Research Article

Diagnosing Chronic Kidney Disease Using Hybrid Machine Learning Techniques

Janani J [a], Sathyaraj R[b]
[a]Student,School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu-14
[b]Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu-14

Abstract: The chronic kidney failure is a serious health issue and if not detected and treated at the early stages, it can be very deadly. Hence the major objective of this paper is to develop a reliable machine learning model which predicts the CKD with a high accuracy rate. The CKD data set is downloaded from the famous UCI ML repository but it suffers from a lot of missing values. To handle the missing values KNN Imputation is used. Feature selection is also performed with the help of information gain as the dataset is huge and hence the cost of modelling can be very high. Various other pre-processing steps like label encoding and Min-max normalization is performed to attain a clean dataset. After pre-processing, various ML algorithms like logistic regression, naïve bayes, artificial neural network and random forest are applied and their performances are compared with the help of various performance metrics. A hybrid of Random Forest and Adaboost algorithm is proposed and it achieves a better accuracy when compared to the other individual component models and hence it can be proved that the proposed hybrid model is much better and accurate in diagnosing CKD.

Keywords: Chronic Kidney Disease (CKD), Machine Learning (ML), Integrated/Hybrid Model

1. Introduction

Chronic kidney disease, also known as the chronic kidney failure, describes the gradual loss of kidney function. The most important role of the kidneys is to filter wastes and excess fluids from the blood. CKD means that the kidney does not work as expected and cannot properly filter blood. Dangerous levels of fluid, electrolytes and wastes can build up in one's body when chronic kidney disease reaches the final stages and this can be very dangerous. Chronic kidney disease occurs when a disease or condition impairs kidney function, leading the kidney damage to worsen over several months or years. In the beginning few stages of chronic kidney disease, one may have only a few signs or symptoms. Chronic kidney disease may not become prominent until the kidney function is significantly impaired.Chronic kidney disease might progress to end-stage kidney failure, which is fatal without artificial filtering (dialysis) or a kidney transplant.There are five stages of CKD and the most dangerous one is stage 5 because, at this stage, the kidneys are unable to do most of their functions. The stages of kidney disease are dependent on how good the kidneys can filter waste and extra fluid out of the blood. The kidneys will still be able to filter out waste from the blood in the early stages of kidney disease. In the later stages, the kidneys must work harder to get rid of waste and may stop functioning altogether. It is difficult to find out the CKD stage of each patient especially at the early stages. Glomerular Filtration Rate (GFR) is the best test to measure the level of kidney function and to determine the stage of chronic kidney disease of a person. It can be calculated from the results of the patient's blood creatinine, age, race, gender, and other such values. CKD has affected the entire world but mainly for the countries that have a low or medium income, it has been very disastrous. About 10% of the population worldwide suffers from the chronic kidney disease, and the number of deaths every year is increasing severely. In the past two decades as per the studies and researches performed by various health organizations, CKD is said to have caused a lot of death and other severe health conditions. The number of people who suffer from end stage renal disease is also becoming high and this considered as the last stage of CKD and for the patients to survive this, kidney transplantation or regular dialysis have to be performed.

2.Significance Of The Study

In India, there is a huge number of the chronic kidney failure cases reported every year which is really very alarming. CKD can gradually destroy the entire function of kidney to filter wastes. If this disease is not diagnosed and treated at the early stages, one might develop a permanent kidney damage. If this disease keeps progressing, dangerous electrolytes can be gathered in high levels in a person's blood and cause the person to fall sick. Almost all the parts of a human body can be damaged by the progression of CKD and this can be a huge risk. Due to CKD, a person might suffer from the problem of other diseases which are equally dangerous like hypertension, anaemia etc. These complications might happen very slowly that a person will not be able to diagnose that he has the chronic kidney disease. Hence early detection of the chronic kidney disease is very important as it prevents the disease from getting worse which can be very deadly. CKD does not show any symptoms or any disease related symptoms in the initial few stages and without taking the test it is not possible to tell is a particular person has the CKD or not. If detected in the early stages, it will be very helpful for the patients as they will be able to get a timely treatment and hence the progression of the disease to the further stages can be stopped. A person is more likely to get the chronic kidney disease if he has a family history of CKD or if he has hypertension or diabetes. If

people get to know about this disease earlier then the treatment process and the recovery process will be much easier. It is hoped that people get to know about this disease sooner with the minimum number of tests possible and also at a low cost. Hence the major motivation behind developing a machine learning methodology to diagnose the chronic kidney disease is the potential risks caused by the progression of CKD to ESRD if not detected at the early stages and also the depressingly increasing number of cases reported every year.

