International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: 6.614
Available Online at www.journalijcar.org
Volume 7; Issue 5(L); May 2018; Page No. 13107-13112
DOI: http://dx.doi.org/10.24327/ijcar.2018.13112.2324

Research Article

# CHRONIC KIDNEY DISEASE PREDICTION USING STACKING

**Ankur Sharma[1*] and Sonal Arora[2]**

[1]DPGITM, Maharshi Dayanand University, Haryana, India
[2]Department of Computer Science and Engineering of DPGITM, Maharshi Dayanand University, Haryana, India

**A R T I C L E I N F O**

***Article History:***

Received 12th February, 2018
Received in revised form 9th March, 2018 Accepted 26th April, 2018
Published online 28th May, 2018

***Key words:***

CKD, Stacking, Bagging, SVM, Naive Bayes, ANN

**A B S T R A C T**

Diagnosing the chronic kidney disease in early stage can have great impact on death and diseases caused due to kidney failure. Its symptoms are shown in patient at last stage or may be never till more than 95% kidney damaged, that makes its treatment much hard or impossible. Data analysis techniques can significantly help in prediction of these diseases. Many classification techniques like ANN, Bayes theorem, SVM etc showed great improvement in this field, the main purpose of this research is to evolve a hybrid technique for improving Accuracy and overall performance, so prediction of disease become fast and efficient.



## INTRODUCTION

In human body two beans shaped, fist sized organs named kidney present back side of abdomen. Its main functioning is to filter toxic elements from our body and maintaining concentration of each element accurate in our body by regulation excretion of this element from body [2]. It also helps to control the BP of our body by producing various hormones.

In Chronic kidney disease due to several reasons our kidney starts damaging, its filtering capacity start decreasing. Due to this the balance of minerals and other components starts changing and many diseases can occur in our body cause death. Symptoms of CKD are [1]:-

1. Fatigue
2. High Blood Pressure
3. Loss in appetite
4. Water electrolyte imbalance
5. Kidney damage
6. Abnormal heart rhythm
7. Failure to thrive
8. Fluid in lungs
9. Insufficient urine production
10. Itching
11. Swelling

Kidney disease can be cured easily if predict in early stage. Main reason for not early detection is because its symptoms are neither visible or appears at last stage till that time 95% of kidney damaged and recovery is not possible [19]. After that patient will survive on dialysis and medicine throughout its life.

Data mining is a growing technology in field of extracting and analysis of large data sets and providing decisions and description of query on the basis of dataset[5]. Various techniques used in Data mining are Classification, Clustering, and Association etc. It is currently used in mainly Business, Research, Medical, Education etc where prediction and description of various problems necessary [2].

Various classification techniques - Artificial Neural network, Bayesian network, SVM, GA, Decision tree etc are used today in health care domain. In which ANN and Decision tree provide highest accuracy [3]. But stability and accuracy of these techniques is not as desired. E.g. ANN takes much time for training, while Decision tree is highly sensible as new value can change structure of tree. To Overcome these problems hybrid model is necessary so the training time can be reduced providing high accuracy.

Ensemble is a process of using more than one classification model to produced one model that have improved accuracy and lowered errors. In this study we evolve a hybrid model using SVM, Naive Bayes model and ANN. This hybrid model uses these models advantages and reduced the errors.

### *Classification Techniques*

Classification is a defined as characterisation, in which Objects, data, ideas are understood, and recognised and classify (based on classes) [19]. In Data mining classification is main technique for predicting results on the basis of prior knowledge. Classification derives model from training that

**Corresponding author:* **Ankur Sharma**
DPGITM, Maharshi Dayanand University, Haryana, India

categories data according to its class and a queried data is processed to identify its class. The process of classifying unknown data set class is called prediction. Various classification techniques used for prediction of disease are ANN, SVM, Decision trees, Decision Rules, Logistic reasoning, Genetic Algorithm etc.

We use hybrid classification algorithm instead of using single classification algorithm. Various techniques for classification are below mentioned:

***Artificial Neural Network (ANN)***

Artificial Neural network are human nervous system artificial implementation. In this model Nodes represent neurons of human body. ANN is group of these interconnected neurons. ANN maps Input patterns to the output patterns by changing these interconnections. ANN is a supervised learning classification technique. In which training data used to creation of model and the weight of connection is updated during training so it will classified data in to correct output.

