International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Two Stage Hybrid Ensemble Classifier Based Diagnostic Tool for Chronic Kidney Disease Diagnosis Using Optimally Selected Reduced Feature Set

Sahil Sharma*[1], Vinod Sharma[1], Atul Sharma[2]

Accepted : 03/04/2018 Published: 29/06/2018

Abstract: This paper presents an idea of applying a two stage hybrid ensemble classifier for improving the prediction accuracy of Machine Learning based automated diagnosis of chronic kidney disease on the basis of values of an optimally selected subset of clinical and physiological parameters fed to it. Chronic kidney disease is a generalized term for various heterogeneous disorders affecting the structure and function of the kidney. It is a disease with high mortality rate. In this paper the authors have proposed a two stage hybrid ensemble technique with very high efficiency. In two stage hybrid ensemble classifier the potential of individual classification algorithms are combined together. In addition to this the authors optimally selected 8 parameters of prime importance from the set of 24 parameters of the dataset used for the study .The parameters (features) selected represent the intersection of the two sets; one containing medically essential parameters arranged in decreasing contribution to the diagnosis and other set containing parameters ranked in decreasing order of their contribution in the Machine Learning classification process. The results depict that the two stage hybrid ensemble is a very efficient method for classification of chronic kidney disease. The results of this ensemble classifier on the optimally selected reduced feature set (with 8 parameters) as well as the complete feature set (with 24 parameters) in terms of various performance metrics are predictive accuracy of (2-class) 100%, sensitivity of 1, precision of 1, specificity of 1 and F-value of 1. The GUI based diagnostic tool developed on the basis of the proposed ensemble can act as a tool for assisting doctors for cross-validating their findings of initial screening of chronic kidney disease using fewer clinical parameters thus helping them to attend to the needs of more patients in less time.

Keywords: *Artificial intelligence; Artificial neural networks; Decision-Tree; Ensemble; K-nearest neighbor; Machine Learning; Medical diagnosis; Support vector machine*

1. Introduction

John McCarthy coined the term ‘‘Artificial Intelligence’’ and defined it as the ‘‘science and engineering of making intelligent machines’’. Artificial Intelligence commonly known as AI is sometimes also referred as “Synthetic Intelligence” [1]. AI is the branch of science and engineering i.e. concerned with the computational understanding of intelligent behaviour and with the creation of artefacts that exhibit such behaviour. Programs which enable computers to function in the ways that make people seem intelligent are called artificial intelligent systems [2].The field of Artificial Intelligence was founded with an intention of imparting a central property of humans i.e. intelligence to machines. Machine Learning is a branch of artificial intelligence which aims at providing computational methods for accumulating, changing and updating knowledge in the intelligent systems. In words of Ethem Alpaydın machine learning can be defined as “Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future or descriptive to gain knowledge from data, or both” [3] .Machine learning can be used effectively for diagnosis, prognosis, and prediction of recurrence of diseases or medical disorders. Life threatening disease like diabetes [4], lung diseases [5], Heart diseases [6] cancer [7] etc can be diagnosed with great accuracy by feeding medical datasets obtained from various sources pertaining to these diseases to Machine Learning based systems that learn from these data sets and predict the future outcome with notable accuracy. The primary motive of machine learning in medicine is developing artificial intelligent systems that can assist a medical doctor in performing expert diagnosis.

Chronic kidney disease is a general term for heterogeneous disorders affecting the structure and function of the kidney [8]. It is a heterogeneous condition, whose clinical manifestations and course depends on the cause and type (pathology), severity, rate of progression, and comorbid conditions [9]. The definition of chronic kidney disease is based on the presence of kidney damage (i.e. albuminuria) or decreased kidney function for 3 months or more [10]. Kidney failure is one of the most serious outcomes of chronic kidney disease, the main reasons being the complications of reduced kidney function. Dialysis and transplantation are the only viable treatment options when the symptoms of kidney failure are severe i.e. end stage renal disease. Complications can occur at any stage, which often lead to death with no progression to kidney failure, and can arise from adverse effects of interventions to prevent or treat the disease [8]. CKD is an internationally recognized public health problem affecting nearly 10% of the world population [11]. The current desired contribution of AI in the medical sciences is the programs that can assist a medical

[1] *Department of Computer Science & IT, University of Jammu, J&K, India*
[2] *Government Medical College, Jammu, J&K, India*
* *Corresponding Author: Email: sahiil91@live.com*

expert in performing expert and more accurate and quick diagnosis. These programs by making use of combination of sciences like statistics and probability try to find out the patterns from the data (i.e. Pattern Recognition) used for training and then make use of these patterns in order to classify the test data into one the possible categories (outcome).

2. Literature Review

The use of machine learning algorithms is day by day increasing in medical domain for solving problems by analyzing and interpreting large volumes of data [4]. A number of researchers in this field have used Machine Learning algorithms in order to solve problems in the field of medicine.

• Igor Kononenko [12] presented a view on the use of Machine learning techniques 1) in the past for the interpretation of medical data 2) For intelligent analysis of medical data in the current scenario and 3) for assistance of physicians in diagnosis of medical disorders, in the future. Integration of machine learning techniques with the existing instrumentations for the acceptance of machine learning in medicine is suggested by the authors.

• Hardik Maniya et al. [13] compared Naïve Bayes classifier and KNN for diagnosis of Tuberculosis, implementation has been done using C language and Weka tool. Medical Dataset used had 19 attributes and 154 instances. The authors classified the patients affected by tuberculosis into two categories (least probable and most probable).The authors achieved nearly 78% accuracy with low false negative.

• R Bharat Rao et al. [14] developed a computer aided diagnosis (CAD) system named lungCAD. The system used classification algorithm for the detection of pulmonary nodules from the CT thorax studies. The clinical approach for the diagnosis of coin lesion used chest x-ray and CT scan. The lungCAD greatly assisted the clinician in order to improve their diagnosis accuracy .LungCAD was also approved by FDA in 2006.

• Abid Sarwar et al. [4] performed a comparative analysis of Artificial neural network, Naïve Bayes and KNN algorithm for the type II diabetes in terms of detection accuracy. The results showed that Artificial neural network with 96% prediction accuracy performs better than Naïve Bayes with 95% and KNN 91%.

• Yasodha et al. [15] did analysis of a database of diabetic patients using weka tool. The authors considered different algorithms such as REP Tree, Bayes Network, J48 and Random Tree classifiers for the study and compared the outputs. The main objective of the study was to develop a Diabetic expert system; inputs being patient's daily glucose rate and insulin dosages the system would predict the patient's insulin dosage for the next day.

• Bekir Karlik [16] did comparison between Backpropagation and Naive Bayes Classifiers to diagnose hepatitis disease. Hepatitis is the general term for inflammation of the liver. The most common causes of hepatitis are the hepatotropic viruses (such as hepatitis A, B, and C) and alcohol abuse. In practice, both of these methods often compete well with more sophisticated classifiers. The performances of proposed methods are selected for each of classification tasks of hepatitis diseases. The over¬all accuracy of diagnosis systems were 98% and 97% respectively.