3.Review Of Related Studies

In the past few years there is a lot of research done on detecting chronic kidney disease with the aid of different kinds of machine learning techniques. Many machine learning algorithms like logistic regression, random forest, support vector machine, k-nearest neighbour, naive Bayes classifier and various neural networks were studied and their performance was compared with the help of various performance metrics and loss which was used for the neural network alone. The dataset for the researches were obtained from various sources like the UCI machine learning repository, King Fahd University Hospital(KFUH) in Khobar etc. In many studies, the main aim was to detect CKD with the least number of predictions and several statistical tests were performed to discard the unwanted attributes. Various feature selection techniques were used as they could be very helpful in reducing the costs. Some of the feature selection techniques used were Correlation-based Feature Selection (CFS), fruit fly optimization algorithm (FFOA), Density based Feature Selection (DFS) and Relief algorithm. Even a Heterogeneous Modified Artificial Neural Network (HMANN) was developed which used ultrasound images to perform various image processing steps with the help of machine learning to detect CKD. There were also a few studies on decision tress to diagnose CKD. In one such study it was proved that the J48 decision tree and random forest achieved highly accurate results when compared to the other kind of tress in machine learning that did not achieve the desirable result in detecting CKD. In 2018 Almarashi A, Alghamdi M, and Mechai I introduced a different kind of detecting technique for CKD by using Artificial Neural Network (ANN). The component parts of the newly developed system were one input layer, one hidden layer, and one output layer. There was also a research which proposed a unique methodology based on Extreme Gradient Boosting (XGBoost) model. In this model three different feature selection methods were utilized.

Among the various algorithms compared, ANN and Random forest achieved the best performance and SVM was the most widely studied algorithm for diagnosing CKD. Logistic Regression and Naïve Bayes algorithms are fairly new when it comes to diagnosing CKD. It can be seen that most of the studies mainly focus on the establishment of models and achieving a high accuracy but there is not enough research on the data pre-processing techniques used. A complete procedure of handling the missing values is not explained in depth. In most of the research papers the rows with the missing values were deleted which could lead to a loss of important data. In other papers the mean or the median methods were used which are not that desirable as they tend to add a lot of unwanted bias and variance to the dataset which could cause a huge problem while developing the machine learning models.Most of the existing work are about individual models of machine learning algorithms and there is not enough research on Adaboost algorithm and integrated models which can achieve a better accuracy.

4.Objectives Of The Study

The major objective of this paper is to diagnose CKD at the early stages with the least possible tests and cost and with a high accuracy rate.The paper also aims to effectively handle the missing values, present in the CKD data set with the help of KNN Imputation. Feature selection is also performed with the help of information gain to find the most important features that play a vital role in detecting CKD. Various machine learning algorithmsare applied and analysed to detect CKD and the best one with the best performance and accuracy rate is found.Adaboost algorithm is also applied as it boosts the performance of the weak classifiers. Finally, the misjudgements generated by the established models are analysed and an integrated model is proposed that combines random forest algorithm and AdaBoost algorithm which can achieve a better accuracy and can thus be an effective and reliable model to detect CKD.

5.Hypotheses Of The Study

The hybrid model will achieve a higher accuracy rate when compared to the individual machine learning models.The use of KNN Imputation to handle the missing values in the dataset and the use of information gain as the feature selection method will further increase the accuracy.

6.The Proposed Methodology

The proposed methodology involves the various steps as follows:

6.1.Data Pre-processing

Data pre-processing in Machine Learning is a very important to convert the raw dataset into a cleaned dataset set that can be desirable to apply variable machine learning algorithms.

6.1.1.Acquire the Dataset

The dataset for predicting the chronic kidney disease is attained from the UCI machine learning repository which is a well-known source for all the machine learning datasets. The CKD dataset has 400 patient records and 25 attributed which are either the symptoms or other attributes related to the disease like hypertension, blood pressure, specific gravity, albumin etc. Among this 400 patient records, 250 patients have the disease and the other 150 of them do not have the disease.

