Journal of Information Hiding and Multimedia Signal Processing ©2023 ISSN 2073-4212
Ubiquitous International Volume **14**, Number **1**, March **2023**

Chronic Kidney Disease Prediction using Ensemble Machine Learning

Md. Minarul Islam Raju

Department of Electronics and Communication Engineering
Hajee Mohammad Danesh Science and Technology University
Dinajpur -5200, Bangladesh
minarulislamraju771@gmail.com

Sumonto Sarker

Department of Electronics and Communication Engineering
Hajee Mohammad Danesh Science and Technology University
Dinajpur -5200, Bangladesh
sumonto@hstu.ac.bd

Md. Mehedi Islam

Department of Electronics and Communication Engineering
Hajee Mohammad Danesh Science and Technology University
Dinajpur -5200, Bangladesh
mehedi@hstu.ac.bd

Received September 2022; revised October 2022

Abstract. *Chronic kidney disease is considered one of the major diseases now-a-days. Most of the people are affected for their irregular lifestyle. Early-stage prediction can reduce it and can suggest a healthy lifestyle. In this study, we predict kidney disease from secondary data using some machine learning algorithms. Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), K Nearest Neighbors (KNN), Stotostical Gradient Deasent (SGD) are used for analysis. We also propose an ensemble machine learning algorithm by stacking RF, SVC, and LR and named RFSVCLR. This algorithm shows better result than others classifiers. Precision, Recall, F1 Score, Accuracy, Cohen Kappa, and ROC is used to evaluate the performance of the algorithms. RFSVCLR shows 99% accuracy with 99% precision, 99% recall, 99% f1 score and 98% Cohen kappa score that is superior to other classifiers.*
Keywords: kidney disease prediction, machine learning, ensemble learning

1. **Introduction.** Diseases of kidney receiving extensive concern globally as the number of victims is rapidly increasing. A devastating kidney disease is chronic kidney disease (CKD) which increases harmful fluids and waste in blood thereby damage the body internally. Such worsening in physical condition leads to renal failure, in consequence death [1] Kidney failure continues to be one of several forms of end-stage organ failure caused by long-term conditions like cardiovascular disease and vision loss. As the only artificial means of keeping the kidneys functional, dialysis is painful, expensive, and arduous. The risk of death from kidney disease is rising annually, affecting millions of people all over the world, says the World Health Organization. That's why it's critical to have an accurate forecast right away so that any precautions or controls can be implemented without delay.

To eliminate the severe effect of this dangerous kidney functional problem, we have to detect the pattern of diseases first. A number of scholars are working on that issue, particularly on CKD dataset using both statistical and ML algorithms. But ML approaches seems relatively good in decision-making in automatic diagnosis of diseases [2]. Classification algorithms such as Support Vector machine (SVM), Neural Network, Random Forest (RF) and Artificial neural Network (ANN) performs well on CKD datasets [3-4].

In this study we predict kidney disease using several machine learning algorithms and to improve the accuracy of all the algorithm, we propose a pipeline including outlier removal, data normalization, imbalance handling. We also propose an ensemble algorithm by stacking RF, SVM, and LR together and named as stacking RSL ensemble classifier. The remaining portion of this paper is organized as follows, section 2 represents the literature review, proposed mechanism in section 3, and result and discussion in section 4 and section 5 represent the conclusion and future work.

2. **Literature Review.** Predicting CKD contains particular interest to researchers. Most of them use ML traditional algorithms as well as ensemble some of them to increase the accuracy of classification. Also, some hybrid algorithms are used to classify the heart disease from open-source datasets that are available in Kaggle and UCI.

[5] Conducting 4,143,535 adults' data from 35 datasets, Matsushita et al., (2020) developed several "CKD Patches" including albuminuria and eGFR, to improve the prediction of risk of CVD mortality by Systematic COronary Risk Evaluation (SCORE) and atherosclerotic CVD (ASCVD) by the Pooled Cohort Equation (PCE). They noticed an improvement in the performance of prediction with the CKD Patch for ASCVD beyond PCE and CVD mortality beyond SCORE in validation datasets.

