Prediction of Chronic Kidney Disease Using Data Mining Feature Selection and Ensemble Method

Sirage Zeynu
Research Scholar, Department of Computer Science & Engineering
Symbiosis Institute of Technology
Pune, India
Sirage.ahmed@sitpune.edu.in

Shruti Patil
Assistant Professor, Department of Computer Science & Engineering
Symbiosis Institute of Technology
Pune, India
Shruti.patil@sitpune.edu.in

Abstract—The failure of the kidney is affected the whole human body and it can be a cause of the seriously ill and cause of deaths. Machine learning and data mining techniques are the most significant role in disease prediction with high-performance rate and used to help decision makers to assemble and understand information. The performance of classification techniques depends on the feature of the data set. To improve the accuracy of classification used feature selection method by reducing the dimensions of the feature and used ensemble or combine a model of the algorithm. In this research K-Nearest Neighbor, J48, Artificial Neural Network, Naïve Bayes and Support Vector Machine classification techniques were used to diagnose Chronic Kidney Disease. To predict chronic kidney disease, build two important models. Namely, feature selection method and ensemble model. To build chronic kidney disease prediction, used Info gain attributes evaluator with ranker search engine and wrapper subset evaluator with the best first engine was used. The result showed that the K-nearest neighbor classifier by using Wrapper Sub set Evaluator with Best first search engine feature selection method has 99% accuracy, J48 with Info Gain Attribute Evaluator with ranker search engine has 98.75, Artificial Neural Network with Wrapper Sub set Evaluator with Best first search engine has 99.5% accuracy, Naïve Bayes with Wrapper Sub set Evaluator with Best first search engine has 99% accuracy, Support Vector Machine with Info Gain Attribute Evaluator with ranker has 98.25% accuracy in prediction of chronic kidney disease compared to other with and without feature section method. The second model building method ensemble model by combing the five heterogeneous classifiers based on a voting algorithm. The effectiveness of the proposed ensemble model was examined by comparison of the base classifier. The experimental result showed that the proposed ensemble model achieved 99% accuracy.

Keywords— Chronic Kidney Disease, Data Mining, Classification Techniques, Feature Selection, Ensemble model, accuracy, prediction

1. INTRODUCTION

Data mining refers to mining important information about the different huge amount of dataset and one of the significant stages in realizing knowledge [1]. Data mining important role in several real-world applications such as business organization, healthcare sector, education, scientific, government sector, and any organization. In the medical domain, data mining is used for mainly disease prediction. Data mining is significant research doings in the field of healthcare sectors to predict and detect disease. There is a requirement of well-organized methodologies for analyzing, predict and detecting diseases [1], [8], [9]. To detect and predict diseases Data mining applications are used for the management of healthcare, health information, patient care system, etc. It also plays a major role in analyzing survivability of a disease [1], [2], [8], [9].

Data mining, classification techniques play a vital role in healthcare domain by classifying, detecting, analyzing and predicting the diseases dataset [6], [10]. The classification algorithm like artificial neural network (ANN), K-nearest neighbor (KNN), naïve Bays, decision tree (J48, C4.5), support vector machine (SVM) etc. Are used to classify, analysis, detected and predict medical datasets.

Feature selection in data mining and machine learning concepts is the key to knowledge discovery, pattern recognition and statistical sciences [6]. The main aim of feature selection to remove some part of the attribute from the data set that is not relevant in that dataset [6], [10]. Removing some feature used to improve the performance accuracy of the

classifier. Feature selection can be grouped into a wrapper and filter methods [6], [10], [12], [17], [19].

Ensemble algorithms are a machine learning algorithm that combines the prediction from heterogeneous machine learning classifiers. Ensemble model is one of the most significant to create great accurate prediction models. The most example of ensemble models used to solve machine learning, data mining, and data science are random forest Bagging, Boosting, stack and vote algorithms.

Chronic kidney disease (CKD) known as a chronic renal failure. Chronic Kidney Disease (CKD) or chronic renal disease gradually serious problem in the world. In which the kidney drops its functionality and it is the cause of the inappropriate functionality of kidney organs [2], [6]. The beginning date of kidney letdown may not be known the exact time, it may not identify as a disease of the patient because it cannot show any symptoms initially [6]. The failure of the kidney is affected the whole human body and it can be a cause of the seriously ill and cause of deaths.

Now a day from the global burden disease project, CKD disease is CKD is rapidly growing through the globe. The statistical report indicates that 90 % increased the loss of life among the patient with chronic kidney disease since 1990 to 2013. CKD disease is the known 13^{th} ranking cause of death in the world [28]. According to kidney international report, CKD was one of the top 5 cause of death in the different country [28] among the top cause of disease. According to the national kidney foundation, 10% of the world population infected by CKD and millions of people die yearly all over the world [29]. The cause of the death is the shortage of treatments and lack of knowledge about kidney disease.

In developing country, most of the kidney patient received treatment after reached in serious cases. This increases the number of CKD patients [28]. CKD can be reduced even can stop by diagnosis before affected and during affected by doing the test like the blood test, urine test, kidney scan and ask doctor other symptoms of kidney disease.

In this study, we have examined the accuracy rate of the methods using feature selection by reducing the dimensionality of the feature and combining heterogeneous classifiers to create the ensemble model.

The rest of this research is organized as follows: section 2 related to the literature review, section 3 methodology, section 4 experimental test result and discussion, section 5 conclusion.

2. LITERATURE REVIEW

Classification techniques, Feature selection, and Ensemble model are the most significant and vital tasks in machine learning and data mining. A lot of research has been conducted to apply data mining and machine learning classification technique, feature selection method and ensemble model on different medical datasets to classify disease datasets. Many of them show good classification accuracy.

