IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 10, 2021, accepted January 15, 2021, date of publication January 22, 2021, date of current version February 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3053763

Prediction of Chronic Kidney Disease - A Machine Learning Perspective

PANKAJ CHITTORA[1], SANDEEP CHAURASIA[1], (Senior Member, IEEE), PRASUN CHAKRABARTI[2,3], (Senior Member, IEEE), GAURAV KUMAWAT[1], TULIKA CHAKRABARTI[4], ZBIGNIEW LEONOWICZ[5], (Senior Member, IEEE), MICHAŁ JASI SKI[5], (Member, IEEE), ŁUKASZ JASI SKI[5], RADOMIR GONO[6], (Senior Member, IEEE), ELŻBIETA JASI SKA[7], AND VADIM BOLSHEV[8]

[1]Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur 303007, India
[2]Department of Computer Science Engineering, Techno India NJR Institute of Technology, Udaipur 313003, India
[3]Data Analytics and Artificial Intelligence Laboratory, Engineering-Technology School, Thu Dau Mot University, Thu Dau Mot 820000, Vietnam
[4]Department of Basic Science (Chemistry), Sir Padampat Singhania University, Udaipur 3136022, India
[5]Department of Electrical Engineering Fundamentals, Faculty of Electrical Engineering, Wroclaw University of Science and Technology, 50-370 Wroclaw, Poland
[6]Department of Electrical Power Engineering, Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava, 708 00 Ostrava, Czech Republic
[7]Faculty of Law, Administration and Economics, University of Wroclaw, 50-145 Wroclaw, Poland
[8]Laboratory of Power Supply and Heat Supply, Federal Scientific Agroengineering Center VIM, 109428 Moscow, Russia

Corresponding author: Michał Jasi ski (michal.jasinski@pwr.edu.pl)

This work was funded by the Chair of Electrical Engineering Fundamentals (K38W05D02), Wroclaw University of Technology, Wroclaw, Poland.

ABSTRACT Chronic Kidney Disease is one of the most critical illness nowadays and proper diagnosis is required as soon as possible. Machine learning technique has become reliable for medical treatment. With the help of a machine learning classifier algorithms, the doctor can detect the disease on time. For this perspective, Chronic Kidney Disease prediction has been discussed in this article. Chronic Kidney Disease dataset has been taken from the UCI repository. Seven classifier algorithms have been applied in this research such as artificial neural network, C5.0, Chi-square Automatic interaction detector, logistic regression, linear support vector machine with penalty L1 & with penalty L2 and random tree. The important feature selection technique was also applied to the dataset. For each classifier, the results have been computed based on (i) full features, (ii) correlation-based feature selection, (iii) Wrapper method feature selection, (iv) Least absolute shrinkage and selection operator regression, (v) synthetic minority over-sampling technique with least absolute shrinkage and selection operator regression selected features, (vi) synthetic minority over-sampling technique with full features. From the results, it is marked that LSVM with penalty L2 is giving the highest accuracy of 98.86% in synthetic minority over-sampling technique with full features. Along with accuracy, precision, recall, F-measure, area under the curve and GINI coefficient have been computed and compared results of various algorithms have been shown in the graph. Least absolute shrinkage and selection operator regression selected features with synthetic minority over-sampling technique gave the best after synthetic minority over-sampling technique with full features. In the synthetic minority over-sampling technique with least absolute shrinkage and selection operator selected features, again linear support vector machine gave the highest accuracy of 98.46%. Along with machine learning models one deep neural network has been applied on the same dataset and it has been noted that deep neural network achieved the highest accuracy of 99.6%.

INDEX TERMS Chronic kidney disease, machine learning, prediction.

I. INTRODUCTION

Chronic kidney Disease (CKD) means your kidneys are damaged and not filtering your blood the way it should. The primary role of kidneys is to filter extra water and waste from your blood to produce urine and if the person has suffered from CKD, it means that wastes are collected in the body. This disease is chronic because of the damage gradually over a long period. It is flattering a common disease worldwide [1]. Due to CKD may have some health troubles. There are many causes for CKD like diabetes, high blood pressure, heart disease. Along with these critical diseases, CKD also

The associate editor coordinating the review of this manuscript and approving it for publication was Haruna Chiroma.

depends on age and gender [2]. If your kidney is not working, then you may notice one or more symptoms like abdominal pain, back pain, diarrhea, fever, nosebleeds, rash, vomiting. There are two main diseases of CKD: (i) diabetes and (ii) high blood pressure [3]. So that controlling of these two diseases is the prevention of CKD. Usually, CKD does not give any sign till kidney is damaged badly. CKD is being increased rapidly as per the studies hospitalization cases increase 6.23 per cent per year but the global mortality rate remains fixed [4]. There are few diagnostic tests to check the condition of CKD: (i) estimated glomerular filtration rate(eGFR) (ii) urine test (iii) blood pressure.

A. EGFR

eGFR value shows that how your kidney cleaning the blood. If your eGFR value is greater than 90, that means the kidney is normal. If eGFR value is less than 60, that means you have CKD [5].

B. URINE TEST

The doctor also asks for urine test for kidney functionality because kidneys make urine. If the urine contains blood and protein [6], that means your kidney is not working properly.

C. BLOOD PRESSURE

Doctor measures blood pressure as Blood pressure range shows how your heart is pumping blood. If eGFR value reaches less than 15, that means the patient has end-stage kidney disease. At this point, there are only available treatments: (i) dialysis and (ii) kidney transplant. Patient's life after dialysis depends on such factors as age, gender, frequency and duration of dialysis, physical movement of the body and mental health [7]. If dialysis is not possible, the doctor has only one solution, i.e., kidney transplantation. However, it is extremely expensive [8].

Therefore, it is critical noteworthiness in early recognition, monitoring and handling of the disease. It is essential to predict the striding of CKD with appropriate accuracy due to its dynamic and secretive nature in the early stages and patient abnormality. Medical treatment of CKD is prescribed by the stage. Anything other than this, it is very imperative to characterize the organization of the infection because it gives a few indications. It underpins the assurance of fundamental intercessions and medications.

Medical treatment is a very significant application area of intellectual intelligent systems [10]. Afterwards, Data mining can play a big role to find out hidden information from the huge patient medical and treatment dataset that doctors frequently obtain from patients to get pieces of knowledge about the symptomatic data and to execute precise treatment plans. Data mining can be categorized as the method of extracting hidden information from a huge dataset. Data mining strategies are connected and utilized broadly in various contexts and areas. Using data mining methods, we may predict, classify, filter and cluster data. The objective states the algorithm processing of a training set containing a set of attributes and targets. Data mining is suitable to mining in data if the dataset is huge but we can also do it with the help of machine learning with a small dataset. The machine learning can also find data analysis and pattern detection [9]. A variety of health dataset is present so machine learning algorithms are best fit to improve the accuracy of diagnosis prediction [11]. As healthcare electronic dataset grows rapidly, machine learning algorithms are becoming more common in healthcare. [12].

Qin *et al.* [13] proposed data assertion and sample diagnosis achievable in CKD diagnosis. KNN is used for data assertion. Six classifiers algorithms used for accuracy of diagnosis: logistic regression, random forest, support vector machine, K-nearest neighbor, naive Bayes classifier and feed-forward neural network. In these classifiers random forest gives better accuracy, i.e., 99.75%.

Vasquez-Morales *et al.* [14] developed a neural network model for risk prediction of Chronic Kidney Disease development on 40000 instances dataset and their model accuracy was 95%.

Chen *et al.* [15] applied three models on the dataset that is provided by UCI. They used KNN, SVM and soft independent modelling of class analogy (SIMCA) for finding the risk calculation of patient using these classifiers. In which the SVM and KNN model attained, the best accuracy of 99.7% and SVM model has the greatest capability to endure noise disturbance.

Because CKD is invasive, costly so that many patients reached at last stages without treatments. So that early detection of this disease remains important. Besides, Amirgaliyev [16] gave the experimental result of SVM machine learning classifier algorithm with accuracy 93%.

Padmanaban and Parthiban [17] suggested that the early detection of CKD for diabetic patients with the help of machine learning classifiers algorithms. They collected data from Chennai based diabetes research center and applied Naive Bayes and Decision tree on the dataset. For finding the accuracy they used Weka tool and concluded that Naïve Bayes classifier achieved the highest accuracy of 91%.

de Almeida *et al.* [18] in their work applied Decision tree, Random Forest, Support Vector Machine (SVM) and also used SVM with linear, polynomial, sigmoid and RBF functions. For their research, they used the MIMIC-II database. They concluded that random forest and Decision tree got the best result in the form of prediction accuracy of 80% and 87% respectively.

Gunarathne *et al.* [19] built a model of various machine learning classifiers algorithm and analysis of which algorithm is best suited to the dataset. They used dataset provided by UCI containing 400 instances and 14 attributes. They concluded that the Multiclass decision forest algorithm was best fitted for the CKD dataset with an accuracy of 99.1%.

Polat *et al.* [20] used SVM algorithm for CKD prediction. For the accurate result, they worked on an important feature. For selecting the correct feature, they used two-approach Wrapper and filter with the SVM algorithm. In the Wrapper, there were the greedy stepwise search engine for classifier

subset evaluator and best first search engine for Wrapper subset evaluator. In filter, there were the greedy stepwise search engine for correlation feature section subset feature and best first search engine for filtered subset evaluator. The results of all techniques were compared and it was found that SVM gave the highest accuracy with filtered subset evaluator, i.e. 98.5%.

Sujata Drall, Gurdeep Singh Drall, Sugandha Singh, Bharat Drall *et al.* [21] worked on CKD dataset given by UCI with 400 instances and 25 attributes. Firstly, data was preprocessed, the missing data was found, filled with 0, then transformed and applied on the dataset. After preprocessing, authors applied algorithm for important attributes and found 5 most important features and then the classification algorithm: Naïve Bayes and K-Nearest Neighbor. The gotten result KNN achieved the highest accuracy.

Almasoud and Ward [22] worked with CKD dataset of 400 instances and 25 attributes. They applied the filter feature selection method on attributes and found that haemoglobin, albumin and specific gravity are feature attributes in CKD dataset. After feature selection, they trained the dataset and validated with 10-fold cross-validation. The gradient boosting algorithm achieved the highest accuracy of 99.1%.

Shankar *et al.* [23] applied three steps on the same UCI dataset: (i) data preprocessing & feature selection (ii), algorithms' accuracy determination and (iii) diet plan suggestion. In the feature selection method, they applied two approaches: one is the Wrapper and the other is the LASSO method. After the feature selection method, 4 classification algorithms were applied: Logistic Regression, Random forest tree K-Nearest Neighbors, Neural Network and Wide and Deep Learning. For diet plan suggestion blood potassium level was used. The blood potassium level was divided into three groups based on its value: Safe Zone, Caution Zone and Danger zone.

Vijayarani and Dhayanand [24] collected kidney function test (KFT) dataset from medical labs, research centres and hospitals. The dataset contained 584 instances and 6 attributes and two classifier applied algorithms: support vector machine (SVM) and artificial neural network (ANN). It was found that ANN achieved the highest accuracy of 87.7%.

Xiao *et al.* [25] used the data of 551 patients and applied 9 machine learning algorithm: XGBoost, logistic regression, lasso regression, support vector machine, random forest, ridge regression, neural network, Elastic Net and K-nearest neighbor. They evaluated accuracy, ROC curve, precision and recall and found that linear model gave the highest accuracy.

Reshma *et al.* [31] used the feature selection technique on CKD Dataset. For feature selection, ACO method was applied. ACO is the meta heuristic algorithm for the feature selection. It is the type of Wrapper method. In their dataset, total 24 attributes were available. After applying feature selection algorithm, 12 features was used for making the model. The Support Vector machine classifiers algorithm was used for building the model.

Deepika *et al.* [32] built a project on prediction of Chronic Kidney Disease based on old dataset of CKD. The dataset had 24 attributes and 1 target variable. For building the model, they applied KNN and Naïve Bayes supervised machine learning algorithm. KNN achieved highest accuracy 97 % and Naïve Bayes achieved 91% accuracy.

Ma *et al.* [33] proposed the deep learning algorithm for predicting the Chronic Kidney Disease s at early stage. The deep neural network was built from Heterogeneous Modified artificial neural network algorithm. For building the model, ultrasound images were used. For comparing the result, there were three different classifiers: Support Vector machine, artificial neural network and multilayer perceptron.

UI Haq *et al.* [34] proposed the machine learning model to predict the diabetes disease at early stage. They concluded that machine learning can play vital role in the healthcare.

Amin *et al.* [35] proposed machine learning model for the prediction of Parkinson's disease at early stage. For building the model, they used SVM classifier. Feature selection algorithms were also applied for extract the important features: Relief and ACO feature selection algorithm.

This research article primarily aims to predict whether a person has Chronic Kidney Disease or not. In this perception, seven different machine learning classifiers were applied on the dataset. All the algorithms were running with both full features and selected features. SMOTE was used for oversampling and all the results were recorded. All the machine learning model results were also compared with one deep neural network algorithm. Deep learning neural network was used with two hidden layers. IBM SPSS Modeler was applied for computational purpose. The contribution reveals the accuracy estimate of 99.6% when applying deep neural network on the dataset.

II. RESEARCH GAP

Until now, in majority of cases full features have been taken into consideration. In this research, feature optimization was carried out, wherein three different feature selection algorithms were applied to find the algorithm most beneficial to extract the important feature for the prediction of Chronic Kidney Disease. As many datasets have imbalanced class, class balancing is needed for increasing the performance of classifier model. In this research SMOTE was used as a class balancer. The highest accuracy of 99.6% was achieved whereas the article [22] provides an accuracy of 99.1% on the same dataset. According to [15] the highest accuracy of the model was 99.7%, but they worked on risk calculation of the patient whereas the main aim of the article is to predict Chronic Kidney Disease.

III. METHODS AND MATERIALS

In this section, the research methodology and a dataset will be discussed.

A. DATASET

Chronic Kidney Disease dataset is used for this research work. Many researchers had also used this dataset [26]. This dataset is being provided by the UC Irvine Machine

Learning Repository and it is available on the UCI website. This dataset contains 400 instances and 24 attributes with 1 target attribute. The target attribute has labelled in two-class to represent CKD or non-CKD. The dataset was collected from various hospitals in 2015. It contains also missing value. The description of all 24 attributes is represented in the table 1 below.