[6] Song et al., (2021) intended to compare the performance of ML algorithms and conventional approaches in predicting acute kidney injury (AKI). They utilized independent samples t-test to determine mean differences in area under the curve (AUC) between ML and LR models.

[7] Utilizing machine learning schemes on 600 clinical records diabetic analysis Centre, Sreeji and Balusamy, (2021) examined the initial forecasting performance of severe kidney ailments known as severe renal ailments for diabetic patients. They found that there is a possibility of arriving at a decision with precision of 90.2% for choice based hierarchical categorization.

[8] Shanthakumari and Jayakarthik, (2021) intended to build a model for ML that employs comorbidity and data on drugs and forecasts population prevalence. They employed ML method in combining ensemble learning for predicting CKD with clinical evidence. The results depicts that the proposed Ensemble Support Vector Machine algorithm performed better on CKD datasets than other ensemble approaches.

[9] Ventrella et al., (2021) aimed to predict how frequently a CKD patient may require to be dialyzed to accelerate strategic planning of treatment. For accurately predicting the time span of dialysis need of a CKD patient, they developed a computational model following a supervised ML algorithm. Result reveals that the occurrence of complete renal failure can happen within the one year rather than later having the test accuracy of 94%, sensitivity of 96%, and specificity of 91%.

[10] To detect CKD, Bhaskara and Suchetha, (2021) developed a computationally efficient Correlational Neural Network (CorrNN) learning model and an automated diagnosis technique. The result depicts that the performance of proposed method surpasses the performance conventional methods bearing prediction accuracy of 98.67%.

[11] Almustafa, (2021) employed a number of classifiers to classify a CKD dataset. Using classifiers such as random tree, K-nearest neighbor (K-NN), decision table (DT),

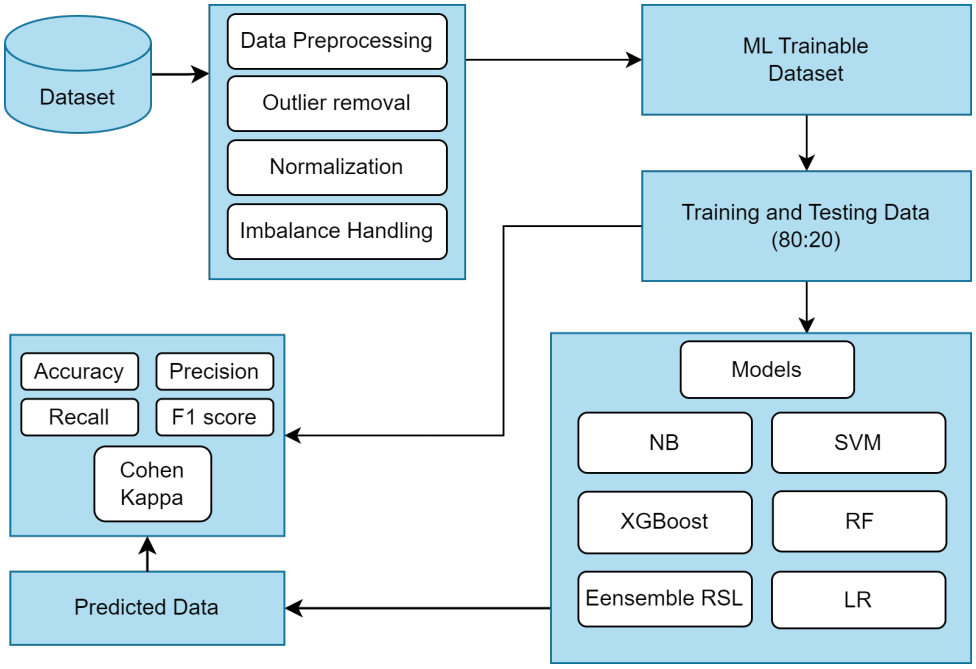

FIGURE 1. Proposed methodology for predicting CKD.

stochastic gradient descent (SGD), J48, and Naïve Bayes, he developed the algorithms. And, based on feature selection, he proposed a prediction model to efficiently forecast CKD cases. Result informs that J48 and decision table classifiers performed well then other classifiers having accuracies of 99%.