Polat, H et al. [6] Diagnosis chronic kidney disease using SVM and effective feature selection methods. They used wrapper and filter feature selection method to reduce the dimensionality of the feature. In their work they improve accuracy by implanting SVM without feature selection the accuracy rate was 97.75%, SVM with the classifier subset evaluator combine with greedy stepwise the accuracy rate was 98%, SVM with the wrapper subset evaluator combine with a best first search engine the accuracy rate was 98.25, SVM with the classifier subset evaluator combine with greedy stepwise the accuracy rate 98.25. And finally, SVM with the filter subset evaluator combine with best first search the accuracy was 98.5.

Bashir, S. et al [15] they proposed ensemble classifier which uses majority Vote Based framework for prediction of heart disease. They used five heterogeneous classifiers used to construct the ensemble model. The classifiers are Naïve Bayes, decision tree based on Gini Index, decision tree based on information Gain, memory based learner and SVM. After experiment using stratified cross-validation show that their MV5 framework has achieved an accuracy 88.5% with 86.96% sensitivity, 90.83% specificity and 88.85 F-Measure and they compare with the base classifiers show to increase the average accuracy of the ensemble model. They involved proposing the ensemble approach. The first approach generates the individual classifier decision and the second approach is combine the individual classifier decision correctly to create the new combine model.

Bashir, S., [33] proposed HMV medical decision support framework using multi-layer classifier for disease prediction. They proposed based on the optimal combination of the heterogeneous classifier to create the ensemble model. The classifiers are Naïve Bayes, Linear Regression, and quadratic discriminate analysis, KNN, SVM, Decision Tree using Gini Index, and Decision Tree using Information Gain. So, their HMV ensemble framework outperforms the other prediction models. HMV framework was proposed based on three modules. The first module was data acquisition and preprocessing. The second module was used to predict unknown class label for test set instances. The third module used to predict and evaluate the proposed HMV ensemble model. After applying all the selected data set the HMV ensemble model achieved highest accuracy disease classification and prediction.

Naghmeh Khajehali et al.[4] were presented by extracting factor affecting for pneumonia patients by using data mining techniques. They proposed modeling by using feature selection and classification with ensemble methods to preprocess, reduce dimensionality and classify the raw data. In their work, the design consists of different stages of preprocessing and used Bayesian Boosting method for constructed which identify factor related to patient LOS in hospital. The construction of modeling based on the data set SVM and ensemble method like AdaBoost, Vote, Stacking, Bayesian Boosting. Among these classifier techniques, Bayesian Boosting used for analysis of data by using 10 fold cross-validation method. In this work the data set was divided into 10 subsets, the training subset participated 10 times. Out of 10 subsets 9 were classified as the training set. The result indicated that the Bayesian Boosting ensemble technique was

scored a better result. The Bayesian boosting ensemble technique, accuracy was 97.17%, which is high performance used to predict pneumonia disease in anticipation of LOS.

Pritom, A. I et al. [12] applied a classification algorithm for Predicting Breast Cancer Recurrence by using SVM, Decision tree, Naïve Bays and C4.5. They enhanced the accuracy of each classifier with the help of effective feature selection methods. They improve the accuracy by using Info Gain attribute with ranker search engine. After implemented on weka tool the recurrence prediction accuracy has SVM achieved 75.75% accuracy, J48 achieved 73.73% and naïve bays achieved 67.17%. These are the original data set without feature selection. After carefully applied feature selection SVM enhanced bay 1.52%, C4.5 enhanced by 2.52% and Naïve Bays enhanced by 9.09%.

Dulhare, U. N. et al. [10] Built classification models, Used feature selection to extract an action rule and predict CKD by using naïve Bayes classifier and one R attribute selector to predict and classify the CKD and none CKD patients. These methods are Naïve bays with the wrapper subset evaluator combine with the best first search. After implemented on weka tool using wrapper subset evaluators combine with the best first search engine; Naïve Bay's classifier achieved 97.5 % accuracy rate.

Many researchers have been conducted different data mining, classification algorithm like KNN [1], [2],[4], [7], [13], [14],[16], [20], [21], [25], ANN [2], [4], [7], [11], [16], [18], [19], [22], [24], Naïve Bays[1, 4, 7,8, 10, 12, 16, 20, 23,25], SVM[2], [5], [6], [12] ,[20], [24], [25] Decision Tree(J48/C4.5) [2], [3], [4], [5], [7], [8], [9], [12], [16], [20], [23], feature selection [6], [10], [12], [17], [19], [21], [26] and ensemble [15], [32] to improve the performance accuracy of algorithm.

Classification algorithms are supervised learning method their class is known to predict the objective class level. The classification used to categorize datasets into training and test set. Data mining, classification is commonly used in healthcare application to classify patient dataset [2]. In most works different machine learning algorithms are used such as an artificial neural network (ANN), K-nearest neighbor (KNN) and decision tree (J48), naïve Bayes, SVM, etc. are used to classify diseases dataset.

2.1. Artificial neural network

Artificial neural network (ANN) also called "neural network', widely applied in the real application based on natural neurons. ANN contains the connected nodes of artificial neurons and interconnects each node by adaptable weights for each node and change its prearrangement throughout message transfer [4]. ANN is learning algorithm, it can learn and adapt to change its structure during information received from the internal and external environment during learning [22]. ANN contains a set of layers to pass a number of messages. The layers are input layer, hidden layers, and an output layer. The hidden layers contain one or more layers with the number of nodes. These three layers are organized to each other, in which weight is linked with each node. ANN is supervised learning which is the input is participating in the network to produce output. The basic operation unit of ANN known as perception. Perceptron can able to classify the dataset into two classes. Perceptron consist of single node carries with weights. Perceptron has three fundamental elements which are link, adder and activation functions.

2.2.K-nearest Neighbor

KNN is a type of supervised learning algorithm and by nature, it is nonparametric [1]. No need to separate linear and nonlinear. KNN is good for a large number of records and fast to train the models. KNN finds objects from the input k-number of objects that are nearest to the exact point query or majority vote. It works founded on the nearby class object that has the shortest distance from request example to the training example. According to [2] KNN is the fastest algorithm in its execution time to build models. KNN gathering all the neighbor objects apply simple majority vote to the prediction query. For each query Xn to be classified x1, x2, … xk are the k instances. The nearest class will be recognized by using different distance measurements such as Euclidean distance, Manhattan distance, Minkowski distance and Hamming distance. The distance formula as follows.

