Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Prediction Of Chronic Kidney Disease Using Data Mining Techniques

Dr. D. Nalini

Assistant Professor, Department of Computer Science, Kurinji College of Arts and Science (Affiliated to Bharathidasan University), Trichy-2

Abstract

Kidney failure disease is being observed as a serious challenge to the medical field with its impact on a massive population of the world. Devoid of symptoms, kidney diseases are often identified too late when dialysis is needed urgently. Advanced data mining technologies can help provide alternatives to handle this situation by discovering hidden patterns and relationships in medical data. The objective of this research work is to predict kidney disease by using multiple machine learning algorithms that are Support Vector Machine (SVM), Multilayer Perceptron (MLP), Decision Tree (C4.5), Bayesian Network (BN) and K-Nearest Neighbour (K-NN). The aim of this work is to compare those algorithms and define the most efficient one(s) on the basis of multiple criteria. The database used is "Chronic Kidney Disease" implemented on the WEKA platform. From the experimental results, it is observed that MLP and C4.5 have the best rates. However, when compared with Receiver Operating Characteristic (ROC) curve, C4.5 appears to be the most efficient.

Keywords: Chronic Kidney Disease, Data Mining, Feature Selection, Support Vector Machine, Multi-Layered Perceptron, Decision Tree (C4.5), Bayesian Network (BN), K-Nearest Neighbour

1. Introduction

Data mining refers to extracting meaning full information from the different huge amount of dataset [1]. It is the process of determining the unseen finding pattern and knowledge from the massive amount of data set. Data mining is significant research doings in the field of medical sciences since there is a requirement of well-organized methodologies for analyzing, predict and detecting diseases [2], [3], [4]. To detect and predict diseases Data mining applications are used for the management of healthcare, health information, patient care system, etc. It also plays a major role in analyzing survivability of a disease [5], [6], [7], [8].

Data mining classification techniques play a vital role in healthcare domain by classifying the patient dataset [9], [10]. Data mining classification technique is used to analyse and predict many diseases. The classification techniques like artificial neural network (ANN), K-nearest neighbor (KNN), naïve Bays, decision tree (J48, C4.5), support vector machine (SVM) etc. Are used by many researchers in the health care area for analysis, detect and predict for a variety of diseases. Feature selection methods, improve the performance accuracy of the

algorithm by reducing the dimensionality of the feature and it can be grouped into a wrapper and filter methods [11]. In developing country, most of the kidney patient received treatment after reached in serious cases. This increases the number of CKD patients [12]. CKD can be reduced even can stop by diagnosis before affected and during affected by doing the test like the blood test, urine test, kidney scan and ask doctor other symptoms of kidney disease.

Most of the works are mainly concentrations of analysis, prediction life danger diseases like chronic kidney disease, cancer disease, diabetic disease, etc. Using feature selection and classification techniques such as KNN, ANN, decision tree (J48/C4.5), SVM, naïve Bays, etc.

2. Related Works

K.R. Lakshmi et al [13] This work is compared to the performance of Bayesian classifier, support vector machine (SVM) classifier and K-nearest neighbour based on accuracy, accuracy and execution time CKD prediction (KNN).

Lambodar Jena et al [14] The main objective of this paper is to use data from this data set to predict the classification of chronic kidney disease in each case accurately.

S.Vijayarani et al [15] The study was to predict kidney disease using vector-based vector machines (SVM) and artificial neural networks (ANN). The aim of this study is to compare the performance of the two algorithms based on accuracy and run-time. From the experimental results it shows that the yield of RNA is better than other algorithms.

Boukenze et al [16] The development of large data sets in the health system is outlined and used in a collection of medical data using three learning algorithms. The goal of this study was to predict kidney disease by using multiple machine learning algorithms, SVM, C4.5 and Bayesian Networks (BNs), and selected efficient.

Yadollahpour, Ali, et al [17] have proposed an adaptive neurofuzzy inference system (ANFIS) for predicting the renal failure timeframe of CKD based on real clinical data, methods used was clinical study records up to 10-year data were collected from newly diagnosed CKD patients.

Sedighi Z. et al. [18] chronic kidney disease is a common disease prevented by early detection and cure. Practical guidance needs classification of kidney disease as a global improvement, data mining, and machine learning techniques supports to discover knowledge in identifying patterns for classification.

3. Proposed Framework for Classification of Chronic Kidney Disease

Figure 1 depicts the proposed framework for the classification of Chronic Kidney Disease using Data Mining techniques like Feature Selection and Classification. The figure 1 depicts the proposed research methodology for the classification of Chronic Kidney Disease. This framework is composed two major stages. 1) Feature Selection: In this stage, the irrelevant, redundant features are removed, 2) Classification of patient into categories Yes and NO category. For the feature selection, Correlation based feature selection with PSO search optimization has utilized. In the classification stage, the evaluation of the feature selection

method has done with the K-Nearest Neighbor (K-NN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT) and Artificial Neural Network (ANN).

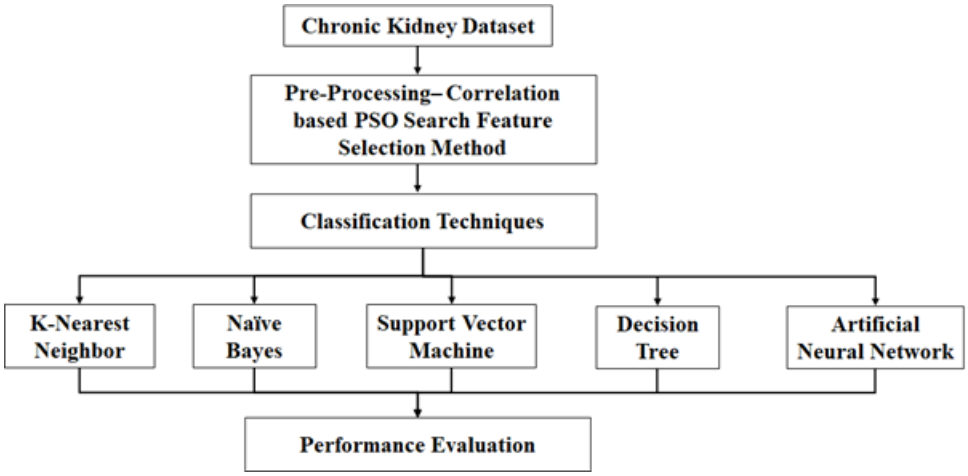

Figure 1: Proposed Framework for the classification of Chronic Kidney Disease using Data Mining techniques

3.1 Pre-Processing Step: Feature Selection Technique

Feature subset selection aims to reduce computing time and improve the results of prediction of machine learning algorithms. This is done by reducing the features/attributes in a dataset that are considered unimportant or unable to contribute positively towards the classification, but it does not create new feature. Fewer features will reduces computing time.

Correlation-based Feature Selection (CFS) is suitable to be applied to multivariate data. CFS works by calculating the interaction between features. CFS evaluates a subset of features taking into account predictive capabilities of each level of redundancy among features and those features. The correlation coefficient is a feature used to calculate the correlation between the subset of features with feature classes and inter correlation among other features.