3. **Proposed Mechanism.** In this analysis, we use five ML algorithms namely Support Vector Classifier (SVC), Random Forest (RF), Logistic Regression (LR), Stochastic Gradient Descent (SGD), K Nearest Neighbors (KNN) and we ensemble RF, LR, SVC (RSL) and also used SMOTE Tomek imbalance data handling techniques to improve the accuracy of prediction.

Following the removal of any outliers and the use of the normalization procedure, the dataset is randomly divided into training and testing portions. The models are then trained using the training data, and the testing data is used to generate predicted values. The outcomes are satisfactory, however there is room for advancement in terms of increasing the algorithms' degree of precision. We rectify the imbalance in the dataset by the application of imbalance data management strategies. This enables the model to produce more accurate predictions and enhances the models' overall performance.

3.1. **Overview of Proposed Methodology.** From data preparation to final evaluation, total methodology is divided into many subsections. All the subsections are describing in below and the overview is shown in figure 1 by a block diagram. After data preprocessing, the total data analysis is divided into two major parts. CKD is predicted on balanced data by different machine learning algorithms. Data are splitted randomly into 80:20 as training and testing data and both of them using separately in the analysis. For evaluation of the performances of the algorithm's accuracy, precision, recall, f1 score, Cohen Kappa and ROC score are used. The performance of the algorithms is shown using a ROC curve and the results of the algorithms are shown by a bar chart.

3.2. **Data Set Description.** The dataset is used for this analysis is the chronic kidney disease dataset that is collected from Kaggle [ref data]. This dataset contains 25 attributes

