Chronic Kidney Disease Prediction using Machine Learning Ensemble Algorithm

Nikhila
PG Student
Department of Electronics and Communication Systems
Centre for PG Studies, Mysuru, Karnataka, India
niki130782@gmail.com

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397144

Abstract*—Chronic Kidney Disease is one among the non-contagious illnesses that affect most of the individual in the world. The main factors of risk for the Chronic Kidney Disease are Diabetes, Heart Ailment, Hypertension. The Chronic Kidney Disease shows no symptoms in the early stages and most of the cases are diagnosed in the advanced stage. This leads to delayed treatment to the patient which may be fatal. Machine learning technique provides an efficient way in the prediction of Chronic Kidney Disease at the earliest stage. In this paper, four ensemble algorithms are used to diagnose the patient with Chronic Kidney Disease at the earlier stages. The machine learning models are evaluated based on seven performance metrics including Accuracy, Sensitivity, Specificity, F1-Score, and Mathew Correlation Coefficient. Based on the evaluation the AdaBoost and Random Forest performed the best in terms of accuracy, precision, Sensitivity compared to Gradient Boosting and Bagging. The AdaBoost and Random Forest also showed the Mathew Correlation Coefficient and Area Under the curve scores of 100%. The machine learning model proposed in this paper will provide an efficient way to prevent Chronic Kidney diseases by enabling the medical practitioners to diagnose the disease at an early stage.

Keywords—AdaBoost, Gradient Boosting, Random Forest, Ensemble Algorithm, Machine Learning.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a serious medical issue in India, with 1 out of 10 individuals are experiencing some type of kidney sickness [1]. Around 1,75,000 new instances of renal failure are enlisted in Asian nation yearly, serious enough to perform dialysis [1]. Universally, 850 million individuals are presently expected to have kidney infections from an assortment of causes and ongoing kidney sickness. The world notices 2.4 million deathsevery year and is currently the sixth most noteworthy developing reason for death [2]. Also, these are the numbers when the cases in India, which is home to world's 17 percent populace, remain to a great extent undocumented and unregistered [2].

Kidneys are one of the vital organs in the lower back, one kidney arranged on one or the other side of the spine. The kidneys main function is to purify blood and remove the waste from the body in the form of urine [1]. Kidney fails to function when it cannot filter the waste materials from the body. Hypertension, obesity and diabetes are the key factors for Chronic Kidney Disease (CKD). Kidney infection is normally called a silent condition [2]. Exactly when a patient begins experience symptoms like weakness, shortness of breath etc. infers that their kidney work has recently tumbled to 25 percent or less.

The Chronic Kidney Disease shows no symptoms in the early phase and most of the cases are left undiagnosed until it reaches the advanced stage. This leads to delayed treatment to the patient which may be fatal**.** Initial diagnosis of the disease is very significant so that the risk may be minimized and proper treatment can be given to the patient at the early stage. The utilization of Artificial Intelligence in the field of clinical investigation is expanding day by day. This can be contributed principally to the improvement in the classification and recognition systems utilized in disease prediction which is capable to give information that guides clinical specialists in early discovery of lethal diseases and in this manner, increment the endurance pace of patients altogether [3]. Previous works for classification of Chronic Kidney Disease used individual classifiers and was evaluated based on single metrics like accuracy. This may lead to reducing the overall performance of the model. On the other hand, ensemble algorithm combines the outcome of several weak learners and the final result is obtained either by averaging or majority voting. The errors made by single weak learners can be overcome by using several weak learners. This paper includes four machine learning ensemble algorithms to build the model with high performance parameters that will help the diagnosis of CKD in a more efficient way.

II. LITERATURE SURVEY

Ahmed J. Aljaaf et al proposed a model for diagnosing CKD using the dataset available in the UCI repository. The dataset had 24 attributes obtained from the blood and the urine test. Only 30% of the attributes were used to build the model. Four supervised machine learning classifiers were used for

predicting the disease achieving a sensitivity of 98.97%, specificity of 100% and AUC of 99.5% [4].

