IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3852~3864
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3852-3864

The role of big data in precision medicine and healthcare monitoring using the MapReduce framework

Meenakshi Sankarasubramanian[1], Meena Chavan[2], Govindan Manoharan Karthik[3], Jhansi Pandiri[4], Arumalla Nagaraju[5], Idimadakala Madhavilatha[6]
[1]Department of Computer Science, Faculty of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur, India
[2]Department of Electronics and Telecommunication, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India
[3]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India
[4]Department of Electronics and Communication Engineering, Aditya University, Surampalem, India
[5]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vijayawada, India
[6]Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Tirupati, India

Article Info

Article history:

Received Jul 3, 2025
Revised Jun 12, 2026
Accepted Jul 9, 2026

Keywords:

Big data
MapReduce framework
Min-max normalization
Precision medicine
Recursive feature elimination
Support vector machine

ABSTRACT

Data analytics has become a cornerstone of precision medicine by enabling doctors and scientists to extract meaningful insights from vast, complex data sets. Most healthcare data are high-dimensional data that not only require longer computational time but also affect the accuracy of analysis. In order to overcome these issues, the map reduce based big data healthcare monitoring framework is proposed. The proposed work comprises preprocessing, map reduce framework, and data classification. The preprocessing can be done using improved min-max normalization, and the big data can be handled using the improved support vector machine (SVM)-recursive feature elimination (RFE) method. Finally, the classification can be done using a deep Q-network (DQN). The performance of the proposed method is analyzed in terms of accuracy, precision, f-measure, and Matthew's correlation coefficient (MCC).

Corresponding Author:

Meenakshi Sankarasubramanian
Department of Computer Science, Faculty of Science and Humanities
RM Institute of Science and Technology
SRM Nagar, Potheri, Kattankulathur–603203, Tamil Nadu, India
Email: meenakshisankar2013@gmail.com

1. INTRODUCTION

The aim of precision medicine is to allow patient-focused choices that fulfil the triple objectives of enhancing patient wellness, enhancing patient satisfaction, and lowering expenses [1]. To reach this objective, precision medicine depends on the examination of big data collections to pinpoint clinically-significant sub-groups and aid in uncovering disease mechanisms. Uncommon diseases present considerable difficulties in identification and management because of their limited frequency and varied manifestations. Despite the incorporation of big data in rare disease genomics, significant challenges remain, such as issues in pinpointing causal variants and applying discoveries in clinical settings. The establishment of large consortia leads to accumulation of vast amounts of data, known as big data, highlighting the need for big data-driven analysis pipeline. Handling large data sets poses challenges, such as storage constraints, computational power needs, and concerns over data security [2].

Individuals with rare diseases frequently encounter obstacles like delayed diagnoses and misdiagnoses, with over 90% of rare conditions having no effective treatments [3]. Artificial intelligence (AI) and machine learning (ML) technologies aid rare disease research by facilitating the examination of large

volumes of genomic and clinical data to recognize disease patterns, foresee treatment results, and create tailored therapies, which enhances diagnostic precision and propels drug development. In the current digital landscape, a significant rise in data has generated numerous opportunities and challenges for governments, industries, and organizations. Recent advancements in technology and networking have led to a growing dependence on data, necessitating robust data security and classification systems. Cyber threats such as attackers, malware, espionage, and fraud are increasingly common, necessitating advanced approaches to protect data integrity and verify systems [4].

Precision medicine's use of ML algorithms has the potential to significantly alter medical procedures. Precision medicine benefits greatly from deep learning (DL) algorithms' capacity to manage complex big data, classify, identify patterns, and make predictions. The identification of small molecules in metabolic phenotyping, prognosis, patient classification, and the creation of individualized treatment plans are only a few of the numerous contributions made by these applications. Methodological and practical challenges, such as unrealistic development environments, the requirement for efficient diagnostic models, validation in a variety of contexts, and smooth integration into current healthcare processes, are the cause of the implementation gap [1].

High dimensionality in big data with precision medicine is one of the critical challenges, as datasets frequently consist of noisy, irrelevant, or redundant features that can delay the effectiveness of predictive models [5]. Selecting the most relevant features can address these challenges by improving the classification accuracy as well as reducing computational necessities. Due to their high computational requirements, conventional ML models like decision trees (DT), support vector machines (SVM), and artificial neural networks (ANNs) face scalability problems with big data. Due to their efficiency and comparable performance, algorithms like k-nearest neighbors (KNN) have exposed capacity for certain classification tasks [6]. By automating feature extraction and classification by neural networks (NN), recent DL developments have significantly enhanced abilities in physical layer modulation classification areas [7], [8]. By addressing the challenges faced in handling big data:

- A new classification model is proposed in this work, and its primary contribution is an improved MapReduce framework, which is used to handle big data.
- The preprocessing is done using improved min-max normalization to remove outliers.
- Here, an improved feature fusion method using an SVM-recursive feature elimination (RFE) is employed in the mapper phase. Particularly, features are extracted using statistical methods, and raw data extracts the precision medicine-oriented attributes to enhance the classification accuracy.

The remainder of this paper is structured as follows. A literature survey presented in section 2. A proposed big data classification model presented in section 3. Experimental validation described in section 4. Finally, a conclusion of the research in section 5.

2. LITERATURE SURVEY

In 2022, Jayasri and Aruna [9] developed a model for evaluating the diabetic patient healthcare dataset by the combination of the multiclass outlier categorization, hierarchical decision association rule (AR) as well as a classification-based MapReduce model. In a MapReduce architecture, the AR-apriori algorithm considered medical data when generating rules. This was used to determine the relationship between the diseases and their symptoms. UCI-ML diabetic datasets with 50 attributes were used to conduct this analysis.

In 2021, Mohammed *et al.* [10] developed a novel optimization technique named E-Bat integrated the bat algorithm (BA) as well as exponentially weighted moving average (EWMA). In this paper, a big data classification was performed by employing the MapReduce model, in which the mapper as well as the reducer function depended on the developed model. Choosing the features was the task of the mapper function, which was developed for the categorization in the reducer phase, utilizing NN. Thus, the E-Bat method, which assisted in using MapReduce, was used to classify big data.