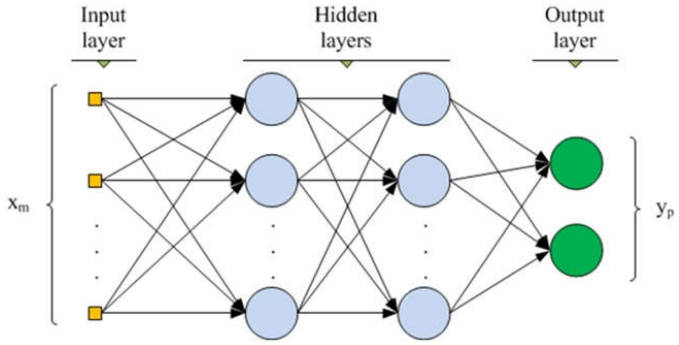


**Fig 1** Multilayer Neural Network

***Support Vector Machine (SVM)***

SVM is mathematical based supervised learning model. It separate data sets into regions (classes) by separation plane this plane is described by mathematical equations. These plains are optimal as they separate all datasets maintaining maximum possible margin between nearest data set from separation plane, so integrating new data set is possible without any ambiguities.

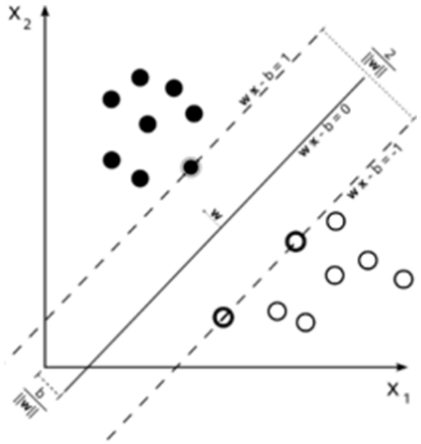


**Fig 2** Linear separation between classes by SVM

***Naive Bayes***

Naive Bayes classifier is based on Bayes theorem, states probability of an event (attribute) based on prior knowledge of condition that may be related to the event. Main assumption of this law is that attributes in data must be exclusive to each other.

Naive Bayes model training is done by evaluating closed form expression in linear time while other techniques in classification used iterative approach for evaluation.

$$p(ci|d) = \frac{p(d|ci)p(ci)}{p(d)}$$

Where

**p(ci|d)** - probability of attribute d that it lies in class ci
**p(d|c)** - probability of instance d given class is ci
**p(ci)** - probability of occurrence of class ci in classes
**p(d)** - probability of instance d occurring

***Ensemble***

Ensemble means a group of items viewed as a whole rather than individual, is a technique is used for combining more than two weak classifiers to classify a data sets. By combining them there advantages also combined hence improve accuracy and stability.

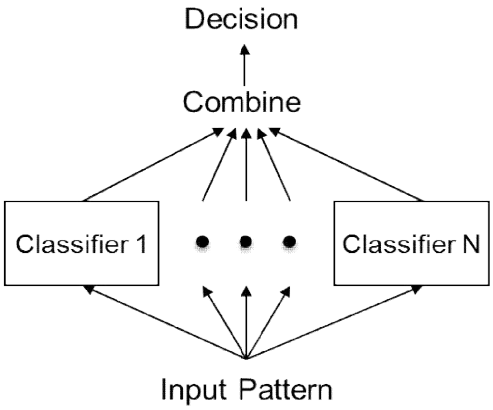


**Fig 3** Ensemble in Data Mining

There are various Ensemble techniques for combine classifiers: -

***Boosting***

Boosting means help to encourage or improve something. It is an ensemble technique increasing power of any algorithm at the very beginning. In this techniques homogenous models are used one model improve performance of its predecessor model hence overall improving model efficiency.

In Data mining ADABoost is the most used boosting algorithm.