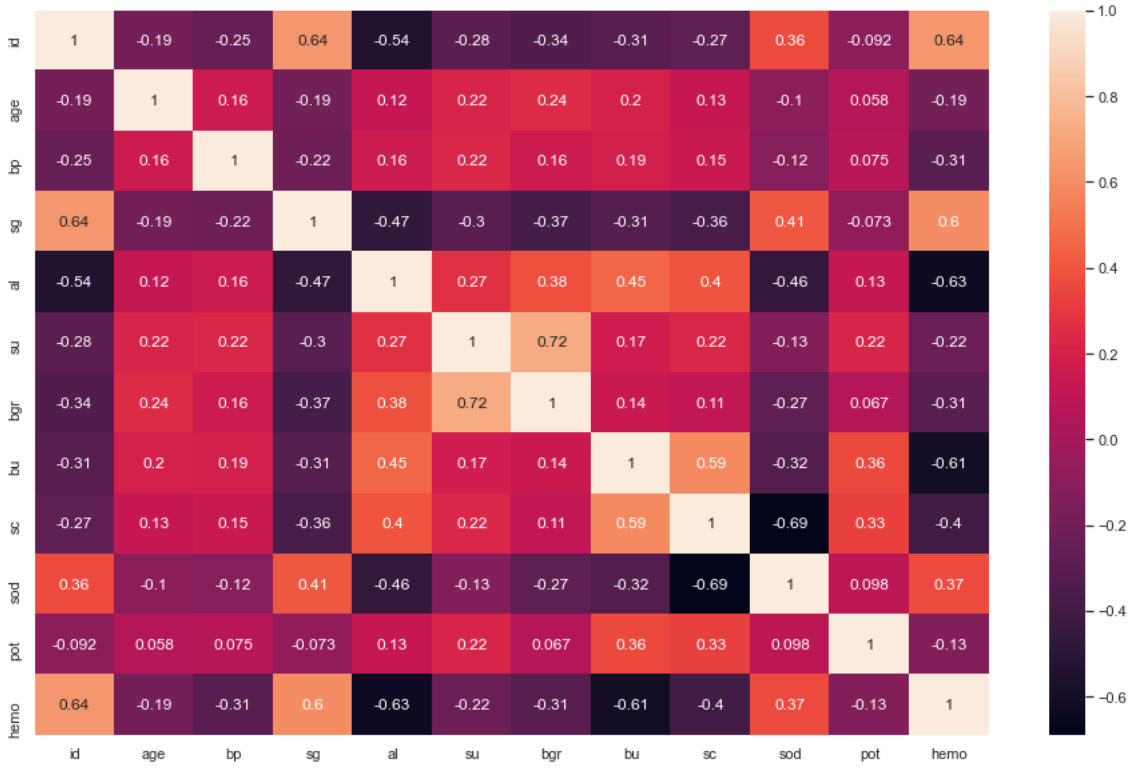

FIGURE 2. Heatmap of the features using Pearson's correlation.

namely age, bp - blood pressure, sg - specific gravity, al – albumin, su – sugar, rbc - red blood cells, pc - pus cell, pcc - pus cell clumps, ba – bacteria, bgr - blood glucose random, bu - blood urea, sc - serum creatinine sod – sodium, pot – potassium, hemo – hemoglobin, pcv - packed cell volume, wc - white blood cell count, rc - red blood cell count, htn – hypertension, dm - diabetes mellitus cad - coronary artery disease, appet – appetite, pe - pedal edema, ane – anemia, class is target. The dataset contains 400 data of kidney disease and it contains many missing values.

3.3. **Correlation among the variables.** The analysis makes use of variables and qualities whose values are associated with one another. In figure 2, the correlation approach developed by Pearson is used to build a heatmap that illustrates the link. The heatmap reveals that there is only a slight association between the different factors.

3.4. **Data preprocessing techniques.** Data preparation deals with converting raw data into a comprehensive form. To run the dataset, the authors preprocessed it to detect missing value, outlier, noisy data and other inconsistences. Popular data preprocessing techniques that are used in this analysis are describe below.

Outlier Detection: This study utilized Turkey fences to detect outliers and extreme values on three distinct quartiles: Q1, Q2, and Q3. The first quartile, often known as Q1, is the value in the data set that contains 25% of the values below it [11]. The third quartile, often known as Q3, is the value that contains 25% of the values above it.

$$lowerlimit = Q1–1.5(Q3–Q1)$$

$$upperlimit = Q3 + 1.5(Q3–Q1)$$

Any values that crosses these upper and lower limit is considered as outlier.

Normalization: In this study, original data is transformed linearly using min-max normalization (range normalization). Assume that minA and maxA are the lowest and

highest values of some attributes A, respectively. In the range [minA, maxA], min-max normalization maps a value of attribute A, d to d'.

$$d' = \frac{d - min_A}{max_A - min_A}$$

Imbalance Data Handling: A dataset is considered as imbalanced dataset when the classification categories of a dataset are not approximately equal. We use SMOTE Tomek to handle the imbalance problem. SMOTETomek is a hybrid approach that employs oversampling and under-sampling techniques to leap up the performance of the classifier model. To ensure a balanced distribution, the SMOTE technique is first utilized to oversample the minority class, after which samples from the majority classes are detected and removed from Tomek Links.

3.5. **Description of algorithms.** In this analysis, we employ five ML algorithms such as Support Vector Classifier (SVC), Random Forest (RF), Logistic Regression (LR), k-Nearest Neighbor (k-NN), and Stochastic Gradient Descent (SGD). Short description of each algorithm is described below:

Support Vector Classifier (SVC): SVC refers to a model which is used to fix pattern recognition problems such as outlier detection and classification. It utilizes the idea of decision planes that apply decision boundaries to optimally distinct data into numerous categories. SVC is relatively better efficient in high dimensional spaces and memory efficient. But, SVC shows severe performance when there is existence of noise in data.

Random Forest (RF): RF is a classifier which contains a range of decision trees on different subsets of a given dataset and employs the average to increase the predictive accuracy of that dataset. RF predicts from every tree and based on the majority votes it predicts the final output rather than relying on single decision tree. For the classification problem, the variables are ranked through their importance. The greater the number of trees in the forest the better the accuracy which limits the negative effect of overfitting.

