# Effective Use of Naïve Bayes, Decision Tree, and Random Forest Techniques for Analysis of Chronic Kidney Disease

**Rajesh S. Walse, Gajanan D. Kurundkar, Santosh D. Khamitkar, Aniket A. Muley, Parag U. Bhalchandra, and Sakharam N. Lokhande**

**Abstract** The researcher is using a classification method for chronic kidney patient analysis of data. Chronic kidney disease data contains 25 attributes and 400 instances. Now, we proposed a best model by applying the decision support system of naïve Bayes, decision tree J48 algorithm, and random forest classifier techniques, and really, this proposed model will be helpful to predict the further CKD as well as not CKD patients on the basis of different parameters. During analysis, the naïve Bayes classifier correctly classifies instances with the 97.50% accuracy, decision tree J48 algorithm finds the correctly classified instances with the 98.33% accuracy, and similarly, random forest is analyzed and giving output with 100% accuracy and 0% of incorrectly classified instances. Therefore, the random forest decision tree classifier algorithm is the best and produces the most accurate and correct results with the 70 percent of the split value (train on a portion of the data and test on the remainder), and the value of ROC area is 1%. The main objective of the research

R. S. Walse (✉) · S. D. Khamitkar · P. U. Bhalchandra · S. N. Lokhande
School of Computational Sciences, Swami Ramanand Teerth Marathwada University, Nanded, India
e-mail: rajeshwalse@gmail.com

S. D. Khamitkar
e-mail: s.khamitkar@gmail.com

P. U. Bhalchandra
e-mail: srtmun.parag@gmail.com

S. N. Lokhande
e-mail: lokhande_sana@rediffmail.com

G. D. Kurundkar
Department of Computer Science, Shri Guru Buddhi Swami Mahavidyalaya, Purna Dist. Parbhani, S.R.T.M. University, Nanded, India
e-mail: gajanan.kurundkar@gmail.com

A. A. Muley
School of Mathematical Sciences, Swami Ramanand Teerth Marathwada University, Nanded, India
e-mail: aniket.muley@gmail.com





T. Sch[illegible] et al. (eds.), *Information and [illegible] Technology for Intelligent Systems*, Smart Innovation, Systems and Technologies 198
https://doi.org/10.1007/978-981-15-7078-0_23

paper is comparative study of NB classifier, DT J48, and RF to analyze chronic kidney disease (CKD) patient's data and to predict how many patients are having CKD. An analysis of how many patients currently have kidney disease and how many people may have this disease in the future has been attributed to it. When analyzing the same algorithm, the decision tree J48 shows that the tree variant can be 100% diagnosed with kidney disease in the future, and the random forest algorithm has analyzed it 100%.

## 1 Introduction

Now, the computer science techniques like optimized association rule mining techniques are used for improved genetic algorithms data mining and machine learning which are used to study the power of various parameters and make predictions of the based on different datasets. Data mining techniques are the process of identifying the hidden patterns from the big and tedious data. This may provide a vital role in the decision making for large data, not only agriculture but also health-related problems. Bharara et al. [1] reviewed to extract for business operations using data mining techniques. Ariff et al. [2] studied RFID based systematic livestock health management system. Jinyin et al. [3] performed a novel cluster center which is the fast determination clustering algorithm. Arasu and Thirumalaiselvi [4] dealt with novel imputation techniques for the effective type of predictions of kidney disease patients. Chuan et al. [5] performed an applied study of Guangdong provincial hospital of traditional China treatment. Guangzhou explored clustering analysis for syndrome evolution peritoneal dialysis patients. Kunwar et al. [6] studied and analyzed chronic in terms of permanent kidney disease harnessing of data mining for classification techniques. Duc Luong [7] applied K-means approach to clustering disease progressions. Güllüoğlu [8] used data mining techniques for segmenting customers' information. Kumar and Lhatri [9] used WEKA which is used for medical-related data classification and to find early disease prediction. Kumar and Lhatri [9], NCBI [10] performed a study on the economics of dialysis in India. J Nephrol [11] studied the occurrence of chronic kidney disease in India, and where are we heading? Ubolthaın et al. [12] performed a diagnostic study of acute kidney injury using the KDIGO guideline approach. In this paper, experiment has been carried out on chronic kidney disease patient based on their relationship attributes, and nowadays, chronic kidney disease patient in India is increasing day by day because of their eating habit and other health issues. Khanna [10] observed that since last ten years number of CKD patients increased tremendously in concerned with this issue there is need of research which will be helpful to the doctors or medical industry. It will be beneficial to make prediction of CKD and not CKD patient based on their other health parameters. Also, to minimize the growth rate of CKD patients and to control further damages of their kidney. Data mining plays an active role in predicting future kidney-related health problems. In this research paper, three algorithms have been analyzed one is NB classifier, J48, and random forest decision tree. Data cleaning by KNN is used to removal of noise and