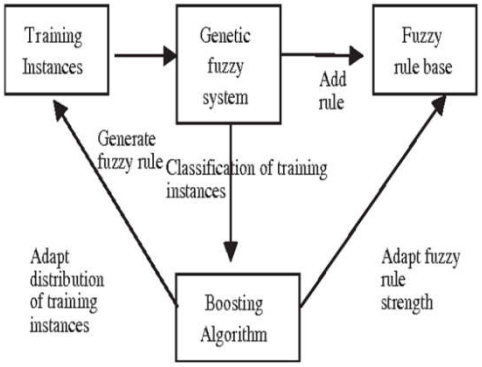


**Fig 4** Boosting

***Bagging***

In Bagging Data sets are divided into subsets and several model prepared using these data sets. The overall output of hybrid model is the average or mean of these models output.

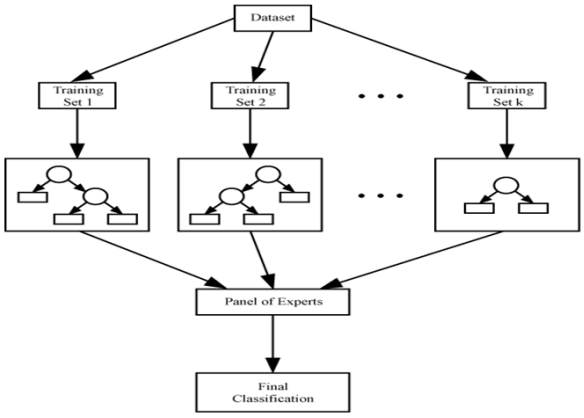


**Fig 5** Bagging

### *Voting*

Voting is similar to bagging except it chooses the final results on the basis of voting from the result. The highest voted model result is taken for hybrid model.

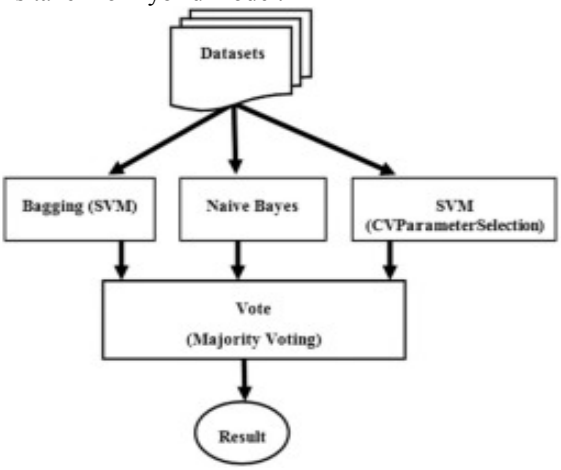


**Fig 6** Voting

### *Stacking*

Stacking means arrange a number of things in a pile. In Data mining Stacking technique, more than one classifiers to classify Dataset and provide a Base (Meta learner) layer for the blending of these models output. Meta layer then takes input the second layer output and provide the learned output as whole model output. Stacking model is also called as Super Learner model.

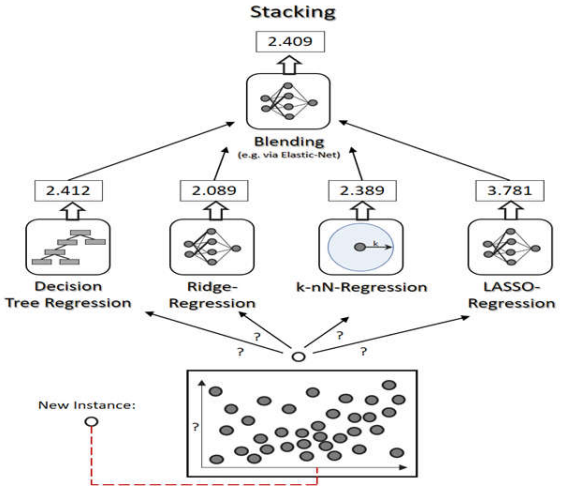


**Fig 7** Stacking

### Literature Survey

In medical domain data mining is a popular and efficient technique for early stage prediction. Many researchers had worked in this field for improvement of data.

