PREDICTION OF CHRONIC AND NON-CHRONIC KIDNEY DISEASE USING MODIFIED DBN WITH MAP AND REDUCE FRAMEWORK

P. Ravikumaran[1], K. Vimala Devi[2] and K. Valarmathi[3]

[1]Dept. of Computer Science and Engineering, Fatima Michael College of Engg & Tech, Madurai- 625020, Tamil Nadu, India
[2]School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Vellore Campus, Vellore- 632014, India
[3]Dept of ECE, P.S.R Engineering College,
Sivakasi- 626140, Tamil Nadu, India

ABSTRACT

Modern medical information comes in the form of an enormous volume of data that is challenging to maintain using conventional methods. The advancement of big data in the medical and basic healthcare societies is facilitated by precision medical data research, which focuses on comprehending early illness, patient healthcare facilities, and providers. It concentrates primarily on anticipating and discovering direct analysis of some of the substantial health effects that have increased in numerous countries. The existing health industry cannot retrieve detailed information from the chronic disease directory. The advancement of CKD (chronic kidney disease) and the methods used to identify the disease is a difficult task that can lower the cost of diagnosis. In this research, a modified MapReduce and pruning layer-based classification model using the deep belief network (DBN) and the dataset used as CKD were acquired from the UCI repository of machine learning. We have utilized the full potentiality of the DBNs by deploying deep learning methodology to establish better classification of the patient's kidney. Finally, data will be trained and classified using the classification layer and the quality will be compared to the existing method.

.

KEYWORDS

Chronic kidney disease, deep belief neural network, MapReduce, Pruning layer.

1. INTRODUCTION

Now a day's knowledge on data is higher and required to process a very large amount of data and time-consuming. Thus, data mining methods can be utilized to classify patients' diseases and one of the major diseases considered is Kidney disorder [8]. CKD is a disorder when kidneys are destroyed and the disposal fluids produced by the physique stay internally that origins health issues. The primary cause is a disease like diabetics or heart problems with heavy blood pressure [6]. Other risky situations triggering CKD involve heart disease, obesity, and family hereditary chronic disease. It may cause very expensive kidney transplantation. So, treatment should be done earlier. On large-scale aggregated datasets, existing simulation approaches aim to attain inaccurate diagnostic identification [9].

David C. Wyld et al. (Eds): AIFZ, SIGPRO, NWCOM, CLSB, MLNLP, DTMN, BDIOT, CSITY - 2022
pp. 185-202, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121615

In recent times, deep learning has received a great deal of prominence from both industry and academia owing to its favourable efficiency in several practical issue [15]. DBN with RBMs are the most essential multi-layer network architectures for deep learning. RBM contains an input and a hidden layer. There have been no neural circuits with same layer and it has a collection of connection weights for input weights as well as hidden neuron. The figure below shows the RBM graphical network, which seems to be a form of energy model.

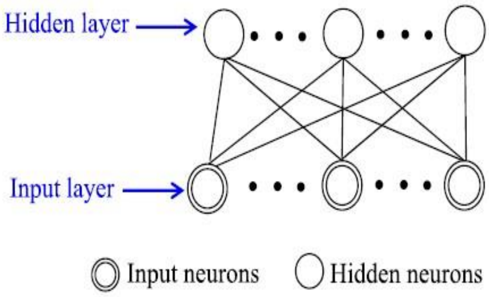


Figure 1. Restricted Boltzmann Machines

DBNs are creative networks trained to retrieve a hierarchical trained feature representation of raw data by optimizing the possibility of the data for training. A single RBM is a key component for deeper architectural design, stacked on top of one another, going to take the previous outcome which is managed as the input following each RBM data type appropriately [11]. Figure 2 shows an explanation of a DBN with stacked RBMs. Label information is used in DBNs to enhance discriminative power for factors learned from the previous RBMs which may not be ideal for variables learned afterwards.

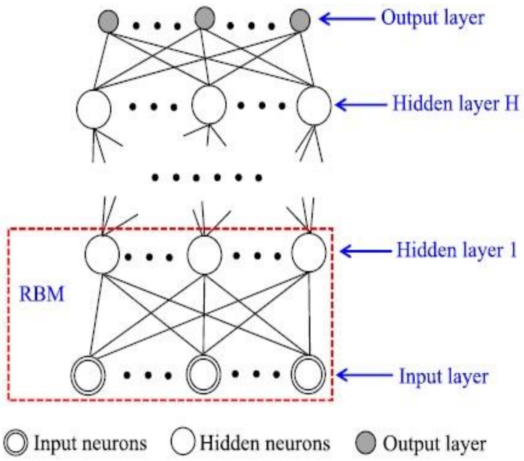


Figure 2. Distributed Belief Network Architecture

The objectives of the current work which are aimed to be achieved are as follows:

- To make effective use of Distributed Belief Network (DBNs) by deploying deep learning.
- To establish better classification of the patient's kidney for protecting them critical conditions.
- To be able to successfully distinguish between non-chronic and chronic kidney diseases with the successful deployment Map Reduce.

To achieve effective results with minimal features along with remarkable connectivity.

2. LITERATURE REVIEW

Jang, Choi et al. [17] developed Deep learning-based radio graphical examination of the hip could predict the osteoporosis. In this development benefit yielded is mere radio graphical examination of hip was sufficient enough to distinguish between the osteoporosis cases and non-osteoporosis without much clinical difficulties

A Temporal modified gradient boost machine was devised by Song, Waitman et al [16]. towards the Longitudinal risk forecasting of the chronic level condition in kidney in patients, who were suffering from diabetes diseases. They utilized data of 14,039 adults, who are suffering type 2 diabetic condition. Higher outcome yielded was ROC of 0.83. However, the pre-defined exclusions in the work could result in wrong false positive values being estimated for diabetic patients with kidney problems, who haven't screened for diabetes readings.

García-Gil et al.,[5] proposed two main pre-processing methods for big data analysis such as homogeneous and heterogeneous ensemble filter to eliminate the noise data. It contains particular prominence in its scalability and performance behavior. This new method is introduced to remove noise data that gives the high quality as well as noise removal clean data also called Smart data. Here, for the experimental evaluation four data sets are used such as HIGGS, Epsilon, SUSY, and ECBDL14. Several stages of class noise have been introduced to examine the cause of deploying such platforms and the development has achieved concerning accuracy using two classifiers like a Knearest neighbor (KNN) and a decision tree (DT) technique. KNN is a noise sensitivity technique if the number of the chosen neighbor is low whereas DT is known as tolerant to noise. Considering the big data issues, the classifier profits from noise treatment even if no additional noise is caused since big data problems involved noise due to incidental homogeneity, accumulation of noisy instances, and spurious correlations. The obtained results showed that noise can be effectively solved in the conceptual methodology. In particular, the homogeneous set is the first effective methodology for attempting to deal with big data noise issues, with low computing moments and encouraging the classifier to achieve greater accuracy.