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{n}(Xi - Yi)2}$$

$$\text{Manhattan} = \sum_{i=1}^{n} |Xi - Yi|$$

$$\text{Minkowski} = \sum_{k}(|Xi - Yi|)^{q}|)^{\frac{1}{q}}$$

2.3.Decision Tree (J48)

A J48 decision tree is an open source Java implementation of C4.5 decision tree algorithm in the weka platform [9]. It is the extension of the earlier ID3 algorithm, which is developed by Ross Quinlan [9]. J48 classification method uses top-down greedy search methods for constructing tree [23]. J48 decision tree produces sorting tree whose, leaf denotes the ending class and the internal attributes represent a possible number of outputs of the branch features [23]. It is a division between information gain and its splitting attributes. Entropy is the measure of the disorder data. Any random variable, entropy is a measure of uncertainty. For any probability p and sample S, entropy can be calculated as:

$$Entropy(S) = \sum_{i=1}^{n}(-\text{pi}\log 2(p_i))$$

Information gain, which is identifying the best attribute for selecting the exact node in a tree. To compute the value of attribute A, which respect to sampling S, the value of attribute called as the value of (A), the value of sample S from Sv where A is an attribute, S is sample and v is value. So we calculate information gain as follows.

$$Gain(S, A) = \text{Entropy(S)} - \sum_{v \in value\ (A)}^{n}\left(\text{Entropy(Sv)}\frac{|Sv|}{|S|}\right)$$

2.4.Support Vector Machine

Support Vector Machine (SVM) is a machine learning, supervised learning algorithm on the base of statistical learning concepts. SVM has the high-performance capability to predict, analyses, regression and classifies dataset [6]. It generates a distinct hyperplane in the descriptor interplanetary of the training data and mixes are classified based on the crosswise of hyperplane located. It is used to predict and analyze the dataset regression and classification techniques [1]. SVM is supervised learning algorithms which are mostly used data mining classification. SVM gives the correct result by associating other classification algorithms. By maximizing the combined between the instances of two classes, it can minimize the error. The benefit of the SVM is that by use of ''kernel trick'', the distance between a particle and the hyperplane can be calculated in a transformed (nonlinear) feature space, lacking the explicit transformation of the original descriptors.

2.5.Naïve Bayes

Naïve Bayes classifier is a probabilistic classification algorithm based on Bayes theorem with independent assumption features. Naïve Bayes classification algorithm performs tasks well and learns quickly in numerous real-world supervised classification problems [1]. Naïve Bayes is used for diagnosis and prediction of the world problem. The Naïve Bayes algorithm requires a less number of training data through classification to predict and evaluate the parameter [21]. The Naïve Bayes classification method used to predict, an associate of each class. For instance the probability for the specified record for the target class. The class which has maximum probability is expected the most likelihood class. Below is Bayes theorem.

$$\mathrm{P(Y/X)} = \frac{P(X/Y).P(Y)}{P(X)}$$

P (X) is similar to whole classes and P (Y) = relative frequency of class Y

3. METHODOLOGY

The proposed research consists of two methods. The first method is constructing a prediction model by using different feature selection method. The second method is constructing a prediction model by using ensemble or combining heterogeneous classifiers.

3.1.Feature selection methods

The feature selection method is used to reduce the dimensionality of the feature and remove irrelevant features from the data set can create a complete model for classification. We used Info Gain Attribute Evaluator feature selection method combine with ranker search method to select most relevant features. Info Gain Attribute Evaluator is evaluating the value of an attribute by measuring the information gain with respect to the classes. Info Gain Attribute Evaluator can binary numeric attributes instead of properly discretizing the features. It can also distribute the missing value across other values in percentage to their mean value for the numeric attribute and frequent value for a categorical attribute or treated as a separate value. Info Gain Attribute Evaluator has a capability of identifying Empty nominal attributes, Missing values, Date attributes, Numeric attributes, and Unary attributes, Binary attributes, Nominal attributes.

The ranker search method used to calculate ranks attribute by their individual evaluator in conjunction with the attribute evaluator like gain ratio and entropy. It has the capability of generating attribute ranking.

The other feature selection we used in our research wrapper subset evaluator combine with the best first search method. Wrapper subset evaluator which, evaluates attribute sets by using learning pattern. It used to cross-validation to estimate the accuracy of learning pattern for a set of attributes. It **w**as capable of determining the Missing class values, the Nominal class, the Binary class, the Date class and the Numeric class and identifies the type of attributes, String attributes, Empty nominal attributes, Missing values, Date attributes, Relational attributes, Numeric attributes, Unary attributes, Binary attributes, Nominal attributes.

Best first search Searches the space of attribute subsets by greedy hill climbing increased with a backtracking capacity. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point).

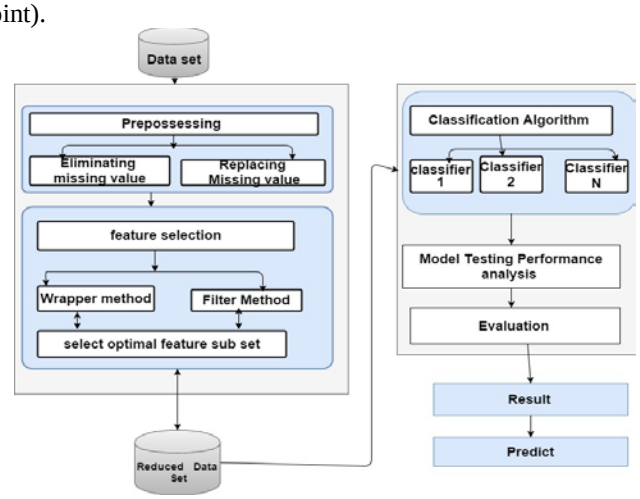


Fig. 1. system architecture for feature selection method

3.2.Ensemble classifiers

Ensemble learning is a machine learning algorithm that creates a number of ensemble prediction models and combines their outputs to increase the performance metrics of the individual algorithm and ensemble with the highest heterogeneous classifier have a tendency to produce the best accuracy rate [32]. The greatest way of using an ensemble classifier to correct errors made by the base classifier [32]. Now a day in machine learning combines classification is very popular with us multiple classifiers instead of one single

classifier. The advantage instead of one classifier algorithm power we can use more than two or more classification algorithm. So the model we build will be more powerful and sophisticated to classify instances from the training set, cross-fold validation or testing set.