B. METHODOLOGY

In this research, we have developed a model to predict CKD disease in patients. The performance of the model was tested on both all attributes and selected features. Among feature selection methods there were Wrapper, Filter and Embedded [27] allowing to select vital features. Classifier algorithms performance was tested on the selected features. IBM SPSS tool is used for preparing the model. The machine learning classifiers such as artificial neural network (ANN), C5.0, logistic regression, linear support vector machine (LSVM), K- nearest neighbors (KNN) and random tree were used for training the model. Each classifier validation and performance matrix were computed. The procedure of this research including five stages: (i) dataset preprocessing, (ii) feature selection, (iii) classifier application, (iv) SMOTE and (v) analyzing the performance of the classifier. Along with machine learning models, a deep neural network was applied for comparing the result of machine learning models and deep neural network. Artificial Neural network classifier was used for this purpose. In this research the significance of two model were checked by statistic testing namely McNemar's test.

C. PREPROCESSING OF DATA

Data preprocessing could be a strategy that is utilized to change over the raw information into a clean dataset. It is a

the basic step to train every machine learning classifier algorithm. This technique concludes such actions as handle missing values, rescaling of the dataset, transform into binary data and standardize of the dataset. When the dataset included attributes with varying scales, rescaling is used to scale the dataset. The binary transformation has been applied to convert the value into 0 and 1. All values of every attribute are considered as 1 for above the threshold and as 0 for below the threshold. Standardized method ensures that each attribute has mean 0 and standard deviation 1.

D. FEATURE SELECTION

Feature selection is needed for trained each machine learning classifier because without removing unnecessary attributes from the dataset result may be affected. The classifier algorithm with feature selection gives better performance and reduce the execution time of the model. For this process, three different feature selection methods were used in this research.

1) FILTER METHOD

The filter is one of the methods to select the appropriate feature. It selects the feature on their integral features without integrating any learning classifier algorithm. This method gives result faster as compared to the wrapper method. The method assigns the score to every attribute based on their statistical correlator between attributes. There are many filter methods are available, but Correlation-based Feature Selection (CFS) method has been used. CFS is the algorithm to select the feature-based on the attribute ranks. It assigns the rank to attribute subset as based on the correlation heuristic evaluation function [28]. The function works on the strategy that creates two class labels, one is correlated to class and low correlated class and selects only correlated label class attributes.

2) WRAPPER METHOD

Wrapper method selects the subset of features based on a precise machine learning algorithm [29]. It used the greedy search method for finding a possible subset of features. The method can be implemented with using any of the following algorithms forward selection, backward elimination and recursive elimination. In the research, we used the forward feature selection method. The forward feature selection iteratively selects the feature. This procedure starts with the null model and work iteratively and add the attribute in each step. The attribute is keeping add in the model until the attribute does not improve model performance.

3) EMBEDDED METHOD

The embedded method is decision tree algorithm for feature selection. It selects the feature in each step works recursively while the tree is growing and split the sample set into a smaller subset. The most common decision tree algorithm are: ID3, C4.5 and CART. There are other available method s creating linear models. The most common methods are LASSO [30] with L1 penalty and Ridge with L2 penalty. In this research LASSO (least absolute shrinkage and selection operator) algorithm has been used. It performs two main tasks: regularization and feature selection. In regularization, it shrinks some feature coefficients to zero that means features are not important for the predictor model.

E. CLASSIFICATION ALGORITHMS

Classification technique is an important feature of supervised learning. Classifiers learn from the training dataset and apply on the testing dataset for finding the target attribute. Below there are classification techniques used in research.

1) ARTIFICIAL NEURAL NETWORK

Artificial neural network is a part of artificial intelligence. It is a type of supervised machine learning. Its structure is the same as the human brain. ANN also have neurons and just like in human all neurons are interconnected to one another, ANN neurons are connected to each other in layers of the network. Neurons there are known as nodes. ANN can solve the problem that has been impossible for human or statistical standards. ANN consists of three layers: input, hidden and output layers. The input layer takes input and weight and passes to hidden layer for performing calculation

TABLE 1. Description of Attributes in the Dataset.

Sr. No	Attribute Name	Description
1	Age	Patient age (It is in years)
2	Bp	Patient blood pressure (It is in mm/HG)
3	Sg	Patient urine specific gravity
4	Al	Patient albumin ranges from 0-5
5	Su	Patient sugar ranges from 0-5
6	Rbc	Patient red blood cells two value normal and abnormal
7	Pc	Patient pus cell two value normal and abnormal
8	Pcc	Patient pus cell clumps two values present and not present
9	Ba	Patient bacteria two values present and not present
10	Bgr	Patient blood glucose random in mg/dl
11	Bu	Patient blood urea in mg/dl
12	Sc	Patient serum creatinine
13	Sod	Patient sodium
14	Pot	Patient potassium
15	Hemo	Patient hemoglobin (protein molecule in red blood cells)
16	Pcv	Patient packed cell volume % of red blood cells in circulating blood
17	Wc	Patient white blood cell counts in per microliter
18	Rc	Patient red blood cell count in million cells per microliter
19	Htn	Patient hypertension two value Yes and No
20	Dm	Patient diabetes mellitus two value Yes and No
21	Cad	Patient coronary artery disease two value Yes and No
22	Appet	Patient appetite two value good and poor
23	Pe	Patient pedal edema two value Yes and No
24	Ane	Patient anemia two value Yes and No
25	Class	Target Variable (CKD or Not)

and finding the hidden structures and patterns. The number of hidden layers can be increased as required. The output layer computes the output. The weight values from the output, i.e. predicted, and actual value were recomputed and the network again restarts for finding the class from the previous learning. Therefore, ANN works based on backpropagation.

2) C5.0

C5.0 is a type of decision tree because it creates the decision tree from the input. The tree has the number of branches. It utilizes the tree structure to model the relationship between features and potential outcomes. At each node of the tree, the attribute of the dataset is chosen. It can handle nominal and numeric features both. C5.0 is the extended version of the C4.5 classification algorithm and uses information entropy concept. Entropy is used for finding the impurity of features. Information entropy is produced based on the calculation of parent and child entropy values. This process is iterative and works until there is no the further split.

3) LOGISTIC REGRESSION

Logistic regression is also a type of supervised learning algorithm. It is a statistical model. The probability of target value is predicted from logistic regression. It is divided the target attribute into two-classes: success or not success. For success, it returns 1 whereas it returns 0 for not succeeding. Logistic regression is represented by equation 1:

$$\boldsymbol{P = 1/(1 + e\text{^}(-(b_0 + b_1 x + b_2 x\text{^}2))} \quad (1)$$

where P is the predicted value, b_0, b_1, b_2 are biases and x is is an attribute. It is used in various field of machine learning application in social sciences and medical arena, for example, for spam detection, diabetes detection, cancer detection, etc. Logistic regression is the advanced version of linear regression. Through this technique, we only concern about the probability of the outcome variable.

4) CHAID

Chi-square automatic interaction detection (CHAID) is a type of decision tree technique. It is used to determine the relationship between variables. Nominal, ordinal and continuous data can be used in CHAID for finding the outcome. For each categorical predictor, all possible cross-tabulation is created in the CHAID model and it process works until the best outcome is attained. The target or dependent variable becomes a root node in the tree, the target variable is split into

two or more parts as per the categories in target variable and child of the root node are created using the statistical method and variable relationship. Such a process will be till leaf nodes of the tree. Ftest is used for the continuous dependent variable and the Chi-square test is used for the categorical dependent variable.

5) LINEAR SUPPORT VECTOR MACHINE (LSVM)

linear support vector machine (LSVM) is the modern particularly fast machine learning algorithm for solving multiclass classification problem for the large dataset based on a simple iterative approach. It is created the SVM model in linear CPU time of the dataset. LSVM can be used for the high dimensional dataset is the sparse and dense format. It is used for solving the large dataset machine learning problems in less expensive computing resource. Support Vector Machine is a supervised classifier algorithm. It is used kernel trick for solving the classification problem. Based on these transformations, ideal edge is found between the possible outputs. SVM is used for the nonlinear kernel, such as RBF. For the linear kernel, LSVM is an appropriate choice. LSVM classifier is sufficient for all linear problems.

6) K- NEAREST NEIGHBORS (KNN)

KNN is a simple type of supervised algorithm. It can be used for both classification and regression problems. However, it is largely used for classification problems. KNN does not use a particular training stage and use all the data for training so that it is a lazy learning algorithm and also it does not consider anything about the underlying data, so that is a non-parametric learning algorithm. KNN stores the whole dataset because it has no model so that there is no learning required. When the new data enter for predicting the outcomes, it compares K – neighbors so that selection of K's value is very important. The distance is calculated between two already label data. The distance helps to find the nearest neighbor of the new data. A Euclidian method is used for finding the distance.

7) RANDOM TREE

The random tree is a type of supervised classifiers. It produces lots of distinct learners. The stochastic process is used to form the tree. It is a type of ensemble learning technique for classification. It works the same as decision tree, but a random subset of attributes uses for each split. This algorithm uses for both classification problems and regression problems. A group of random trees is known as a forest. The random trees classifier takes the input feature set and classifies input for every tree in the forest. The output of the random tree selects from the majority of votes. In the tree, every leaf node holds a linear model. The bagging training algorithm is used to train the model.

F. VALIDATION METHOD OF CLASSIFIERS

The dataset was divided into parts: training dataset and testing dataset. IBM SPSS modeller was used for the partition and prediction of the result. The training dataset contains 50% of the data and remaining data is considered as the testing data. The type tool of IBM SPSS was applied for changing the type of attributes. The performance evaluation matrix was received for each classification algorithm.

TABLE 2. Confusion Matrix.

	Positive (1)	Negative (0)
Positive (1)	TP	FN
Negative (0)	FP	TN

G. PERFORMANCE EVALUATION MEASURE

Various evaluation matrices were used for checking the performance of the classifier. For this purpose, the confusion matrix was used. It is a 2*2 matrix due to two classes in the dataset. The confusion matrix gives two types of correct prediction of the classifier and two types of incorrect prediction of the classifier. The confusion matrix is presented in Table 2.

1) CONFUSION MATRIX DESCRIPTION

TP: True Positive means output as positive such that predicted result is correctly classified.

TN: True Negative means output as negative such that predicted result is correctly classified.

FP: False Positive means output as positive such that predicted result is incorrectly classified.

FN: False Negative means output as negative such that predicted result is incorrectly classified.

2) CLASSIFICATION ACCURACY

Classification accuracy shows the correct rate of prediction results. It computes from the confusion matrix. The classification accuracy is found by equation 2:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{2}$$

3) CLASSIFICATION ERROR

Classification error shows the incorrect rate of prediction results. It computes from the confusion matrix. The classification error is found by equation 3:

$$Error = \frac{FP + FN}{TP + TN + FP + FN} * 100 \tag{3}$$

4) PRECISION

Precision is an important model performance evaluation matrix. It is the fraction of related instances among the total retrieved instances. It is a positive predicted value. The precision is calculated as follows in equation 4:

$$Precision = \frac{TP}{TP + FP} * 100 \tag{4}$$

5) RECALL

Recall is also an important model performance evaluation matrix. It is the fraction of related instances among the total number of retrieved instances. The recall is calculated as follows in equation 5:

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100 \tag{5}$$

6) F-MEASURE

It is also known as F Score. F-measure is calculated so as to measure the accuracy of test. It is calculated from the precision and recall by equation 6:

$$F - Measure = 2^* \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

7) ROC AND AUC

The performance of the classification model is measured from the Receiver operating a characteristic curve (ROC). ROC is a graph that is created for true positive rate vs. false positive rate at different classifications threshold. The entire area under the ROC curve is known as area of the curve (AOC). It gives a collective measure of performance across all achievable classification's threshold.

8) GINI COEFFICIENT

It is also known as GINI index. It is a measure of statistical distribution. It is used to measure the inequality amongst values of attributes. It is also possible to say that it calculates the impurity of a particular attribute in the form of degree or probability.

H. SMOTE

Synthetic Minority Oversampling Technique (SMOTE) is used for oversampling the minority class. It is also known as a balancer. It takes the whole dataset as input but works only on minority class. It increases the percentage of minority class. SMOTE used KNN for finding new instances. It does not make any change in the majority cases. The new examples are not simply duplicating of existing minority cases. Instead, the calculation takes tests of the component space for each target class and its closest neighbors and then produces new models that join attributes of the objective case with the highlights of its neighbors. This methodology builds the highlights accessible for each class and makes tests progressively broad.

The mathematical symbols used in this research is shown in the table 3.

I. STATISTICS TEST FOR MODEL COMPARISON

For the purpose of comparing two models, McNemar's test was applied on the predicted output of two models. The McNemar's test is used to determine whether there are differences on bipolar dependent variables between two related groups. In this test 2*2 contingency matrix is formed of two groups and p value is calculated. For this purpose, significance level $\alpha = 0.05$ is considered. If $p < \alpha$, we can reject the null hypothesis. If $p > \alpha$, we fail to reject the null hypothesis. If p value is less than α, it means both models show a significant difference as regards the hypothesis. However, if $p > \alpha$, the difference would not be regarded as statistically significant.

IV. EXPERIMENTAL AND RESULT

The result of this research including all outcomes and classification models from different perception will be discussed in this section. IBM SPSS model is shown in figure 1. First, the performance of different machine learning algorithm viz. an artificial neural network, logistic regression, C5.0, CHAID, random tree, K-nearest neighbors and linear support vector machine have been checked on all features. Second, part feature selection algorithm CFS, forward Wrapper and LASSO have been applied on the dataset to find the important features. Third, the performance of all above-mentioned classification algorithm on important features was checked. Fourth, SMOTE filter was also applied to the dataset and the result of classifiers were checked. Various tools were used. Weka tool was used for CFS and Forward method. r studio was used for LASSO. IBM SPSS Modeler was used for the performance of classifiers. Deep neural network was built in IBM SPSS modeler. ANN was used with 2 hidden layers for building deep neural network. Twelve nodes were used in hidden layer 1 and eight nodes were in hidden layer 2.