In 2024, Huang *et al.* [11] suggested a deep neural network (DNN)-based stock prediction model enhanced for the big data systems used by financial organizations and internet-based financial facilities. Based on big data technologies, this work offered existing tendencies in processing as well as analysis platforms through the world, presenting insights into a modularly calculated big data stage for commercial banks. Also, it proposed a convolutional neural network (CNN) model, focusing on their experimental parameter optimization, principles, and structures. Finally, the proposed stock prediction model demonstrates better performance than conventional methods with respect to accuracy, verifying its efficiency and focusing on its dependable arrangement with experimental results.

In 2024, Wang *et al.* [12] focused on the dynamics among blockchain, network trust, user identity, information security, and big data in the digital economy. A combined model of digital identity attributes is

developed, with three core and nine additional dimensions, and a delegated proxy system enabling cooperation between government and private blockchain alliances. This paper also created a digital identity model with three main dimensions and nine supporting dimensions, which introduced a proxy-based model among governmental and non-governmental blockchain organizations. It redesigns an extensive big data supply chain model and creates a secure blockchain digital identities procedure under the zero-trust model, which offers new security governance solutions for the digital economy, metaverse, and digital government on an international scale.

In 2023, Al Banna *et al.* [13] proposed to examine the mental health consequences of the COVID-19 pandemic. The study was focused on depression, over the X data analysis. This work obtained an impressive high accuracy in identifying depressive tweets by employing an innovative pipeline, which combines CNN and long short-term memory (LSTM).

In 2023, Jaiswal *et al.* [14] proposed big data fusion methods and ensemble learning for breast cancer risk prediction as well as classification. A novel I-extreme gradient boosting (XGBoost) method was proposed in this work, which contains a data preprocessing phase, feature extraction techniques, and an identified target role to improve the accuracy of diagnosing breast cancer. The suggested model showed superior performance in terms of higher accuracy than other classification models like DT, random forest (RF), and SVM.

In 2025, Thatha *et al.* [15] proposed decisive red fox-based feature selection (DRFBPS) which was developed in Python, using data preprocessing and red fox optimization for feature selection. It analyzed risk factors and made predictions, validated through metrics like precision, area under the curve (AUC), F1-score, recall, accuracy, and error rate. The system demonstrated its potential in healthcare analytics, offering accurate, reliable predictions. Overall, DRFBPS proved to be a promising tool for heart disease risk prediction.

In 2025, Palanisamy *et al*. [16] developed hybrid framework combining local heterogeneity-oriented (LHO) for feature extraction and heterogeneous graph embedding-similarity network fusion (HGE-SNF) for robust disease prediction. Developed in MATLAB R2023b, it was tested on datasets for COVID-19. It attained high performance with an AUC up to 0.995 and a sensitivity/specificity of above 98%. A comparative analysis showed its superiority over conventional models.

In 2025, Qi [17] integrated ML methods with cloud-based internet of things (C-IoT) systems for health data security. It used a lightweight encryption block and helped diagnose health issues based on the patient's historical database. It attained approximately 91% accuracy via ANN algorithms. The system provided a secure and efficient way to manage health data. This expansion represented a step forward in the modernization of society 5.0.

In 2025, Manikandan *et al.* [18] developed deep convolutional neural network (DCNN)-LSTM, which automated diverse healthcare functions like infrastructure management, patient monitoring, and disease prediction. Moreover, diverse data types such as alphanumeric, numeric, signals, video, and images are analyzed by this model to accurately find out the abnormal conditions and predict the diseases. For medical personal, the real-time alerts were also generated by this system, which ensures timely intervention. In addition, 96% categorization accuracy was attained by this DCNN-LSTM model.

In 2025, Jameil and Al-Raweshidy [19] developed a real-time healthcare monitoring system, which utilizes the digital twin. This model utilized the internet of things (IoT) for data collection, scalable data processing using cloud computing, and accurate health prediction using advanced algorithms. This seamless integration was realized via both digital and physical architectures, including sensing equipment, multi-objective algorithms, cloud-based storage, real-time data synthesis, and intuitive dashboard analytics. A XGBoost hybrid model was introduced, which was the combination of both merges both XGBoost and MLP models, that addresses the challenges related to historical and real-time data analysis.

3. PROPOSED BIG DATA CLASSIFICATION FRAMEWORK

In this proposed work, a novel big data healthcare model is proposed based on the MapReduce framework model under three phases:

i) Initially, the input healthcare data is preprocessed using improved min-max normalization.
ii) Secondly, the MapReduce framework is employed to handle big data. The mapper phase contains an improved feature fusion technique using improved SVM-RFE and the reducer phase merging the fused features.
iii) Lastly, the fused features are input into the deep Q-network (DQN) classification model to obtain the final classification output. The entire big data classification process is depicted in Figure 1.

3.1. Preprocessing phase

Preprocessing is the primary step for processing numerous data for classification. In this paper, a healthcare dataset referred to as D_{Big}, which is used for big data classification. The data is normalized using

the modified min-max normalization method [20]. This technique adjusts the values of the dataset based on its minimum and maximum values. By standardizing the smallest values to 0 and the largest values to 1, this approach ensures that all other data points are proportionally scaled in the range [0-1]. In (1) is used to represent the formula for min-max normalization.

$$\widehat{N} = \frac{x - x_{min}}{xmin_{max} * MAD} \quad (1)$$

In (1), $\widehat{N}$ denotes the normalized data, x represents the input value, x_{min} signifies the smallest value in the dataset, and x_{max} indicates the largest value in the dataset. The MAD represents the median absolute deviation as (2) of the input data, which is less sensitive to outliers and applies to both normal and non-normal distributed data.

$$MAD = Median(|x_i - \tilde{x}|) \quad (2)$$

Here, $\tilde{x}$ represents the mean of the input data. This normalization process ensures uniformity and standardization of the data, facilitating more effective analysis and modeling. The preprocessed data PD_{big} is partitioned into multiple data, such as $(PD_{big1}, PD_{big2}, \ldots, PD_{bigN})$ during this preprocessing phase. Each data is subjected to the MapReduce framework.