Logistic Regression (LR): LR is a well-performing supervised ML approach for predicting the likelihood of a binary outcome. In logistic regression regularization is necessary to minimize overfitting, especially when there are a limited number of training instances or a high number of parameters to learn. Logistic regression can be applied for classifying multiclass problems. LR cannot solve non-linear problems as it contains linear decision surface.

K-Nearest Neighbors (k-NN): k-NN is a commonly used classification algorithm which is employed in different applications. K-NN is developed based on the concept that the forecasted value of the example is perhaps similar to those of neighbors. The k-NN algorithm describes a metric in the predictor's vector space, plots all applicants to a position in this space and assesses subsequent probability through the relative amount of good risks among the k-nearest points in the training set.

Stochastic Gradient Descent (SGD): SGD is a simple efficient optimization algorithm utilized to determine the values of coefficients of functions which lessen a cost function. It can be applied to big datasets because the update to the coefficients is executed for every training case, rather than at the end of examples.

Stacking Support Vector Machine, Random Forest, and Logistic Regression (RSL): The term "ensemble approach" refers to a group of methods that combine the

best features of numerous learning algorithms or models into a single predictive algorithm. Overall, the model's performance exceeds that of the individual basis learners. We use stacking ensemble procedure to generate the new model named RSL. The SVC is used as the base model that ensemble with another two algorithm LR and RF and finally construct ensemble RSL. It improves the accuracy and the classification reports of the classification with the help of the three algorithms.

3.6. **Performance measure techniques.** To determine the performance of ML and DL algorithms, this study employed eight performance measure techniques such as Accuracy, Precision, Recall, F-1 Score, Specificity, Cohen Kappa, AUC, and ROC.

Accuracy: The number of correctly classified data instances divided by the total number of data instances are called accuracy. Although accuracy is one of the most basic performance measures, it can sometimes provide false outcome, especially for imbalanced dataset. Mathematically, $Accuracy = (TP + TN)/(TP + TN + FP + FN)$

Precision: Precision for binary classification is defined as the number of TP divided by the number of TP and FP. Precision performs precisely on imbalance data when the goal is to reduce FP. Even, whether the rate of FP is high, precision is a good metric to use. Mathematically, $Precision = TP/(TP + FP)$

Recall: Recall is also known as True Positive Rate (TPR) or Sensitivity. Generally, recall is calculated as the number of TP divided by the number of TP and FN. In case of reducing FN from imbalanced dataset, recall is appropriate. Mathematically, $Recall = TP/(TP + FN)$

F-1 Score: The harmonic mean of Precision and Recall is called F-1 score. To identify the model is applicable or not, only accuracy is not enough. The model will make sense only when both Precision and Recall are high. That is why, F1-Score is calculated to compare two classifiers' performance. Its range is in between [0, 1] and higher the value of f-1 scores, a more sensible model we get. Mathematically, $F - 1Score = 2PR/(P + R)$

Cohen Kappa: We occasionally encounter a multi-class classification or an unbalanced dataset. In those circumstances, metrics like accuracy, precision or recall sometimes don't give us accurate performance. Cohen's kappa statistic is an excellent metric for dealing with both multi-class and imbalanced class issues. Mathematically, $CohenKappa = (Po–Pe)/(1–Pe)$

Where, Po is the observed agreement and Pe is the expected agreement

Receiver Operating Characteristics (ROC): ROC curve is a technique for visualizing, organizing and selecting classifiers based on their performance. It is a probability curve that plots the FPR on the X-axis and the TPR on the Y-axis at various threshold values.

