Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 7, Issue 10, October -2020

A Comparative Study of Building Disease Classification Model through Supervised Machine Learning Algorithms for HealthCare Data

Anjali Choudhary[1], Shrwan Ram[2]

Computer Science Department, M.B.M. Engineering College Jodhpur

Abstract - *With the emerging new disease patterns, new technologies like machine learning and data analytics are proving to provide promising solutions in early detection of symptoms, decoding various patterns, predicting various responses to drugs, etc. These are proving to be very helpful to biomedical professionals, the healthcare industry, and patients. Machine learning can be used to develop models for the prediction of chronic diseases.*
In this paper, machine learning techniques will be compared using the benchmark datasets. The different types of data classification methods and techniques are available such as Decision Tree, k-Nearest Neighbor, Support Vector Machine, Naive Bayes, Logistic Regression, and Linear Discriminant algorithms. The objective of the thesis work is to do the comparative study and evaluation of supervised machine learning methods with the help of reduced healthcare datasets collected. It is shown that the accuracy of the SVM classifier is better than the others.

Keywords - *Supervised Machine learning, Chronic Kidney Disease, Classification Techniques, Decision tree, k-nearest neighbor, Support vector machine, Naïve bayes, Logistic regression, Linear discriminant.*

I. INTRODUCTION

Chronic kidney disease (CKD) is commonly referred to as disorders that affect kidney structure and heterogeneous function. Chronic kidney disease (CKD) is a global health issue with high morbidity and mortality rates and it causes other diseases. Since there are no visible symptoms during the early phases of CKD, patients often fail to identify the disease. Early CKD diagnosis allows patients to obtain timely treatment to increase the progression of this disease [1]. Machine learning models can effectively help clinicians achieve this aim due to their fast and accurate recognition performance. Machine learning techniques for diagnosing CKD or non-CKD are used. In addition to having less pathological tests with less expense and less time, general knowledge of people must be increased. CKD data set contains 13 attributes plus one attribute for class (binary). It contains 400 samples to two different classes ("CKD" - 250 cases; "NOTCKD" - 150 cases). Out of 14 attributes, 8 are numeric and 6 are nominal.
This paper aims to predict the patient's status in Chronic Kidney Disease (CKD) or non Chronic Kidney Disease (non-CKD). To predict the value in machine learning classification algorithms have been used. Classification models that have been built with different classification algorithms will predict the CKD and non-CKD status of the patient [3]. Also, a comparative study among these algorithms based on prediction accuracy is performed. Further, K-fold Cross-Validation is used to generate randomness in the data. These models have applied to a recently collected CKD dataset downloaded from the kaggle with 400 data records and 14 attributes. The results of the different models are compared. From the comparison, it has been observed that the model with the Support Vector Machine algorithm performed best with an accuracy of 98.30% for reduced attributes.

II. LITERATURE REVIEW

Rajesh Misir[1], Malay Mitra[2], Ranjit Kumar Samanta[2], "A Reduced Set of Features for Chronic Kidney Disease Prediction". In this paper author focused on predicting CKD or non-CKD with reasonable accuracy using fewer features. As suggested from our results, we may more concentrate on those reduced features for identifying CKD and thereby reduces uncertainty, saves time, and reduces costs.

Dr. S. Vijayarani[1], Mr.S.Dhayanand[2], "Data Mining Classification Algorithms For Kidney Disease Prediction". The main objective of this research work is to predict kidney diseases using classification algorithms such as Naïve Bayes and Support Vector Machine. This research work mainly focused on finding the best classification algorithm based on the classification accuracy and execution time performance factors. SVM performs better than Naïve Bayes.

Made Satria Wibawa[1, a], I Made Dendi Maysanjaya[2,b], I Made Agus Wirahadi Putra[1], "Boosted Classifier and Features Selection for Enhancing Chronic Kidney Disease Diagnose". This study developed a machine learning method using ensemble learning and feature selection to improve the quality of CKD diagnosis. The used data in this study are 24 attributes including signs, symptoms, and risk factors that may appear due to CKD. K-Nearest Neighbour algorithm (kNN), Naive Bayes, and Support Vector Machine (SVM) are used as base classifier. Based on the evaluation parameters, the best combination is achieved by the kNN classifier with a 0.981 of accuracy rate.