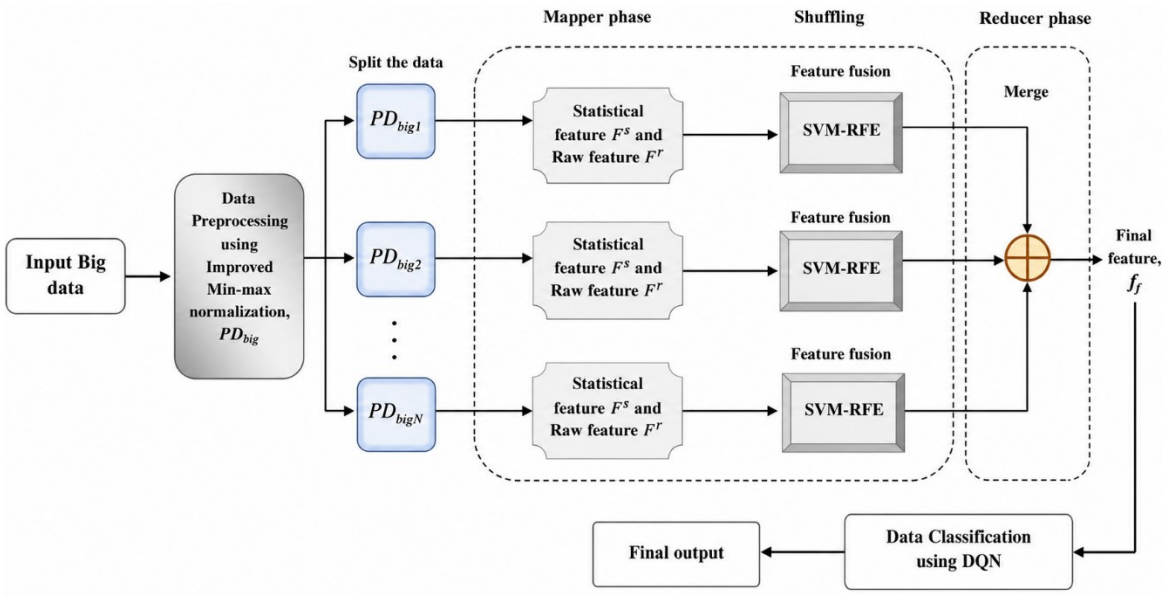

Figure 1. Big data classification process

3.2. MapReduce framework for handling big data

In this second phase, MapReduce framework is applied to efficiently process and handle the big data. The MapReduce framework is split into two main phases such as mapper phase and reducer phase. The mapper phase consists of improved feature fusion using a deep residual network, and the reducer phase focuses on merging the fused features to generate the final feature. Figure 2 depicts the architecture of MapReduce framework. The step-by-step process of the MapReduce framework is structured as follows:

i) Step 1: to splitting the original input data, PD_{big}
ii) Step 2: the divided data is inputted into the mapper phase, here an improved feature fusion technique using an SVM-RFE. This process improves feature extraction for better representation.
iii) Step 3: subsequently, the shuffling process is used to the intermediate output from the mapper phase.
iv) Step 4: finally, the fused features are merged or concatenated in the reducer phase. These features are subjected to the classification phase.

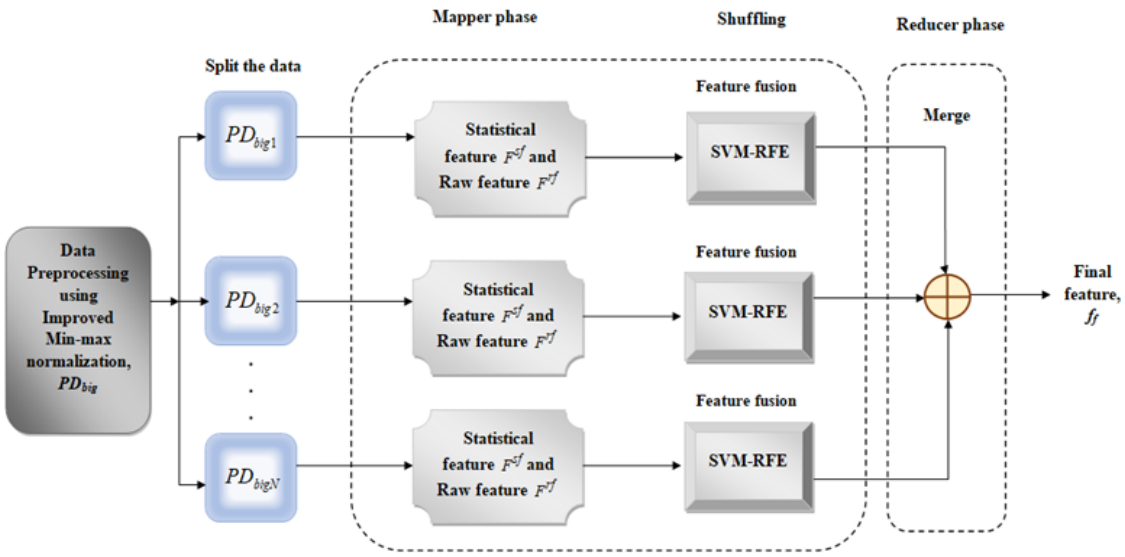

Figure 2. Architecture of MapReduce framework

3.2.1. Mapper phase

In general, the MapReduce algorithm is excellent for mining huge amounts of datasets that are larger than a petabyte and cannot be physically kept in memory. In order to detect the characteristics of massive collections of health information, the MapReduce function is employed in this study. Accordingly, the input data is split and distributed to the mappers, where parallel processing is applied during the feature extraction stages.

The big data is handled using the MapReduce framework. MapReduce [21] is a programming approach used in a distributed system for computing [22] to handle enormous amounts of data. It's employed for managing data that cannot be accommodated in physical memory. These two operations are carried out in two phases, each of which is followed by a data transfer between cluster nodes. Data in the type of "key, value" pairs is processed in parallel throughout these phases. From the dataset, the map function gets the input and produces intermediary output. The map function's output is organized to be used as input for the reducer function. This involves data exchanges among maps and reduces operations. The values are gathered at each node while the reduction function is executed for a certain key. The reduce function also generates the final result in the form "key, value" [21]. MapReduce functions are employed in this study to discover the features of big data [22]. Accordingly, the mapper phase includes the extraction of features like raw features and statistical features.

i) Raw features: these features represent the original input data and are symbolized by the raw data is used to analyze the original parameters of the patient health record and precision medicine, which relies on analysis of large data sets to identify clinically-relevant sub-groups and contribute to discovering disease mechanisms. Advancing precision medicine by enabling measurement of high information content, clinically relevant parameters: heart rate, respiration rate, blood pressure, temperature, myocardial infarction, stroke, and body mass index (BMI).

ii) Statistical features: the derived features contain median, mean, skewness, kurtosis, and standard deviation (SD) features. These features are symbolized by F^{sf}. After deriving the features in the mapper, the reducer phase combines the features from all the mappers and provides the combined feature set F that is defined as $F = [F_{rf}, F_{sf}]$.

iii) Feature selection by using improved SVM-RFE: the feature selection process is a crucial aspect of ML model development, especially in scenarios where the dimensionality of the dataset is high. In the proposed method, the raw feature contains the attributes of the original data which is high-dimensional. Here, an improved SVM-RFE [23] is gaining traction for its ability to identify the most discriminative features while considering various factors.

