A Hybrid Analytic Model for the Effective Prediction of Different Stages in Chronic Kidney Ailments

P. Antony Seba
Indian Institute of Information Technology Kottayam
Bibal Benifa JV (benifa.john@gmail.com)
Indian Institute of Information Technology

Research Article

Keywords: Random Forest (RF), AdaBoost (AB), Voting Classifier, CKDstage, Predictive Analytics

Posted Date: June 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-550159/v1

A HYBRID ANALYTIC MODEL FOR THE EFFECTIVE PREDICTION OF DIFFERENT STAGES IN CHRONIC KIDNEY AILMENTS

P. Antony Seba[a], J.V. Bibal Benifa[b],

a, b Indian Institute of Information Technology Kottayam

[b]Corresponding author email: benifa.john@gmail.com

ABSTRACT

Chronic Kidney Disease (CKD) is a gradual loss of kidney function over the period of time and it is irrevocable once functionality reaches the critical state. Detecting the various stages of CKD helps to reduce the progression of the disease. Accurate prediction of CKD stages is one of the urgent needs in the medical industry and it can be effectively done by adopting Machine Learning (ML) techniques. The primary objective of the present research is to develop an effective classification model for the accurate prediction of CKD stages based on the patient's health profile as well as the clinical test reports. Here, a hybrid ML strategy is employed that integrates Random Forest (RF) and AdaBoost (AB) techniques through a voting classifier (VC). The standard CKD dataset with 400 tuples and 25 parameters is used for the proposed investigation. The Modification of Diet in Renal Disease (MDRD) equation is used to extract an additional feature known as "estimated Glomerular Filtration Rate (*eGFR*)" for the prediction of the CKD stage. Pre-processing is carried out on the CKD dataset to fill the missing values by considering the skewness of the parameters and the issue of data leakage is also well addressed. Medically important features are considered and Correlation analysis is carried out to select the appropriate features for the model building process.The proposed Hybrid Ensemble Model (HEM) aids in lowering the bias and variance. HEM model efficiency is assessed using the performance metrics such as cross validation score (CVS), accuracy, precision, recall, F1 measure, Mean Squared Error (MSE), bias and variance and it is compared with the state-of-the-art classification schemes. The outcomes of the analysis reveal that the proposed HEM ensures that the CKD stage prediction is more accurate with 99.16%, 100%, 100% in reduced feature set I, set II, set III and with cross validation score of 97.85%, 99.28%, and 99.64% with reduced features set I, set II and set III respectively.

Keywords: Random Forest (RF), AdaBoost (AB), Voting Classifier, CKDstage, Predictive Analytics.

1. INTRODUCTION

Chronic Kidney Disease (CKD) is a progressive loss of kidney function over a specific time period [1]. Kidney is an organ that maintains the balance of minerals and electrolytes in a human body. CKD develops gradually and it is not possible to reverse its state to initial condition but can be controlled [2]. The diagnosis of CKD starts with ethnicity, heredity, blood pressure (*bp*), diabetes level (*su*), potassium level (*pot*), blood urea nitrogen (*bun*), serum creatinine (*sc*) and Glomerular Filtration Rate (*gfr*) [3]. However, in general the blood as well as urine samples are preferred for the testing purposes.

The stages of CKD shall be identified by the way of accurate measurement of the various levels of kidney function. Creatinine level of the patient (i.e.)., the amount of unwanted substances in the blood would determine the *eGFR* value, and it should be noted that the lesser *eGFR* results in the greater risk of kidney disease. The factors such as *age*, *sc*, *race* and *gender* have a great impact on the *eGFR* value [4]. The rate of kidney ailments varies among various demographic groups and the higher creatinine level lowers the *eGFR* value. Various ranges of *eGFR* values are used to fix the stages of kidney ailments. In general, the *eGFR* level < 60 mL/min prevails continuously for more than three months or Albumin Creatinine Ratio (ACR) > 30 mg/g is considered to be a key symptom of CKD. The five stages of CKD can be identified by calculating the *eGFR* levels using the MDRD study equation. CKD Stage 1 is classified as minimal kidney disease for which the *eGFR* > 90 mL/min. Mild decrease in kidney function is classified as Stage 2 CKD and the *eGFR* is in the range of 60 - 89 mL/min. Stage 3A is the moderate stage of CKD with *eGFR* about 45-59 mL/min and Stage 3B is the moderate stage of CKD with *eGFR* of 30 - 44 mL/min. Stage 4 is the severe CKD with *eGFR* level is 15-29 mL/min and Stage 5 is end stage renal disease with *eGFR* < 15 mL/min.

Most of the research works reported herein so far utilized the benchmark dataset available in the University of California Irvine (UCI) repository for the effective prediction of CKD using statistical as well as ML algorithms [5]. It should be noted that very few works have been reported the prediction of severity levels of CKD [6] [7]. In this article, a multi-classifier prediction model is proposed for the effective prediction of CKD as well as its severity by extracting *eGFR* as a main feature in the dataset [8]. Further, the data leakage problem is also well addressed and correlation analysis is carried out to select the appropriate features to enhance the performance of the proposed HEM [9] [10]. In addition, certain medically important features are extracted manually to form a new subset to train the HEM [1] [2].

2. RELATED WORK

For the past two decades, several research works were reported for effective classification and prediction of various diseases at the early stages. Many data mining techniques and ML algorithms are utilized for prediction in healthcare applications viz., Support Vector Machine (SVM) algorithm is utilized by many researchers to detect the presence of diabetes [11], Alzheimer disease [12] etc., based on the clinical reports and laboratory test records of the patients. Probabilistic Neural Network (PNN) algorithm was used by Dessai et al. (2013) towards heart disease prediction [13]. Also, the digitization of medical records (i.e.)., Electronic Health Records (EHR) leads to accurate prediction of diseases using Artificial Intelligence (AI) techniques, which in turn helps the medical practitioners to react towards effective treatment to their patients. It was inferred that a significant research gap exists towards the accurate prediction of CKD and its severity by considering the data leakage aspects as well [14] – [18].

Accurate prediction of the progression of CKD stage is really a challenging task because the disease will not indicate any symptom at the early stages. The medical treatment and diet prescription are completely depending on the severity of CKD and its rate of progression. Physicians solely depend on the clinical and laboratory test records of the CKD patients for the identification of different stages. The patient's demographic information and clinical test reports such as blood pressure, blood sugar, serum creatinine, coronary artery disease, race, albumin creatinine ratio and especially the *eGFR* play a vital role in the prediction of various stages of the CKD [4]. Currently, *eGFR* is used to estimate the level of kidney function and with the continuous estimation of this measure for a stipulated period of time the stage of the disease is defined.

The benchmark dataset for CKD analysis is usually large and probably with missing values and hidden features, which may be essentially required for accurate prediction [5]. Preliminary study on the prediction of various stages of CKD shall be carried out by generating the complete dataset by considering the above stated attributes including race, gender and *eGFR* of the patients using an online GFR Calculator [19]. El-Houssainy et al. (2019) used efficient data mining techniques to extract hidden information from the patient data and hence the accuracy of prediction of CKD stage was improved [6]. Here, four classifiers were applied towards the accurate prediction of CKD stages and observed that PNN yields better results for predicting the severity of CKD.

Elhoseny et al. (2019) introduced a hybrid density-based feature selection with Ant Colony Optimization (ACO) algorithm to eliminate the redundant features prior to the classification in a benchmark CKD dataset [20]. The proposed intelligent framework consists of pre-processing, optimal feature selection and classification of CKD. Here, the application of ACO in the classification process involves a structural scheme, generation and pruning of rules, and a heuristic function to enhance the predictive results with pheromone update. The proposed hybrid algorithm is simulated in the MATLAB environment by considering the clinical features that influence CKD and the performance metrics such as False Positive Rate, False Negative Rate, accuracy, specificity, sensitivity, F-score and Kappa value that are evaluated and compared with other existing classifiers. This algorithm is efficient towards the identification of CKD with a significant invention in classification accuracy using fewer features.