• Huda Yasin[17] proposed a method for investigating factors which are more pervasive for the risk of hepatitis C virus. The dataset has been obtained from the machine learning warehouse of University of California. The authors compared the proposed method with nearly 20 classification techniques which includes Naïve Bayes, GRNN, and CART etc. and proved the proposed method is having the highest accuracy rate of 89.6%.The proposed method worked by using only 37% of the total fields depicting low feature complexity.

• L.C. van der Gaag et al. [18] developed a decision-support system for patient-specific therapy selection for esophageal cancer. The system predicts the correct stage of cancer, which helps the oncologist to start with the correct treatment plan for the patient. The kernel of the system is a probabilistic network that describes the presentation characteristics of cancer of the esophagus and the path physiological processes of invasion and metastasis. Results showed that for 85% of the patients, the network predicted the correct cancer stage.

• Babak Sokouti et.al. [19] proposed Levenberg–Marquardt feedforward MLP neural network (LMFFNN) in order to classify cervical cell images obtained from 100 patients including healthy, low-grade intraepithelial squamous lesion and high-grade intraepithelial squamous lesion cases. This neural network along with extracted cell image features is a new model for cervical cell image classification. Based on the results, cervical cell images were classified successfully with 100% correct classification rate using the proposed method. Moreover, the rates of sensitivity and specificity were calculated as 100 % using LMFFNN method. It was shown there was a good agreement between the expert decision and values gained from the ANN model.

• Shivajirao M.Jadhav et al. [20] proposed a system that used the ECG recordings for the detection of Athymias in the human heart, by training a multilayer perceptron (MLP) artificial neural network(ANN) on an ECG data set. The ECG dataset has been extracted from University of California at Irvine (UCI) data repository and it contains 452 instances with 279 attributes. The proposed system classifies the patterns into two classes 1) normal and 2) abnormal classes. The data set has also been used to train a modular artificial neural network (ANN). The system with MLP model showed 86.67% classification accuracy and 93.75% sensitivity and Modular ANN showed a classification accuracy of 93.1%.

• T.Manju et al. [21] proposed a hybrid system based on multi-layer feed forward neural network (MLFFN) and genetic algorithm (GA) for assisting medical doctors in predicting the heart disease. The cardiac arrest (heart attack) is a major cause of death in the world, its major causes are smoking, high blood pressure, unhealthy diet, and obesity and diabetes .The data set used in the study was collected from university of California at Irvine (UCI) repository and consists of data of 270 patients. The ANN is trained using back propagation and feed forward neural network. Weight optimization is done using genetic algorithm. The weights are associated with each connection in the neural network nodes. The accuracy of the system on training dataset came out to be 79.7% and on testing accuracy 89.67%.

• S.Vijayarani et al. [22] compared Support Vector Machine (SVM) and Artificial Neural Network (ANN) in order to predict kidney disease. The performance metrics used were accuracy and execution time. Analysis of results showed that ANN performed better in terms of classification accuracy whereas SVM required less execution time as compared to ANN. The authors considered ANN to be better than SVM considering both the performance metrics together.

• Andrew Kusiak et al. [23] elicited knowledge about the interaction between many of measured parameters and kidney dialysis patient survival using data pre-processing, data transformations and data mining approach. Two different data mining algorithms were employed for extracting knowledge in the form of decision rules which in turn were used by a decision making algorithm, which predicts survival of new unseen patients.

Important parameters identified by data mining were interpreted for their medical significance. The approach presented in this manuscript reduced the cost and effort for selecting patients for clinical studies. Patients can be selected based on the prediction results and the most significant parameters discovered.

Various other research works and studies have also been carried out for performing machine learning based diagnosis of diseases like cancer [24, 25], diabetes [26, 27], heart diseases [28, 29], kidney disease [29-31].

3. Methods and Materials

3.1. Chronic Kidney Disease data

The medical dataset used to carry out this research has been obtained from the UCI data repository [32].The dataset used was created from data obtained from Dr. P. Soundarapandian of Apollo Hospitals, Tamil Nadu and contains data of 400 people from the southern part of India with their ages ranging between 2-90 years. There are in total twenty four parameters, most of which are clinical in nature and a few are physiological ones. On the basis of these 24 attributes each instance is assigned one of the two class-labels i.e. suffering or not suffering with chronic kidney disease. Table 1 summarizes the various parameters chosen and their allowed values.

Among the twenty four parameters a few are of prime importance for diagnosis of CKD. In a patient suffering from CKD the values of serum creatinine and blood urea are usually elevated. Specific gravity of urine remains fixed due the inability of kidneys to dilute or concentrate urine. Blood Pressure can be the result of CKD or in some cases high blood pressure may itself lead to CKD over time. Long standing diabetes could also be a reason of CKD. The Hemoglobin level falls and the patient is most of the times anaemic. Table 2 enlists all the twenty four parameters in descending order of their medical relevance as well as in terms of machine learning classification process for the diagnosis of CKD.

3.2. Methodology

The work carried out in this paper is an extension to the preliminary comparative analysis that the authors carried out earlier [33]. There are in total twenty five different classifiers that the authors considered for the development of the two stage hybrid ensemble classifier. Most of the candidate classifiers are majorly based on the following techniques: Decision-Tree, support vector machine (SVM), K-nearest neighbor, Artificial Neural Networks and Ensemble methods. These algorithms were selected for the analysis & study because of their popularity in the recent relevant literature & good performance even in fewer amounts of training data. A short description about the selected algorithms for study is given below.

3.2.1. Decision Tree Classifiers

DT classifiers classify data by making use of tree structure algorithms [34]. The underlying algorithm begins with the training samples and corresponding class labels. The training set is partitioned recursively based on a feature value into subsets. Each internal node represents a test on attribute; each edge (branch) represents an outcome of the test. A decision tree classifier identifies the class label of an unknown sample by following path root to the leaves, which represent the class label for that sample. The feature (attribute) i.e. selected as the root node is the one that best divides the training data. Fig.1 illustrates a decision tree in which nodes specify conditional attributes (features) F={F1,F2,.....Fi}, branches which show the values of features vi,h i.e. the hth range for ith feature and leave nodes which represent decisions C={C1,C2,.....Ci} having binary values Vdi={0,1}.

