IJAMEC
e-ISSN: 2147-8228

INTERNATIONAL JOURNAL OF APPLIED MATHEMATICS ELECTRONICS AND COMPUTERS

www.dergipark.org.tr/ijamec

International Open Access

Volume 08
Issue 04

December, 2020

Research Article

Comparison analysis of decision tree and ensemble models in the classification of chronic kidney diseases

Olawumi Olasunaknmi[a,*], ***Odunayo Olanloye***[b], , ***Abdulquadri Adegbiji*** [c]

[a]*Ladoke Akintola University of Technology, PMB 4000, Ogbomoso, Oyo State, Nigeria*
[b]*Bowen University, PMB 214, Iwo, Osun State, Nigeria*
[c]*Student, Ekiti State University, Ekiti State, Nigeria*

ARTICLE INFO

Article history:
Received 10 September 2020
Accepted 26 October 2020

Keywords:
Ensemble Classifier
Decision trees
Kidney
MATLAB

ABSTRACT

The world is now in the era of big data and processing, and exploring the data has become one of the significant challenges. Hence, researchers have done a lot to analyse these data in the health sector to enhance disease detection and classification using artificial intelligence and ML principles. Kidney disease is one of the terrible conditions in which its late detection has sent many people to untimely graves. ML classifiers have been employed in many dimensions to classify heart disease, but, existing works have not explored the variants of each method for selection of best model parameters. An attempt is being made in this research to study the behaviour of three (3) variants each from two(2) tree-based models in the classification of Kidney Disease. Three of the variants are Complex, Medium and Simple models of Decision tree classifier and the other one are Boosted, Bagged and RUSBoosted of Ensemble Classifiers. Using MATLAB for implementation, the model performance established that the accuracy of Ensemble Classifier (Bagged tree model) is the best, concerning the speed, Decision tree (Complex and Simple tree models have the same and highest value). Hence, the two are the best. In terms of training time, Decision tree(Simple tree) has the least time and therefore the best.

1. Introduction

The kidney is one of the major organs of the body that is responsible for the removal of waste products and excess water from the blood. Its malfunctioning often leads to various health challenges such as High Blood Pressure, Anemia, the disorderliness of the cholesterol etc. Hence, it has become one of the most chronic mono-communicable diseases epidemics globally [7]. Chronic Kidney Disease (CKD) is responsible for the death of millions of people across the globe. According to medical experts, its leading cause includes high blood pressure and diabetics. The disease affects the performance of renal arteries by gradually reducing its function capability for months or years. The death rate traced to this disease is becoming too alarming and worrisome. So, it has become a significant problem in the health sector across the globe, and it is a leading cause of morbidity and mortality in various countries as it accounts for 60% of the death record around the world [12].

In this era of advance technology, various means or tools are available for collection of a series of data. Despite the availability of data in the medical field, diagnosis of kidney diseases still poses a severe threat, and it has become an extreme challenge. Past research works have attempted to find possible solutions to this problem. Artificial intelligence has become so popular because of its ability to solve multidimensional real-life issues making use of the various tools. It has to do with a scientific method of creating intelligent machines used to solve real-life problems in a human-like fashion. It is concerned with developing computer systems that can store knowledge and effectively use the experience to solve future issues and accomplish task [1]. An intelligent system should be able to think and act humanly.

One significant area of artificial intelligence is Machine

[*] Corresponding author. E-mail address: *wumirose@gmail.com*

DOI: 10.18100/ijamec.792863

Learning(ML). It is the act of training AI models and applying the knowledge acquired to solve future problems. The types of ML models include supervised, unsupervised and reinforcement models. Decision tree and Ensemble classifiers are the product of supervised learning algorithms. This research work evaluates the performance of the models in the classification of CKD using MatLab for model implementation and analysis.