Gabriel et al. (2019) introduced a Neural Network (NN) based classifier to predict the risk of developing CKD in the Colombian demographic by considering two population groups (people diagnosed with & without CKD) [21]. This model predicted the likely course of the medical condition of CKD using the test dataset with more accuracy. The accuracy of prediction is justified by an example Case-Based Reasoning (CBR) with appropriate explanation. The demographic data and medical care information obtained through previous diagnoses of about 20,000 people with CKD and 20,000 people without CKD are used as test dataset for training and validating the proposed NN-CBR twin system model. Considering this dataset with larger features, the proposed NN model with 5 layers (including 3 hidden layers) predicts the risk of developing CKD with an accuracy of 95 %.

Bilal Khan et al.(2020) carried out an experimental analysis on various ML techniques with an objective to determine the best classifier (i.e.)., to predict CKD or NOTCKD accurately from the dataset of kidney patients acquired from UCI repository [22]. The dataset has been pre-processed to supplant the missing values with the mean of the existing values. Seven ML techniques such as Naïve Bayes (NB), Logistic Regression (LOG), Multi-Layer Perceptron (MLP), J48 Decision Tree (DT), SVM, NB Tree and Composite Hypercube on Iterated Random Projection (CHIRP) have been employed on the UCI dataset. The outcomes are examined by N-fold cross validation procedure utilizing the classification error rates, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error and Root Relative Squared Error. From the overall analysis, it is observed that CHIRP outperforms in terms of reduced error rates and improved accuracy (99.75%).

Jiongming Qin et al. (2020) proposed an integrated classifier model that combines LOG and RF by perceptron over the CKD dataset obtained from the UCI ML repository to improve the prediction accuracy [23]. Initially, the dataset was tuned using K-Nearest Neighbour (KNN) imputation, where the numerical missing values are filled by the median and the categorical missing values are filled with the mode of K-samples to maintain similar physiological measurements among people with similar physical conditions. Further, LOG, RF, SVM, KNN, NB and Feed Forward Neural Network (FFNN) are evaluated using complete and tuned CKD dataset for optimal feature selection. The independent model or the integration of one model with the other which yields better performance has been identified and misjudgement analysis is carried out on such models. As LOG computes optimum adjusted r-squared value while RF contributes in the reduction of Gini Index, the experimental results show that the integrated model with signum activation function performs well in CKD diagnosis by achieving an average accuracy of 99.83 %.

Hosseinzadeh et al. (2020) have utilized smart multimedia medical devices and sensors for remote monitoring of kidney function [6]. The authors proposed an IoT based model for effective prediction of CKD and its severity using the multimedia data acquired from various IoT devices. DT classifier is adopted for the prediction of CKD and its stages by appropriately selecting the features based on the clinical observations and using the results of previous studies for CKD prediction. The performance of J48 classifier is compared with SVM, MLP and NB classifiers and it has been proved that J48 is more accurate and sensitive with specificity. Further the feature sets, which have more influencing parameters on CKD are selected to reduce the execution time for prediction.

Krishnamurthy et al. (2020) developed a ML model using the dataset, which consists of patient's details who are having simultaneous presence of two or more diseases [24]. The dataset obtained from Taiwan's National Health Insurance Research Database is analysed using various classifiers including Convoluted Neural Networks (CNN) and observed that the CNN performed well as a result of a 5-fold-cross-validation process used for the assessment of performance metrics. Lot of research works are reported in the literature for the prediction of CKD as presented in Table 1.

Table 1. Comparison of related works

Reference		Mohamed Elhoseny et al. [20]	Gabriel et al. [21]	Bilal Khan et al. [22]	Jiongming et al. [23]	Hossein zadeh et al. [7]	Surya et al. [24]	El-Houssainy et al. [6]	Ananda nadarajah et al. [25]
Dataset		UCI	Colombian Dataset	UCI	UCI	UCI + sensor data	Insurance Claim data - Taiwan	UCI	UCI
Feature Extraction		X	X	X	X	X	X	X	X
Data Leakage Handled		X	X	X	X	X	X	X	X
Data Cleaning		X	X	✓	KNN	X	Dropped	X	X
Statistical /Model based Feature Selection		Statics	model	✓	Regression, RF	X	Tree based	✓	Common Spatial Pattern
Classifier(s)		D-ACO, ACO, PSO,Olex GA	NN-CBR, SVM, RF	NBTree, J48, SVM, LR, MLP, NB, CHIRP	LOG, RF, SVM, KNN, NB, FNN, Integgrated LOG and RF	DT, SVM, MLP, NB	LR, DT, RF, XGBoost, AB, LightGBM, CNN, BLSTM	PNN, MLP, SVM, RBF	KNN, LDA
Best Classifier		D-ACO	NN	CHIRP	Integgrated LOG and RF	DT	CNN	PNN	KNN
Predicting CKD Stages		X	X	X	X	✓	X	✓	X
Performance measures	Cross Validation Score	X	X	X	X	X	✓	✓	X
	Accuracy	95	95	99.75	99.83	97	90.5	96.7	98.81
	Precision and Recall	X	✓	✓	X	X	✓	✓	✓
	Bias and Variance	X	X	X	X	X	X	X	X

It is inferred from the literature that rather than using a single algorithm, if multiple classifiers are grouped, prediction accuracy will be improved with the combined effects of overall classifiers as shown in Table 1. The prediction becomes better with proper assigning of weights via voting process. In this work, a HEM is proposed to make CKD stage prediction and additionally required features are extracted using MDRD equation to predict the severity of CKD.

Exploratory data analysis is carried out on the training dataset to prevent data leakage. Further, the missing values are filled by evaluating the skewness of the variables. Two reduced subsets of features have been formulated using correlation analysis and medically important features are selected from the training dataset. The ML algorithms such as RF, AB and Voting classifiers are applied for better classification and prediction of the stages of CKD.

3. PROPOSED METHODOLOGY

In the proposed work, it is planned to utilize the benchmark dataset from the UCI repository. The dataset has been splitted into training and test datasets in the ratio of 70:30 respectively in a stratified manner. The training dataset is pre-processed to build a HEM by combining the RF and AB ensemble classifiers and the test dataset is used to validate the trained model towards accurate prediction of CKD stages. Further, the trained model is evaluated statistically with the performance metrics such as MSE, bias, variance, precision, recall, F1 measure, accuracy and cross validation scores. The complete model has been built gradually from the data collection to performance analysis and the outcome of the same is depicted in Figure 1. The procedure adopted in the building blocks of the model is presented in the following sections.