A. RESULT WITHOUT FEATURES SELECTION

In this subsection, the full features of the dataset were used and the result was tested on all seven machine learning classification algorithms with 50% of training data and 50% of testing data. The comparison matrix was created for all algorithms. With the resultant matrix, three graphs were also created for checking the variation of various classifiers. The first graph provides a comparison of all classifier's accuracy, precision and recall. The second one contains the variation of AUC and the third one includes the variation of F-measure. The comparison of all classifiers showed that the C5.0 algorithm achieved the highest accuracy, i.e. 96.10%. The value of all parameters of C5.0: precision was 92.40%, recall was 97.30%, F-measure was 94.80%, AUC was 97.80% and GINI index was 0.96. The artificial neural network was trained on 3 hidden layers. The ANN achieved accuracy of 94.63%, precision of 93.24% and recall of 92%. The logistic regression achieved accuracy of 71.71%, precision of 56.48% and recall of 98.6%. The CHAID algorithm achieved accuracy of 96%, precision of 93.50% and recall of 92%. The LSVM with Penalty L1 and Lambda 0.5 achieved accuracy of 92.2%, precision of 83.90% and recall of 97.33%. The LSVM with Penalty L2 and Lambda 0.5 achieved accuracy of 94.63%, precision of 90% and recall of 96%. The KNN gave the worst result for this dataset with a K value of 5: accuracy of 64.39%, precision of 59.01% and recall of 96%. The random tree algorithm achieved accuracy of 90.73%, precision of 83.34% and recall of 93%. As a result, ANN achieved the highest AUC and C5.0 achieved F-measure. The result of all classifiers is

TABLE 3. Description of Used Mathematical Symbol.

Symbol	Name
p	Significant Value
α	Significant Level
P	Predicted Value
b_0, b_1, b_2	Bias
x	Attribute
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
accuracy	Classification Accuracy
Error	Classification Error
Precision	Classification precision rate
Recall	Classification recall rate
F-Measure	Classification F1 Score

as follows in table 4. The comparison of precision, recall and accuracy is described in figure 2. The comparison of AUC is described in figure 3. The comparison of F-Measure is described in figure 4.

B. RESULT OF CORRELATION-BASED FEATURE SELECTION (CFS)

In this subsection, the important features were selected by CFS algorithm to pass in the classifier algorithms for predicting the outcomes. Six most important features were used for finding the outcomes such as bp, pc, pe, ane, pcv and rbc. As per the CFS algorithm, bp and pc are the most important factors for predicting Chronic Kidney Disease. The result of CFS algorithm is shown in figure 5. The performance of all seven classifiers was described in table 5. The LSVM with Penalty L2 and Lambda 0.5 achieved the highest accuracy for selected features from the CFS algorithm with 95.12% accuracy, 93.34% precision and 93.34% recall. C5.0 and CHAID achieved an accuracy of 92.68%. The C5.0 algorithm achieved 85.71% precision and 96% recall. The CHAID algorithm achieved 92.68% accuracy, 96.87% precision and 83% recall. The ANN algorithm achieved 91.71% accuracy, 89.19% precision and 88% recall. The logistic regression algorithm achieved the lowest accuracy of 51.22%, 96.87% precision and 92.54% recall. The LSVM with Penalty L1 and Lambda 0.5 achieved 93.66% accuracy, 87.8% precision and 96% recall. The KNN achieved for this dataset with a K value of 5 accuracy of 53.17%, precision of 97.05% and recall of 100%. The random tree algorithm achieved 87.80% accuracy,82.05% precision and 85% recall. As from the result, LSVM with penalty L1 achieved the highest AUC. The comparison of precision, recall and accuracy is described in figure 5. The comparison of the GINI index is shown in figure 6. The comparison of AUC is described in figure 7.

C. RESULT OF WRAPPER FORWARD FEATURE SELECTION AND CLASSIFICATION

In this subsection, the important features were selected by Wrapper forward feature selection algorithm to pass in the classifier algorithms for predicting the outcomes. Six most important features were used for finding outcomes such as hemo, htn, dm, cad, pe, al. As per the CFS algorithm, hemo and htn are the most important factors for predicting Chronic Kidney Disease. The result of the Wrapper algorithm is shown in figure 9. The result of all classifier algorithm performance is described in table 6. The C5.0 achieved the highest accuracy with the Wrapper algorithm, namely 96.1% accuracy, 98.55% precision and 90.67% recall. ANN, CHAID and the random tree also gave a good result. The ANN algorithm achieved 94.63% accuracy, 90% precision and 96% recall. The CHAID algorithm achieved 94.63% accuracy, 93.24% precision and 92% recall. The random tree algorithm achieved 94.63% accuracy, 93.24% precision and 92% recall. The logistic regression algorithm achieved 78.54% accuracy, 98.55% precision and 100% recall. The LSVM with Penalty L1 and Lambda 0.5 achieved 94.15% accuracy, 88.89% precision and 96% recall. The LSVM with Penalty L2 and Lambda 0.5 achieved 93.66% accuracy, 87.80% precision and 96% recall. The KNN gave the worst result for this dataset with a K value of 5 76: 10% accuracy, 95.58% precision and 95.58% recall. As from the result, LSVM achieved the highest AUC. The comparison of precision, recall and accuracy is described in figure 10. The comparison of the GINI index is shown in figure 11. The comparison of AUC is described in figure 12.

D. RESULT OF LASSO FEATURE SELECTION

In this subsection, the important features were selected by LASSO feature selection algorithm to passe in the classifier algorithms for predicting the outcomes. Six most important features were used for finding outcomes such as rbc, pc, al, ba, su, pcc. As per the LASSO FS algorithm, rbc and pc are the most important factors for predicting Chronic Kidney Disease. The result of the LASSO FS algorithm is shown in figure 13. The result of algorithm performance for all seven classifiers is described in table 7. LSVM and CHAID achieved the highest accuracy of 97.07%. LSVM with both penalty L1 and L2 achieved 97.07% accuracy, 98.59% precision and 93.33% recall. The CHAID algorithm achieved

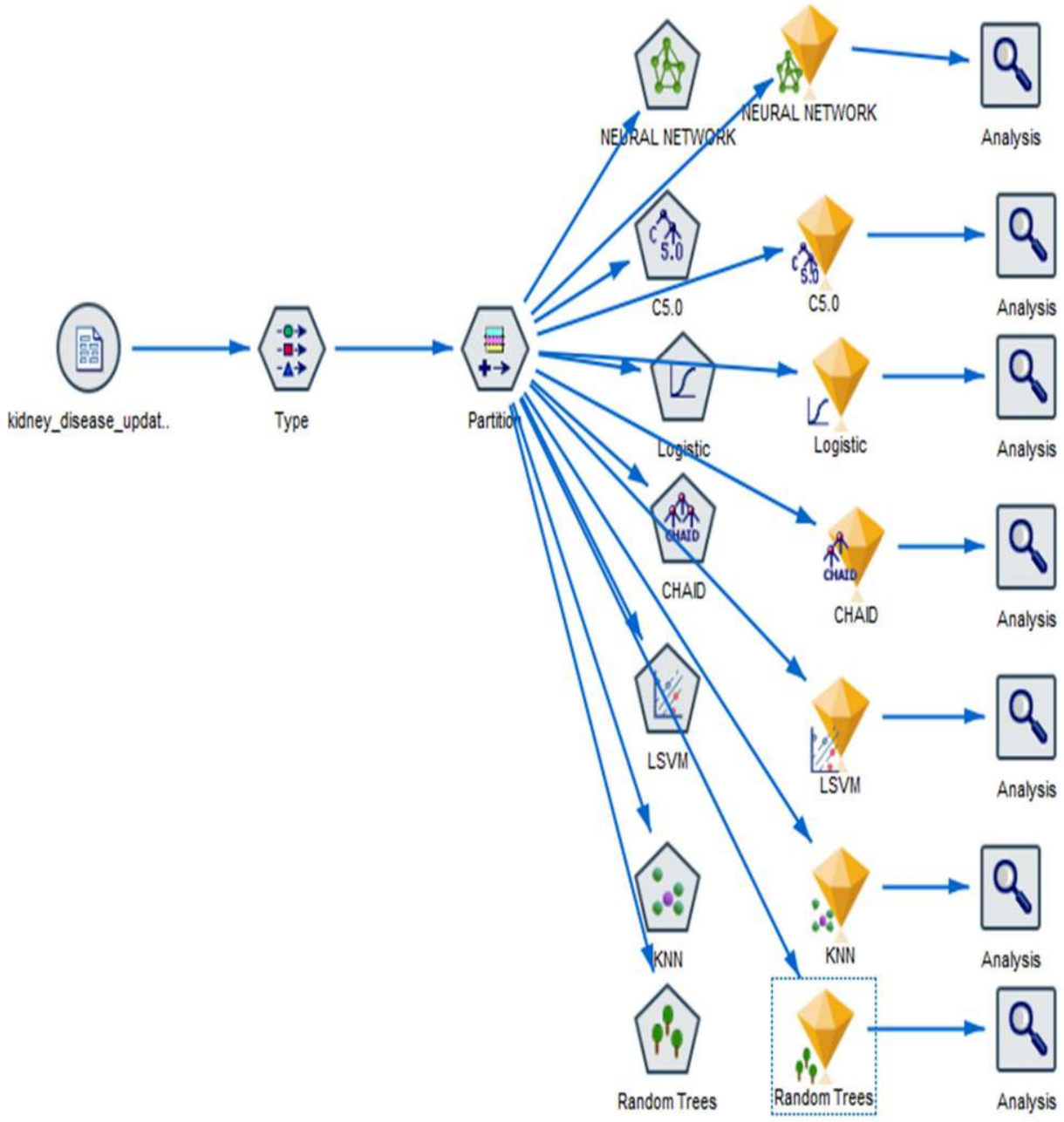


FIGURE 1. IBM SPSS model for kidney disease prediction.

97.07% accuracy, 100% precision and 92% recall. The ANN algorithm achieved 94.63% accuracy, 90% precision and 96% recall. The random tree algorithm achieved 90.24% accuracy, 80.90% precision and 96% recall. The logistic regression algorithm achieved 74.15% accuracy, 80.23% precision and 100% recall. The random tree algorithm achieved 88.78% accuracy, 78.26% precision and 96% recall. The KNN gave the worst result for this dataset with a K value of 5: 56.59% accuracy, 92% precision and 100% recall. As from the result, LSVM achieved the highest AUC. The comparison of precision, recall and accuracy is described in figure 13. The comparison of the GINI index is shown in figure 14. The comparison of AUC is described in figure 15 shows.

E. RESULT OF SMOTE

As the above result, it was observed that the highest accuracy achieved on the selected features was given by LASSO feature selection method. Thus, SMOTE technique was applied on full features and on selected features given by LASSO regression method. The performance of ANN, CHAID, LSVM and Random Tree was checked by SMOTE. These classification algorithms were performed very well in all

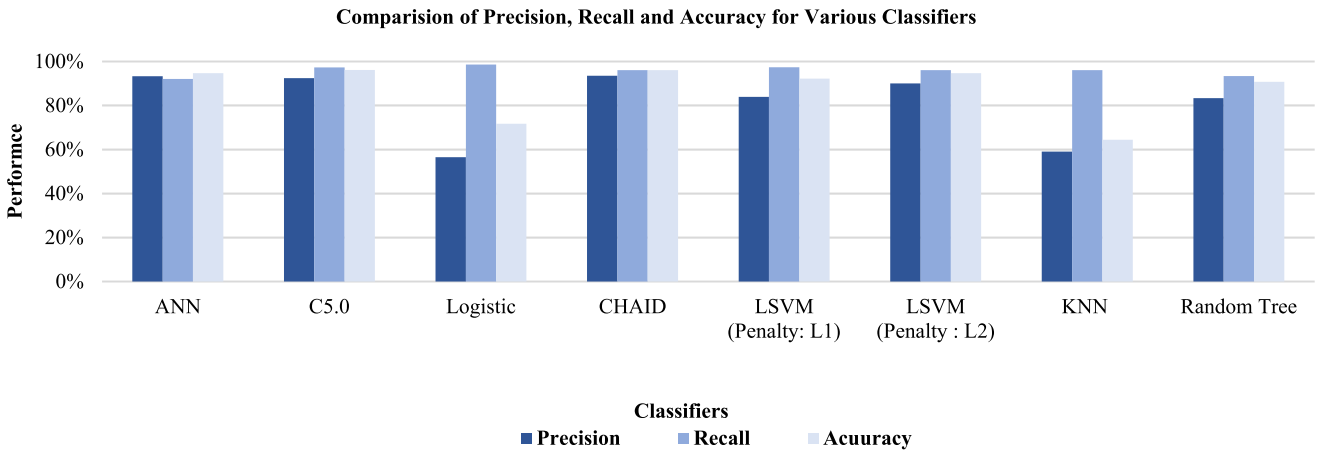


FIGURE 2. Comparison of precision, recall and accuracy for all classifiers without feature selections.

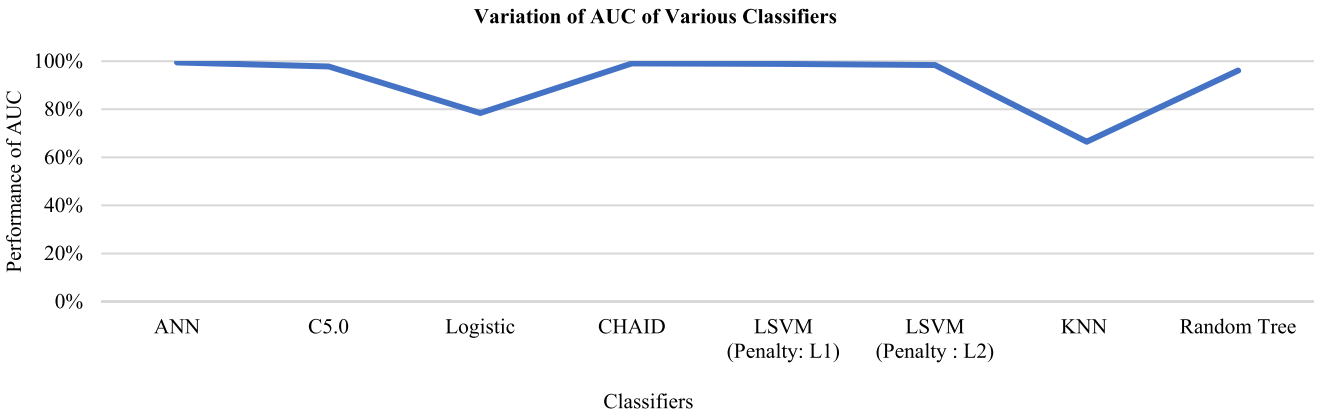


FIGURE 3. Performance of Area under curve for all classifiers without feature selections.