Particle Swarm Optimization (PSO) is based on the social behavior associated with bird's flocking for optimization problem. A social behavior pattern of organisms that live and interact within large groups is the inspiration for PSO. The PSO is easier to lay into operation than Genetic Algorithm. It is for the motivation that PSO doesn't have mutation or crossover operators and movement of particles is affected by using velocity function. In PSO, every particle alters its own flying memory and its partner's flying inclusion keeping in mind the end goal to flying in the search space with velocity.

The best-fit particle of the entire swarm influences the position of each particle. Each individual particle $j \in [1 \ldots m]$ where $m > 1$, has current position in search space s_j, a current velocity u_j and a personal best position $p_{b,j}$ where j is the smallest value determined by objective function o. By using $p_{b,j}$ the global best position G_b is calculated, which is the buck value obtained by comparing all the $p_{b,j}$

The $p_{b,j}$ is calculated by using the formula

$$p_{b,j}=\begin{cases} p_{b,j} \text{ if } (y_j) > p_{b,j} \\ y_j \text{ if } f(y_j) \leq p_{b,j} \end{cases}$$

The formula used to calculate Global Best Position G_{best} is

$$G_b = \{\min\{p_{b,j}\}, \text{where } j \in [1, \ldots, m] \text{where } m > 1$$

Velocity can be updated by using the formula

$$u_j^{j+1} = wu_j(t) + s_1 i_1[y_j(t) - y_j(t)] + d_2 i_2[g(t) - y_j(t)]$$

where $u_i(t)$ is the velocity and w, s_1 and s_2 are used supplied co-efficient. The i_1 and i_2 are random values $y_j(t)$ is the individual best solution, g(t) is the swarm's global best candidate solution. $wu_j(t)$ is known as inertia component. Inertia component value lies between 0.8 and 1.2. Lower the values of inertia component, it speeds up the convergence of swarm to optima. But higher value encourages the exploration of entire search space. $s_1 i_1[y_j(t) - y_j(t)]$ is known as cognitive component.

The following steps are done in PSO algorithm:

1. Initialize each particle in the population with random positions and velocities.

2. Repeat the following steps until stopping criterion is met.

i. for each particle

{

Calculate the fitness function value;

Compare the fitness value:

If it is superior to the best fitness value pbest, then current value is assigned pbest value;

}

ii. Best fitness value particles among all the particles are selected and assign it as gbest;

iii. for each particle

{

Calculate particle velocity;

Change the position of the particle;

}

3.2 Classification Techniques

Classification is a very important data mining task, and the purpose of classification is to propose a classification function or classification model (called classifier).The classification model can map the data in the database to a specific class. Classification construction methods

include: Decision Tree, Naive Bayes, ANN, K- NN, Support Vector Machine, Rough set, Logistic Regression, Genetic Algorithms (GAs) / Evolutionary Programming (EP).

Naïve Bayes Classification Technique

Naive Bayes is a strategy for assessing probabilities of individual variable qualities, given a class, from preparing information and to then permit the utilization of these probabilities to order new elements, which is a term in Bayesian insights managing a straightforward probabilistic classifier taking into account applying Bayes' hypothesis (from Bayesian measurements) with strong (guileless) autonomy assumptions. In basic terms, a strong Bayes classifier expect that the nearness (or nonappearance) of a specific feature of a class is disconnected to the nearness (or nonattendance) of some other element. The Naive Bayesian classifier, fills in as taking after inference:

Step 1: Let T be a training set of tuples and their related class names. Each tuple is spoken to by a m-dimensional attribute vector, A = (a1, a2, … .. , am), m estimations made on the tuple from m properties, individually, X1, X2, … , Xm.

Step 2: Suppose that there are n classes D1, D2, … . , and Dn. Given a tuple, A, the classifier will anticipate that A has a place with the class having the most noteworthy back likelihood, adapted on A. That is, the guileless Bayesian classifier predicts that tuple A has a place with the class Tj if and just if

$P\left(\left(D_j\middle|A\right) > P\left((D_k|A)\right)\right.$ for $1 \le k \le n, k \ne j$ (7)

The boost P(Dj|A). The class Dj for which P(Dk|A) is amplified is known as the most extreme posterior hypothesis. By Bayes' hypothesis (Next condition)

$$P\left(D_j\middle|A\right) = \frac{P\left(A\middle|D_j\right)P(D_j)}{P(A)} \quad (8)$$

Step 3: Since P(A) is consistent for all classes, just (P(Dj|A) = P(A |Dj)P(Dj)) should be amplified.

Step 4: Based on the supposition is that properties are restrictively free (i.e., no reliance connection between attributes), the registering of P(A|Dj) utilizing the accompanying condition:

$$P\left(A\middle|D_j\right) = \prod_{i=1}^{m} P\left(a_i\middle|D_j\right) \quad (9)$$

Diminishes the calculation cost by Equation (P(Dj|A) = P(A |Dj)P(Dj), just numbers the class appropriation. On the off chance that Xi is unmitigated, P(Ai|Dj) is the no. of tuples in Dj having esteem Ai for Xi separated by |Dj, T| no. of tuples of Dj in T. Also, if Xi is persistent esteemed, P(Ai|Dj) is typically processed in view of Gaussian circulation with a mean μ and standard deviation σ and P(Ai|Dj) is:

$$g\left(x, \mu, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (10)$$

$$P\left(D_j\middle|A\right) = g\left(a_i, \mu_{Dj}, \sigma_{Dj}\right) \quad (11)$$

Where μ is the mean and σ is the difference. On the off chance that a property estimation doesn't happen with each class esteem, the likelihood will be zero, and a posteriori likelihood will likewise be zero.

K-Nearest Neighbour Classification Method

The k-nearest neighbors algorithm is one of the most used algorithms in machine learning. It is a learning method bases on instances that does not required a learning phase. The training sample, associated with a distance function and the choice function of the class based on the classes of nearest neighbors is the model developed. Before classifying a new element, we must compare it to other elements using a similarity measure. Its k-nearest neighbors are then considered, the class that appears most among the neighbors is assigned to the element to be classified. The neighbors are weighted by the distance that separate it to the new elements to classify The proper functioning of the method depends on the choice of some number of parameter such as the parameter k which represents the number of neighbors chosen to assign the class to the new element, and the distance used.

Decision Tree Classification Method

The decision tree is a structure that includes root node, branch and leaf node. Each internal node denotes a test on attribute, each branch denotes the outcome of test and each leaf node holds the class label. The topmost node in the tree is the root node. The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48.

Support Vector Machine

Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features.

Artificial Neural Network Classification Method

Artificial Neural Network (ANN): is a collection of neuron like processing units with weight connections between the units. It maps a set of input data onto a set of appropriate output data. It consists of 3 layers: input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weights wji of respective connections from input layer. The output Yj is function of $Yj = f(\Sigma wji\ xi)$ Where f is a simple threshold function such as sigmoid or hyperbolic tangent function

4. Result and Discussion

4.1 Dataset Description

The benchmark Chronic Kidney Disease dataset is considered for the classification of kidney disease using data mining in this research work. Table 1 gives the detailed description of the CKD dataset.