4. **Results and Analysis.** Most of the algorithm perform well to classify the kidney disease using the secondary dataset. Performance of the algorithms are tabulated below based on different performance measure techniques. In dataset the performance of SVC, KNN and RF is 98%, LR and SGD are 97%, proposed ensemble RSL shows 99% accuracy after handling the imbalance of the dataset using SMOTETomek imbalance data handling techniques. Besides the accuracy the precision, recall, f1 score of our proposed method is better than other benchmark algorithms. SVC and KNN shows 98% precision, LR and

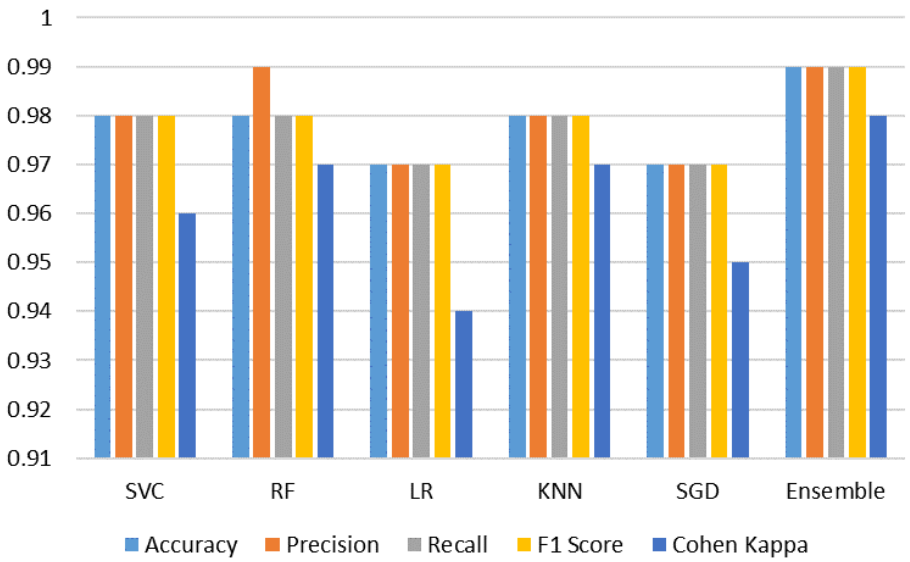

FIGURE 3. Accuracy and classification matrices report of different models.

SGD shows 98% precision, RF and ensemble RSL shows 99% precision. All the accuracy including precision, recall, f1 score, sensitivity, Cohen Kappa and AUC score is tabulated below.

TABLE 1. Accuracy and scores of models on datasets.

Methods	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Cohen Kappa**
SVC	0.98	0.98	0.98	0.98	0.96
RF	0.98	0.99	0.98	0.98	0.97
LR	0.97	0.97	0.97	0.97	0.94
KNN	0.98	0.98	0.98	0.98	0.97
SGD	0.97	0.97	0.97	0.97	0.95
Ensemble	0.99	0.99	0.99	0.99	0.98

The table 1 represent the values of different performance measure techniques on balanced data using SMOTETomek.

Fig.3 is representing the accuracy, precision, recall, f1 score, Cohen kappa of different algorithms from the values that are represented in table 1. The performance is also show by a ROC curve that is in fig 4. Proposed ensemble RSL shows better performance than others. Due to overlap of the accuracy the curve is now shown properly.

5. **Conclusion and Future Work.** Kidney disease is one of the major diseases of human body. Early-stage production can give a better change of solving this problem by the medical experts. This study proposes an ensemble RSL model that can predict the kidney disease more accurately than benchmark machine learning algorithms. The proposed RSL shows the superiority than another classifier in terms of accuracy, precision, recall, f1 score and Cohen kappa score. Our proposed ensemble RSL helps the medical domain analyst to predict kidney disease more accurately than the previous ways.

Health care is another important area where DL algorithms play an important role. It's possible that using DL algorithms will improve the outcome. Dimensionality reduction