2. Literature Review

2.1. ML Algorithms

Decision tree: is used for regression and classification based on a set of rules. It is easy to interpret the rules that guide a decision and incorporate numeric data because of its non-parametric nature. [6] submitted that decision trees are robust in training data in that classification is faster once the decision rules are defined. Notable disadvantages include overfitting when used on big data; this eventually affects the final result obtainable. Again, Decision trees can not predict the minimum and the maximum limit of the response variable in the training data.

Decision tree solvable problems are practically categorised into three parts depending on the level of complexity of the problem, which is determined by the number of splits. The three categories include simple, medium and complex decision trees.

A simple decision tree has a maximum split of four with a few numbers of splits; a medium tree has a maximum of 20 splits but with fewer numbers of leaves; while a complex tree has a maximum of 100 splits with many leaves [15]. The variants of the decision trees are shown in Figure 1.

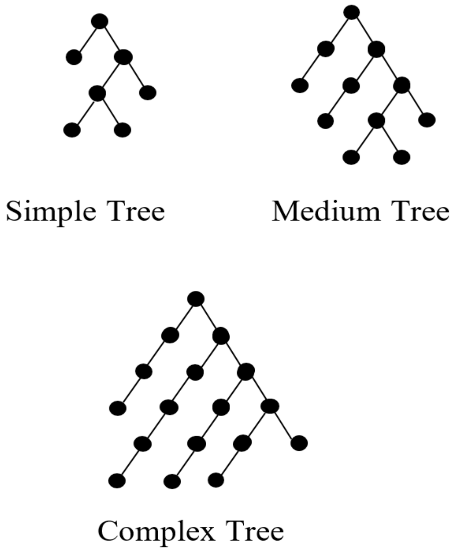


Figure 1. Variants of Decision Trees

Ensemble Classifier: As depicted in Figure 2, this classifier combines different types of individual decisions by weighted or unweighted voting to carry out classification of other examples. The word ensemble means "union of parts". In most cases, the regular classifier mainly used for prediction is weak to a certain extent, they are prone to errors, and such error reduces the accuracy of the result obtained from time to time. [2] submitted that ensemble classifiers generated from various base classifiers perform better than any constituent classifier. The base classifier may differ in the algorithm used, hyperparameters representations or training set. The significant advantage of ensemble classifier is that it reduces bias and variance.