inconsistent data with data integration technique with the combination of multiple types of data. To evaluate the data, we have used secondary data, and it is retrieved from UCI machine learning repository [13]. J Nephrol [11] with increasing life period and the frequency of lifestyle disease, the US has seen a 30% considerable growth in the widespread presence of CKD in the last decade. Unfortunately, from India, there is no longitudinal study and limited data on the incidence of CKD. At present, the living standard of the people and the daily consumption of food are adversely affecting their health, especially their everyday living, which is increasing the number of kidney diseases in India every day. His anatomy also depended on the diet of people 40 years ago or older, but today, kidney disease is not only limited to people with diabetes or hypertension, but it has many causes. Chemical cereals, vegetables, and fruits are the result of all these things, The loss of kidney function increased slowly due to the daily in appropriate chemical mixed food and further, it may causes to the damage of kidney failure. According to reference of J Nephrol [11], unfortunately, from India, there is no longitudinal study of CKD and limited data. So because of all of the above, we have tried to analyze acute kidney disease by using naïve Bayes, decision tree J48, and random forest algorithm unprocessed learning technique. Indeed, the purpose of our research is to use our research to analyze kidney disease or whether it can cause kidney disease in the future.

## 2 Methodology

See Fig. 1.

## 3 Result and Discussion

We are using a chronic kidney failure disease dataset, in dataset training database perfection for the NB, J48, and RF decision tree, and select some parameters such as 1. RBC count, 2. hypertension (BP), 3. diabetes M., 4. coronary disease, 5. appetite, 6. pedal edema, and 7. anemia. We are using WEKA tool for classifying data using algorithms decision tree and naïve Bayes classifier.

In this study, secondary data of 400 observatins has been extracted for anaysis purpose [13]. The data was obtained after cleaning and removing missing values for further analysis, the data contains 25 attributes in the dataset with class (CKD and Not-CKD), and class distribution is 63% for CKD and 37% for not CKD. Performance of NB, decision tree, and RF: naïve Bayes: The performance criteria values. Accuracy refers: summary shows the correctly classified instances 117, and its accuracy near about 97.5%, incorrectly classified instances is 3, and its accuracy is 2.5%. J48 decision tree: The performance criteria values. Accuracy refers: summary shows the correctly classified instances 118, and its accuracy near about 98.33%, incorrectly classified instances is 2, and its accuracy is 2%. Random forest: The performance