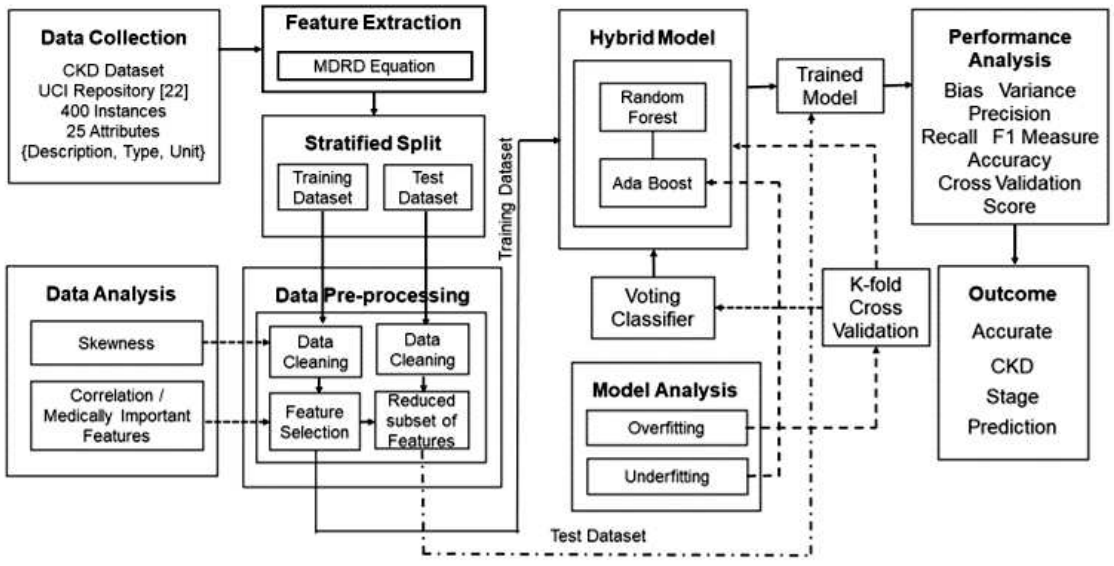

Figure 1. CKD Stage Prediction – Model Building Process

3.1 Data Collection

The CKD dataset has been obtained from UCI machine learning repository. It has 400 instances, 25 attributes with 11 numeric and 14 categorical attributes to predict either “ckd” or “notckd”. The dataset consists of laboratory test records of real cases and the attributes are either

quantitative or qualitative and each attribute is represented with three entities {description, type, unit} as shown in Table 2. Few attributes such as gender, race and estimated glomerular filtration rate are missing in this dataset, which are recommended as essential features by Kidney Disease Improving Global Outcomes (KDIGO) guidelines to predict the stages of "ckd" patients [4].

Table 2. CKD Dataset (UCI Repository)

Sl.No.	Attribute	Description	Type	Units/Range
1	*age*	Age	numerical	age in years
2	*bp*	Blood Pressure	numerical	bp in mm/Hg
3	*sg*	Specific Gravity	nominal	(1.005,1.010, 1.015, 1.020, 1.025)
4	*al*	Albumin	nominal	(0,1,2,3,4,5)
5	*su*	Sugar	nominal	(0,1,2,3,4,5)
6	*rbc*	Red Blood Cells	nominal	(normal, abnormal)
7	*pc*	Pus Cell	nominal	(normal, abnormal)
8	*pcc*	Pus Cell clumps	nominal	(present, notpresent)
9	*ba*	Bacteria	nominal	(present, notpresent)
10	*bgr*	Blood Glucose Random	numerical	bgr in mgs/dl
11	*bu*	Blood Urea	numerical	bu in mgs/dl
12	*sc*	Serum Creatinine	numerical	sc in mgs/dl
13	*sod*	Sodium	numerical	sod in mEq/L
14	*pot*	Potassium	numerical	pot in mEq/L
15	*hemo*	Hemoglobin	numerical	hemo in gms
16	*pcv*	Packed Cell Volume	numerical	mL in volume
17	*wc*	WBC Count	numerical	wc in cells/cumm
18	*rc*	Red Blood Cell Count	numerical	rc in millions/cmm
19	*htn*	Hypertension	nominal	(yes,no)
20	*dm*	Diabetes Mellitus	nominal	(yes,no)
21	*cad*	Coronary Artery Disease	nominal	(yes,no)
22	*appet*	Appetite	nominal	(good,poor)
23	*pe*	Pedal Edema	nominal	(yes,no)
24	*ane*	Anemia	nominal	(yes,no)
25	*class*	Class	nominal	(notckd, ckd)

3.2 Feature Extraction

As per the guidelines of KDIGO, the CKD and its stages are predicted with the *eGFR* using the patient's demographic information and clinical reports. The feature *eGFR* is extracted with the help of existing attributes '*age*' and '*sc*' and populated demographic attributes '*race*' and '*gender*'. The *eGFR* is estimated using Equation (1), which is known as the MDRD equation [19].

$$eGFR = 175 * (S_{cr})^{-1.154} * (age)^{-0.203} * 0.742\ [if\ female] * 1.212\ [if\ black] \quad (1)$$

Where, S_{cr} is the standardized *sc*. The dataset is enhanced with newly extracted features and multiclass labels as shown in Table 3. The CKD dataset includes the newly extracted features and instances that are relabelled as {notckd, stage1, stage 2, stage 3, stage 4, stage 5} against {notckd, ckd} as per KDIGO guidelines.

Table 3. CKD Dataset with newly Extracted Features

Sl.No	Attribute	Description	Type	Units/Range
1	*gender*	Gender	nominal	(female, male)
2	*race*	Race	nominal	(black, other)
3	*egfr*	estimated Glomerular Filtration Rate	numeric	mL/min/1.73m^2
4	*class*	Class	nominal	(notckd, stage1, stage2, stage3, stage4, stage5)

3.3 Stratified Split of Dataset

The benchmark CKD dataset has to be splitted into training and test datasets to envisage the model to learn and then to validate itself. Splitting the CKD dataset randomly leads to imbalanced distribution of classes (i.e., CKD stages) which affects the performance of the model for the minority classes. Therefore, the dataset is splitted in a stratified way (in the ratio of 70:30) to maintain an equal distribution of classes. The training (70 %) and test datasets (30 %) are to be pre-processed separately to avoid the data leakage problem.

3.4 Data Pre-processing

Exploratory data analysis is the preliminary investigation on data to identify the factors that cause the progress of CKD through patterns, anomalies, statistical information and graphical representations of the dataset. The enhanced dataset with additional extracted features after the stratified splitting process is utilized for data tuning. The pre-processing is carried out on the training as well as test datasets independently to address the issues of data leakage [9]. The fine-tuned training dataset is used for model building and the pre-processed test dataset is utilized to validate the learnt model.

a). Data Cleaning

Data Cleaning is done to handle the missing values in the dataset through univariate data analysis. To quantify the average value, the distribution of data around this average value and the overall degree of asymmetry over the full range of observed values are considered. An estimate of central tendency is used to identify the midpoint of data distribution and to measure the metrics

used (mean, median and mode) to fill the missing values [21]. The mean is the weighted average of all values (n), based on the relative frequency as expressed in Equation (2).

$$Mean = \sum_{i=1}^{n} f_i x_i \quad (2)$$

where, f_i is the relative frequency and it is assumed to be $1/n$. Median is the midpoint of the distribution as expressed in Equation (3). For ungrouped data,

$$Median = \begin{cases} \frac{n+1}{2}, & n\,is\,odd \\ \frac{n}{2}, & n\,is\,even \end{cases} \quad (3)$$

The values of mean, median and mode are equal for symmetrical distributions and the mean is strongly influenced by the extreme values, whereas median is more robust and less sensitive to outliers. For asymmetrical distribution, median lies between mode and mean. Skewness is a statistic measure used to reveal the asymmetry of a probability distribution. A measure of skewness (Equation 4) indicates the degree of symmetry in a dataset. For the more skewed distribution, the higher variability of the measures exists and it leads to unreliable data.

$$skewness = \frac{\sum f(x_i - \bar{x})^3}{s^3} \quad (4)$$

Where, 's' is standard deviation, x_i ($i = 1, \ldots, n$) is the univariate data and $\bar{x}$ is the mean of the univariate data. For both positively and negatively skewed data,

$$mean > median > mode \text{ and } mean < median < mode$$

If the data has outliers or highly skewed then median is preferred over mean to handle the missing data, otherwise mean is preferred. If the data is categorical, then mode is used. If $-1 > skewness > 1$, then the data distribution is highly skewed. If $-1 < skewness < -0.5$ or $0.5 < skewness < 1$, then the data distribution is moderately distributed. If $-0.5 < skewness < 0.5$, then the distribution is approximately symmetric. The skewness for exact normal distribution is fixed as zero. Measures of central tendency have been adopted to handle the missing values as per the skewness range. If a variable is normally distributed, then 'mean' is used for imputation. If the variable exhibits skewness, then median is used for imputation for quantitative attributes and for qualitative variables, mode is used for imputation. A sample data instance before and after handling of missing values is shown in the Tables 4 and 5.