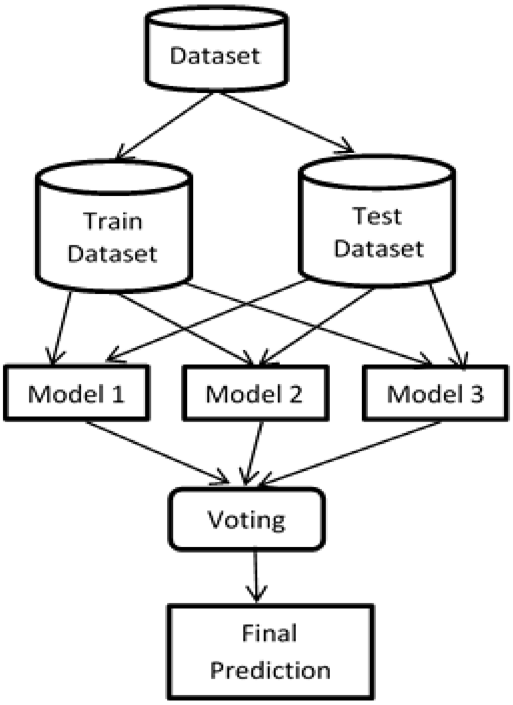


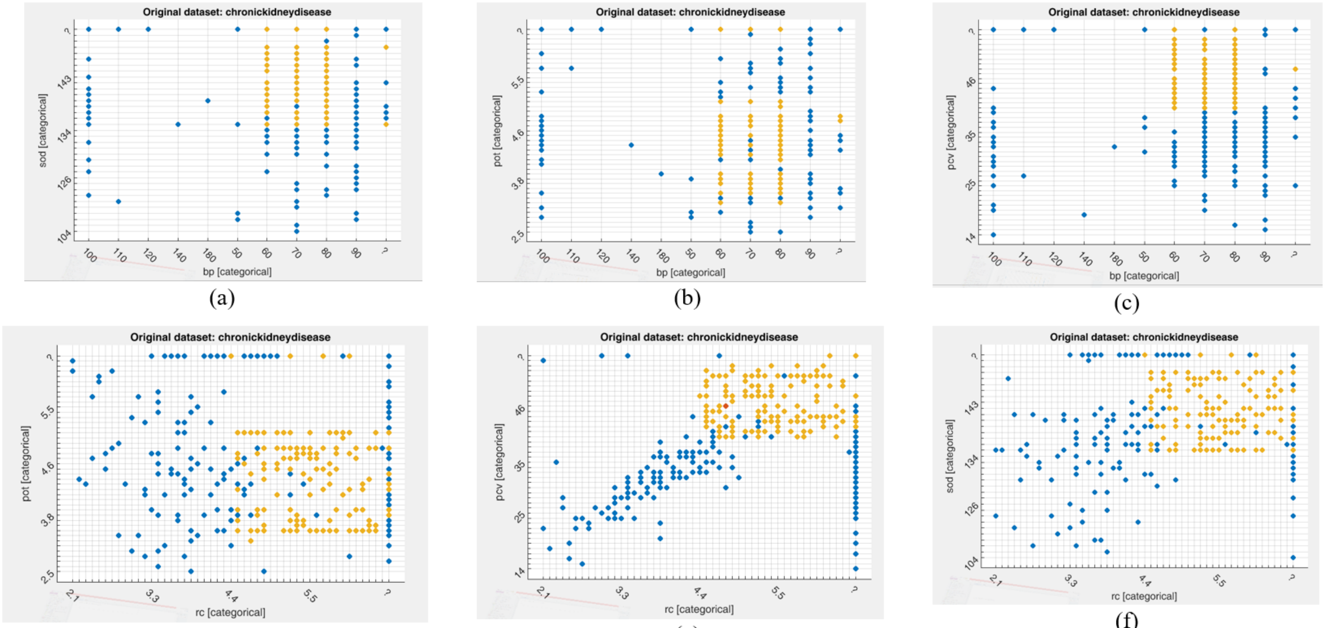
Figure 2. Basic Outline of Ensemble Techniques [2]

Boosted trees create an ensemble of medium decision tree making use of Adaboost algorithm. When compared with others in this same group, it has relatively short execution time and uses little memory, but in most cases, it requires more ensemble members. Bagged tree, on the other hand, is an ensemble classifier for complex decision tree. It has a relatively low speed, requires large memory space but produces a more accurate result. Rusboosted tree is also an ensemble classifier applicable mostly, where there is skewed data with a loot off observations in one class.

2.2. Related Research

[12] presented a research article titled diagnosis of CKDs using ML algorithm: The research attempted to reduce the diagnosis time and improve the diagnosis accuracy using classification algorithms. The algorithm used includes Back Propagation Neural Network Radial Basis Function and Random Forest [14]. The result shows that the Radial Basis Function gives a better result.

[3] presented a research work titled Missing Data Classification of CKD. The research compared different algorithms that deals with missing data. The algorithm used



(a) (b) (c)

(d) (e) (f)

Figure 3. 2D Plot of Features in the Dataset Blood Pressure-mm/Hg **vs** Sodium- mEq/L; (b) Blood Pressure-mm/Hg **vs** Potassium-mEq/L Blood Pressure-mm/Hg **vs** Packed Cell Volume; (d) Red BloodCell (RBC)-millions/cmm **vs** Potassium-mEq/L (e) RBC-millions/cmm **vs** Packed Cell Volume; (f) RBC-millions/cmm **vs** Sodium- mEq/L

includes K-Neural Network Classifier, Bayes Classifier, Decision tree and Support Vector Machine (SVM). Decision tree appears to be the best classifiers among others.