Radha Mothukarri *et al*[3] focus there research work for prediction in renal disorder(Diabetes Mellitus) disease with various data mining techniques - Bayesian networks, ANN, SVM, Decision trees etc and found that Decision tree algo C4.5 (extension of ID3 algo) is much efficient and accurate than others. It execute in minimum time.

M.A. Muslim *et. al*. [4] improves the performance of j48 algo up to 1.5% by pessimistic pruning. Pruning is used to identify and remove branching that is not needed now. So decrease space requirements and execution time of algo.

Gopika and Dr. M Vanitha [5] increased the prediction accuracy of chronic kidney disease by improving fuzzy classification algo to Hybrid Fuzzy C-Means. It is an improvement of FCM with Euclidean space.

Bharti yadav *et. al*. [6] proposed a healthcare system using Artificial Neural network with supervised learning. They used Data set of 276 instances each having 11 common attributes, and got accuracy of 98.36% using Matlab software. This system helps people early detection and prediction of various common diseases and also reduced the visit of health care. This system is much helpful in poor areas where health care system is not so good.

Taban Eslami and Fahad Saeed [7] research for the Attention Deficit Hyperactivity Disorder (ADHD) brain disease in children. They used EROS technique to computing similarity between two multivariate time series along with K-NN algo. They use public dataset provided by ADHD-200 consortium and designed a scheme named J-Eros which picks optimal values from KNN training data.

M Nikihil kumar *et. al*. [9] conduct experiments for prediction of Coronary Artery Disease. They experiment various data mining techniques and conclude that Decision tree C4.5 algo implementation in weka J48 is provided best accuracy of 56.76 % on training data set.

Jovelin M Lapatase *et. al* [10] research for getting the relationship cost of living vs. quality of healthcare. They analyse data of 36 Asian cities. From this data they generate patterns using Minitab Software and produced relationship using regression data mining technique. They conclude that area where cost of living is high more people depends on the government healthcare programs but the quality of these is not good so various system can be provided to overcome deficiency of healthcare quality.

Sai Prasad Pothuraju *et. al*. [14] work on improvement in prediction of Cardiotocography disease. They improve pre-processing technique by Symmetric Minority over sampling technique (SMOTE) which makes the dataset balanced and hence improvement is observed in IBK algorithm from 96.89 to 98.05%. They used UCI repository public dataset having 2126 instances with 23 attributes.

A. Suresh *et.al* [17] developed an Assessment model for health care system. They pre-processed data using outlier detection with k-Means clustering technique. Data set used in their study

is Pima Indian Dataset. Proposed model provide accuracy of 98.79 % with this data set.

***Proposed Approach***

In this paper we use Stacking ensemble technique for hybrid model. In our proposed model the classifier layer contains SVM and Naive Bayes models for data classification. Meta classifier used for our model is Artificial Neural Network Multilayer perceptron model. We are using publically available UCI repository of chronic kidney disease Data set for working. This Data set contains 400 instances each have 25 attributes of nominal and integer type. The attribute and its description are given in table 1.

We use Weka 3.9.2 version for our experiments. Weka is a powerful tool for Machine learning algo creation and verification purpose. We split data set into two parts 66% training instances used for training data model and rest is used for testing the model. Testing is done using 10 fold Cross validation method.

***Proposed model work as follows***

1. Random data subset is formed from the training data set.
2. These data set is provided to the classifiers model SVM and Naive Bayes
3. Classifier produced output classification.
4. This classified output is then processed by Meta Layer classifier, this layer also called blend layer.
5. Main function of this layer is to choose the best data among the output of the previous layer
6. Resulted output is then tested with training data set

***Performance is measured in terms of***

1. Accuracy - number of correct prediction of data made from total no of prediction made
2. Time taken in building model
3. Time taken in testing data
4. F-Measure - combination of precession and recall
5. Precession - Positive predictive values

First we pre-process data set and replace the empty values with default observational value. This pre-processed data set is used for classification experiment.