Table 4. Sample Data Instance before Imputation

(a) Numerical attributes

Id	age	bp	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	egfr
166	27	60	76	44	4	127	4	NaN	NaN	NaN	NaN	14

(b) Categorical attributes

sg	al	su	rbc	pc	pcc	ba	htn	dm	cad	appet	pe	ane	gender	race
NaN	NaN	NaN	NaN	NaN	Not-present	Not-present	no	no	no	poor	yes	yes	female	other

Table 5. Sample Data Instance after Imputation

(a) Numerical attributes

Id	age	bp	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	egfr
166	27	60	76	44	4	127	4	12.8	41	8000	4.8	14

(b) Categorical attributes

sg	al	su	rbc	pc	pcc	ba	htn	dm	cad	appet	pe	ane	gender	race
1.02	0	0	normal	normal	Not-present	Not-present	no	no	no	poor	yes	yes	female	other

After imputation, the attributes are pulled towards normal distribution to some extent. It leads to less bias and high variance and turns to be an overfitting model. In this work, the issue of overfitting is handled by validation and bagging techniques.

b). Feature selection

Feature selection is a process of evaluating the relationship among the attributes in the enhanced CKD training dataset to handle the issue of underfitting and to improve accuracy by reducing the features in model building. The reduced subset of features thus formed is being utilized in training the hybrid model as well as to be used along with the test dataset to validate the learned model. Unsupervised feature selection process is adopted in which the class attribute is not considered for finding the relationship among other variables and to remove redundant attributes. The relevant features have been identified using Pearson's correlation (r_{xy}) for numerical attributes and Spearman's correlation (ρ) for categorical attributes with threshold > 0.5

and threshold > 0.6 for both cases. These statistical approaches illustrate the degree of correlation between attributes which may be positive, negative or null [10]. The r_{xy} summarizes the strength and direction of the linear association between the numerical attributes as given in Equation (5).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(yi - \bar{y})^2}} \quad (5)$$

The 'ρ' summarizes the relationship among the categorical variables as shown in Equation (6). The training dataset is visualized through scatter plot (Figure 2) and heat maps (Figures 3(a) & 3(b)) to aid in the selection of features appropriately in both correlation analysis respectively.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \quad (6)$$

Where, d_i is the difference between paired ranks. The redundant attributes are removed in this process of statistical analysis by considering the coefficients r_{xy} and ρ. From Figure 2, it is observed that while '*hemo'* increases '*pcv'* also increases as per the enhanced CKD training dataset and these two attributes are highly positively correlated with the strength of 0.83. Since, both are redundant attributes, the attribute '*pcv*' is dropped and '*hemo*' is considered for model building.

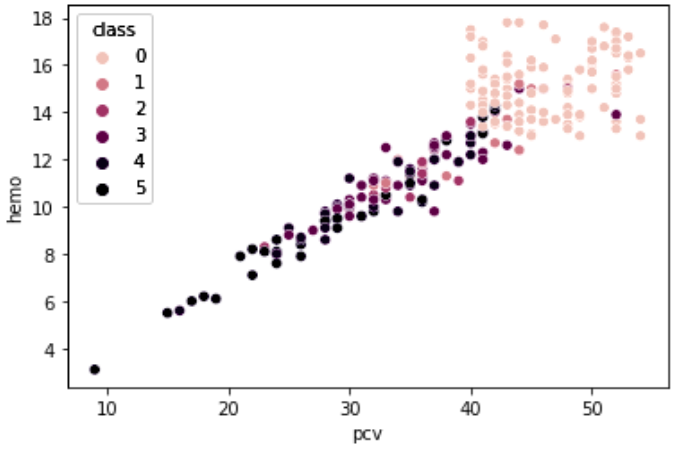

Figure 2. Scatter Plot – Pearson's Correlation between '*pcv*' and '*hemo*'