[5] used SVM and Neural Network algorithm for classification and prediction of CKD on the dataset obtained from UCI ML repository Ten-fold validation tests were used to determine the performance of the classification algorithms. The algorithms classified patients into chronic disease and non-chronic disease groups, and the work concluded that SVM performed better than Neural Network.

[10] implemented K Nearest Neighbor J48, Artificial Neural Network, Naïve Bayes and SVM for classification. The result showed that K-nearest neighbour is the best with an accuracy of 99.5%.

[13] used SVM and Artificial Neural Network to predict Kidney Disease and compared the performances of the two algorithms based on accuracy and execution time. The researcher established that the accuracy of ANN is better than the SVM in the prediction of CKD.

[11] built a feature model that deals with comparative analysis of predicting accuracy of chronic kidney dataset. It made use of stepwise regression classification (SRC) model built around the Random Forest Classification algorithm. The researcher obtained the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), Root Relative. Square Error (RRSE) and Kappa values and affirmed that Multilayer Perceptron (MLP) and SVM (SVM) performed better than others.

[9] researched the development of an ensemble approach to chronic kidney disease diagnosis using K-nearest neighbour, Naive Bayes and decision tree. The result confirmed that the ensemble approach outperformed decision tree algorithm with 95% and 89.2% accuracy for DT and Ensemble Classifier respectively.

[8] in research on detection of Chronic disease using Ensemble Classifier, adopted AdaBoost Classifier, Bagging and Random subspace Ensemble Learning for CKD diagnosis and, concluded that ensemble learning outperformed the individual classifiers.

[4] proposed an Ensemble model with feature selection technique for classification of CKD. Implementation of the model was through a combination of two similar trained model - single-rule classification and conditional inference tree with 92 and 93.75% accuracy respectively. They compared the accuracies with that of SVM, ANN and DT, which produced an accuracy of 63, 83, and 91% respectively.

Table 1. Training Time, Accuracy and Prediction Speed for each Model

	Model Name	**Accuracy (%)**	**Speed(obs/sec)**	**TrainingTime(secs)**	
1.1	Complex tree	95	4200	0.9163	Decision tree
1.2	Medium tree	95	4100	0.9028	
1.3	Simple tree	95	4200	0.8577	
2.1	Boosted tree	94.5	570	9.4414	Ensemble classifier
2.2	Bagged tree	98.8	360	9.3910	
2.3	RUSBoosted tree	75.5	580	8.5940	

[16] presented a research work titled prediction of CKD using data mining features selection and ensemble method. The result showed that the ensemble method produced a more accurate result.

The literature established that researchers had not done enough to compare the behaviour of the variants of Decision tree and Ensemble classifiers when they are being used to solve classification problems. Therefore, this research work compares the performance of six models, three each from Decision tree and Ensemble classifier.

3. Material and Methods

This research work compared the performances of decision tree and ensemble classifier in the classification of CKD. The dataset obtained from UCI repository(source) contained some missing, and duplicated values, hence, we pre-processed the data. It has 25 attributes, some which were explored and visualised as shown in Figure3 a, b, c, d and e.

Moreover, of the entire dataset, the instances of at least 2% correlated features were selected and split into two (2) - 80% meant for training and 20% for testing. Again, for model training, the 80% train set was cross-validated using 10- fold validation. ROC Curve, Confusion matrix, and Parallel Coordinate Plot were used to visualise the test performance of the model.

3.1. Receiver Operating Characteristics (ROC)

ROC is used to measure the performance of the classifier. It indicates at different classes, the True Positive Rate (TPR) against the False Positive Rate (FPR) in the classification outcome. for instance, an FPR value of 0.4 indicate that the classifier incorrectly assigns 40% observations to the positive class. At the same time, 0.8 TPR implies that the classifier classifies 80% instances appropriately into the positive class to which they originally belong. The ROC curve with no misclassified result appears as a right angle to the left of the plot, and the curve of a classifier with many misclassified instances appears as a line at an angle of 450. The area under the curve determines the overall performance of the classifier, therefore, the larger the area, the better the performance of the classifier.