Abdelaziz et al., [1] Designed an example of cloud-based health services to predict CKD. Cloud computing and also IoT acts an important portion in health care. IoT big data smartphone prediction involves sending large amounts of data to CKD in attempt that would save them in cloud computing. The estimation of harmful diseases like cloud-based CKD –IoT is known to be a big issue facing health care stakeholders. Cloud computing helps patients estimate CKD anyplace at any period in smart cities. The author presented an Intelligent Hybrid Model to Predict CKD-IoT utilizing two intelligent, (linear regression) LR and NN techniques. LR is commonly applied to estimate the major risk factors for CKD. NN can be used to calculate CKD. The results show that when it comes to predicting CKD, the hybrid smart model is 97.8 percent accurate. A hybrid smart model has also been used on Cloud Services as a cloud computing system for predicting CKD in order to assist people in smart cities. The suggested framework is 64% higher than many of the models alluded to in similar works.

Ismail et al., [19] Summarized the implications of big healthcare analytics and relevant applications for both EMR and sensor data. To overcome the issues of irregularity and sparsity in health care data processing algorithms and systems of health care analytics and applications are presented. The proposed outcome depends on an additional layer called middleware which is located between sources of heterogeneous data as well as the Map-reduce Hadoop cluster. The result revealed that the issues that arrived with heterogeneous data have been solved. Furthermore, the availability to incorporate the system with deep learning models that add the capability to identify a particular disease in patients. The proposed framework will be employed in several applications using different optimization methods to minimize the processing time.

Koti & Alamma, [7] The use of health care databases in data analytics, as well as the methods for analysing big data, were discussed. Massive amounts of data are generated in wellbeing organisations with respect to audio, video, text, and (Electronic Health Records) EHR. The big data analytics takes into account the eruption of data to attain significant information that aid throughout the development of effective decisions. After evaluating large data sets of healthcare services in their results, data analysis can improve the operation by using effective methods to achieve the outcomes.

Lokeswari et al., [10] Proposed a new technique designed to improve the above benefits about scalability through maximizing the amount of nodes in the Hadoop cluster and analyzed the efficacy of the classification techniques such as Naïve Bayes, decision tree, and K-nearest neighbor. These parallel methods could be applied in many other biotechnological fields where big dataset forecasting is critical. It offer advantages such as reduction training time, reduced memory requirements, and reduced execution time. There are numerous issues at stake when running parallel algorithms of data mining in cloud systems. It's crucial to split data between processors in a way that minimises computational dependence, communication, proper coordination, overlay interaction, workload averaging among nodes in architectures, and disc IO costs. Running parallel data mining methods upon on Apache Hadoop Map Reduce platform will help with some of these problems. It increases productivity while lowering computational cost, training time, prediction accuracy, and IO availability.

Ahmad et al., [2] A methodology with two major stages, such as categorization modelling and system development, was proposed. Data collection, preparation, grouping, classification, and rule extraction are all part of categorization modelling. Patients with kidney failure have the potential to progress to the chronic stage. A slow decline in kidney function over 3 months characterises CKD, resulting in a severe termination of function (kidney). The goal is to provide such a doctor with a decision-making tool for diagnosing patients with kidney failure. The scheme showed the impact of predicting whether or not due to renal disease have chosen to access the chronic renal progression of the disease. Initially, the processed rules were used to develop the system. This study showed a method that accurately identified a CKD occurrence depends on various factors with a 98.34 percentile accuracy rate. This scheme is utilized to help the physician in accurately measuring the chronic illness of kidney diseases in humans.

Sahoo et al., [13] A probabilistic data collection and correlation analysis of the collected data has been designed. Analysis of the medical fields and expectations about future medical problems are still in the insightful phase. The Data Analysis Framework was also cloud-enabled, and it is the optimal way to analyse structured and unstructured information generated by health-care management solutions. Performance analysis of the proposed processes is carried out and uses extensive simulations that provide 98 percent accuracy in the cloud environment. And retains 90 percent of the CPU as well as bandwidth utilization to shorten the length of analysis. Also, the cloud-based MapReduce model is often used as system architectures for our data analytics. Our method can be used for a variety of health or patient remote monitoring applications, including cardiovascular predictive modelling or tumor severity classification, as per the researchers. The new design will be validated in the realworld healthcare sector using real-time analysis systems such SPARK.

Ed-daoudy & Maalmi, [4] Focused on the application of the distributed machine learning algorithm to stream health data activities absorbed via Kafka concepts to stream processing. First of all, utilizing Spark rather than Hadoop Map Reduce which is restricted to real-time computation, we convert the standard decision tree method into such a parallel, distributed, scalable and fast DT. Secondly, to forecast health status, this method is adopted to transmitting information coming across distributed sources of a different illness. The system predicts health status based on a variety of input data, transmits an warning notification to healthcare professionals, and data is stored in a database system for evaluation of health data and reporting streams. DTs tested the output against data machine learning methods such as WEKA. Finally, to demonstrate the efficiency of the proposed architecture, performance assessment variables like throughput and execution time are measured.

After reviewing all the works reviewed above, the scope and Motivation of this work is, Studies on the precision medical data have focused on understanding early onset of illness as well as the patience health care providers and centre, which had led to the progression of big data in the medical and fundamental healthcare societies. As a result, the concentration on expecting and investigating the direct impact of some substantial health effects have raised in several countries. The existing methodologies in the health sector can't obtain useful information from the any kind of chronic disease directory. The advancement of CKD (chronic kidney disease) and the methods used to identify the disease had also become a daunting task in lowering the diagnosis cost. Some of the identified research gaps that were identified are as follows:

- Several Deep Neural network-based behavioral systems are still in developmental phase as the understanding of the unique behavior of patient's chronic conditions need to get improvise a lot.
- One among the necessary functions in medical care is processing of the data, which should be supported by the big data analytics platform. However, many of the traditional algorithms were found ineffective in handling vast data.
- The prediction performance of the model needs to be improved and better concentration on significant feature is required in order to deal with the sensitive data in the medical care.