Table 1: Description of the Chronic Kidney Disease dataset

Sl.No	Feature Name	Description
1	Age	Age -age in years
2	bp	blood pressure - bp in mm/Hg
3	sg	Specific gravity - sg - (1.005,1.010,1.015,1.020,1.025)
4	al	Albumin - al - (0,1,2,3,4,5)
5	su	Sugar - su - (0,1,2,3,4,5)
6	rbc	Red Blood cell - rbc - (normal,abnormal)
7	pc	pus cell - pc - (normal,abnormal)
8	pcc	pus cell clumps - pcc - (present,notpresent)
9	ba	bacteria - ba - (present,notpresent)
10	bgr	blood glucose random - ba - (present,notpresent)
11	bu	blood urea –bu in mgs/dl
12	sc	serum creatinine - sc in mgs/dl
13	sod	sodium - sod in mEq/L
14	pot	potassium - pot in mEq/L
15	hemo	haemoglobin - hemo in gms
16	pcv	packed cell volume
17	wc	white blood cell count - wc in cells/cumm
18	rc	red blood cell count - rc in millions/cmm
19	htn	hyptertension - htn - (yes,no)
20	dm	diabetes mellitus - dm - (yes,no)
21	cad	coronary artery disease - cad - (yes,no)
22	appet	appetite - appet - (good,poor)
23	pe	pedal edema - pe - (yes,no)
24	ane	anemia - ane - (yes,no)
25	class	class class - (ckd,notckd)

4.2 Number of Features obtained

Table 2 depicts the number of features obtained by using correlation-based feature selection with PSO search method.

Table 2: Number of Features obtained by Feature Selection

Sl.No	Feature Name
1	Blood Pressure
2	Specific Gravity
3	Albumin
4	Red Blood Cells
5	Pus Cell
6	Packed Cell Volume(numerical)
7	Hypertension
8	Diabetes Mellitus
9	Appetite
10	Pedal Edema
11	Anemia

4.3 Performance Metrics

Table 3 depicts the performance metrics considered for evaluating the feature selection technique and classification techniques.

Table 3: List of Performance Metrics

Metrics	Equation
Accuracy	$\frac{TP + TN}{TP + FN + TN + FP}$
True Positive Rate (TPR)	$\frac{TP}{TP + FN}$
False Positive Rate (FPR)	$\frac{FP}{FP + TN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F-Measure	$2.\ \frac{Precision.\ Recall}{Precision+Recall}$
Mean Absolute Error (MAE)	$\frac{1}{N}\sum_{I=1}^{N}\lvert\hat{\theta}_i - \theta_i\rvert$
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}$
Relative Absolute Error (RAE)	$\frac{\sum_{I=1}^{N}\lvert\hat{\theta}_i - \theta_i\rvert}{\sum_{I=1}^{N}\lvert\bar{\theta}_i - \theta_i\rvert}$

Root Relative Squared Error (RRSE)	$\sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{\sum_{i=1}^{N}(\bar{\theta}_i - \theta_i)^2}}$

Table 4 depicts the performance analysis of the Naïve Bayes classification technique for original dataset and feature selection processed dataset. From the table 4, the pre-processed dataset performs well in all aspect when it is compared with the original dataset using Naïve Bayes classification method.

Table 4: Performance analysis of the Naïve Bayes classification techniques for original dataset and pre-processed dataset

Performance Metrics	Type of Dataset	
	Original Dataset	**Pre-Processed Dataset**
Accuracy	78.2278 %	**96.25 %**
Kappa Statistic	0.5672	**0.9274**
MAE	0.0879	**0.0296**
RMSE	0.2691	**0.145**
RAE	45.8877 %	**8.7118 %**
RRSE	87.3506 %	**35.2013 %**
TPR	0.782	**0.963**
FPR	0.172	**0.021**
Precision	0.812	**0.966**
F-Measure	0.791	**0.963**
ROC Area	0.892	**1.000**

Table 5 depicts the performance analysis of the ANN classification technique for original dataset and feature selection processed dataset. From the table 5, the pre-processed dataset performs well in all aspect when it is compared with the original dataset using Artificial Neural Network classification method.

Table 5: Performance analysis of the ANN classification techniques for original dataset and pre-processed dataset

Performance Metrics	Type of Dataset	
	Original Dataset	**Pre-Processed Dataset**
Accuracy	70.8861 %	**96.25 %**
Kappa Statistic	0.3688	**0.9278**
MAE	0.1165	**0.0248**
RMSE	0.3209	**0.157**
RAE	60.3793 %	**7.2977 %**
RRSE	104.1767 %	**38.1254 %**
TPR	0.709	**0.963**
FPR	0.347	**0.015**

Precision	0.6885	**0.969**
F-Measure	0.684	**0.964**
ROC Area	0.693	**0.978**

Table 6 depicts the performance analysis of the Support Vector Machine classification technique for original dataset and feature selection processed dataset. From the table 6, the pre-processed dataset performs well in all aspect when it is compared with the original dataset using Support Vector Machine classification method.

Table 6: Performance analysis of the SVM classification techniques for original dataset and pre-processed dataset

Performance Metrics	**Type of Dataset**	
	Original Dataset	**Pre-Processed Dataset**
Accuracy	82.0253 %	**98 %**
Kappa Statistic	0.625	**0.9609**
MAE	0.2397	**0.2267**
RMSE	0.3248	**0.2802**
RAE	124.2287 %	**66.6271%**
RRSE	105.4194 %	**68.0246%**
TPR	0.820	**0.980**
FPR	0.181	**0.016**
Precision	0.813	**0.980**
F-Measure	0.813	**0.980**
ROC Area	0.808	**0.982**

Table 7 depicts the performance analysis of the K-Nearest Neighbor classification technique for original dataset and feature selection processed dataset. From the table 7, the pre-processed dataset performs well in all aspect when it is compared with the original dataset using KNN classification method.

Table 7: Performance analysis of the KNN classification techniques for original dataset and pre-processed dataset

Performance Metrics	**Type of Dataset**	
	Original Dataset	**Pre-Processed Dataset**
Accuracy	74.4304 %	90.5 %
Kappa Statistic	0.4341	0.8067
MAE	0.1023	0.0633
RMSE	0.3198	0.2517
RAE	53.0184 %	18.6164 %
RRSE	103.813 %	61.0936 %
TPR	0.744	0.905

FPR	0.327	0.108
Precision	0.73	0.908
F-Measure	0.7165	0.892
ROC Area	0.705	0.898

Table 8 depicts the performance analysis of the J48 decision tree classification technique for original dataset and feature selection processed dataset. From the table 8, the pre-processed dataset performs well in all aspect when it is compared with the original dataset using J48 Decision tree classification method.

