apps can be developed.

There are lots of mobile apps, smart devices like smartwatches, smart bands, and shirts which increase healthcare facilities in the mobile healthcare system. A variety of wearable sensors have been developing to deal with both physical and social interactions practically. Combining artificial intelligence with healthcare systems extends the healthcare facilities up to the next level. Dimitrov [5] conducted a systematic survey on mobile internet of things in the devices which allow business to emerge, spread productivity improvements, lock down the cost, and intensify customer experience and change in a positive way. Similarly, Monteith et al. [6] performed a paper-based survey on clinical data mining to analyze different data sources to get psychiatry data and optimized precedence opportunities for psychiatry.

One of the machine learning algorithms named artificial neural network (ANN) is based on three-layer architecture. Kellmeyer [7] introduced a way to secure big brain data from clinical and consumer-directed neurotechnological devices using ANN. But this model needs to be trained on a huge amount of data to get accurate results. Jiang et al. [8] designed and developed a wearable device with multisensing capabilities including audio sensing, behavior monitoring, and environment and physiological sensing that evaluated speech information and automatically deleted raw data. Tested students were split into two groups, those with excessive scores or in excessive score. Participants were required to wear the device to make sure of the authenticity of the data. But one of the major challenges to enable IoT in the device is safe communication.

Yang et al. [9] invented an IoT enabled wearable device for mental well-being and some external equipment to record speech data. This portable device would be able to recognize motion, pressure, monitoring, and physiological status of a person. There are lots of technologies that produce tracking data, such as smartphones, credit cards, websites, social media, and sensors offering benefits. Monteith and Glenn [10] elaborated some kind of generated data using human made algorithm, searching for disease symptoms, hit disease websites, sending/receiving healthcare e-mail, and

sharing health information on social media. Based on perceived data, the system predicted automated decision-making without the involvement of user to maintain security.

Considering all the above issues, there is a need for proper treatment of a disordered person. Mood of the patient is one of the parameters to detect his/her mental health. Public mood is hugely reflected in the social media as almost everyone uses social media in this modern era. Goyal [11] introduced a procedure in which tweets are filtered out for specific keywords from saved databases regarding food price crisis. Data is trained using two algorithms, K nearest neighbor and Naïve Bayes for unsupervised and supervised learning, respectively. Cloud storage is the best option to store huge amounts of unstructured data. Kumar and Bala [12] proposed functionalities of Hadoop for automatic processing and repository of big data. MongoDB is a big data tool for analyzing statistics related to world mental healthcare. Dhaka, P., and Johari [13] presented a way of implementation of big data tool 'MongoDB' for analyzing statistics related to world mental healthcare. The data is further analyzed using genetic algorithms for different mental disorders and deployed again in MongoDB for extracting final data.

But all of the above methods are useless without the user involvement. De Beurs et al. [14] introduced expert-driven method, intervention mapping, and scrum methods which may help to increase the involvement of the users. This approach tried to develop user-focused design strategies for the growth of web-based mental healthcare under finite resources. Turner et al. [15] elaborated in their article that the availability of the big data is increasing twice in size every two year for use in automated decision-making. Passos et al. [16] believed that the long-established connection between doctor and patient will change with the establishment of big data and machine learning models. ML algorithm can allow an affected person to observe his fitness from time to time and can tell the doctor about his current condition if it becomes worst. Early consultation with the doctor could prevent any bigger loss for the patient.

If the psychiatric disease is not predicted or handled earlier, then it enforces the patient to involve into many illegal activities like suicide as most of the suicide attempts are related to mental disorder. Kessler et al. [17] proposed meta-analysis that focused on suicide incidence within 1 year of the self-harm using machine learning algorithm. They analyzed the past reports of suicide patients and concluded that any prediction was impossible to be made due to short duration of psychiatric hospitalizations. Although a number of AI algorithms are used to estimate patient disease by observing past data, the focus of all studies was related to suicide prediction by setting up a threshold. Defining a threshold is a very crucial point or sometimes even impossible to be predicted. Cleland et al. [18] reviewed many studies but were unable to discover principles to clarify threshold. Authors used a random-effects model to generate a meta-analytic ROC. On the basis of correlation results, it is stated that depression prevalence is mediating factor between economic deprivation and antidepressant prescribing.

Another side effect of mental disease is drug addiction. Early drug prediction is possible by analyzing user data. Opioid is a swear type of drug. Hasan et al. [19] explored the Massachusetts All Payer Claim Data (MA APCD) dataset and examined how naïve users develop opioid use disorder. A popular machine learning algorithm is tested to predict the risk of such type of dependency of patent. Perdue et al. [20] predicted ratio of drug abusers by comparing Google trends data with monitoring the future (MTF) data; a well-structured study was made. It is concluded that Google trends and MTF data provided combined support for detecting drug abuse.

### 3. Mental Illness and Its Type

*3.1. Depression and Bipolar Disorder.* Bipolar disorder is also known as the worst form of depression. In Table 1, Bauer et al. [1] conducted a survey to check the bipolar disorder in adults. Data is collected from 187 older adults and 1021 younger adults with excluded missing observations. The survey contained 39 questions which took 20 minutes to complete. Older adults with bipolar disorder were addicted to the internet less regularly than the younger ones. As most of the healthcare services are available only online and most digital tools and devices are evolved, the survey has some limitations that it did not contain any question about technology usage in older adults. There is a need for proper treatment of a disordered person. Mood of the patient is one of the parameters to detect his/her mental health. Table 1 describes another approach of personality assessment using machine learning algorithm that focused on other aspects like systematic fulfillment and argued to enhance the validity of machine learning (ML) approach. Coming with technological advancement in the medical field will promote personalized treatments. A lot of work has been done in the field of depression detection using social networks.

The main goal of personalized psychiatry is to predict bipolar disorder and improve diagnosis and optimized treatment. To achieve these goals, it is necessary to combine the clinical variables of a patient as Figure 1 describes the integration of all these variables. It is now impossible to manage data in mental healthcare with database management traditional tools as data is in terabytes and petabytes now. So, there is a high need to introduce big data analytics tools and techniques to deal with such big data in order to improve the quality of treatment so that overall cost of treatment can be reduced throughout the world.

MongoDB is one of the tools to handle big data. The data is further analyzed using genetic algorithms for different mental disorders and deployed again in MongoDB for extracting final data. This approach of mining data and extracting useful information reduced overall cost of treatment. It provides the best results for clinical decisions. It helps doctors to give more accurate treatment for several mental disorders in less time and at low cost using useful information extracted by big data tool Mongo DB and genetic algorithm.

In Table 1, some of the techniques are handled and stored huge amount of data.

TABLE 1: Types of mental illness and role of big data.

Authors	Discipline(s) reviewed	Keywords used to identify papers for review	Methodology	Number of papers reviewed	Primary findings
Bauer et al. [1]	Bipolar disorder	Bipolar disorder, mental illness, and health literacy	Paper-based survey	68	47% of older adults used the internet versus 87% of younger adults having bipolar disorder
Dhaka and Johari [13]	Mental disorder	Mental health, disorders, and using MongoDB	Genetic algorithm and MongoDB tool	19	Analyzing and storing a large amount of data on MongoDB
Hill et al. [4]	Mental disorder	Mental health, collaborative computing, and e-therapies	(i) Online CBT platform (ii) Collaborative computing	33	(i) Developing smartphone application (ii) For mental disorder (iii) For improving e-therapies
Kumar and Bala, [12]	Depression detection through social media	Big data, Hadoop, sentiment analysis, social networks, and Twitter	Sentimental analysis and save data on Hadoop	14	Analyzing twitter users' view on a particular business product
Kellmeyer [7]	Big brain data	Brain data, neurotechnology, big data, privacy, security, and machine learning	(i) Machine learning (ii) Consumer-directed neurotechnological devices (iii) Combining expert with a bottom-up process	77	(i) Maximizing medical knowledge (ii) Enhancing the security of devices and sheltering the privacy of personal brain data
De Montjoye [2]	Mobile phone and user personality	Personality prediction, big data, big five personality prediction, Carrier's log, and CDR	(i) Entropy: detecting different categories (ii) Interevent time: frequency of call or text between two users (iii) AR coefficients: to convert list of call and text into time series	31	Analyzing phone calls and text messages under a five-factor model
Furnham [21]	Personality disorder	Dark side, big five, facet analysis, dependence, and dutifulness	Hogan 'dark side' measure (HDS) concept of dependent personality disorder (DPD)	34	All of the personality disorders are strongly negatively associated with agreeableness (a type of kind, sympathetic, and cooperative personality)
Bleidorn and Hopwood [3]	Personality assessment	Machine learning, personality assessment, big five, construct validation, and big data	(i) Machine learning (ii) Prediction models (iii) K-fold validation	65	Focusing on other aspects like systematic fulfillment and arguing to enhance the validity of machine learning (ML) approach

Using MongoDB tool, researchers are working to predict mental condition before severe mental stage. So, some devices introduced a complete detection process to tackle the present condition of the user by analyzing his/her daily life routine. There is a need for reasonable solutions that detect disable stage of a mental patient more precisely and quickly.

**3.2. Personality Disorder.** Dutifulness is a type of personality disorder in which patients are overstressed about the disease that is not actually much serious. People with this type of disorder tend to work hard to impress others. A survey was conducted to find the relationship between normal and dutifulness personalities. Other researchers are working on the most interesting and unique method of tremendous interest to

check the personality of a person just by looking at the way he or she is using the mobile phone. This approach provides cost-effective and questionnaire-free personality detection through mobile phone data that performs personality assessment without conducting any digital survey on social media. To perform all nine main aspects of the constructed validation in real time is not easy for the researchers. This examination, like several others, has limitations. This is just a sample that has implications for generalization when it is used in the near-real-time scenario which may be tough for the researchers.

#### 4. Effects of Mental Health on User Behavior

Mental illness is upswing in the feelings of helplessness, danger, fear, and sadness in the people. People do not

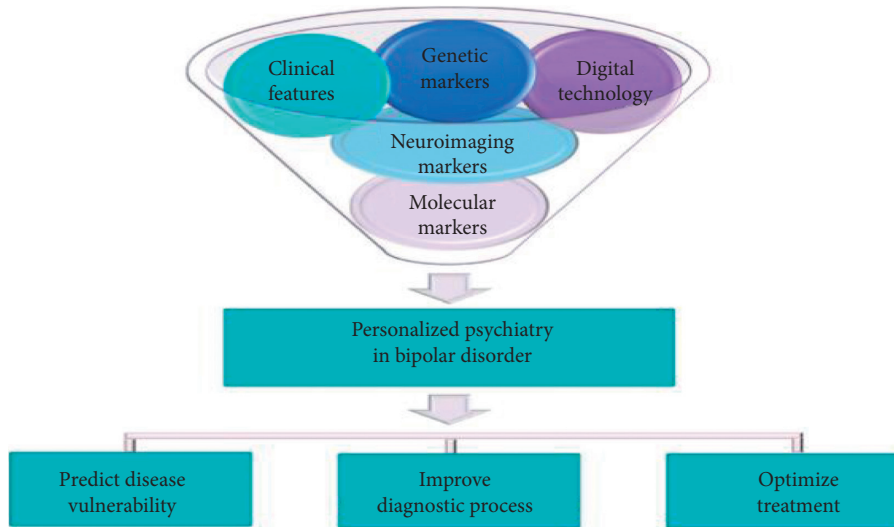


FIGURE 1: Goals of personalized treatment in bipolar disorder [22].

understand the current situation so this thing imposes psychiatric patients to illegal activities. Table 2 described some issues that appear because of mental disorder like suicide, drug abuse, and opioid use as follows.

**4.1. Suicide.** Suicide is very common in underdeveloped countries. According to researchers, someone dies because of suicide in every 40 seconds all over the world. There are some areas in the world where mental disorder and suicide statistics are relatively larger than other areas.

Psychiatrists say that 90% of people who died by suicide faced a mental disorder. Electronic medical records and big data generate suicide through machine learning algorithm. Machine learning algorithms can be used to predict suicides in depressed persons; it is hard to estimate how accurately it performs, but it may help a consultant for pretreating patients based on early prediction. Various studies depict the fact that there are a range of factors such as high level of antidepressant prescribing that caused such prevalence of illness. Some people started antidepressant medicine to overcome mental affliction. In Table 1, Cleland et al. [18] explored three main factors, i.e., economic deprivation, depression prevalence, and antidepressant prescribing and their correlations. Several statistical tools could be used like Jupyter Notebook, Pandas, NumPy, Matplotlib, Seaborn, and ipyleaflet for creation of pipeline. Correlations are analyzed using Pearson's correlation and  $p$  values. The analysis shows strong correlation between economic deprivation and antidepressant prescribing whereas it shows weak correlations between economic deprivation and depression prevalence.

**4.2. Drug Abuse.** People voluntarily take drugs but most of them are addicted to them in order to get rid of all their problems and feel relaxed. Adderall, Divinorum, Snus, synthetic marijuana, and bath salts are the novel drugs. Opioid is a category of drug that includes the illegitimate drug heroin. Hasan et al. [19] compared four machine learning

algorithms: logistic regression, random forest, decision tree, and gradient boosting to predict the risk of opioid use disorder. Random forest is one of the best methods of classification in machine learning algorithms. It is found that in such types of situations random forest models outperform the other three algorithms specially for determining the features. There is another approach to predict drug abusers using the search history of the user. Perdue et al. [20] predicted ratio of drug abusers by comparing Google trends data with monitoring the future (MTF) data; a well-structured study was made. It is concluded that Google trends and MTF data provided combined support for detecting drug abuse.

Google trends appear to be a particularly useful data source regarding novel drugs because Google is the first place where many users especially adults go for information on topics of which they are unfamiliar. Google trends not to predict heroin abuse; the reason may be that heroin is a relatively uniquely dangerous than other drugs. According to Granka [23], internet searches can be understood as behavioral measures of an individual's interest in an issue. Unfortunately, this technique was not going to be very convenient as drug abuse researchers are unable to predict drug abuse successfully because of sparse data.

## 5. How Data Science Helps to Predict Mental Illness?

Currently, there are numerous mobile clinical devices which are established in patients' personal body networks and medical devices. They receive and transmit massive amounts of heterogeneous fitness records to healthcare statistics structures for patient's evaluation. In this context, system learning and data mining strategies have become extremely crucial in many real-life problems. Many of those techniques were developed for health data processing and processing on cellular gadgets.

There is a lot of data in the world of medicine as data is coming from different sources like pharmacy and patient's

TABLE 2: Side effects of mental illness and their solution through data science.

Authors	Side effects of mental disorder	Tools/techniques	Primary findings
Kessler et al. [17]	Suicide and mental illness	Machine learning algorithm	Predicting suicide risk at hospitalization
Cleland et al. [18]	Antidepressant usage	Clustering analysis based on behavior and disease	Identifying the correlation between antidepressant usage and deprivation
Perdue et al. [20]	Drug abuse	(i) Google search history (ii) Monitoring the future (MTF)	Providing real time data that may allow us to predict drug abuse tendency and respond more quickly
Hasan et al. [19]	Opioid use disorder	(iii) Feature engineering (iv) Logistic regression (v) Random forest (vi) Gradient boosting	Suppressing the increasing rate of opioid addiction using machine learning algorithms

history and from nonproviders (cell phone and internet searches). Big data needs to be interpreted in order to predict future data, estimate hypothesis, and conclude results. Psychiatrists should be able to evaluate results from research studies and commercial analytical products that are based on big data.

**5.1. Artificial Intelligence and Big Data.** Big data collected from wearable tracking devices and electronic records help to store accumulating and extensive amounts of data. Smart mobile apps support fitness and health education, predict heart attack, and calculate ECG, emotion detection, symptom tracking, and disease management. Mobile apps can improve connection between patients and doctors. Once a patient's data from different resources is organized into a proper structure, artificial intelligence (AI) algorithm can be used. After all, AI recognizes patterns, finds similarity between them, and makes predictive recommendations about what happened with those in that condition.

Techniques used for healthcare data processing can be widely categorized into two classes: nonartificial intelligence systems and artificial intelligence systems. Although non-AI techniques are less complex, but they are suffering from a lack of convergence that gives inaccurate results as compared to AI techniques. Contrary to that, AI methods are preferable than non-AI techniques. In Table 3, Dimitrov [5] combined artificial intelligence with IoT technology in existing healthcare apps so that connection between doctors and patients remains balanced. Disease prediction is also possible through machine learning. Figure 2 shows hierarchical structure of AI, ML, and neural networks.

One of the machine learning algorithms named artificial neural network (ANN) is based on three-layer architecture. Kellmeyer [7] introduced a way to secure big brain data from neurotechnological devices using ANN. This algorithm was working on a huge amount of data (train data) to predict accurate results. But patients' brain diseases are rare so training models on small data may produce imprecise results. Machine learning models are data hungry. To obtain accurate results as an output, there is a need of training more data with distinct features as an input. These new methods cannot be applicable on clinical data due to the limited economy resources.

**5.2. Prediction through Smart Devices.** Various monitoring wearable devices (Table 3) are available that continuously capture the finer details of behaviors and provide important cues about fear and autism. This information is helpful to recognize mental issues of the user of those devices. Victims were monitored continuously for a month. High level computation performed on the voice requires high complexity data as well as high computational power which leads to a huge pressure on the small chip. In order to overcome power issues, relatively low frequency was chosen.

Yang et al. [9] invented an audio well-being device and conducted a survey in which participants have to speak more than 10 minutes in a quiet room. The first step is to choose the validity of the sample by completing some questions (including STAI, NEO-FFI, and AQ) to the participants. In order to determine whether they are suitable for the experiment or not, a test was conducted based on an AQ question. There was a classification algorithm applied on the AQ data. This type of device has one advantage; it perfectly worked on long-term data instead of low-term one but they used offline data transfer instead of real time.