Table 1. Various parameters & their allowed values

Parameter	**Allowed Values**
Age	Discrete Integer Values
Blood pressure	Discrete Integer Values
Specific gravity	Nominal Values(1.005,1.010,1.015,1.020,1.025)
Albumin	Nominal Values(0,1,2,3,4,5)
Sugar	Nominal Values(0,1,2,3,4,5)
Red blood cells	Nominal Values(Normal, Abnormal)
Pus cell	Nominal Values(Normal, Abnormal)
Pus cell clumps	Nominal Values(Present, Not-Present)
Bacteria	Nominal Values(Present, Not-Present)
Blood glucose random	Discrete Integer Values
Blood urea	Discrete Integer Values
Serum creatinine	Numeric Values
Sodium	Discrete Integer Values
Potassium	Numeric Values
Hemoglobin	Numeric Values
Packed cell volume	Discrete Integer Values
WBC count	Discrete Integer Values
RBC count	Numeric Values
Hypertension	Nominal Values(Yes, No)
Diabetes mellitus	Nominal Values(Yes, No)
Coronary artery disease	Nominal Values(Yes, No)
Appetite	Nominal Values(Good, Poor)
Pedal edema	Nominal Values(Yes, No)
Anemia	Nominal Values(Yes, No)
Parameter	Allowed Values
Age	Discrete Integer Values
Blood pressure	Discrete Integer Values
Specific gravity	Nominal Values(1.005,1.010,1.015,1.020,1.025)

Decision Tree Variants used:

3.2.1.1. Simple Tree

In this variant of decision tree, maximum number of splits is taken to be four and gini‘s diversity index is used as split criterion.

3.2.1.2. Medium Tree

In this variant of decision tree, maximum number of splits is taken to be twenty and gini‘s diversity index is used as split criterion.

3.2.1.3. Complex Tree

In this variant of decision tree, maximum number of splits is taken to be hundred and gini‘s diversity index is used as split criterion.

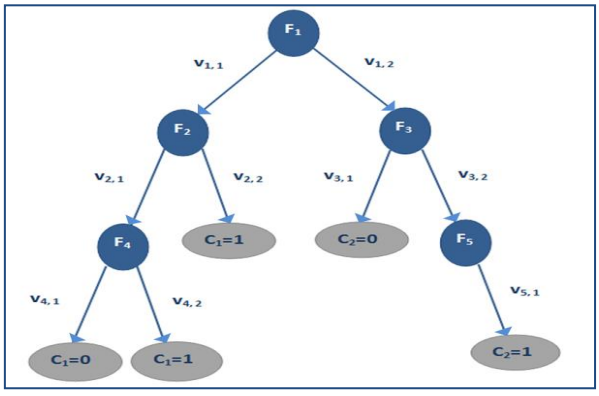

Fig. 1. Sample Decision Tree

3.2.2. Support Vector Classification (SVC)

SVC revolves around the perception of a “margin”-either side of a hyperplane that divides two data classes. Maximizing the margin creates the largest possible distance among the hyperplane and the instances on either side of the hyperplane reduce an upper bound on the anticipated generalization error. It works on two types of data i.e. linearly separable data and linearly Non-separable data. In case of linearly separable data only one hyperplane is needed for separating the data but in the case of latter more than one

hyperplanes are needed.

SVM variants used:

3.2.2.1. Linear SVM

In this variant of SVM, Linear Kernel Function has been used with kernel scale set to 1.

3.2.2.2. Quadratic SVM

In this variant of SVM, Quadratic Kernel Function has been used with kernel scale set to 2.

3.2.2.3. Cubic SVM

In this variant of SVM, Cubic Kernel Function has been used with kernel scale set to 2.5.

3.2.2.4. Fine Gaussian SVM

In this variant of SVM Gaussian Kernel Function has been used with kernel scale set to 1.2.

3.2.2.5. Medium Gaussian SVM

In this variant of SVM Gaussian Kernel Function has been used with kernel scale set to 4.9.

3.2.2.6. Coarse Gaussian SVM

In this variant of SVM Gaussian Kernel Function has been used with kernel scale set to 20.

3.2.3. Discriminant analysis (DA)

DA classifiers work under the assumption that different classes generate data based on different Gaussian distributions. In the training phase the Gaussian distribution parameters for each class are estimated by the fitting function and in order to predict the classes (class-labels) of new data, the trained classifier finds the class with the smallest misclassification cost. There are mainly two types of discriminant analysis classifiers namely – Linear Discriminant Analysis Classifier (LDA) and Quadratic Discriminant Analysis Classifier. The Quadratic Discriminant Analysis Classifier can be considered as the generalization of LDA. In this study Linear and Quadratic Discriminant Classifiers have been used and in both the models diagonal covariance is used for regularization.

3.2.4. K-nearest neighbour (KNN)

KNN is a classification technique which classifies the test objects on the basis of number of closest training examples. It is also termed as a lazy-learning algorithm. KNN is a non-parametric algorithm which means that it does not assume anything on the underlying data distribution. In this, the Euclidean distance is calculated between the test data and every sample in the training data followed by classifying the test data into a class in which most of k-closest neighbours of training data belong to. K is usually a very small positive integer. As the Value of K increases it becomes increasingly difficult to distinguish between the various classes. Cross-validation and other heuristic techniques are used to choose an optimal value of K.

KNN variants used:

3.2.4.1. Fine KNN

In this variant of KNN, number of neighbours has been taken as one, Euclidean distance has been used as distance metric and equal distance weight has been used.

3.2.4.2. Medium KNN

In this variant of KNN, number of neighbours has been taken as ten, Euclidean distance has been used as distance metric and equal distance weight has been used.

3.2.4.3. Coarse KNN

In this variant of KNN, number of neighbours has been taken as hundred, Euclidean distance has been used as distance metric and equal distance weight has been used.

3.2.4.4. Cosine KNN

In this variant of KNN, number of neighbours has been taken as ten, Cosine distance has been used as distance metric and equal distance weight has been used.

3.2.4.5. Cubic KNN

In this variant of KNN, number of neighbours has been taken as ten, Minkowski distance has been used as distance metric and equal distance weight has been used.

3.2.4.6. Weighted KNN

In this variant of KNN, number of neighbours has been taken as ten and Euclidean distance has been used as distance metric and squared inverse distance weight has been used.

3.2.5. Artificial neural network (ANN)

ANN is a methodology inspired by the biological network of neurons. It is a powerful data-modelling tool capable of capturing, representing and simulating complex relationships between inputs and outputs by performing multiple parallel computations. These are analytical tools which try to emulate ‘‘learning’’ process of the cognitive system and the neurobiological functions of the human brain. In ANN, the neurons are grouped into different layers, an input layer, one or more hidden layers, and an output layer. Fig.2 shows a neural network with two hidden layers. Learning is achieved by repeatedly adjusting the numerical weights associated with the interconnecting edges between different artificial neurons. In addition to this an activation function is used that converts a neuron’s weighted input to its output activation. In this study two versions of Feed Forward Back-Propagation Neural Network (FFBPNN) have been used. One of them uses Levenberg-Marquardt (LM) back propagation training function along with gradient descent weight and bias learning function and other uses gradient descent training function along with gradient descent weight and bias learning function.