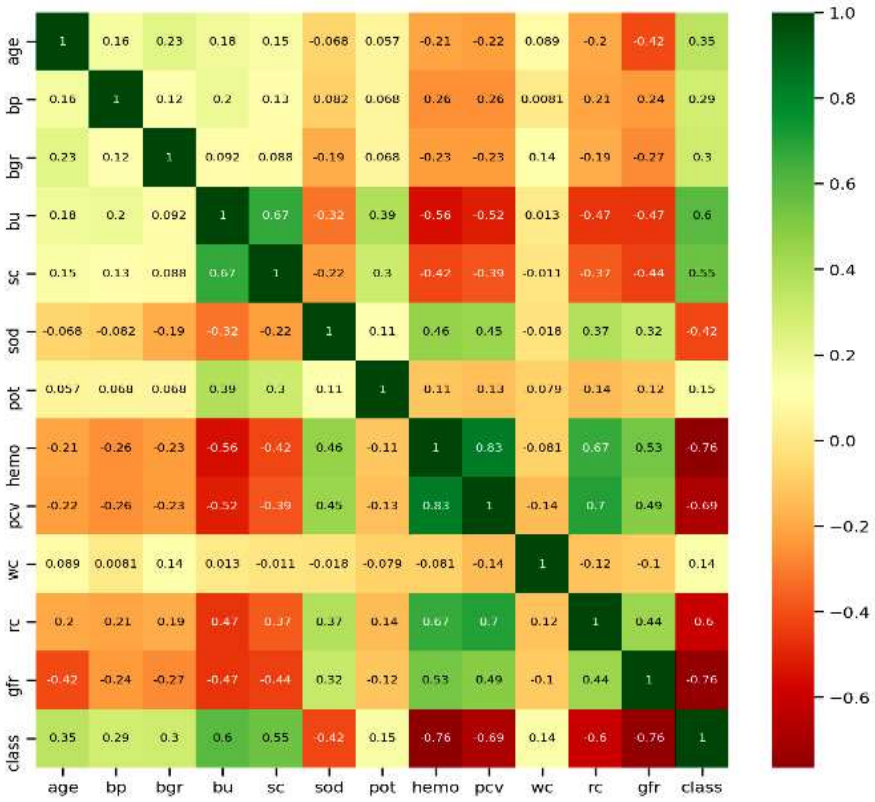

Figure 3 (a). Heat Map –Pearson's Correlation Coefficient Matrix

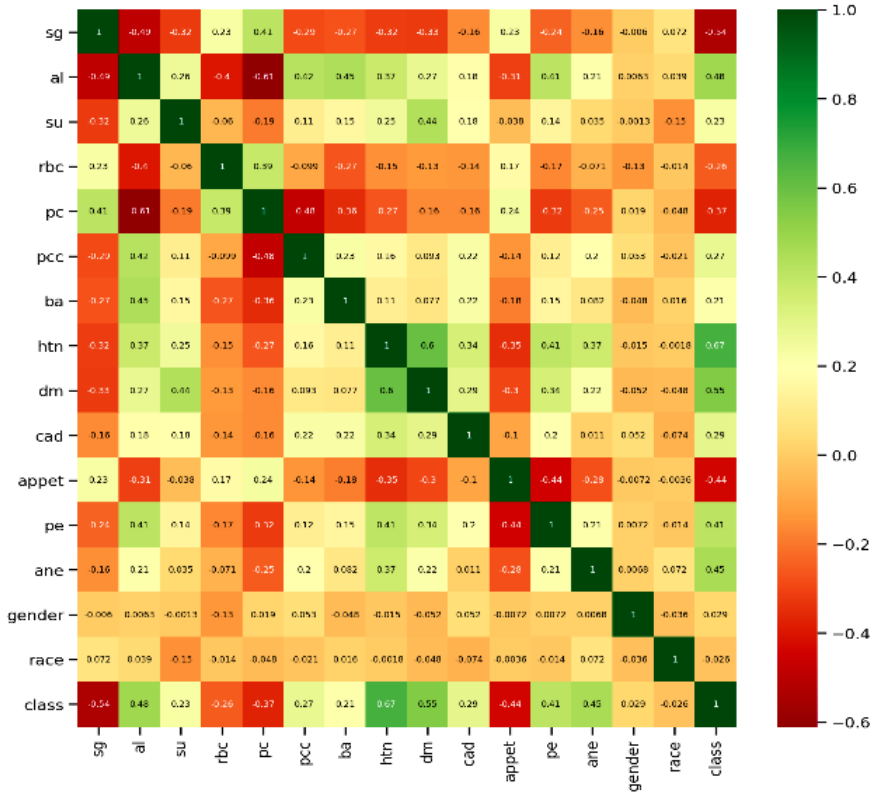

Figure 3 (b). Heat Map - Spearman's Correlation Coefficient Matrix

Many ML algorithms cannot be operated directly on categorical data. Therefore, the categorical attributes are subjected to label encoding which is a binary representation before fitting into a ML model. Another set of features have been selected manually, which is based on medically important features [1], [2]. As per the clinical reports, patients with diabetes, blood pressure and coronary artery diseases are more likely to get CKD. The urine and blood samples are tested to track the presence and the progression of CKD. The amount of albumin is checked in the urine samples and eGFR is calculated from the blood samples. The eGFR is estimated using the attributes *age*, *sc*, *gender* and *race* of the patients. Therefore, the attributes *su*, *bp*, *cad, al, age, sc, gender, race* and *eGFR* are considered as medically important features to be selected for model building.

3.5 Ensemble Classifiers

a.Random Forest Classifier

In this work, RF algorithm is used to identify the stages of CKD by analysing the patient's medical reports especially the laboratory test reports [18]. This algorithm is utilized in the context of classification to predict the desired results as it is a supervised classification algorithm. Overfitting is the major issue while applying traditional statistical models in medical analysis. RF algorithm is not only handling the overfitting issue optimally but also handles the missing values and the categorical values effectively. In general, RF is applied to extract the relevant features from the training dataset, while forming the classifier. Prediction is the next stage of RF algorithm once the classifier is generated. The RF algorithm is applied on the enhanced CKD dataset, from which it generates a number of decision trees using the reduced set of features. The base learner for RF is a DT that would be generated as a parallel ensemble method as shown in Figure 4. The motivation of parallel generation is to reduce the variance by aggregation.

A threshold is fixed while generating the DTs as if the number of trees is more it may slow down the training process and also decrease the performance. At the same time, if more features are considered to increase the depth of the tree, the algorithm faces the challenge of overfitting. RF adopts aggregation principle to overcome the overfitting issue and to improve the prediction accuracy. Also, if more features are considered in each node, the model cannot learn enough for the correct prediction and it is an example of underfitting. The numerical and categorical features are to be handled appropriately to overcome the issues of overfitting and underfitting and hence to improve the performance.

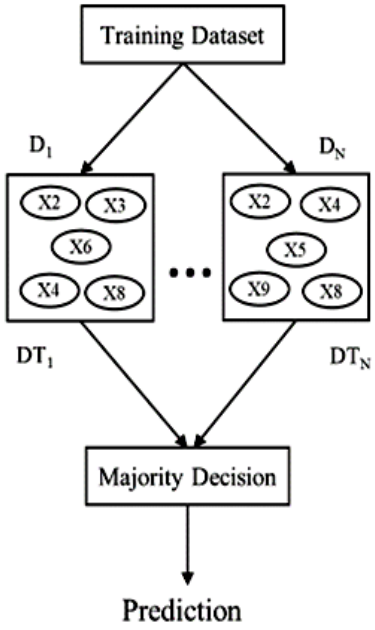

Figure 4. Random Forest

The training dataset consists of '*m*' data instances, {*X1* to *Xm*} with F reduced features. RF collects *N* samples randomly from the training dataset, in which each sample (D_i, $i = 1 \ldots N$) consists of R rows and K reduced features with replacement as given in Equation (7) where, K equals sqrt(F). From each sample, a DT is constructed, which is represented as DT_i, ($i = 1 \ldots N$). The root of each DT is determined by Gini or Entropy equations (Equations 8 & 9).

$$D_N = \begin{bmatrix} X2 \rightarrow bp2\ hemo2\ htn2\ Class2 \\ X4 \rightarrow bp4\ hemo4\ htn4\ Class4 \\ X5 \rightarrow bp5\ hemo5\ htn5\ Class5 \\ X8 \rightarrow bp8\ hemo8\ htn8\ Class8 \\ X9 \rightarrow bp9\ hemo9\ htn9\ Class9 \end{bmatrix} \quad (7)$$

$$Gini = 1 - \sum_{i=1}^{c} (P_i)^2 \quad (8)$$

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (9)$$

Where, '*c*' is the number of classes and p_i is the probability of class c_i. Each DT gives a prediction of $\hat{y}_i$. The algorithm bootstraps and aggregates the results of each DT as given in Equation (10) and hence the high variance is reduced to low variance.

$$prediction\ \hat{y} = \mathrm{mode}(\hat{y}_1 \ldots \hat{y}_N) \quad (10)$$

b).AdaBoost

AB classifier is another ensemble classifier, in which multiple classifier algorithms are grouped and the final prediction is based on the combined effect of all those algorithms. The AB works on iterative principle, for improving the accuracy of prediction in the successive iterations based on the accuracy achieved in the previous training with proper selection of training set. The accuracy is achieved based on the selection of training set and appropriate consideration of weightage for each classifier. In AB, each classifier is learnt sequentially by fitting the appropriate classifiers and analysing the data for errors and hence, the DT's are constructed at every step to improve the accuracy from the previous error. An iterative approach has been adopted to learn from the mistakes of weak classifiers, and turn them into strong ones by adjusting the weights [17]. AB aims to decrease the bias in each and every successive iterations by better modelling. The motivation for serial generation of ensemble methods is to reduce the bias by adjusting the weights of correctly classified and misclassified labels as presented in Figure 5.

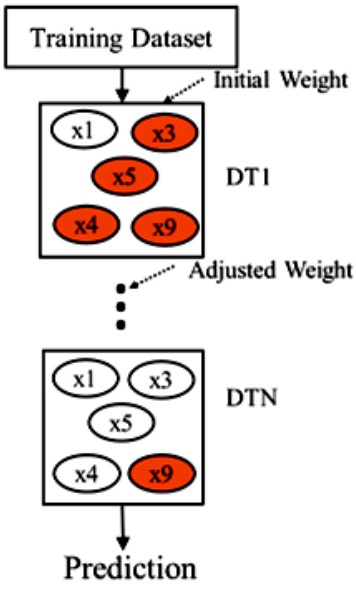

Figure 5. AdaBoost

In this work, stagewise additive modelling with a multiclass exponential loss function (SAMME) is used by the AB classifier for CKD stage prediction [26]. The iterative process continues till the required number of DT's is constructed with adjusted weights using another sub-sample of the dataset where all the misclassified data instances are considered for the prediction.

c. Voting Classifier

The hybrid model is proposed (i.e.)., ensemble is enforced to make prediction based on different classifiers. A Voting Classifier (VC) shall be used to boost the performance of the other

ensemble classifiers to achieve the desired accuracy level of overall classifier. Voting is the principle of grouping by weightage the outcomes of multiple classifiers in the model. In the proposed model, a voting classifier is used to combine both RF and AB classifiers to balance the bias and variance levels to boost the performance. In the voting classifier, hard voting outperforms since it predicts the class with the largest sum of votes from models.