6.1.2.Import all the crucial libraries

In order to perform data pre-processing using Python, various predefined Python libraries must be imported. All the libraries have a certain task to do when it comes to machine learning. In this diagnosis, various libraries such as numpy, pandas, matplotlib, sklearn etc. have been imported and all these libraries have a certain task to perform.

6.1.3.Import the dataset

The dataset which has been collected for the machine learning project is imported. While doing the dataset importing process, one more important thing has to be done which involves extracting dependent and independent variables. In this dataset, classification is the dependent variable and all the other features are independent variables.

6.1.4.Feature Selection

A Machine Learning model can suffer from the issue of overfitting if the number of features become identical or huge. To prevent this from happening it is important to reduce the number of features in the dataset. Feature selectionrefers to reducing the number of features in a dataset when building a machine learning model. Another advantage of decreasing the number of input attributes is the decrease in the cost of building the ML model and in a few situations it might even improve the accuracy of the model. There are various ways to perform feature selection and in this paper Information Gain is used. Information gain is chosen over the other techniques as it comes under filter technique. The feature selection using the filter techniques choose the intrinsic properties of the features measured with the help of univariate statistics other than using the cross-validation performance. When compared to the other feature selection methods, filter methods are speedier and less computationally costlier than wrapper methods. While handling with a huge dataset with a lot of features, it is comparatively less costly and easier to make use of the filter methods. Information gain of every variable with respect to the target variable is found to determine the most important features that have a vital role in the prediction process. The information gain of a feature will be anything from zero to one. The features with the highest information gain will be retained as they are the most important ones and the features with the least information gain will be discarded. The Information gain of each input attribute or the feature dataset with respect to the output attribute ie. Classification is given below:

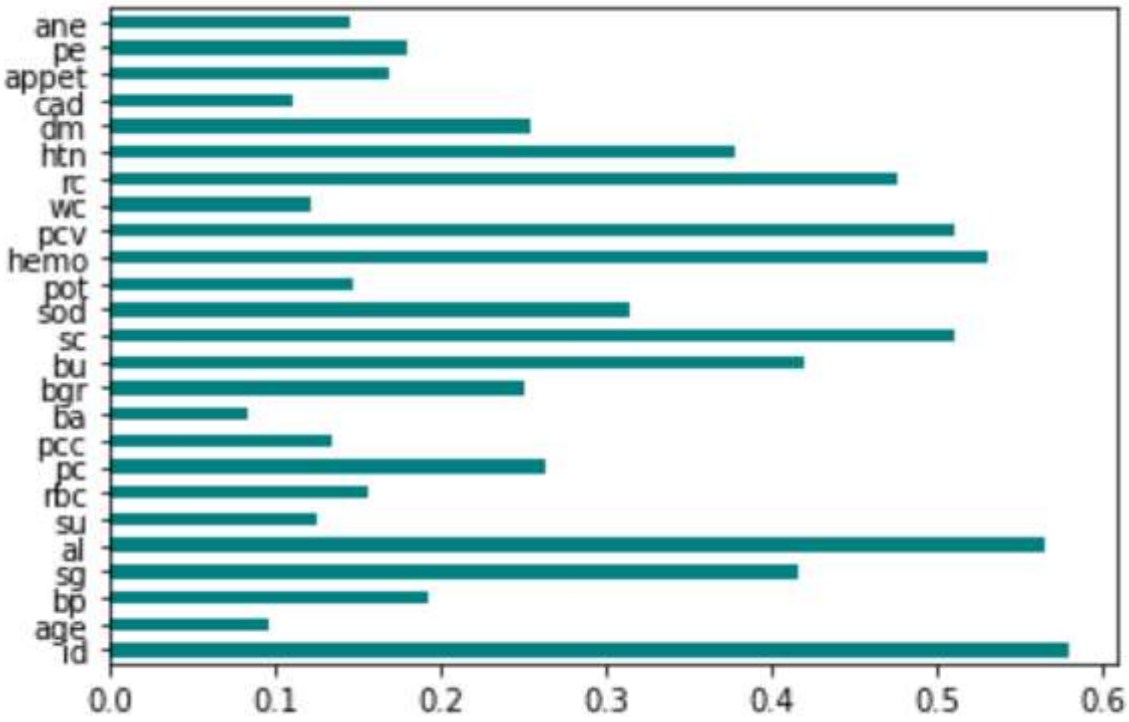

Figure 1. Information Gain of all features with respect to target feature

6.1.5.Identify and manage missing values

Next it is very essential to manage the missing values in a dataset as it will cause a huge issue later while applying the machine learning algorithms. Hence it is essential to handle missing values that are there in the dataset. The CKD dataset has various missing values as some of the patients might forget or miss to fill in certain values in the real life. Hence, KNN imputation is going to be used to handle missing values as it has proven to be effective in experiments. Every missing value can then be handled by replacing them with the mean value of the k nearest neighbours, to do so in general Euclidean distance metric is used in default.

6.1.6.Encoding the categorical data

Categorical data in a dataset are the ones that have certain categories like in this dataset there are categorical variables like packed cell volume, hypertension and classification. These categorical data might cause a huge challenge while building the model as machine learning only deals with numbers and mathematics. So it is essential to convert these categorical variables into numbers. There are various techniques for this purpose and in this paper Label Encoding of sklearn library is going to be used.

6.1.7.Feature Scaling

In feature scaling all the input attributes are converted to value of common range or distribution so that it becomes easier to compare all the variables on common grounds to build a reliable model. In this dataset, it can be noticed that the albumin and packed cell volume columns do not have the same range or distribution of values. Packed cell volume has a higher range when compared to the albumin's range and hence a proper result will not be achieved as PCV dominates the albumin. Hence it becomes very important to perform feature scaling to the dataset. In this paper Min-Max Normalization is going to be used to perform feature scaling. This technique converts a feature or observation value with distribution value in the range 0 to1.

6.1.8.Splitting the dataset

All the dataset must be split into the training subset and the testing subset to proceed with the prediction. The training subset is used to perform training whereas the testing subset is used to perform testing. One is aware of the results in the training subset but one is not aware of the prediction result or output of the testing subset. The CKD dataset is going to be split in the ratio 70:30.

6.2.Developing individual ML models and evaluating them

The various machine learning algorithms applied for prediction are:

- Logistic Regression
- Naïve Bayes
- Artificial Neural Network
- Random Forest

The performance of these classifiers are analysed based on various metrics such as Accuracy, Precision, Recall, F-measure and loss and the best classifier for diagnosing chronic kidney disease is found.

6.3.Establishing the Integrated/Hybrid Model

Once the individual ML algorithms are analysed and compared for the misjudgements, a hybrid model is developed to further increase the accuracy. A hybrid of Random Forest and Adaboost algorithm will be proposed. Random Forest is chosen over ANN as it acquires a high accuracy and it is also very compatible with AdaBoost algorithm. ANN also suffers from the problem of unexplained behaviour of network so it is not used for the hybrid model. Boosting is the process of improving the performance of weak classifiers with the help of an ensemble technique. This is done by creating a new model that keeps solving the problems and issues of the previous few models. Until a desirable result is attained, this process of creating a new model keeps continuing to create a highly accurate model. The final equation for hybrid model is given below where$f_m$is the weak classifier and $\theta_m$is the weight that corresponds to every classifier.

$$F(x)= \text{sign}(\sum_{m=1}^{M} \theta m f m(x))$$

Usually AdaBoost uses Decision Tree Classifier as default weak learner for training purpose but any type of machine learning algorithms can be utilised as long as it accepts the parameter known as weights. In this research Random Forest is used as the base classifier as it is much better than decision tree. So the f_m in the formula of the integrated model will be the random forest algorithm. The hybrid model will diagnose the CKD more effectively and will achieve a higher accuracy when compared to the individual models.