Step 1: initialization

The feature selection process begins by initializing the original feature set F_s containing features indexed from 1 to E, representing the total number of features in the dataset. Additionally, the output

feature ordered set F_R is initialized to eventually contain the selected features in descending order of importance.

Step 2: looping procedure

- Obtain a training set with the candidate feature set: in each iteration, the current feature set F_s is used to train the SVM classifier, ensuring consideration of different feature subsets.
- Train SVM classifier to obtain weight vector W: following training, a weight vector $W$is obtained from the SVM classifier, representing the importance of each feature for the classification task.
- Calculate ranking criteria score R_k: in the conventional approach, the ranking criteria score is calculated by using (3). Were, k ranges from 1 to the number of features in F_s.

$$R_k = W_k^2 \tag{3}$$

This approach solely relies on the square of each feature's weight, potentially missing other important factors. In (4) defines the proposed approach, here, additional factors, such as the interquartile mean (μ) and SD (σ), are incorporated [24]. This accounts for both linear and non-linear correlations among features, overcoming the limitations of the conventional method.

$$R_k = \left\{W_k^2 + \left[\frac{W_k}{|W_k|+1} * \frac{|\mu+(W_k)-\mu-(W_k)|}{\sigma+(W_k)+\sigma-(W_k)}\right]\right\} \tag{4}$$

- Find the smallest ranking criteria score feature: the feature with the smallest ranking criteria score is identified using (5), indicating the feature contributing the least to the classification task.

$$arg\ \min_k R_k \tag{5}$$

- Update feature set F_R: the selected feature is added to the feature ordered set F_R by taking the union with the set F_P, containing the features selected in this iteration.
- Remove feature from F_s: finally, the selected feature is removed from the current feature set F_s, $F_s = \frac{F_s}{F_P}$ to ensure subsequent iterations consider a reduced set of features. Thus, this method fully exploits the classification performance of each feature subset, ensuring no potentially useful information is overlooked. By considering both linear and non-linear correlations among features, it provides a more comprehensive feature ranking. Through the incorporation of additional factors into the ranking criteria, it effectively addresses the limitations of the conventional approach, resulting in a more robust feature selection process. Finally, the Improved SVM-RFE algorithm offers a refined approach to feature selection, which is represented by F_R, enhancing the effectiveness of the process in healthcare monitoring applications by considering a wider range of factors and criteria. The parameters of proposed SVM-RFE are a step function as 0.10 and n_feature_to_select as 10.

3.2.2. Reducer phase

In this work, the reducer phase is a critical step in the MapReduce framework, which is responsible for combining the extracted features through the mapper phase. The reducer phase's output is a final feature set, f_f which is subjected to the subsequent phase. Thus, the MapReduce framework ensures scalability and effectiveness in handling big data, whereas usage of Improved SVM-RFE enhances feature fusion quality.

3.2.3. Data classification

After performing feature fusion, the merged output f_f is given as an input to the phase of classification. Here, the DQN [25] classifier is used for detection. The main advantage of using the DQN classifier is that it can effectively detect diabetes by increasing the speed of training, thereby achieving powerful results. DQN [25] is a Q-learning mechanism that uses CNN to measure the action value function (Q-function) as well as being the basis for the DQN architecture and training procedures. The advantage of exploiting DQN is that it utilizes the replay hypothesis to mitigate the issues based on data distribution that occur in the network. The architecture includes various layers, such as input layers, convolution layers, dropout layers, pooling layers, flattening layers, dense layers, and output layers. The data augmentation outcome D_o is taken by the input layer and the output is generated in the dimension of [10 × 1]. In addition, two different layers of convolution are exploited, and the result generated by the input layer is presented to the first layer of convolution and thereafter the output is achieved with dimension of [8 × 32]. Also, the second layer receives the output of the prior convolution layer as input and produces output with size

[6 × 64]. The next layer is a dropout layer that decreases the problem of overfitting. Besides, the dropout layer takes the output produced by the second convolution layer as input and the result is generated with size of [6 × 64]. Furthermore, the output of the dropout layer is presented to the maximum pooling layer and the outcome attained is in the dimension of [6 × 64], which is fed to the flattening layer. The layer result generated from flattening layer has the dimension of [1 × 384]. In addition, the Q-learning structure with the loss function is (6).

$$K_o(\upsilon_o) = L\left[\left(\phi + \gamma \underset{\theta'}{max}\widehat{M}(\kappa', \theta'; \upsilon_o^-) - M(\kappa, \theta; \upsilon_o)\right)^2\right] \quad (6)$$

Here, ϕ symbolizes reward, γ indicates discount factor, υ_o specifies the network parameter at iteration o, θ denotes action, and κ implies the network state. The outcome achieved by the flattened layer is given to dense layer. Here, the first and second dense layers produce a result with a size [1 × 128]. Thus, the outcome of the second dense layer is given to third dense layer. This layer measures the output with dimension [1 × 2]. Thus, the output produced by the DQN classifier is represented as O_d.

In this classifier, the convolution layer employs a set of kernels or feature map to extract vital feature from data. Pooling layers reduce the feature's spatial dimension, which ensures computational effectiveness, followed by fully connected layers. Finally, a softmax activation is employed in the output layer and the predicted output is 0 and 1. According to the medical dataset, the value of "0" specifies "normal" and "1" specifies "abnormal". The hyperparameters of the proposed DQN model is given in Table 1. With these hyperparameters, model's generalization and robustness can be increased.

Table 1. Hyperparameters of DQN

Category	Hyperparameter	Value
Q-network parameters	Hidden layer	2
Training parameters	Nodes per layer	50
	Activation function	Softmax
	Optimizer	Adams
	Learning rate	0.001
Other parameters	Batch size	32
	Epochs	50

4. RESULTS AND DISCUSSION

4.1. Simulation procedure

The proposed big data classification was simulated using PYTHON, specifically "Version 3.7". The processor utilized was 11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40GHz 2.42 GHz, and the installed RAM size was 16.0 GB. Further, the big data classification was analyzed using the diabetes health indicators dataset [26].