Although it has different sensors, adding up garbage data to the sensors is a very obvious thing. This is an application that offers on-hand record management using mobile/tablet technology once security and privacy are confirmed. To increase the reliability of IoT devices, there is a need to increase the sample size with different age groups in real time environment to check the validity of the experiment.

There are a lot of technologies that effectuate tracking data like smartphones, credit cards, social media, and sensors. This paper discussed some of the existing work to tackle such data. In Table 3, one of the approaches is human made algorithm; searching for disease symptoms hits disease websites, sending/receiving healthcare e-mail, and sharing health information on social media through this kind of data. These are some examples of activities that perform key rules to produce medical data.

**5.3. Role of Social Media to Predict Mental Illness.** Constant mood of the patient is one of the parameters to detect his/her mental health. According to Lenhart, A. et al. [25] studied almost four out of five internet users of social

TABLE 3: Data analytics and predicting mental health.

Authors	Tool/technology	Methodology	Purpose of finding	Strength	Weakness
Dimitrov [5]	(i) Sensing technology	Emergence of medical internet of things (mIoT) in existing mobile apps	Providing benefits to the customers	(i) Achieving improved mental health	Adding up garbage data to the sensors
	(ii) Artificial intelligence		(i) Avoiding chronic and diet-related illness (ii) improving cognitive function	(ii) improving lifestyles in real time decision-making	
Monteith et al. [6]	Survey based approach	Clinical data mining	Analyzing different data sources to get psychiatry data	Optimized precedence opportunities for psychiatry	N/A
Kellmeyer [7]	Neurotechnology	(i) Machine learning (ii) Consumer-directed neurotechnological devices (iii) Combining expert with a bottom-up process	Enhancing the security of devices and sheltering the privacy of personal brain	Maximizing medical knowledge	Model needs huge amount of training data as brain disease is rarely captured
Yang et al. [9]	Long-term monitoring wearable device with internet of things	Well-being questionnaires with a group of students	Developing app-based devices linked to android phones and servers for data visualization monitoring and environment sensing	Perfectly working on long-term data	Offline data transfer instead of real time
Monteith and Glenn [10]	Automated decision-making	Hybrid algorithm that combines the statistical focus and data mining	Tracking day-to-day behavior of the user by automatic decision-making	Automatically detecting human decision without any input	How to ignore irrelevant information is a key headache
De Beurs et al. [14]	Online intervention	Expert-driven method Intervention mapping Scrum	Increasing user involvement under limited resources	Standardizing the level of user involvement in the web-based healthcare system	Deciding threshold for user involvement is problematic
Kumar and Bala [12]	Hadoop	Doing sentimental analysis and saving data on Hadoop	Analyzing twitter users' view on a particular business product	Checking out popularity of a particular service	Usage of two programming languages needs experts
Goyal [11]	KNN and Naïve Bayes classifier	Text mining and hybrid approach combining KNN and Naïve Bayes	Opinion mining of tweets related to food price crisis	Cost-effective way to predict prizes	Data needs to be cleaned before training

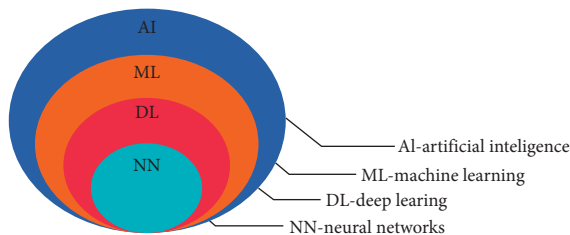


FIGURE 2: AI and ML [24].

media. In Table 3, researchers used twitter data to get online user review that helps the seeker to check out popularity of a particular service or purchase a product. In order to collect opinion of people on Airtel, they did analysis on it. Filter of the keyword is done using *Filter by content* and *Filter by location*. First of all, special character, URL, spam, and short

words are removed from the tweets. Secondly, remaining words from the tweets are then tokenized and TF-IDF score is calculated for all the keywords. After cleaning of data, classification algorithm named K nearest neighbor and Naïve Bayes algorithm were applied on the text in order to extract feature. Location filters work on specific bounding filter. Although hybrid recommendation system is providing 76.31% accuracy of the result, then Naïve Bayes is 66.66%. At the end, automated system is designed for opinion mining.

There is another point of consideration that Tweeter has unstructured data so handling such a huge amount of unstructured data is a tedious task to take up. Due to lack of schema structure, it is difficult to handle and store unstructured data. There is a need for storage devices to store an insignificant amount of data for processing. Cloud storage is the best option for such a material. The entire program is designed in Python so that it could be able to

catch all possible outcomes. Hadoop works on cloud computing that helps to accomplish different operations on distributed data in a systematic manner. Success rate of the above approach was around 70% but authors have done these tasks using two programming languages. Python code for extraction tweets and Java is used to train the data which required expert programmers on each language. It will help doctors to give more accurate treatment for several mental disorders in less time and at low cost. Infusing this approach provides predetection of depression that may preserve the patient to face the worst stage of mental illness.

#### 5.4. Key Challenges to Big Data Approach

- (i) Big data has many ethical issues related to privacy, reusability without permission, and involvement of the rival organization.
- (ii) To work in diverse areas, big data requires collaboration with expert people in the relative field including physicians, biologists, and developers that is crucial part of it. Data mining algorithms can be used to observe or predict data more precisely than traditional clinical trials.
- (iii) People may feel hesitant to describe all things to the doctors. One of the solutions to estimate the bad mental illness before time is automated decision-making without human input as shown in Table 3 . It collects data from our behavior that is unsophisticated to the digital economy. Key role of digital providence must be inferred in order to understand the difficulties that technology may be responsible for people with mental illness.
- (iv) There are many security issues while discussing sensitive information online as data may be revealed so a new approach to provide privacy protections as well as decision-making from the big data through new technologies needs to be introduced.
- (v) Also, if online data is used to predict user personality, then keeping data secured and protected from hacker is a big challenge. A lot of cheap solutions exist but they are not reliable from a user's perspective especially.
- (vi) Major challenges for enabling IoT in the device is communication; all of the above methods are useless without the user involvement. User is one of the main parts of the experiment especially if the user's personal or live data is required. Although many web-based inventions related to mental health are being released, the actual problem of active participation by end users is limited. In Table 3, an expert-driven method is introduced that is based on intervention mapping and scrum methods. It may help to increase the involvement of the users. But if all the users are actively involved in the web-based healthcare system, then it becomes problematic.
- (vii) When deciding on the level of user involvement, there is a need to decide about user input with the

accessibility of resources. It required an active role of technological companies and efficient time consumption. Further research should provide direction on how to select the best and optimized user-focused design strategies for the development of web-based mental health under limited resources.

## 6. Conclusions

Big data are being used for mental health research in many parts of the world and for many different purposes. Data science is a rapidly evolving field that offers many valuable applications to mental health research, examples of which we have outlined in this perspective.

We discussed different types of mental disorders and their reasonable, affordable, and possible solution to enhance the mental healthcare facilities. Currently, the digital mental health revolution is amplifying beyond the pace of scientific evaluation and it is very clear that clinical communities need to catch up. Various smart healthcare systems and devices developed that reduce the death rate of mental patients and avert the patient to associate in any illegal activities by early prediction.

This paper examines different prediction methods. Various machine learning algorithms are popular to train data in order to predict future data. Random forest model, Naïve Bayes, and k-mean clustering are popular ML algorithms. Social media is one of the best sources of data gathering as the mood of the user also reveals his/her psychological behavior. In this survey, various advances in data science and its impact on the smart healthcare system are points of consideration. It is concluded that there is a need for a cost-effective way to predict intellectual condition instead of grabbing costly devices. Twitter data is utilized for the saved and live tweets accessible through application program interface (API). In the future, connecting twitter API with python, then applying sentimental analysis on 'posts,' 'liked pages,' 'followed pages,' and 'comments' of the twitter user will provide a cost-effective way to detect depression for target patients.

### Data Availability

The authors will provide the data used for the experiments, if requested.

### Conflicts of Interest

There are no conflicts of interest regarding the publication of this paper.

### Acknowledgments

The authors are thankful to Prince Sultan University for the financial support towards the publication of this paper.

### References

- [1] R. Bauer, T. Glenn, S. Strojilovich et al., "Internet use by older adults with bipolar disorder: international survey results,"



- International Journal of Bipolar Disorders*, vol. 6, no. 1, p. 20, 2018.
- [2] Y.-A. De Montjoye, J. Quoidbach, F. Robic, and A. Pentland, "Predicting personality using novel mobile phone-based metrics," in *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pp. 48–55, Berlin, Heidelberg, April 2013.
  - [3] W. Bauer and C. J. Hopwood, "Using machine learning to advance personality assessment and theory," *Personality and Social Psychology Review*, vol. 23, no. 2, pp. 190–203, 2019.
  - [4] C. Hill, J. L. Martin, S. Thomson, N. Scott-Ram, H. Penfold, and C. Creswell, "Navigating the challenges of digital health innovation: considerations and solutions in developing online and smartphone-application-based interventions for mental health disorders," *British Journal of Psychiatry*, vol. 211, no. 2, pp. 65–69, 2017.
  - [5] D. V. Dimitrov, "Medical internet of things and big data in healthcare," *Healthcare Informatics Research*, vol. 22, no. 3, pp. 156–163, 2016.
  - [6] S. Monteith, T. Glenn, J. Geddes, and M. Bauer, "Big data are coming to psychiatry: a general introduction," *International Journal of Bipolar Disorders*, vol. 3, no. 1, p. 21, 2015.
  - [7] P. Kellmeyer, "Big brain data: on the responsible use of brain data from clinical and consumer-directed neurotechnological devices," *Neuroethics*, vol. 11, pp. 1–16, 2018.
  - [8] L. Jiang, B. Gao, J. Gu et al., "Wearable long-term social sensing for mental wellbeing," *IEEE Sensors Journal*, vol. 19, no. 19, 2019.
  - [9] S. Yang, B. Gao, L. Jiang et al., "IoT structured long-term wearable social sensing for mental wellbeing," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3652–3662, 2018.
  - [10] S. Monteith and T. Glenn, "Automated decision-making and big data: concerns for people with mental illness," *Current Psychiatry Reports*, vol. 18, no. 12, p. 112, 2016.
  - [11] S. Goyal, "Sentimental analysis of twitter data using text mining and hybrid classification approach," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 2, no. 5, pp. 2454–132X, 2016.
  - [12] M. Kumar and A. Bala, "Analyzing twitter sentiments through big data," in *Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 2628–2631, New Delhi, India, March 2016.
  - [13] P. Dhaka and R. Johari, "Big data application: study and archival of mental health data, using MongoDB," in *Proceedings of the 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp. 3228–3232, Chennai, India, March 2016.
  - [14] D. De Beurs, I. Van Bruinessen, J. Noordman, R. Friele, and S. Van Dulmen, "Active involvement of end users when developing web-based mental health interventions," *Frontiers in Psychiatry*, vol. 8, p. 72, 2017.
  - [15] V. Turner, J. F. Gantz, D. Reinsel, and S. Minton, "The digital universe of opportunities: rich data and the increasing value of the internet of things," *IDC Analyze the Future*, vol. 16, 2014.
  - [16] I. C. Passos, P. Ballester, J. V. Pinto, B. Mwangi, and F. Kapczynski, "Big data and machine learning meet the health sciences," in *Personalized Psychiatry*, pp. 1–13, Springer, Cham, Switzerland, 2019.
  - [17] R. C. Kessler, S. L. Bernecker, R. M. Bossarte et al., "The role of big data analytics in predicting suicide," in *Personalized Psychiatry*, pp. 77–98, Springer, Cham, Switzerland, 2019.
  - [18] B. Cleland, J. Wallace, R. Bond et al., "Insights into antidepressant prescribing using open health data," *Big Data Research*, vol. 12, pp. 41–48, 2018.
  - [19] Hasan M. M., M. Noor-E-Alam, Patel M. R., Modestino A. S., Young G. Sanchez L. D., A Novel Big Data Analytics Framework to Predict the Risk of Opioid Use Disorder. 2019.
  - [20] R. T. Perdue, J. Hawdon, and K. M. Thames, "Can big data predict the rise of novel drug abuse?" *Journal of Drug Issues*, vol. 48, no. 4, pp. 508–518, 2018.
  - [21] A. Furnham, "A big five facet analysis of sub-clinical dependent personality disorder (dutifulness)," *Psychiatry Research*, vol. 270, pp. 622–626, 2018.
  - [22] E. Salagre, E. Vieta, and I. Grande, "Personalized treatment in bipolar disorder," in *Personalized Psychiatry*, pp. 423–436, Academic Press, Cambridge, MA, USA, 2020.
  - [23] L. Granka, "Inferring the public agenda from implicit query data," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Boston, MA, USA, July 2009.
  - [24] V. Sinha, 2019, <https://www.quora.com/What-are-the-main-differences-between-artificial-intelligence-and-machine-learning-Is-machine-learning-a-part-of-artificial-intelligence>.
  - [25] Lenhart A., Purcell K., Smith A., Zickuhr K., Social media & mobile internet use among teens and young adults. Millennials, Pew Internet & American Life Project, Washington, DC, USA, 2010.

## Research Article

# Application of Big Data Fusion Based on Cloud Storage in Green Transportation: An Application of Healthcare

Li Qin Hu <sup>1</sup>, Amit Yadav <sup>2</sup>, Asif Khan,<sup>3</sup> Hong Liu,<sup>4</sup> and Amin Ul Haq<sup>3</sup>

<sup>1</sup>Chengdu Neusoft University, Department of Information Management, Chengdu 611844, Sichuan, China

<sup>2</sup>Chengdu Neusoft University, Department of Information and Software Engineering, Chengdu 611844, Sichuan, China

<sup>3</sup>Department of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China

<sup>4</sup>Chengdu University of Technology, Department of Human Resources Management, Chengdu, Sichuan, China

Correspondence should be addressed to Li Qin Hu; [huliqin@nsu.edu.cn](mailto:huliqin@nsu.edu.cn) and Amit Yadav; [amit@nsu.edu.cn](mailto:amit@nsu.edu.cn)

Received 7 January 2020; Accepted 6 May 2020; Published 29 June 2020

Academic Editor: Rodziah Binti Atan

Copyright © 2020 Li Qin Hu 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 21st century, transportation brought great convenience to people, but at the same time, automobile transportation is the major factor causing greenhouse gas emissions and climate change. Movements of the world towards green environments, there is hike in use and production of electric vehicles (energy vehicles). However, with the continuous growth in the number of energy vehicles, it is necessary for the government to provide strong support in the construction of charging piles. Real-time and effective management has become a practical problem for the relevant departments which needs to be solved. This paper uses the information research method to fuse the huge amount of heterogeneous data generated by the charging pile resultant to the new energy electric vehicle in the vehicle network and introduces cloud computing as its storage module to facilitate the storage and related expansion of the big data. This paper proposes a system scheme of heterogeneous data fusion based on cloud computing for the acquisition, storage, and fusion of heterogeneous data in the vehicle network. After testing the results, it shows that the system is stable and effective in practical application, which can meet the design requirements of the system. What is the significance of analyzing big data of charging point? Considering from the supply side, obtaining the user's charging behaviour data is helpful to build a digital map of the charging pile of new energy vehicles, connect the service information between the vehicle enterprises and the charging pile enterprises, and provide the most comprehensive and effective real-time charging information covering the widest range of vehicles, which can solve many problems of information asymmetry in the current charging information service.

## 1. Introduction

The importance of green transportation embodies not only the concept of sustainable development but also the impact of climate stability on human health [1]. China's new energy vehicles have a strong momentum of development. They have been developed rapidly in terms of model matching, technology research and development, and new energy vehicle consumer market and made breakthroughs in the fields of enterprises, technology, and market. However, they also face great challenges: weak industrial scale effect, high cost and price, short battery life, battery problems, and charging convenience problems in pure electric vehicles [2].

The first thing to be solved for electric vehicles is the battery problem, which lies in the weight and service life. The convenience of electric vehicle charging is also an important factor for mass production. A car with an ordinary tank capacity of 50 liters can fill up a tank of oil in 5 or 6 minutes. Whether the charging time of electric vehicles can be controlled within a few minutes is uncertain. A certain electric vehicle can only be charged 70% in 10 minutes at a special charging station, and it will take 6-7 hours to be fully charged at a household 220 V socket. Many cars of the people in China are parked in the underground parking lot or the parking space of the community, with few matching charging plugs. It is a good way to set up a charging station in

the gas station, but the battery charging time should be guaranteed, and the service life of the battery after repeated charging is also unknown. Therefore, the construction of charging piles will play an important role in the normal operation of new energy electric vehicles, and how to reasonably build charging piles has become an urgent problem to be solved by the government [3, 4].

In the long run, big data is expected to change the competitive ecology of the new energy vehicle market. On the one hand, big data is conducive to understanding customers' consumption preferences and realizing customized product services. In the future, automobile manufacturing is expected to present a new format of "hardware + software + services"; on the other hand, it is expected to connect big data with the power battery traceability and retirement platform, and the full life cycle management of the power battery is also expected to achieve a perfect combination of digitalization and informatization [5].

The establishment of a long-term mechanism for the healthy development of new energy vehicles is inseparable from the integration of industrial chain resources. The real value of big data lies in data analysis, deep mining data value, providing high-quality digital services, and promoting value achievements conducive to the development of the industry, so as to achieve the long-term development goal. Data acquisition is the premise of mining the big data value of new energy vehicles. Only by promoting the data sharing of new energy vehicles, can we play a greater market value and provide better services for users. However, it is difficult to realize the interconnection between industries at this stage. Charging difficulty is one of the important problems restricting the development of new energy vehicles, but it is not only the number of charging piles but also closely related to the charging service of charging pile enterprises. Now, there are many enterprises providing charging services. Charging data is the core data of the pile enterprises, so it is difficult to integrate the relevant data of the pile enterprises [6].