El-Houssainy A. and RadyAyman S. Anwar have classified chronic kidney disease using 4 algorithms mainly Probabilistic Neural Networks(PNN), Radial Basis Function (RBF), Support Vector Machine (SVM)algorithms and Multilayer Perceptron (MLP). The model was built based on the different stages of CKD which was grouped according to the Glomerular Filtration Rate (GFR) measured values. PNN gave the best classification result as compared with other classification algorithm with an overall accuracy of 96.7%. The Execution time of MLP was 3s whereas the execution time of PNN was 12s [5].

Anusorn Charleonnan et al. built a machine learning technique for classifying chronic kidney disease. Four machine learning classifiers were used including Decision Tree, Logistic Regression (LR), K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). The performances of these were compared to choose the best classifier. SVM proved to be the best classifiers among others with sensitivity of 99% [6].

Indu Yekkala et al. in their research used machine learning ensemble algorithm technique along with feature selection technique to predict Heart diseases. The ensemble algorithm included Bagging, AdaBoost and Random Forest. Using the PSO method 6 attributes with low rank was removed and the experiment was conducted on the remaining 7 attributes that scored high rank. Bagging obtained an overall performance of 100% when compared with other classifiers [7].

Pramila Arulanthu and Eswaran Perumal in their research used feature Selection methods to reduce the number of features used to classify the Kidney disease. The different algorithms used were Jrip, SMO, IBK and Naïve Bayes. The result obtained from reduced features was compared with the result obtained from original dataset [12].

Guozhen Chen et al. Adaptive hybridized Deep Convolutional Neural Network (AHDCNN) has been proposed for the early identification of Kidney infection effectively and adequately. Characterization innovation productivity relies upon the part of the informational collection. To improve the accuracy of the arrangement framework by reducing the features has been obtained by utilizing CNN. These elevated level properties help to fabricate a regulated tissue classifier that segregates between the two kinds of tissue [13].

III. MATERIALS AND METHODS

The scope of this paper is to make use of Machine Learning ensemble algorithms to predict chronic kidney disease. The data set used for building the model is taken from the UCI repository [8]. The dataset involves information of 400 patients with 25 attributes including the class. The dataset consists of data collected from blood test and urine test and also some of the general information such as age, appetite. Out of the 400 patients, 250 patients were diagnosed by CKD and 150 patients were healthy. The dataset is divided into training set and testing set. The Training set is utilized to prepare the model with different Machine Learning ensemble algorithms. The hyper parameters of each of the ensemble classifiers are tuned to get the best parameters that will provide the best model for predicting the chronic kidney disease in patient. The trained model is then used on the testing dataset. The model is assessed based on the performance of each model in terms of accuracy, sensitivity, specificity, precision, F-score, ROC-AUC and Mathew Correlation Coefficient.

A. Missing Values

CKD dataset available in the UCI Repository is raw and needs some data preprocessing techniques before applying it to the model. The CKD dataset consist of missing values in many of the features. Figure 1 shows the count of missing values in some of the attribute. All the missing values are replaced by mean for numerical attributes and by mode for categorical attributes.

B. Feature Scaling

The dataset consist of attributes having different range. Such data cannot be applied to the machine learning model. Therefore it requires rescaling which ensures that all the features fall under the same scale. MinMax scaling technique is used to scale the attibutes in the range 0 and 1.

C. Training and Testing Dataset

The dataset is split into training dataset of 70% which includes 280 patient details and 30% testing dataset which includes 120 patient details. The 70% of the training set is further split into.

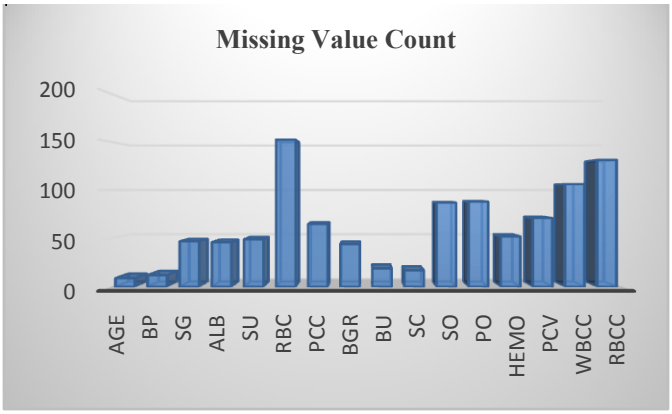

Figure 1 Missing Value Count

training set and validation set using the cross-validation technique A 10-fold cross-validation is used, which splits the training set into 10 folds. In each fold, one cluster is held as validation data and the remaining nine groups are used to train the model. For each fold the evaluation score is retained. Finally, the mean of all the evaluation score for the 10-fold cross validation is calculated.