3.2. Confusion Matrix (CM)

CM is also used to measure the performance of the classifier. The row shows the true class and column shows predicted class. If the cells in the diagonal are green and of high percentage, then the performance of the classifier is classifier is quite good, and it indicates that it classifies observations correctly. But red cells of high percentage shows that the classifier performed poorly, meaning that numerous data points are misclassified.

3.3. Parallel Coordinate Plot (PCT)

PCT Compares the features of several individual observations on a set of numerical variables. Values are plotted as a series of lines connected across each axis. It enables comparison of features of different samples with several variables and visualisation data in 3D form.

4. Results and Discussion

Table 1 shows the results obtained from training and testing of the selected models for classification of CKD. The bagged tree has 98.8%, the highest accuracy, followed by the Complex, Medium and Simple tree with a 95% level of accuracy in each case. Boosted and RUSBoosted tree comes last with 94.5% and 75.5% level of accuracy, respectively

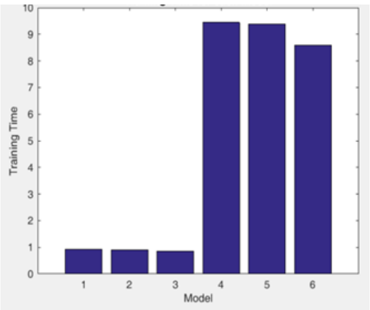


Figure 4. Plot of Training Time for each Model

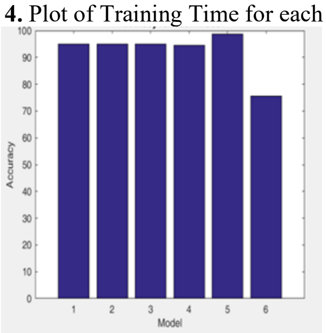


Figure 5. Plot of Accuracy for each Model

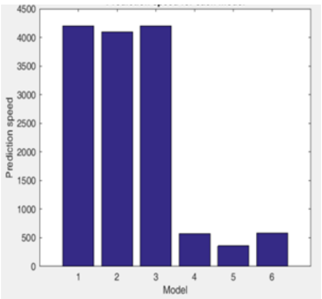


Figure 6. Plot of Prediction Speed for each Model

The results of each variant of the decision tree and ensemble classifiers are explained further in Figure 4 5, and 6. In term of accuracy, the Bagged tree is the best classifier. Complex and Simple tree each has a speed of 4200obs/sec while Medium, RUSBoosted, Boosted and Bagged tree each has a speed of 4100, 580, 570 and 360 obs/sec, respectively. Hence, in term of speed, Complex and Simple trees are the best classifiers.

In term of training time, Simple tree has the least 0.8577 seconds where Medium, Complex, RUSBoosted, Bagged, the Boosted tree has training times of 0.9028, 0.9163, 8.5940, 9.3910 and 9.4414 seconds respectively. These imply that the Simple tree is the best classifier with the least training time. In addition to the accuracy, training time and prediction speed, confusion matrix, Figure7 below shows the ROCAUC and prediction outcome for each of the models.

With the Maximum number of splits set to twenty(20), Complex and Medium tree have lesser falsely classified instances and AUC score (0.97) and therefore outperforms Simple Tree which has AUC score of 0.96. With the maximum number of split set as 20, 30 learners and 0.1 learning rate in each case, Bagged Tree produced just four falsely classified instances and highest approximate AUC score (1.0) while Boosted and RUSBoosted tree with 0.97 and 0.92 AUC Score respectively. The parallel coordinate plot (PCP) in Figure 5b further expressed the relationship between various features and also identified useful predictors used in the classification task. Hence, the test dataset and the misclassified points are shown at a glance.

Literatures reviewed indicated that ensemble classifiers are better than individual classifiers- [8], [16]. Apart from supporting these results, the proposed model shows some improvement compared to the existing models.