There were works Comito, Talia et al. [3] ; Comito, Talia et al. [18]; and Comito, Falcone et al. [3] reported in the literature that concerned about the various issues pertaining to the mobile applications like better routing, resource delegation, and management of the energy. However, the current work will only be considering the benchmark dataset, which doesn't require any data gathering. Thus, deployment of mobile devices was not required and usage of it was not done.

3. PROPOSED METHODOLOGY

The overall flow of the methodology has been explained in the figure 3. Initially, the input dataset has been taken. Here, the used dataset is Chronic Kidney Dataset (CKD). We proposed a novel MR (Map Reduce) and pruning layer-based classification model using the deep belief network (DBN). The novelties of the work carried out are discussed below.

The remarkable connectivity features are selected through a pruning deep belief network algorithm. DBNs are then stacked by plenty of the restricted Boltzmann machines (RBMs). The bottom layer is mainly used to retrieve the input data vector and transfer of the input data to the hidden layer is done through RBM, that is, the output of the lower layer RBM goes to the input of the higher layer RBM. This special case based-energy generation model serves as a learning model for randomly distributed data and then the pruned feature will be given as an input for the DBN's RBM. In Back propagation, we will incorporate Map and Reducer to handle the huge volume of data.

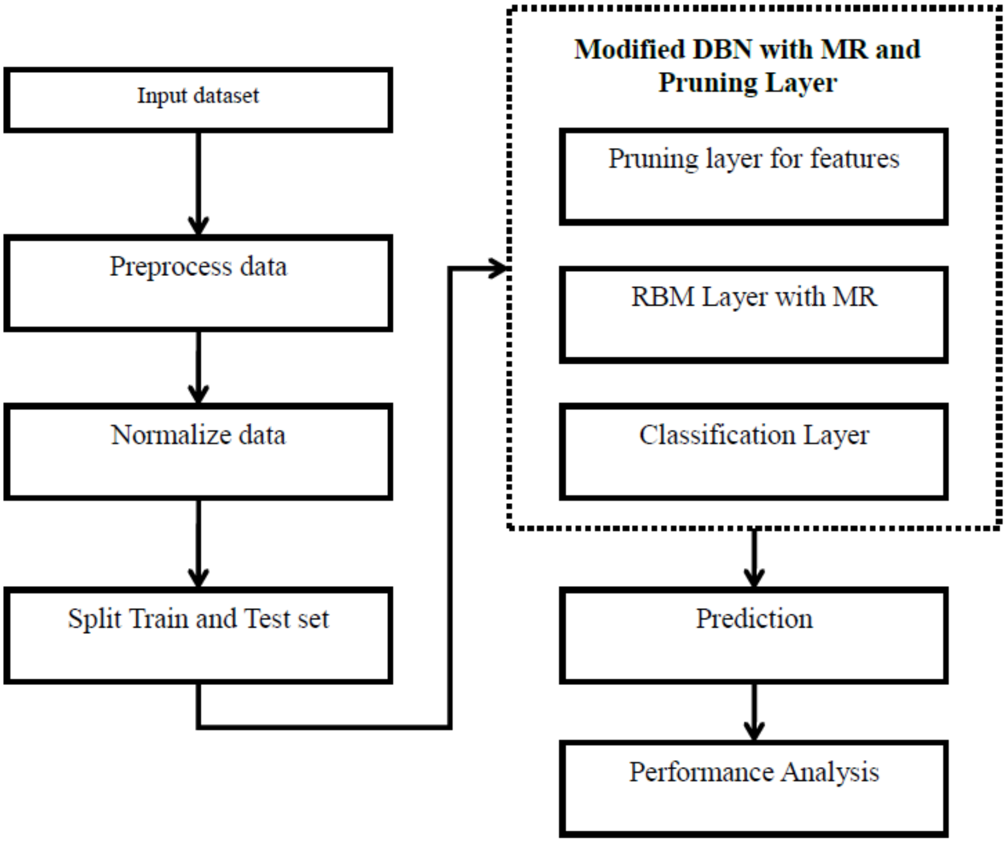


Figure 3. Overall flow of Modified DBN with MR and pruning

The given data are pre-processed and then the data are normalized using standard scalar. It normalizes features of the dataset through scaling to unit variance and this process is very usual in pre-processing step. It also avoids features with large variances from employing a large influence during model training. Then the normalized features are split as training and test set. The next step is feature pruning which took relevant features and sent to the next layer. This creates a DBN layer and that pruned features are sent to the RBM layer with map Reduce. During this process, classification is done in the RBM which predicts the chronic and non-chronic disease, and the performance analysis was done to know the accurate prediction of the given data.

Algorithm 1

Modified Deep belief network (DBN) with MR and pruned layer

Input: Number of epoch e, internal hidden number of layersN_l , hyper-parameters, hidden nodes, v visible node.

Output: The trained DBN model

Step 1: Pruning

```
While $N_l \neq 0$
{
```

```
Initialize the hyper-parameters for the network based on the configuration of the model at the master node

        Divide the training set into subsets
        Send Parameter information to each worker node from the master node
        Allocate jobs to worker node (Subset transfer) $W_n$
        for i = 1 to $W_n$
{
                for j = 1 to e
{

Calculate Gibbs sampling to determine the estimated w, b, and c gradients
Save the average parameter results to the master node.

}
Transfer the trained network to the master node.
}

Create next layer of RBM

Save the average parameters to global parameters that were calculated from each worker node.

}
```

Step 2: θ *updation*

```
Initialize all hyper parameter values for pruned features and create hidden and visible nodes for the training dataset.

for i in h:
{
        Execute activation function using equation (7)
}

for j in v:

{
        Execute activation function using equation (8)
}

Execute equation (11, 12, 13) with activation function to return learned w,b, c($\theta$) value
```

Step 3: Mapping

```
//Mapping function for each key-value pairs
Execute equation 11, 12, 13 to calculate the initial $\theta$ value
```

Step 4: Reducing

```
//Reducing function for each key-value pairs
Execute step 2 to update $\theta$ value.
```

Compute mean squared error (MSE)
Store all MSE and corresponding θ value

Step 5: Prediction
Test data with the trained model
Return predicted data.