Big data application realizes the interconnection of cars, people, and piles, integrates the data of charging piles into mobile phones and car machines, and provides intimate data services for new energy owners at all times. Rich fun in use will also have many benefits for the promotion of new energy vehicles. Driving behavior is a very typical application scenario of big data. The quality of driving behavior not only affects the level of energy consumption but also affects the driver's driving portrait, including driving habits, charging habits, emergency response, and other scenarios [7, 8].

At present, many enterprises are fully mining the potential value of big data in the field of new energy vehicles, for example, through monitoring the operation data, improving the user's use behaviour, and realizing the intelligent health management of the whole vehicle and the power battery; through OTA system, realizing the upgrading of the BMS and ensuring the charging safety of the vehicle; building an intelligent interactive system between the BMS on the edge and the cloud, making the functions of the BMS on the edge more accurate; building a high-precision calculation model, eliminating the estimation deviation to the greatest extent, and realizing high-precision electricity pool

capacity calculation; building the system model of battery capacity and battery safety to realize safety early warning; and analyzing the attenuation data of the power battery, realizing rapid classification of echelon utilization, and optimizing the whole life cycle management of the power battery [9, 10].

The integration of massive heterogeneous data in the Internet of Vehicles is an important technical means to build a green city and green transportation. With the support of Internet technology, valuable information can be quickly and accurately extracted from massive traffic data, and the foresight, initiative, timeliness, and coordination of traffic management can be greatly improved. Under the background of the rapid development of the Internet of Vehicles, heterogeneous data fusion in the Internet of Vehicles based on cloud computing will certainly play an important role in the improvement of road traffic management, so as to make green traffic more "sustainable" [5].

## 2. Big Data Foundation of New Energy Vehicles

The traditional data analysis field is mainly based on structured data such as table data, which are relatively solid. With the rise of computers, Internet of Things, and other technologies, a large number of unstructured data, such as images, sounds, and videos, began to emerge, and the data scale also showed an explosive growth trend. To fully understand the basis of big data and accurately grasp the characteristics of big data will help to mine the intrinsic value of massive data [11].

*2.1. Big Data Features.* In 1890, Hollerith, an American statistician, invented an electric machine to make statistics of American census data and completed the expected work for eight years in one year, which is considered as the earliest application example of the big data method. At the end of the 20th century, human society stepped into the era of computer and Internet, and data also entered a period of explosive growth. In 2008, the global data volume was only 0.49 ZB, while in 2017, the global data volume was 21.6 ZB, 44 times of that in 2008. Researchers predict that, by 2020, the global data volume will reach 35 ZB. Take today's famous Internet enterprises as an example, Google needs to process nearly 100 Pb of data every month, and Taobao's daily online transaction data reach 10 TB [3].

However, big data is not only a large-scale data set but also it is biased to simply define big data in terms of quantity. Laney [12], an analyst at Meta Group, put forward that there are three major challenges in big data management in the future: volume, velocity, and variety. On this basis, some researchers supplement the two concepts of veracity and value depth, forming the "5V" characteristics of big data. The "5V" characteristics of big data are mainly reflected in its processing, calculation, and storage process. However, the traditional technology is not competent for big data analysis and processing nor can it realize real-time online calculation of big data. At the same time, the traditional data processing technology is mostly based on structured data, which cannot

deal with unstructured data such as text, pictures, and media. The development of big data processing technology is an effective way to solve the current data processing needs [12].

*2.2. Big Data Processing Technology.* Before the emergence of modern big data processing architecture, technicians used the MPI (Message Passing Interface) programming model and method to process large-scale data. MPI is a kind of high-performance parallel message passing interface, which is the main data programming and computing carrier at that time. It can make full use of hardware resources for parallel computing and is widely used in physical, meteorological, and other fields [13].

Due to the lack of good architecture support, low degree of automation, complex programming, and heavy tasks of programmers, researchers developed Hadoop MapReduce processing system. MapReduce is mainly for parallel processing of large-scale data. It was first developed by Google's research team for internal employees to process data. After that, the technical team of Apache Nutch expanded MapReduce to Hadoop MapReduce, an open-source parallel computing framework system based on Java language. With its outstanding functions of task scheduling, data recovery, and system optimization, it has become the mainstream big data processing system, which is widely used in academia and industry [14].

MapReduce is designed for offline batch processing of data. When online rapid data processing is required, MapReduce efficiency is low. The Spark big data processing system developed in 2013 absorbed the advantages of Hadoop MapReduce, greatly improved the parallel computing performance, made up for the shortcomings of the latter in data real-time computing, and made the modern big data analysis technology more complete. At the beginning of Spark, Scala, a professional functional programming language, was used as the development language, which restricted the use and promotion of Spark. Many common programming languages (such as Python and R) support the addition of functions, as well as the update of the data structure dataset. Spark is gradually accepted by the majority of data researchers [15].

In addition to Spark, Flink from Europe is also a commonly used parallel big data processing system. Flink supports both streaming and batch computing and has rich data conversion interfaces. Different from Spark, Flink has a unique storage management mechanism, which can save a lot of computing space. At the same time, it can automatically optimize the program to avoid redundant result cache. It provides a variety of programming language interfaces such as Java, Scala, and Python to further facilitate the use of users; Flink also provides table computing, complex event processing, and other big data computing libraries, which can be integrated with other mainstream processing systems. Users can choose corresponding processing systems flexibly and easily according to their actual needs [16].

The aforementioned big data processing systems can be divided into four categories according to the processing objects and processing forms: batch processing system,

streaming real-time processing system, real-time interactive query system, and graph data processing system.

*2.3. National Monitoring and Management Platform for New Energy Vehicles.* The premise of the combination of big data technology and new energy vehicles is to establish a big data platform to efficiently collect massive data resources. In order to solve the safety problems of new energy vehicles in China, improve the supervision of the new energy vehicle industry, and promote the development of the new energy vehicle industry, the Ministry of Industry and Information Technology established the national monitoring and management platform for new energy vehicles (hereinafter referred to as the national platform) in Beijing in 2016. By 2019, the number of vehicles connected to the national platform has exceeded 2.2 million. It is estimated that 7 million vehicles will be connected in 2020 and 80 million in 2025. The establishment of the national platform plays an important supporting role for the government to strengthen the safety supervision of new energy enterprises and vehicles [17].

The national platform architecture is mainly based on Linux system and Java programming language and is built with Hadoop system. Hadoop is the mainstream big data processing architecture at present. There are many precedents at home and abroad that adopt Hadoop architecture to build big data platform, covering medical, banking, rail transit, power system, and other fields. Its mode has been very mature [14, 15].

The existing data types of the national platform are mainly divided into static data and dynamic data. Static data, also known as file data, consist of basic vehicle information, such as license plate number, vehicle VIN number, vehicle manufacturer, vehicle type, and sales area. The data types of dynamic data are divided into online real-time running data and offline storage historical data. The difference between the two types of data lies in different stages and storage locations. The real-time operation data are the current transfer data, which are constantly updated and replaced and stored in the real-time cache so that the staff can monitor the safety of vehicle operation. The replaced data will be converted into historical data and stored in a dedicated server for researchers to call and check [15].

There are three kinds of data frame intervals of real-time operation data: 1 s, 10 s, and 30 s. According to the requirements of GB/T 32960, the data items are mainly collected from the following systems: power battery system, motor drive system, vehicle control system, and other parts. The data of the power battery system mainly include the total voltage and current of the battery system, SOC, cell voltage, and characteristic point temperature of the battery system. The data of the motor drive system mainly include motor voltage and current, speed, torque, and temperature. Vehicle control system data mainly include vehicle speed, gear information, accelerator pedal travel, and GPS position. In addition, there are air conditioning information, tire pressure status, and other information data [6].

Based on the principle of privacy protection, new energy vehicles in the private sector only transmit complete

monitoring data in the event of failure warning. In the field of public transport, new energy buses, taxis, and logistics vehicles all transfer complete data around the clock to ensure the safety of public transport. The national platform mainly performs the industry regulatory responsibilities, while the researchers use the operation data to analyze and study the battery system, driving behaviour, vehicle energy consumption, charging behaviour, etc., so as to promote the overall development of the new energy automobile industry [18].

*2.4. Foreign Development Status.* At present, global energy and environmental systems are facing huge challenges. As a major player in oil consumption and carbon dioxide emissions, revolutionary changes are needed. At present, the global new energy vehicle development has reached a consensus. In the long run, pure electric drive, including pure electric and fuel cell technology, will be the main technical direction of new energy vehicles. In the short term, hybrid electric and plug-in hybrid power will be an important transition route. At present, the development of global new energy vehicles still faces some common problems, such as breakthroughs in key technologies, transformation of the automobile industry, construction of infrastructure, and consumer acceptance [19]. Specific to each country, it should be said that the main leaders in the development of new energy vehicles are the United States, Japan, and some European countries. These countries started much earlier than China, and their development focuses on each [19].

The United States has long focused on strategies to reduce oil dependence and ensure the safety of new energy. It has taken the development of new energy vehicles as an important measure to fundamentally get rid of oil dependence in the transportation field and determined the strategic position of new energy vehicles in the form of laws and regulations. As early as the Clinton period, the United States proposed plans to improve fuel economy, and hybrid was the main technical solution at the time. In the Bush era, it became a pursuit of zero emissions and zero oil dependence. The technical solution was mainly hydrogen fuel-cell vehicles. Later, there was a plan to achieve 20% oil replacement and savings in ten years. The main measure was biomass fuel. After the international financial crisis, the Obama administration will vigorously develop electric vehicles as an important part of the implementation of the new energy strategy. It has proposed a total of 4 billion US dollars in power batteries and plans for the development and industrialization of electric vehicles. Focus on power electric vehicles [20].

Compared with the United States and Japan, Europe focuses more on greenhouse gas reduction strategies. Meeting increasingly stringent carbon dioxide emission restrictions has become a major driving force for the development of new energy vehicles in Europe. The development of new energy vehicles in Europe in the early days was mainly based on biomass fuels, natural gas, and hydrogen fuels. At the beginning of this century, a 23% oil

replacement target was proposed by 2020. Recently, Europe has paid great attention to electric vehicles. For example, Germany attaches great importance to the development of electric vehicles driven by pure electric power, focusing on pure electric power, and put forward the industrialization and marketization goals of 2012, 2016, and 2020, respectively [21].

### 3. Overall System Design

Aiming at the data obtained by the new energy electric vehicle management system, it takes a lot of manpower and financial resources to find the charging pile [22]. This paper builds a system hardware platform, including the heterogeneous data fusion of the vehicle network and the massive heterogeneous data application of the cloud storage [16]. In this paper, three ways of data collection and storages are used, i.e., data acquisition, data storage, and the data fusion display. The data acquisition end is responsible for data collection and data uploading; the data storage end uses cloud storage, which is responsible for classified storage of the vehicle network heterogeneous data; and the data fusion display layer communicates with the data storage layer to realize the display of related heterogeneous data fusion. Its complete architecture is shown in Figure 1.

The mobile terminal is also the data acquisition end used for the Internet of Vehicles. It is mainly a smart device with built-in sensors and Android operating systems, such as smart phones and smart rear-view mirrors. The server and database side use a distributed system of three hosts, including distributed file system (HDFS), nonrelational database MongoDB, and relational database MySQL. The hardware device of the data fusion display refers to the laptop or desktop computer. Later realizes the application of related data fusion by remotely calling is of three kinds, i.e., heterogeneous data of text, picture, and video stored on the server.

### 4. Data Acquisition Module

One advantage of the Android operating system is that it has a rich application for web application integration, and users can easily develop it according to their own needs. This module is the bottom layer of the entire IOT heterogeneous data fusion system [16]. The sensing layer of the IOT three-tier architecture is the data source of the entire system, mainly responsible for collecting real-time GPS data information, picture information, and video information of vehicles. The vehicle data collected by the Android data module laid the foundation for the subsequent implementation of vehicle management [23].

In the whole system, a smart device with an Android operating system is used as a data acquisition end. The smart device itself comes with a variety of sensors, such as GPS positioning systems, cameras, and wireless network cards. The real-time GPS position information of the vehicle is obtained by the sensor provided by the device, and the camera and the video recording function can be performed by using the camera. Heterogeneous data collected

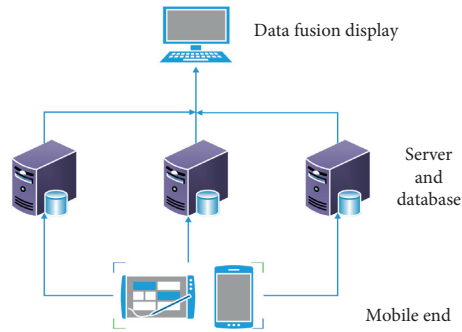


FIGURE 1: System overall architecture diagram.

throughout the process can be uploaded to the cloud storage database server via wireless network or data traffic. The specific Android data acquisition end architecture is shown in Figure 2.

The main function of the Android data acquisition end is to help the user registration management module of the rights management, obtain the real-time GPS text information, and upload it to the positioning information module of the MySQL relational database. Select picture from the photo stored place and upload the picture data to the real-time. MongoDB nonrelational database image upload module and real-time video capture, and upload it to the HDFS for storage video upload module [24].

### 5. Cloud Storage Data Module

This module is mainly used to save the data uploaded by the Android data acquisition module in real time. As shown in the cloud storage model of Figure 3, the state of data-related storage is represented. The main purpose is to classify and store the three heterogeneous data of text, picture, and video and select the appropriate storage system for related storage according to its characteristics. For the text data, the relational database MySQL was selected, and for the image data, the nonrelational database MongoDB was selected. The video data were stored in the distributed file system (HDFS) because they occupy more memory.

*5.1. MySQL Text Data Storage.* The text data uploaded by the data acquisition end are stored by the MySQL database, and the main uploaded text information is GPS real-time location information. There are 7 fields in the uploaded data, as shown in the GPS data field in Table 1, where the main fields are specifically distinguished. A few portions of storage data in the MySQL are shown in Table 2.

The data obtained by the Android client are first uploaded to the Tomcat server, and then the data in the Tomcat server are transferred to the MySQL database. The content stored in the MySQL database has ID (self-growth ID), Longitude (longitude), Latitude, Time, Serial number, Mac (IMSI code), Remark (IMEI code).

The upload process is real time. As long as the Android client application is started, the acquired GPS text data are uploaded to the corresponding Tomcat server in real time, and the data acquired in the Tomcat server are also uploaded

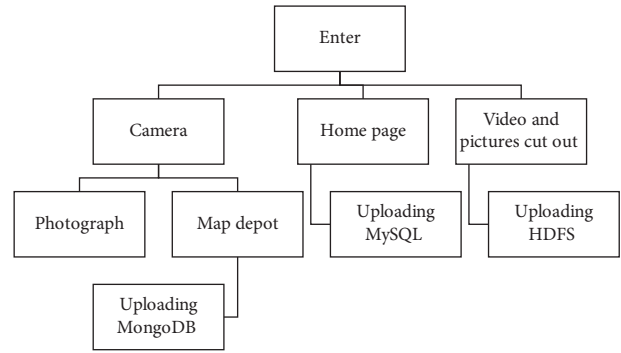


FIGURE 2: Android data acquisition end architecture.

Saas Software services	User system management	Vehicle path tracing	Real-time vehicle location
Paas Application service	MySQL	MongoDB	HDFS
Paas Application runtime	JAV	Node.js Tomca	Vmware

FIGURE 3: Cloud storage model.

TABLE 1: GPS data fields.

Number	Field	Type	Description
0	ID	Int	Label
1	Longitude	String	Longitude
2	Latitude	String	Latitude
3	Time	String	Time
4	Serial number	String	SIM serial number
5	Mac	String	Device IMEI
6	Remark	String	SIM IMSI

to the corresponding MySQL relational database in real time. The data stored in the MySQL database can be used for path backtracking, as well as for related applications such as positioning. In the whole system, a smart device with an Android operating system is used as a data acquisition end. The smart device itself comes with a variety of sensors, such as GPS positioning systems, cameras, and wireless network cards. The real-time GPS position information of the vehicle is obtained by the sensor provided by the device, and the camera and the video recording function can be performed by using the camera. Heterogeneous data collected throughout the process can be uploaded to the cloud storage database server via wireless network or data traffic.

*5.2. MongoDB Image Data Storage.* The image storage module uses a MongoDB distributed nonrelational database cluster built by three hosts. The distributed cluster adopts the

TABLE 2: Table storage status in the MySQL database.