III. RESEARCH METHODOLOGY

3.1. Chronic Kidney Disease data

The dataset used to research this manuscript has been obtained from the Kaggle. There are in total fourteen parameters, most of which are clinical and the rest are physiological. Table 1 summarizes various parameters. In the preprocessing of the data the missing values were dealt with by replacing numeric and discrete integer values by attribute mean of all the instances with the same class-label as that of the instance under consideration and nominal values were replaced using attribute mode.

Table. 1: Various attributes and their type, class, and allowed values.

Attribute	Type	Class	Values
Blood pressure	Numerical	Predictor	in mm/Hg
Specific gravity	Nominal	Predictor	1.005,1.010,1.015,1.020,1.025
Albumin	Nominal	Predictor	0,1,2,3,4,5
Sugar	Nominal	Predictor	0,1,2,3,4,5
Red blood cells	Nominal	Predictor	Normal, Abnormal
Blood urea	Numerical	Predictor	in mgs/dl
Serum creatinine	Numerical	Predictor	in mgs/dl
Sodium	Numerical	Predictor	in mEq/L
Potassium	Numerical	Predictor	in mEq/L
Haemoglobin	Numerical	Predictor	in gms
WBC count	Numerical	Predictor	in cells/cumm
RBC count	Numerical	Predictor	in millions/cumm
Hypertension	Nominal	Predictor	Yes, No
Class	Nominal	Target	Ckd, Non-Ckd

3.2. Candidate classification techniques

The authors analyzed six different classifiers majorly based on the following techniques: Decision-Tree, K-nearest neighbor (KNN), Support vector machine (SVM), Naive Bayes, Logistic Regression, and Linear Discriminant. These techniques were selected for the analysis and study because of their popularity in the recent relevant literature. A brief description of the selected techniques has been given below:

1) Decision Tree: In these types of structures the algorithms are mirror images in form of a tree. The various parts of the tree are represented as various elements of algorithms. The root nodes represent the functions, the edges represent the results, the leaves represent the classification of samples, etc. The algorithms define the training samples and designate class labels. While working on unknown samples the decision tree designates the labels to unknown samples and places that at leaves. The separation of various samples is based on the functions assigned to various data samples. These functions can be values like myopic tests, G-statistics, etc. Generalization of any sample value and using measures to reduce complexity are options available in the decision tree method. The various examples of decision tree algorithms are ID3, CART, etc.

2) K-Nearest Neighbours: The k nearest neighbor classifier which is based on the theory that identical or near-identical samples lie close to each other, is one of the simplest algorithms of machine learning. The basis of classification is using Euclidean distance between the test data and each sample in the training data for separation of each sample data and forming sets. The K-Closest value data or samples are grouped. Due to simplicity KNN is also called lazy-learning. The data distribution is free of any assumption and thus non-pragmatic algorithm. However, ask is a very tiny positive integer, it becomes difficult to differentiate between the different groups as the value of K increases. It can be made easier by selecting the optimal value of K, cross-validation is used along with other techniques.

3) Support Vector Machine (SVM): Support vector machine is a supervised machine learning algorithm that determines the hyper planes which separating two groups of data and it is an underlying algorithm that revolves around the concept of "margin". Support vector machine classifiers based on the theory of statistical learning and the concept of structural risk minimization. The support vector machine classifies the data into two categories i.e. linearly separable and linearly non-separable data. Linearly separable data requires only one hyperplane for data separation while linearly non-separable data requires more than one hyperplane. Maximizing the margin produces the largest possible hyperplane gap.

4) Naive Bayes: Naive Bayes supervised machine learning algorithm is useful for very large data sets. And it is very simple to create. Ii is a Bayes Theorem-based classification technique with an assumption of independence among predictors. It is sophisticated classification methods are known to outperform. In a Naive Bayes classifier, the existence of a certain feature in a class is unrelated to the presence of any other feature.