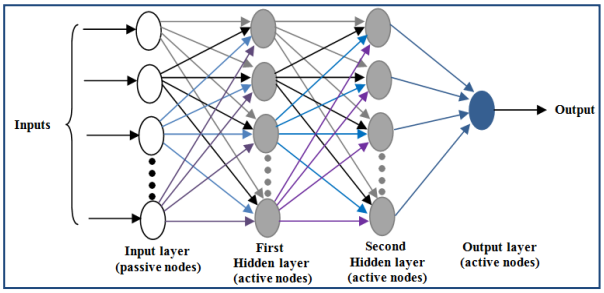

Fig. 2. A Multilayer Neural Network

3.2.6. Ensemble method (EM)

In this method potentials of various individual classifiers are fused together. Using Ensemble method increases the performance by combining the classifying ability of individual classifiers and the chances of misclassifying a particular instance are reduced significantly, this provides a greater accuracy to the overall classification process.The different learners can be combined in a number of ways. They can work in parallel on all of the inputs, and their outputs can be combined in some way. If an instance gets wrongly classified by an individual classifier, the error is corrected by the right classification done by other individual classifiers. Alternatively, a multistage combination will train the base learners on different subsets of the input data. For example, the AdaBoost algorithm first trains an initial learner, and then trains subsequent learners on data that the first learner misclassifies. This way, the weaknesses of each base-learner are made up for by the next learner [35]. Fig.3 illustrates the general working of an ensemble method in which all individual classifiers work in parallel.

Ensemble variants used:

3.2.6.1. Boosted Trees

In this variant AdaBoost method is used, decision tree is the learner

type with maximum number of splits being twenty and number of learners used is thirty.

3.2.6.2. Bagged Trees

In this variant bagging method is used, decision tree is the learner type and number of learners used is thirty.

3.2.6.3. Subspace Discriminant

In this variant Subspace method is used, discriminant analysis is the learner type, number of learners used is thirty and subspace dimension is twelve.

3.2.6.4. Subspace KNN

In this variant Subspace method is used, nearest neighbours is the learner type with number of learners being thirty and subspace dimension is twelve.

3.2.6.5. RUSBoosted Trees

In this variant RUSBoost method is used, decision tree is the learner type, number of learners is thirty, maximum number of splits is twenty and learning rate is 0.1.

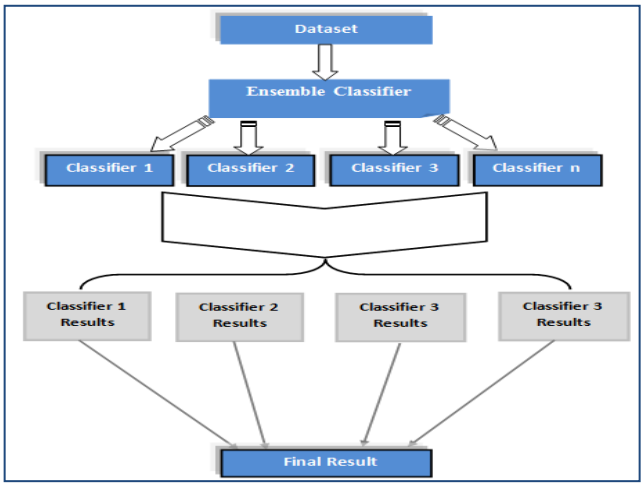

Fig. 3. General methodology of an ensemble method

3.3. Two stage Hybrid Ensemble Classifier

The ensemble classifier proposed in this study is based on using classification potential of different individual classifiers collectively. If some of these individual classifiers in turn are themselves ensemble methods, the new method becomes a two stage Hybrid Ensemble classifier. Fig. 4 illustrates the proposed method graphically. The values of all the performance metrics of twenty five different candidate classifiers were initially evaluated and analyzed after applying them to optimally selected reduced feature set. The values of all the performance metrics of twenty five different candidate classifiers were initially evaluated and analyzed after applying them to optimally selected reduced feature set. In order to select individual classifiers to be used in the ensemble method, the sensitivity and specificity i.e. the true positive rate and true negative rate respectively of each classifier were analyzed. The motive behind was to select two set of classifiers; one that would ensure high true positive rate (sensitivity) and other with high true negative rate (specificity). Based on the results of values of these two performance metrics of all the candidate classifiers, the authors selected: *Ensemble using bagged trees, Linear SVM and Ensemble using boosted trees Classifier*. The reason for their selection being that Ensemble using bagged trees has true negative rate (specificity) of 100% (i.e. 1), Linear SVM has true positive rate (sensitivity) of 99.2% (i.e. 0.992) and Ensemble using boosted trees to make up for the comparatively low true negative rate (specificity) of the Linear SVM, as it has reasonably both high true positive rate of 98%(i.e. 0.98) & true negative rate of 100%(i.e. 1). Inclusion of the Ensemble using boosted trees ensures that the cases in which conflict may arise between class labels assigned by the other two classifiers due to comparatively low true negative rate of Linear SVM of 85.33 % (i.e. 0.8533) will be dealt with and the chances of wrong classification will be minimized. In general, a Two Stage Ensemble Classifier can be represented as:

$$C_{TEC}=\text{Mode}\,(C_1\ \ C_2\ \ C_3\ \ .....\ \ C_n\ )$$

Where,

C_{TEC} = Class-label assigned by the Two stage hybrid ensemble classifier.

C_i = Class-label assigned by every i^{th} individual classifier.

In particular, the two stage ensemble classifier used in context of this study can be represented as :

$$C_{TEC}=\text{Mode}\,(\,C_{E.Bagged}\ \ C_{LSVM}\ \ C_{E.Boosted})$$

Where,

C_{TEC} = Class-label assigned by the Two stage hybrid ensemble classifier.

$C_{E.Bagged}$ = Class-label assigned by Ensemble classifier using bagged trees classifier.

C_{LSVM} = Class-label assigned by Linear SVM classifier.

$C_{E.Boosted}$ = Class-label assigned by Ensemble using boosted trees classifier

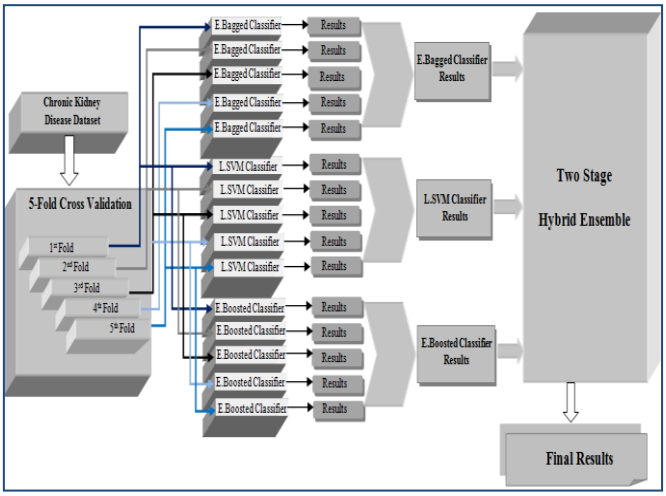

Fig. 4. Graphical representation of Classification using Two Stage Hybrid Ensemble Classifier