ID	Longitude	Latitude	Time	Serial number	Mac	Remark
2	102.732794	25.05347	2018-12-26 10:27:49	8906116886010600124	864032030290728	460019775446990
3	102.732794	25.05347	2018-12-26 10:27:55	8906116886010600124	864032030290728	460019775446990
4	102.732794	25.05347	2018-12-26 10:28:01	8906116886010600124	864032030290728	460019775446990
5	102.732794	25.05347	2018-12-26 10:28:04	8906116886010600124	864032030290728	460019775446990
6	102.732794	25.05347	2018-12-26 10:28:09	8906116886010600124	864032030290728	460019775446990
7	102.732794	25.05347	2018-12-26 10:28:16	8906116886010600124	864032030290728	460019775446990
8	102.732794	25.05347	2018-12-26 10:28:19	8906116886010600124	864032030290728	460019775446990
9	102.732794	25.05347	2018-12-26 10:28:24	8906116886010600124	864032030290728	460019775446990
10	102.732794	25.05347	2018-12-26 10:28:33	8906116886010600124	864032030290728	460019775446990
11	102.732794	25.05347	2018-12-26 10:28:36	8906116886010600124	864032030290728	460019775446990
12	102.732794	25.05347	2018-12-26 10:28:45	8906116886010600124	864032030290728	460019775446990
13	102.732794	25.05347	2018-12-26 10:28:47	8906116886010600124	864032030290728	460019775446990
14	102.732794	25.05347	2018-12-26 10:28:55	8906116886010600124	864032030290728	460019775446990
15	102.732794	25.05347	2018-12-26 10:28:57	8906116886010600124	864032030290728	460019775446990
16	102.732805	25.053522	2018-12-26 10:29:54	8906116886010600124	864032030290728	460019775446990
17	102.732794	25.05347	2018-12-26 10:29:59	8906116886010600124	864032030290728	460019775446990
18	102.732794	25.05347	2018-12-26 10:30:04	8906116886010600124	864032030290728	460019775446990

form of Sharing + Replica Sets. Sharing is used to add related machines to slice large files for storage. Replica Sets ensure that each shard node has automatic backup and automatic failover capabilities [24].

In the cloud storage system build, the operating system of all nodes is CentOS-7-x86\_64-DVD-1161.iso. The MongoDB cluster consists of three servers: Server A: 192.168.118.100, Server B: 192.168.118.101, and Server C: 192.168.118.102. In Table 3, the MongoDB architecture diagram is shown [24].

TABLE 3: MongoDB architecture.

	Server A	Server B	Server C
Replica Set 1	Mongod shard 1-1	Mongod shard 1-2	Mongod shard 1-3
Replica Set 2	Mongod shard 2-1	Mongod shard 2-2	Mongod shard 2-3
3 Config	MongodConfig 1	MongodConfig 2	MongodConfig 3
3 Mongos	Mongos 1	Mongos 2	Mongos 3

5.3. *HDFS Video Data Storage.* Install a fully distributed cluster of Hadoop to store video data uploaded by the data acquisition end. Because the HDFS is a distributed file system used in common hardware devices, HDFS is highly fault-tolerant and can be deployed on low-cost hardware. It provides high-throughput features for accessing application data and is suitable for applications with very large data sets [25]. So, the video storage module uses a three-host Hadoop fully distributed cluster. The cluster-related node allocation status is shown in Figure 4.

The sHadoop (2.5.2) cluster configuration can be divided into two steps. The first step is configured on Hadoop252. The second step uses the SCP command to copy the configuration file to the slave 01 and slave 02 subnodes.

Three virtual hosts are created through VMware Workstation Pro for platform testing. Hadoop clusters are assigned as Hadoop252 as the management node of HDFS, slave01 as the management node of yarn, hadoop252, slave01, slave02 host installed processes as storage data nodes, yarn Node Manager node as shown in Figure 4.

In order to upload video data to the HDFS in real time, it needs to be mounted by NFS (Network File System). The main function of NFS is to achieve file sharing across networks, which allows computers on the network to share resources over a TCP/IP network. A local NFS client application can transparently read and write files located on a remote NFS server just as if it were a local file. This article

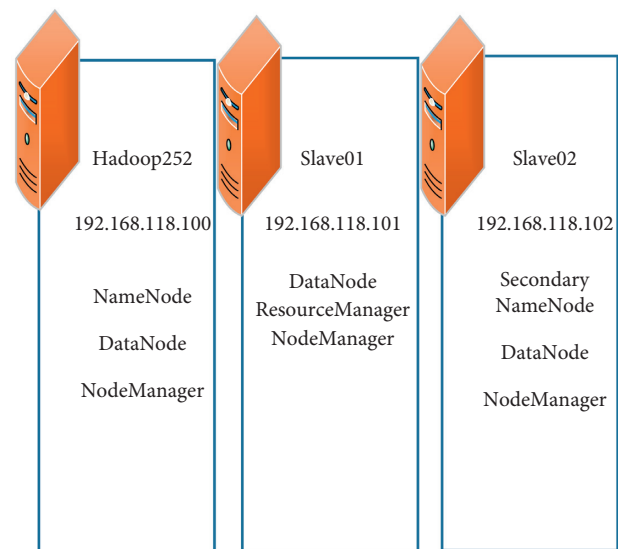


FIGURE 4: Hadoop distributed cluster.

mainly uses NFS to mount HDFS to a local directory. File sharing can be performed between the two directories. The display content is the same. Therefore, the Android client video data are uploaded to the HDFS and uploaded to the local directory in real time. In this way, the video data are uploaded to the HDFS in real time [26].

## 6. Data Fusion Display Module

With the rapid development of the Internet, many information systems related to the Internet of Vehicles have gradually shifted from C/S architecture to Web application form based on B/S architecture. This module is mainly developed with the B/S architecture. The development process uses the relevant functions of the server and then implements the corresponding functions on the browser side through the related call work. The main modules for data fusion implementation include login registration, real-time location, and path backtracking [16].

**6.1. Node.js Real-Time Location.** Real-time location display application is developed through the JavaScript language on the Node.js platform, mainly based on Baidu Map as the result of the system display page and also the page that collects user information and interacts with the user. JavaScript calls Baidu Map through the Baidu Map API to add custom function components to meet the needs of users. Baidu Map API is a set of application interface based on Baidu Map service provided free for developers. Users can introduce Baidu Map API in JavaScript code by using `<script>` tag to introduce Baidu Map API in the page. Good result data are presented in a graphical interface on the map.

When the Android client wirelessly transmits the location information, picture information, and video information to the server, the server performs extraction, analysis, and processing. The vehicle is monitored in real time according to the location information uploaded by the client, and the real-time image is displayed in the background of Baidu Map in combination with the uploaded real-time image, and the real-time location information and corresponding picture information are displayed when the corresponding vehicle is clicked.

**6.2. Java Web Path Backtracking.** The dynamic display of the page is inevitable using Ajax technology, which is a web development technology for creating interactive web applications, which can dynamically refresh the display of a certain part of the page. This part of the path backtracking is through Ajax technology, dynamically displayed to the user.

The application implementation process is mainly carried out from the following three aspects, namely, data reading, loading data, and map page.

**6.2.1. Data Reading.** Querying the serial number, time, and GPS latitude and longitude of the device dynamically stored in the MySQL database and providing the data needed to draw the vehicle trajectory.

**6.2.2. Loading Data.** This application is developed through Java Web. By processing the query data received by the page and processing the processed data as the filtering condition of the SQL query data, the queried record is encapsulated into a Java class and processed and transmitted to the page for processing. Map development use.

**6.2.3. Map Page.** Through the Baidu Map API call, secondary development for Baidu Map, draw the vehicle historical running track according to the query data, complete the page layout in the browser, collect the query conditions, and query the request to send and respond to the data. Processing, complete filtering of the location information of the vehicle and the serial number of the device from the returned JSON object, and finally implementing the path backtracking function.

## 7. Conclusion

In this paper, the massive heterogeneous data generated by the charging piles corresponding to the new energy electric vehicles in the Internet of Vehicles are fused, and cloud computing is introduced as its storage module, which is convenient for dealing with the storage and related expansion of massive data. For the problem of heterogeneous data acquisition, storage, and fusion in the Internet of Vehicles, a system scheme of heterogeneous data fusion method in the Internet of Vehicles based on cloud computing is proposed. After testing, the system runs stably and effectively in practical application and can meet the design requirements of the system.

In the future, if the taxi demand data, charging pile data, and vehicle operation data can be fully connected, then the vehicle operation, charging pile information, vehicle residual power, order distribution, etc. will be comprehensively optimized and upgraded, which will be of great benefit to the whole travel ecology. Data will drive continuous innovation in R&D, products, manufacturing, and supply chain and business model, and automobile will build a new industrial ecosystem around big data. At the same time, massive data drive the rise of computing and analysis platform, and vehicle enterprises urgently need to build computing soft power to win future differentiated competition.

## Data Availability

The authors confirm that the data supporting the findings of this study are available within the article.

## Conflicts of Interest

The authors declare that there are no conflicts of interest.

## Authors' Contributions

Li Qin Hu, Amit Yadav, and Asif Khan were responsible for conceptualization, methodology, and writing and preparing the original draft. Hong Liu was responsible for writing and preparing the original draft, validation, formal analysis, and supervision of the study. Amin Ul Haq worked on facts visualization.

## References

- [1] Sustainable Development Goals, "17 goals to transform our world," 2018, <http://www.un.org/sustainable-development/sustainable-development-goals/>.



- [2] M. Friman, J. Huck, and L. E. Olsson, "Transtheoretical model of change during travel behavior interventions: an integrative review," *International Journal of Environmental Research and Public Health*, vol. 14, no. 6, p. 581, 2017.
- [3] Y. Shi, X. Ming, and D. Yin, "Research on architecture of automotive product data management and digital process application cloud platform," *Machine Design & Research*, no. 1, p. 37, 2017.
- [4] M. Xiao, *Applied Research of Intelligent Traffic Information Collection and Fusion Technology*, East China Jiaotong University, Nanchang, China, 2016.
- [5] M. Li, "Models and suggestions for the integration development of "Internet+transportation"," *Journal of Xinyang Teachers College (Philosophy and Social Sciences Edition)*, vol. 1, pp. 61-65, 2017.
- [6] J. R. Wen, S. Zheng, and H. Lu, "Semi-structured data storage schema selection," 2017, <https://patents.google.com/patent/US20060053127>.
- [7] Y. Cao, H. Song, O. Kaiwartya et al., "Mobile edge computing for big-data-enabled electric vehicle charging," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 150-156, 2018.
- [8] I. Semanjski and S. Gautama, "Smart city mobility application-gradient boosting trees for mobility prediction and analysis based on crowdsourced data," *Sensors*, vol. 15, no. 7, pp. 15974-15987, 2015.
- [9] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599-616, 2009.
- [10] K. Zhou, H. Wang, and C. Li, "Cloud storage technology and its application," *ZTE Technologies*, vol. 16, no. 4, pp. 24-27, 2016.
- [11] R. Scotini, I. Skinner, F. Racioppi, V. Fusé, J. Bertucci, and R. Tsutsumi, "Supporting active mobility and green jobs through the promotion of cycling," *International Journal of Environmental Research and Public Health*, vol. 14, no. 12, p. 1603, 2017.
- [12] D. Laney, "3D data management: controlling data volume, velocity and variety," *META Group Research Note*, vol. 6, no. 70, p. 1, 2001.
- [13] A. Khan, J. P. Li, F. Hasan, and R. Alam, "Real world image analysis with CBIR in cloud database," *Advances in Computer Science and Information Technology (ACSIT)*, vol. 1, no. 3, pp. 140-143, 2014.
- [14] H. Ding, D. Zhang, and Z. Luo, "Research on massive data processing platform based on hadoop, 2011 power communication management and smart, grid communication technology forum," 2016.
- [15] Q. Zeng, *Cloud-based Storage Technology Based on Heterogeneous Hadoop in Vehicle environment*, Nanjing University of Posts and Telecommunications, Nanning, China, 2016.
- [16] E. Zhuang, *Research on Data Fusion of Internet of Things*, Shijiazhuang Tiedao University, Shaoxing, China, 2017.
- [17] N. Xiao, *Design and Implementation of Vehicle Positioning System in Vehicle Network Environment*, Beijing Jiaotong University, Beijing, China, 2016.
- [18] D. Zhai and C. Chen, *Detailed Explanation of Android Project Development*, Mechanical Industry Press, Beijing, China, 2015.
- [19] Z. Han, *Research on the Design and Key Technologies of Expressway Traffic Information Platform Based on the Internet of vehicles*, School of Transportation, Jilin University, Changchun, China, 2016.
- [20] Y. Jiang, *Android System Principle and Practical application*, Beijing Institute of Technology Press, Beijing, China, 2015.
- [21] N. Yuan, *Design and Development of Vehicle Remote Monitoring System Based on Android Smart phone*, Chongqing University, Chongqing, China, 2014.
- [22] H. Gao, "Design and implementation of C/S mode video surveillance system," *Journal of Chengdu University of Information Technology*, vol. 4, pp. 386-389, 2014.
- [23] N. Peng, *Design and Implementation of a Blog System Based on Node JS*, Dalian University of Technology, Dalian, China, 2016.
- [24] G. E. Tara and C. Pu, *MongoDB Combat Architecture, Development and Management*, Tsinghua University Press, Beijing, China, 2017.
- [25] Y. Li, *Research on Web Services Technology Based on REST Architecture*, Wuhan University of Technology, Wuhan, China, 2016.
- [26] N. Zhang, *Research on Real-Time Data Processing Based on Data Fusion*, Wuhan University of Technology, Wuhan, China, 2016.

## Review Article

# Biomedical Relation Extraction Using Distant Supervision

Nada Boudjellal <sup>1</sup>, Huaping Zhang <sup>1</sup>, Asif Khan <sup>1</sup>, and Arshad Ahmad <sup>1,2</sup>

<sup>1</sup>School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

<sup>2</sup>Department of Computer Science, University of Swabi, Anbar, Pakistan

Correspondence should be addressed to Huaping Zhang; kevinzhang@bit.edu.cn

Received 20 March 2020; Revised 22 May 2020; Accepted 27 May 2020; Published 16 June 2020

Academic Editor: Shaukat Ali

Copyright © 2020 Nada Boudjellal 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.

With the accelerating growth of big data, especially in the healthcare area, information extraction is more needed currently than ever, for it can convey unstructured information into an easily interpretable structured data. Relation extraction is the second of the two important tasks of relation extraction. This study presents an overview of relation extraction using distant supervision, providing a generalized architecture of this task based on the state-of-the-art work that proposed this method. Besides, it surveys the methods used in the literature targeting this topic with a description of different knowledge bases used in the process along with the corpora, which can be helpful for beginner practitioners seeking knowledge on this subject. Moreover, the limitations of the proposed approaches and future challenges were highlighted, and possible solutions were proposed.

## 1. Introduction

Information extraction (IE) is the task of getting structured information out of unstructured or semistructured text, where the goal is to extract the relevant data found in a massive amount of text in a structured format which can be used by an end-user or other computer systems (i.e., databases or search engines) [1, 2]. Given, for example, the sentence “William Shakespeare was born in 1564; he wrote of The Tragedy of Romeo and Juliet,” information extraction can discover the following information:

BornIn (William Shakespeare, 1564)

WrittenBy (The Tragedy of Romeo and Juliet, William Shakespeare)

With the growth of the Internet and thus the expansion of the amount of data coming with it, the need for information extraction systems has been growing exponentially.

Medical domain has its share of data expansion with more than 30 million citations of biomedical literature found in PubMed [3] and an endless amount of electronic health records (EHR); this makes it hard for biomedical researchers to discover facts about a specific biomedical entity (i.e., gene, protein, disease, etc.) automatically and timely. Thus, it is

critical to harvest information and knowledge from unstructured medical data using information extraction systems.

Two of the most important subfields of IE are (1) named entity recognition and (2) relation extraction. The former focuses on extracting relevant entities from the text, while the latter deals with discovering and disambiguating semantic relationships between those entities. The focus of this work will be on relation extraction.

Relation extraction from the biomedical literature is an essential task for building a biomedical knowledge graph, which can provide useful and structured information for the healthcare research community. The methods used for this task can be categorized into four groups: (1) rule-based methods [4, 5]; (2) supervised methods [6, 7]; (3) unsupervised [8]; and (4) minimally supervised methods (semi-supervised [9] and weakly supervised are its examples). Although rule-based and supervised methods can achieve high accuracy results, the first is considered nowadays old fashioned because of the enormous effort spent in hand-crafting rules, while the second is expensive in matters of time and cost spent in labelling data mainly in the biomedical field. Therefore, recent work on relation extraction focused on using minimal supervision methods to tackle the

biomedical relation extraction task to minimize the human intervention and, as a result, reduce the cost and time of labelling along with human error. Distant supervision is one of those promising approaches that aim to do all that while keeping good performance.

Much work has been done for RE featuring Distant Supervised Learning, mainly for general-domain data. Readers can refer to [10] for a detailed review of methods, knowledge bases, and dataset used for general-domain RE using distant supervision with a mention of some work done for the biomedical domain, besides metrics of evaluation used for this task which will not be covered in this paper. For biomedical RE, Zhou et al. [11] presented work conducted prior to 2014 regarding binary and complex biomedical RE. To the best of our knowledge, there has been no work surveying biomedical relation extraction using distant supervision. This paper targets the literature addressing the subject of biomedical RE in a distant supervised setting.

The main contributions of this work are as follows:

- (i) It overviews the topic of biomedical RE using DS in a simple, comprehensible format providing a generalized architecture
- (ii) It presents a review of papers addressing the subject of using distant supervision for biomedical relation extraction and discusses the methods used regarding this topic and which datasets were implied in the experiments
- (iii) It identifies the limitations of those methods and proposes some solutions
- (iv) This work can be considered as a reference for beginners aiming to indulge in the subject of Distant Supervision for biomedical RE

*1.1. Selection of Papers.* The papers in this work were selected after performing a search in four different relevant sources of research papers (Scopus, Web of Science, IEEE Xplore Digital Library, and ACM Digital Library). All the years were included in the search query. After getting the results of each query in each of the four libraries, they were filtered according to the scope of this paper, and then the duplicates were eliminated. The final set of papers is listed in Table 1. Figure 1 shows the propagation of published papers about biomedical relation extraction using distant learning through the years. It is observed that the number of publications regarding biomedical RE using distant learning is increasing since 2017, which shows somehow the need for distant learning in biomedical text mining and information extraction wise.