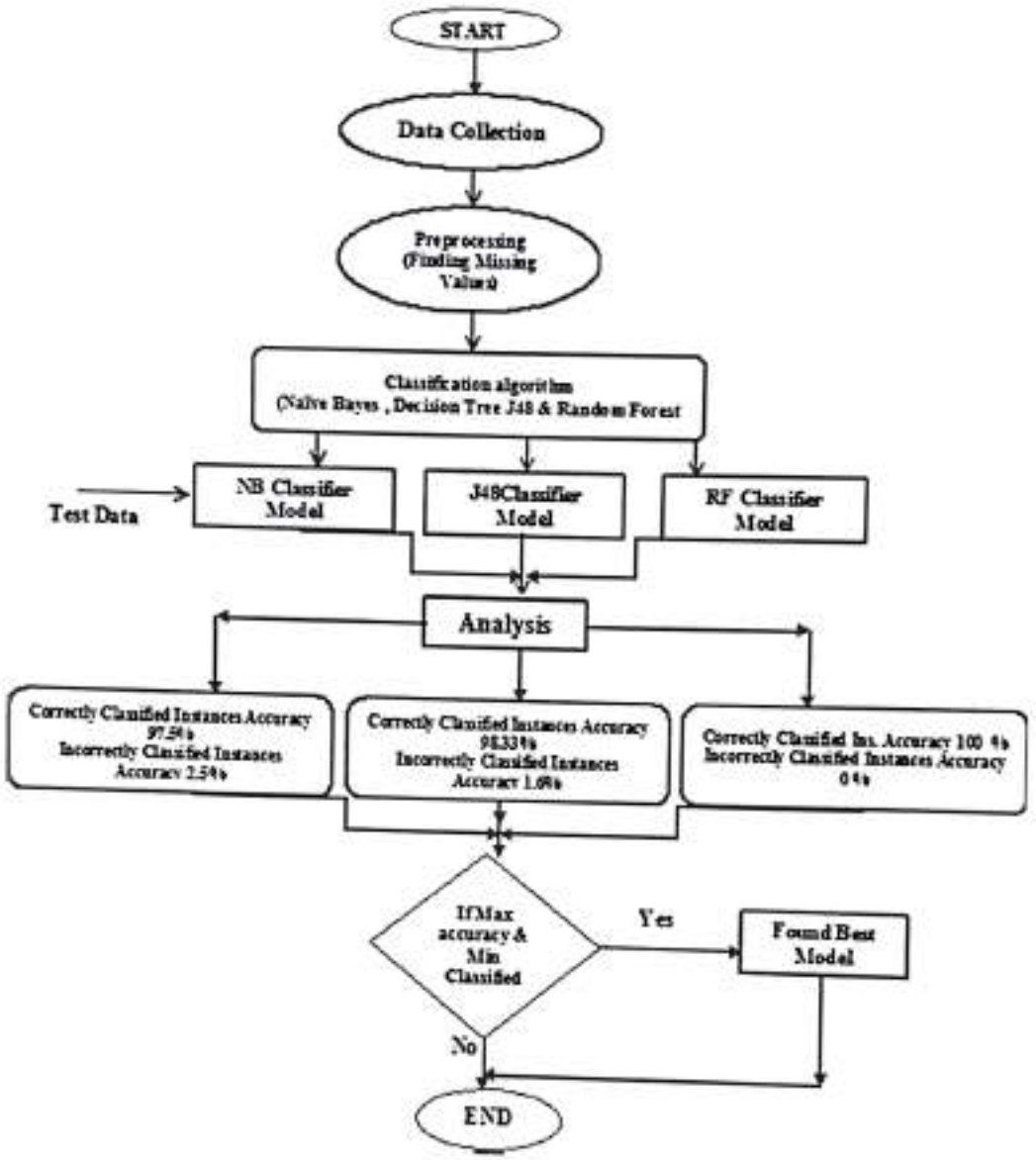


Fig. 1 Adopted methodology flowchart

criteria values. Accuracy refers: summary shows the correctly classified instances 120, and its accuracy near about 100%, incorrectly classified instances is 0, and its accuracy is 0%.

The result of experiment to be compared of naïve Bayes and decision tree (J48) with random forest is established on the basis of performance in terms of high accuracy with a minimum period processing. The following algorithm is to analyze through data; the results and analysis of all three algorithms are as follows (Fig. 2).