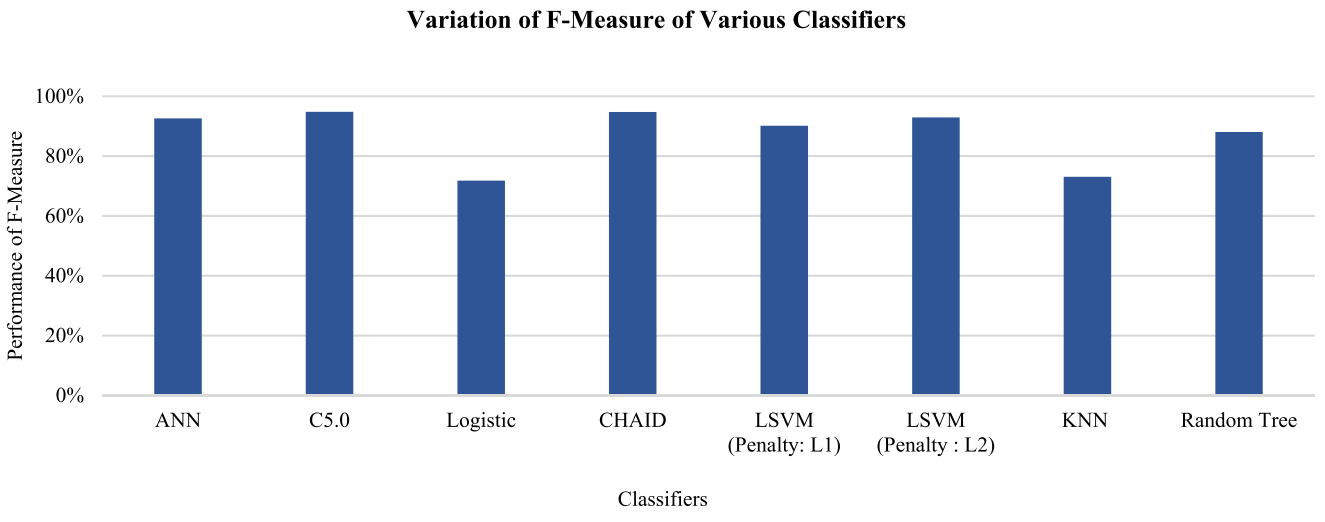


FIGURE 4. Performance of F-Measures for all classifiers without feature selections.

experiments. As the above result, Logistic regression and KNN were not performed well on this dataset, so these two classification techniques performance were not checked with SMOTE technique.

1) SMOTE WITH SELECTED FEATURES

The main purpose of this experiment was to increase the performance of the model and to achieve higher accuracy in this model. As per the expectation, this experiment

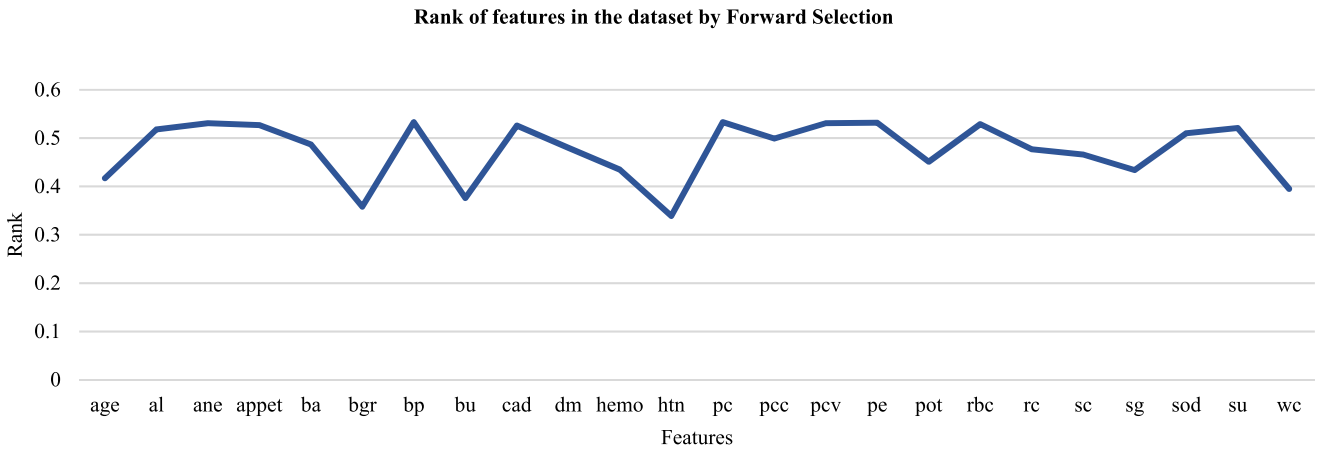


FIGURE 5. Dataset Feature importance using CFS algorithm.

TABLE 4. Performance of Classifiers Without Feature Selection.

Classifiers	Precision	Recall	F-Measure	AUC	GINI Coefficient	Accuracy
ANN	93.24%	92.00%	92.61%	99.50%	0.99	94.63%
C5.0	92.40%	97.30%	94.80%	97.80%	0.96	96.10%
Logistic	56.48%	98.60%	71.80%	78.40%	0.567	71.71%
CHAID	93.50%	96%	94.74%	99.10%	0.981	96.00%
LSVM (Penalty: L1, Lambda: 0.5)	83.90%	97.33%	90.12%	98.90%	0.978	92.20%
LSVM (Penalty: L2, Lambda: 0.5)	90.00%	96.00%	92.90%	98.40%	0.968	94.63%
KNN	59.01%	96%	73.09%	66.50%	0.329	64.39%
Random Tree	83.34%	93%	88.05%	96.10%	0.922	90.73%

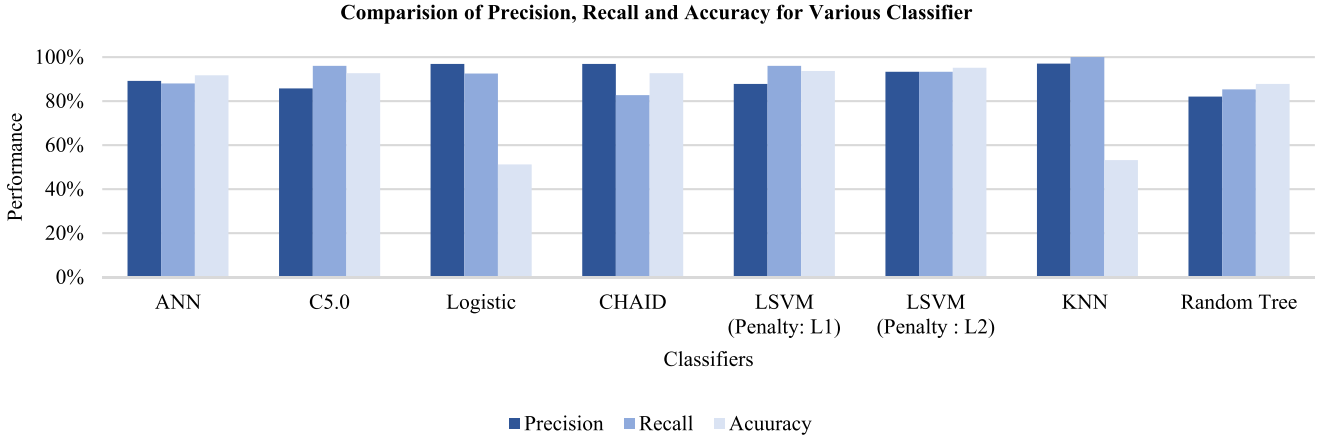


FIGURE 6. Comparison of precision, recall and accuracy for all classifiers after correlation-based feature selection.

gave the highest accuracy as compared to without SOMTE in LASSO's selected features model. Linear Support Vector Machine (LSVM) achieved the highest accuracy with 98.46%. LSVM with penalty L1 and L2 gave the same result i.e. 98.46% accuracy, 98.59% precision and 97.22% recall. Table 8 shows the result of classification model with

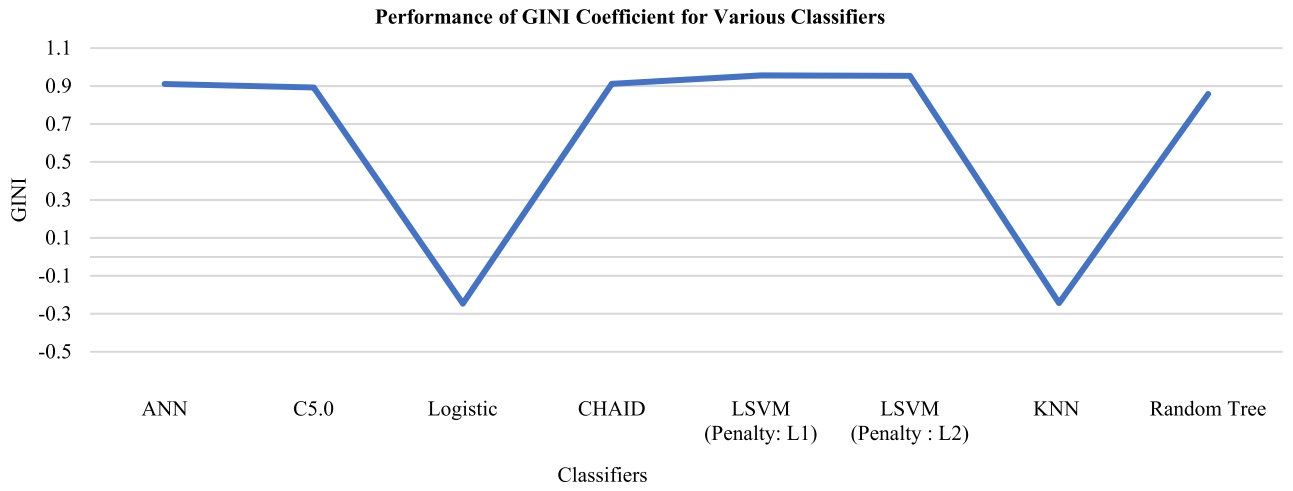


FIGURE 7. Performance of GINI index for all classifiers after correlation-based feature selection.

TABLE 5. Performance of Classifiers After Correlation-Based Feature Selection.

Classifiers	Precision	Recall	F-Measure	AUC	GINI Coefficient	Accuracy
ANN	89.19%	88.00%	88.59%	95.50%	0.91	91.71%
C5.0	85.71%	96.00%	90.57%	94.60%	0.89	92.68%
Logistic	96.87%	92.54%	94.65%	37.70%	-0.246	51.22%
CHAID	96.87%	83%	89.20%	95.60%	0.912	92.68%
LSVM (PenaltyL1)	87.80%	96.00%	91.72%	97.80%	0.956	93.66%
LSVM (PenaltyL1)	93.34%	93.34%	93.34%	97.70%	0.954	95.12%
KNN	97.05%	100%	98.50%	37.90%	-0.243	53.17%
Random Tree	82.05%	85%	83.66%	92.90%	0.858	87.80%

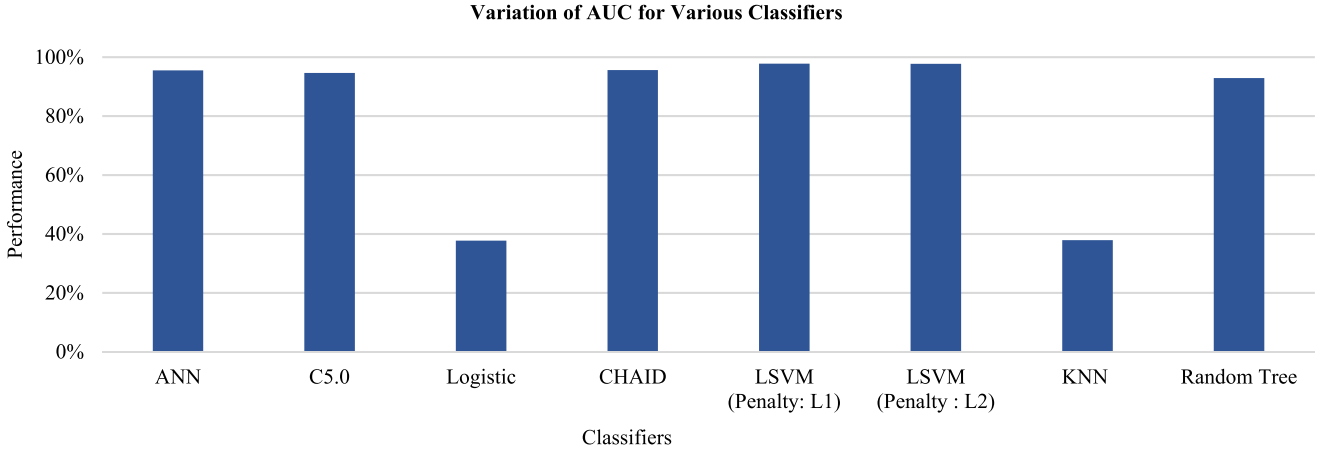


FIGURE 8. Performance of area under the curve for all classifiers after correlation-based feature selection.

SMOTE and LASSO. It can be noted that all the algorithms were performed better with SMOTE than without SMOTE. After the LSVM, CHAID achieved 97.95% accuracy, 95.49% precision and 99% recall. ANN achieved 91.92% accuracy, 84.34% precision and 95.89% recall. C5.0 achieved 88.72% accuracy, 77.17% precision and 98.61% recall. The Random

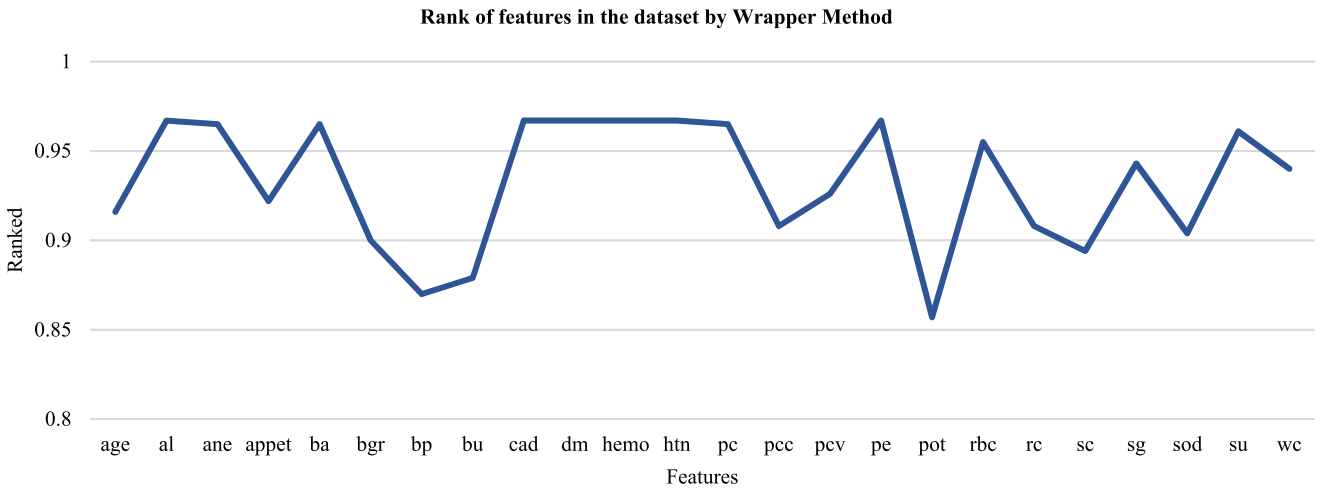


FIGURE 9. Dataset feature importance using CFS algorithm.