4.2. Dataset description

This dataset was derived from the "CDC's BRFSS2015" survey. Also, it encompasses 3 files that were described:

i) "Diabetes _ 012 _ health _ indicators _ BRFSS2015.csv:" this dataset consists of 253,680 survey responses, and it utilizes three distinct classes: "0 indicates no diabetes or only during pregnancy, 1 for prediabetes, and 2 for diabetes." It was imbalanced and comprises 21 feature variables.

ii) "diabetes _ binary _ 5050split _ health _ indicators _ BRFSS2015.csv:" this dataset contains 70,692 survey responses. It had an equivalent 50-50 split of respondents with "no diabetes and with either prediabetes or diabetes." It incorporates 2 classes, including "0 for no diabetes, and 1 for prediabetes or diabetes." It was a balanced dataset and consisted of 21 feature variables.

iii) "diabetes _ binary _ health _ indicators _ BRFSS2015.csv:" this dataset comprises 253,680 survey responses. It includes 2 classes like "0 for no diabetes, and 1 for prediabetes or diabetes." This dataset has 21 feature variables and it was a class imbalance dataset.

In this investigation, the total amount of data utilized was 48,000 and it consists of two distinct classes, such as normal (label 0) and abnormal (label 1). The normal class includes 23,900 data and abnormal class contains 24,100 data. Here, the dataset is balanced to avoid overfitting and generalization issues.

4.2.1. Performance analysis

Further, the proposed big data classification model was compared with the conventional techniques, such as KNN [6], optimal recurrent neural network (RNN) [27], link network (LinkNet), bidirectional long

short-term memory (Bi-LSTM), CNN, and LeNet, respectively. Moreover, it was examined with respect to specificity, false negative rate (FNR), accuracy, Matthew's correlation coefficient (MCC), and other performance metrics. Additionally, the ablation and statistical evaluation and computational time were analyzed. Initially, the proposed method is analyzed with various learning rate with 80% training data which is given in Figure 3. By observing the figure, the accuracy, f-measure, precision, and MCC is high for learning rate 0.001 when compared to other learning rates. The proposed DQN is evaluated with various learning rate such as 0.1, 0.01, 0.05, and 0.001. The hyperparameter tuning is done for a learning rate 0.001 which gives better accuracy and precision.

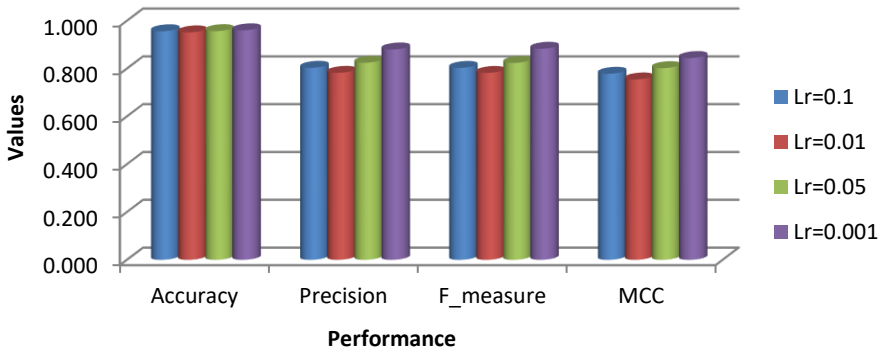

Figure 3. Performance analysis of the proposed method with various learning rates

4.2.2. Comparative analysis on classification performance with various training data

The comparative evaluation of the DQN approach in the realm of big data classification exposes fascinating contradictions when contrasted against existing strategies, KNN [6], optimal RNN [27], LinkNet, Bi-LSTM, CNN, and LeNet. The assessments are conducted by varying the training data to 60%, 70%, 80%, and 90%. Within this comparative study, the efficiency of big data classification depends upon its capability to achieve greater positive and neutral ratings while minimizing negative measure scores. For 60% of the training data, among the models, the Bi-LSTM, CNN, LinkNet, and optimal RNN scored lesser accuracy rate of 0.90, followed by LeNet and KNN, which accomplished the accuracy of 0.91. Nevertheless, the proposed DQN surpasses the traditional methodologies. With an accuracy score of 0.93, the proposed scheme establishes significant enhancements, particularly with larger training data. Regarding 60% training data, the proposed DQN scheme outperforms all other conventional methodologies, exhibiting the greatest f-measure of 0.79. These observations are described in Figure 4.

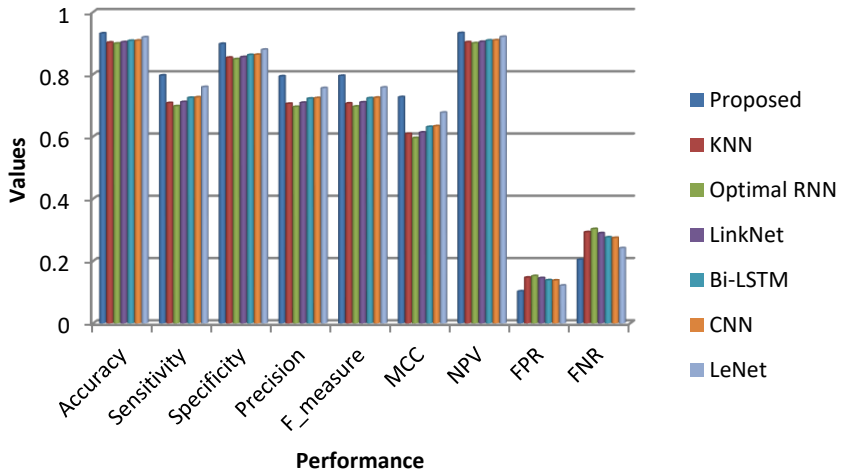

Figure 4. Comparative analysis of proposed method with various existing algorithms with 60% training data

More particularly, the proposed DQN scheme exhibits the least false positive rate (FPR) at 0.081 at 70% training data in Figure 5. In contrast, the existing approaches obtained higher FPR rates with KNN [6] at 0.11, optimal RNN [27] at 0.13, LinkNet at 0.1, Bi-LSTM at 0.11, CNN at 0.11, and LeNet at 0.12, respectively. Additionally, the proposed DQN scheme surpasses the traditional methodologies in terms of neutral metrics across all training data. In particular, the precision achieved using the proposed method is 0.833 in training data 70%, which is extremely higher than KNN [1] (0.762), optimal RNN [2] (0.726), LinkNet (0.78), Bi-LSTM (0.766), CNN (0.77), and LeNet (0.761), respectively. The employment of improved SVM-RFE can accelerate training process and reduce training time as well as computational cost. It can help preserve important features and reduce information loss, leading to better feature representations.