3.1. Pruning

Pruning is the first step in emerging distributed architecture. That master node sets up a network specification and parameters with the total number of neurons inside the visible and hidden layers, prejudice of the input and hidden layers, amount of epoch, and the learning rate, etc. The dataset is then divided into several subsets at random and broadcasted to all staff nodes, including a copy of all specified parameters. Based on the training setup, the amount of spits is measured dynamically. Afterward, each task requires the sampling methodology (Gibbs) to evaluate the estimated gradient w, b, and c over his or her portion of the break for each epoch. We regularly measure the gathered parameters for each task and modify the global parameter configuration, which is modified w, b, and c, as well as their cumulative estimated gradient, in the intermediary storage approach. The masters acquire a copy of the training set and initialize this for another layer of RBM because after training is completed. The training of the resting stage of RBM is close to that of the lowermost stage RBM except the input is modified. The modified DBM is developed using the spark algorithm from algorithm 1. The distribution of the learning process is achieved using the data-parallel method, which is focused on the methodology suggested above. We keep a copy of the original model on each workforce machine, and then method various training subsets for each worker. Depending on the synchronous variable averaging method, the effects are averaged and the model parameters are synchronized.

3.2. Map and Reduce Function

MapReduce provides a scheduling algorithm for performing distributed computing on multiple computers.

The system administrator modifies a Master Controller procedure, a sequence of mappers and reducers activities on multiple systems in the scheme. One MapReduce job is computation, which comprises of two stages involving the map and reduces operations. The map principle specifies how the input is segmented into a sequence of subset that are divided into pairs and assigned to mappers. In this regard, each mapper employs the user-specific mapping designed for each input pair and outputs a set of middle pairs that are indicated to the map computer systems' local discs. The source code sends these middle pairs to the master, who is in charge of informing the reducers about these locations.

When the reducers read all of the transitional pairs in digital format, they organise and assemble the keys. To manage all values from each individual key and create a modern significance with seperate key, each reducer uses a user-defined reduce mechanism. All reducers' resulting key-value pairs are acquired as findings and transferred to the output file. Finally, the system of MapReduce completes all of the challenges in comparison. The MapReduce framework can thus be used for highlevel segmentation as well as data processing.

3.3. Probabilistic graph framework:

It is a type of generative stochastic computer program which an understand a probability distribution throughout its inputs. It is made up of both visible and hidden nodes. A RBM is additionally constrained by the elimination of visible-visible and hidden-hidden links. Figure 1 depicts a graphical representation of an RBM. Its probability is described as,

$$p(k,h,\theta) = \frac{e^{-E(k,h,\theta)}}{Q} \quad (1)$$

$$Q = \sum_{\bar{k}} \sum_{\bar{h}} e^{-E(\bar{k},\bar{h},\theta)} \quad (2)$$

where $e^{-E(k,h,\theta)}$ signifies the energy function
θdenotes the parameters
Qindicates the normalizing factor and also known as the partition function
$k, h$is the two vector variables denoting visible and hidden nodes

The energy function is computed as follows

$$E(k,h,\theta) = -b^T k - c^T h - h^T wk \quad (3)$$

Where $b_{j\,and}\ c_j$ are the offsets and w_{ij} denotes the connection weight.
The likelihood $p(k)$denotes the probability allotted to visible vector x, which is added overall feasible configuration of hidden nodes that is

$$p(k) = \sum_{\bar{h}} p(k,\bar{h}) = \sum_{\bar{h}} \frac{e^{-E(k,h)}}{Q} \quad (4)$$

Negative log-likelihood is the optimal solution, which is calculated as,

$$L_l(\theta, Tr_s) = -\sum_{k\epsilon Tr_s} \log P(k,\theta) \quad (5)$$

Where $\theta = \{b,c,w\}$ and Tr_s is the training dataset. According to Bayesian statistics, the issue is to estimate the parameters governing the model through minimizing the negative log-likelihood $\log P(k)$ that is

$$\min_{\theta} L_l(\theta, Tr_s) \quad (6)$$

The comprehensive procedure for solving the dual issue of maximum likelihood that used a stochastic gradient descent framework will be reviewed in the following paragraph.

3.4. RBM Learning Method

Based on the specific structure of the RBM, visible as well as hidden nodes are linearly separable from each other. Where k_j and $h_i \in \{0,1\}$are the probabilistic versions of the regular neuron kernel function, the probabilistic version of the regular neuron activation functions is provided by

$$p(h_i = 1\,|k) = \frac{e^{c_i + w_i k}}{1 + e^{c_i + w_i k}} = L_{sig}(c_i + w_i k) \quad (7)$$

$$p(h_i = 1\,|k) = \frac{e^{c_i + w_i k}}{1 + e^{c_i + w_i k}} = L_{sig}(c_i + w_i k) \tag{7}$$

$$p(k_i = 1\,|h) = \frac{e^{b_i + w_i^T k}}{1 + e^{b_i + w_i^T k}} = L_{sig}(e^{b_i + w_i^T k}) \tag{8}$$

Where w_j is j-th column of w, L_{sig} is the logistics sigmoid function defined by

$$L_{sig}(k) = \frac{e^k}{1+e^k} = \frac{1}{1+e^{-k}} \tag{9}$$

$$-\frac{L_{sig}\ logp(k)}{L_{sig}\ w} = \exp_{ed}[k.h] - \exp_{\ddot{ed}}[k.h],$$
$$-\frac{L_{sig}\ logp(k)}{L_{sig}\ w} = \exp_{ed}[k] - \exp_{\ddot{ed}}[k],$$
$$-\frac{L_{sig}\ logp(k)}{L_{sig}\ w} = \exp_{ed}[h] - \exp_{\ddot{ed}}[h], \tag{10}$$

Where $\exp_{\ddot{ed}}$ is the expectation over k below the experiential distribution $\ddot{ed}$ and is $\exp_{ed}$ the expectation below the distribution of model. It is usually tough to compute this gradient function, though $\exp_{\ddot{ed}}[k.h]$ can be computed easily. The evaluation of $\exp_{ed}[k.h]$ is more difficult, as the actual expectation over all feasible configurations of the input x is costly to evaluate. In most cases, the expectations are calculated using a set amount of unbiased specimens. As shown in the equation, unbiased illustrations of $\exp_{ed}[k.h]$ can be obtained by computing modified Gibbs sampling, where k^0 is a training example from the training dataset, also referred as the order level in Gibbs sampling. The visible and hidden node vectors generated after n-th Gibbs sampling are k^n and h^n, accordingly.

Gibbs sampling is time-consuming. CD (contrastive divergence) learning was adopted as an effective learning method. Two elements of CD learning were used to accelerate the sampling process. After single n-steps of Gibbs sampling, initialise the Markov chain with both the training and obtained sample. The modified conditions for every parameter are given through,

$$w = \epsilon(k^0.h^0 - k^1.h^1) \tag{11}$$
$$\text{b}=\epsilon(k^0.k^1) \tag{12}$$
$$\text{c}=\epsilon(h^0.h^1) \tag{13}$$

where ϵ describes the learning rate. The training issued of pseudo-code for RBM is demonstrated in algorithm 1 after describing the above discussion. Finally, data will be trained and classified using the classification layer and the quality will be compared to the existing method.