The ensemble classification model is to combine varied classifier that is different on result individual [15]. The methods have changed the training process in order to generate classifier model that generate output in different classification results [15]. The main advantage of ensemble method combines the individual classifier rules will strong prediction as compared to the individual classifier rules. The principle of ensemble model combines a heterogeneous individual classifier together and produce superior predictive power.

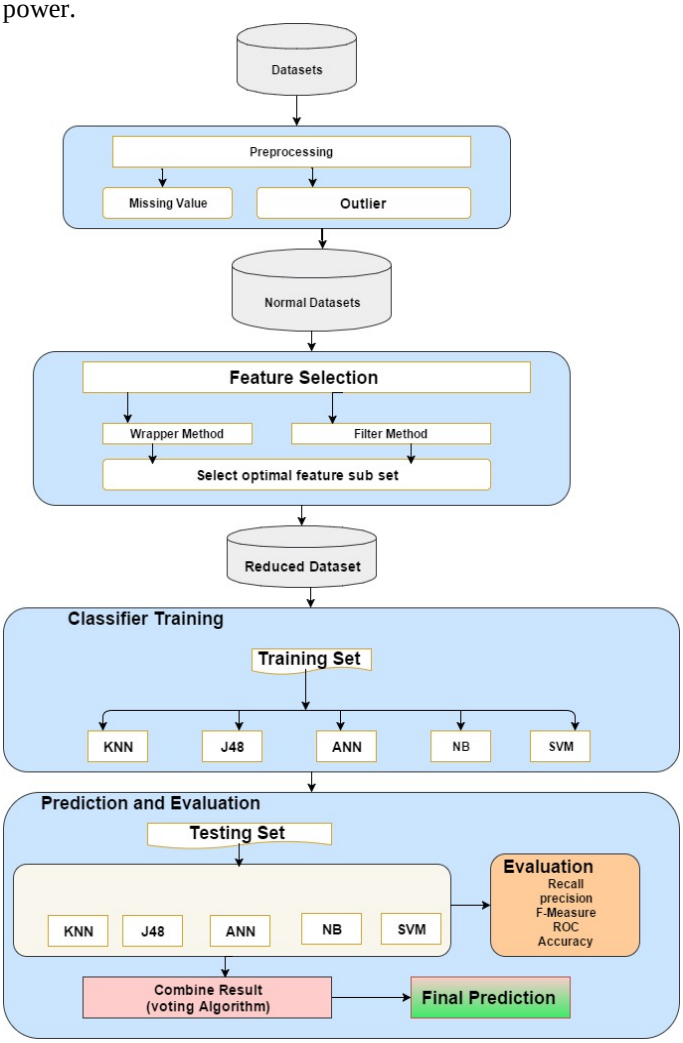


Fig 2. architecture of proposed ensemble model

The general methodology of the proposed system, the dataset is taken from the UCI machine learning repository. The data set is prepared inside weka in the form of .arff files. We use weka replaces all missing values for nominal and numeric attributes in a dataset with the modes and means of the training data. Then reduce the dimensionality of the feature using feature selection methods. After reducing feature we get the optimal feature subset. The reduced subset dataset used for proposed work. The reduced subset dataset is selected relevant features. After performed feature selection on each attribute value individually, then pass it to the base classifier. The base classifiers are KNN, J48, ANN, NB, and SVM. The data are divided into training and testing set. Training data set used to train the base classifier. Testing set used to evaluate and predict the diseases. Apply ensemble vote algorithm to combine the classifier to produce improved results and then make a final prediction is achieved. The models are evaluated by performance metric. It used to evaluate results on the basis of accuracy, precision, recall, F-Measure, and ROC. The general proposed system architecture is shown in figure 2

4. EXPERIMENTAL TEST RESULTS

4.1.Dataset

The dataset was collected from UCI machine learning repository [27]. The dataset contains 400 instances with 24 attributes and 1 class attributes. These attributes are presented in the following table. The dataset contains 400 instances (250 CKD, 150 notCKD) and number of Attributes: 24 + class = 25 (11 numeric, 14 nominal)

TABLE I. *THE ATTRIBUTE OF CHRONIC KIDNEY DISEASE*

No	Attributes	Type of Attribute	Explanation
1	age	numerical	age
2	bp	numeric	blood pressure
3	sg	nominal	specific gravity
4	al	nominal	albumin
5	su	nominal	sugar
6	rbc	nominal	red blood cells
7	pc	nominal	pus cell
8	pcc	nominal	pus cell clumps
9	ba	nominal	bacteria
10	bgr	numeric	blood glucose random
11	bu	numeric	blood urea
12	sc	numeric	serum creatinine
13	sod	numeric	sodium
14	pot	numeric	potassium
15	home	numeric	hemoglobin
16	pcv	numeric	packed cell volume
17	wc	numeric	white blood cell count
18	rc	numeric	red blood cell count
19	htn	nominal	hypertension
20	dm	nominal	diabetes mellitus
21	cad	nominal	coronary artery disease
22	appet	nominal	appetite
23	pe	nominal	pedal edema
24	ane	nominal	anemia
25	class	nominal	class

[a] The attribute of chronic kidney disease

4.2.Performance metrics

Confusion matrix: It is a table that is used to refer to the performance of learning algorithm by computing the performance metrics. Confusion matrix shows correctly and incorrectly predictive made by classification model compared to the actual outcomes or targeted value in the dataset.