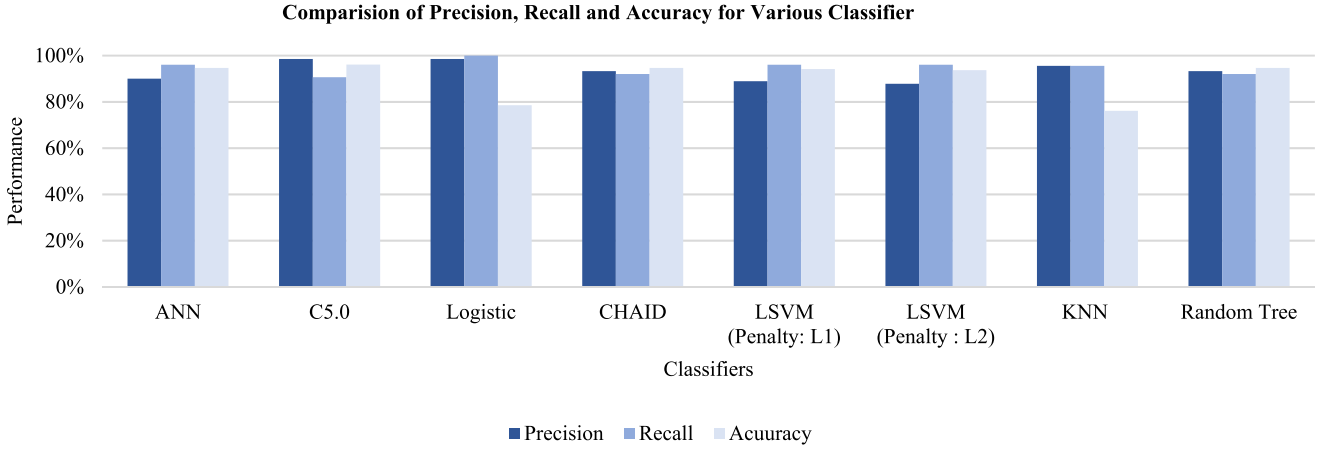


FIGURE 10. Comparison of precision, recall and accuracy for all classifiers after Wrapper feature selection.

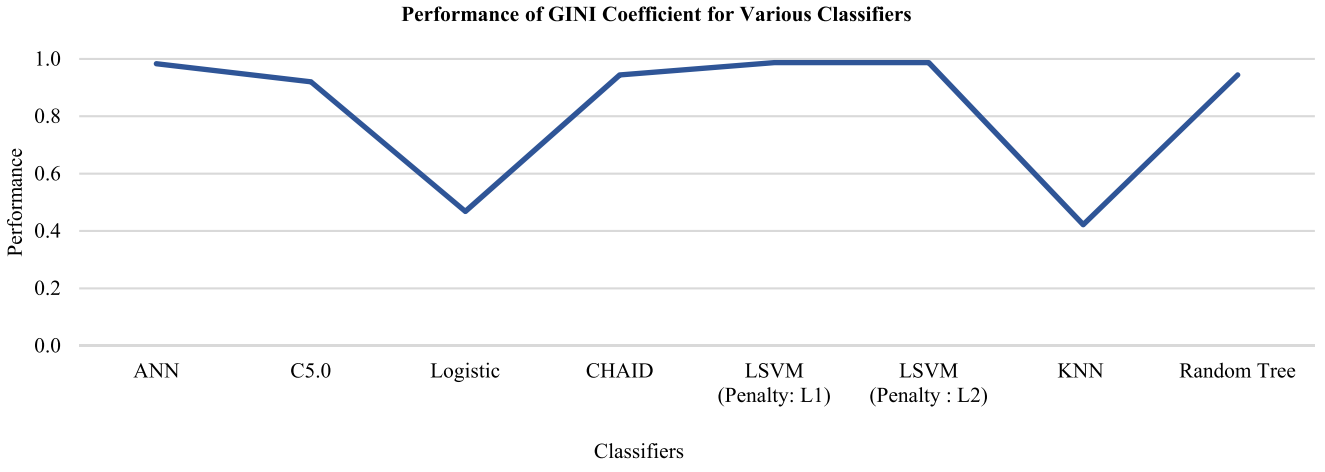


FIGURE 11. Performance of GINI index for all classifiers after Wrapper feature selection.

tree achieved 89.23% accuracy, 78.02% precision and 99% recall. The comparison graph of precision, recall and accuracy for all algorithms is shown in figure 17. The comparison graph of AUC for all algorithms is shown in figure 18. The comparison graph of F-Measure for all algorithms is shown in figure 19.

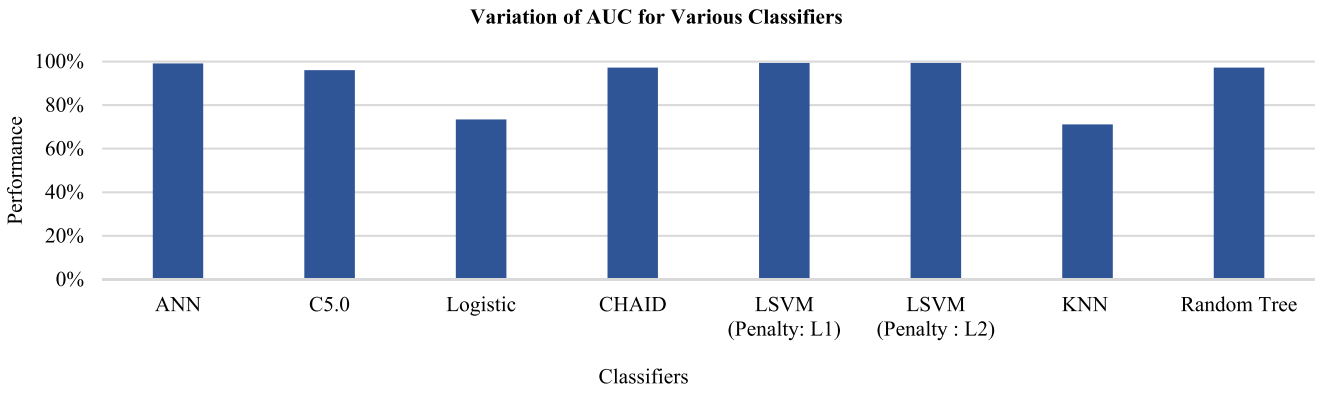


FIGURE 12. Performance of area under the curve for all classifiers after Wrapper feature selection.

TABLE 6. Performance of Classifiers After Wrapper Method Feature Selection.

Classifiers	Precision	Recall	F-Measure	AUC	GINI Coefficient	Accuracy
ANN	90.00%	96.00%	92.90%	99.10%	0.98	94.63%
C5.0	98.55%	90.67%	94.44%	96.00%	0.92	96.10%
Logistic	98.55%	100.00%	99.27%	73.40%	0.468	78.54%
CHAID	93.24%	92%	92.61%	97.20%	0.944	94.63%
LSVM (PenaltyL1)	88.89%	96.00%	92.30%	99.30%	0.987	94.15%
LSVM (PenaltyL2)	87.80%	96.00%	91.72%	99.30%	0.987	93.66%
KNN	95.58%	95.58%	95.58%	71.10%	0.422	76.10%
Random Tree	93.24%	92%	92.61%	97.20%	0.944	94.63%

TABLE 7. Performance of Classifiers After Lasso Feature Selection.

Classifiers	Precision	Recall	F-Measure	AUC	GINI Coefficient	Accuracy
ANN	80.90%	96.00%	87.80%	91.20%	0.83	90.24%
C5.0	77.42%	96.00%	85.71%	91.30%	0.83	88.29%
Logistic	80.23%	100.00%	89.03%	71.80%	0.437	74.15%
CHAID	100.00%	92%	95.83%	97.20%	0.945	97.07%
LSVM (Penalty: L1, Lambda: 0.5)	98.59%	93.33%	95.89%	98.70%	0.974	97.07%
LSVM (Penalty: L2, Lambda: 0.5)	98.59%	93.33%	95.89%	98.80%	0.976	97.07%
KNN (K=1)	92.00%	100%	95.83%	43.10%	-0.137	56.59%
Random Tree	78.26%	96%	86.22%	91.40%	0.828	88.78%

2) SMOTE WITH FULL FEATURES

As after the LASSO, full features model performance was satisfactory. So also SMOTE was applied on models with full features and all 5 classifiers performed well in this case. LSVM with penalty L2 achieved the highest accuracy with 98.86%, precision of 96.67% and recall of 100%.

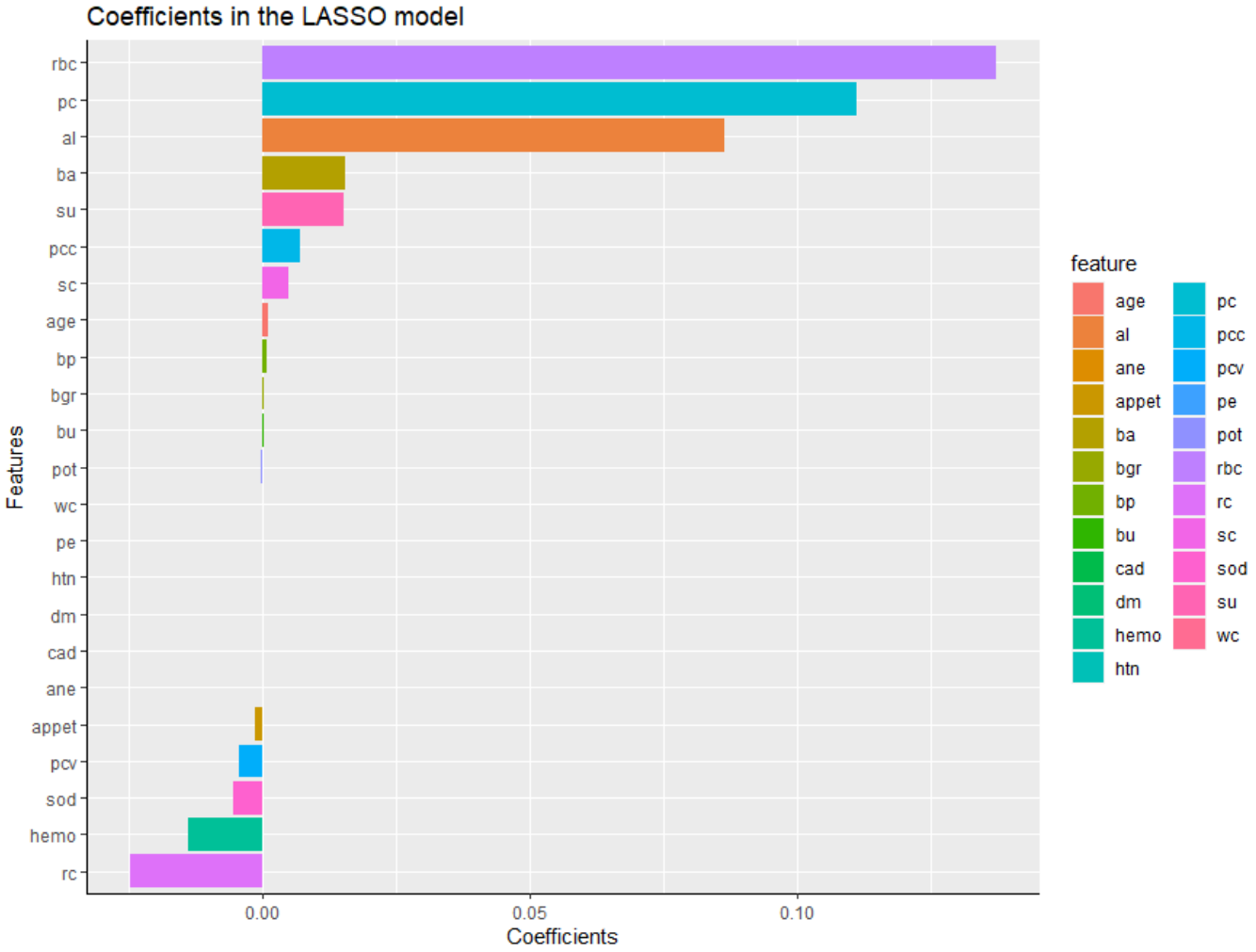


FIGURE 13. Dataset feature importance using LASSO regression algorithm.

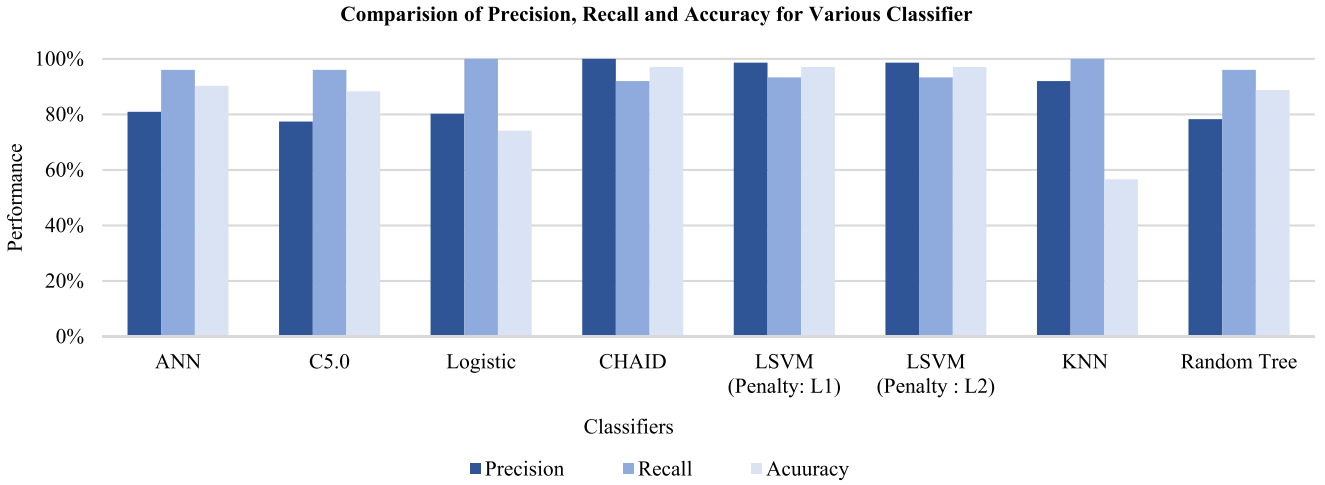


FIGURE 14. Comparison of precision, recall and accuracy for all classifiers after LASSO feature selection.