**Table 1** Attributes of CKD Dataset

| Attribute Name | Description | Type / Values |
|---|---|---|
| Age | Age | Numerical |
| BP | Blood Pressure (mm/Hg) | Numerical |
| SG | Specific Gravity | Nominal (1.005, 1.010, 1.015, 1.020, 1.025) |
| AL | Albumin | Nominal (0-5) |
| SU | Sugar | Nominal (0-5) |
| RBC | Red Blood Cell | Nominal (normal, abnormal) |
| PC | Pus Cell | Nominal (normal, abnormal) |
| PCC | Pus Cell Clumps | Nominal (present, not present) |
| BA | Bacteria | Nominal (present, not present) |
| BGR | Blood Glucose (mgs/dl) | Numerical |
| BU | Blood Urea (mgs/dl) | Numerical |
| SC | Serum Creatinine (mgs/dl) | Numerical |
| SOD | Sodium (mEq/dl) | Numerical |
| POT | Potassium (mEq/dl) | Numerical |
| HEMO | Haemoglobin (g) | Numerical |
| PCV | Packed Cell Volume | Numerical |
| WC | White Blood cell count (cell/cumm) | Numerical |
| RC | Red Blood cell count (millions/ cmm) | Numerical |
| HTN | Hypertension | Nominal (Y/N) |
| DM | Diabetes Mellitus | Nominal (Y/N) |
| CAD | Coronary artery disease | Nominal (Y/N) |
| APPET | Appetite | Nominal (good, poor) |
| PE | Pedal Edema | Nominal (Y/N) |
| ANE | Anaemia | Nominal (Y/N) |
| CLASS | Classification group | Nominal (CKD/notCKD) |

Figure 8 shows the unprocessed data. We processed data by replacing missing values with mean values.

The processed data is shown on figure 9.

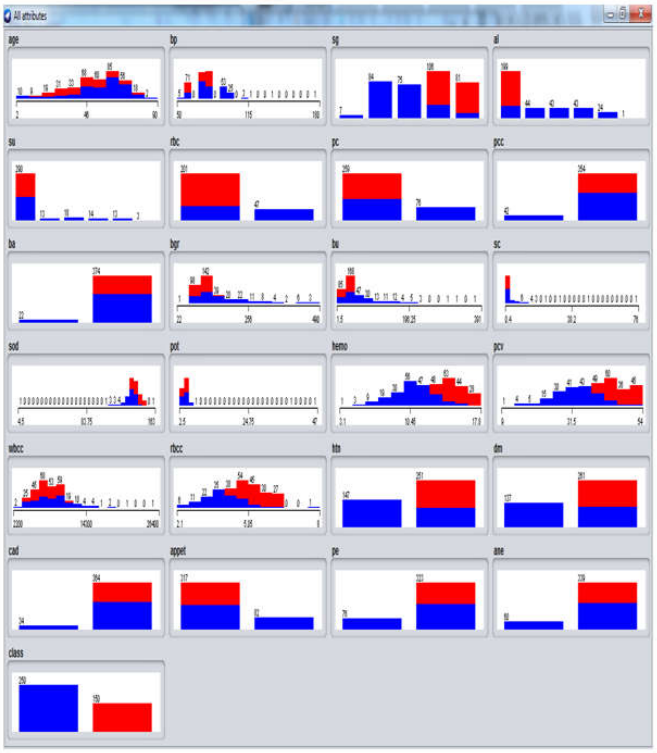

**Fig 8** Data set attributes before pre-processing

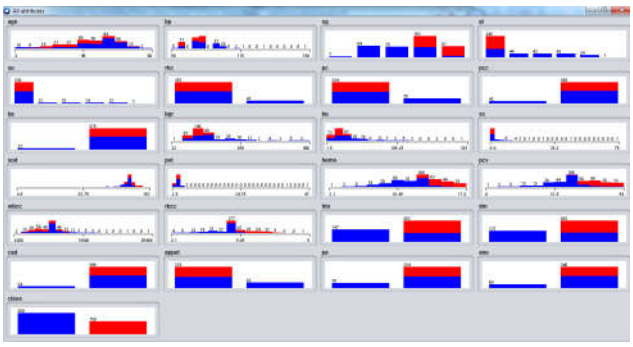

**Fig 9** Data set after pre-processinig

## RESULT AND DISCUSSION

**Table 2** Performance MEASURES of Classification techniques

| Algo / Test | SVM | Naive Bayes | ANN | Hybrid Model |
|---|---|---|---|---|
| Accuracy | 97.0588 | 94.8529 | 97.058 | 98.5294 |
| Time taken to Build | 0.08 | 0.02 | 2.66 | 0.23 |
| Time Taken to Test | 0.01 | 0.03 | 0.001 | 0.002 |
| Precision | 0.973 | 0.955 | 0.973 | 0.986 |
| Recall | 0.971 | 0.949 | 0.971 | 0.985 |
| F-Measure | 0.971 | 0.949 | 0.971 | 0.985 |