3.4. Optimally Selected Reduced Feature Set

The medical dataset considered for this study consists of twenty-four parameters (features). In the course of this study each individual parameter was considered and evaluated in terms of its contribution towards the results, i.e. in diagnosing whether a patient suffers from the chronic kidney disease or not , both in terms of their medical relevance and how much their contribution is in the Machine Learning classification process. For ranking various parameters in accordance with their importance in terms of medical relevance, help from the medical experts from the concerned field was sought. On the other hand in order to rank all the parameters in terms of their contribution in the Machine Learning classification process ranker method of attribute evaluation using 5-fold cross validation in WEKA benchmark (version 3.6.13) was used. It is a filter approach and it ranks the attributes with respect to their information gain. This ranking is independent of a specific learning algorithm. The list of all twenty four parameters arranged in descending order in terms of their contribution towards the diagnosis of chronic kidney disease in terms of Medical relevance and Machine learning classification process as well are shown in Table 2.

Table 2. Parameters arranged in descending order of their relevance

In Terms of Medical Relevance	In Terms of Machine Learning Relevance
Serum Creatinine Most Relevant	Hemoglobin
Blood Urea	Serum Creatinine
Blood Pressure	Packed Cell Volume
Hemoglobin	Specific Gravity
Specific Gravity	Albumin
Hypertension	Hypertension
Diabetes Mellitus	Diabetes Mellitus
Pedal Edema	RBC Count
Albumin	Blood Urea
Age	Blood Glucose Random
Sodium	Sodium
Potassium	Blood Pressure
Packed Cell Volume	Appetite
RBC Count	Pus Cell
Coronary Artery Disease	Potassium
Appetite	Pedal Edema
WBC Count	Red Blood Cells (in urine)
Sugar	Sugar
Pus Cell	Anemia
Pus Cell Clumps	Age
Red Blood Cells (in Urine)	WBC Count
Bacteria	Pus Cell Clumps
Anemia	Coronary Artery Disease
Blood Glucose Random Least Relevant	Bacteria

Table 3. Optimally Selected Reduced Feature-Set

Parameters	Allowed Values
Albumin	Nominal Values
Diabetes Mellitus	Nominal Values
Hypertension	Nominal Values
Specific Gravity	Nominal Values
Blood Pressure	Discrete Integer values
Blood Urea	Discrete Integer values
Haemoglobin	Numeric Values
Serum Creatinine	Numeric Values

After this an optimal subset of 8 parameters was extracted from the complete set of 24 parameters; these are such parameters which meet the requirements both in the terms of medical relevance and machine learning classification process. Every parameter (feature) that is part of optimally selected reduced feature set was evaluated on two fronts; one being their importance in medical relevance along with the cost incurred in clinical test to obtain their value and the other front being the contribution of that parameter in the machine learning classification process. All the Eight parameters included in the reduced feature set stand high both in terms of medical relevance and classification process. All the parameters used in the reduced feature set are given in table 3. Two (*blood pressure, hypertension)* out of the eight parameters *does not require any clinical tests* .The reduced number of parameters requiring clinical tests means considerable reduction in the cost incurred to the patient.

4. Implementation and Results

All the twenty five candidate classifiers and the two stage hybrid ensemble were applied to the complete as well as the reduced feature set using 5–fold cross validation. Cross-Validation is used to give a good estimate of the predictive accuracy of the final classifier trained with all the data. The procedure includes selecting a number of folds to partition the data set followed by the following steps:

1. Partition the data into k disjoint sets or folds
2. For each fold:
   a. Classifier is trained using the out-of-fold observations
   b. Model performance is assessed using the in-fold data.
3. The average test error over all folds is calculated

Afterward this methodology of each classifier differs but the last step i.e. common to all the classifiers is assigning a class-label to every single instance of the dataset. For the implementation of individual classifiers as well as the two stage hybrid ensemble MATLAB 2016a was used. The feature set was imported to the MATLAB environment from a Microsoft excel-sheet. The performance metrics used for the evaluation of results are predictive accuracy, sensitivity, precision, specificity and F-score These performance metrics are explained below:

1. *Predictive accuracy* of Z% shows that the classifier is able to classify Z% of instances correctly.
2. *Sensitivity* can be defined as the ability of a test to correctly detect cases which do have the condition. Mathematically, it can be defined as the number of true positives divided by the number of true positives and false negatives.
3. *Precision* can be defined as the number of correctly classified positive examples divided by the total number of examples labelled by the system classified as positive [36] i.e. total number of true positives divided by the total number of true positives and false positives.
4. *Specificity* can be defined as the ability of a test to correctly detect cases which do not have the condition. Mathematically, it can be expressed as the number of true negatives divided by the number of true negatives and false positives.
5. F-value is a measure of a test's accuracy. It considers both the precision and the recall (sensitivity) of the test to compute the value. The F- value can be interpreted as a weighted average of the precision and recall. Mathematically it can be defined as the twice the product of precision and recall divided by the sum of precision and recall.

True positives are the correctly recognized class examples, true negatives are correctly recognized examples that do not belong to the class, whereas false positives are examples that were incorrectly assigned to the class and false negatives are those examples that were not recognized as class examples [36]. Fig. 5 shows the predictive accuracy of all the individual classifiers and as well as that of the two stage hybrid ensemble classifier and table 4 lists them all along with their values of sensitivity, precision and specificity in terms of complete as well as reduced feature set whereas Fig.6 and Fig.7 illustrate them graphically. As it can be seen from the results, the two-stage hybrid ensemble classifier outperformed all the individual classifiers on both the complete as well as the reduced feature set.