3.6 Hybrid ensemble model

A hybrid ensemble VC combines heterogeneous collection of weak learners to create a single model. Ensemble bagging techniques and cross validations are used to reduce the variance. Since the data is imbalanced, ensemble boosting technique is used while building the model to reduce the bias [17]. The bagging and boosting models are integrated through voting classifier. The intention is to make the hybrid model more flexible i.e., with less bias and less data sensitive i.e., less variance. RF classifier trains the encompassed classifiers in parallel with random subset of data while AB trains the encompassed classifiers in sequential, where each classifier learns by the experience of previous classifier. In machine learning terminology, these ensemble methods are termed as bagging and boosting respectively. In this research work, AB is combined with RF to make a trade-off between overfitting and underfitting issues while training the CKD dataset with reduced set of features, which includes relevant new features to predict the stages of CKD accurately. This proposed hybrid ensemble classification model combines the predictions from bagging and boosting models with k-fold cross validation to correctly classify the new data instance by reducing the bias and variance. Multiple RF and AB classifiers with different parameters are generated and fed to the ensemble voting classifier to predict the class of the new instance with the largest sum of votes from all RF and AB models. The steps followed to predict CKD stages using the hybrid ensemble model are described below. The metrices like CVS, accuracy, precision, recall, F1-measure are considered for performance measurement.

Input: UCI Repository - CKD Dataset

1. Extract the feature *eGFR* using MDRD equation.
2. Apply stratified split in the enhanced CKD Dataset in the ratio of 70:30, in which the 70 percent dataset is used to train the model.
3. Thoroughly examine the training data to identify trends, patterns, and relationships.
    a. Handle missing data using measures of central tendency.

b. Select three sets of features by applying Pearson's correlation coefficient for numerical data and Spearman's correlation for categorical data with different thresholds and features based on medical importance.
c. Apply label encoding to the categorical variables.

4. Apply the steps 3a and 3c to the 30 percent test dataset separately with the reduced feature subsets towards validation of the learned model.
5. Normalize the training dataset as well as the test dataset.
6. Build a RF classifier to predict the stages of CKD or notCKD
7. Build AB classifier to predict the stages of CKD or notCKD
8. Generate five RF models with different parameter values.
    i. M1: bootstrap=True, max_features =" auto", criterion =" gini", n_estimators=50, random_state=0
    ii. M2: bootstrap=False, max_features =" sqrt", criterion =" entropy", n_estimators=100, random_state=0
    iii. M3: bootstrap=False, max_features =" auto", criterion =" entropy", n_estimators=100, random_state=0
    iv. M4: bootstrap=True, max_features =" sqrt", criterion =" entropy", n_estimators=50, random_state=0
    v. M5: bootstrap=True, max_features ="log2", criterion ="gini", n_estimators=100, random_state=0
9. Generate five AB models with different parameter values
    i. M6: DecisionTreeClassifier (max_depth=4), n_estimators =100, algorithm =" SAMME"
    ii. M7: DecisionTreeClassifier (max_depth=3), n_estimators =100, algorithm =" SAMME"
    iii. M8: DecisionTreeClassifier (max_depth=2), n_estimators =100, algorithm =" SAMME"
    iv. M9: DecisionTreeClassifier (max_depth=2), n_estimators = 200, algorithm =" SAMME"
    v. M10: DecisionTreeClassifier (max_depth=3), n_estimators = 200, algorithm =" SAMME"
10. Apply cross fold validation with splits as 10 to each model in step 6 to step 9.
11. Build a hybrid VC with the generated two models in step 8 and step 9 with k-

fold=10.

12. Validate the RF, AB and the hybrid models with the test dataset.
13. Evaluate the performance of the model using the MSE error, bias, variance, cross validation score, accuracy, precision, recall and F1 measure in predicting the stages of CKD or notCKD.

Output: Prediction of stages of CKD

The proposed hybrid ensemble voting classifier yields better accuracy in all three reduced feature sets with less MSE, low bias and variance when compared with the RF and AB.

4. Results and Discussion

4.1 Experimental Setting

The experimental setup requires Python 3.8 and packages scikit for machine learning, sklearn, numpy, pandas, matplotlib, seaborn for data analysis and visualization.

4.2 Experimental Evaluation

In this work, the feature *eGFR* is extracted with existing features; *age*, *sc* and populated features; *gender* and *race* as required by the MDRD equation and the instances are labelled as per KDIGO guidelines in accordance with *eGFR* value. In the enhanced CKD dataset, the class distribution is as follows: “notckd” instances are 38 %, CKD - stage1 instances are 6 %, stage2 instances are 5 %, stage3 instances are 19 %, stage4 instances are 14 % and stage5 instances are 18% respectively. The entire dataset is applied with stratified split and hence, the same percentage of classes is maintained in the training as well as test datasets.

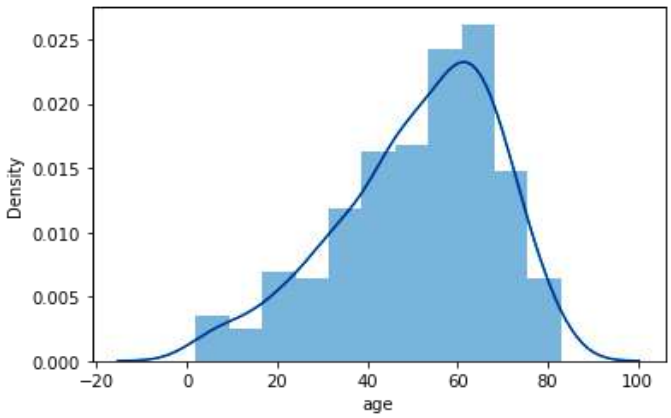

Figure 6. Actual distribution of age

The data pre-processing (i.e.)., filling missing values and feature selection is done separately in the training dataset to prevent data leakage, where the knowledge of test dataset does not leak

into the training dataset and vice versa. The missing values are handled in the test dataset separately. The same features which have been selected in the training dataset are being considered in the test dataset. This results in correct estimation of the model's performance when an unseen data is tested for predictions. The univariate analysis is done on the training dataset to know the data distribution, and for the attribute "age" the distribution is presented in Figure 6. Moreover, sample training data instances with skewness values are shown in Table 6.

Table 6. Skewness of the Numerical Attributes

Sl.No.	Attribute	Skewness
1	*age*	-0.654868
2	*bp*	2.054112
3	*bgr*	1.813749
4	*bu*	2.971817
5	*sc*	6.135109
6	*sod*	-0.613384
7	*pot*	9.896213
8	*hemo*	-0.391718
9	*pcv*	-0.536475
10	*wc*	1.786081
11	*rc*	-0.186372
12	*eGFR*	1.410340

The table clearly shows that all the numerical attributes in the training dataset exhibit skewness towards either left or right. Therefore, the mean of each numerical attribute is calculated and presented in Table 7. The missing values in numerical attributes are replaced with the mean of the corresponding variables.

Table 7. Mean Value for Numerical Attributes

Sl.No.	Numerical Attributes	Mean
1	*age*	55
2	*bp*	70
3	*bgr*	119.5
4	*bu*	42
5	*sc*	1.3
6	*sod*	138
7	*pot*	4.4
8	*hemo*	12.8
9	*pcv*	41
10	*wc*	8000
11	*rc*	4.8
12	*gfr*	49

The mode for each categorical attribute is calculated and shown in Table 8. The missing values for the categorical attributes are filled with the corresponding mode.

Table 8. Mode of Categorical Attributes

Sl.No.	Categorical Attributes	Mode
1	*sg*	1.02
2	*al*	0
3	*su*	0
4	*rbc*	normal
5	*pc*	normal
6	*pcc*	Not present
7	*ba*	Not present
8	*htn*	no
9	*dm*	no
10	*cad*	no
11	*appet*	good
12	*pe*	no
13	*ane*	no
14	*gender*	female
15	*race*	other

After imputation the attributes are pulled towards normal distribution to some extent. This leads to less bias and high variance. The change in the distribution of the variable *wc* (i.e.)., white blood cells count before and after imputation is shown in Figure 7.