**Table 3** Error result of Classification Technique

| Algo / Test | SVM | Naive Bayes | ANN | Hybrid Model |
|---|---|---|---|---|
| Mean Absolute Error | 0.0294 | 0.0552 | 0.0234 | 0.0226 |
| RMS Error | 0.1715 | 0.2187 | 0.117 | 0.1176 |
| Relative Absolute Error | 6.3003 | 11.8173 | 5.0055 | 5.2338 |
| Root Relative Squared Error | 35.7951 | 45.6391 | 24.4144 | 24.5433 |

From results the overall accuracy of this model increased from 97.058 to 98.5882% While other errors are decreased 0.0294, 0.0552 to 0.0244(Mean absolute error), F-Measure and recall also increased in proposed model.

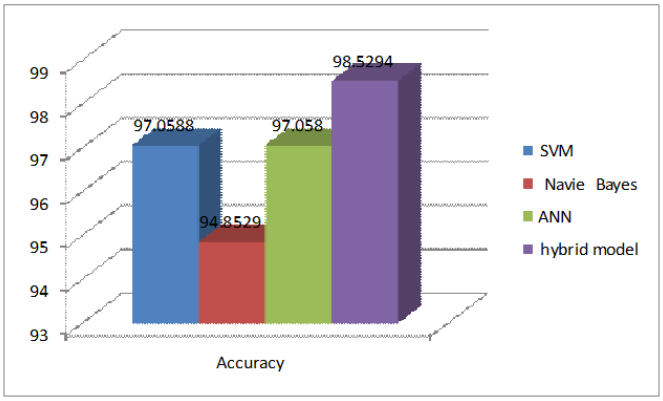


**Fig 3** Accuracy

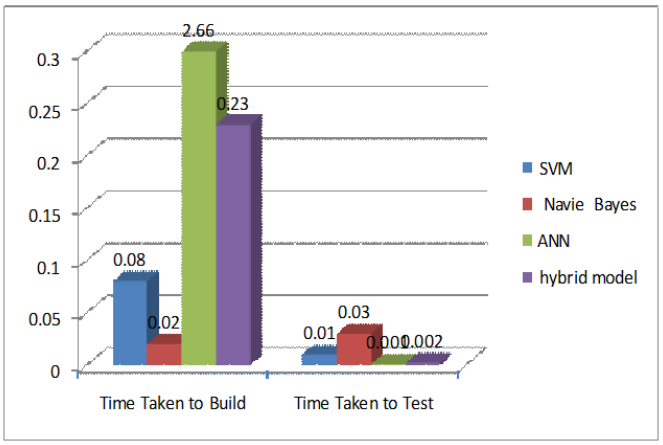


**Fig 4** Time Taken to Build and Test Model

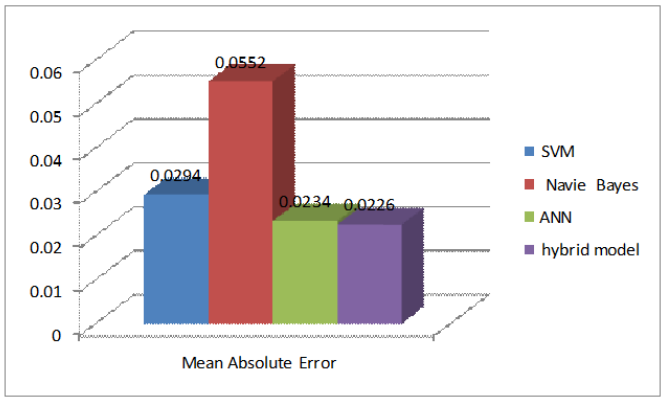


**Fig 5** Absolute Error

Results state that the accuracy of proposed model increased and the property of Meta layer (base learner ANN) are also combined with weak classifiers (SVM, Naive Bayes).