4. EXPERIMENTAL RESULTS

4.1. Dataset Description

This study made use of a dataset called CKD, which was uploaded to the UCI machine learning repository in 2015. This dataset contains 25 attributes, 14 of which are nominal and 11 of which are numerical. [8] The characteristics and their description of the features that cope with our research which is outlined. A maximum inst

ance of the dataset will be used for prediction model development of which has been labeled for chronic and non-chronic kidney disease.

4.2. Performance Metrics

The performance metrics are described below for the proposed system to measure the efficiency of the given research [12].

TP (**True positive**): The province where several instances are categorized as precise as true.

FP (**False positive**): The entity in which the amount of scenarios is assembled as precise as they were not correct.

FP (**False Negative**): The province in which the number correctly classified is categorized as false as being probably true.

TN (**True negative**): The situation that the amount of data is classified as false because they were untrue.

Accuracy:

It is a metric of the computation sector of the formulation that represents the systematised error. The difference among the potential outcome and the real value is often caused by low accuracy. This means that the machine evaluates the exceptional input variables several times using the same procedure, and the results are consistent. The amount of real outcomes in the total is known as accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (14)$$

Precision:

It's a metric of algebraic variance that describes random error.

$$Precision = \frac{TP}{TP+FP} \qquad (15)$$

Sensitivity:

It is known as the true optimistic rate, recall, or detection possibility in some fields, is a metric that measures the percentage of true positives that are correctly identified [9]

$$Sensitivity = \frac{TP}{TP+FN} \qquad (16)$$

Specificity:

In some fields, sensitivity is named as the true positive rate, recall, or probability of detection, measures the percentage of actual positive that is expected.

$$Specificity = \frac{TP}{TP+FP} \qquad (17)$$

Recall:

It is the total of all true positives in the positive category.

$$Recall = TP + FN \tag{18}$$

Root Mean Square Error RMSE: It is one of the most widely used types for assessing the performance of predictions. It demonstrates how far assumptions fall from the true value measured using the Euclidean distance. RMSE is also used in supervised learning applications as RMSE utilizes and requires true measurement at each predicted data point. RMSE is computed as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y\ (i) - y\ (i))^2}{N}} \tag{19}$$

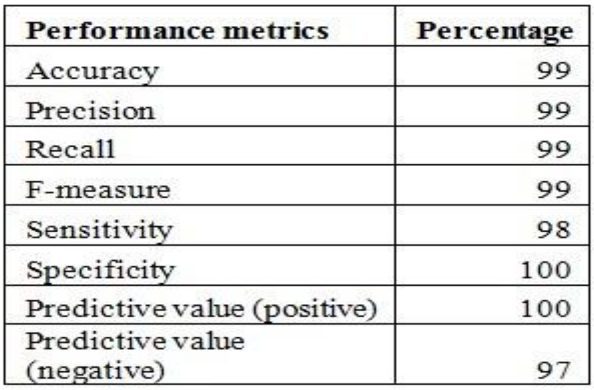
where, $y\ (l)$ indicates the true value and $y\ (i)$ represents the prediction value.

$$\text{Predictive value (positive)} = \frac{TP}{TP+FN} \tag{20}$$

$$\text{Predictive value (negative)} = \frac{TN}{FN+TN} \tag{21}$$

The performance metrics are described with the formula which is used for the simulation performance. The comparative analysis results as explained as below
.

Table 1. Comparison of proposed method in terms of performance metrics



Performance metrics	Percentage
Accuracy	99
Precision	99
Recall	99
F-measure	99
Sensitivity	98
Specificity	100
Predictive value (positive)	100
Predictive value (negative)	97

Table 1 depicts the values of accuracy, precision, recall, f-measure, sensitivity, and specificity of the proposed work which shows the maximum value for the prediction of chronic kidney disease.

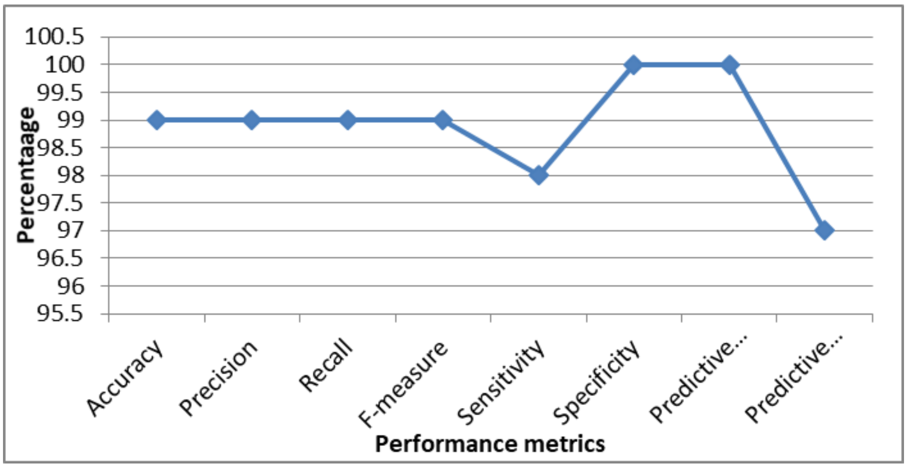
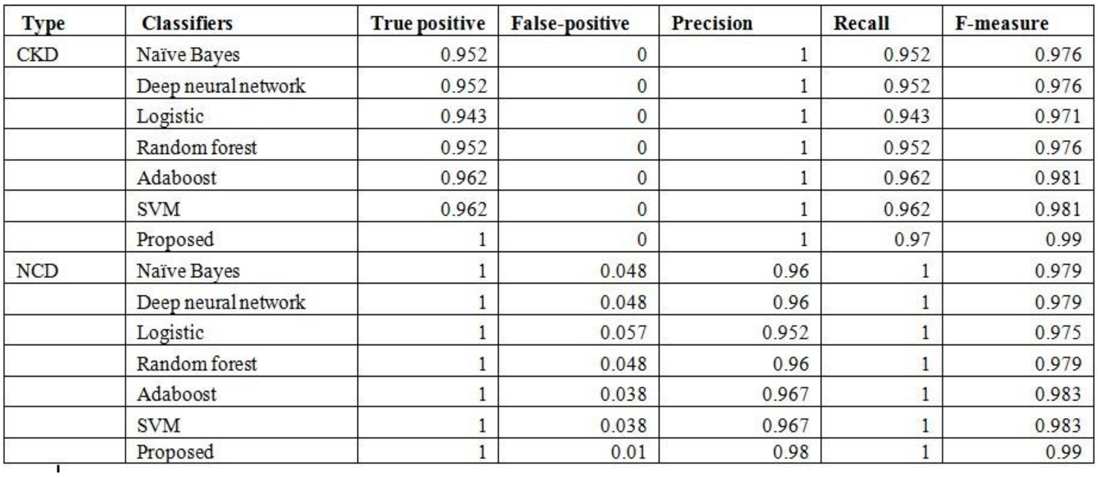


Figure 4. Comparative analysis of the proposed system

Figure 4 depicts the value of all of the proposed performance metrics. The x-axis depicts performance metrics, while the y axis depicts the percentage.