TABLE II. *CONFUSION MATRIX*

b. The aconfusion matrix

In our experiment, there are two predicted classes: "CKD" and "not CKD". If we were predicting the presence of a disease, for example, "CKD" would mean they have the chronic kidney disease, and "not CKD" would mean they don't have the chronic kidney disease.

In this experiment for calculating the performance following metrics have been used.

True Positive (TP): the predicted indicates positive occurrences correctly classified as positive outputs that predict they have CKD.

True Negative (TN): The predicted indicates negative instances correctly classified as negative outputs. That predict not CKD they do have chronic kidney disease.

False Positive (FP): the predicted CKD, but they don't actually have the chronic kidney disease the prediction indicates negative instances wrongly classified as positive outputs

False Negative (FN): the predicted not CKD but they actually do have chronic kidney disease. It indicates positive instances wrongly classified as negative output

Accuracy: accuracy implies the capability of classification algorithm to predict of the classes of the dataset.it indicates how the classifiers are correctly classified.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Recall: recall is also called sensitivity that retrieved relevant instances.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Precision: based on a measure of relevant it retrieved information that is relevant instances.

Confusion Matrix				
	Positive	Negative	Target value	
Positive	TP	FN	Positive Predictive value	$\frac{TP}{TP+FP}$
Negative	FP	TN	Negative predictive Value	$\frac{TN}{FN+TN}$
	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	Accuracy=$\frac{TP+TN}{TP+FP+FN+TN}$	
	Precision	Recall		

$$\text{Precision} = \frac{TP}{TP+FP}$$

F-Measure: it is also called F-score. It is a measure of a test accuracy. This is a biased mean of the recall and precision

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

Receiver Operating Characteristics (ROC): It is used for evaluating the test result. It is mostly represented by a graph that visualized the performance of classification algorithm all the threshold values. It is generated by plotting the true positive rate against the false positive rate. It is used to visualize comparison of a classification model and it shows the tradeoff between the true positive rate and the false positive rate. The area under the ROC curve is a measure of the accuracy of the model. ROC curve lies always between 0 and 1or 0<ROC<1. If the model is near to 1 it is better to model. The area under ROC curve is called AUC (area under the curve).in ROC cure we having the X axis is represented by false positive rate and true positive is represented by true positive rate or recall.

TABLE III. *PERFORMANCES OF CLASSIFIERS WITH AND WITHOUT FEATURE SELECTION*

classification by using with and with feature selection methods	Precision	Recall	F-Measure	Accuracy%
KNN Without feature selection	0.985	0.985	0.985	98.5
KNN with InfoGainAttributeEval with ranker (selected 20 Attribute)	0.988	0.988	0.988	98.75
KNN with InfoGainAttributeEval with ranker (selected 15 Attribute)	0.98	0.98	0.98	98
WrapperSubsetEval with Best first search engine (selected 8 attribute)	0.99	0.99	0.99	99
J48 Without feature selection	0.967	0.968	0.967	96.75
J48 with InfoGainAttributeEval with ranker (selected 20 Attribute)	0.987	0.988	0.987	98.75
J48 with InfoGainAttributeEval with ranker (selected 15 Attribute)	0.987	0.988	0.987	98.75
J48 WrapperSubsetEval with Best first search engine (selected 7 attribute)	0.973	0.973	0.972	97.25
ANN Without feature selection	0.978	0.978	0.978	97.75
ANN with InfoGainAttributeEval with ranker (selected 20 Attribute)	0.981	0.98	0.98	98
ANN with InfoGainAttributeEval with ranker (selected 15 Attribute)	0.976	0.975	0.975	97.5
ANN with WrapperSubsetEval with Best first search engine (selected 8 attribute)	0.995	0.995	0.995	99.5
NB Without feature selection	0.951	0.945	0.946	94.5
NB with InfoGainAttributeEval with ranker (selected 20 Attribute)	0.952	0.948	0.948	94.75
NB with InfoGainAttributeEval with ranker (selected 15 Attribute)	0.946	0.94	0.941	94
NB with WrapperSubsetEval with Best first search engine (selected 9 attribute)	0.99	0.99	0.99	99
SVM Without feature selection	0.979	0.978	0.978	97.75
SVM with InfoGainAttributeEval with ranker (selected 20 Attribute)	0.983	0.948	0.983	98.25
SVM with InfoGainAttributeEval with ranker (selected 15 Attribute)	0.979	0.978	0.978	97.75
SVM with WrapperSubsetEval with Best first search engine (selected 8 attribute)	0.98	0.98	0.98	98

Feature selection which includes info gain attribute evaluator combine with ranker search engine and WrapperSubsetEval with Best first search engine. To reduce the dimensionality of the dataset and improve the accuracy of CKD prediction. All the used methods can produce a new dataset by lower dimensional than the original dataset. The summary of all selected classifier results value with and without using feature selection methods are shown in the following tables.

The first feature selection method is infoGainAttributeEval evaluator with ranker search engine reduce dataset dimension to 20 attributes for each classifier. The second reduced dimension method reduces from the reduced dataset by using infoGainAttributeEval evaluate and ranker search engine reduce dataset dimension to 15 attributes for each classifier. The third feature selection method is WrapperSubsetEval evaluator with Best First search engine reduced dataset dimension to 8 attributes. K-nearest neighbor's classifier can select an appropriate value of K based on cross-validation by calculating the distance weighting. We used all the 25 features for KNN without feature selection. For J48, WrapperSubsetEval evaluator with Best First search engine reduced dataset dimension to 7 attributes. For ANN, WrapperSubsetEval evaluator with Best First search engine reduced dataset dimension to 7 attributes. For Naïve Bayes, WrapperSubsetEval evaluator with Best First search engine reduced dataset dimension to 8 attributes. For SVM, WrapperSubsetEval evaluator with Best First search engine reduced dataset dimension to 8 attributes. In this study, WEKA (version 3.8) was used for feature selection and we used Java (NetBeans 8.0.1) and weka jar file to build the models. The experimental result of each classifier with and without feature selection method is shown in Table III.