In Figure 3, decision tree shows the exact prediction of chronic kidney patients based on attributes relations, and here, decision tree J48 algorithm shows the prediction on the basis of hemoglobin and its basic laboratory values, i.e., if hemoglobin level is ≤12.9, then respective attributes like serum creatinine level is ≤1.1, and

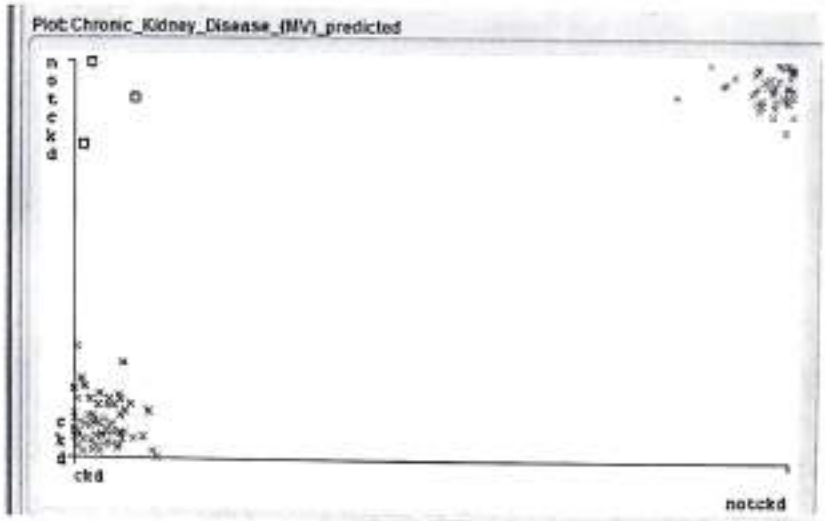


Fig. 2 Naïve Bayes: Visualize classifier error: Class: CKD and not CKD

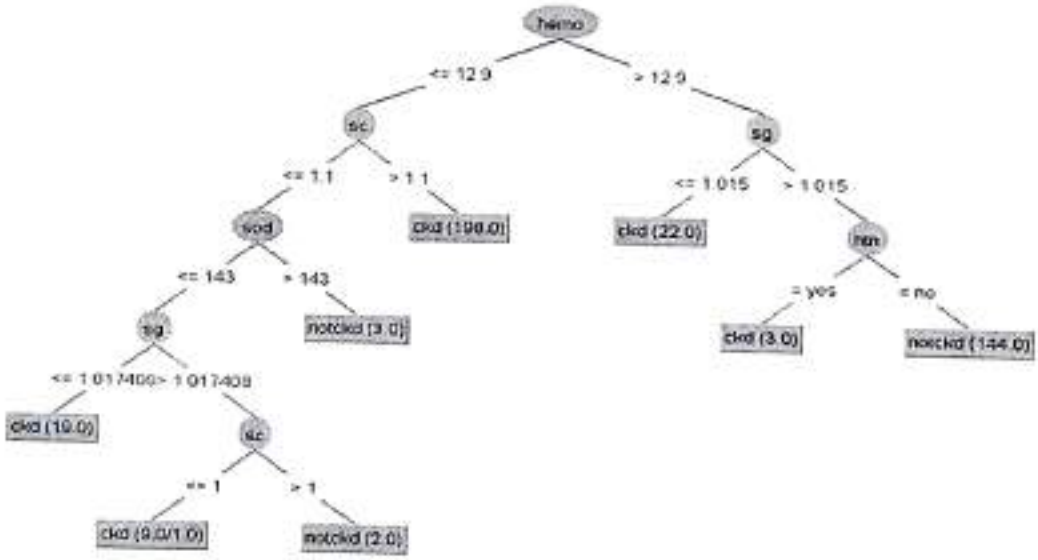


Fig. 3 Performance accuracy by class and confusion matrix by decision tree

then, again respective condition of sodium level and sugar level checks shows that the predictive result particular patient is having CKD or not and also shows that the particular patient is having CKD or not CKD on the basis of their hypertension. This will allow kidney patients who are currently in a state to know what caused the kidney disease, and those who have not had a kidney disease will see if they can develop kidney disease in the future, so they will not need to perform additional tests and save money (Fig. 4).