The remaining part of the paper is as follows: an overview of Distant Supervised Learning for RE is given in Section 2. In Section 3, the authors discussed the research done in biomedical relation extraction using distant supervision. Section 4 provides insight into possible limitations of presented literature and future challenges and directions. Finally, Section 5 concludes the paper.

## 2. Distant Supervised Learning for Relation Extraction

Distant supervision (DS) is an alternative way to generate labelled data automatically while making use of an available knowledge base (KB) [20], which can be general- or specific-domain KB to extract seed examples that will be used to train the model. Distant supervision allows the generation of an extensive training set with a minimum effort.

DS has been used for the task of relation extraction (RE) and was introduced first by Mintz et al. [24], who used it to create a large dataset for Freebase RE. In their work, the authors assumed that any sentence featuring a pair of entities that corresponds to a knowledge base entry is more likely to express a relation between those entities. Since most of the papers tackling the topic of relation extraction using distant supervision were inspired by Mintz et al.'s work, a generalized architecture of their method is presented in Figure 2.

The elements of this method are explained briefly in what follows.

*2.1. Knowledge Base.* According to the study [15], identifying a knowledge base that comprises the target relations is an essential matter in distant supervision since the annotation is supervised by the chosen knowledge base instead of manual annotation. In some approaches, the KB can be used to perform two tasks: the first is the identification of entities participating in the target relations, by using it as a lexicon; the second task is the extraction of positive examples of those relations. These knowledge bases can be a database or an ontology, and they are available—mostly all—freely for the biomedical domain [16]. Existing knowledge bases are mostly topic-oriented, focusing on one type of entities or relations such as the Protein Data Bank (PDB) [25], which contains a description of large biological molecules (proteins) along with their description and 3D structure.

*2.2. Corpus.* Choosing a compatible corpus with selected knowledge base can have a positive impact on the overall accuracy of the classifier. In the biomedical domain, the corpus consists of full-text biomedical research articles or just abstracts, mostly from PubMed, or online medical webpages data. Distant supervision involves large corpora [16].

*2.3. Generation of Training Examples.* After identifying the desired entities in the corpus, the assumption mentioned earlier is used to extract all candidate positive examples; i.e., take into consideration all the sentences mentioning two pairs of entities that express a relation in the knowledge base, which means that noisy data will be generated since not every sentence expresses the relation that links those pairs of entities in the KB.

One fallout of this assumption is that it can generate false positive, i.e., two entities may appear in the same sentence and correspond to an entry in our selected knowledge base,

TABLE 1: Selected papers with their date of publication.

ID	Title	Date of publication
1	Literature mining of protein-residue associations with graph rules learned through distant supervision [12]	2012
2	Improving distantly supervised extraction of drug-drug and protein-protein interactions [13]	2012
3	Relation extraction from biomedical literature with minimal supervision and grouping strategy [14]	2014
4	Using Distant Supervised Learning to identify protein subcellular localizations from full-text scientific articles [15]	2015
5	Extracting microRNA-gene relations from biomedical literature using distant supervision [16]	2017
6	A semi-automated entity relation extraction mechanism with weakly supervised learning for Chinese medical webpages [17]	2017
7	Distant supervision for relation extraction beyond the sentence boundary [18]	2017
8	HighLife: higher-arity fact harvesting [19]	2018
9	Using distant supervision to augment manually annotated data for relation extraction [20]	2019
10	Chemical-induced disease relation extraction via attention-based distant supervision [21]	2019
11	Distant supervision for treatment relation extraction by leveraging MeSH subheadings [22]	2019
12	CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision [23]	2019

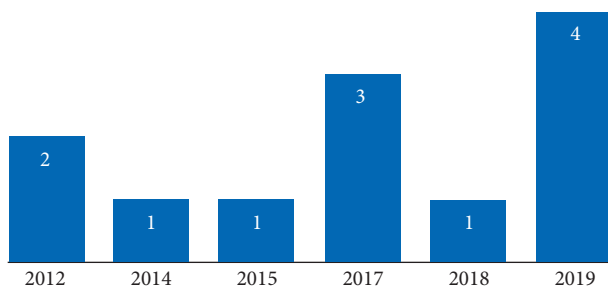


FIGURE 1: Propagation of published papers through the years from 2012 to 2019.

but they do not express that relationship in reality. An example to explain this point is as follows.

Saying that a KB of disease-virus pairs contains the relation: *CausedBy* (COVID-19, SARS-CoV-2).

COVID-19 is a disease caused by the SARS-CoV-2 virus

COVID-19 is continuing its spread worldwide, while scientists are trying their best to find a vaccine for the SARS-CoV-2

From the above sentences, it can be seen that although the second sentence mentions both entities COVID-19 and SARS-CoV-2, it clearly does not express the *CausedBy* relation as it is expressed in the first sentence. To overcome the problem of false positives resulting from this assumption, some authors tend to apply some changes to it, and that is what will be explained in Section 3.

**2.4. Features Extraction.** In their method, Mintz et al. considered two types of features:

- (1) Syntactic features: they are part of speech tags, dependency paths connecting the pair of entities
- (2) Lexical features: they describe words before, between, and after the pair of entities, for example, their POS tags

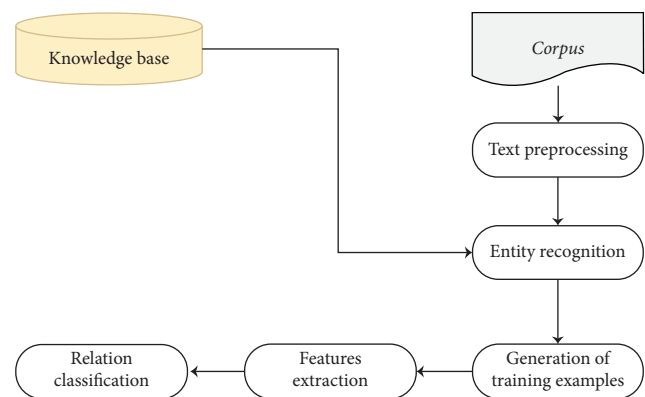


FIGURE 2: General system architecture of relation extraction using distant supervision, according to Mintz et al.

Each method used what comes better with it from those features for feature selection can have a significant impact on classification performance.

**2.5. Relation Classification.** In most cases, the relation extraction is considered a binary classification problem where the output is true or false. The next section will present the different classification methods used for RE in a distant supervised setting.

### 3. Methods and Approaches

As was mentioned previously, most approaches regarding relation extraction under distant supervision are inspired by Mintz et al. [24]; however, they differentiate from it in some points, namely, the classifier model they choose or how they handle the noise caused by their assumption. Table 2 gives an overview of the relations targeted in each selected paper, with a mention of the KB and corpora used in each, besides the results got in the RE task and NER task if available.

In the remainder of this section, the different classifying methods used for the RE task in biomedicine are presented with a description of the way the authors handled the noisy data if available.

TABLE 2: An overview of the relations targeted by each method with a mention of the resources used and the results obtained.

Paper	Relation type	Knowledge base	Corpora	NER results	RE results
[12]	Protein-residue	Protein Data Bank (PDB) [25]	PubMed abstracts	Evaluated on 3 gold corpora only for amino acid/mutation entities: Nagel et al. $F$ -measure = 93.28%/mutation finder: development ( $F$ -measure = 89.32%) and test corpora ( $F$ -measure: 88.04%) LEAP-FS corpus: $F$ -measure = 86.56%	0.84 $F$ -measure (silver corpus) 0.79 $F$ -measure (gold corpus)
[13]	Drug-drug protein-protein	IntAct database [26], KUPS database [27], DrugBank [28]	The five corpora of Pyysalo et al. [29]. The corpus of Segura-Bedmar et al. [30]	Not mentioned	Drug-drug (DDI) $F$ -score = 61.19 PPI $F$ -score = 78.0 on LLL corpus
[14]	Gene-brain regions	UMLS Semantic Network [31]	10,000 randomly selected full-text articles from Elsevier Neuroscience corpus	$F1 = 0.8$ (for 300 manually examined examples)	$F1$ -score = 0.468, recall = 0.459, precision = 0.477 (for 259 manually labelled sentence out of 30,000)
[15]	Protein-location	UniProtKB (Swiss-Prot) [32]	43,000 full-text articles from the Journal of Biological Chemistry	Not mentioned	$F1 = 0.61$ , $R = 0.49$ , $P = 0.81$ (sentence level) accuracy = 0.57 (RL instance level)
[16]	microRNA-gene	TransmiR database (nonhuman entries) [33]	IBRel-miRNA corpus	Evaluated on 3 corpora: Bagewadi corpus [34] ( $F = 0.919$ miRNA/ $F = 0.677$ gene), miRTex [35] ( $F = 0.941$ miRNA/ $F = 0.795$ genes), and TransmiR ( $F = 0.687$ miRNA/ $F = 0.361$ genes)	Evaluated on 3 corpora: Bagewadi corpus ( $F = 0.532$ ), miRTex ( $F = 0.383$ ), and TransmiR ( $F = 0.413$ )
[17]	Related symptoms, related diseases, related examination, complications, and related treatment	Not mentioned	Medical websites	Not mentioned	Accuracy = 91.87%, recall = 91.58%, $F1$ -score = 0.8908
[18]	Gene-drug	Gene Drug Knowledge Database (GDKD) [36]	Biomedical literature from PubMed Central	Not mentioned	Automatic evaluation best average test accuracy in fivefold cross-validation (single sentence: 88, cross sentence: 87.5) manual evaluation (precision = 71 for single sentence and 61 for cross sentence)
[19]	n-arity relations: Treats, ReducesRisk, Causes, Diagnoses	474 seed facts from online medical portals uptodate.com, drugs.com	Encyclopaedic articles and PubMed scientific publications	Not mentioned	Treats avg. precision: 0.86, ReducesRisk avg. $P$ : 0.82, Causes avg. $P$ : 0.80, and Diagnoses avg. $P$ : 0.89
[20]	Protein-protein, protein-location	IntAct database, UniProt database	Medline, literature found in IntAct database	Not mentioned	PPI (PCNN $F$ -score = 56.8 BiLSTM $F$ -score = 50.4) PLOC (PCNN $F$ -score = 54.5 BiLSTM $F$ -score = 60.4)
[21]	Chemical-disease	Comparative Toxicogenomics Database (CTD Database) [37]	PubMed abstracts	Not mentioned	Intrasentence level: best $F$ -score = 60.8; intersentence level: best $F$ -score = 22.8

TABLE 2: Continued.

Paper	Relation type	Knowledge base	Corpora	NER results	RE results
[22]	Binary treatment relation	UMLS database, SemMedDB [38]	PubMed abstracts for which there exist both the therapeutic use and the therapy medical subject headings (MeSH) subheadings	Not mentioned	PR-AUC: logistic regression: 82.86 BiLSTM:81.18 BiLSTM-NLL:81.38
[23]	Human disease-gene, tissue-gene, and protein-protein in different species	Genetics Home Reference (GHR) [39], UniProtKB, KEGG maps [40], STRING [41]	PubMed, full-text articles from PMC in BioC XML format [42]	Not mentioned	Adjusted area under the precision-recall curve (AUPRC): disease-gene: 0.86/tissue-gene: 0.19

*3.1. Graph-Based Approach.* Graph-based approach has been used by [12, 14] to extract protein-residue and gene-brain regions, respectively.

Ravikumar et al. [12] applied a dictionary lookup method on a compiled dictionary from BioThesaurus database [43] to extract protein entities while using defined patterns and regular expressions for amino acids and mutations entities extraction. After extracting positive examples, i.e., sentences containing pairs corresponding to entries of Protein Data Bank (PDB) [25], the authors constructed their silver corpus composed of 1728 PubMed abstracts related to proteins and divided it to training, development, and testing corpora. Later on, they used the graph-based rule induction method to learn the protein-residue relation rules from the training set. This method consists of calculating the union of all shortest-dependency paths binding a pair of entities then use it as an event rule. To extract relations from test sentences, they perform subgraph matching, i.e., search for a subgraph within the test sentence dependency graph that is similar to an event rule graph. To show their method efficiency, they tested it on golden corpora, i.e., manually annotated and their automatically generated silver corpus. They found that their distant supervised method for automatic generation of training data performed better than cooccurrence baseline methods. To address the false positives problem, the authors used a rule ranking strategy by ranking the rules according to their precision PRC ( $r_i$ ) (where  $r_i$  is a rule); according to the authors, rules with higher PRC ( $r_i$ ) tend to produce less false positives. This method helped in enhancing the precision of extracted relations.

After annotating the selected articles with brain and gene entities (using Brain dictionary and a tagger, respectively), Liu et al. [14] applied their grouping strategy consisting of creating parse trees of selected sentences and developing a set of heuristic rules to find parallel entities. Their next step is to extract features, which are the same syntactic and lexical features used in [24]. To generate training examples, they used a tool to get knowledge from the UMLS Semantic Network. Then, for each pair of entities, they designed an undirected graphical model that defines a conditional probability for extraction using the feature vector of sentences containing the pair of entities. In the end, the model,

given a pair of entities, predicts the relation type, whether it is a gene expression or other expression. The authors argue that grouping strategy performs better since it can discover more relations that are not available in the knowledge base; therefore, the recall will be higher. They tested their model at sentence level as well as corpus level.

*3.2. Machine Learning Classifiers.* Following Mintz et al., Zheng and Blake [15] used UniProtKB (specifically Swiss-Prot) knowledge base to detect protein and subcellular locations entities and to abstract positive examples. In their work, they considered using both lexical and syntactic features. For lexical features, only one was used (namely, the sequence of words between a pair of entities); as for syntactic ones, the dependency paths between entities were used. Then, they applied a binary Support Vector Machine classifier to classify protein-location relations. For the evaluation task, they used a manual approach by testing the predictions of the classifier manually and held out test. According to the authors, one of their work limitations is using the KB as a lexicon for NER, which makes the task of finding relations featuring entities not included in the KB an impossible mission.

In their work, Bobi et al. [13] used five corpora presented by Pyysalo et al. [29] for the Protein-Protein Interaction (PPI) extraction task. The features used in their work are bag of words and n-grams as lexical features while using dependency paths as syntactic ones. They used rich feature vectors along with an SVM classifier named LibLINEAR for their RE. They applied the same process for drug-drug relation instances using the DrugBank database. To solve noise issue, they presented an “autointeraction filtering” constraint that removes any pair containing entities referring to the same object in real world, i.e., for the relation instance  $r \langle e_1, e_2 \rangle$ , if  $e_1$  is identical to  $e_2$ , then this pair is labelled as negative.

Junge and Jensen [23] introduced a scoring method called CoCoScore to score the certainty of a relationship between a pair of entities in a sentence, i.e., it gives a score to considered positive examples generated using distant supervision. The logistic regression classifier scores give a score between 0 and 1 as a prediction whether the input example is

positive or negative; then the CoCoScore aggregates all the scores computed by the classifier over the whole dataset to get the final decision. They tested their method on three types of relations (see Table 2) and found that their scoring strategy gave a better performance than baseline methods.

Another way to alleviate noise in DS data is multi-instance learning (MIL), which, differently from traditional DS, instead of labelling each instance individually, it labels a bag of instances. Lamurias et al. [16] use a variant of MIL called sparse multi-instance learning (sMIL) for microRNA-gene RE task. This algorithm assumes that the bags are sparse, i.e., only a few instances are positive, which is true for distant supervision where false positives can occur. A bag is considered positive if it covers at least one positive instance; otherwise, it is negative. Features of each instance were learned and converted into a bag of words; then, an SVM classifier was implemented. The authors compared their method to supervised learning algorithms and found that it performed better on their automatically annotated corpus.

Where the previous literature focused only on extracting relations in single sentences, the authors of [18] worked on RE on an intersentence level. Similar to previous papers, they used a knowledge base (namely, Gene Drug Knowledge Database (GDKD) [36]) for their distant learning approach. After annotating the gene and drug entities using an existing tagger, and because they are working with cross sentence RE, the authors selected the pair of entities with minimal span, i.e., there is no overlapping cooccurrence of the same pair where the distance between those entities is smaller. In order to extract features on intra- and intersentence levels, they used a document graph where nodes represent words while edges characterize relations within and cross sentences (e.g., adjacency relations). The minimal span candidates mentioned earlier were filtered to leave only pairs that are within or less than three successive sentences. These candidates constitute the positive training examples, which will be fed, along with generated features and negative examples to a logistic regression classifier. The model was tested automatically using a fivefold cross validation and manually by asking experts to judge the correctness of 450 instances. Both evaluations showed the validity of their approach.

*3.3. Deep Learning Approaches.* Deep learning approaches showed their effectiveness since their appearance, so it is no surprise to see them used along with distant supervision techniques. Where neural networks need a huge amount of labelled data, using distant supervision to generate that data presents a profitable option.

The authors of [20] worked on augmenting manually labelled data for RE using DS. They focused on protein-protein and protein-location relations; therefore, IntAct and UniProt databases were used, respectively, to get training examples for each relation type. To reduce the noise, the authors used the heuristics chosen by [44]; some are applied to positive examples such as closest pairs and trigger words, while some are applied on negative examples such as high-confidence patterns heuristic. A full explanation of that heuristic can be found in their paper. For the classification

task, they chose two types of neural networks: PCNN (CNN based) and BiLSTM, which performed better when given more information about the input such as POS tag and entity type. To achieve their study objective, they used transfer learning to combine distant supervision generated data and manually labelled data.