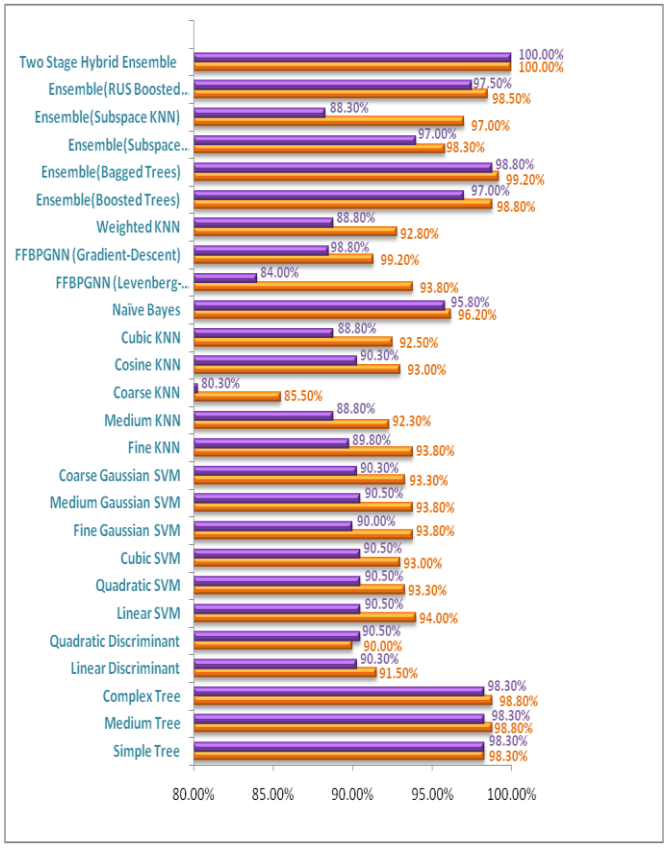

Fig. 5. Predictive accuracy of all candidate classifiers

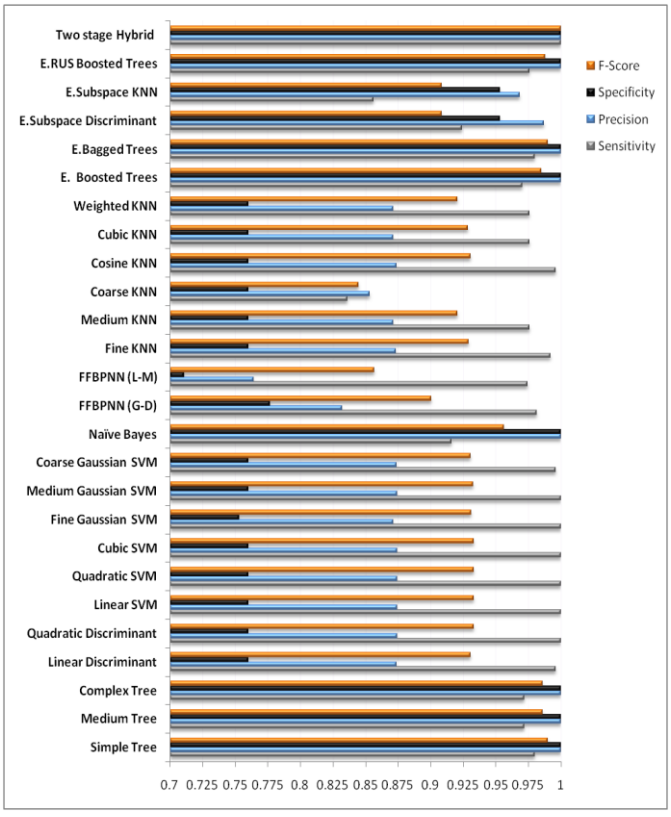

Fig. 6. Graphical representation of Specificity, Precision, Specificity, F-Value values classifiers on complete feature-set.

On both the feature sets the two stage hybrid ensemble achieved a predictive accuracy of 100%, sensitivity of 1, precision of 1 and specificity of 1. Among the individual classifiers ensemble method using bagged trees performed best with predictive accuracy of 99.2% along with sensitivity of 0.992 , precision of 1 and specificity of 1 on complete feature set and predictive accuracy of 99.2 % along with sensitivity of 0.984 , precision of 1 and specificity of 1 on reduced feature set. The efficiency of the optimally selected reduced feature set can also be seen from the fact that the performance of most of the individual classifiers improved when they were trained using reduced feature set instead of complete feature set. As the dimensionality of data increases, classification problems become significantly harder i.e. a high number of features can lead to lower classification accuracy. The classification accuracy achieved with reduced feature sets is often significantly better than with the complete feature set [37]

A GUI based diagnostic tool based on the two stage hybrid ensemble classifier is developed that can be used to predict whether a patient is suffering from chronic kidney disease or not when it is fed with all the 8 attributes from user through a user friendly GUI (Graphical User Interface).The development of this diagnostic tool is done using MATLAB 2016a. Out of 8 parameters that the user needs to enter as input in GUI based diagnostic tool four are numeric and the rest are nominal values. The diagnostic tool in execution is shown in Fig.8.

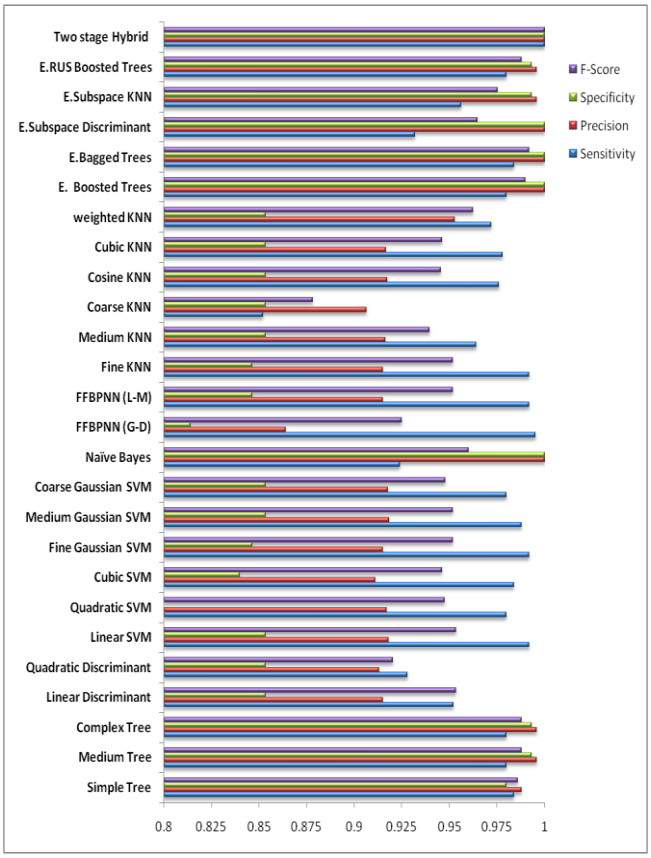

Fig. 7. Graphical representation of Specificity, Precision, Specificity, F-Value values classifiers on reduced feature-set.