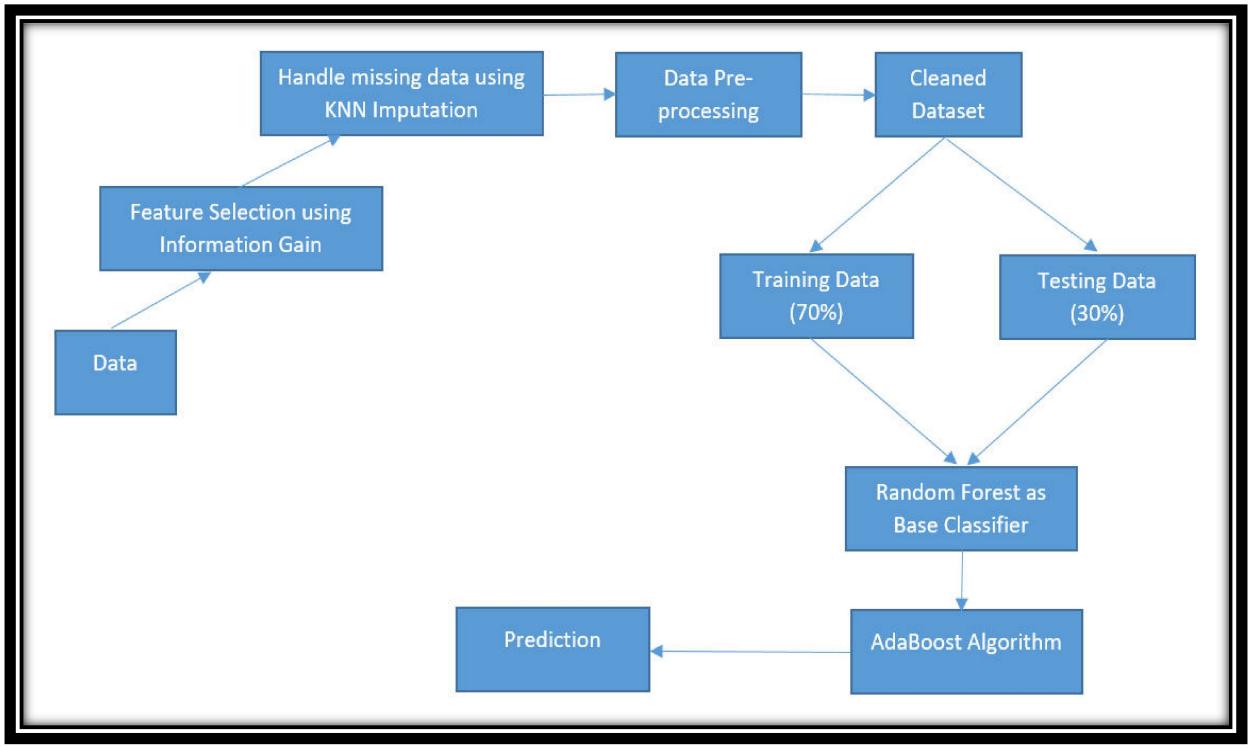

Figure 2. Architecture diagram of the hybrid model

7.Results and Discussion

The result of all the machine learning algorithms were compared and analysed with the help of various performance metrics. The experiments were conducted using Python 3.6.7 programming language and it can be done either using the Jupyter Notebook web application or Google collab. Many libraries were used for the implementation and one such important library is Sciket-learn, which is a very useful library to develop ML models. Various performance metrics in the confusion matrix are considered in this research. The experimental results of all the developed models is given in the table given below.

Table 1. Performance table

Classifiers	Accuracy	Precision	Recall	F-Measure	Loss
Logistic Regression	0.97701	0.95349	1.0	0.97619	
Naïve Bayes	0.94253	0.89130	1.0	0.94253	
Artificial Neural Network	1.0	1.0	1.0	1.0	0.01262
Random Forest	0.98851	0.97619	1.0	0.98795	
Integrated model	1.0	1.0	1.0	1.0	