Noisy labels were also considered by the authors of [22] to reduce it; they used the method of modifying the loss function to be noise resistant. Their work was a bit different from traditional DS, for they used MeSH subheadings to extract relevant articles to their study, i.e., the selected PubMed articles containing both Therapy and Therapeutic Use subheadings. The existence of both subheadings in an article indicates implicitly the existence of treatment relation. They use the UMLS database along with MeSH terms to extract positive example and in a mostly similar way the generated negative examples. In their experiments, the authors used two types of classifiers: logistic regression and BiLSTM-NLL which is a variant of BiLSTM with a loss function resistant to noise. Same as the study in [18], Precision-Recall Area under the Curve (PR-AUC) metric was used to compute the performance of the system since it is more suitable for unbalanced data, i.e., ratio of positive and negative samples is not 1 : 1.

The study in [21] combined both intra- and intersentence level relation extraction to extract a document-level RE. Training examples for their chemical-disease relation extraction task were generated with the aid of the Comparative Toxicogenomics Database using a multi-instance learning (MIL) paradigm. While aligning facts from the KB to PubMed dataset, a fact can be present in many single sentences; therefore, a bag of single-sentence level is created. The other scenario is that a fact is not present in any single sentence. Thus, a bag of cross-sentence level contains the nearest mentions of this pair of entities. An attention-based neural network was used for single-sentence level to minimize the noise by automatically weighting the generated instances where relevant ones get higher weights, while a stacked autoencoder neural network was proposed for intersentence level. Then, results from both classifiers were combined to get the document-level relations.

Liang et al. [17] proposed a method to extract relations between medical entities and their attributes located in different webpages within the same website. To achieve their goal, they first designed a visual labelling tool where the user can choose the entity and its attribute, whether it is on the same page or on a separate one; then patterns will be generated, and data will be extracted. The authors mentioned using weak supervision to extract training examples without mentioning which knowledge base they used for each relation they claimed they targeted. At the end, they used a CNN to extract relations.

*3.4. n-Arity Relation Extraction.* Limited work has been done for n-arity biomedical RE due to the complexity of the biomedical text and the complexity of complex relations themselves.

Ernst et al. [19] tackled this problem. Their method was applied for both newswire and biomedical data. They used

seed facts as a source of distant supervision. Each seed fact was used to deduct pattern trees from dependency graphs that were used to get fact candidates. False candidates were then eliminated using a constraint reasoning comprising a set of hand-crafted constraint rules. This step only leaves what they called salient trees, which express a highly confident n-arity fact and consequently increasing the precision. Named Entity annotation was performed using a set of resources; for biomedical data, which is the focus of this review, UMLS was used as the primary source of medical NE. The annotation was applied to a corpus that incorporates a group of PubMed biomedical literature, medical portal, and encyclopedic articles. Then, a number of 474 seed facts varying from binary to quinary were manually extracted for four types of relations (namely, Treats, ReducesRisk, Causes, and Diagnoses). To evaluate the performance of their suggested method, the authors used CrowdFlower Platform for Crowdsourcing according to which they achieved an average precision of 0.83.

#### 4. Limitations, Future Challenges, and Directions

This section states some limitations of the literature using distant supervision for RE in biomedicine, future challenges, and how it can be improved.

As entity recognition is a necessary step that cannot be skipped before relation extraction, it affects the performance of relation extraction [45]. If the entities' annotation has a high error rate, the accuracy of training examples generation will decline since some instances will be missing, and as a result, the whole process of relation extraction will suffer from inefficiency. To overcome this problem, more work should be done to enhance the accuracy and precision of NER. Aside from NER, the size of corpora was also a problem for researchers; Lamurias et al. [16] stated that having a larger corpus can lead to a flexible classifier for more instance structures can be taken into consideration, hence, more accuracy and precision.

The scarcity of golden data (manually annotated) makes the task of evaluation hard. That can be seen through some papers such as [14, 18] wherein the former, the authors manually labelled 259 sentences out of 30,000, while in the latter, only 450 instances were manually judged whether it is correct or not.

One problem that can occur while using the Knowledge base as a lexicon for entity recognition is that it is impossible to extract relations featuring entities that do not exist in the KB for all the generated instances will only contain entities of the KB, and that is what happened with [15]. Using machine-learning classifiers to annotate entities can solve this issue since the ML classifier is not bound with specific terms.

Since most biomedical knowledge bases are topic-oriented, i.e., focus on a specific entity or relation (drug or protein database [46]), it makes it difficult to generalize. However, this does not infer the fact that that databases with multientity types do not exist. One promising database is the UMLS database, which includes multiple concepts and links them with its semantic network.

Almost all discussed methods only focus on single sentence binary relations; though for a complicated domain such as healthcare, it is essential to spend more efforts on the extraction of n-arity relations, i.e., relations with more than two entities.

Considering the complex nature of biomedical text, devoting more work to extracting n-arity relations on an intersentence level can improve enormously the biomedical relation extraction, especially when under a distant supervised environment, which can permit achieving good performance with less cost and time.

#### 5. Conclusion

Over the last decade, Distant Supervised Learning is growing towards being of great importance for information extraction tasks in the biomedical area, especially for the task of relation extraction. The work done on this subject shows the efficiency of this method despite the challenges facing researchers which vary from the availability of structured medical knowledge resources to the complex nature of medical literature that is entirely different from other domains, besides the importance of high precision and accuracy in this area that requires great efforts to achieve it.

This paper gives an overview of the distant supervision method for RE, which is believed to be of some help to beginner practitioners seeking general knowledge about this subject. It discusses the different approaches used to tackle the biomedical RE in a distant supervised setting where three types of classification used by researchers are distinguished (graph-based, machine learning, and deep learning classifiers). Finally, it sheds light on some limitations of the proposed methods and suggests some solutions to be conducted in the future work.

#### Data Availability

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

#### Conflicts of Interest

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

#### Acknowledgments

This research work was funded by the National Science Foundation of China (Grant no. 61772075), Scientific Research Project of Beijing Educational Committee (Grant no. KM201711232022), and Beijing Institute of Computer Technology and Application (Grant by Extensible Knowledge Graph Construction Technique Project). The authors are thankful to them for their financial support.

#### References

- [1] A. Téllez-Valero, M. Montes-y-Gómez, and L. Villaseñor-Pineda, "A machine learning approach to information





- extraction,” *Lecture Notes in Computer Science*, vol. 3406, pp. 539–547, 2005.
- [2] C. C. Aggarwal and C. X. Zhai, *Mining Text Data*, Springer Science + Business Media, Berlin, Germany, 2012.
  - [3] PubMed, <https://pubmed.ncbi.nlm.nih.gov/>.
  - [4] K. E. Ravikumar, M. Rastegar-Mojarad, and H. Liu, “BEL-Miner: adapting a rule-based relation extraction system to extract biological expression language statements from biomedical literature evidence sentences,” *Database*, vol. 2017, 2017.
  - [5] A. Ben Abacha and P. Zweigenbaum, “Automatic extraction of semantic relations between medical entities: a rule based approach,” *Journal of Biomedical Semantics*, vol. 2, no. 5, pp. 1–11, 2011.
  - [6] O. Frunza and D. Inkpen, *Extraction of Disease-Treatment Semantic Relations from Biomedical Sentences*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010.
  - [7] Y. Guan, J. Yang, X. Lv, and J. Wu, “Clinical relation extraction with Deep learning,” *International Journal of Hybrid Information Technology*, vol. 9, no. 7, pp. 237–248, 2016.
  - [8] C. Quan, M. Wang, and F. Ren, “An unsupervised text mining method for relation extraction from biomedical literature,” *PLoS One*, vol. 9, no. 7, Article ID e102039, 2014.
  - [9] R. Xu and Q. Wang, “A semi-supervised approach to extract pharmacogenomics-specific drug-gene pairs from biomedical literature for personalized medicine,” *Journal of Biomedical Informatics*, vol. 46, no. 4, pp. 585–593, 2013.
  - [10] A. Smirnova and P. Cudré-Mauroux, “Relation extraction using distant supervision,” *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–35, 2019.
  - [11] D. Zhou, D. Zhong, and Y. He, “Biomedical Relation Extraction: From Binary to Complex,” *Computational and Mathematical Methods in Medicine*, vol. 2014, Article ID 298473, 18 pages, 2014.
  - [12] K. E. Ravikumar, H. Liu, J. D. Cohn, M. E. Wall, and K. Verspoor, “Literature mining of protein-residue associations with graph rules learned through distant supervision,” *Journal of Biomedical Semantics*, vol. 3, no. 3, 2012.
  - [13] T. Bobi, R. Klinger, P. Thomas, and M. Hofmann-apitius, “Improving distantly supervised extraction of drug-drug and protein-protein interactions,” in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 35–43, Madison, WI, USA, April 2012.
  - [14] M. Liu, Y. Ling, Y. An, X. Hu, A. Yagoda, and R. Misra, “Relation extraction from biomedical literature with minimal supervision and grouping strategy,” in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 444–449, Belfast, UK, November 2014.
  - [15] W. Zheng and C. Blake, “Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles,” *Journal of Biomedical Informatics*, vol. 57, pp. 134–144, 2015.
  - [16] A. Lamurias, L. A. Clarke, and F. M. Couto, “Extracting microRNA-gene relations from biomedical literature using distant supervision,” *PLoS One*, vol. 12, no. 3, Article ID e0171929, 2017.
  - [17] Y. Liang, C. Xing, and Y. Zhang, “A semi-automated entity-relation extraction mechanism with weakly supervised learning for Chinese medical webpages,” The series Lecture Notes in Computer Science (LNAI) and Lecture Notes in Bioinformatics, vol. 10219, Springer Science + Business Media, Berlin, Germany, pp. 44–56, 2017.
  - [18] C. Quirk and H. Poon, “Distant supervision for relation extraction beyond the sentence boundary,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1171–1182, Valencia, Spain, April 2017.
  - [19] P. Ernst, A. Siu, and G. Weikum, “HighLife: higher-arity fact harvest,” in *WWW ’18: Proceedings of The Web Conference*, pp. 1013–1022, Lyon, France, April 2018.
  - [20] P. Su, G. Li, C. Wu, and K. Vijay-Shanker, “Using distant supervision to augment manually annotated data for relation extraction,” *PLoS One*, vol. 14, no. 7, pp. 1–17, 2019.
  - [21] J. Gu, F. Sun, L. Qian, and G. Zhou, “Chemical-induced disease relation extraction via attention-based distant supervision,” *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–14, 2019.
  - [22] T. Tran and R. Kavuluru, “Distant supervision for treatment relation extraction by leveraging MeSH subheadings,” *Artificial Intelligence in Medicine*, vol. 98, pp. 18–26, 2019.
  - [23] A. Junge and L. J. Jensen, “CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision,” *Bioinformatics*, vol. 36, no. 1, pp. 264–271, 2020.
  - [24] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, p. 1003, Stroudsburg PA USA, 2009.
  - [25] H. M. Berman, J. Westbrook, Z. Feng et al., “The Protein Data Bank,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2003.
  - [26] S. Kerrien, B. Aranda, L. Breuza et al., “The IntAct molecular interaction database in 2012,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D841–D846, Jan. 2012.
  - [27] X.-W. Chen, J. C. Jeong, P. Dermeyer, and “KUPS,” “KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions,” *Nucleic Acids Research*, vol. 39, pp. D750–D754, 2011.
  - [28] C. Knox et al., “DrugBank 3.0: a comprehensive resource for “Omics” research on drugs,” *Nucleic Acids Research*, vol. 39, pp. D1035–D1041, 2011.
  - [29] S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski, “Comparative analysis of five protein-protein interaction corpora,” *BMC Bioinformatics*, vol. 9, no. SUPPL. 3, 2008.
  - [30] I. Segura-Bedmar, P. Martínez, and D. Sánchez-Cisneros, “The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts,” in *CEUR Workshop Proceedings*, vol. 761, pp. 1–9, Magdeburg, Germany, September 2011.
  - [31] Unified Medical Language System (UMLS), “National library of medicine,” <https://www.nlm.nih.gov/research/umls/index.html>.
  - [32] The UniProt Consortium, “Reorganizing the protein space at the universal protein resource (UniProt),” *Nucleic Acids Research*, vol. 40, no. D1, pp. D71–D75, 2012.
  - [33] J. Wang, M. Lu, C. Qiu, Q. Cui, and “TransmiR,” “TransmiR: a transcription factor-microRNA regulation database,” *Nucleic Acids Research*, vol. 38, no. suppl\_1, pp. D119–D122, 2010.
  - [34] S. Bagewadi, T. Bobić, M. Hofmann-Apitius, J. Fluck, and R. Klinger, “Detecting miRNA mentions and relations in biomedical literature,” *F1000Research*, vol. 3, p. 205, 2014.
  - [35] G. Li, K. E. Ross, C. N. Arighi, Y. Peng, C. H. Wu, and K. Vijay-Shanker, “miRTex: a text mining system for miRNA-gene relation extraction,” *PLoS Computational Biology*, vol. 11, no. 9, Article ID e1004391, 2015.

- [36] R. Dienstmann, I. S. Jang, B. Bot, S. Friend, and J. Guinney, "Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors," *Cancer Discovery*, vol. 5, no. 2, pp. 118–123, 2015.
- [37] A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wieggers, and C. J. Mattingly, "Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks," *Nucleic Acids Research*, vol. 37, pp. D786–D792, 2009.
- [38] H. Kilicoglu, D. Shin, M. Fiszman, G. Roseblat, and T. C. Rindfleisch, "SemMedDB: a PubMed-scale repository of biomedical semantic predications," *Bioinformatics*, vol. 28, no. 23, pp. 3158–3160, 2012.
- [39] C. Fomous, J. A. Mitchell, and A. McCray, "'Genetics home reference': helping patients understand the role of genetics in health and disease," *Public Health Genomics*, vol. 9, no. 4, pp. 274–278, 2006.
- [40] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research*, vol. 45, no. D1, pp. D353–D361, Jan. 2017.
- [41] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. D1, pp. D808–D815, 2012.
- [42] D. C. Comeau, C.-H. Wei, R. Islamaj Doğan, and Z. Lu, "PMC text mining subset in BioC: about three million full-text articles and growing," *Bioinformatics*, vol. 35, no. 18, pp. 3533–3535, 2019.
- [43] H. Liu, Z.-Z. Hu, J. Zhang, and C. Wu, "BioThesaurus: a web-based thesaurus of protein and gene names," *Bioinformatics*, vol. 22, no. 1, pp. 103–105, 2006.
- [44] G. Li, C. Wu, and K. Vijay-Shanker, "Noise reduction methods for distantly supervised biomedical relation extraction," in *Proceedings of the BioNLP 2017*, pp. 184–193, Vancouver, Canada, August 2017.
- [45] J. Jiang, "Information extraction from text," in *Mining Text Data*, pp. 11–41, Springer, New York, NY, USA, 2013.
- [46] M. H. Saier, V. S. Reddy, D. G. Tamang, and Å Västermark, "The transporter classification database," *Nucleic Acids Research*, vol. 42, no. D1, pp. D251–D258, 2014.

## Research Article

# Towards Energy-Efficient Framework for IoT Big Data Healthcare Solutions

**Chong Feng** <sup>1</sup>, **Muhammad Adnan**,<sup>2</sup> **Arshad Ahmad** <sup>1,3</sup>, **Ayaz Ullah**,<sup>3</sup>  
and **Habib Ullah Khan**<sup>4</sup>

<sup>1</sup>*School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China*

<sup>2</sup>*Department of Computer Science, Abdul Wali Khan University, Mardan, Pakistan*

<sup>3</sup>*Department of Computer Science, University of Swabi, Anbar, Swabi, Pakistan*

<sup>4</sup>*Department of Accounting & Information Systems, Qatar University, Doha, Qatar*

Correspondence should be addressed to Chong Feng; [fengchong@bit.edu.cn](mailto:fengchong@bit.edu.cn)

Received 22 February 2020; Revised 23 May 2020; Accepted 26 May 2020; Published 12 June 2020

Academic Editor: Iván García-Magariño

Copyright © 2020 Chong Feng 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 aim of the Internet of things (IoT) is to bring every object (wearable sensors, healthcare sensors, cameras, home appliances, smart phones, etc.) online. These different objects generate huge data which consequently lead to the need of requirements of efficient storage and processing. Cloud computing is an emerging technology to overcome this problem. However, there are some applications (healthcare) which need to process data in real time to improve its performance and require low latency and delay. Fog computing is one of the promising solutions which facilitate healthcare domain in terms of reducing the delay multihop data communication, distributing resource demands, and promoting service flexibility. In this study, a fog-based IoT healthcare framework is proposed in order to minimize the energy consumption of the fog nodes. Experimental results reveal that the performance of the proposed framework is efficient in terms of network delay and energy usage. Furthermore, the authors discussed and suggested important services of big data infrastructure which need to be present in fog devices for the analytics of healthcare big data.

## 1. Introduction

Nowadays the Internet of things (IoT) paradigm has been utilized in various healthcare domains for analyzing the real-time data and recommendations upon it. IoT is playing an important role in those applications specifically which involve ubiquitous sensors and actuators communicating through wireless sensor network (WSN) towards solutions of many problems. Most of the applications today highly demand for faster processing of generated data [1]. Furthermore, many issues arise such as data volume, velocity, and variation due to the utilization of sensors, mobility, and geographic distribution in addition to the requirements for accuracy, security, quality of service (QoS), and operational costs [2].