The result obtained established the accuracy levels of decision tree and ensemble classifier as 95% and 98.8% respectively. This is an improvement on [9] with an accuracy level of 89.2% and 95% for decision tree and ensemble classifier respectively. The result obtained in this research work is also better than that of [4] where the decision tree and ensemble classifier accuracy values obtained are 92% and 94.25% respectively.

Again, to justify each model in term of computational resources exerted at run-time, the research examined the speed and time of execution of the classifiers as made available in Table1 which was not made available in existing research.

5. Conclusion

In this research work, six (6) models were used to clarify CKD. The accuracy inferred that Bagged tree model is the best with 98.8% level of accuracy. In term of Speed, Complex and Simple tree algorithms are the same with the same value of 4200 obs/sec. And in terms of training time, a simple tree is the best since it has the least training time of 0.8577 seconds.

The performance of the models was compared and established that in term of accuracy Ensemble Classifier (Bagged tree model) is the best as regards the speed, Decision tree(Complex and Simple tree models have the same and highest value). Hence, the two are the best. In term of training time, Decision tree(Simple tree) has the least execution and therefore the best. The instrument used for the implementation is MATLAB ML Classifier.

References

[1] A. F. Kana, “Introduction to Artificial Intelligence Lecture Note”, 2016.

[2] A. Tarun. Towards Data Science. https://tpwardsdatascience.com/advanced_ensemble_class, 2019.

[3] A. Wala and A. Noora, Abdulrahman, “Missing Data Classification of Chronic Kidney Disease”, International Journal of Data Mining and Knowledge Management process (IJDKP), vol 7., pp. 5-6, November 2017.

[4] A.K. Shrivas and K. S. Sanat, “A proposed ensemble model with feature selection technique for classification of chronic kidney disease”, International Journal of Engineering and Advance Technology (IJEAT), vol 9, 2019. DOI: DOI: 10.35940/ijeat.A2207.129219

[5] B. Basma, M. Hajar and H. Abdelkrim, “Performance of Data Mining Technique to Predict In Health Care In Health Care Case Study: Chronic Kidney Failure Disease”, International Journal of Database Management System.(IJDMS), vol 8, June , 2016.

[6] H. Ned, “Intoduction to Decision tree and Random Forest”, American Museum of Natural History’s Center for Biodiversity and Conservation, 2019.

[7] K. S. Sanjay, M. Adeel, F. Ahmad and J. Vivekanand, “A Clinical Database of Kidney Disease”, BMC Nephrology, vol 13, pp. 1471- 2369, 2012.

[8] M. D. Basar and A. Akan, "Detection of chronic kidney disease by using ensemble classifiers”, 10th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, 2017, pp. 544-547.A.K.

[9] O. A. Jongbo , A.O. Adetunbi, B.O. Ogunrinde, B.B. Ajisafe, “Development of an ensemble approach to chronic kidney disease diagnosis”, Scientific Africa, 2020. DOI: https://doi.org/10.1016/j.sciaf.2020.e00456

[10] S. Jyoti, R.C. Gangwar and M. Molute, “A novel detection for Kidney Disease using Improved Support Vector Machine” International Journal of Latest Trends in Engineering and Technology vol 8, pp 114–121, 2015. DOI: http://dx.doi.org 10.21172//.81.015.

[11] S. P. Senthil and P. Anitha, “Comparison of feature selection methods for chronic kidney dataset usind data mining classification analytical model” International Research Journal of Engineering and Technology (IRJET), 6 (2), 2019.

[12] S. Ramya and S. Radha, “Diagnosis of Chronic Kidney Disease using Machine Learning algorithm” International Journal of Innovative research in Computer and Communication Engineering, vol 4. January, 2016.

[13] S. Vijayarani and S. Dhayanand, “Kidney disease prediction using SVM and ANN algorithm”, International Journal of Computing and Business Research (IJBCR), vol 6, 2015.

[14] Z. Sirage and P. Shruti, “Prediction of chronic kidney disease using data mining features selection and ensemble method”, WSEAS Transactions on Information Science Applications, vol 15, pp 168- 176, 2018.