D. Classifiers

Once the pre-processing of the dataset is completed the next step is to train different machine learning ensemble

algorithm using the training dataset and test the model on the testing dataset and evaluate the model performance.

1) Bagging

Bootstrap Aggregation ordinarily known as Bagging is a sort of ensemble algorithm that arbitrarily picks some instances from the training set with substitution [9]. In Bagging, bootstrap samples are obtained from the training dataset collection and the classifier is set up with every model. The outcome from each classifier is consolidated, and the final result is obtained from the process of majority voting. Examination shows that bagging can be used to upgrade the overall performance of a weak classifier preferably [9].

2) AdaBoost

Adaptive Boosting generally called as AdaBoost is another ensemble classifier. The regular over fitting issue present in various Machine Learning systems can be diminished using AdaBoost. AdaBoost works by picking base classifiers and improve its performance by identifying the misclassified cases from the training datasets in an iterative method. Equal weight-age are assigned to all training samples and a weak classifier is chosen. After each iteration, the base classifiers are applied to the training dataset and increase the weights of the misclassified characteristics. The cycle is iterated n times, each time applying base classifier on the training set with updated weights. In the last model, proposed approach consolidates the output for each weak classifier either by majority voting or averaging [7].

3) Random Forest

Random Forest helps in clinical applications for better accuracy by combining a group of weak classifiers like Decision Tree. It produces N number of Decision trees by using randomly picked attributes as their information. In Random Forest, the bias is not changed, but the number of trees increases. The outcomes from all the trees can be picked by casting a vote or averaging [7].

4) Gradient Boosting

As opposed to Random Forest, this model continuously creates decision trees using gradient decent to minimize the loss function. A final forecast is made using a weighted dominant part vote of the whole decision trees [10]. Gradient boosting invalidates the over-fitting issue and manages the bias. [11].

E. Performance Metrics

In order to estimate the performance of chronic kidney disease model using machine learning some of the performance metrics are utilized from the confusion table. Table I shows the Confusion matrix with Accuracy, Positive predicted value and negative predicted value.

Table I Confusion matrix for CKD

Confusion Matrix		Actual Values: CKD =1	Actual Values: No CKD= 0	Accuracy= (TP+TN)/(TP+FP+FN+TN)	
Observed Values	CKD =1	TP	FP	Positive Predictive Value	TP/(TP+FP)
	No CKD =0	FN	TN	Negative Predictive value	TN/(FN+TN)

True Positive (TP) = Samples correctly predicted as having CKD.

False Positive (FP) = Samples falsely predicted as having CKD.

False Negative (FN) = Samples Falsely predicted as not having CKD.

True Negative (TN) = Samples correctly predicted as not having CKD.

The Different metrics used in evaluating the model are as follows. Equation (1), (2), (3), (4), (5) and (6) are used to calculate the different metrics.

Accuracy

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \quad (1)$$

Precision (Positive Predictive Value)

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (2)$$

1) Sensitivity (Recall)

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (3)$$

2) Specificity

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \quad (4)$$

3) F1-Score

$$\text{F1Score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (5)$$

4) *Mathew Correlation Coefficient (MCC))*

MCC provides a balanced result in case the dataset is imbalanced. It takes into account all the four parameters in the confusion matrix [14].

$$\text{MCC} = \frac{(\text{TP}*\text{TN})-(\text{FP}*\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (6)$$

IV. RESULTS AND DISCUSSIONS

For this study, the dataset used was Chronic Kidney Disease dataset available in the UCI repository. The dataset

had 400 records out of which 250 records had CKD and 150 records NoCKD. The missing values in the dataset were handled using the mean method for numerical values and mode method for categorical values. The labels were encoded using the label encoding method. Min-Max normalizing technique was used to scale all the attributes in the range 0 and 1. The pre-processed dataset was split into training and testing dataset. The training set consists of 280 records and testing set consists of 120 records. Four ensemble algorithms like Bagging, Gradient Descent, Random Forest and AdaBoost were used. The model was trained with the training set using 10-Fold Cross validation technique. The Hyper parameters of each of the model were tuned to get the best performance. The performance of each of the model was evaluated using different metrics like Accuracy, recall, specificity, precision, f1-score, MCC and ROC-AUC curve. Table II shows the confusion matrix of various ensemble algorithms on test dataset.