Table 2. Comparison of existing classifiers and proposed system with the type of disease prediction [8]



Type	Classifiers	True positive	False-positive	Precision	Recall	F-measure
CKD	Naïve Bayes	0.952	0	1	0.952	0.976
	Deep neural network	0.952	0	1	0.952	0.976
	Logistic	0.943	0	1	0.943	0.971
	Random forest	0.952	0	1	0.952	0.976
	Adaboost	0.962	0	1	0.962	0.981
	SVM	0.962	0	1	0.962	0.981
	Proposed	1	0	1	0.97	0.99
NCD	Naïve Bayes	1	0.048	0.96	1	0.979
	Deep neural network	1	0.048	0.96	1	0.979
	Logistic	1	0.057	0.952	1	0.975
	Random forest	1	0.048	0.96	1	0.979
	Adaboost	1	0.038	0.967	1	0.983
	SVM	1	0.038	0.967	1	0.983
	Proposed	1	0.01	0.98	1	0.99

Table.2 the quantity of the true positive rate is calculated as the median of all positive values within each epoch. For each technique, the positive result rate is computed in the same way. Precision indicates how many tuples' metrics the classifier classified as chronic kidney disease using structure we developed. For F-measure, the harmonic mean of accuracy and recall is well founded. The designed methodology has the greatest accuracy, recall, and true positive valuation of any systems available in the market.

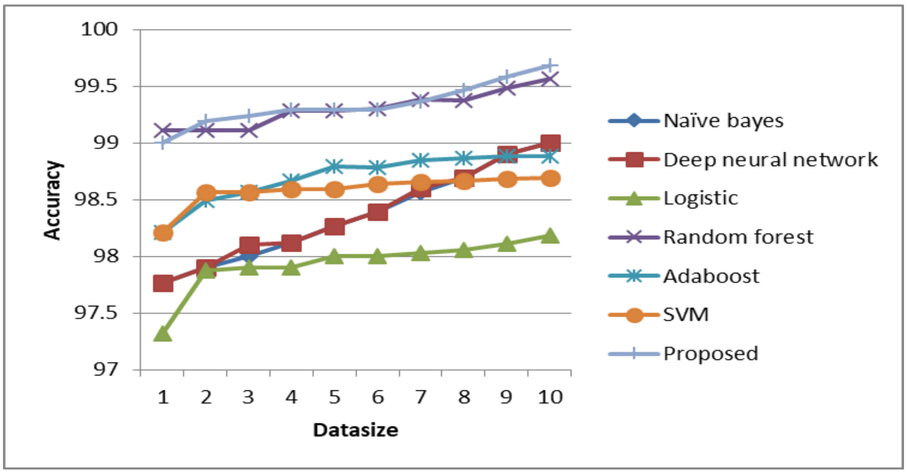


Figure 5. Comparative analysis of accuracy for proposed and existing models

Figure 5 depicts the accuracy measures for the proposed as well as previous methodologies. The x-axis signifies size of data (gigabytes), and the y-axis reflects accuracy (percentage). The results revealed that the proposed method outperforms other methods with the better result of accuracy.

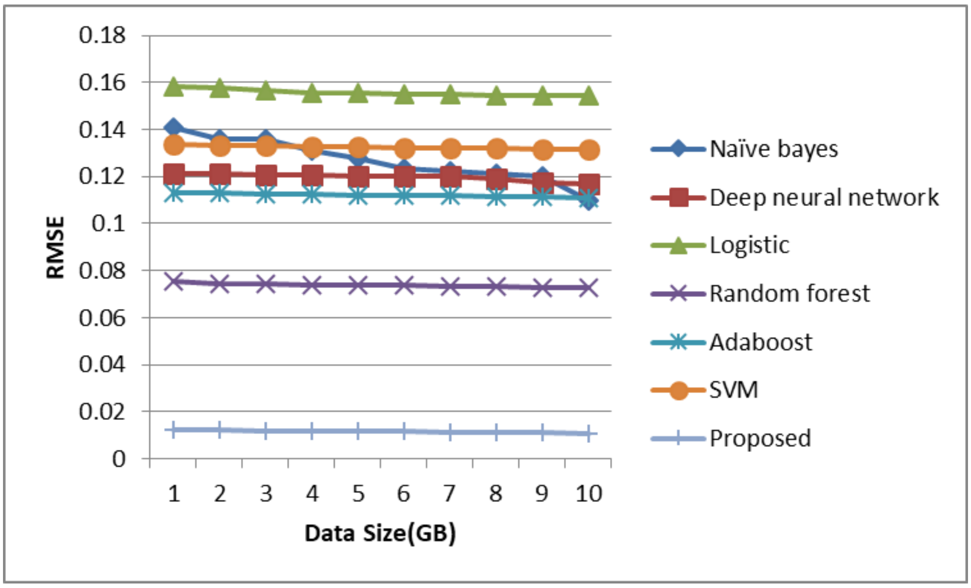


Figure 6. Comparisons of RMSE value

Figure 6 presents a comparison of proposed and current RMSE value techniques. The data size is represented on the x-axis, and the RMSE value is depicted on the y-axis. The results demonstrate that the systems algorithms are superior; the RMSE value is low, indicating better performance.