Table 8: Performance analysis of the J48 classification techniques for original dataset and pre-processed dataset

Performance Metrics	**Type of Dataset**	
	Original Dataset	**Pre-Processed Dataset**
Accuracy	81.2658 %	96.25 %
Kappa Statistic	0.593	0.9275
MAE	0.1226	0.0459
RMSE	0.2478	0.1416
RAE	63.535 %	13.4814 %
RRSE	80.4331 %	34.3786 %
TPR	0.813	0.963
FPR	0.233	0.022
Precision	0.8045	0.966
F-Measure	0.798	0.963
ROC Area	0.785	0.985

Figure 2 depicts the graphical representation of the performance analysis on Accuracy in % of the original dataset and pre-processed dataset using NB, ANN, KNN, SVM and J48 classification techniques. From the figure 2, it is clear that the reduced dataset gives more accuracy than original dataset when using classification methods.

Figure 3 depicts the graphical representation of the performance analysis on Kappa Statistics in % of the original dataset and pre-processed dataset using NB, ANN, KNN, SVM and J48 classification techniques. From the figure 3, it is clear that the reduced dataset gives more accuracy than original dataset when using classification methods.

Figure 4 depicts the graphical representation of the performance analysis on Mean Absolute Error (MAE) of the original dataset and pre-processed dataset using NB, ANN, KNN, SVM and J48 classification techniques. From the figure 4, it is clear that the reduced dataset gives less Mean Absolute Error (MAE) than original dataset when using classification methods.

Figure 5 depicts the graphical representation of the performance analysis on Root Mean Squared Error (RMSE) of the original dataset and pre-processed dataset using NB, ANN, KNN,

SVM and J48 classification techniques. From the figure 5, it is clear that the reduced dataset gives less RMSE than original dataset when using classification methods.

Figure 6 depicts the graphical representation of the performance analysis on Relative Absolute Error (RAE) in % of the original dataset and pre-processed dataset using NB, ANN, KNN, SVM and J48 classification techniques. From the figure 6, it is clear that the reduced dataset gives less RAE than original dataset when using classification methods.

Figure 7 depicts the graphical representation of the performance analysis on Root Relative Squared Error (RRSE) in % of the original dataset and pre-processed dataset using NB, ANN, KNN, SVM and J48 classification techniques. From the figure 7, it is clear that the reduced dataset gives less RRSE than original dataset when using classification methods.

Figure 8 depicts the graphical representation of the performance analysis on True Positive Rate (TPR) of the original dataset and pre-processed dataset using NB, ANN, KNN, SVM and J48 classification techniques. From the figure 8, it is clear that the reduced dataset gives more TPR than original dataset when using classification methods.

Figure 9 depicts the graphical representation of the performance analysis on False Positive Rate (FPR) of the original dataset and pre-processed dataset using NB, ANN, KNN, SVM and J48 classification techniques. From the figure 9, it is clear that the reduced dataset gives less FPR than original dataset when using classification methods.

Figure 10 depicts the graphical representation of the performance analysis on Precision of the original dataset and pre-processed dataset using NB, ANN, KNN, SVM and J48 classification techniques. From the figure 10, it is clear that the reduced dataset gives more precision than original dataset when using classification methods.

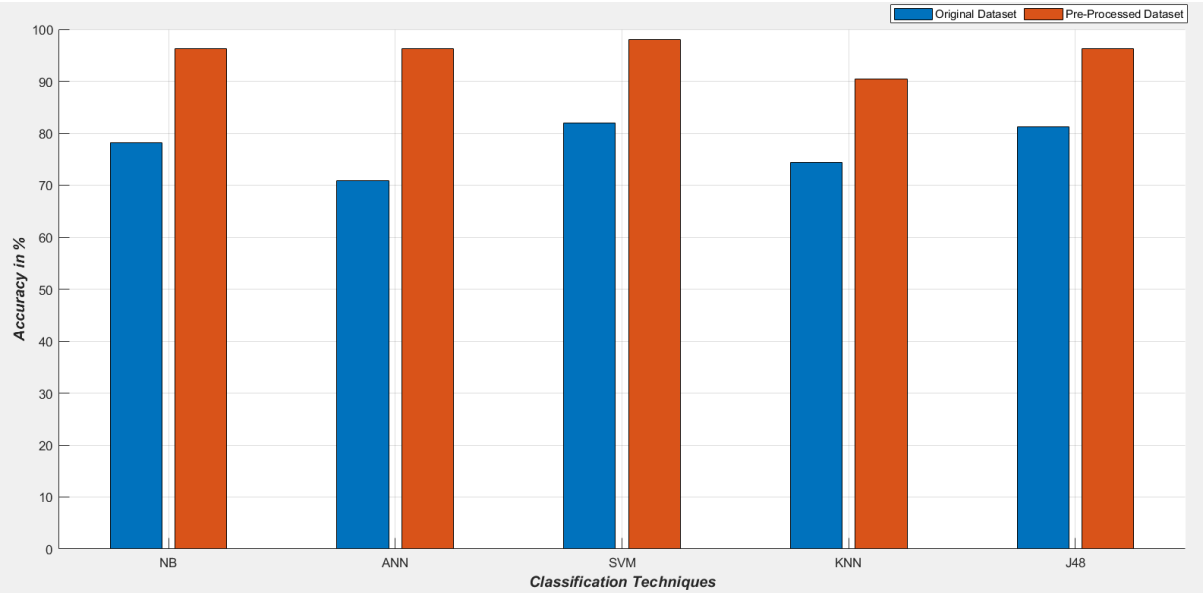

Figure 2: Performance Analysis on Accuracy in % of the original dataset and pre-processed dataset using NB, ANN, KNN, SVM and J48 classification techniques