Similarly, the comparative analysis of the proposed method with various existing methods is given in Figures 6 and 7. When the training data increases, the performance also increases. From the figures, it is observed that the accuracy increases from 0.931 to 0.966 when the training data increases for the proposed framework. Likewise, the F-measure and precision are also boosted as 0.793 to 0.900 and 0.795 to 0.906, respectively, which is higher than the existing methodologies such as KNN [6], optimal RNN [27], LinkNet, Bi-LSTM, CNN, and LeNet. The existing methodologies reach a maximum of 0.873 precision and 0.879 F-measure for Bi-LSTM. Because of improved normalization, the proposed DQN framework attains higher performance in all training data percentages. The robust scaling technique accurately identifies the outliers in extreme values also. The DQN classifier has delivered the lower FNR and FPR ranges to verifying that their frameworks are detecting the diseased and non-diseased cases most accurately without any misclassification is occurred in their healthcare monitoring system.

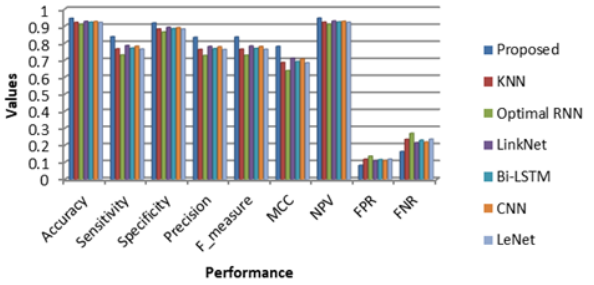

Figure 5. Comparative analysis of proposed method with various existing algorithms with 70% training data

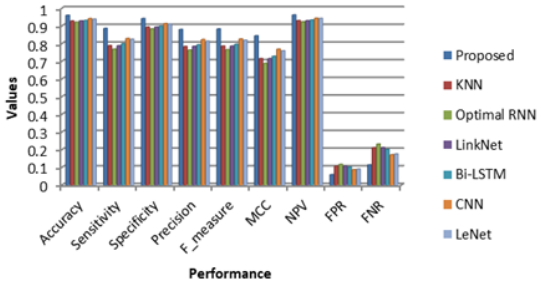

Figure 6. Comparative analysis of the proposed method with various existing algorithms with 80% training data

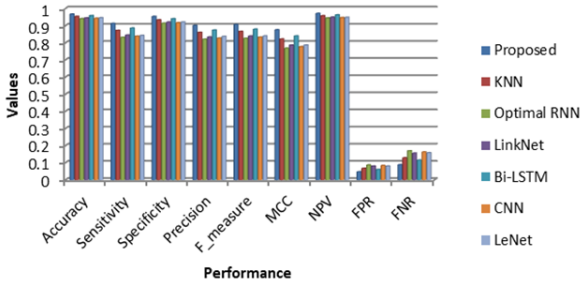

Figure 7. Comparative analysis of the proposed method with various existing algorithms with 90% training data

4.3. Statistical analysis on accuracy

The statistical assessment illustrated in Table 2 offers an exhaustive comparison of the efficacy of the proposed DQN scheme against existing approaches, including KNN [6], optimal RNN [27], LinkNet, Bi-LSTM, CNN, and LeNet for a big data classification framework. For the maximum statistical measure, the proposed scheme exhibited a superior accuracy of 0.967, signifying that it accurately categorizes the big data. Comparatively, KNN [6], optimal RNN [27], LinkNet, Bi-LSTM, CNN, and LeNet obtained the accuracies of 0.954, 0.940, 0.945, 0.958, 0.942, and 0.946, exhibited lesser accuracy values compared to the proposed approach. In addition, the worst performance is accomplished using the optimal RNN [27] of 0.906. The proposed approach can be more robust to shifts in data distribution, enhancing the model's performance. From the statistical analysis, it is observed that the model's accuracy is consistently above 95%, with a peak

at 96.7%. The mean and median values are almost close which indicates a symmetric distribution and the lower SD suggests moderate variability across runs. The performance analysis is executed for 10 runs under 90% training data (different subsets).

Table 2. Statistical evaluation of accuracy

Statistical metrics	Proposed method	KNN	Optimal RNN	LinkNet	Bi-LSTM	CNN	LeNet
Mean	0.949	0.928	0.919	0.924	0.933	0.925	0.932
Median	0.952	0.924	0.915	0.928	0.927	0.934	0.929
Std-Dev	0.016	0.021	0.018	0.017	0.021	0.016	0.013
Min	0.931	0.902	0.898	0.903	0.907	0.908	0.919
Max	0.967	0.954	0.940	0.945	0.958	0.942	0.946

4.4. Ablation study

An ablation study thoroughly eliminates parts of a technique to establish which parts are significant for the system's outcome. Table 3 presents the ablation assessment on proposed DQN with existing feature fusion for big data classification. The proposed model attained the peak negative predictive value (NPV) of 0.9622, while the proposed with existing feature fusion scored the precision of 0.784. Moreover, the least FNR is attained using the proposed DQN model is 0.114, which is exceptionally lesser than that proposed with existing feature fusion. This signifies that the proposed scheme reduced the error values compared to conventional methodologies.

Table 3. Ablation analysis

Metrics	Proposed method without preprocessing	Proposed method with conventional SVM-RFE	Proposed method
Accuracy	0.9012	0.8762	0.9601
Sensitivity	0.7456	0.6945	0.8859
Specificity	0.9131	0.8813	0.9417
Precision	0.7619	0.7849	0.8800
F-measure	0.7456	0.6947	0.8829
MCC	0.7614	0.6655	0.8441
NPV	0.9230	0.8988	0.9622
FPR	0.0769	0.1011	0.0583
FNR	0.2186	0.3055	0.1141

4.5. Computational time analysis

Determining and assessing the amount of processing time required for a method to solve a problem is known as computational time assessment. The proposed method has a computational time of 20.369 s when compared to other methodologies like KNN [6], optimal RNN [27], LinkNet, Bi-LSTM, CNN, and LeNet. The KNN [1] has the worst performance of high computational time of 25.656 s. From Table 4, it is observed that the proposed framework with improved normalization and improved feature fusion obtained better performance in terms of computational time.