The performance metric the classifiers are shown in table 3 without and used with feature selection. For each CKD and not CKD class's precision, recall, F-Measure were present used weighted average. From table 3 the accuracy value of CKD prediction for KNN classifier on reduced (selected 8 attributes from 25) dataset by wrapperSubSetEva and Best First search engine feature selections are most acceptable it has the highest weighted average value of precision, recall, F- measure and

[c.] precision, recall, F-measure and accuracy of the classifier with and without feature selection

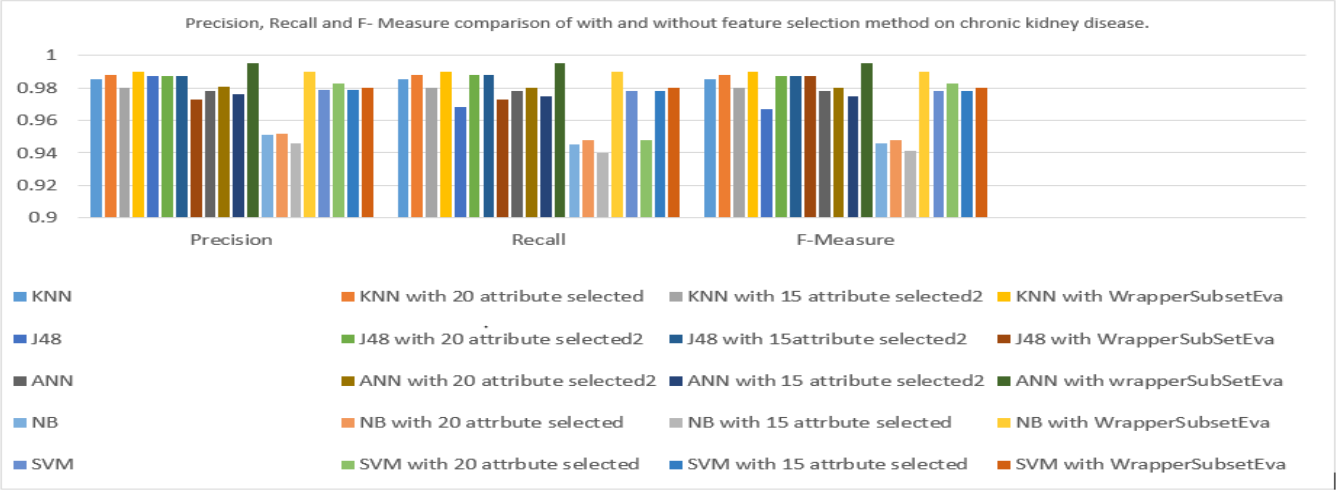
accuracy. KNN without used feature selection has accuracy 98.5% in prediction CKD. KNN with InfoGainAttributeEval with ranker (selected 15 Attribute) has list accuracy rate 98% it reduced the accuracy compared to the normal data set because the removed attributes are important in this method. KNN with InfoGainAttributeEval with ranker (selected 20 Attribute) has accuracy 98.75% it improved to from 98.5% to 98.75% the normal data.

From table 3 the accuracy value of CKD prediction for J48 classifier on the reduced dataset by InfoGainAttributeEval with ranker (selected 15 or 20 Attribute) feature selections are most acceptable it has the highest weighted average value of precision, recall, F- measure. Both were achieved the same accuracy 98.75%. J48 without used feature selection has the least accuracy 96.5% in prediction CKD. The reduced the dimension by WrapperSubsetEval with Best first search engine (selected 7 attributes), the J48 classifier has an accuracy rate of 97.25. It is higher than the accuracy rate of 25 attributes of the dataset.

ANN classify without used feature selection method has accuracy rate (97.75%) in the prediction of CKD. After applied ANN with InfoGainAttributeEval with ranker (selected 20 Attribute), feature selected method was 98%. ANN with used ANN with InfoGainAttributeEval with ranker (selected 15 Attribute) feature selection method has least accuracy rate (97.5). It is lower than the accuracy rate of 25 dimensions of the dataset. Finally, ANN classifier on CKD dataset whose dimension has reduced to by using ANN with WrapperSubsetEval with Best first search engine (selected 8 attributes) has the highest accuracy rate (99.5).

Naïve Bayes classify without used feature selection method has accuracy rate (94.5%) in the prediction of CKD. After applied naïve Bayes by InfoGainAttributeEval with ranker (selected 20 Attribute) feature selected method was 94.75%. Naïve Bayes by used InfoGainAttributeEval with ranker (selected 15 Attribute) feature selection method has least accuracy rate (94). It is lower than the accuracy rate of 25 dimensions of the dataset. Finally, Naïve Bayes classifier on CKD dataset whose dimension has reduced to by using WrapperSubsetEval with Best first search engine (selected 9 attributes) has the highest accuracy rate (99.5).

Fig. 3. Precision, Recall, and F- Measure comparison of with and without feature selection method on chronic kidney disease



SVM classify without used feature selection method has lower accuracy rate (97.75%). The accuracy rate of SVM by using InfoGainAttributeEval with ranker search engine (selected 20 Attribute) feature selected method has the highest accuracy (98.25%). SVM by used InfoGainAttributeEval with ranker (selected 15 Attribute) feature selection method has least accuracy rate (97.75%). It was the same as 25 dimension dataset. Finally, SVM classifier on CKD dataset whose dimension has reduced to by using WrapperSubsetEval with Best first search engine (selected 9 attributes) has the highest accuracy rate (98%). However, it is lower than the 20 dimension of data set.