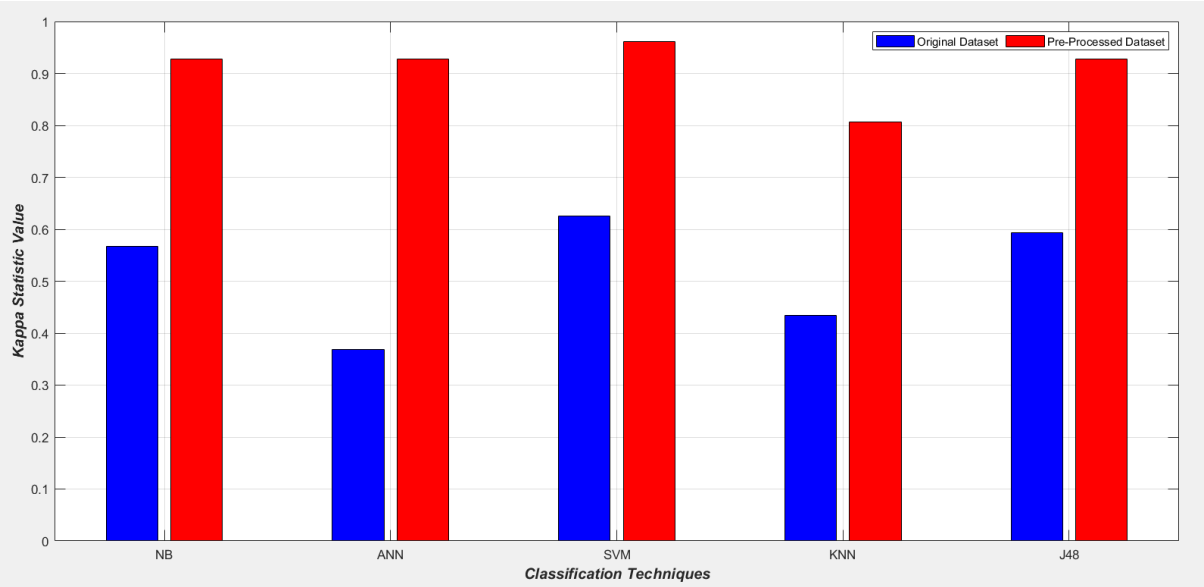

Figure 3: Performance analysis on the Kappa Statistic of the original dataset and pre-processed dataset using NB, ANN, SVM, KNN and J48 classification techniques

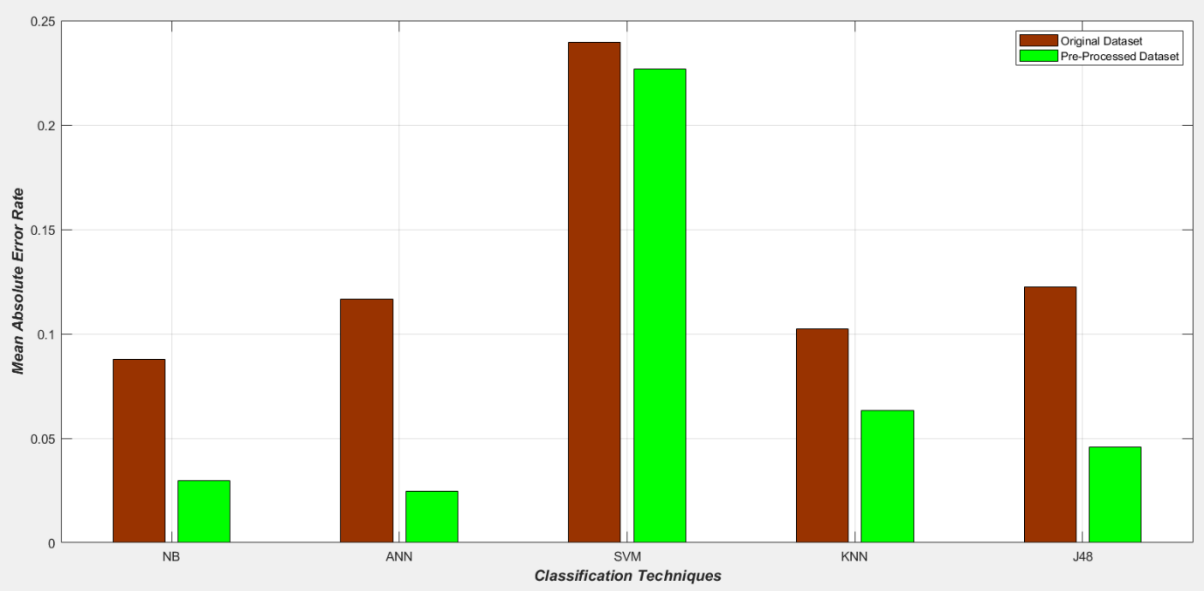

Figure 4: Performance Analysis on the Mean Absolute Error of the original dataset and pre-processed dataset using NB, ANN, SVM, KNN and J48 classification techniques

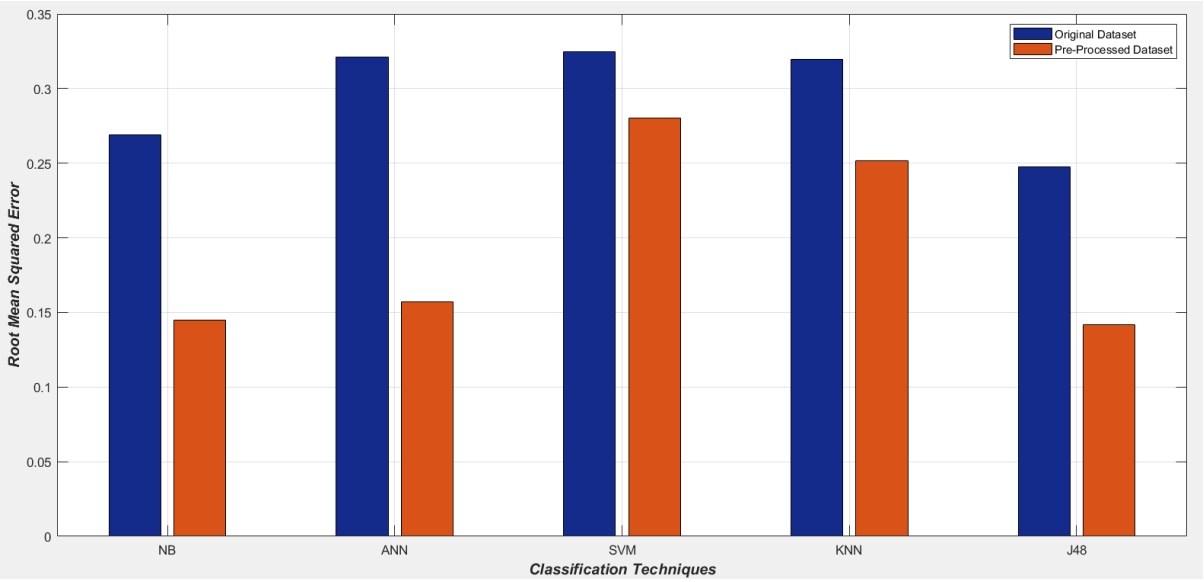

Figure 5: Performance Analysis on the Root Mean Squared Error of the original dataset and Pre-processed dataset using NB, ANN, SVM, KNN and J48 classification techniques