Table 9 shows the result of classification model with SMOTE and full features. It should be noted that all the algorithms performed better with SMOTE than without SMOTE. After the LSVM with penalty L2, CHAID achieved the highest accuracy. CHAID achieved 97.25% accuracy, 91.93% precision and 100% recall. ANN achieved 96.47% accuracy, 98.14%

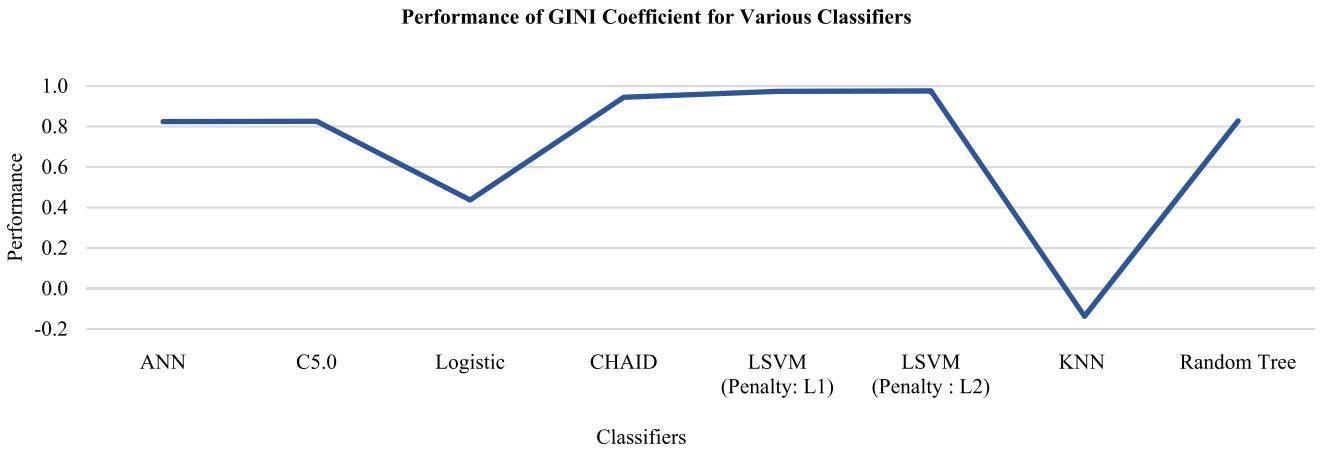


FIGURE 15. Performance of GINI index for all classifiers after LASSO feature selection.

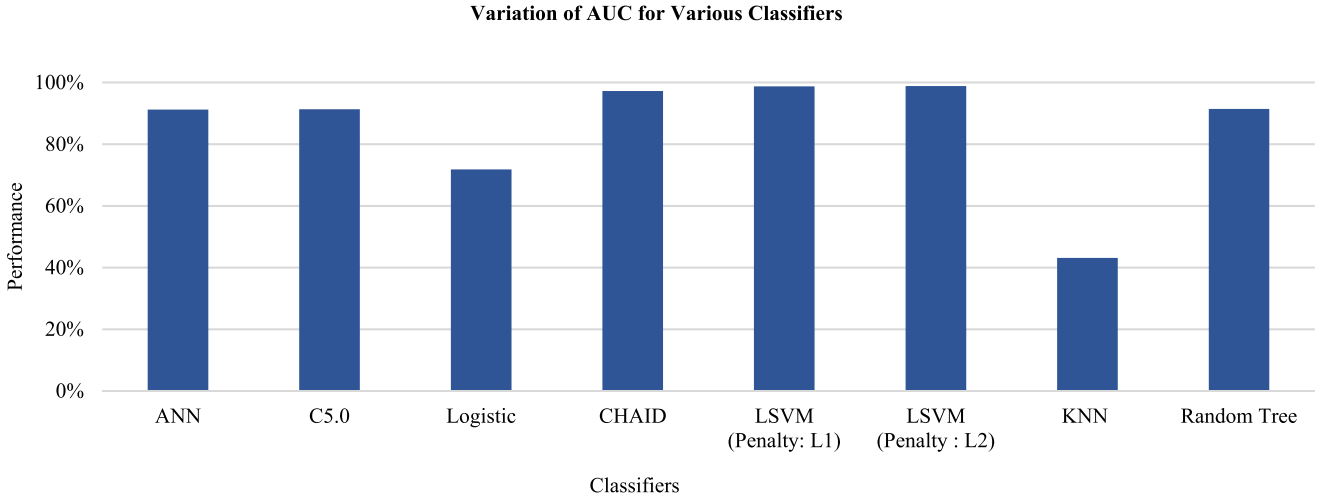


FIGURE 16. Performance of area under the curve for all classifiers after LASSO feature selection.

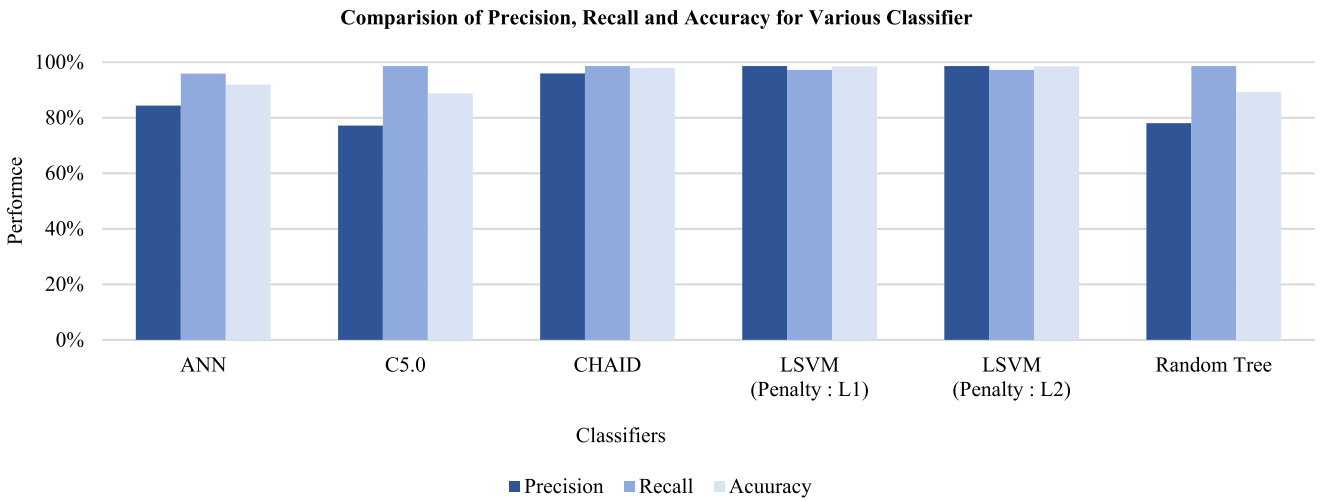


FIGURE 17. Comparison of precision, recall and accuracy for all classifiers with SMOTE and selected features.

precision and 91.38% recall. LSVM with penalty L1 achieved 96.53% accuracy, 91.04% precision and 100% recall. ANN achieved 96.47% accuracy, 98.14% precision and 91.38% recall. C5.0 achieved 96.45% accuracy, 96.61% precision and 93.44% recall. The Random Tree achieved 91.43% accuracy, 84.72% precision and 94% recall. The comparison graph of

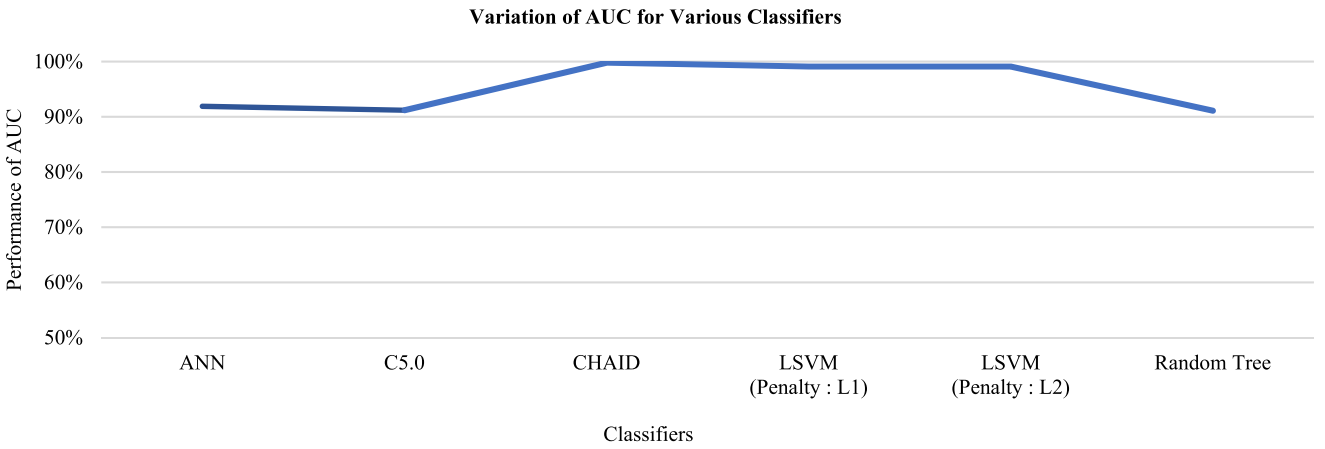


FIGURE 18. Performance of AUC for all classifiers with SMOTE and selected features.

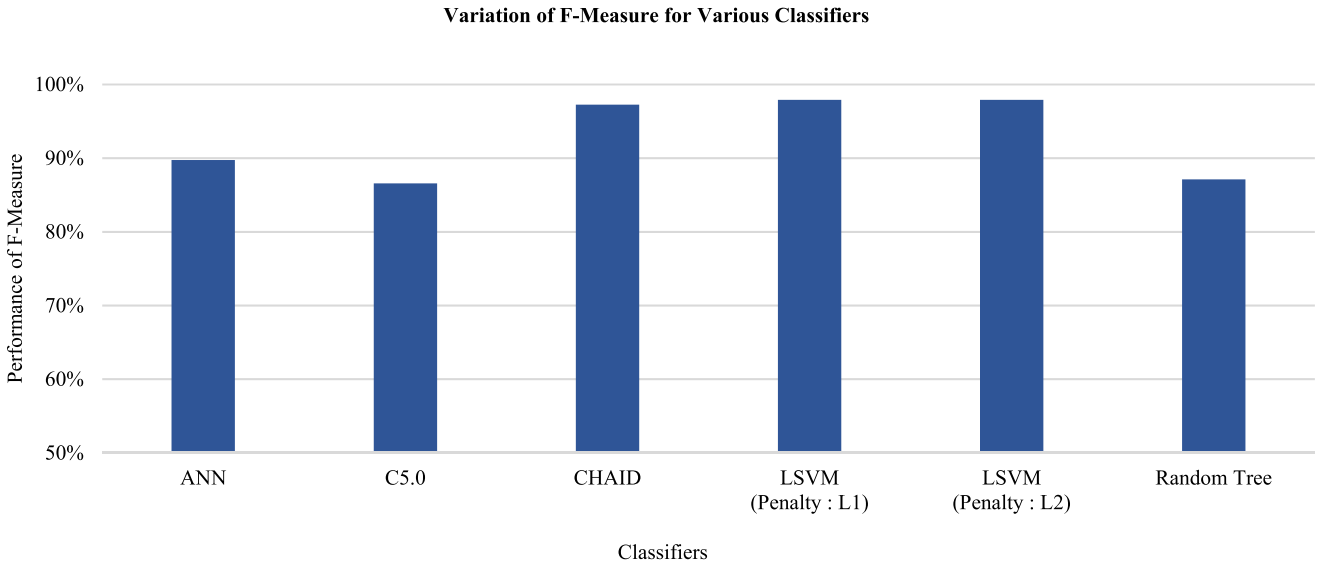


FIGURE 19. Performance of F-Measure for all classifiers after LASSO feature selection and SMOTE.

TABLE 8. Performance of Classifiers With SMOTE and Selected Features.

Classifiers	Precision	Recall	F-Measure	AUC	GINI Coefficient	Accuracy
ANN	84.34%	95.89%	89.74%	91.90%	0.84	91.92%
C5.0	77.17%	98.61%	86.58%	91.20%	0.83	88.72%
CHAID	95.94%	99%	97.26%	99.80%	0.997	97.95%
LSVM (Penalty: L1, Lambda: 0.5)	98.59%	97.22%	97.90%	99.10%	0.981	98.46%
LSVM (Penalty: L2, Lambda: 0.5)	98.59%	97.22%	97.90%	99.10%	0.981	98.46%
Random Tree	78.02%	99%	87.11%	91.10%	0.823	89.23%

precision, recall and accuracy for all algorithms is shown in figure 20. The comparison graph of AUC for all algorithms is shown in figure 21. The comparison graph of F-Measure for all algorithms is shown in figure 22.