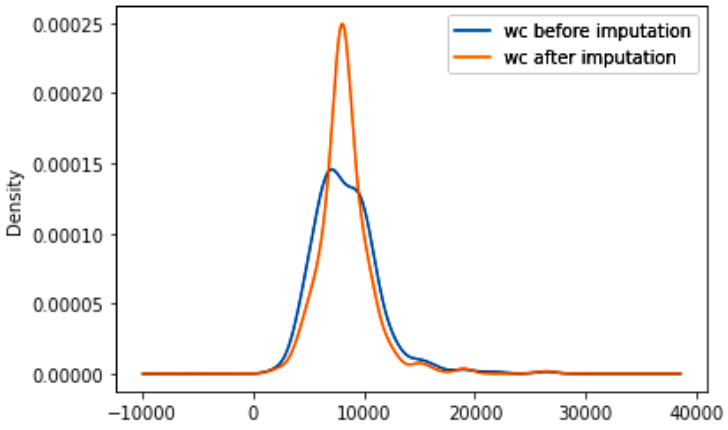

Figure 7. *wc*'s Distribution Before and After Imputation

The Pearson's and Spearman's correlation matrices are used to determine the relationship among quantitative and qualitative attributes to select the features for model building by dropping

the redundant features. The features selected with medical importance and with Pearson and Spearman correlation coefficients are shown in Table 9.

Table 9. Reduced feature set

Sl. No	Type	Correlation Coefficient	Threshold	Selected features
I	Unsupervised	Pearson and Spearman	> 0.5	*age, bp, sg, su, rbc, pcc, ba, bgr, bu, sod, pot, wc, htn, cad, appet, pe, ane, gender, race*
II		Pearson and Spearman	> 0.6	*age, bp, sg, al, su, rbc, pcc, ba, bgr, bu, sod, pot, hemo, wc, htn, dm, cad, appet, pe, ane, gender, race, gfr*
III	Medically important	Clinical factors: Blood and Urine test		*age, bp, al, su, sc, cad, gender, race, gfr*

The hybrid model is built with these reduced set of features presented in Table 9. RF, a bagging technique is used to reduce the variance. As the enhanced dataset itself imbalanced, the bias is found to be invariably high. Subsequently, AB boosting technique is used to reduce the bias.

RF and AB classifier models with different parameters are generated and k-fold cross validation is applied to each model for testing and to verify how accurately a new unseen data is classified. With the selected features subset I, the AdaBoost gives an accuracy of 94.16 % with MSE 0.052 with less bias and variance of values 0.009 and 0.043 respectively. The RF gives an accuracy of 91.7 % with MSE of 0.096 with low variance about 0.058 and a bias of 0.037 respectively. Similarly, the bias, variance and MSE values with the selected feature subsets II and III are estimated along with other performance metrices and the results are summarized in Table 10.

Table 10. Metric comparisons on models with different set of features

	Reduced Features Subset I			Reduced Features Subset II			Reduced features Subset III		
Metrics	**Random Forests**	**Ada Boost**	**Hybrid Model**	**Random Forests**	**Ada Boost**	**Hybrid Model**	**Random Forests**	**Ada Boost**	**Hybrid Model**
MSE	0.096	0.052	0.015	0.030	0.060	0.004	0.006	0.081	0.005
Bias	0.037	0.009	0.002	0.008	0.015	0.000	0.002	0.004	0.002
Variance	0.058	0.043	0.013	0.023	0.045	0.004	0.004	0.076	0.003
Cross Validation Score (%)	91.7	95	97.85	98.21	94.64	99.28	99.28	94.64	99.64
Accuracy (%)	94.16	94.16	99.16	99.16	94.16	100	99.16	94.16	100
Precision (%)	92.41	75	99.30	99.30	75	100	97.91	75	100
Recall (%)	88.04	83.33	97.61	97.61	83.33	100	97.61	83.33	100
F1 measure (%)	89.80	77.77	98.36	98.36	77.77	100	97.60	77.77	100

The two heterogeneous models are combined together as a HEM through a voting classifier to improve the accuracy with low bias and variance trade-off. This hybrid model yields 100 % accuracy in predicting the severity of the CKD stages for the feature subsets II and III and the other estimated performance metrices are shown in Table 9. The error rates obtained using all the subsets are represented in Figure 8.

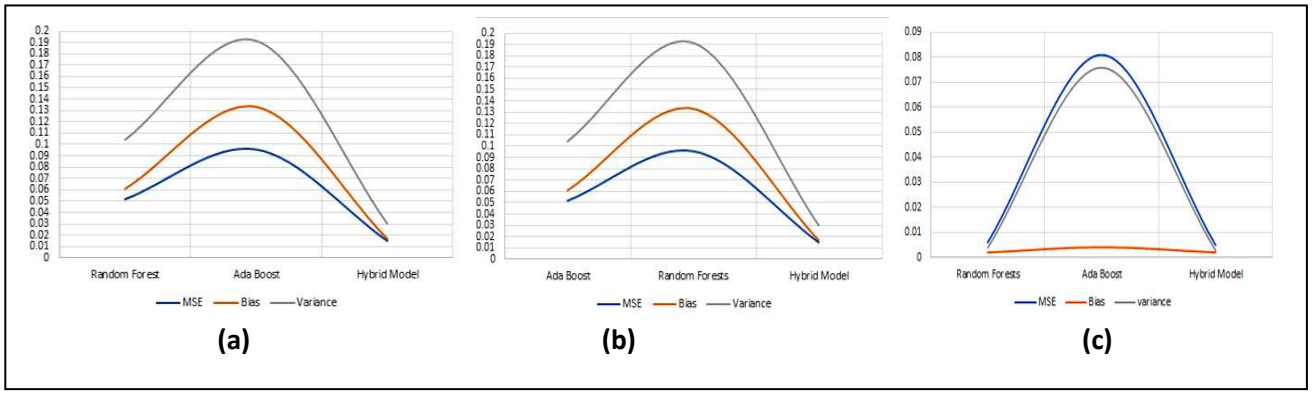

Figure 8. Error rate (a) Reduced Features Subset I (b) Reduced Features Subset II (c) Reduced Feature Subset III

The models are estimated using the mean of the cross-validation scores, where the proposed HEM predicts the stages of CKD with an accuracy of 97.85 % with reduced features subset I, 99.28% with reduced features subset II and 99.64 % with reduced features subset III 99.64 % respectively. While comparing the Precision and Recall, it is evident that the occurrence of False Negative and False Positive errors are nil and the F1 measure is well balanced in the proposed HEM with respect to the reduced feature subsets II and III.

5. Conclusions

For the accurate prediction of CKD stages through ML, the issue of data leakage is eliminated during the data pre-processing stage itself and the HEM classifier is made to learn with low bias and variance. The bagging approach decreases the variance and the boosting approach reduces the bias. The reduced set of features are obtained through correlation analysis and another set of features gathered by manual selection, which are medically important for the CKD analysis. It has been given independently as input to the HEM classifiers, thus enabling the model to learn for the accurate prediction of stages of CKD. The outcomes of the classifiers have been validated using the test dataset and the learning process of the hybrid model controls the False Negative and False Positive errors thereby the model detects as many patients with CKD stages as possible. Since, the data leakage issue is handled appropriately, the proposed model will have the same accuracy and variance in the production environment. Further, the collection of more samples with timely clinical reports would lead to better prediction accuracy as well.

Declarations:

- There is no conflict of interest between the authors(s).
- The software code/ data will be provided once the formal acceptance is received.
- No funding received for the proposed research work.
- This work does not perform any direct research on human/organisms and hence ethical clearance not required. Only the public dataset is used for research.