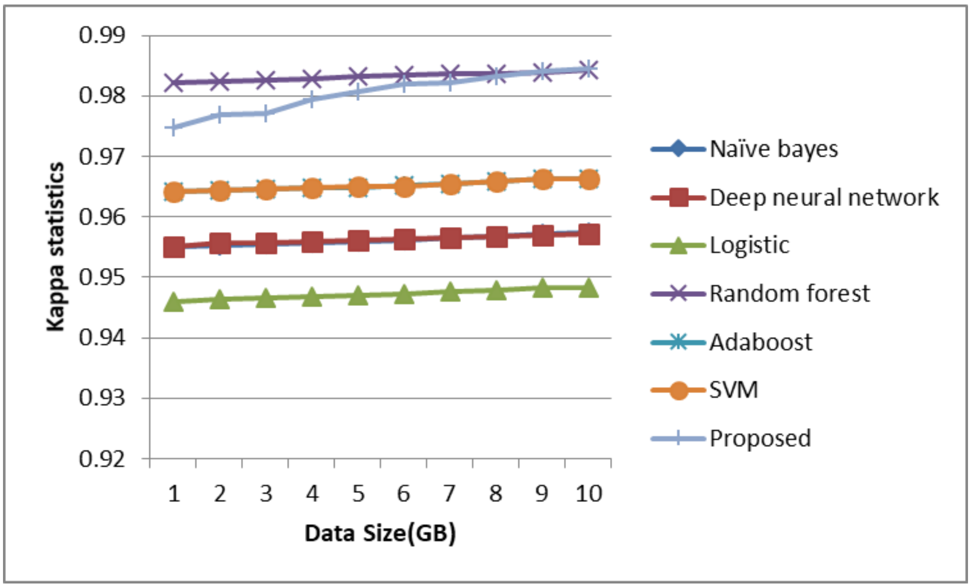
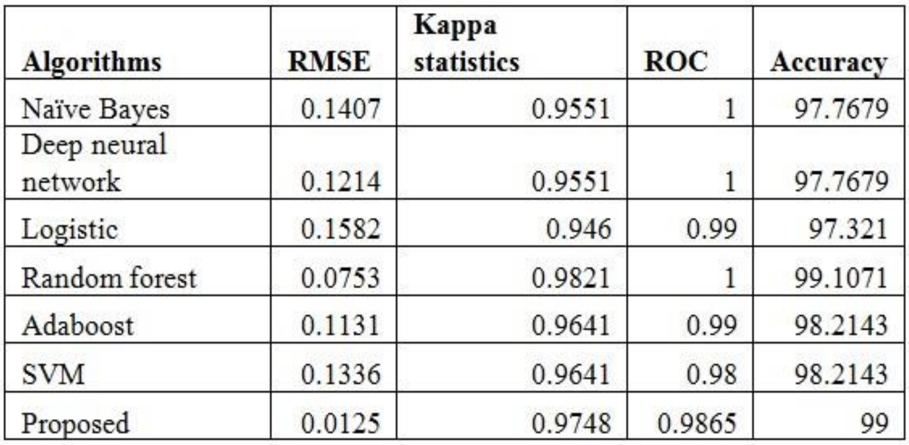


Figure 7. Comparative analysis of Kappa Statistics

Figure 7 represents a kappa statistics-based comparative survey of various methods. The data size is defined by the x-axis, and the value of kappa is indicated by the y-axis. In this case, the proposed kappa value rises as data size increases, whereas the random forest decreases as data size rises.

Table 3. Comparison of proposed and existing methods of RMSE, ROC, and kappa statistics



Algorithms	RMSE	Kappa statistics	ROC	Accuracy
Naïve Bayes	0.1407	0.9551	1	97.7679
Deep neural network	0.1214	0.9551	1	97.7679
Logistic	0.1582	0.946	0.99	97.321
Random forest	0.0753	0.9821	1	99.1071
Adaboost	0.1131	0.9641	0.99	98.2143
SVM	0.1336	0.9641	0.98	98.2143
Proposed	0.0125	0.9748	0.9865	99

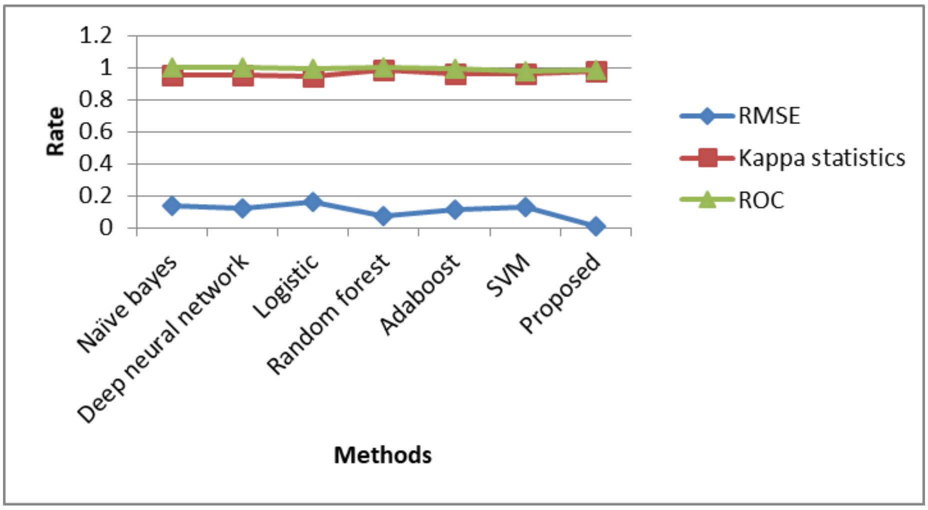


Figure 8. Comparison of RMSE, ROC, and kappa statistics

Table.3 and Figure 8 show that as the amount of epochs increases, method loss decreases. The training set's loss function is set to zero, and the test set's loss function is slightly greater than zero. The RMSE deals with the data's outcome, supplying a severity despite its negative and positive standard errors described in the prediction methods. Kappa statistics show the true percentage of raters who agree on the prediction of CKD with a score of 0.97, indicating nearly perfect agreement. On the ROC curve, the area under plot curve represents false positive versus true positive. We might have had the entire distance taken up by the curve, so the valuation for our proposed system is 98 percent and the precision is 99 percent, indicating a higher level of accuracy.