Table 4. Computation time analysis (in seconds)

Algorithm	Computational time (in seconds)
Proposed	20.369
KNN	25.656
Optimal RNN	22.762
LinkNet	22.065
Bi-LSTM	24.656
CNN	24.879
LeNet	23.184

5. CONCLUSION

Ultimately, a novel big data classification model was introduced and successfully utilized in this paper. This process started with the suggested MapReduce framework, which includes mappers and reducers for managing large data sets. The input data underwent normalization in the preprocessing stage and was subsequently introduced into the mapper phase, which encompassed both raw features and statistical features for extracting characteristics. The extracted features are subsequently input into the SVM-RFE for feature selection, which interacts with and is combined during the reducer phase. Ultimately, the combined feature

was used as input for the DQN classification model to classify big data. This approach provides an accurate disease classification along with better healthcare monitoring that was proven by the experimental outcomes. However, this work focuses only on health care monitoring and disease classification, so the data security and privacy concerns will be discussed in the future to provide a safe healthcare monitoring to the patient. While handling big data in healthcare, class imbalance is a vital problem which will be handled in future.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Meenakshi Sankarasubramanian	✓	✓	✓			✓			✓		✓		✓	
Meena Chavan		✓		✓		✓				✓				
Govindan Manoharan	✓							✓		✓			✓	
Karthik Jhansi Pandiri						✓				✓				
Arumalla Nagaraju		✓							✓		✓			
Idimadakala Madhavilatha			✓		✓		✓			✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

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

DATA AVAILABILITY

The data used in this study are available from publicly accessible sources and were utilized in compliance with relevant data usage policies. Processed data supporting the findings of this study are available from the corresponding author, [MS], upon reasonable request. Due to privacy and ethical concerns, some raw healthcare datasets cannot be shared publicly.

REFERENCES

[1] 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, 2017, doi: 10.1109/TBME.2016.2573285.

[2] F. Faye *et al.*, "Time to diagnosis and determinants of diagnostic delays of people living with a rare disease: results of a rare barometer retrospective patient survey," *European Journal of Human Genetics*, vol. 32, no. 9, pp. 1116–1126, 2024, doi: 10.1038/s41431-024-01604-z.

[3] M. Wojtara, E. Rana, T. Rahman, P. Khanna, and H. Singh, "Artificial intelligence in rare disease diagnosis and treatment," *Clinical and Translational Science*, vol. 16, no. 11, pp. 2106–2111, 2023, doi: 10.1111/cts.13619.

[4] P. Kaufmann, A. R. Pariser, and C. Austin, "From scientific discovery to treatments for rare diseases- the view from the national center for advancing translational sciences - office of rare diseases research," *Orphanet Journal of Rare Diseases*, vol. 13, no. 1, 2018, doi: 10.1186/s13023-018-0936-x.

[5] K. Haritha, M. V. Judy, K. Papageorgiou, V. C. Georgiannis, and E. Papageorgiou, "Distributed fuzzy cognitive maps for feature selection in big data classification," *Algorithms*, vol. 15, no. 10, 2022, doi: 10.3390/a15100383.

[6] A. S. Tarawneh, E. S. Alamri, N. N. Al-Saedi, M. Alauthman, and A. B. Hassanat, "CTELC: a constant-time ensemble learning classifier based on KNN for big data," *IEEE Access*, vol. 11, pp. 89791–89802, 2023, doi: 10.1109/ACCESS.2023.3307512.

[7] Z. Cai, J. Wang, and M. Ma, "The performance evaluation of big data-driven modulation classification in complex environment," *IEEE Access*, vol. 9, pp. 26313–26322, 2021, doi: 10.1109/ACCESS.2021.3054756.

[8] F. Hang, L. Xie, Z. Zhang, W. Guo, and H. Li, "Research on the application of network security defence in database security services based on deep learning integrated with big data analytics," *International Journal of Intelligent Networks*, vol. 5, pp. 101–109, 2024, doi: 10.1016/j.ijin.2024.02.006.