Figure 4, cleary shows that, there are 2 incorrectly classified instances and it is 1.66 %. It has been observed that one error is contained in both the classes. The research

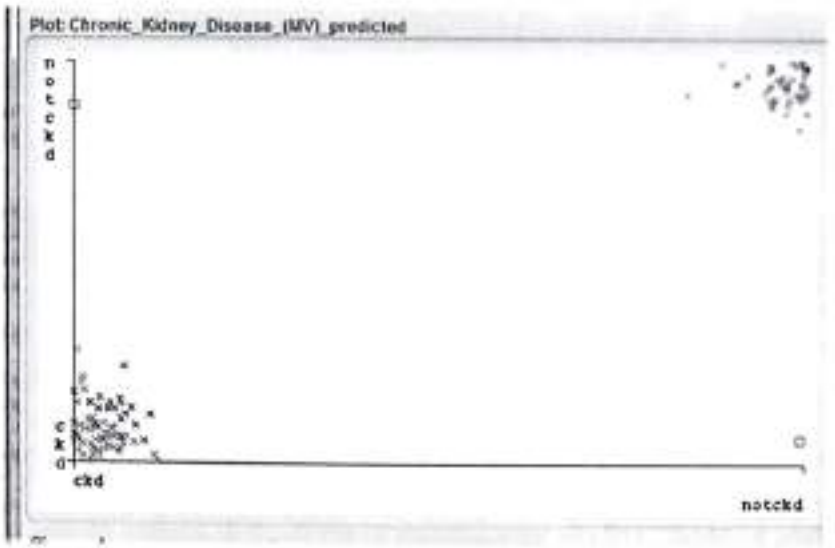


**Fig. 4** Decision tree J48 algorithm to visualize classifier error

aim is to minimize the classified error for correct and maximize the prediction accuracy for further treatment to the patients. Random forest: classifier (misclassification ROC value is 1%); therefore, the random forest is the best classifier algorithm as comparative naïve Bayes and decision tree J48 classifier algorithm, and therefore, these are designed novel model or technique for a classifier for the best prediction. The Random Forest model gives 100 % accuracy for full data set, splitted data set in terms of train set (70 %) data set. Two classes are correctly classified and its performance is measured through Kappa statistics and receiver operating characteristic curve value 1.

Figure 5 displays result 0 incorrectly classified instances, i.e., 0%, the graphical result shows the zero classified errors in both class a and b. The aim of the research

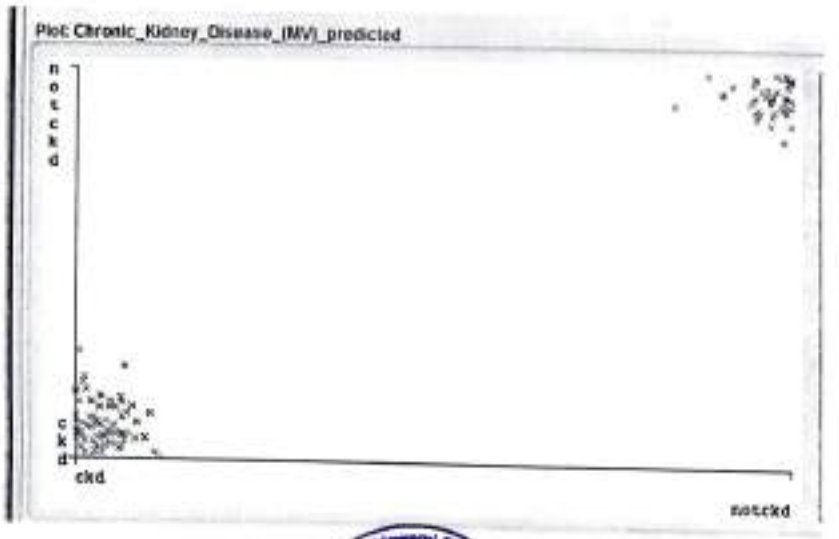


**Fig. 5** Random forest visualize classifier error CKD and not CKD Class

is to minimize the classified error for correct and maximize the prediction accuracy for further treatment to the patients. It will show in the confusion matrix

$$a\ b <== \text{classified}, 76\,0|\ a = \text{CKD}, 0\,44|\ b = \text{not CKD}$$