Table II Confusion Matrix of Ensemble Classifiers

		Predicted Value				
		BAGGING			ADABOOST	
		No CKD	CKD		NO CKD	CKD
Actual value	NO CKD	36	1		37	0
	CKD	0	83		0	83
	NO CKD	37	0		37	0
	CKD	0	83		2	81
		RANDOM FOREST			GRADIENT BOOSTING	

As shown in table II, bagging classifier wrongly classified 1 patient as having NoCKD. Gradient Boosting Classifier falsely classified 2 patients as having CKD. AdaBoost and Random Forest perfectly classified all the patients. Table III shows the performance of the classifier based on Accuracy, Sensitivity, Specificity and Precision. Figure 2 shows the performance of various classifiers.

Table III Accuracy, Sensitivity, Specificity and Precision of Various Ensemble Classifiers

	Accuracy	Sensitivity	Specificity	Precision
Bagging	0.991666	1	0.988095	0.972973
AdaBoost	1	1	1	1
Gradient Boosting	0.983333	0.9759	1	1
Random Forest	1	1	1	1

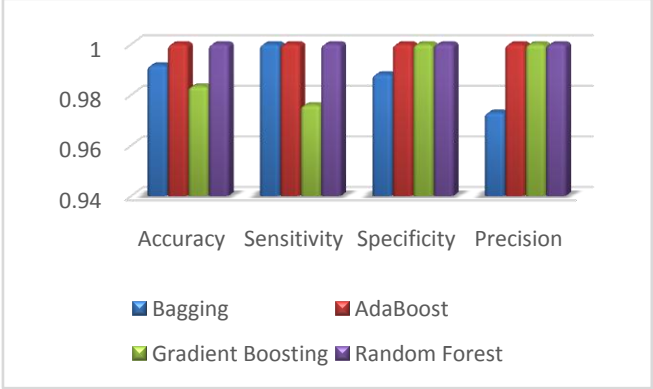

Figure 2: Performance of various Ensemble Classifiers

As shown in figure 2, Bagging has as accuracy of 99.166% whereas Gradient Boosting has an accuracy of 98.33%. AdaBoost and Random Forest has an accuracy of 100%. In terms of sensitivity also called as recall, Gradient Boosting has a sensitivity of 97.59% which is the lowest when compared to sensitivity of other classifiers which is 100%. In terms of specificity, Bagging obtained a specificity of 98.8% whereas other classifiers obtained a specificity of 100%. In terms of precision, Bagging observed a precision of 97.29% whereas other classifiers obtained 100% precision. Table IV shows the performance of ensemble classifiers based on F1-score, Area under Curve (AUC) and Mathew Correlation Coefficient (MCC). Figure 3 shows the performance of the ensemble classifiers based on F1-score, AUC and MCC

Table IV Performance of Ensemble Classifiers based on F1-Score, AUC and MCC

	F1-Score	AUC	MCC
Bagging	99.4	98.6	98.05
AdaBoost	100	100	100
Gradient Boosting	98.78	98.8	96.22
Random Forest	100	100	100

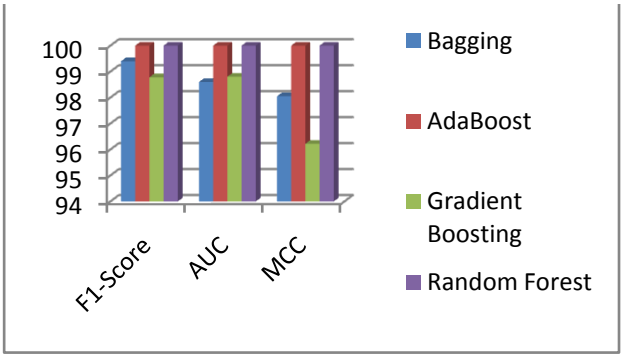

Figure 3: Performance of Ensemble Classifiers based on F1-Score, AUC and MCC

As shown in Figure 3, the F1-Score for bagging is 99.4% whereas for Gradient Boosting it is 98.78%. AdaBoost and Random Forest obtained a F1-Score of 100%. The Area under Curve for Bagging is 98.6% and for Gradient Boosting of 98.8%. Random Forest and AdaBoost obtained an AUC of 100%. The Mathew Correlation Coefficient of Bagging was observed as 98.05% and that of Gradient Boosting is 96.22%. A 100% of MCC was observed in both Random Forest and AdaBoost.