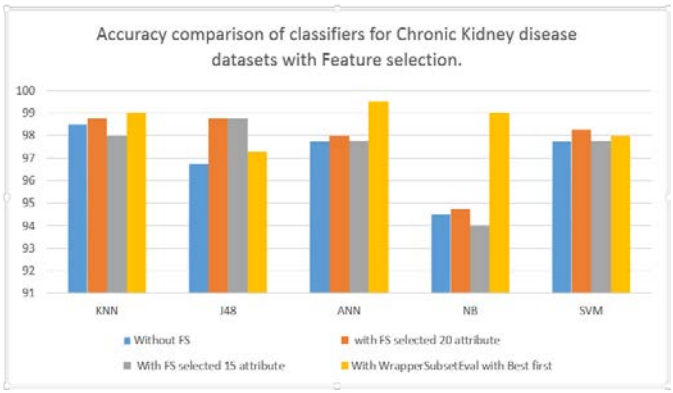


Fig.4. Accuracy comparison of classifiers for Chronic Kidney disease datasets with Feature selection

Our ensemble classifier model performs on CKD dataset as shown table. It performs the accuracy, precision, recall, F-Measure, true positive rate and ROC comparison in performance of the individual classifier. The ensemble model achieved the high accuracy rate in CKD.

TABLE IV. PERFORMANCES OF CLASSIFIERS WITHOUT AND WITH FEATURE SELECTION AND ENSEMBLE METHODS

Classifier	Precision	Recall	F-Measure	Accuracy%
KNN	0.985	0.985	0.985	98.5
KNN after FS	0.98	0.98	0.98	99
J48	0.967	0.968	0.967	96.75
J48 after FS	0.987	0.988	0.987	98.75
ANN	0.978	0.978	0.978	97.75
ANN after FS	0.995	0.995	0.995	99.5
NB	0.951	0.945	0.946	94.5
NB after FS	0.99	0.99	0.99	99
SVM	0.979	0.978	0.978	97.75
SVM after FS	0.983	0.948	0.983	98.25
Ensemble model	0.99	0.99	0.99	99

[a.] comparison of Ensemble model with other methods

The proposed ensemble model outperforms the other base classifier and used with most of our proposed feature selection method. The proposed ensemble model comparison of heterogeneous base classifier without and with feature selection method based on performance metrics. The proposed ensemble model has used the reduced dimension that is we used in feature selection method.it used to reduce the computation cost the training time. Feature selection is used to reduce the cost and execution time using training and test set. Because the dimension gets to reduce the cost and execution time lower and get high accuracy. From table 4 that ensemble models produce the highest accuracy level for CKD dataset when comparing the base classifier. Our ensemble model achieved high accuracy of 99%, 99% precision, 99% recall and 99% F-Measure. Table 4 shows a comparison of accuracy precision, recall, and F-Measure with the base classifier with feature selection and without feature selection. The analysis of the results shows that ensemble model has achieved the highest accuracy, precision, recall, and F-Measure.

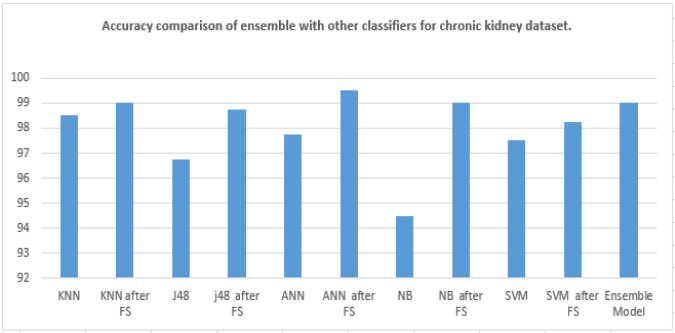


Fig.5. Accuracy comparison of the ensemble with other classifiers for chronic kidney disease dataset

5. CONCLUSION

In this study feature selection method and ensemble method has been utilized on data set of CKD dataset to improve the accuracy rate of the classifiers. Different feature selection evaluator has been used for each classifier. For feature selection method InfoGainAttributeEval with ranker search engine and WrapperSubsetEval with Best first search engine have been used. These methods have been used both proposed feature selection method and ensemble model to improve the accuracy of machine learning classifiers. The accuracy rate of KNN, J48, ANN, NB, and SVM classifier on CKD dataset has been compared to its accuracy, precision, recall, and F-Measure on a reduced dataset which has been used WrapperSubsetEval with Best first search engine and InfoGainAttributeEval evaluator for feature selection method. The experimental result shows that after reducing the dimension of the dataset the accuracy of the classifier has been improved. The accuracy rate of KNN classification reduced dataset by WrapperSubsetEval with Best first search engine was 99%, which is more than the original dataset and other feature selection methods. The accuracy rate of J48 classification reduced dataset by InfoGainAttributeEval with ranker search engine was 98.75. Which was more than the original and other feature selection method. The accuracy of ANN classification on the reduced dataset by WrapperSubsetEval with Best first search engine was 99.5%, which was the highest accuracy among all other methods. The accuracy rate of Naïve Bayes classification on reduced data set by WrapperSubsetEval with Best first search engine was 99 %, which was high accuracy rate compared to the original dataset and other feature selection methods. The accuracy of SVM classification on the reduced dataset by InfoGainAttributeEval with ranker search engine was 98.25%,

which was more than the original data set and other feature selection methods. The methods have been improved the other performance methods like precision, recall, F-Measure and true positive rate and reduce False Positive rate. The ensemble model experimental result shows that proposed ensemble models have achieved the highest accuracy CKD classification and prediction for CKD dataset. The accuracy of ensemble classification on reduced CKD dataset was 99%, which was the highest accuracy compared to the base classifiers.

REFERENCES

[1] K. Chandel, V. Kunwar, S. Sabitha, T. Choudhury, and S. Mukherjee, "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques," CSI Trans. ICT, vol. 4, no. 2–4, pp. 313–319, Dec. 2016.