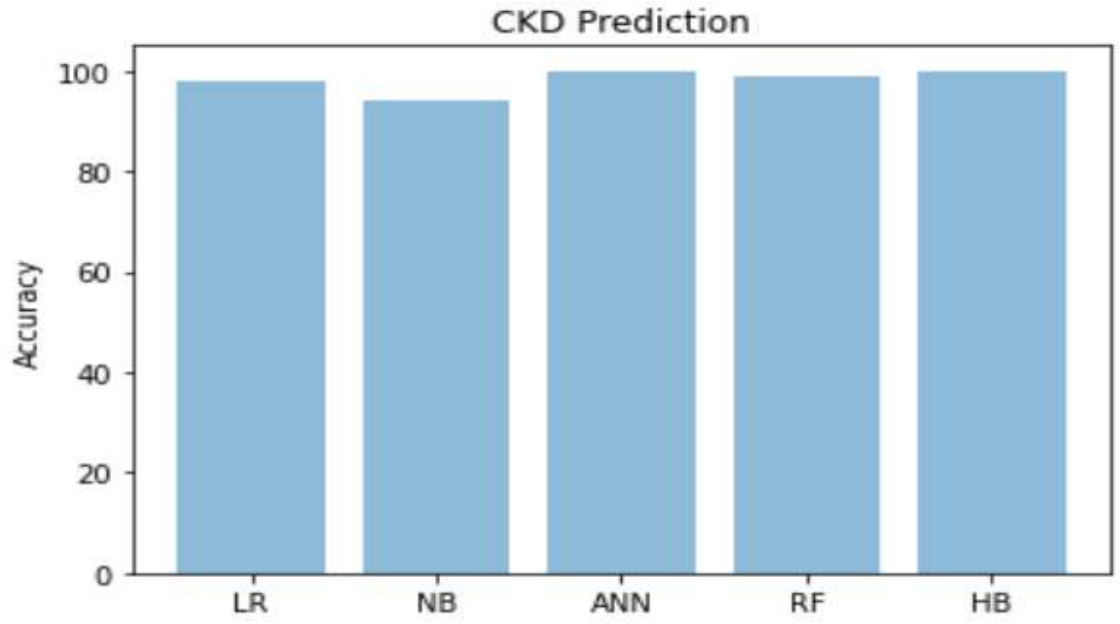

Figure 3. Accuracy graph of the models

The results prove the feasibility of the proposed methodology. The feature selection technique of Information Gain was used to extract the most important features which have a vital role in the diagnosis of the chronic kidney disease. It was found that specific features like specific gravity, albumin, serum creatinine, haemoglobin, packed cell volume, red blood cell count and hypertension were more important than the other features as they had a higher information Gain. With the help of KNN Imputation, LOG, NB, ANN and RF were able to attain a better performance than the cases when random imputation, deleting rows with missing values or mean and mode

imputation were used. These methods are not preferable as they add a bias and variance to the model. KNN Imputation handled the missing values by replacing them with the mean of k nearest neighbours. In the experiment imputer optimization was performed and it was found that 5 was most optimum value for k for the CKD dataset. From the evaluation results, all the models have a great performance against detecting CKD with a good accuracy. From the table it can be seen that Random forest and Artificial neural network achieved the best performance when compared to the other models. To further increase the efficiency and reliability of the individual ML algorithms, an integrated model is proposed which combines Random Forest and AdaBoost algorithm and it achieved a higher accuracy. Random Forest is chosen over ANN as it more compatible with AdaBoost algorithm and ANN also suffers from the problem of unexplained behaviour of the network. Whenever ANN produces a probing solution, it does not explain the solution as in why and how they are produced. This reduces the belief in the ANN. Adaboost is generally used to improve the performance of the decision tree algorithm and if Random Forest is used instead of decision tree it further improves the performance.

8.Conclusion

This research paper has examined the ability to detect CKD using machine learning methodologies. The aim was successfully achieved by applying and analysing various ML algorithms like logistic regression, random forest, naive Bayes classifier, and artificial neural network and the performance of these algorithms were compared. KNN imputation was used to handle the missing values and Information Gain was used as the feature selection method. By analyzing the problems and shortcomings of the individual ML models, a hybrid model was proposed that combines the AdaBoost and Random Forest algorithms which was able to achieve a better accuracy rate. Hence it can be speculated that this hybrid methodology can be used in the practical diagnosis of CKD and it can achieve a desirable effect. It can also be noted that this methodology might be useful to the clinical data of the other diseases in actual medical diagnosis. In the future, a larger dataset can be used by attaining a more number of patient records from various hospitals and health organisations to improve the accuracy of the prediction. It is hoped that the efficiency of the system will be more and more accurate with an increase in the size and quality of the dataset.

References

Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C. and Chen, B., 2019. A machine learning methodology for diagnosing chronic kidney disease. IEEE Access, 8, pp.20991-21002.

Almasoud, M. and Ward, T.E., 2019. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. International Journal of Soft Computing and Its Applications, 10(8).

Ahmad, M., Tundjungsari, V., Widianti, D., Amalia, P. and Rachmawati, U.A., 2017, November. Diagnostic decision support system of chronic kidney disease using support vector machine. In 2017 second international conference on informatics and computing (ICIC) (pp. 1-4). IEEE.