5) Logistic Regression: Logistic regression is a very effective modeling method. In this modeler, the response variable is linear and it assumes that in the coefficients of the predictor variables. The response variable cannot be explicitly modeled by linear regression since it is discrete. It is mostly used for predicting dependent variables that are binary or multi-class. Instead of predicting the point estimate of the event itself, the model is therefore designed to predict the odds of its occurrence. Also, the modeler must select the correct inputs and determine their functional relationship to the response variable, based on his or her familiarity with data and data analysis. In a two-class issue, odds greater than 50 % would indicate that the case is allocated to the class designated as "1" and "0".

6) Linear Discriminant: Linear discriminant is the algorithm in which the data sample is already provided that belongs to a particular class. The classifiers based on the concept of different Gaussian distribution and different classes generate data. The parameters for each class of Gaussian distribution are calculated during the training process with the help of the fitting function and qualified classifiers detect the lowest misclassification cost to predict the classes of the new data.

3.3. Methodology

First, for each classifier that is chosen to train the system. Initially, all attributes corresponding to each event, i.e. input data, are defined and only one attribute is used from these to represent a decision on the given problem, i.e. output data. For the input attributes, unique values are specified. 5-fold cross-validation was used by scientists. A fair estimate of the accuracy of the final classifier trained with all the data is given by Cross-Validation. The procedure is the following:

1) Partition the data into k disjoint sets or folds
2) For each fold:
   a) Classifier is trained using out-of-fold observations.
   b) Model performance is assessed using the in-fold data.
3) The average test error overall folds are calculated.

Afterward, this methodology of each classifier differs but the last step i.e. common to all the classifiers is assigning a class to every single instance of the dataset.

IV. EXPERIMENTAL SETUP

All six classifiers were applied to the same dataset using MATLAB Version: 9.8.0.1417392 (R2020a) and the results were obtained and analyzed in the terms of predictive accuracy, sensitivity, specificity & error rate. The predictive accuracy of Z percent indicates that almost Z percent of instances can be identified correctly by the classifier. Sensitivity can be defined as the ability to identify cases that have the condition correctly. It can be defined mathematically as the number of true positives split by the number of true positives and false negatives. It is possible to describe accuracy as the ability to correctly diagnose cases that do not have the disorder. It can be represented mathematically as the number of true negatives separated by the number of true negatives and false positives. A classifier's error rate or misclassification rate. The Classification Matrix indicates the frequency of predictions that are accurate and wrong. It compares the actual values in the test dataset with those in the trained model that is expected.

Table 2: Confusion Matrix

Predicted	Classified as Healthy (0)	Classified as not Healthy (1)
Actual Healthy (0)	TP	FN
Actual not Healthy (1)	FP	TN

Table 2 shows the results of the Classification Matrix for all six models. The rows represent predicted values while the columns represent actual values. The left-most columns show values predicted by the models. The diagonal values show correct predictions. For Classification, this work constructed the Confusion Matrix for the frequency of correct and incorrect predictions. From the confusion matrix, the Specificity, Sensitivity, Accuracy Rate, and Error rate have been calculated. For measuring the accuracy rate and Error Rate, the following mathematical model is used.

Measures of Performance evaluation

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN} \qquad \text{Specificity} = \frac{TN}{FP+TN}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \qquad \text{Error rate} = \frac{FP+FN}{TP+FP+TN+FN}$$

V. RESULTS

Predictive accuracies of all six classifiers are shown in figure 5.1. Results show that the Support vector machine performed best among all the classifiers in terms of three out of four performance metrics with a predictive accuracy of

98.30%, a sensitivity of 0.9612, specificity of 0.9959, error rate of 1.75%. The three performance metrics in which it performed better than the rest of all classifiers are predictive accuracy, sensitivity, specificity, and error rate. Logistic regression performed better than the support vector machine in terms of specificity. The overall performance of the support vector machine is better than all the other classifiers considered for this study.