The research, the kidney dataset, processed with different attributes (25), which contains 400 rows, i.e., instances and 25 attributes, means columns. The researcher has select every attributes to display type of attributes, the type means nominal, how many missing values present in the dataset for each attribute viz instances, how many distinct values are present in the dataset, distinct means different values, if we select attribute is shown nom- in front of attribute-nom means nominal type. The numeric data gives summary of the overall data in the form of descriptive statistics. Also, if data is qualitative, then it treated as an attribute class, and it shows in the form of label count and its weight in the form of true/false or yes/no (Tables 1 and 2).

It observed that among the classification of CKD and not CKD patients, the precision level of the RF algorithm is best to as compared to naïve based algorithm and decision tree J48 algorithm. Also, the receiver operating characteristics curve (ROC curve) area gives of classifying CKD and not CKD patients obtained by random forest, i.e., 100% accuracy. Further, it is observed that it is more precise compared to naïve based and decision tree J48 algorithm.

**Table 1** Correctness by class values

| | Class | TP rate | FP rate | Precision | Recall | *F*-measure | MCC | ROC area | PRC area |
|---|---|---|---|---|---|---|---|---|---|
| Naïve based | CKD | 0.961 | 0.000 | 1.000 | 0.961 | 0.980 | 0.948 | 0.999 | 1.000 |
| | Not CKD | 1.000 | 0.039 | 0.936 | 1.000 | 0.967 | 0.948 | 0.999 | 0.998 |
| | Weight avg. | 0.975 | 0.014 | 0.977 | 0.975 | 0.975 | 0.948 | 0.999 | 0.999 |
| Decision tree | CKD | 0.987 | 0.023 | 0.987 | 0.987 | 0.987 | 0.964 | 0.991 | 0.993 |
| | Not CKD | 0.977 | 0.013 | 0.977 | 0.977 | 0.977 | 0.964 | 0.991 | 0.972 |
| | Weight avg. | 0.983 | 0.019 | 0.983 | 0.983 | 0.983 | 0.964 | 0.991 | 0.985 |
| Random forest | CKD | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | Not CKD | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | Weight avg. | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

**Table 2** Summary of classifier model (Train set data)

| Sr. no. | Particulars | Naïve Bayes | Decision tree (J48) | Random forest tree |
|---|---|---|---|---|
| 1 | Test mode: 70% train, 30% test | 70.0% train | 70.0% train | 70.0% train |
| 2 | To build model | 0.02 s | 0.05 s | 0.19 s |
| 3 | Test model on test split | 0.02 s. | 0.02 s. | 0.01 s. |
| 4 | Correctly classified instances | 117 | 118 | 120 |
| 5 | Incorrectly classified instances | 3 | 2 | 0 |
| 6 | Kappa statistic | 0.9469 | 0.9641 | 1 |
| 7 | Mean absolute error | 0.0251 | 0.0238 | 0.0485 |
| 8 | Root mean squared error | 0.1544 | 0.1268 | 0.1061 |
| 9 | Relative absolute error | 5.3604% | 5.0892% | 10.3666% |
| 10 | Root relative squared error | 32.0255% | 26.298% | 22.009% |
| 11 | Total number of instances | 120 | 120 | 120 |