V. CONCLUSION

Chronic Kidney Disease (CKD) is one of the diseases that affect the people in large numbers. As the symptoms of CKD are not visible in the early stages many a times the disease is only detected when it has reached an advanced stage. This may lead to failure of the kidney and hence death. Machine learning classifiers provide an efficient way to predict the disease at an early stage. Ensemble classifiers combine the predicted output of various classifiers which further enhance the performance of the model. The four-ensemble algorithm like Bagging, Random Forest, AdaBoost and Gradient Boosting were used. The performance of these classifiers was evaluated using different metrics. Based on Accuracy the AdaBoost and Random Forest performed better with 100% Accuracy. But since the dataset was slightly imbalanced accuracy cannot be the only parameter to be considered for evaluation. Based on Precision, Bagging showed 97.29% and AdaBoost, Gradient Boost and Random Forest showed 100%. The F1-Score and AUC of 100% for AdaBoost and Random Forest was better compared to Bagging and Gradient Boost. Based on the evaluation AdaBoost and Random Forest was the best classifier when compared with Bagging and Gradient Boosting.

REFERENCE

[1] N. Health, "World Kidney Day 2019: Important aspects for Chronic Kidney Disease in Modern time," *Narayana Health Care*, Mar. 14, 2019. https://www.narayanahealth.org/blog/world-kidney-day-2019-important-aspects-for-chronic-kidney-disease-in-modern-time/ (accessed May 12, 2020).

[2] "World Kidney Day 2019: CKD is 6th deadliest disease worldwide causing 2.4 million deaths per year; here's how to reduce risk of renal ailments," *Firstpost*. https://www.firstpost.com/india/world-kidney-day-2019-ckd-is-6th-deadliest-disease-worldwide-causing-2-4-million-deaths-per-year-heres-how-to-reduce-risk-of-renal-ailments-6256331.html (accessed May 12, 2020).

[3] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Dec. 2018, pp. 1–4, doi: 10.1109/CCAA.2018.8777449.

[4] A. J. Aljaaf *et al.*, "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, Jul. 2018, pp. 1–9, doi: 10.1109/CEC.2018.8477876.

[5] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Inform. Med. Unlocked*, vol. 15, p. 100178, Jan. 2019, doi: 10.1016/j.imu.2019.100178.

[6] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *2016 Management and Innovation Technology International Conference (MITicon)*, Oct. 2016, p. MIT-80-MIT-83, doi: 10.1109/MITICON.2016.8025242.

[7] I. Yekkala, S. Dixit, and M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization," in *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, Aug. 2017, pp. 691–698, doi: 10.1109/SmartTechCon.2017.8358460.

[8] "UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set." https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease (accessed May 12, 2020).

[9] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Inform. Med. Unlocked*, vol. 16, p. 100203, Jan. 2019, doi: 10.1016/j.imu.2019.100203.

[10] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 211, Nov. 2019, doi: 10.1186/s12911-019-0918-5.

[11] R. Islam and Md. A. Shahjalal, "Soft Voting-Based Ensemble Approach to Predict Early-Stage DRC Violations," in *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2019, pp. 1081–1084, doi: 10.1109/MWSCAS.2019.8884896.

[12] P. Arulanthu and E. Perumal, "Predicting the Chronic Kidney Disease using Various Classifiers," in *2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, Dec. 2019, pp. 70–75, doi: 10.1109/ICEECCOT46775.2019.9114653.

[13] G. Chen *et al.*, "Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform," *IEEE Access*, vol. 8, pp. 100497–100508, 2020, doi: 10.1109/ACCESS.2020.2995310.

[14] M. A. U. H. Tahir, S. Asghar, A. Manzoor, and M. A. Noor, "A Classification Model For Class Imbalance Dataset Using Genetic Programming," *IEEE Access*, vol. 7, pp. 71013–71037, 2019, doi: 10.1109/ACCESS.2019.2915611.