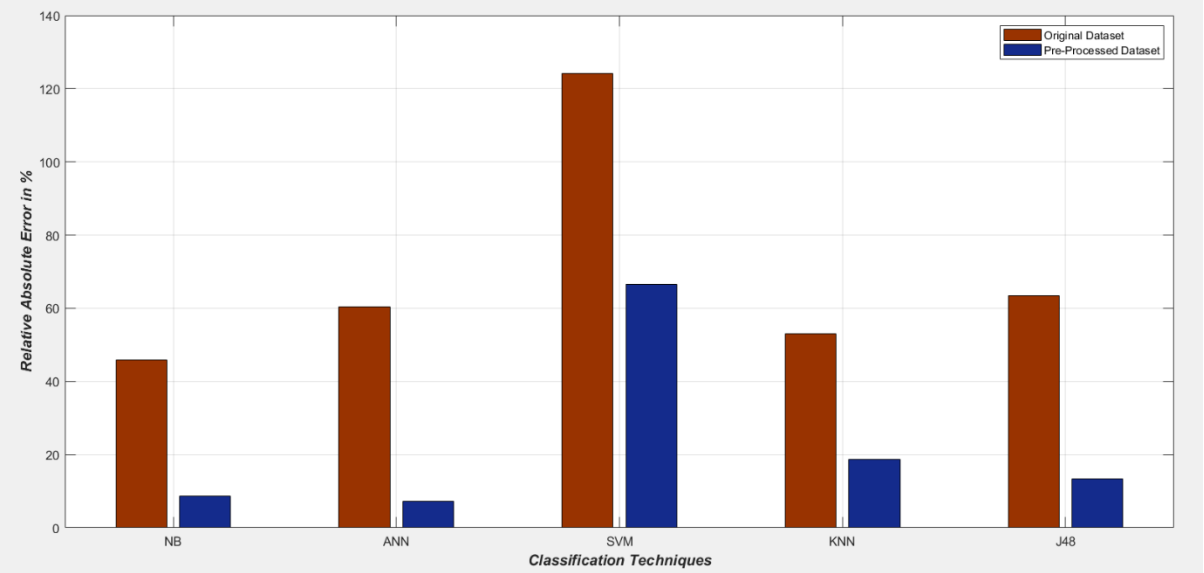

Figure 6: Performance analysis on the Relative Absolute Error (RAE) in % of the original dataset and pre-processed dataset using NB, ANN, SVM, KNN and J48 classification techniques

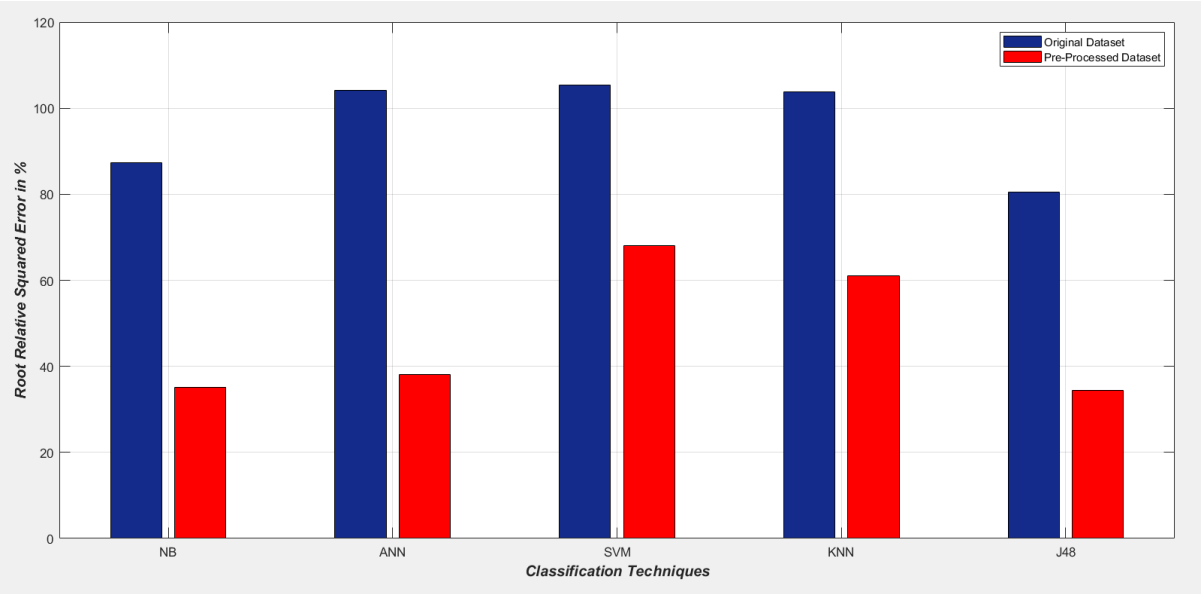

Figure 7: Performance analysis on the Root Relative Squared Error in % of the original dataset and pre-processed dataset using NB, ANN, SVM, KNN and J48 classification techniques

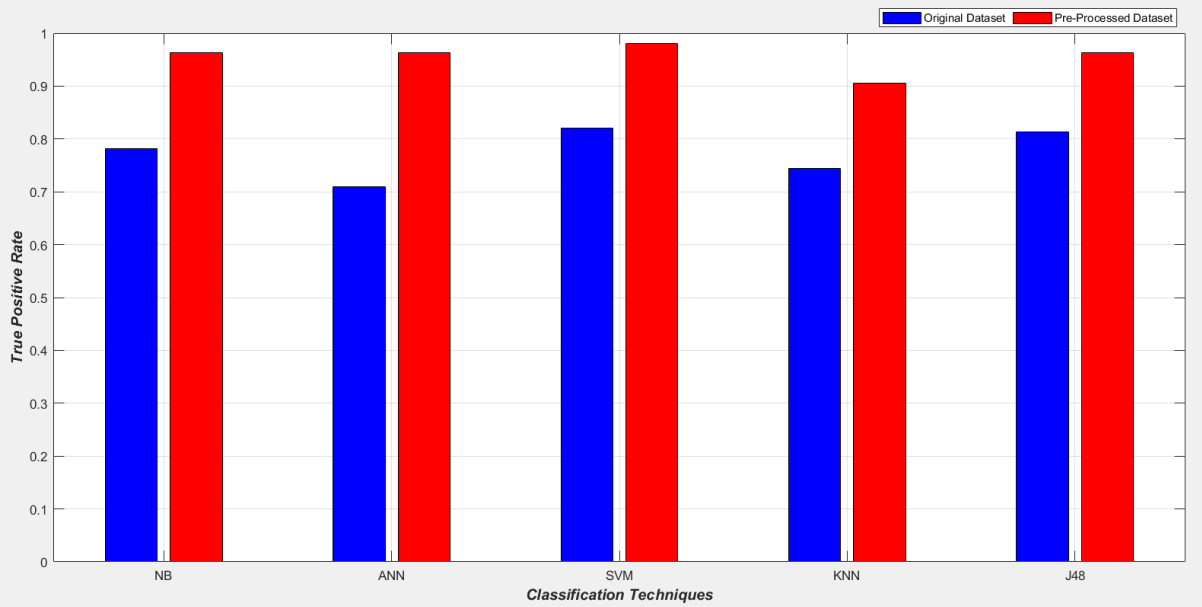

Figure 8: Performance analysis on the True Positive Rate of the original dataset and pre-processed dataset using NB, ANN, SVM, KNN and J48 classification techniques