## 4 Conclusion

The ckd data is analyzed and predicted for diagnosed patients using data mining classifiers algorithm of naive Bayes, decision tree J48, and random forest algorithms. The performance of these algorithms is compared using Weka tools. The final obtained result shows that the naïve Bayes is the best truthful classifier with 100% accuracy, i.e., correctly classified instances as compared with decision tree J48 having 98.33% accuracy and naïve Bayes algorithm having 97.5% of efficiency. For research work, some of the attributes were measured RBC count, HP, diabetes mellitus, CAD, appetite, pedal edema, anemia, etc. Now future, this kind of research will be helpful to the doctors or medical industry for prediction of CKD and not CKD patient based on their other health parameters, to minimize the growth rate of CKD patients and to control further damages of the kidney. Data mining plays an active role in predicting future kidney-related health problems. In this paper, these algorithms have been analyzed. We have tried to analyze acute kidney disease by using naïve Bayes, decision tree J48, and random forest algorithm unprocessed learning technique. Indeed, the purpose of our research is to use our research to analyze kidney disease, or whether it can cause kidney disease in the future, this will allow kidney patients who are currently in a state to know what caused the kidney disease, and those who have not had a kidney disease will see if they can develop kidney disease in the future, so they will not need to perform additional tests and save money.

## 5 Acknowledgement

Authors are grateful to the UCI ML repository is provided that all types of necessary database WEKA for providing such a reliable tool to extract and analyze knowledge from the database.

## References

1. Bharara, S., Sai Sabitha, A., Bansal, A.: A review on knowledge extraction for business operations using data mining. In: 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, pp. 512–518. IEEE (2017)
2. Ariff, M.H., Ismarani, I., Shamsuddin, N.: RFID based systematic livestock health management system. In: 2014 IEEE Conference on Systems, Process and Control (ICSPC 2014), pp. 111–116. IEEE (2014)
3. Jinyin, C., et al.: A novel cluster center fast determination clustering algorithm. Appl. Soft Comput. **57**: 539–555 (2017)
4. Arasu, S.D., Thirumalaiselvi, R.: A novel imputation method for effective prediction of coronary Kidney disease. In: 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), pp. 127–136. IEEE (2017)
5. Chuan, Z., Ying, T., Li, B., Yuqun, Z., Fuhua, L.: Application of clustering analysis to explore syndrome evolution law of peritoneal dialysis patients. In: 2013 IEEE International Conference on Bioinformatics and Biomedicine, pp. 23–26. IEEE (2013)
6. Kunwar, V., et al.: Chronic Kidney Disease analysis using data mining classification techniques. In: 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence). IEEE (2016)
7. Duc Luong, T.A.: K-Means Approach to Clustering Disease Progressions IEEE Keywords. Department of the Computer Science & Engineering, University at Buffalo, Buffalo, NY, USA (2017)
8. Güllüoğlu, S.S.: Segmenting customers with data mining techniques. In: 2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC), pp. 154–159. IEEE (2015)
9. Kumar, N., Lhatri, S.: Implementing WEKA for medical data classification and early disease prediction 978-1-50. In: 3rd IEEE International Conference on Computational Intelligence and Communication Technology (IEEE-CICT 2017). Department of Computer (2017)
10. Khanna, U.: The economics of dialysis in India. Ind. J. Nephrol. **19**(1), 1 (2009)
11. Varma, P.P.: Prevalence of chronic kidney disease in India—Where are we heading? Ind. J. Nephrol. **25**(3), 133 (2015)
12. Uboltham, I., Prompoon, N., Pan-Ngum, W.: AKIHelper: acute kidney injury diagnostic tool using KDIGO guideline approach. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1–6. IEEE (2016)
13. https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

Co-ordinator
IQAC
Shri Guru Buddhiswami Mahavidyalaya
Purna (Jn) Dist. Parbhani - 431511 (M.S.)

Shri Guru Buddhiswami Shikshan Prasarak Sanstha's Shri Guru Buddhiswami Mahavidyalaya Purna (Jn.) Estd.1983

PRINCIPAL
Shri Guru Buddhiswami Mahavidyalaya
Purna (Jn.) Dist.Parbhani