Al Imran, A., Amin, M.N. and Johora, F.T., 2018, December. Classification of chronic kidney disease using logistic regression, feedforward neural network and wide & deep learning. In 2018 International Conference on Innovation in Engineering and Technology (ICIET) (pp. 1-6). IEEE.

Tekale, S., Shingavi, P., Wandhekar, S. and Chatorikar, A., 2018. Prediction of chronic kidney disease using machine learning algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 7(10), pp.92-96.

Alassaf, R.A., Alsulaim, K.A., Alroomi, N.Y., Alsharif, N.S., Aljubeir, M.F., Olatunji, S.O., Alahmadi, A.Y., Imran, M., Alzahrani, R.A. and Alturayeif, N.S., 2018, November. Preemptive diagnosis of chronic kidney disease using machine learning techniques. In 2018 international conference on innovations in information technology (IIT) (pp. 99-104). IEEE.

Ma, F., Sun, T., Liu, L. and Jing, H., 2020. Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. Future Generation Computer Systems, 111, pp.17-26.

Amirgaliyev, Y., Shamiluulu, S. and Serek, A., 2018, October. Analysis of chronic kidney disease dataset by applying machine learning methods. In 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-4). IEEE.

Sobrinho, A., Queiroz, A.C.D.S., Da Silva, L.D., Costa, E.D.B., Pinheiro, M.E. and Perkusich, A., 2020. Computer-aided diagnosis of chronic kidney disease in developing countries: A comparative analysis of machine learning techniques. IEEE Access, 8, pp.25407-25419.

Almansour, N.A., Syed, H.F., Khayat, N.R., Altheeb, R.K., Juri, R.E., Alhiyafi, J., Alrashed, S. and Olatunji, S.O., 2019. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in biology and medicine, 109, pp.101-111.

Ogunleye, A. and Wang, Q.G., 2018, June. Enhanced XGBoost-based automatic diagnosis system for chronic kidney disease. In 2018 IEEE 14th International Conference on Control and Automation (ICCA) (pp. 805-810). IEEE.

Almarashi, A., Alghamdi, M. and Mechai, I., 2018. A new mathematical model for diagnosing chronic diseases (kidney failure) using ANN. Cogent Mathematics & Statistics, 5(1), p.1559457.

Pasadana, I.A., Hartama, D., Zarlis, M., Sianipar, A.S., Munandar, A., Baeha, S. and Alam, A.R.M., 2019, August. Chronic kidney disease prediction by using different decision tree techniques. In Journal of Physics: Conference Series (Vol. 1255, No. 1, p. 012024). IOP Publishing.

Muslim, M.A., Herowati, A.J., Sugiharti, E. and Prasetiyo, B., 2018, March. Application of the pessimistic pruning to increase the accuracy of C4. 5 algorithm in diagnosing chronic kidney disease. In Journal of Physics: Conference Series (Vol. 983, No. 1, p. 012062). IOP Publishing.

Wibawa, M.S., Maysanjaya, I.M.D. and Putra, I.M.A.W., 2017, August. Boosted classifier and features selection for enhancing chronic kidney disease diagnose. In 2017 5th international conference on cyber and IT service management (CITSM) (pp. 1-6). IEEE.

Lestari, A., 2020. Increasing Accuracy of C4. 5 Algorithm Using Information Gain Ratio and Adaboost for Classification of Chronic Kidney Disease. Journal of Soft Computing Exploration, 1(1), pp.32-38.

JerlinRubini, L. and Perumal, E., 2020. Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm. International Journal of Imaging Systems and Technology, 30(3), pp.660-673.

Elhoseny, M., Shankar, K. and Uthayakumar, J., 2019. Intelligent diagnostic prediction and classification system for chronic kidney disease. Scientific reports, 9(1), pp.1-14.

Komal, K.N., Tulasi, R.L. and Vigneswari, D., 2019. An ensemble multi-model technique for predicting chronic kidney disease. International Journal of Electrical and Computer Engineering, 9(2), p.1321.

Kayaalp, F., Basarslan, M.S. and Polat, K., 2018, April. A hybrid classification example in describing chronic kidney disease. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4). IEEE.