In recent years, cloud computing technology has been widely adopted in IoT-enabled healthcare applications to provide scalability, data analysis, and reliability [3]. The

geographic distribution of cloud data centers where processing of data collected from sensors requires transmission through multihop distances affects the delay-sensitive healthcare applications. Moreover, healthcare applications environments are heterogeneous in nature, so managing the cloud of resource allocations for uneven and uncertain data loads coming from healthcare solutions is a very complex task [4]. There are certain issues regarding cloud computing commonly reported in the literature as follows [4]:

- (i) Sending huge amount of data to the virtual computing platform causes significant overhead in terms of time, throughput, energy consumption, and cost
- (ii) The cloud may be physically located as very far away, so it cannot service IoT application with reasonable latency and throughput

- (iii) Data centers may be overloaded to process large amount of big data in real time and may lead to facing challenges, i.e., capacity, security, and analytics
- (iv) Cloud computing is hard to accommodate analytic engines for efficient processing of big data

Fog computing is one of the promising solutions for healthcare applications to explore lightweight and computing resources close to the IoT data source [5]. Fog nodes are equipped with computational infrastructure, services, and management models to execute data processing closer to the edge of network and provide opportunities in the form of reducing latency, reducing the need for multihop data communication, distributing resource demands, and promoting service flexibility. Healthcare data are delay sensitive which need to be process in real time to make timely decisions on critical patient's health. In one way, fog computing helps in reducing delays but in the other way, fog nodes consume power and energy because most of the time, these nodes remain active to process healthcare data. In order to overcome this problem, this study proposed a framework with a clustering method helping to minimize the energy consumption.

Big data is considered as a large amount of structure, unstructured, and semistructured data which are continuously generated and received by the hospitals [6]. According to the current organization understandings, one way to deal with the big data is to apply analytics to their big data and get useful insights out of it [7]. Based on the literature review, we discussed and suggested important services of big data infrastructure which need to be present in fog devices for analysis of healthcare data.

The rest of the paper is organized as follows: the state of the art is discussed in Section 2. An overview of fog computing platform is given in Section 3. A framework for healthcare solution is proposed and a hospital case scenario is explained in Section 4. Experimental setup, parameters, and results are discussed in Section 5. The big data analytics and its infrastructure containing important components are discussed in detail in Section 6. Finally, the paper is concluded in Section 7.

## 2. Related Work

This section describes the state of the art in order to understand the relevant studies carried out in the literature. The authors selected the most relevant articles for the literature review which correspond to the healthcare applications. These are discussed as follows.

Gia et al. [8] proposed a health monitoring system based on fog computing, and this system includes facilities, i.e., data mining, storing, and notification at the edge of architecture. In this study, the authors explored how the ECG extraction is arranged. Template-based technique known as feature extraction is used to analyse the ECG signals. Bandwidth usage and service delivery are found to be efficient in the experimental results.

Doukas and Maglogiannis [9] presented the online data management where processing of IoT-enabled pervasive applications is handled by the cloud. The proposed prototype is able to receive patient data from the IoT devices and finally process them in the cloud. Issues related to security are observed among the entities during communication. Important features of the prototype, i.e., Representational State Transfer (REST) API-based access; scalability, and interapplication interoperability, are considered.

According to the Renta et al. [10], healthcare data received from the IoT distributed devices are focused to be stored in remote cloud. Data management system consists of IoT devices which collect the patient data in real time. This study revealed that data stored in cloud processed quickly and subscribed users can get quick notification during emergency. The alert system is also presented based on the predefined health rules and users' reactions.

Chen et al. [11] considered the security aspects of medical data shared through cloudlet data collected through encryption. A trust model is proposed to identify reliable entities to share the data i.e. hospital, doctors, chambers etc. Moreover, the trust model also used to connect medical professionals and patients. During data sharing, data are segmented into three parts and stored in the cloud. The intrusion detection system (IDS) remains active throughout the whole process to prevent malicious attacks.

Mahmud et al. [12] proposed a framework to perform data analysis and visualization in order to predict health shocks based on predefined data set. This framework depends on cloud platform and inclusion of Amazon web services (AWS), geographical information system (GIS), and fuzzy rule-based summarized techniques. The framework provides the opportunity to classify health shocks with accuracy using a data model. Moreover, it can explain the casual factors of health with the help of linguistic rules.

Zhang et al. [13] introduced Health-CPS, a cyberphysical system, aimed at providing convenient and efficient healthcare service to patients. The system depends on the cloud computing and data analytics to solve various big data-related issues of healthcare applications. The proposed system consists of layers, i.e., data collection layer, data management layer, and data-oriented service layer. The collection of data is performed in the unified standard which supports distributed storage and parallel processing.

Fazio et al. [14] designed an e-health Remote Patient Monitoring (RPM) system using cloud platform called FIWARE. Their main focus is to speed up the development of RPM availing the facilities from FIWARE. Patients are assisted to optimize the responsibilities of medical professionals. The implementation of FIWARE cloud to the RPM enhanced modularity, scalability, and efficiency.

Peddi et al. [15] proposed a e-health calorie system based on the cloud. The system can classify accurately different food objects from the meal and compute the overall calories. This system is able to do computation offloading from mobile e-health applications to the cloud. Cloud platform provides accurate outcome with tolerable latency, after the resources are managed by broker entity in the cloud. The

broker is used to manage dynamic cloud allocation mechanism in real time if there is demand.

Jindal [16] proposed a technique to calculate heart rate with the help of embedded sensors and photoplethysmography signals in smart phone. This proposal consists of three steps of data processing. This technique is required to be associated with cloud to select accurate PPG signals through deep learning mechanisms and classify the signals into estimated heart rates. The TROIKA dataset is used for the evaluation of this technique. They concluded that this technique brought accurate heart rate predictions.

Muhammad et al. [17] explained healthcare solution for voice pathology monitoring users. As a cloud-based platform, this proposed solution used a voice pathology detection system that incorporates local binary pattern on voice signal produced through MelSpectrum technique. Pathology conduction is done by the machine learning classifier. Results showed that with the merging of cloud platform, the accuracy and accessibility of the healthcare solutions improved.

Gupta et al. [18] discussed a cloud-based IoT-enabled predictive physical activity analysis model. Embedded sensors, cloud computing, and XML web services are used in this model for the faster, secure, and efficient data collection, processing, and communication. Different perspectives were taken, i.e., service adaptation, prediction analysis, and efficiency for the evaluation of this model. The proposed model can send an alert to the responsible person to notify the abnormality.

Real-time health issues of aged and disable people were discussed by Hossain and Muhammad [19]. The proposed Health-IoT can monitor track and store healthcare data for treatment. The Health-IoT framework is able to collect ECG data from smart phones and sensors. The data can be transferred to the cloud where doctors can access and assess it. Data analytics are applied on the data to find out any errors in data and to detect abnormality.

According to the work of Gia et al. [20], the remote monitoring of cardiac patients can be made possible at low cost through their fog-based health monitoring system. The system includes energy-efficient IoT sensors and smart gateways. ECG, body temperature, and respiration rate data are collected by the sensors and are sent to the gateways through wireless for autoanalysis and notifications. Furthermore, it can also help in visualizing the outcome in an efficient way.

Fog-based healthcare framework is proposed by Ahmad et al. [21]. The framework is considered as the middle layer between cloud platform and end IoT devices. Cloud access security broker (CASB) is used with the framework to enhance data privacy and security of healthcare data. The framework is able to aggregate data from several sources with decent cryptographic assessment.

Chakraborty et al. [22] proposed a fog-based computation platform where latency-sensitive health data are considered. In their proposed programming model, geographically distributed large-scale healthcare application is handled. Evaluation of this model is done through processing heart rate healthcare data. The fog-based healthcare

solution improved data accuracy, service delivery, and data consistency.

Service-oriented architecture of fog computing is discussed by Dubey and Constant [23] where validation and evaluation of raw health data are sensed through IoT devices. The proposed system with resource-constrained embedded computing instances is able to conduct the data mining and data analysis. Furthermore, these instances are also capable of identifying important patterns from the health data and forwarding them to the cloud for further storage and usage. The main theme of this study is to highlight the big data processing with low-power fog resources.

Negash et al. [5] implemented a smart e-health gateway of fog computing for IoT-enabled health-care services. All the smart gateways that were distributed geographically were used to manage IoT devices connected with the patients. Clusters of gateways were formed to perform data analytics and configurations. The proposed system was responsible for monitoring patient's movement independently.

Rahmani et al. [24] proposed an e-health gateway system where the smart gateways are placed in correct places to offer real-time storage, processing, and analytics. The system overcomes the issues related to the mobility, energy, and reliability. The authors also developed a prototype called UTGATE which is based on the concept of smart e-health gateway. System performance is evaluated through the IoT-based early warning score (EWS).

### 3. Fog Computing Platform

This section aims to give an overview of fog computing and its components. Furthermore, fog computing is shown with the help of diagram to help readers better understand it.

The environment of fog computing consists of fog nodes which perform diverse computational tasks at the edge of the network as illustrated in Figure 1. There can be many fog levels that are arranged in a network to form in hierarchically distributed way. Each fog node is equipped with memory, storage, network bandwidth, and processing cores. In a particular hospital scenario, the sensors deployed in hospital collect the data and forward to the fog nodes for further processing. In fog nodes, the resources such as memory, storage, cores, bandwidth, etc. are virtual and can be shared through MCI (Micro Computing Instances) [25].

As we assume, all the fog nodes in healthcare solution are always active to perform computation of delay-sensitive healthcare data, so an energy-efficient network needs to be designed to minimize the energy consumption of the whole network. Therefore, we aim to propose an energy-efficient fog-based solution for healthcare.

### 4. IoT Healthcare Solution and Proposed Framework

In this section, a fog-based energy-efficient wireless sensor network healthcare solution is proposed. Based on the literature review, a healthcare solution framework is discussed. We aim to focus on energy-efficient wireless sensor network fog-based architecture where the energy usage of sensors is

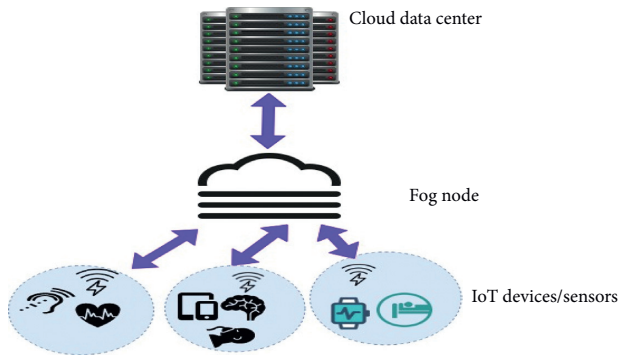


FIGURE 1: General fog computing environment.

to be minimized in order to prolong the network lifetime of sensor nodes at bottom layer. Furthermore, big data collected by these sensors are to be analyzed at the fog nodes. Literature review leads us to provide the answer to the following key research questions:

- (i) RQ1: How to increase the lifetime of the sensor network use in the IoT healthcare applications?
- (ii) RQ2: What are the important services that need to be present in fog nodes to provide analytics?

As Figure 2 depicts, the fog computing environment consists of special networking devices called fog nodes which perform various computational tasks at the edge of network. Every fog node has the capability of providing services, i.e., processing, memory, and storage and network bandwidth. We place these devices in the middle layer of our proposed architecture.

In the bottom layer, wearable sensors are deployed on the body of the patients and used to collect a vast amount of data. These wearable sensors monitor and collect patient's physiological data in the form of ECG, blood oxygen, and other health-related information. Deployed sensors help the patients to reduce their inconvenience of regular visits to the doctor [26].

Wearable sensors use in the bottom layer could have limited power, memory, processing, and communication, so we aim to implement a clustering method which is used to maximize the network lifetime of wireless sensor network [27].

The middle layer is composed of fog nodes which are composed of processing, storing, memory, and network bandwidth capabilities. As the healthcare data collected by the wearable devices at the bottom layer can be increased in size, there is a need to carry out data mining and analytics on such big data. Fog node at the middle layer can process the raw data collected from the bottom layer and carries out analytics [28].

The important healthcare data analyzed in the fog node (middle layer) are processed immediately; otherwise, top layer consisting of cloud computing is responsible for further storage and processing [27].

Regarding the proposed framework of this study, it is necessary to design energy-efficient sensor network in the bottom layer where energy usage of healthcare sensor nodes

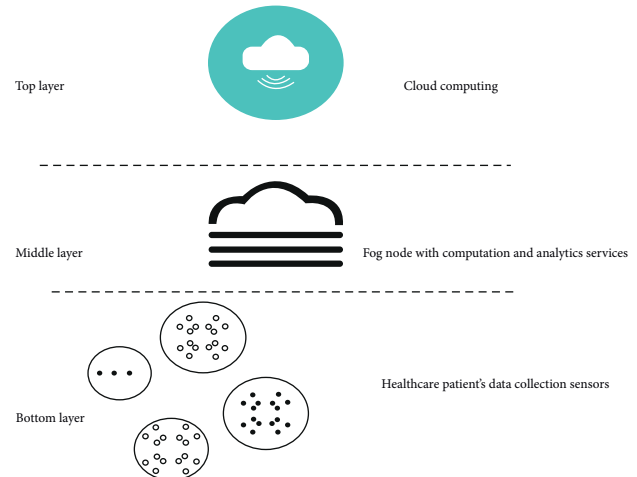


FIGURE 2: The three layers in the proposed framework.

can be minimized before it is transmitted to the middle layer such as fog node for further analytics. Therefore, a hospital scenario has been discussed with the clustering technique to minimize the energy usage of sensor nodes inside the hospital.

**4.1. Case Study: Hospital Scenario.** This section aims to address the RQ1 as aforementioned. The authors discussed a hospital scenario as a case study which help readers to understand the useful implementation of fog nodes inside the hospital and how clustering method could help in energy-efficient network.

Clustering is a technique used to increase the network lifetime and energy efficiency. Sensor nodes can be organized into groups called clusters. Each cluster contains cluster members and cluster head (CH), where cluster members send the data packets to the cluster head, and cluster head aggregates and gathers the data and finally forwards the data to the base station. Comparing both types of sensor nodes in terms of energy usage, cluster heads consume more energy.

In WSN, sensor nodes are small in size, having low power, communication, and computing properties. WSN (Wireless Sensor Network) contains thousands of nodes which can be spatially distributed in various locations to monitor physical, environmental, medical, etc. conditions. Lifetime of sensor nodes depends on the batteries inside these sensors, and it is impractical to change the batteries of every sensor node due to the huge network scale. Therefore, it is important to consider the efficient energy usage of these networks before designing any topology [27].

Regarding the proposed framework, it is necessary to design energy-efficient sensor network in the bottom layer where energy usage of healthcare sensor nodes can be minimized before it is transmitted to the middle layer, i.e., fog node, for analytics.

Figure 3 illustrates various kinds of sensors and IoT devices used inside the hospital in our proposed framework. These sensors collect patient's health data, and the collected

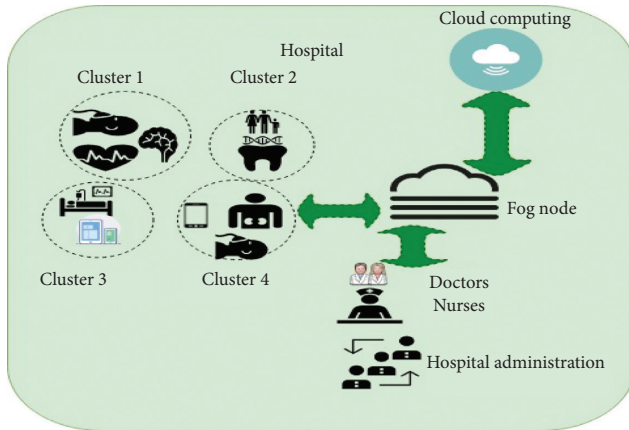


FIGURE 3: Hospital scenario where the clustering method is performed.

data are sent to the fog node for analytics. The data needed on immediate basis are processed by the fog node and can be easily accessed by the doctors, health professionals, and administration staff in the hospital for decisions. The data not needed immediately is sent to the cloud for future use.

It must be noted that all the IoT devices and sensors are battery-powered and should be utilized in an efficient way. Furthermore, some of the IoT devices have built-in batteries so it is not an easy task to recharge or replace the batteries every time, especially in hospital scenario. Also, it is to be noted that IoT devices deployed in hospital send data continuously; in this way, energy utilization increases with the number of transmissions too. To solve this issue, we have proposed a framework based on clustering technique, which will group the number of IoT devices into clusters. The IoT device with high energy and processing capability will be cluster head (CH). Other IoT devices near the CH will become members in respective clusters. When a CH gathers data from other cluster heads, it performs aggregation and forwards to the fog node. The IoT nodes in every floor of the hospital can organize as a cluster.

We know that healthcare data are delay sensitive. These data need to be processed on time to be analyzed. To solve this problem, fog node is placed in the edge of the network to process the data immediately. The healthcare data processed by the fog node can be further accessed by the doctors, healthcare professionals, and administration staff inside the hospital. Furthermore, certain issues arise by using the cloud computing discussed as follows:

- (i) Sending huge amount of data to the virtual computing platform causes significant overhead in terms of time, throughput, energy consumption, and cost
- (ii) The cloud may be physically located as very far away, so it cannot service IoT application with reasonable latency and throughput
- (iii) Data centers may be overloaded to process large amount of big data in real time and may lead to facing challenges, i.e., capacity, security, and analytics

- (iv) Cloud computing is hard to accommodate analytic engines for efficient processing of big data

To overcome these challenges, data analytics is performed at the edge of the network called fog. This fog node can be placed near where the data are generated [29–32].

## 5. Experimental Setup

This section discusses the details about the parameters used for experiments. Furthermore, the authors briefly explain the experimental setup and the results they achieved after the simulations.

A proposed IoT healthcare fog-based solution is compared with cloud-based solution using IFogSim [33]. Although there are various simulators used today for simulation of fog computing, due to the high availability of its source code on GitHub, a lot of target audience, and its easy graphical user interface (GUI), we decided to use IFogSim in our simulations. IFogSim has the ability to integrate various resource management techniques which can be further customized depending on the research area. It is high-performance simulator and its association with CloudSim makes it more useful. CloudSim is very efficient tool in simulation of cloud-based environments [33]. The performance of the proposed solution is compared to the cloud-based solution in terms of network delay and energy usage. Simulation parameters are given in Table 1.