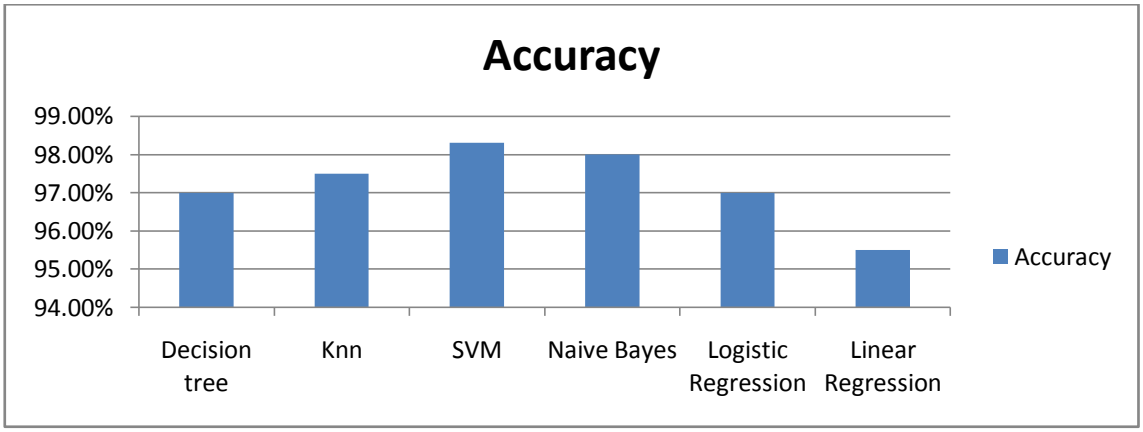
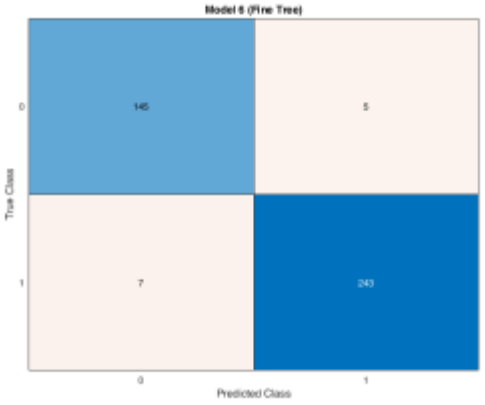
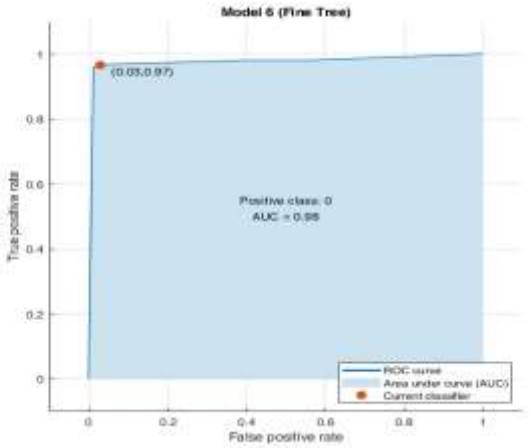


Fig. 5.1 Predictive accuracy of all classifiers

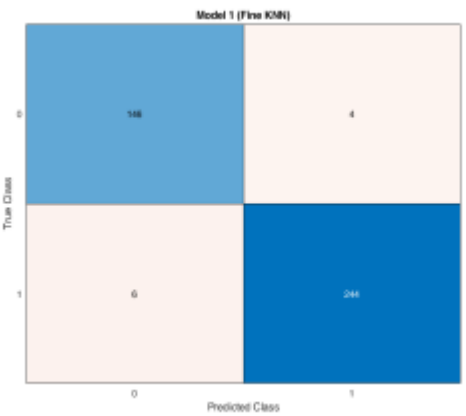
For the six classifiers, the various ROC curves are presented. The region under the curve tests the classifier's overall capacity to differentiate between the two classes. It is evident from the figure that, with 1 and 1 respectively, SVM and Linear discriminant have the largest AUC. Naive Bayes, decision tree and logistic regression, on the other hand, with 0.99,0.98 and 0.98 respectively AUC. kNN has the lowest AUC at 0.97. According to the ROC curve and the AUC SVM and Linear discriminant are the best classifiers. The confusion matrix of all six classifiers and the graphical representation of the ROC curve is shown below:



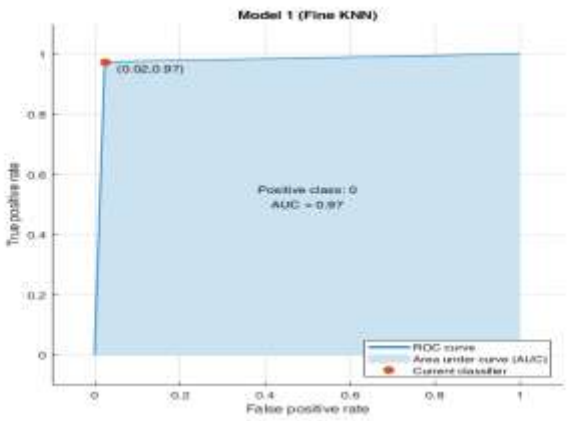
(a) Confusion Matrix of Decision tree



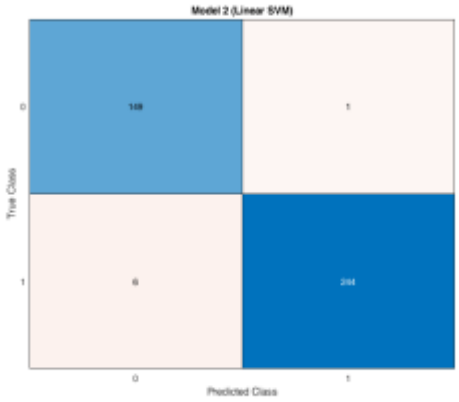
(b) ROC curve of Decision tree



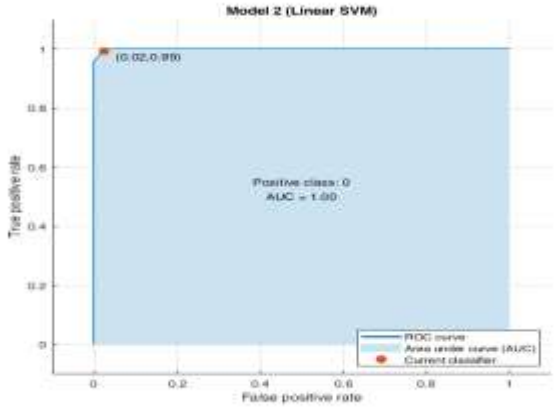
(a) Confusion Matrix of k-NN



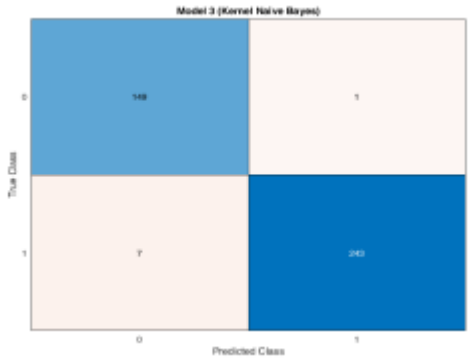
(b) ROC curve of k-NN



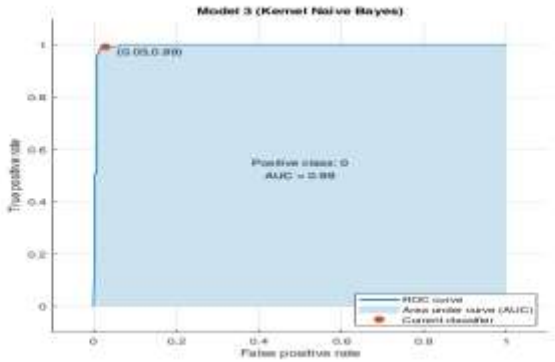
(a) Confusion Matrix of SVM



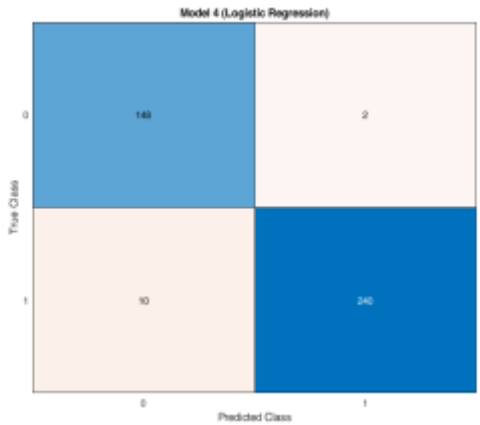
(b) ROC curve of SVM



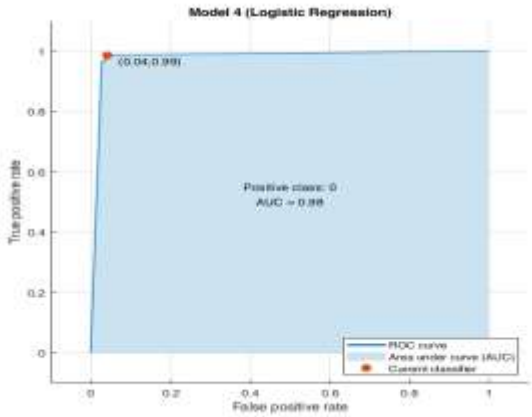
(a) Confusion Matrix of Naive Bayes



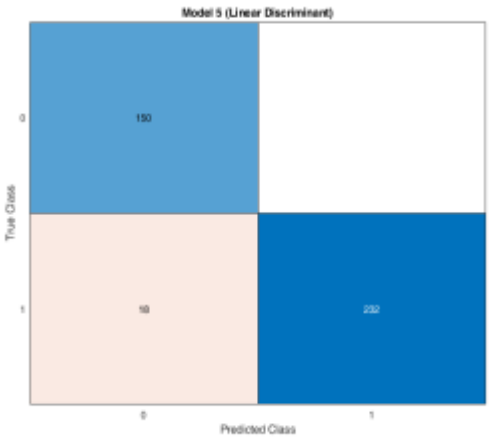
(b) ROC curve of Naive Bayes



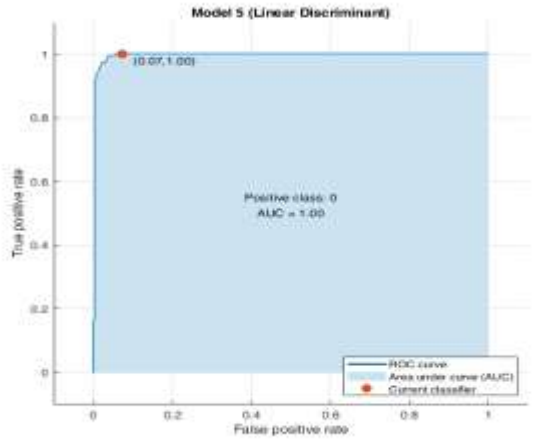
(a) Confusion Matrix of Logistic Regression



(b) ROC curve of Logistic Regression



(a) Confusion Matrix of Linear Discriminant



(b) ROC curve of Linear Discriminant

We compare confusion matrix and ROC curve of all six classifiers and we find out that SVM gives better accuracy of 98.30%, and better ROC curve with AUC = 1 which is far better than accuracies and ROC curve of the decision tree, k-nearest neighbor, support vector machine, naive bayes, logistic regression, and linear discriminant.

Table 5.1 Sensitivity, Specificity, Error rate and Accuracy values of all classifiers

Algorithms	Sensitivity	Specificity	Error rate	Accuracy
Decision tree	0.9539	0.9798	3.0%	97.0%
k-nearest neighbor	0.9605	0.9838	2.5%	97.5%
Support Vector Machine	0.9612	0.9959	1.75%	98.3%
Naive Bayes	0.9551	0.9959	2.0%	98.0%
Logistic Regression	0.8928	1.0000	4.5%	95.5%
Linear Discriminant	0.9367	0.9917	3.0%	97.0%

V. CONCLUSION AND FUTURE SCOPE

A common word that encompasses multiple heterogeneous kidney diseases is Chronic Kidney disease. Five to ten percent of the population suffers from this disease worldwide. Chronic Kidney Disease is a health problem in the world. Most cases of Chronic Kidney Disease in underdeveloped and developing countries go undiagnosed or are later diagnosed; this is one of the key reasons why, relative to developed countries where most patients go through regular check-up and diagnosis, a higher percentage of these cases are from developing and underdeveloped nations. More than 80% of all patients seeking kidney failure treatment are in wealthy countries with free access to health care and large elderly populations, according to the report [6]. For timely and reliable diagnosis of Chronic Kidney Disease, machine learning-based software can be used to help doctors check the outcomes of their diagnosis in a reasonably short time, thereby allowing a doctor to treat and diagnose more patients in less time compared to the scenario where he/she has to go through the diagnostic process entirely manually. In the future course of this study, by testing such hybrid or ensemble techniques, a subset of features can be extracted from the complete medical data set of chronic kidney disease of fourteen parameters (features) without affecting the efficiency of the classification process, so that the financial burden can be extracted from the complete medical data set of chronic kidney disease.