Contributions:

P.Antony Seba- Modelling and Evaluation

Bibal Benifa- Design and Manuscript Preparation

V Ramachandran- Data Collection

References

1. Robert Thomas, Abbas Kanso, John R. Sedor, "Chronic Kidney Disease and Its Complications", Primary Care: Clinics in Office Practice, Vol. 15, No 2, pp. 329-344, https://doi.org/10.1016/j.pop.2008.01.008., May 2008.
2. Himanshu Kriplani, Bhumi Patel and Sudipta Roy, "Prediction of Chronic Kidney Diseases Using Deep Artificial Neural Network Technique", Computer Aided Intervention and Diagnostics in Clinical and Medical Images, Lecture Notes in Computational Vision and Biomechanics, Vol. 31, pp. 179-187, Springer, https://doi.org/10.1007/978-3-030-04061-1_18, January 2019.
3. Hsueh-Lu Chang, Chia-Chao Wu, Shu-Pei Lee, Ying-Kai Chen, Wen Su and Sui-Lung Su, "A predictive model for progression of CKD", Medicine, Vol. 98, No. 26, doi:10.1097/MD.0000000000016186, June 2019.
4. https://kdigo.org/wp-content/uploads/2018/08/KDIGO-Txp-Candidate-GL-Public-Review-Draft-Oct-22.pdf
5. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease (accessed on 14th Februry, 2020).
6. El-Houssainy A. Rady and Ayman S. Anwar, "Prediction of Kidney Disease Stages using Data Mining Algorithms", Open Access Journal on Informatics in Medicine Unlocked, https://doi.org/10.1016/j.imu.2019.100178, Vol.15, April 2019.
7. Mehdi Hosseinzadeh, Jalil Koohpayehzadeh, Ahmed Omar Bali, Parvaneh Asghari, AlirezaSouri, Ali Mazaherinezhad, Mahdi Bohlouliand Reza Rawassizadeh, "A Diagnostic Prediction Model for Chronic Kidney Disease in Internet of Things Platform", Journal of Multimedia Tools and Applications, Springer Science Business Media, https://doi.org/10.1007/s11042-020-09049-4, July 2020.
8. Samina Khalid, Tehima Khalil, Shamila Nasreen, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning", 2014 science and information conference, IEEE, pp. 372-378, August 2014.
9. Saravanan N, Sathish G, Balajee J M, "Data Wrangling and Data Leakage in Machine Learning for Healthcare", Journal of Emerging Technologies and Innovative Research, Vol. 5, No. 8, pp. 553-557, August 2018.
10. Sunil Kumar and IIyoung Chong, "Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States", International

Journal of Environmental Research and Public Health, Vol. 15, No. 12, https://doi.org/10.3390/ijerph15122907, December 2018.

11. Shaoze Cui, Dujuan Wang, Yanzhang Wang, Pay-Wen Yu, and Yaochu. Jin, “An improved support vector machine-based diabetic readmission prediction”, Computer Methods and Programs in Biomedicine, Vol. 166, pp. 123-135, https://doi.org/10.1016/j.cmpb.2018.10.012, November 2018.
12. Saruar Alam, Goo-Rak Kwon, Ji-In Kim, and Chun-Su Park, “Twin SVM-Based Classification of Alzheimer's Disease Using Complex Dual-Tree Wavelet Principal Coefficients and LDA”, Journal of Healthcare Engineering, Volume 2017, Article ID 8750506, 12 pages, https://doi.org/10.1155/2017/8750506, August 2017
13. Shrinivas D Dessai, Indira Fal Dessai and Linganagouda Kulkarni, “Intelligent Heart Disease Prediction System Using Probabilistic Neural Network”, International Journal of Advanced Trends in Computer Science and Engineering, Vol. 2, No. 5, pp. 22-28, September 2013.
14. Jamshid Norouzi, Ali Yadollahpour, Seyed Ahmad Mirbagheri, Mitra Mahdavi Mazdeh, and Seyed Ahmad Hosseini, “Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System”, Computational and Mathematical Methods in Medicine, Volume 2016, Article ID 6080814, 9 pages, https://doi.org/10.1155/2016/6080814, February 2016.
15. Srinivasa R. Raghavan, Vladimir Ladik, and Klemens B. Meyer, “Developing Decision Support for Dialysis Treatment of Chronic Kidney Failure,” IEEE Transactions on Information Technology in Biomedicine, Vol. 9, No. 2, pp. 229-238, doi: 10.1109/TITB.2005.847133, June 2005.
16. Jason Roy, Haochang Shou, Dawei Xie, Jesse Y. Hsu, Wei Yang, Amanda H. Anderson, J. Richard Landis, Christopher Jepson, Jiang He, Kathleen D. Liu, Chi-yuan Hsu and Harold I. Feldman, “Statistical Methods for Cohort Studies of CKD: Prediction Modeling”, Clinical Journal of the American Society of Nephrology, Vol. 12, No. 6, pp. 1010-1017, doi**:** 10.2215/CJN.06210616, June 2017.
17. JafarTanha, Yousef Abdi, Negin Samadi, Nazila Razzaghi and Mohammad Asadpour, “Boosting methods for multi-class imbalanced data classification: an experimental review”, Journal of Big Data, Vol.7, No.1, pp. 1-47, https://doi.org/10.1186/s40537-020-00349-y, September 2020.

18. Xin Han, Xiaonan Zheng, Ying Wang, Xiaoru Sun, Yi Xiao, Yi Tang and Wei Qin, “Random forest can accurately predict the development of end-stage renal disease in immunoglobulin a nephropathy patients”, Annals of Translational Medicine, Vol. 7, No. 11, doi: 10.21037/atm.2018.12.11, June 2019.
19. https://www.kidney.org/professionals/kdoqi/gfr_calculator
20. Mohamed Elhoseny, K. Shankar and J. Uthayakumar, “Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease”, Scientific Reports, Open Access Journal, Springer Nature, July 2019.
21. Gabriel R. Vasquez-Morales, Sergio M. Martinez-Monterrubio, Pablo Moreno-Ger and Juan A. Recio-Garcia, “Explainable Prediction of Chronic Renal disease in the Colombian Population Using Neural Networks and Case-Based Reasoning”, IEEE Access, Special Section on Data-enabled Intelligence for Digital Health, Vol. 7, pp. 152900-152910, October 2019.
22. Bilal Khan, Rashid Naseem, Fazal Muhammad, Ghulam Abbas and Sunghwan Kim, “An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy”, IEEE Access, Vol. 8, pp. 55012-55022, March 2020.
23. Jiongming Quin, Lin Chen, Yuhua Liu, Chuanjun Liu, Changhao Feng and Bin Chen, “A Machine Learning Methodology for Diagnosing Chronic Kidney Disease”, IEEE Access, doi: 10.1109/ACCESS.2019.2963053, February 2020.
24. Surya Krishnamurthy, Kapeleshh K S, Erik Dovgan, Mitja Luštrek, Barbara GradišekPiletič, Kathiravan Srinivasan, Yu-Chuan Li, Anton Gradišek and Shabbir Syed-Abdul, “Machine Learning Prediction Models for Chronic Kidney Disease using National Health Insurance Claim Data in Taiwan”, medRxiv, The Preprint Server for Health Sciences, Cold Spring Harbour Laboratory, https://doi.org/10.1101/2020.06.25.20139147, July 2020.
25. Anandanadarajah Nishanth and Tharmarajah Thiruvaran, “Identifying important attributes for early detection of Chronic Kidney Disease”, IEEE Reviews in Biomedical Engineering, Vol. 11, pp. 208-216, doi: 10.1109/RBME.2017.2787480, December 2017.
26. Ji Zhu, HuiZou, SaharonRosset and Trevor Hastie,” Multi-class AdaBoost”, Statistics and its Interface, Vol. 2, No 3, pp. 349-360, https://dx.doi.org/10.4310/SII.2009.v2.n3.a8, January 2009.

****