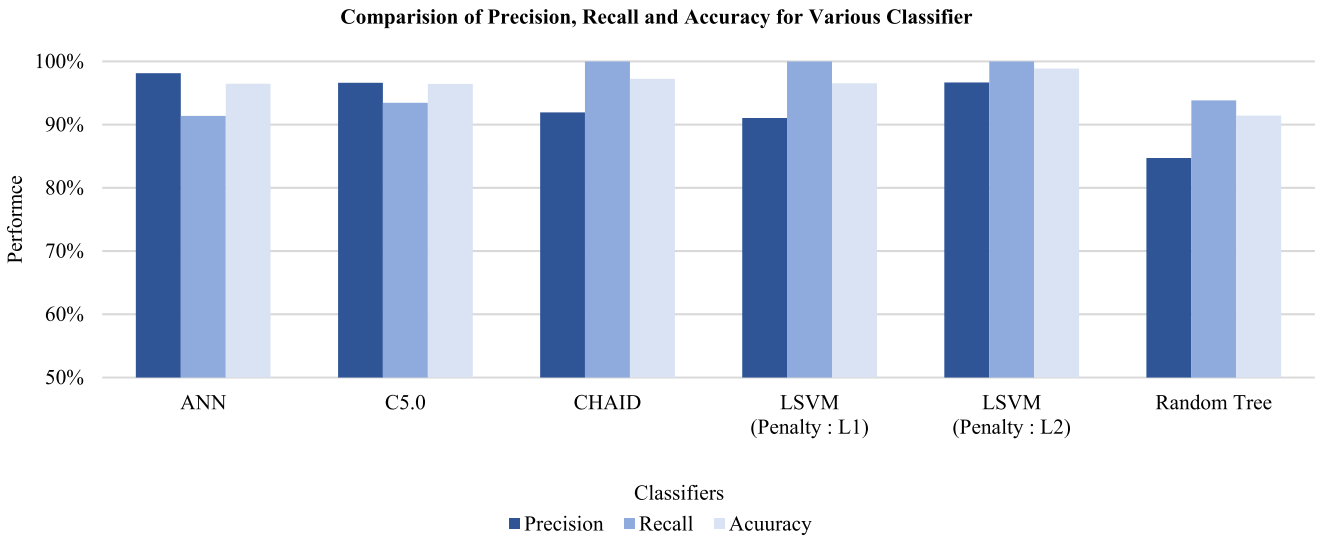


FIGURE 20. Comparison of precision, recall and accuracy for all classifiers with SMOTE and full features.

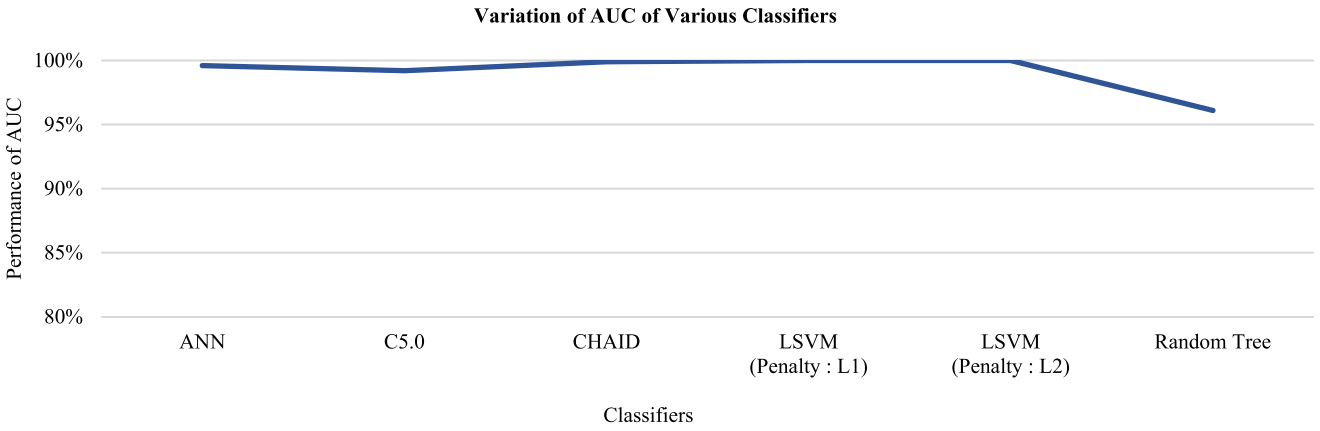


FIGURE 21. Performance of AUC for all classifiers with SMOTE and full features.

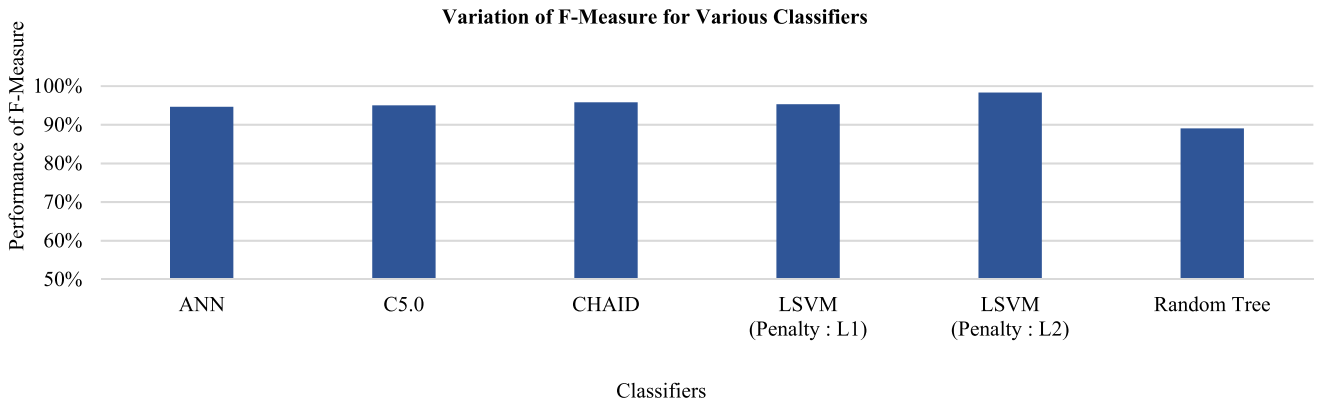


FIGURE 22. Performance of F-Measure for all classifiers after LASSO feature selection and SMOTE.

F. COMPARISION MATRIX OF ALL EXPERIMENTS

It was observed that ANN, C5.0, CHAID, LSVNM and Random tree performed well on the considered CKD dataset. The Logistic regression and KNN have not given outcomes as expected. So, the comparison table has been created for five best-performed algorithms of all different technique type.

TABLE 9. Performance of Classifiers With SMOTE and Full Features.

Classifiers	Precision	Recall	F-Measure	AUC	GINI Coefficient	Accuracy
ANN	98.14%	91.38%	94.64%	99.60%	0.99	96.47%
C5.0	96.61%	93.44%	95.00%	99.20%	0.98	96.45%
CHAID	91.93%	100%	95.79%	99.90%	0.999	97.25%
LSVM (Penalty: L1, Lambda: 0.5)	91.04%	100.00%	95.31%	100.00%	1.0	96.53%
LSVM (Penalty: L2, Lambda: 0.5)	96.67%	100.00%	98.30%	100.00%	0.99	98.86%
Random Tree	84.72%	94%	89.05%	96.10%	0.921	91.43%

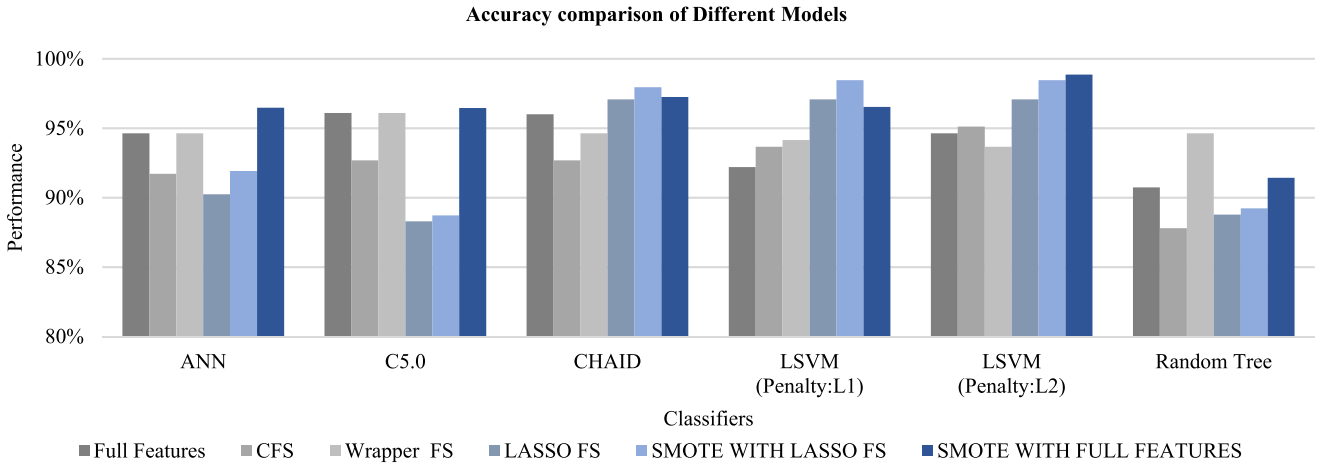


FIGURE 23. Comparison of all classifier models.

TABLE 10. Classifier Performance in Various Models.

Classifiers	Full Features	CFS	Wrapper FS	LASSO FS	SMOTE WITH LASSO FS	SMOTE WITH FULL FEATURES
ANN	94.63%	91.71%	94.63%	90.24%	91.92%	96.47%
C5.0	96.10%	92.68%	96.10%	88.29%	88.72%	96.45%
CHAID	96.00%	92.68%	94.63%	97.07%	97.95%	97.25%
LSVM (Penalty: L1)	92.20%	93.66%	94.15%	97.07%	98.46%	96.53%
LSVM (Penalty: L2)	94.63%	95.12%	93.66%	97.07%	98.46%	98.86%
Random Tree	90.73%	87.80%	94.63%	88.78%	89.23%	91.43%

The result is described in table 10. The accuracy comparison graph is shown in figure 23.

G. PERFORMANCE OF LSVM IN ALL TECHNIQUES

As the above result, LSVM with penalty L2 gave a better result in all techniques is it was discussed previously. In this section, the performance of LSVM will be discussed. The table 11 shows the result of LSVM in all techniques. Along with the table, there is the graph on the table data. Figure 24 shows the comparison of LSVM in all techniques.

H. VALIDATE MACHINE LEARNING MODEL

To validate the findings above, the study includes results from another data set found at The Cancer Imaging Archive

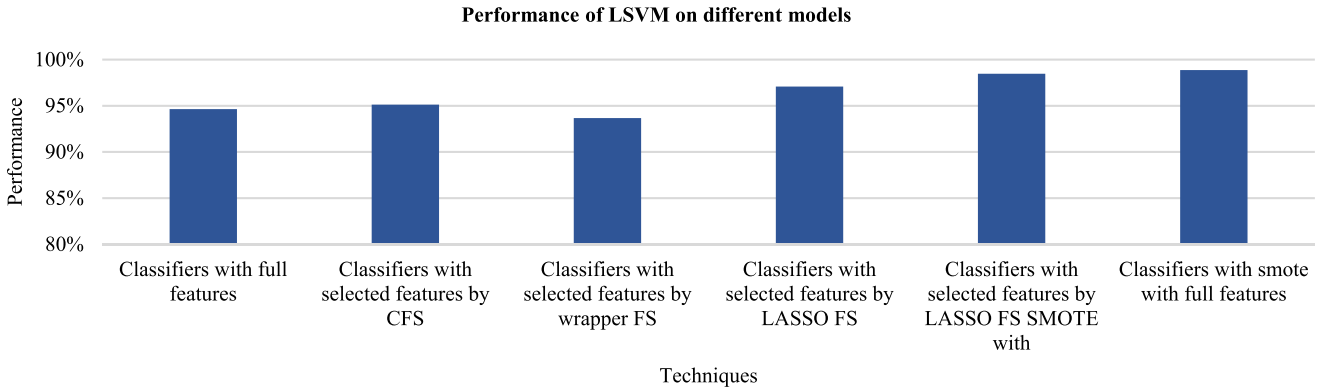


FIGURE 24. Comparison of LSVM in different models.

TABLE 11. Performance of LSVM in all Machine Learning Models.

Different Techniques	Accuracy
Classifiers with full features	94.63%
Classifiers with selected features by CFS	95.12%
Classifiers with selected features by Wrapper FS	93.66%
Classifiers with selected features by LASSO FS	97.07%
Classifiers with selected features by LASSO FS SMOTE with	98.46%
Classifiers with smote with full features	98.86%

(TCIA). The dataset has 210 instances of kidney disease patient. It contains 48 attributes and 1 target variable. The dataset was used on the same models applied earlier. These findings are given below and are, in general, comparable to earlier results with no significant differences observed. Though, the outcome of applying machine learning models largely dependens on the specific dataset, the experiments above validate earlier findings, namely, SMOTE with full features result. Table 12 shows the result of both datasets.

I. PERFORMANCE OF DEEP NEURAL NETWORK

In this research work, artificial neural network was used for machine learning and deep neural network-based analysis. For machine learning artificial neural network used only single hidden layer, but for the usage of the artificial neural network as a deep neural network more hidden layers can be added. So as to test the performance of machine learning classifier algorithms, one deep neural network model was built and the results were noted. In some cases, the deep neural network gave strong result and important features were extracted by itself, that is, no feature selection algorithm was required. The same dataset was used for building a deep neural network. It was noted that a deep neural network achieved the highest accuracy of 99.6% and it was better than other machine learning models.

V. DISCUSSION OF MACHINE LEARNING MODELS

All the machine learning models but Logistic and KNN classifiers give satisfactory result and have the negligible difference between precision and recall values. In comparison with them precision for Logistic and KNN classifiers is low whereas recall is high. It indicates that these two classifiers give many False positive results due to unbalanced dataset. Logistic and KNN algorithms have not enough capacity to distinguish between positive class and negative class as the related AUC score is very low. Along with AUC, the GINI coefficient is also not satisfactory. Hence, Logistic and KNN are not suitable for the prediction of CKD. In all cases LSVM with L1 and L2 penalty has the best precision, recall, AUC score and GINI coefficient and the model achieved the highest accuracy in majority cases.

VI. PERFORMANCE COMPARISON OF MACHINE LEARNING MODEL AND DEEP NEURAL NETWORK

All the machine learning model results was discussed in the table 10 and according to it LSVM with penalty L2 performed best in SMOTE with full features and achieved the highest accuracy of 98.46%. As discussed, the deep neural network achieved the highest accuracy from among all models with 99.6%. In order to compare the performance of two models, McNemar's test was applied. For this test, the highest accuracy was achieved by machine learning model, i.e., LSVM

TABLE 12. Validation of Machine Learning Model.

Classifiers	Dataset 1 (UCI) Result (SMOTE with Full Features)	Dataset 2 (TCIA) Result (SMOTE with Full Features)
ANN	91.92%	95.20%
C5.0	88.72%	93.48%
CHAID	97.95%	94.12%
LSVM (Penalty: L1, Lambda: 0.5)	98.46%	80.95%
LSVM (Penalty: L2, Lambda: 0.5)	98.46%	91.67%
Random Tree	89.23%	94.00%

with SMOTE for all features and a deep neural network was taken and their significant value was noted. The p value of this test was 0.29 and it is greater than significant level ($\alpha = 0.05$) and, hence, we would reject hypothesis.