5. Conclusion

In this paper, a modified DBN with MR and pruning layer has been proposed to manage the massive dataset samples. The methodology utilizes a Map-Reduce algorithm incorporated with a modified DBN to receive the predicted outcomes efficiently with trained persistent DBN with CKD datasets. The standard scalar approach is used for normalizing the data in the preprocessing steps to predict CKD. This proposed methodology predicts chronic and non-chronic kidney disease. Thus, the combination of the modified DBN classification with the MR platform is proficient for giving better outcomes. The efficacy was calculated in terms of accuracy, recall, sensitivity, specificity, ROC, kappa statistics, and also RMSE which are very crucial in the medical field. Our proposed methodology gave rise to the following advantages, which was able to effectively utilize the DBN by achieving efficient classification performance characteristics as compared in the above sections with minimal and remarkable connectivity between the features to distinguish between the non-chronic as well as the chronic kidney diseases in human beings. Our method was also found to hold good in terms of considered big data. The limitations that our work had was that deploying the benchmark dataset instead of data gathering, which could have yielded more practical applicability. Furthermore, we also desire to make use of various mobile devices by also taking into the consideration various challenges pertaining to the mobile data collection operations.

REFERENCES

[1] Abdelaziz, A., Salama, A. S., Riad, A. M., & Mahmoud, A. N. (2019): A Machine Learning Model for Predicting of Chronic Kidney Disease Based Internet of Things and Cloud Computing in Smart Cities. https://doi.org/10.1007/978-3-030-01560-2_5 pp. 93–114

[2] AHMAD, M., TUNDJUNGSARI, V., WIDIANTI, D., AMALIA, P., & RACHMAWATI, U. A. (2022): FUZZY LOGIC-BASED SYSTEMS FOR THE DIAGNOSIS OF CHRONIC KIDNEY DISEASE. DOI: 10.1155/2022/2653665[3]

[3] Comito, C., D. Talia and P. Trunfio (2011). An energy-aware clustering scheme for mobile applications. 2011 IEEE 11th International Conference on Computer and Information Technology, IEEE.

[4] Ed-daoudy, A., & Maalmi, K. (2019): A new Internet of Things architecture for real-time prediction of various diseases using machine learning on big data environment. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0271-7

[5] García-Gil, D., Luengo, J., García, S., & Herrera, F. (2019): Enabling Smart Data: Noise filtering in Big Data classification. Information Sciences, 479, 135–152. https://doi.org/10.1016/j.ins.2018.12.002

[6] Khamparia, A., Saini, G., Pandey, B., Tiwari, S., Gupta, D., & Khanna, A. (2020). KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network. Multimedia Tools and Applications, , 35425–35440. https://doi.org/10.1007/s11042-019-07839-z pp (47–48)

[7] Koti, M. S., & Alamma, B. H. (2019): Predictive analytics techniques using big data for healthcare databases. Smart Innovation, Systems and Technologies, 105, https://doi.org/10.1007/978-98113-1927-3_71, pp 679–686.

[8] Kriplani, H., Patel, B., & Roy, S. (2019): Prediction of chronic kidney diseases using deep artificial neural network technique. In Lecture Notes in Computational Vision and Biomechanics. Springer Netherlands. https://doi.org/10.1007/978-3-030-04061-1_18. pp. 179–187

[9] Larson, E. (1991): Medicare: A Strategy for Quality Assurance. In Journal of Nursing Care Quality (Vol. 5, Issue 4). https://doi.org/10.1097/00001786-199107000-00013

[10] Lokeswari, Y. V., Jacob, S. G., & Ramadoss, R. (2019): Parallel Prediction Algorithms for Heterogeneous Data: A Case Study with Real-Time Big Datasets. Advances in Intelligent Systems and Computing,. https://doi.org/10.1007/978-981-13-1882-5_46. pp 529–538

[11] Merzenich, M. M., Nahum, M., & Van Vleet, T. M. (2013): Neuroplasticity: introduction. Progress in Brain Research, https://doi.org/10.1016/B978-0-444-63327-9.10000-1,pp 14–36

[12] Ramani, R., Vimala Devi, K., & Ruba Soundar, K. (2020): MapReduce-based big data framework using modified artificial neural network classifier for diabetic chronic disease prediction. Soft Computing,. https://doi.org/10.1007/s00500-020-04943-3, pp 16335–16345

[13] Sahoo, P. K., Mohapatra, S. K., & Wu, S. L. (2016): Analyzing Healthcare Big Data with Prediction for Future Health Condition. IEEE Access,. https://doi.org/10.1109/ACCESS.2016.2647619, pp 9786–9799

[14] Wang, Y., Pan, Z., Yuan, X., Yang, C., & Gui, W. (2020): A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network. ISA Transactions, https://doi.org/10.1016/j.isatra.2019.07.001, pp 457–467.

[15] Zhang, K., & Chen, X. W. (2014): Large-scale deep belief nets with mapreduce. IEEE Access, 2,. https://doi.org/10.1109/ACCESS.2014.2319813. pp 395–403

[16] Song, X., L. R. Waitman, S. Alan, D. C. Robbins, Y. Hu and M. J. J. m. i. Liu (2020). "Longitudinal risk prediction of chronic kidney disease in diabetic patients using a temporal-enhanced gradient boosting machine: retrospective cohort study." 8(1): e15510.

[17] Jang, R., J. H. Choi, N. Kim, J. S. Chang, P. W. Yoon and C.-H. J. S. r. Kim (2021). "Prediction of osteoporosis from simple hip radiography using deep learning algorithm." 11(1): 1-9.

[18] Comito, C., D. J. P. Talia and M. Computing (2017). "Energy consumption of data mining algorithms on mobile phones: Evaluation and prediction." 42: 248-264

[19] Ismail, A., Shehab, A., & El-Henawy, I. M. (2019): Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges and Recommendations. https://doi.org/10.1007/978-3-030-01560-2_2, pp. 27–45

AUTHORS

P. Ravikumaran, An active teacher and researcher scholar of Dr. K. Vimala Devi affiliated to Anna University, Chennai, Tamil Nadu. Having 17 years of experience in teaching out of which 7 years in research. Published about 15 research papers in International/National Journals and conferences

Dr. K. Vimala Devi, An active teacher and researcher. Having above 25 years of experience in teaching out of which 15 years in research. Guiding 10 research scholars for Ph. D in the areas of Computer Networks, Network Security, Text mining, Image Processing , Software Testing, Cloud Computing and Big Data. Published about 120 research papers in International/National Journals and conferences, out of which 17 journal papers indexed in ACM portal and 30 indexed in Scopus and 15 IEEE conf. publications. Reviewer in IEEE Communications letter, IJCS and IJNM of John Wiley publications.

Dr. K. Valarmathi, An active teacher and researcher. Having above 25 years of experience in teaching out of which 15 years in research. Guiding 10 research scholars for Ph. D in the areas of Computer Networks, Genetic algorithims, Neural Networks, Fuzzy logic Security, Text mining, Image Processing , Cloud Computing and Big Data. Published about 114 research papers in International/National Journals and conferences.