ACKNOWLEDGEMENT

We would like to thank Dr. N.C Barwar, Professor and Head of the Department, Computer Science and Engineering, MBM College, for his motivation without which we would not be able to write this paper. We are grateful to our family and friends for their support.

REFERENCES

[1] Rajesh Misir1, Malay Mitra2, Ranjit Kumar Samanta2, "A Reduced Set of Features for Chronic Kidney Disease Prediction", [1]Department of Computer Science, Vidyasagar University, Medinipur, [2]Department of Computer Science and Application, Expert Systems Laboratory, University of North Bengal, Darjeeling, West Bengal, India, February|2017.

[2] Dr. S. Vijayarani[1], Mr.S.Dhayanand[2], "Data Mining Classification Algorithms For Kidney Disease Prediction", Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India[1,2], International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August|2015.

[3] Made Satria Wibawa[1, a], I Made Dendi Maysanjaya[2,b], I Made Agus Wirahadi Putra[1,c], "Boosted Classifier and Features Selection for Enhancing Chronic Kidney Disease Diagnose", [1]Department of Information System, STMIK STIKOM Bali, Bali, Indonesia - 80226 [2]Department of Informatics Engineering Education, Universitas Pendidikan Ganesha, Bali, Indonesia – 81116, January|2018.

[4] Sahil Sharma[1], Vinod Sharma[2], Atul Sharma[3], "Performance-Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis", Department of Computer Science & IT University Of Jammu Jammu, India.

[5] K.R.Lakshmi[1], Y.Nagesh[2,] and M.VeeraKrishna[3], "Performance Comparison Of Three Data Mining Techniques For Predicting Kidney Dialysis Survivability", [1]Director, IERDS, Maddur Nagar, Kurnool, Andhra Pradesh, India [2]Dept. of Computer Science, Assosa University, Ethiopia [3]Dept of Mathematics, Rayalaseema University, Kurnool, Andhra Pradesh, India, International Journal of Advances in Engineering & Technology, ISSN: 22311963, March|2014.

[6] V Jha, G Garcia-Garcia, K. Iseki, et.al. "Chronic kidney disease: global dimension and perspectives "Lancet. , pp. 260-272, 2013.

[7] S.Dilli Arasu, Dr. R.Thirumalaiselvi, "Review of Chronic Kidney Disease based on Data Mining Techniques", [1]Bharath Institute of Higher Education and Research (BIHER), [2]Government Arts College for Men (Autonomous)-Nandanam, Chennai -600 035, Tamil Nadu, India, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13498-13505.

[8] Sirage Zeynu, Shruti Patil, "Survey on Prediction of Chronic Kidney Disease Using Data Mining Classification Techniques and Feature Selection", Department of Computer Science & Engineering Symbiosis Institution of Technology Pune, India, International Journal of Pure and Applied Mathematics Volume 118 No. 8 2018, 149-156 ISSN: 1311-8080; ISSN: 1314-3395.

[9] AN Ramesh[1], C Kambhampati[2], JRT Monson[1], PJ Drew[1], "Artificial intelligence in medicine", [1]The University of Hull Academic Surgical Unit, Castle Hill Hospital, Cottingham, UK [2]Department of Computer Science, University of Hull, UK, pp. 334–338, 2004.

[10] Vikas Chaurasia[1], Saurabh Pal[1,*], B.B. Tiwari[2], "Chronic Kidney Disease: A Predictive model using Decision Tree", 1Dept of MCA, VBS Purvanchal University, Jaunpur, UP, India 2Dept. of Electronics, VBS Purvanchal University, Jaunpur, UP, India, International Journal of Engineering Research and Technology. ISSN 0974-3154 Volume 11, Number 11 (2018), pp. 1781-1794.