[9] N. P. Jayasri and R. Aruna, "Big data analytics in health care by data mining and classification techniques," *ICT Express*, vol. 8, no. 2, pp. 250–257, 2022, doi: 10.1016/j.icte.2021.07.001.
[10] M. S. Mohammed, P. S. Rachapudy, and M. Kasa, "Big data classification with optimization-driven MapReduce framework," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, vol. 25, no. 2, pp. 173–183, 2021, doi: 10.3233/KES-210062.
[11] S. Huang, "Big data processing and analysis platform based on deep neural network model," *Systems and Soft Computing*, vol. 6, 2024, doi: 10.1016/j.sasc.2024.200107.
[12] F. Wang, Y. Gai, and H. Zhang, "Blockchain user digital identity big data and information security process protection based on network trust," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, 2024, doi: 10.1016/j.jksuci.2024.102031.
[13] M. H. Al Banna *et al.*, "A hybrid deep learning model to predict the impact of COVID-19 on mental health from social media big data," *IEEE Access*, vol. 11, pp. 77009–77022, 2023, doi: 10.1109/ACCESS.2023.3293857.
[14] V. Jaiswal, P. Saurabh, U. K. Lilhore, M. Pathak, S. Simaiya, and S. Dalal, "A breast cancer risk predication and classification model with ensemble learning and big data fusion," *Decision Analytics Journal*, vol. 8, 2023, doi: 10.1016/j.dajour.2023.100298.
[15] V. N. Thatha *et al.*, "Optimized machine learning mechanism for big data healthcare system to predict disease risk factor," *Scientific Reports*, vol. 15, no. 1, pp. 1–12, 2025, doi: 10.1038/s41598-025-98721-6.
[16] S. Palanisamy, V. Thangaraju, J. Kandasamy, and A. O. Salau, "Towards precision in IoT-based healthcare systems: a hybrid optimized framework for big data classification," *Journal of Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-025-01243-1.
[17] K. Qi, "Advancing hospital healthcare: achieving IoT-based secure health monitoring through multilayer machine learning," *Journal of Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-024-01038-w.
[18] R. Manikandan, S. Arunprakash, R. A. Alsowail, and T. Pandiaraj, "A novel wireless sensor network deployment for monitoring and predicting abnormal actions in medical environment and patient health state," *Alexandria Engineering Journal*, vol. 119, pp. 149–167, 2025, doi: 10.1016/j.aej.2025.01.064.
[19] A. K. Jameil and H. Al-Raweshidy, "A digital twin framework for real-time healthcare monitoring: leveraging AI and secure systems for enhanced patient outcomes," *Discover Internet of Things*, vol. 5, no. 1, 2025, doi: 10.1007/s43926-025-00135-3.
[20] A. J. Mohammed, "Improving classification performance for a novel imbalanced medical dataset using SMOTE method," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3161–3172, 2020, doi: 10.30534/ijatcse/2020/104932020.
[21] S. M. Mahmoud and R. S. Habeeb, "Analysis of large set of images using the MapReduce framework," *International Journal of Modern Education and Computer Science*, vol. 11, no. 12, pp. 47–52, 2019, doi: 10.5815/ijmecs.2019.12.05.
[22] P. P. Anchalia, "Improved MapReduce k-means clustering algorithm with combiner," in *UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014*, 2014, pp. 386–391, doi: 10.1109/UKSim.2014.11.
[23] Y. Zhang, Q. Deng, W. Liang, and X. Zou, "An efficient feature selection strategy based on multiple support vector machine technology with gene expression data," *BioMed Research International*, vol. 2018, no. 1, 2018, doi: 10.1155/2018/7538204.
[24] K. Luo, G. Wang, Q. Li, and J. Tao, "An improved SVM-RFE based on F-statistic and mPDC for gene selection in cancer classification," *IEEE Access*, vol. 7, pp. 147617–147628, 2019, doi: 10.1109/ACCESS.2019.2946653.
[25] H. Sasaki, T. Horiuchi, and S. Kato, "A study on vision-based mobile robot learning by deep Q-network," in *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan, SICE 2017*, 2017, doi: 10.23919/SICE.2017.8105597.
[26] A. Teboul, "Diabetes health indicators dataset," Kaggle.com, 2019. [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
[27] R. Vinoth and J. P. Ananth, "Rider chicken optimization algorithm-based recurrent neural network for big data classification in spark architecture," *Computer Journal*, vol. 65, no. 8, pp. 2183–2196, 2022, doi: 10.1093/comjnl/bxab053.

BIOGRAPHIES OF AUTHORS

Meenakshi Sankarasubramanian is currently working assistant professor, Department of Computer Science, Faculty of Science and Humanities, SRM Institute of Science and Technology, Potheri, Kattankukulathur, Tamil Nadu 603203, India. Her research interests include artificial intelligence, data science, and machine learning. She has 18 years of working experience in higher education institutions and 10 years of research experience. The author is an active member of IEEE and IAENG and life member of ISTE. She can be contacted at email: drsmeenakshicse@gmail.com.

Dr. Meena Chavan currently working as associate professor at Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune. She has 20.6 years of teaching experience and 2.2 years of industry experience. Three patents are granted, and three patents are published on her name. She has done B.E. in Electronics Engineering in 2001 from Shivaji University, Kolhapur, Maharashtra and M.E. in Electronics-VLSI in 2008 from Bharati Vidyapeeth Deemed to be University, Pune, and Maharashtra. She did her Ph.D. (Electronic and Computer Engineering) in 2019 from Pacific Academy of Higher Education and Research University, Udaipur, India. Her research interests are in the field of speech processing, image processing, VLSI design, embedded systems and IoT, and power electronics. She has published more than 50 papers. She has been guiding PG students and Ph.D. scholars. She can be contacted at email: mschavan@bvucoep.edu.in.

Dr. Govindan Manoharan Karthik received the B.E. in Computer Science and Engineering from SACS MAVMM Engineering College, Madurai, M.E. in Computer Science and Engineering from PSNA College of Engineering and Technology, Dindugal, in 2003 and 2005 respectively. He has completed Ph.D. in Information and Communication Engineering from Anna University, Chennai. He is active life member of ISTE and IEEE. He is having 17 years of teaching experience in more than five engineering colleges in India. His primary research interests lie in the areas of data mining, big data analytics, time series data, web mining, modelling, and complexity analysis. Currently, he is working as associate professor in School of Computer Science Engineering, Vellore Institute of Technology, Vellore, India. He can be contacted at email: karthik.gm@vit.ac.in.

Jhansi Pandiri pursuing Ph.D. in Vignan's Foundation for Science, Technology and Research in IoT and sensor networks. Currently, she working as an assistant professor in Aditya University, Surampalem. She completed post-graduation in the stream of Embedded Systems in Aditya College of Engineering and Technology. She is active member of ISTE and IETE. She published book with title "Essential of wireless networks" by REST Publishers. She completed various NPTEL courses. She can be contacted at email: jhansi.pandiri8@gmail.com.

Arumalla Nagaraju pursuing Ph.D. in Computer Science and Engineering From Koneru Lakshmaiah Education Foundation, Andhra Pradesh earned his master's degree in Information Technology from Vignan University, Andhra Pradesh, bachelor degree in Information Technology from Vignan's Engineering College affiliated to Jawaharlal Nehru Technological University, Kakinada. He is currently working as an assistant professor in the Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, Andhra Pradesh 522302, India. He has more than 13 years of teaching experience, he has 12 research publications in journals 4 papers indexed by Scopus. He has guided 15 UG projects and 4 PG projects. He has filed 1 Indian patent been granted and 1 Australian patent. He has organized one national workshop and one national conference. He has received "Best teacher award 2014-15" from Chalapathi Institute of Engineering and Technology. He has life membership in Indian Society for Technical Education and other professional memberships are CSI, IAENG, IACSIT, UACEE, and CSTA. He completed 14 online certification courses from reputed universities and organizations like Coursera, NPTEL, Microsoft, and IBM. He researches interests are artificial intelligence, machine learning, and deep learning. He can be contacted at email: raju051285@gmail.com.

Idimadakala Madhavilatha working as assistant professor in Sri Venkateswara College of Engineering, Karakambadi road Tirupati. She graduated in Narayana Engineering College Nellore and got PG in SVIST Kadapa and pursuing Ph.D. in Computer Science and Engineering at Mohan Babu University. She has more than 6 years of industry experience and more than 8 years of teaching experience. She has published more than 10 papers in international journals. She can be contacted at email: madhavi.mf@gmail.com.