Table 4. Sensitivity, Precision, Specifity and F- values classifiers

Method	Classifier Name	Sensitivity		Precision		Specificity		F-Score	
		Reduced F-set	Complete F-set	Reduced F-set	Complete F-set	Reduced F-set	Complete F-set	Reduced F-set	Complete F-set
Decision Trees	Simple Tree	0.9840	0.9800	0.9879	1	0.9800	1	0.9859	0.9898
	Medium Tree	0.9800	0.9720	0.9959	1	0.9933	1	0.9878	0.9858
	Complex Tree	0.9800	0.9720	0.9959	1	0.9933	1	0.9878	0.9858
Discriminant Analysis	Linear Discriminant	0.9520	0.9960	0.9150	0.8736	0.8533	0.7600	0.9468	0.9307
	Quadratic Discriminant	0.9280	1	0.9130	0.8741	0.8533	0.7600	0.9204	0.9328
Support Vector Machine	Linear SVM	0.9920	1	0.9180	0.8741	0.8533	0.7600	0.9535	0.9328
	Quadratic SVM	0.9800	1	0.9170	0.8741	0.8000	0.7600	0.9474	0.9328
	Cubic SVM	0.9840	1	0.9110	0.8741	0.8400	0.7600	0.9460	0.9328
	Fine Gaussian SVM	0.9920	1	0.9151	0.8710	0.8466	0.7530	0.9519	0.9310
	Medium Gaussian SVM	0.9880	1	0.9182	0.8740	0.8533	0.7600	0.9518	0.9327
	Coarse Gaussian SVM	0.9800	0.9960	0.9176	0.8736	0.8533	0.7600	0.9477	0.9307
ANN	FFBPNN (G-D)	0.9953	0.9811	0.8640	0.8320	0.8142	0.7765	0.9250	0.9004
	FFBPNN (L-M)	0.9920	0.9744	0.9151	0.7640	0.8466	0.7107	0.9519	0.8564
KNN	Fine KNN	0.9920	0.9920	0.9151	0.8732	0.8466	0.7600	0.9519	0.9288
	Medium KNN	0.9640	0.9760	0.9163	0.8712	0.8533	0.7600	0.9395	0.9206
	Coarse KNN	0.8520	0.8360	0.9063	0.8530	0.8533	0.7600	0.8783	0.8444
	Cosine KNN	0.9760	0.9960	0.9172	0.8736	0.8533	0.7600	0.9456	0.9307
	Cubic KNN	0.9780	0.9760	0.9166	0.8714	0.8533	0.7600	0.9463	0.9287
	Weighted KNN	0.9720	0.9760	0.9529	0.8714	0.8533	0.7600	0.9623	0.9207
Naive Bayes	Naive Bayes	0.9240	0.9160	1	1	1	1	0.9600	0.9560
Ensemble	Boosted Trees	0.9800	0.9700	1	1	1	1	0.9898	0.9847
	Bagged Trees	0.9840	0.9800	1	1	1	1	0.9919	0.9898
	Subspace Discriminant	0.9320	0.9240	1	0.9870	1	0.9800	0.9648	0.9544
	Subspace KNN	0.9560	0.8560	0.9958	0.9680	0.9933	0.9533	0.9754	0.9085
	RUS Boosted Trees	0.9800	0.9760	0.9959	1	0.9933	1	0.9878	0.9878
	Two Stage Hybrid	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

5. Conclusion

Chronic Kidney disease is a disease with high mortality rate. Five to ten percent of the population worldwide suffers from this disease. Chronic kidney disease is a worldwide health crisis. A majority of the cases are not timely diagnosed or remain undiagnosed in developing and underdeveloped nations majorly due to poor doctor-patient ratio and poverty; this is one of the prime reasons that higher percentage of these cases are from developing and underdeveloped nations in comparison to developed nations as majority of people in developed nations go through routine check-up and diagnosis. More than 80% of all patients who receive treatment for kidney failure are in affluent countries with universal access to health care and large elderly populations [38]. The cost incurred by the clinical tests a patient has to go through acts as a deterrent for them to visit a doctor in order to get timely and regular check-up in developing and underdeveloped nations. Reducing the number of parameters required for the diagnosis to be done by the classification system without hampering its performance, one major problem i.e. of the cost incurred in going through a number of clinical tests can be addressed to a great extent .The reduced number of parameters means fewer clinical tests a patient has to go through and fewer the clinical tests taken less will be the cost incurred. Keeping this in mind all the parameters added to the optimally selected reduced feature set were also evaluated in terms of the cost incurred to a patient by going through the clinical test that provides the value for that parameter, in addition to its medical relevance and considering its role in classification process done by classifier. It can also be seen from the results of individual classifiers that their classification performance improved in case of optimally reduced feature set as compared to the complete feature set. The GUI based diagnostic tool using two stage hybrid ensemble classifier developed by the authors can result in timely and accurate diagnosis of this disease by assisting doctors in cross checking their diagnosis findings in relatively short time with minimal number of clinical tests required and thus helping a doctor to attend and diagnose more number of patients as compared to the scenario where he has to go through the diagnosis process entirely manually. There were some missing values in the dataset that were dealt with by replacing numeric and discrete integer values by attribute mean of the all the instances with the same class-label as that of the instance under consideration and nominal values were replaced using attribute mode. In the future, this study could be extended to perform multi-stage diagnosis of chronic kidney disease by including the GFR (glomerular filtration rate) as a parameter.

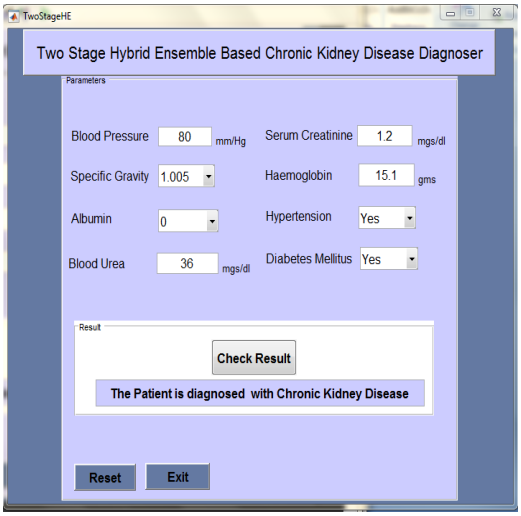

Fig. 8. CKD Diagnostic tool in execution.

References

[1] David Poole, Alan Mackworth, Randy Goebel, Computational Intelligence: A Logical Approach (New York: Oxford University Press, 1998) p.1.

[2] A.N Ramesh, C Kambhampati, J.R Monson, PJ Drew, Artificial intelligence in medicine. (Ann R Coll Surg Engl, 2004), pp. 334–338.

[3] Ethem Alpaydın, Introduction to Machine Learning (The MIT Press,

Cambridge, Massachusetts ,London, England, 2010) p. 3.

[4] Abid Sarwar, Vinod Sharma, Comparative analysis of machine learning techniques in prognosis of type II diabetes in AI & Society (Springer Verlag, 2013).

[5] R.Bharat Rao, Jinbo Bi, Nancy Obuchowski and David Naidich, LungCAD: A Clinically approved Machine Learning System for Lung Cancer Detection in International conference on knowledge discovery and data mining (San Jose, California, USA, 2007),ACM 978-1-59593-609-7/07/0008.

[6] T.Manju, K.Priya and R.chitra, Heart Disease Prediction System using Weight Optimized Neural Network, in International Journal of Engineering and Computer Science, Vol-2, No.3 pp.781-788, 2013.

[7] B. Sokouti, S. Haghipour, and A. D. Tabrizi, A framework for diagnosing cervical cancer disease based on feedforward MLP neural network and ThinPrep histopathological cell image featuresin Neural Comput. Appl., Vol. 24, No. 1 (2014) , pp. 221–232,.