Both healthcare solutions are compared and investigated. After careful investigation, our proposed fog-based big data healthcare solution performs well. In the following, we aim to discuss the measured parameters used in the simulations.

**5.1. Network Delay.** In the experiment, it is observed that average network delay becomes high in cloud healthcare solution because the same communication link shared in cloud by multiple healthcare applications reduces the bandwidth. Furthermore, we observed an increase in the network congestion and a higher round trip time. In contrast, network delay found in fog-based healthcare solution was low as there were multiple communication links present between data source and proximate computing components. Also, the cluster head node is responsible for controlling the data flow to reduce the network delay as mentioned in Figure 4.

**5.2. Energy Usage.** Single virtual machines (VM) in cloud-based healthcare solution execute applications where in fog-based healthcare solution multiple MCIs (micro computing instances) execute an application collectively. MCI used in fog-based healthcare solution consumes fewer amounts of energy and is lightweight as compared to the VM of cloud-based healthcare solution. It is observed from the experiment that the overall energy usage of MCIs was less than the VMs, even in the case of increasing the number of applications load on MCIs as shown in Figure 5.

TABLE 1: Simulation parameters.

Parameter	Value
Simulation duration	400 sec
Cloud data center	
Network latency	10 ms
Energy consumption of VMs	10–15 megaJoules
Average VMs per server	10–15
Fog cluster	
Network latency	10 ms
Energy consumption of MCIs	2-3 megaJoules
Average MCIs per server	3–10
Network size (fog- and cloud-based solution)	25 × 25 m

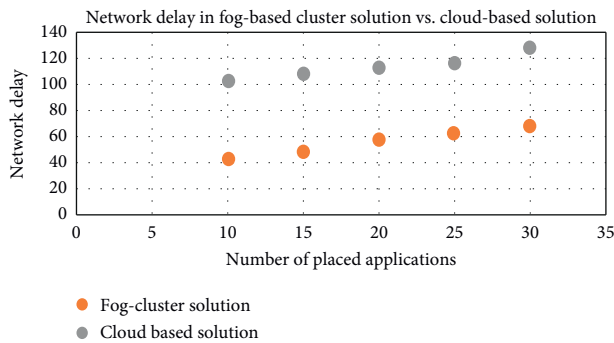


FIGURE 4: Network delay in fog-based and cloud-based solution.

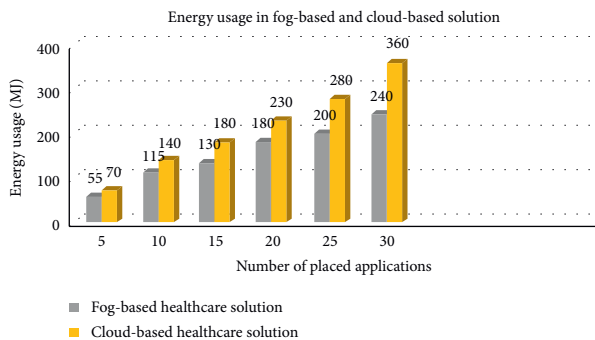


FIGURE 5: Energy usage in fog- and cloud-based solutions.

## 6. Big Data Analytics

The authors aim to give a brief discussion about big data, its infrastructure with important components which must be present to better analyse the big data. Moreover, how to use data analytics in fog devices is discussed. In the end of the section, our hospital scenario is discussed in the context of big data analytics in fog devices.

Large organizations, companies, and research centers receive terabytes of data from various sources, i.e., social media, e-mails from customers, survey responses, phone calls, web server logs, and IoT nodes. Big data is considered as large amount of structure, unstructured and semi-structured data, which are continuously generated and

received by the organizations [34]. According to the current organization understanding, one way to deal with the big data is to apply analytics to their big data and get useful insights out of it. Big data is a kind of advanced analytics where complex applications such as predictive models and statistical algorithms are involved. Big data help to examine a large amount data and extract/uncover hidden patterns from it. Big data can be analyzed in batch mode and streamline mode. It means that for some applications, data are analyzed and results are generated on store-and-process paradigm basis [7].

**6.1. Big Data Analytics Infrastructure.** In this section, we aim to answer the research questions 2 as mentioned above. According to Tang et al. [35], a typical big data infrastructure consists of the following components and layers. The big data platforms have the capabilities of integrating, managing, and applying efficient computational processing to the data. Furthermore, these platforms are used to optimize complex manipulations of large amount of data and considered as big data execution engine. Handling big data with traditional databases is impossible due to its performance/cost towards processing:

- (i) Data management helps any organization to produce data in high-quality format for efficient analysis. Steps included in data management are cleaning, removing anomalies, and transformations to the desired format. Once the steps are performed, the organization must create master data management program for the objective of best analysis.
- (ii) Storage is another important component where a large amount of data is stored because traditional warehouses are not good with storing unstructured and semistructured data.
- (iii) Analytics core functions include data mining that help to examine a large amount of data and discover patterns from it. This information can be further analyzed to help answer complex questions.
- (iv) Presentation is the displaying of information in intuitive and graphical form that help organizations to extract insights for decision making.

**6.2. Data Analytics in the Fog Devices.** In the context of cloud computing, data generated by the IoT devices and sensors are collected and transferred to the cloud for further processing and storage. Although it works well, it poses some challenges; for example, applications require real-time processing, shortening communication time and its cost. Fog computing helps to process the data before they are transferred to the cloud and provides many benefits in the form of shortening communication time and cost and minimizing the need of huge data storage. In short, it is the best solution for all IoT applications [36].

Fog devices are capable of providing low latency and context awareness while cloud providing globalization, so some applications achieve their goals using fog computing



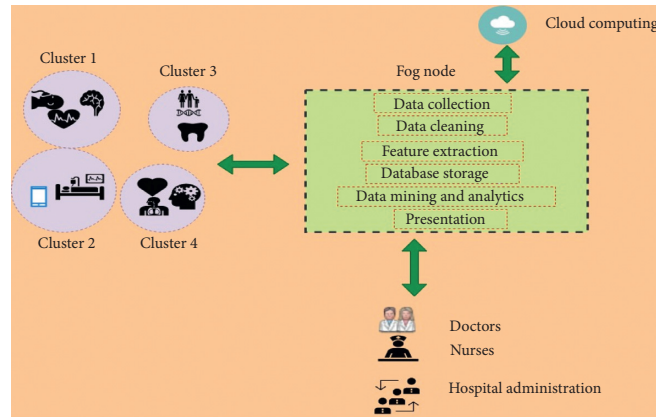


FIGURE 6: Data analytics in fog node.

and cloud computing in the form of localization and globalization [36]. All the fog devices need to have data interfaces, integration with the cloud, handling of incoming continuous data with streaming analytics, and reliable network architecture for moving the real-time processing functions to the edge. Less-time sensitive data can be moved to the cloud for long-term storage and historical analysis. Machine learning and visualization functions need to be applied for improving the performance of IoT applications [6].

**6.3. Big Data Analytics.** Based on the literature review regarding data analytics in fog nodes, we aim to discuss hospital scenario in the context of big data analytics.

As illustrated in Figure 6, according to the literature, the authors suggest six important features that must be present in the fog nodes used for data analytics. Data are collected by the fog nodes coming from various sensors deployed in hospital. Cleaning the data process involves the identification of inaccurate or errors in data and removing them to avoid full storage. Feature extraction is the ability to reduce the raw data into more managed groups for processing. Also, it helps to reduce the number of redundant hospital data for analysis by the doctors, nurses, and administration staff. In the above scenario, data mining is applied with the help of machine learning algorithm to extract and find data from disparate systems. In addition, it supports providing services in healthcare system such as identifying unnecessary utilization of high-cost services, e.g., imaging tests and emergency department, understanding the flow of patients through the hospital, identifying the patients diagnosed for diabetes, etc. [37]. Finally, the data are presented in graph or text format which help the doctors, nurses, or other hospital's management staff in making useful decisions to diagnose the patients [37].

## 7. Conclusions

In this study, the performance of the proposed solution is compared to the cloud-based solution in terms of network delay and energy usage. The parameters used in the experiment were network delay and energy usage. The average

network delay becomes high in cloud-based healthcare solution because the same communication link shared in cloud by multiple healthcare applications reduces the bandwidth. Furthermore, we observed an increase in the network congestion and a higher round trip time. In contrast, network delay found in fog-based healthcare solution was low as there were multiple communication links present between data source and proximate computing components.

Regarding energy usage parameter, we observed that the overall energy usage of MCIs was less than the VMs even in the case of increasing the number of applications load on MCIs. An overview of big data analytics and its infrastructure is discussed. According to the literature review, fog computing supports many applications. Here, we need to describe how our energy-efficient framework can be supported by other domains. The first typical example is smart home where many devices inside the home are connected which require high computing power; if our proposed framework is properly implemented inside the smart home, it would be more helpful in minimizing the energy cost of the network. Another domain to implement the proposed framework in vehicles is called vehicular ad hoc network in which fog nodes are responsible for receiving/sending data from vehicles or other fog nodes, to help in prolonging the lifetime of network through clustering method. In the future, we plan to simulate other parameters in experiments such as instances cost of cloud- and fog-based solutions, to investigate CPU utilization after varying the number of sensors. This study was focused on proposing energy-efficient framework; in the future, we would like to add an important functionality such as secure connection between end devices in order to bring a desired performance of fog computing. We also have a plan to do simulation using another simulator which supports fog computing and compare the results from both simulators. And last but not the least, mathematical modeling of the proposed framework will be introduced.

## 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 regarding the publication of this paper.

## Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2017YFB1002101).

## References

- [1] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, "The internet of things for health care: a comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.
- [2] F. Fernandez and G. C. Pallis, "Opportunities and challenges of the internet of things for health-care: systems engineering perspective," in *Proceedings of the 4th International Conference on Wireless Mobile Communication and Health-Care-Transforming Health-Care through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*, IEEE, Athens, Greece, 2014.
- [3] M. Hassanaliheragh, A. Page, T. Soyata et al., "Health monitoring and management using internet-of-things (IoT) sensing with cloud-based processing: opportunities and challenges," in *Proceedings of the IEEE International Conference on Services Computing*, IEEE, New York, NY, USA, July 2015.
- [4] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, Helsinki, Finland, August 2012.
- [5] B. Negash, T. N. Gia, and A. Anzanpour, "Leveraging fog computing for health-care iot," *Fog Computing in the Internet of Things*, pp. 145–169, Springer, Berlin, Germany, 2018.
- [6] H. Dubey, "Fog data: enhancing telehealth big data through fog computing," in *Proceedings of the ASE Bigdata & Socialinformatics*, pp. 1–6, Kaohsiung Taiwan, October 2015.
- [7] B. Javadi, B. Zhang, and M. Taufer, "Bandwidth modeling in large distributed systems for big data applications," in *Proceedings of the 15th International Conference on Parallel and Distributed Computing, Applications and Technologies*, IEEE, Hong Kong, China, December 2014.
- [8] T. N. Gia, M. Jiang, A. Rahmani, T. Westerlund, P. Liljeberg, and H. Tenhunen, "Fog computing in healthcare internet of things: a case study on ECG feature extraction," in *Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, pp. 356–363, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, Liverpool, UK, 2015.
- [9] C. Doukas and I. Maglogiannis, "Bringing IoT and cloud computing towards pervasive health-care," in *Proceedings of the Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pp. 922–926, IEEE, Palermo, Italy, July 2012.
- [10] T. R. Rentas, S. Sotiriadis, and E. G. M. Petrakis, "Health-care sensor data management on the cloud," in *Proceedings of the 2017 Workshop on Adaptive Resource Management and Scheduling for Cloud Computing (ARMS-CC'17)*, pp. 25–30, ACM, New York, NY, USA, 2017.
- [11] M. Chen, Y. Qian, J. Chen, K. Hwang, S. Mao, and L. Hu, "Privacy protection and intrusion avoidance for cloudlet-based medical data sharing," *IEEE Transactions on Cloud Computing*, 2016.
- [12] S. Mahmud, R. Iqbal, and F. Doctor, "Cloud enabled data analytics and visualization framework for health-shocks prediction," *Future Generation Computer Systems*, vol. 65, pp. 169–181, 2016.
- [13] Y. Zhang, M. Qiu, C. Tsai, M. M. Hassan, and A. Alamri, "Health-CPS: health-care cyber-physical system assisted by cloud and big data," *IEEE Systems Journal*, vol. 11, no. 1, pp. 88–95, 2015.
- [14] M. Fazio, A. Celesti, F. G. Marquez, A. Glikson, and M. Villari, "Exploiting the FIWARE cloud platform to develop a remote patient monitoring system," in *Proceedings of the 2015 IEEE Symposium on Computers and Communication (ISCC)*, IEEE, Larnaca, Cyprus, 2015.
- [15] S. V. B. Peddi, P. Kuhad, A. Yassine, P. Pouladzadeh, S. Shirmohammadi, and A. A. N. Shirehjini, "An intelligent cloud-based data processing broker for mobile e-health multimedia applications," *Future Generation Computer Systems*, vol. 66, pp. 71–86, 2017.
- [16] V. Jindal, "Integrating mobile and cloud for PPG signal selection to monitor heart rate during intensive physical exercise," in *Proceedings of the 2016 IEEE/ACM International Conference on Mobile Software Engineering and Systems (MOBILESoft)*, IEEE, Austin, TX, USA, May 2016.
- [17] G. Muhammad, S. M. M. Rahman, A. Alelaiwi, and A. Alamri, "Smart health solution integrating IoT and cloud: a case study of voice pathology monitoring," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 69–73, 2017.
- [18] P. K. Gupta, B. T. Maharaj, and R. Malekian, "A novel and secure IoT based cloud centric architecture to perform predictive analysis of users activities in sustainable health centres," *Multimedia Tools and Applications*, vol. 76, no. 18, pp. 18489–18512, 2017.
- [19] M. S. Hossain and G. Muhammad, "Cloud-assisted industrial internet of things (IIoT)—enabled framework for health monitoring," *Computer Networks*, vol. 101, pp. 192–202, 2016.
- [20] T. N. Gia, M. Jiang, V. K. Sarker et al., "Low-cost fog-assisted health-care IoT system with energy-efficient sensor nodes," in *Proceedings of the 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, IEEE, Valencia, Spain, June 2017.
- [21] M. Ahmad, M. B. Amin, S. Hussain, B. H. Kang, T. Cheong, and S. Lee, "Health fog: a novel framework for health and wellness applications," *The Journal of Supercomputing*, vol. 72, no. 10, pp. 3677–3695, 2016.
- [22] S. Chakraborty, S. Bhowmick, P. Talaga, and D. P. Agrawal, "Fog networks in health-care application," in *Proceedings of the 2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, IEEE, Brasilia, Brazil, October 2016.
- [23] H. Dubey and N. P. Constant, "Fog data: enhancing telehealth big data through fog computing," *Proceedings of the ASE Bigdata & Socialinformatics*, vol. 14, pp. 1–6, 2015.
- [24] A. M. Rahmani, T. N. Gia, B. Negash et al., "Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach," *Future Generation Computer Systems*, vol. 78, pp. 641–658, 2018.
- [25] R. Mahmud, S. N. Srirama, K. Ramamohanarao, and R. Buyya, "Quality of experience (QoE)-aware placement of applications in fog computing environments," *Journal of Parallel and Distributed Computing*, vol. 132, pp. 190–203, 2019.

- [26] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: a review," *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1321–1330, 2014.
- [27] J. Wang, Y. Gao, W. Liu, A. Sangaiah, and H.-J. Kim, "An improved routing schema with special clustering using PSO algorithm for heterogeneous wireless sensor network," *Sensors*, vol. 19, no. 3, p. 671, 2019.
- [28] H. Cao and M. Wachowicz, "An edge-fog-cloud architecture of streaming analytics for internet of things applications," *Sensors*, vol. 19, no. 16, p. 3594, 2019.
- [29] M. Satyanarayanan, P. Simoens, Y. Xiao et al., "Edge analytics in the internet of things," *IEEE Pervasive Computing*, vol. 14, no. 2, pp. 24–31, 2015.
- [30] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.
- [31] L. M. Vaquero and L. Rodero-Merino, "Finding your way in the fog," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 27–32, 2014.
- [32] S. Yi, L. Cheng, and Q. Li, "A survey of fog computing: concepts, applications and issues," in *Proceedings of the 2015 Workshop on Mobile Big Data*, Hangzhou, China, June 2015.
- [33] H. Gupta, A. Vahid Dastjerdi, S. K. Ghosh, and R. Buyya, "iFogsim: a toolkit for modeling and simulation of resource management techniques in the internet of things, edge and fog computing environments," *Software: Practice and Experience*, vol. 47, no. 9, pp. 1275–1296, 2017.
- [34] F. Mehdipour, N. Hamid, and B. Javadi, "Energy-efficient big data analytics in datacenters," *Advances in Computers*, vol. 100, pp. 59–101, 2016.
- [35] B. Tang, Z. Chen, T. Wao et al., "A hierarchical distributed fog computing architecture for big data analysis in smart cities," in *Proceedings of the ASE Bigdata & SocialInformatics*, Kaohsiung Taiwan, October 2015.
- [36] F. Bonomi, "Fog computing and its role in the internet of things," in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, Helsinki, Finland, August 2012.
- [37] HealthIT Analytics, "Data mining, big data analytics in health-care: what's the difference?," 2020, <https://healthitanalytics.com/news/data-mining-big-data-analytics-in-health-care-whats-the-difference>.
- [38] Sensors Facilitate Health Monitoring, 2019, <https://www.fierceelectronics.com/components/sensors-facilitate-health-monitoring>.