[2] B. Boukenze, A. Haqiq, and H. Mousannif, "Predicting Chronic Kidney Failure Disease Using Data Mining Techniques," in Advances in Ubiquitous Networking 2, Springer, Singapore, 2017, pp. 701–712.

[3] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," Biomed. Signal Process. Control, Feb. 2017.

[4] R. Ani, G. Sasi, U.R. Sankar, & O.S. Deepa, (2016, September). "Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification – I EEE Conference Publication."[Online]. Available: http://ieeexplore.ieee.org/abstract/document/7732224/?reload=true. [Accessed: 15-Dec-2017].

[5] L.-C. Cheng, Y.-H. Hu, and S.-H. Chiou, "Applying the Temporal Abstraction Technique to the Prediction of Chronic Kidney Disease Progression," J. Med. Syst., vol. 41, no. 5, p. 85, May 2017.

[6] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods," J. Med. Syst., vol. 41, no. 4, p. 55, Apr. 2017.

[7] P. Pangong and N. Iam-On, "Predicting transitional interval of kidney disease stages 3 to 5 using data mining method," in 2016 Second Asian Conference on Defence Technology (ACDT), 2016, pp. 145–150.

[8] 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.

[9] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," Procedia Comput. Sci., vol. 82, no. Supplement C, pp. 115–121, Jan. 2016.

[10] U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naive Bayes classifier," in 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, pp. 1–5.

[11] N. Borisagar, D. Barad, and P. Raval, "Chronic Kidney Disease Prediction Using Back Propagation Neural Network Algorithm," in Proceedings of International Conference on Communication and Networks, Springer, Singapore, 2017, pp. 295–303.

[12] A. I. Pritom, M. A. R. Munshi, S. A. Sabab, and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," in 2016 19th International Conference on Computer and Information Technology (ICCIT), 2016, pp. 310–314.

[13] S. Mishra, P. Chaudhury, B. K. Mishra, and H. K. Tripathy, "An Implementation of Feature Ranking Using Machine Learning Techniques for Diabetes Disease Prediction," in Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, New York, NY, USA, 2016, p. 42:1–42:3.

[14] D. Zufferey, T. Hofer, J. Hennebert, M. Schumacher, R. Ingold, and S. Bromuri, "Performance comparison of multi-label learning algorithms on clinical data for chronic diseases," Comput. Biol. Med., vol. 65, no. Supplement C, pp. 34–43, Oct. 2015.

[15] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, "MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble," Arab. J. Sci. Eng., vol. 39, no. 11, pp. 7771–7783, Nov. 2014.

[16] T.R. Baitharu, & S.K. Pani, (2016). "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset- ScienceDirect." [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050916306263. [Accessed: 15-Dec-2017].

[17] Z. Sedighi, H. Ebrahimpour-Komleh, and S. J. Mousavirad, "Feature selection effects on kidney disease analysis," in 2015 International Congress on Technology, Communication and Knowledge (ICTCK), 2015, pp. 455–459.

[18] D.M. Filimon, & A. Albu, (2014, May "Skin diseases diagnosis using artificial neural networks - IEEE Conference Publication." [Online]. Available: http://ieeexplore.ieee.org/abstract/document/6840059/. [Accessed: 15-Dec-2017].

[19] F. Ahmad, N. A. M. Isa, Z. Hussain, M. K. Osman, and S. N. Sulaiman, "A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer," Pattern Anal. Appl., vol. 18, no. 4, pp. 861–870, Nov. 2015.

[20] [20] H. Asri, H. Mousannif, H. Al Moatassime, & T. Noel, (2016). "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis ScienceDirect." [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050916302575. [Accessed: 15-Dec-2017].

[21] B.L Deekshatulu, & P. Chandra, (2013 "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm ScienceDirect." [Online].Available: http://www.sciencedirect.com/science/article/pii/S2212017313004945. [Accessed: 15-Dec-2017].

[22] S. Ramya, and N. Radha. "Diagnosis of chronic kidney disease using machine learning algorithms." International Journal of Innovative Research in Computer and Communication Engineering vol. 4, no. 1 pp. 812-820. 2016.

[23] R. Dhruvi, R. Yavnika, & R. Nutan, " Prediction of Probability of Chronic Diseases and Providing Relative Real-Time Statistical Report using data mining and machine learning techniques". International Journal of Science, Engineering, and Technology Research (IJSETR) vol. 5, no. 4. 2016.

[24] S. Vijayarani, S. Dhayanand, and M. Phil. "Kidney disease prediction using SVM and ANN algorithms." International Journal of Computing and Business Research (IJCBR) vol. 6, no. 2, 2015.

[25] N. Chetty, S. V. Kunwar, and S. D. Sudarsan. "Role of attributes selection in the classification of Chronic Kidney Disease patients." In Computing, Communication and Security (ICCCS), 2015 International Conference on, pp. 1-6. IEEE, 2015.

[26] http://www.datasciencecentral.com/profiles/blogs/python-resources-for-top-data-mining-algorithms

[27] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[28] J. Radhakrishnan, and M. Sumit, "KI Reports and World Kidney Day." Kidney international reports vol.2, no. 2, pp. 125-126, Mar. 2017.

[29] https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease#_ENREF_1

[30] P. Ahmad, Q. Saqib, and S. Q. A. Rizvi. "Techniques of data mining in healthcare: a review." International Journal of Computer Applications Vol. 120, no. 15 Jan. 2015.

[31] B. R. Sharma, K. Daljeet, and A. Manju. "Review on Data Mining: Its Challenges, Issues and Applications." International Journal of Current Engineering and Technology vol. 3, no. 2 jun. 2013.

[32] Bashir, S., Qamar, U., Khan, F. H., & Naseem, L. (2016). HMV: a medical decision support framework using multi-layer classifiers for disease prediction. Journal of Computational Science, 13, 10-25.