VII. CONCLUSION

This article objects to predict Chronic Kidney Disease based on full features and important features of CKD dataset. For feature selection three different techniques have been applied: correlation-based feature selection, Wrapper method and LASSO regression. In this perception, seven classifiers algorithm were applied viz. artificial neural network, C5.0, logistic regression, CHAID, linear support vector machine (LSVM), K-Nearest neighbors and random tree. For each classifier, the results were computed based on full features, selected features by CFS, selected features by Wrapper, selected features by LASSO regression, SMOTE with selected features by LASSO, SMOTE with full features. It was observed that LSVM achieved the highest accuracy of 98.86% in SMOTE with full features. All classifiers algorithms performed well on features selected by LASSO regression with SMOTE and without SMOTE. SMOTE with full features gave the best result for all 5 classifiers. In this research, a total of 7 classifiers were used. However, Logistic and KNN did not give suitable results and it was why they were not used in SMOTE. As per the result, it is concluded that SMOTE is a best technique for balancing a dataset. It is noted that SMOTE gave better results with selected features by LASSO regression as compare to without SMOTE on LASSO regression model. LSVM achieved the highest accuracy in all experiments as compared to other classifiers algorithms.

REFERENCES

[1] Q.-L. Zhang and D. Rothenbacher, "Prevalence of chronic kidney disease in population-based studies: Systematic review," *BMC Public Health*, vol. 8, no. 1, p. 117, Dec. 2008.

[2] W. M. McClellan, D. G. Warnock, S. Judd, P. Muntner, R. Kewalramani, M. Cushman, L. A. McClure, B. B. Newsome, and G. Howard, "Albuminuria and racial disparities in the risk for ESRD," *J. Amer. Soc. Nephrol.*, vol. 22, no. 9, pp. 1721–1728, Aug. 2011.

[3] M. K. Haroun, "Risk factors for chronic kidney disease: A prospective study of 23,534 men and women in Washington County, Maryland," *J. Amer. Soc. Nephrol.*, vol. 14, no. 11, pp. 2934–2941, Nov. 2003.

[4] W. D. Souza, L. C. D. Abreu, L. G. D. SilvaI, and I. M. P. Bezerra, "Incidence of chronic kidney disease hospitalisations and mortality in Espírito Santo between 1996 to 2017," Wisit Cheungpasitporn, Univ. Mississippi Medical Center, Rochester, MN, USA, Tech. Rep., 2019, doi: 10.1371/journal.pone.0224889.

[5] W. Mula-Abed, K. A. Rasadi, and D. Al-Riyami, "Estimated glomerular filtration rate (eGFR): A serum creatinine-based test for the detection of chronic kidney disease and its impact on clinical practice," *Oman Med. J.*, vol. 27, no. 4, pp. 339–340, 2012.

[6] A. S. Levey, D. Cattran, A. Friedman, W. G. Miller, J. Sedor, K. Tuttle, B. Kasiske, and T. Hostetter, "Proteinuria as a surrogate outcome in CKD: Report of a scientific workshop sponsored by the national kidney foundation and the US food and drug administration," *Amer. J. Kidney Diseases*, vol. 54, no. 2, pp. 205–226, Aug. 2009.

[7] S. Gerogianni, "Concerns of patients on dialysis: A research study," *Health Sci. J.*, vol. 8, no. 4, pp. 423–437, 2014.

[8] J. R. Chapman, "What are the key challenges we face in kidney transplantation today?" *Transplantation Res.*, vol. 2, no. S1, pp. 1–7, Nov. 2013.

[9] T. Xiuyi and G. Yuxia, "Research on application of machine learning in data mining," in *Proc. IOP Conf., Mater. Sci. Eng.*, 2018, doi: 10.1088/1757-899X/392/6/06220.

[10] B. Zupan, A. J. Halter, and M. Bohanec, "Qualitative model approach to computer assisted reasoning in physiology," in *Proc. Intell. Data Anal. Med. Pharmacol. (IDAMAP)*, Brighton, U.K., 2018, pp. 1–7.

[11] A. Dhillon and A. Singh, "Machine learning in healthcare data analysis: A survey," *J. Biol. Today's World*, vol. 8, no. 2, pp. 1–10, Jan. 2018.

[12] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A review of challenges and opportunities in machine learning for health," in *Proc. AMIA Joint Summits Transl. Sci.*, 2020, p. 191.

[13] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2020.

[14] G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, and J. A. Recio-Garcia, "Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning," *IEEE Access*, vol. 7, pp. 152900–152910, 2019.

[15] Z. Chen, X. Zhang, and Z. Zhang, "Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models," *Int. Urol. Nephrol.*, vol. 48, no. 12, pp. 2069–2075, Jun. 2016.

[16] Y. Amirgaliyev, S. Shamiluulu, and A. Serek, "Analysis of chronic kidney disease dataset by applying machine learning methods," in *Proc. IEEE 12th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2018, pp. 1–4.

[17] K. R. A. Padmanaban and G. Parthiban, "Applying machine learning techniques for predicting the risk of chronic kidney disease," *Indian J. Sci. Technol.*, vol. 9, no. 29, Aug. 2016.

[18] L. Kilvia De Almeida, L. Lessa, A. Peixoto, R. Gomes, and J. Celestino, "Kidney failure detection using machine learning techniques," in *Proc. 8th Int. Workshop ADVANCEs ICT Infrastructures Services*, 2020, pp. 1–8.

[19] W. Gunarathne, K. D. M Perera, and K. A. D. C. P Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in *Proc. IEEE 17th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2017, pp. 291–296.

[20] H. Polat, H. D. Mehr, and A. Cetin, "Huseyin polat1 & homay danaei mehr1 & aydin cetin," *J. Med. Syst.*, vol. 41, no. 4, pp. 1–11, Apr. 2017, doi: 10.1007/s10916-017-0703-x.

[21] S. Drall, G. S. Drall, S. Singh, and B. B. Naib, "Chronic kidney disease prediction using machine learning: A new approach," *Int. J. Manage., Technol. Eng.*, vol. 8, pp. 278–287, May 2018.
[22] M. Almasoud and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, pp. 89–96, 2019.
[23] S. Shankar, S. Verma, S. Elavarthy, T. Kiran, and P. Ghuli, "Analysis and prediction of chronic kidney disease," *Int. Res. J. Eng. Technol.*, vol. 7, no. 5, May 2020, pp. 4536–4541.
[24] S. Vijayarani1 and S. Dhayanand, "Kidney disease prediction using SVM and ANN algorithms," *Int. J. Comput. Bus. Res.*, vol. 6, no. 2, pp. 1–12, Mar. 2015.
[25] J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *J. Transl. Med.*, vol. 17, p. 119, Dec. 2019.
[26] M. S. Gharibdousti, K. Azimi, S. Hathikal, and D. H. Won, "Prediction of chronic kidney disease using data mining techniques," in *Proc. Ind. Syst. Eng. Conf.*, K. Coperich, E. Cudney, H. Nembhard, Eds., 2017, pp. 2135–2140.
[27] E. M. Karabulut, S. A. Ozel, and T. Ibrikci, "A comparative study on the effect of feature selection on classification accuracy," *Procedia Technol.*, vol. 1, pp. 323–327, Jan. 2012.
[28] A. Wosiak and D. Zakrzewska, "Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis," *Complexity*, vol. 2018, Oct. 2018, Art. no. 2520706.
[29] N. A. Nnamoko, F. N. Arshad, D. England, J. Vora, and J. Norman, "Evaluation of filter and wrapper methods for feature selection in supervised machine learning," in *Proc. 15th Annu. Postgraduate Symp. Converg. Telecommun., Netw. Broadcast.*, Liverpool, U.K., 2014, pp. 2–33.
[30] J. M. Pereira, M. Basto, and A. F. D. Silva, "The logistic lasso and ridge regression in predicting corporate failure," *Procedia Econ. Finance*, vol. 39, pp. 634–641, Jan. 2016.
[31] P. G. Scholar, "Chronic kidney disease prediction using machine learning," *Int. J. Eng. Res. Technol.*, vol. 9, no. 7, pp. 137–140, 2020.
[32] B. Deepika, "Early prediction of chronic kidney disease by using machine learning techniques," *Amer. J. Comput. Sci. Eng. Survey*, vol. 8, no. 2, p. 7, 2020.
[33] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network," *Future Gener. Comput. Syst.*, vol. 111, pp. 17–26, Oct. 2020.
[34] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Aliss, "Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data," *Sensors*, vol. 20, no. 9, p. 2649, May 2020.
[35] U. H. Amin, J. Li, Z. Ali, M. H. Memon, M. Abbas, and S. Nazir, "Recognition of the Parkinson's disease using a hybrid feature selection approach," *J. Intell. Fuzzy Syst.*, vol. 39, no. 1, pp. 1–21, Jul. 2020.

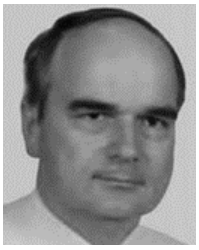
PANKAJ CHITTORA received the B.Tech. degree in information technology from the National Institute of Technology, Durgapur, in 2010, and the M.Tech. degree in computer science and engineering in 2012. He is currently pursuing the Ph.D. degree from the Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur. He has ten years of experience as a Faculty with the Computer Science and Engineering Department. He is doing research in the field of healthcare. His current research interests include data science, machine learning, database, programming, information security, and the analysis of algorithms. He has delivered various subject including data structure, algorithms analysis, theory of computation, compiler construction, and others.

SANDEEP CHAURASIA (Senior Member, IEEE) is currently a Professor with the Department of CSE, School of Computing and I.T., Manipal University Jaipur, Jaipur. He has more than 12 years of rich experience in academics and one year in industry. He has more than more than 30 publications in international/national journals/conference proceedings. His research interests include machine learning, soft computing, algorithms, and artificial intelligence. He is associated with machine learning for more than seven years. He is currently working in the area of application of machine/deep learning in natural language processing like semantic analysis and lexical analysis. He is also guiding four Ph.D. students in the area of NLP, intrusion detection, and food adulteration using AI techniques. He is also an active member of special interest group and initiative by MIR labs to connect the researchers and professional across the globe. He is a Senior Member of LMCSI and MACM, and a member of Machine Intelligence Research Labs, USA. He is also member of reviewer board of various journals and technical program committee of several reputed conferences.

PRASUN CHAKRABARTI (Senior Member, IEEE) received the Ph.D. (Engg.) degree from Jadavpur University, in 2009. He is currently an Executive Dean (Research and International Linkage) and also an Institute Distinguished Senior Chair Professor with the Department of Computer Science and Engineering, Techno India NJR Institute of Technology. He has several publications, books and 31 filed Indian patents in his credit. He has supervised ten Ph.D. candidates successfully. On various research assignments, he has visited Waseda University, Japan, (2012 availing prestigious INSA-CICS travel grant), University of Mauritius, in 2015, Nanyang Technological University Singapore, in 2015, 2016, and 2019, Lincoln University College Malaysia, in 2018, the National University of Singapore, in 2019, Asian Institute of Technology, Bangkok, Thailand, in 2019, and ISI Delhi, in 2019. He is a Fellow of IETE, ISRD, U.K., IAER, London, AE(I), and CET(I).

GAURAV KUMAWAT received the M.Tech. degree from Pacific University. He is currently pursuing the Ph.D. degree from the Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur. He has a total experience of 14 Years as a Faculty with the CSE Department. His research interests include C programming, object-oriented programming using C++, core java, advance java, unix programming, web programming, operating systems, computer architecture, information security, programming analysis, and data analysis.

TULIKA CHAKRABARTI received the Ph.D. (Sc.) degree from the Indian Institute of Chemical Biology, Jadavpur University, in 2013. She is currently an Assistant Professor (Senior Grade) with Sir Padampat Singhania University, Udaipur. She has several publications, books and filed patents to her credit. She is a national merit scholarship holder in both 10th and 12th grade. She has visited NUS, NTU, Lincoln University College, Malaysia, and AIT, Bangkok, on several academic assignments.



ZBIGNIEW LEONOWICZ (Senior Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from the Wroclaw University of Science and Technology, in 1997 and 2001, respectively, and the Habilitation degree from the Bialystok University of Technology, in 2012. Since 1997, he has been with the Electrical Engineering Faculty, Wroclaw University of Technology. He also received the two titles of a Full Professor from the President of Poland, in 2019, and the President of the Czech Republic. Since 2019, he has been a Professor with the Department of Electrical Engineering, where he is currently the Head of the Chair of electrical engineering fundamentals.

MICHAŁ JASI SKI (Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from the Wroclaw University of Science and Technology, in 2016 and 2019, respectively. Since 2018, he has been with the Electrical Engineering Faculty, Wroclaw University of Technology, where he is currently an Assistant Professor. He has authored or coauthored more than 60 scientific publications. His research interests include using big data in power system especially in point of power quality.

ŁUKASZ JASI SKI received the degree in electrical engineering from the Wroclaw University of Technology, Poland, and the Ph.D. degree in dyscypine of automation, electronics and electrical engineering from the Wroclaw University of Technology. In 2020, he started working as a designer of electrical installations in the design office. His research interest includes big data in power systems, especially in point of power quality.

RADOMIR GONO (Senior Member, IEEE) received the M.Sc., Ph.D., Habilitate Ph.D., and Professor degrees in electrical power engineering in 1995, 2000, 2008, and 2019, respectively. Since 1999, he has been with the Department of Electrical Power Engineering, VSB–Technical University of Ostrava, Czech Republic, where he is currently a Professor and the Vice Head of the department. His current research interests include electric power systems reliability, the optimization of maintenance, and renewable energy sources.

ELŻBIETA JASI SKA received the Ph.D. degree from Pozna University Technology, in 2013. Since 2019, she has been with the Faculty of Law, Administration and Economics, University of Wroclaw, where she is currently an Assistant Professor. She has authored or coauthored more than 70 scientific publications. Her research interests include sustainable development, corporate socially responsible, renewable energy resources, and virtual power plants.

VADIM BOLSHEV received the M.S. degree in electrical engineering from Orel State Agrarian University, in 2012, and the Ph.D. degree in electrical engineering in 2020. He gained experience in the industry as an electrician, an electrical engineer, a chief engineer from 2010 to 2018. Since 2018, he has been a Researcher with the Laboratory of Power and Heat Supply, Federal Scientific Agroengineering Center VIM. His scientific activity is to develop methods and tools aimed at improving power supply efficiency including the development of methods and devices for monitoring power quality and the technical state of power supply system elements.