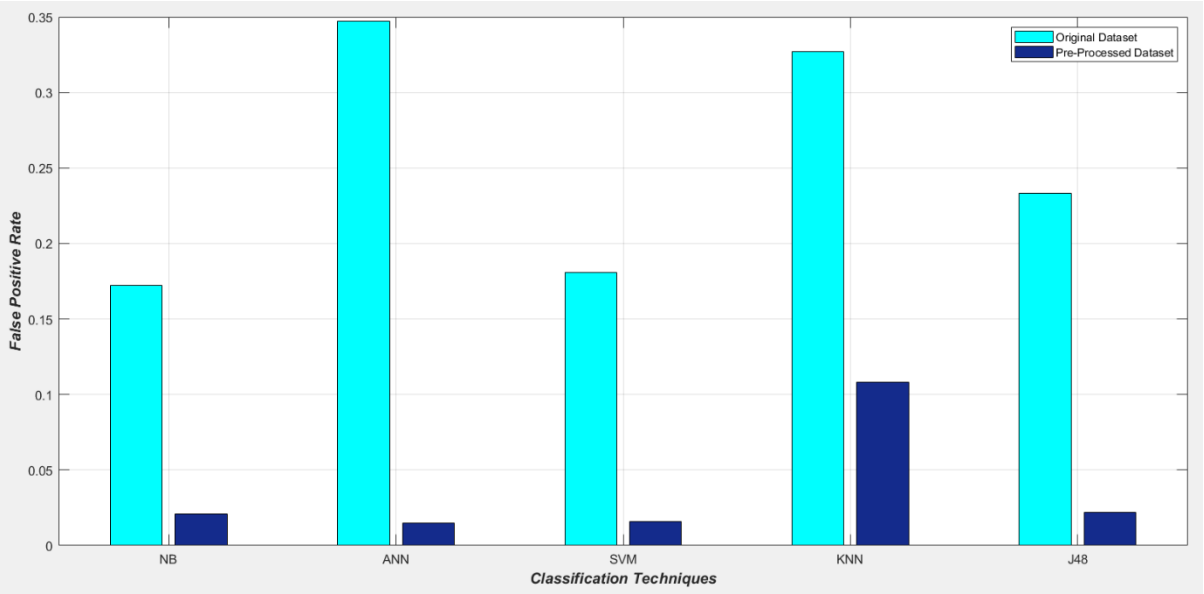

Figure 9: Performance analysis on the False Positive Rate of the original dataset and pre-processed dataset using NB, ANN, SVM, KNN and J48 classification techniques

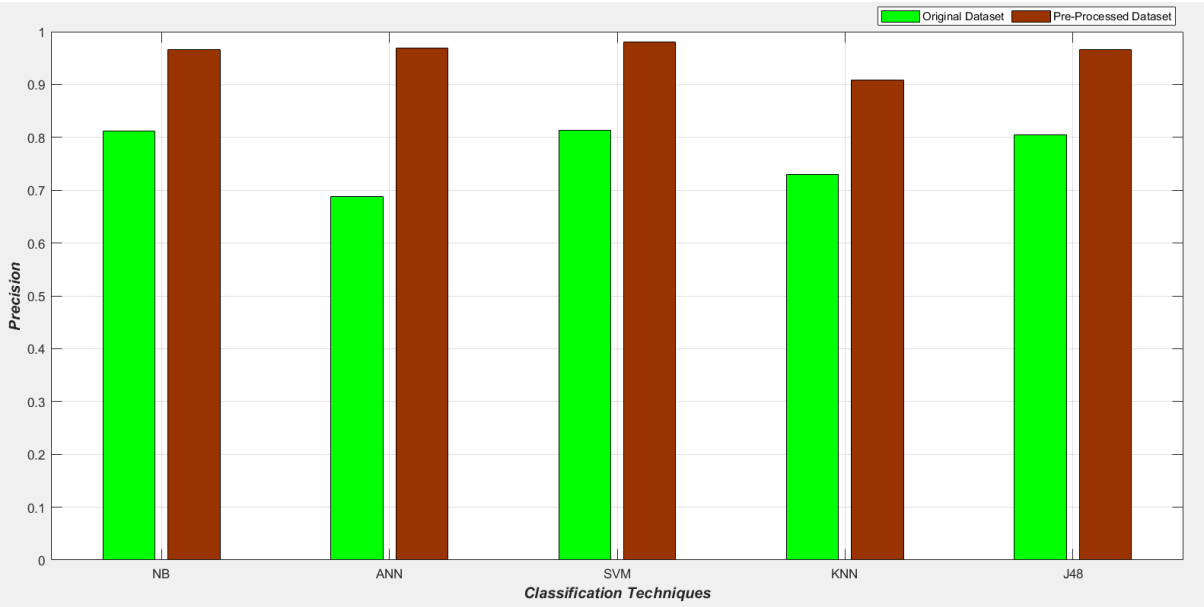

Figure 10: Performance Analysis on the Precision of the original dataset and pre-processed dataset using NB, ANN, SVM, KNN and J48 classification techniques

5. Conclusion

In this study feature selection method and ensemble method have been used on CKD dataset to improve the accuracy of the classifiers. For feature selection method Correlation based Feature Selection method combine with Particle Swarm Optimization (PSO) search have been used. These methods have been used both proposed feature selection method to improve the accuracy of machine learning classifiers. The accuracy rate of KNN, J48, ANN, NB, and SVM classifier on CKD dataset has been compared to its accuracy, on a reduced dataset which has been used Correlation based Feature Selection method combine with Particle Swarm

Optimization (PSO) search. The experimental result shows that after reducing the dataset the accuracy of the classifier has been improved.

In the future, we will integrate the real-time patient data combine with recorded dataset by storing patient data the recorded containing will stored in database to improve chronic kidney disease and other disease prediction system by using hybrid algorithm. In the future, we suggest that including some visualization method like logical representation of the useful knowledge base transformation into visualize to the user.

Reference

[1] K. Chandel, V. Kunwar, S. Sabitha, T. Choudhury, and S. Mukherjee, A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques, CSI Trans. ICT, vol. 4, no. 24, pp. 313319, Dec. 2016.

[2] B. Boukenze, A. Haqiq, and H. Mousannif, Predicting Chronic Kidney Failure Disease Using Data Mining Techniques, in Advances in Ubiquitous Networking 2, Springer, Singapore, 2017, pp. 701712.

[3] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation, Biomed. Signal Process. Control, Feb. 2017.

[4] R. Ani, G. Sasi, U.R. Sankar, & O.S. Deepa, (2016, September). Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification I EEE Conference Publication.[Online]. Available: http://ieeexplore.ieee.org/abstract/document/7732224/?reload=true. [Accessed: 15-Dec-2017].

[5] L.-C. Cheng, Y.-H. Hu, and S.-H. Chiou, Applying the Temporal Abstraction Technique to the Prediction of Chronic Kidney Disease Progression, J. Med. Syst., vol. 41, no. 5, p. 85, May 2017.

[6] H. Polat, H. D. Mehr, and A. Cetin, Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods, J. Med. Syst., vol. 41, no. 4, p. 55, Apr. 2017.

[7] P. Pangong and N. Iam-On, Predicting transitional interval of kidney disease stages 3 to 5 using data mining method, in 2016 Second Asian Conference on Defence Technology (ACDT), 2016, pp. 145150.

[8] K. R. A. Padmanaban and G. Parthiban, Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease, Indian J. Sci. Technol., vol. 9, no. 29, Aug. 2016.