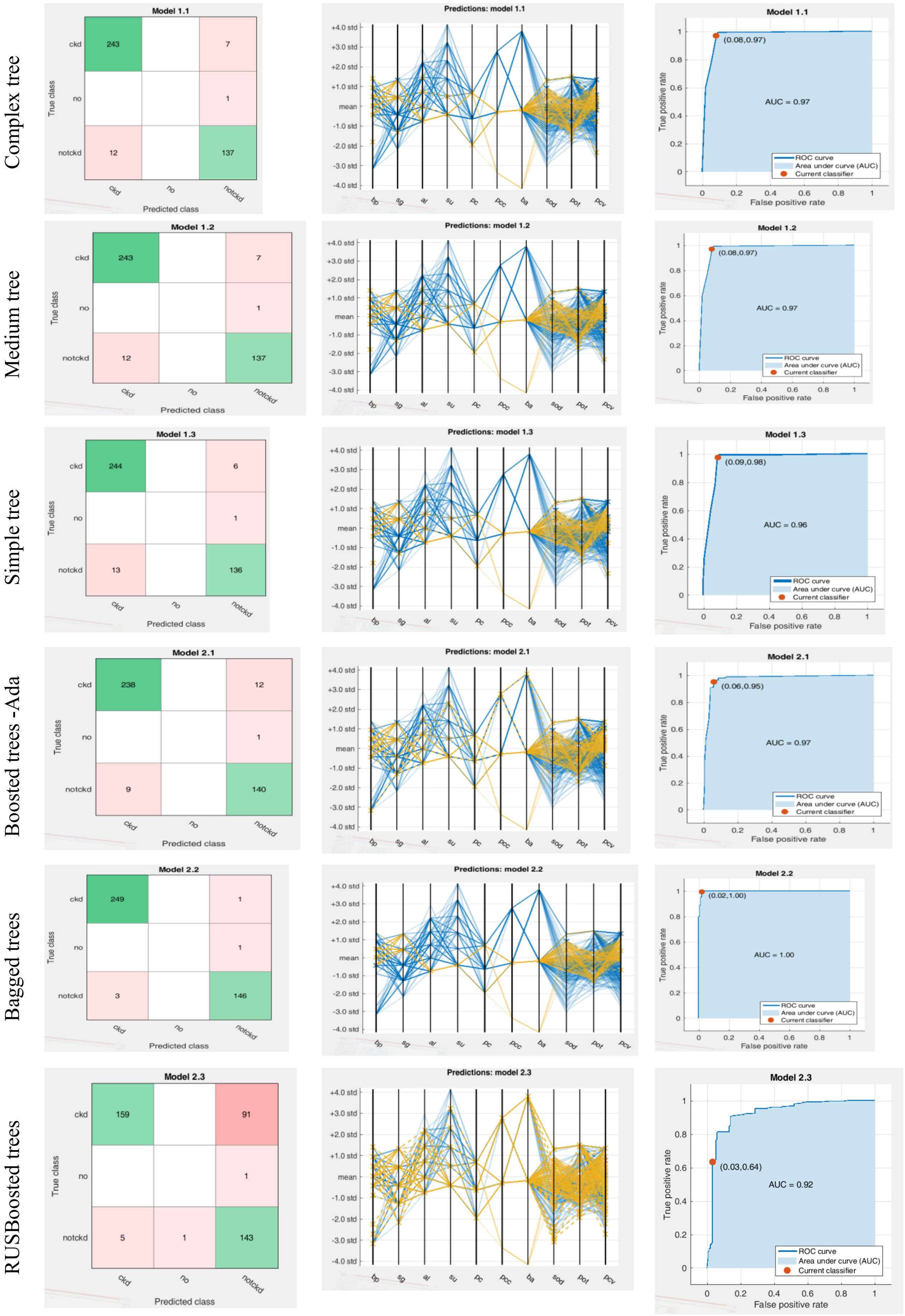


Figure 7. Confusion Matrix, ROCAUC and Prediction Outcome for each Models