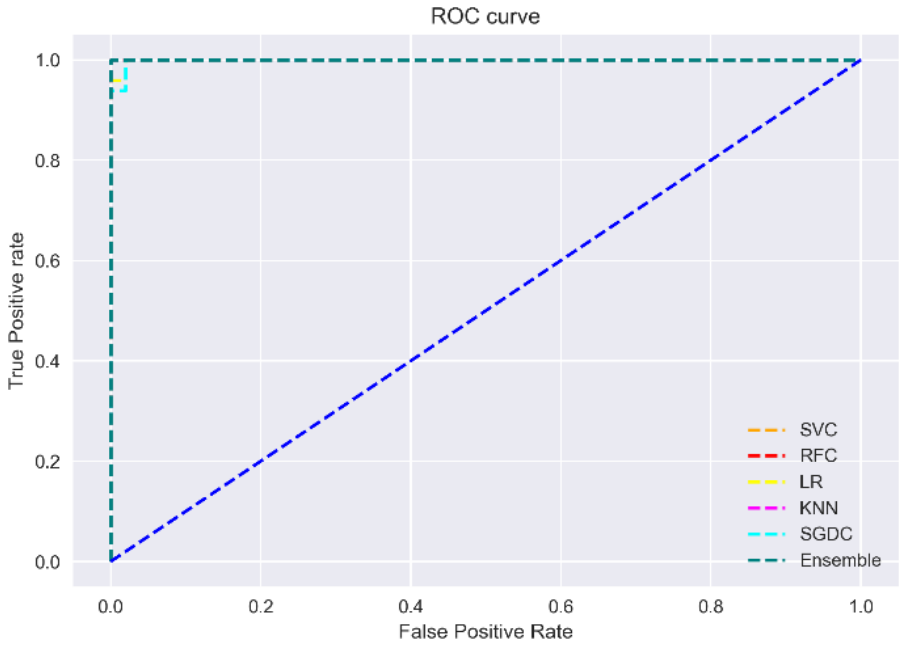

FIGURE 4. ROC curve of different classifiers using SMOTETomek.

and feature selection could be useful in this field to produce more accurate results. It's possible that using a combination of methods (such as a hybrid or an ensemble) would result in more precise forecasts. We'd like to categorize the illness as a multiclass problem so that we can determine its severity.

REFERENCES

[1] Arora, M., and Sharma, E. A. (2016). Chronic kidney disease detection by analyzing medical datasets in weka. International Journal of Computer Application, 6(4), 20-26.

[2] Bhaskar, N., and Suchetha, M. (2021). A computationally efficient correlational neural network for automated prediction of chronic kidney disease. IRBM, 42(4), 268-276.

[3] Huang, J., Jin, T., Liang, M., and Chen, H. (2021). Prediction of heat exchanger performance in cryogenic oscillating flow conditions by support vector machine. Applied Thermal Engineering, 182, 116053.

[4] Zhou, Y., Yu, Z., Liu, L., Wei, L., Zhao, L., Huang, L., and Sun, S. (2022). Construction and evaluation of an integrated predictive model for chronic kidney disease based on the random forest and artificial neural network approaches. Biochemical and Biophysical Research Communications, 603, 21-28.

[5] Matsushita, K., Jassal, S. K., Sang, Y., Ballew, S. H., Grams, M. E., Surapaneni, A., ... and Coresh, J. (2020). Incorporating kidney disease measures into cardiovascular risk prediction: Development and validation in 9 million adults from 72 datasets. EClinicalMedicine, 27, 100552.

[6] Song, X., Liu, X., Liu, F., and Wang, C. (2021). Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. International Journal of Medical Informatics, 151, 104484.

[7] Balusamy, B. (2021). A novel algorithm for prediction of chronic kidney risks using machine learning schemes. Materials Today: Proceedings.

[8] Shanthakumari, A. S., and Jayakarthik, R. (2021). Utilizing support vector machines for predictive analytics in chronic kidney diseases. Materials Today: Proceedings.

[9] Ventrella, P., Delgrossi, G., Ferrario, G., Righetti, M., and Masseroli, M. (2021). Supervised machine learning for the assessment of chronic kidney disease advancement. Computer Methods and Programs in Biomedicine, 209, 106329.

[10] Bhaskar, N., and Suchetha, M. (2021). A computationally efficient correlational neural network for automated prediction of chronic kidney disease. IRBM, 42(4), 268-276.
[11] Almustafa, K. M. (2021). Prediction of chronic kidney disease using different classification algorithms. Informatics in Medicine Unlocked, 24, 100631.