[8] Andrew S Levey, Josef Coresh, Chronic kidney disease in Lancet (2012), pp. 165–80.

[9] A S Levey, R Atkins, J Coresh, E. P Cohen, A.J Collins, K.U Eckardt, M. E Nahas, B. L Jaber, M Jadoul, A Levin, N. R Powe, J. Rossert, D. C Wheeler, N. Lameire, and G. Eknoyan, Chronic kidney disease as a global public health problem: approaches and initiatives - a position statement from Kidney Disease Improving Global Outcomes in Kidney Int., Vol. 72 (2007) pp. 247–259.

[10] L.A Stevens, A.S Levey, Current status and future perspectives for CKD testing in American Journal of Kidney Diseases(2009), pp. 17-26.

[11] WorldKidneyDay:ChronicKidneyDisease(2015);http://www.world kidneyday.org/faqs/chronic-kidney-disease/

[12] Igor Kononenko, Machine Learning for Medical Diagnosis: History, State of the Art and perspective in Artificial Intelligence in Medicine, Vol.23, No. 1 (Elsevier, , 2001).

[13] Hardik Maniya, Mosin I. Hasan, Komal P. Patel, Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis in International Conference on Web Services Computing (ICWSC,2011).

[14] R.Bharat Rao, Jinbo Bi, Nancy Obuchowski and David Naidich, LungCAD: A Clinically approved Machine Learning System for Lung Cancer Detection in International conference on knowledge discovery and data mining (San Jose, California, USA, 2007) ACM 978-1-59593-609-7/07/0008.

[15] P Yasodha, M Kannan, Analysis of Population of Diabetic Patient Database in WEKA Tool in International Journal of Science and Engineering Research, Vol.2, No.5 (May 2011).

[16] Bekir Karlik, Hepatitis Disease Diagnosis using Backpropagation and the Naive Bayes Classifiers in Journal of Science and Technology, Vol.1, No.1, (2011).

[17] Huda Yasin, Tahseen A.Jilani, Madiha Danish, Hepatitis-C Classification using Data Mining Techniques in International Journal of Computer Applications, Vol.24, No.3 (June 2011).

[18] L.C van der Gaag, S Renooij, C.L.M. Witteman, B.M.P Aleman and B.G Taal, Probabilities for a probabilistic network: a case study in oesophageal cancer in Artificial Intelligence in Medicine, Vol. 25, No 2 (Elsevier, , 2002), pp. 123–148.

[19] B. Sokouti, S. Haghipour, and A. D. Tabrizi, A framework for diagnosing cervical cancer disease based on feedforward MLP neural network and ThinPrep histopathological cell image features in Neural Comput. Appl.,Vol. 24, No. 1 (2014), pp. 221–232.

[20] Shivajirao M. Jadhav , Sanjay L. Nalbalwar and Ashok A. Ghatol, Artificial Neural Network Models based Cardiac Arrhythmia Disease Diagnosis from ECG Signal Data in International Journal of Computer Applications, Vol.44 ,No 15 (2012), pp. 8-13.

[21] T Manju, K Priya and R chitra, Heart Disease Prediction System using Weight Optimized Neural Network in International Journal of Engineering and Computer Science, Vol-2, No.3 (2013), pp.781-788,.

[22] S.Vijayarani and S.Dhayanand, Kidney disease prediction using Support Vector Machine and Artificial Neural Network algorithms in International Journal of Computing and Business Research (IJCBR), Vol. 1, No. 3 (2015), pp.1765-1771.

[23] Andrew Kusiak, Bradley Dixonb and Shital Shaha, Predicting survival time for kidney dialysis patients: a data mining approach in Computers in Biology and Medicine, Vol. 35 (Elsevier, 2005), p. 311–327.

[24] Abid Sarwar, Jyotsna Suri, Mehbob Ali and Vinod Sharma, Novel benchmark database of digitized and calibrated cervical cells for artificial intelligence based screening of cervical cancer in Journal of Ambient Intelligence and Humanized Computing (Springer-Verlag Berlin Heidelberg 2016), DOI 10.1007/s12652-016-0353-8.

[25] D Lavanya, K Usha Rani, Ensemble Decision Tree Classifier For Breast Cancer Data in International Journal of Information Technology Convergence and Services, Vol.2, No.1(2012), pp. 17-24.

[26] N.H Barakat, A.P Bradley and M.N.H Barkat, Intelligible support vector machines for diagnosis of diabetes mellitus in IEEE Transactions on Information Technology in BioMedicine(2009).

[27] B.M Patil, R.C Joshi, Durga Toshniwal, Association rule for classification of type-2 diabetic patients in IEEE-Second International Conference on Machine Learning and Computing (2010), p.67. DOI 10.1109/ICMLC.

[28] S Bhatia, P Prakash, G.N Pillai, SVM based Decision Support System for Heart Disease Classification with Integer-coded Genetic Algorithm to select critical features in Proceedings of the World Congress on Engineering and Computer Science, San Francisco, (USA,2008), pp. 34–38.

[29] My Chau Tu, Dongil Shin, Dongkyoo Shin, Effective Diagnosis of Heart Disease through Bagging Approach in 2nd International Conference on Biomedical Engineering and Informatics(2009).

[30] B. Boukenze, A. Haqiq, & H. Mousannif, Predicting Chronic Kidney Failure Disease Using Data Mining Techniques in Advances in Ubiquitous Networking 2, Springer Singapore(2017) , pp. 701-712.

[31] Lambodar Jena and Narendra Ku. Kamila, Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney Disease in International Journal of Emerging Research in Management and Technology, Vol. 4, No. 11(2015), pp. 110-118.

[32] Rubini,L.Jerlin,UCIMachineLearningRepository[http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease].Karaikudi,TamilNadu: Algappa University, Department of Computer Science and Engineering(2015).

[33] Sahil Sharma, Vinod Sharma, Atul Sharma, Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis in International Journal of Modern Computer Science (IJMCS), Vol.4, Issue 3(2016), pp. 11-16.

[34] J. R Quinlen, Introduction of Decision Trees in Machine Learning, vol. 1(1986), pp. 81-106

[35] Yoav Freund, Robert E Schapire, A decision-theoretic generalization of on-line learning and an application to boostingl in Proceedings of the Second European Conference on Computational Learning Theory, EuroCOLT '95, (London, UK, Springer-Verlag, 1995), pp. 23-37

[36] Marina Sokolova, Guy Lapalme, A systematic analysis of performance measures for classification tasks in Information Processing and Management,Vol.45, Issue-4(Elsevier, 2009), pp.427-437.

[37] A Janecek, W N W Gansterer, M Demel, and G Ecker, On the Relationship Between Feature Selection and Classification Accuracy in *Fsdm*, Vol. 4(2008), pp. 90–105.

[38] V Jha, G Garcia-Garcia, K Iseki, et al., Chronic kidney disease: global dimension and perspectives in *Lancet*, Vol. 382 (July 2013), pp. 260-272.