In this model SVM and Naive bayes are weak classifiers. There classification processing is fast, While ANN has best classification accuracy but take long time to build model with large number of attributes. This hybrid model used fast classification processing of SVM and Bayes model having less accuracy of 97.0588 and improves its accuracy using ANN model. Since at Meta layer output of two classifier is used to classify so ANN processing time reduced and hence decrease overall model building time to 0.23 sec.

## CONCLUSION

Different classification techniques are studied and reviewed there performance in Chronic kidney disease data set prediction are analysed. Various methods of Data mining have different power in various data sets. In this study the proposed method provide highest accuracy of 98.5582% with least Mean absolute error of 0.0244 which makes the model much stable. Therefore stacking ensemble method provides best accuracy and performance than single classifier.

### *Future Work*

In this study we focus only on chronic kidney disease and for small data sets (400 instances). In future this method can also be applied for other data set as well and its accuracy for large data sets will be analysed.

Here we used classifiers ANN, SVM and Naive Bayes, in future other classifiers such as decision tree, decision rules or combination of these can be used for classifier layer and other

combination can be used for base Meta layer. Also in future new model can be deduced from this which will predict for all type data sets with similar accuracy.

## References


1. Radha Mothukuri, K Gurnadha Guptha, Performance Prediction of Chronic Kidney Disease using various Data Mining Techniques, Dec 2017
2. M A Muslim, A J Herowati,E Sugiharti, B Presitiyo, Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease, 2017
3. S. Gopika, Dr. M. Vanitha, Efficiency of data mining techniques for predicting kidney disease , Nov 2017
4. Bharti Yadav, Shilpi Sharma, Ashima Kalra, Supervised learning techniques for prediction of diseases, April 2018
5. Himanshu Das, Bighnraj Naik, H.S. Behera, Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach, April 2018
6. Taban Eslami, Fahad Saeed, Similarity based classification of ADHD using singular value Decomposition, 2018
7. M Nikhil Kumar, K.V.S. Koushik, K.Deepak, Prediction of heart diseases Using Data Mining and machine learning Algorithms and tools, 2018
8. Jovelin M lapates, Sales G. Aribe Jr., Jennifer P. Barroso and beulah Joy K. damasco , Health quality and cost of Living in Asian cities, 2017
9. Husna Aydadenta, Adiwijaya , On the classification techniques in data mining for microarray data classification, 2004
10. Thulasi Bikku, Alapati Padma Priya, A novel algorithm for clustering and feature selection of high dimensional dataset, Jan 2018
11. Dhanashri Gujar, Rashmi Biyani, Tejaswini Bramhane, Snehal Bhosale, Tejaswita P. Vaidya, Disease prediction and Doctor Recommendation System
12. Sai Prasad Pothuraju, M. SreeDevi, Vinay kumar Andey, Ravi kumar Tirandasu, Data mining approach for accelerating the classification accuracy of cardiotocography, March 2018
13. Mr. Chala Beyene, Prof. Pooja Kamat, Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Technique., 2018
14. Jothi, R.Sudha, Heart Disease Prediction System Using Naive Bayes, March 2018
15. Suresh, R. Kumar, R. Varatharajan, Health care Data Analysis using evolutionary algorithm, March 2018
16. Miss Sheetal A. Tayade, Prof. Parag D. Thakare, Review on knowledge discovery and analysis in healthcare using Data mining, March 2018
17. Shivani S. Waghade, Prof. Aarti M. Karandikar, A comprehensive study of healthcare Fraud detection based on machine learning,2018


**How to cite this article:**

Ankur Sharma and Sonal Arora (2018) 'Chronic Kidney Disease Prediction Using Stacking', *International Journal of Current Advanced Research,* 07(5), pp. 13107-13112. DOI: http://dx.doi.org/10.24327/ijcar.2018.13112.2324

*******