[9] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, Performance Analysis of Data Mining Classification Techniques to Predict Diabetes, Procedia Comput. Sci., vol. 82, no. Supplement C, pp. 115121, Jan. 2016.

[10] U. N. Dulhare and M. Ayesha, Extraction of action rules for chronic kidney disease using Naive Bayes classifier, in 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, pp. 15.

[11] N. Borisagar, D. Barad, and P. Raval, Chronic Kidney Disease Prediction Using Back Propagation Neural Network Algorithm, in Proceedings of International Conference on Communication and Networks, Springer, Singapore, 2017, pp. 295303.

[12] A. I. Pritom, M. A. R. Munshi, S. A. Sabab, and S. Shihab, Predicting breast cancer recurrence using effective classification and feature selection technique, in 2016 19th International Conference on Computer and Information Technology (ICCIT), 2016, pp. 310314.

[13] Lakshmi, K.R., Nagesh, Y., & VeeraKrishna, M, Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability, International Journal of Advances in Engineering & Technology, Mar., Vol.7, Issue 1, Page No: 242-254, 2013.

[14] Lambodar Jena, Narendra Ku. Kamila, Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney Disease, International Journal of Emerging Research in Management & Technology ISSN: 2278-9359(Vol 4, Issue-11), IJERMT, 2015.

[15] Vijayarani, S., & Dhayananad, M., S Kidney Disease Prediction using SVM and ANN algorithms, International Journal of Computing and Business Research (IJCBR), Volume 6 Issue 2 March 2015.

[16] Boukenze, Basma, Abdelkrim Haqiq, and Hajar Mousannif. "Predicting Chronic Kidney Failure Disease Using Data Mining Techniques." Advances in Ubiquitous Networking 2. Springer, Singapore, 2017. 701-712.

[17] Yadollahpour, Ali, et al. "Designing and implementing an ANFIS based medical decision support system to predict chronic kidney disease progression." Frontiers in physiology 9 (2018).

[18] Sedighi, Zeinab, Hossein Ebrahimpour-Komleh, and Seyed Jalaleddin Mousavirad. "Featue selection effects on kidney desease analysis." 2015 International Congress on Technology, Communication and Knowledge (ICTCK). IEEE, 2015.

[20] Subhashini, M., & Gopinath, R., Mapreduce Methodology for Elliptical Curve Discrete Logarithmic Problems – Securing Telecom Networks, International Journal of Electrical Engineering and Technology, 11(9), 261-273 (2020).

[21] Upendran, V., & Gopinath, R., Feature Selection based on Multicriteria Decision Making for Intrusion Detection System, International Journal of Electrical Engineering and Technology, 11(5), 217-226 (2020).

[22] Upendran, V., & Gopinath, R., Optimization based Classification Technique for Intrusion Detection System, International Journal of Advanced Research in Engineering and Technology, 11(9), 1255-1262 (2020).

[23] Subhashini, M., & Gopinath, R., Employee Attrition Prediction in Industry using Machine Learning Techniques, International Journal of Advanced Research in Engineering and Technology, 11(12), 3329-3341 (2020).

[24] Rethinavalli, S., & Gopinath, R., Classification Approach based Sybil Node Detection in Mobile Ad Hoc Networks, International Journal of Advanced Research in Engineering and Technology, 11(12), 3348-3356 (2020).

[25] Rethinavalli, S., & Gopinath, R., Botnet Attack Detection in Internet of Things using Optimization Techniques, International Journal of Electrical Engineering and Technology, 11(10), 412-420 (2020).

[26] Priyadharshini, D., Poornappriya, T.S., & Gopinath, R., A fuzzy MCDM approach for measuring the business impact of employee selection, International Journal of Management (IJM), 11(7), 1769-1775 (2020).

[27] Poornappriya, T.S., Gopinath, R., Application of Machine Learning Techniques for Improving Learning Disabilities, International Journal of Electrical Engineering and Technology (IJEET), 11(10), 392-402 (2020).

[28] Periyasamy, R., and Nalini, D. (2015). Binary back propagation based lift association mining for heart disease and stroke identification. International journal of applied engineering research, 10(6), pp. 16071-16087.

[29] Periyasamy, R., and Nalini, D. (2015). Lloyd Minkowski based K -Means clustering for effective diagnosis of heart disease and stroke. International review on computers and software, 10(6).

[30] Periyasamy, R., and Nalini, D. (2015). Pairwise proximity clustering for medical disease identification. International journal of advanced computer science and technology, 5(1), pp. 1-10.

[31] Periyasamy, R., and Nalini, D. (2015). A clinical heart disease decision supportive optimized mining method for effective disease diagnosis. International journal of computer applications, 126 (9).

[32] Jayasimman, L., Geetha Dhanalakshmi, V. (2020). Design of Enhanced Dynamic Resource Allocation Framework for Heterogeneous Cloud Environment, IJITEE, 9(3), 1563-1568.

[33] Jayasimman, L., Geetha Dhanalakshmi, V. (2018). A Study on Spatial-Temporal Load Balancing Approach in Cloud Computing, JCSE, 6(11).

[34] Geetha Dhanalakshmi, V., Jayasimman, L. (2018). A Review on Dynamic Resource Allocation strategies for Cloud Computing, IJSRCSAMS, 7(4).

[35] Geetha Dhanalakshmi, V., Jayasimman, L. (2018). Resource Provisioning for Ensuring QoS in Virtualized Environments, JCSE, 6(4).

[36] James Manoharan, J., Hari Ganesh, S. (2016). A framework for enhancing the efficiency of k-means clustering algorithm to avoid formation of empty clusters. Middle-East J. Sci. Res (MEJSR).

[27] James Manoharan, J., Hari Ganesh, S., Sathiaseelan, JGR. (2016). Outlier detection using enhanced k-means clustering algorithm and weight-based center approach, Int. J. Comput. Sci. Mobile Comput, 5(4), 453-464.

[38] James Manoharan, J., Hari Ganesh, S. (2016). INITIALIZATION OF OPTIMIZED K-MEANS CENTROIDS USING DIVIDE-AND-CONQUER METHOD, ARPN Journal of Engineering and Applied Sciences, 11(2), 1086-1091.

[39] James Manoharan, J., Ganesh, S. H., Felciah, M. L. P., & Shafreenbanu, A. K. (2014, February). Discovering Students' Academic Performance Based on GPA Using K-Means Clustering Algorithm. In 2014 World Congress on Computing and Communication Technologies (pp. 200-202). IEEE.

[40] Vijilesh, VG., Hari Ganesh, S., James Manoharan, J. (2018). An Enhancing the Performance of High Utility Itemset Mining using Utility Information Record, International Journal of Pure and Applied Mathematics, 118(17), 257-272.

[41] Selvaramalakshmi, P., Hari Ganesh, S., James Manoharan, J. (2017). A Novel PSS Stemmer for String Similarity Joins, 2017 World Congress on Computing and Communication Technologies (WCCCT), 147-150, IEEE.