IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|

Chronic Kidney Disease Prediction Using MachineLearning

Parimal Meshram, Mohammad Umair, Raashid Umar, Paritosh Vaidya, Prof. A. U. Chaudhari

Under Graduate Student, Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology and Research, Amravati., India

Under Graduate Student, Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology and Research, Amravati., India

Under Graduate Student, Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology and Research, Amravati., India

Under Graduate Student, Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology and Research, Amravati., India

Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology & Research, Amravati, Maharashtra, India.

ABSTRACT: Chronic renal disorder (CKD) or chronic renal disease has become a serious issue with a gentle rate. A person can only live without kidneys for a median time of 18 days, which makes a large demand for a kidney transplant and Dialysis. it's important to own effective methods for early prediction of CKD. Machine learning methods are productive in CKD prediction. This work proposes a workflow to predict CKD status supported clinical data, incorporating data prepossessing, a missing value handling method with collaborative filtering and attributes selection. Out of the 11 machine learning methods considered, the additional tree classifier and random forest classifier are shown to lead to the best accuracy and minimal bias to the attributes. The research also considers the sensible aspects of information collection and highlights the importance of incorporating domain knowledge when using machine learning for CKD status prediction.

KEYWORDS: Chronic Kidney Disease, Chronic Renal Disease, Machine Learning, Classification Algorithms, Extra Tree Classifier, Random Forest Classifier.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a serious medical problem in India, and 1 in 10 suffers from some form of kidney disease. About 1,75,000 new cases of kidney failure are introduced into the Asian nation each year, critical enough to perform dialysis. In total, 850 million people are currently expected to develop kidney infection due to various causes and ongoing kidney disease. The world records 2.4 million deaths each year and is currently the sixth leading cause of death. The kidneys are one of the most important organs in the lower back, one kidney arranged on one side or the other. The main function of the kidneys is to filter the blood and remove waste products from the body in the form of urine. The kidneys fail to function properly when they cannot filter out waste products. High blood pressure, obesity and diabetes are important components of Chronic Kidney Disease (CKD). Kidney infection is often referred to as a silent condition. Especially when the patient begins to feel symptoms such as weakness, shortness of breath etc. means that its kidney function has recently dropped to 25 percent or less. Chronic Kidney Disease does not show any symptoms in the first stage and most cases are left untreated until they reach the advanced stage. This leads to delays in treatment of the patient which can be dangerous to health. Early diagnosis is very important to reduce the risk and the patient is given appropriate treatment early. The use of Artificial Intelligence in the field of medical research is increasing day by day. This can contribute significantly to the development of diagnostic and diagnostic systems used in the prediction of diseases that can provide information that directs medical professionals to the early detection of deadly diseases and thus, increase the patient's endurance rate as a whole. Previous diagnostic tests for Chronic Kidney Disease were used in each category and evaluated based on a single mathematical accuracy. This can lead to a reduction in the overall performance of the model.

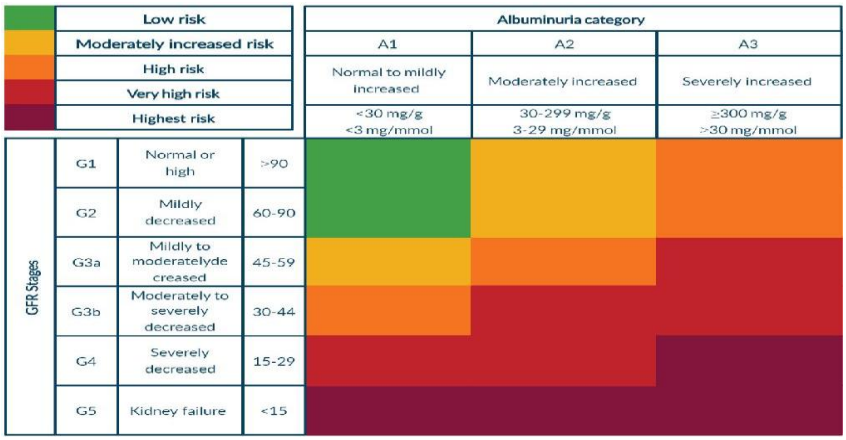
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|

They are located just beneath the rib cage,one on each side of the spinal cord. Daily, the kidneys filter about 120 to 150 quarts of blood to produce about 1 to 2quarts of urine. The basic function of the kidneys is to remove waste products and excess fluid from the body through urine. The production of urine calls for highly complex steps of excretion and re-absorption. The kidney functions normal in stage 1 and is minimally reduced in stage 2 but the majority of casesare at stage 3 (see Fig. 1).

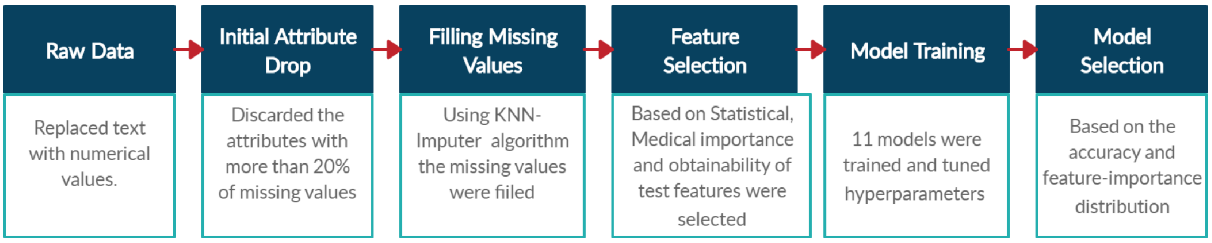


(Fig. 1Heat map of the severity.)

The use of machine learning methods to predict CKD has been evaluated based on multiple data sets. Among them, a database from the UCI repository [7] (referred to as the UCI data set thereafter) is identified as a benchmark database. As with most related work, this function looks at the specified benchmark database.

When analyzing CKD-related clinical data, if there are cases with missing attributes then the method of managing missing values should be determined based on the random method of the missed method. In addition, the UCI data set [7] has 400 scenarios which is a small amount compared to the 25 seam sample.

As such, this function demonstrates limitations in handling missing values when analyzing CKD data, suggests a new way of handling missing values and introduces different methods based on UCI data. In addition, this work also highlights the importance of statistical analysis and background information of features when making predictions based on clinical data related to CKD.



(Fig. 2. Proposed workflow.)

II. RELATED WORK

A. J. Hussain and team found 0.995 accuracy in predicting CKD in the early stages using multi-layer visibility while including pre-processing of data set by sensory networks to supplement missing values. Work flow involves discarding external factors, selecting seven relevant attributes in mathematical analysis, and disposing of high-correlated attributes in key component analysis (PCA).[8] In the said task, the algorithm for filling in the missing value has a significant impact on the accuracy of the trained models.

However, due to the use of the Neural Network of only 20 attributes with 260 fully completed data conditions, the accuracy of predicting shortages is reduced slightly.[8] The slightly lower accuracy of the model results in the use of fixed conditions instead of missing values.In our work, however, it is shown that random random data loss is complete

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|

according to the Little's MCAR algorithm (see Methodology section). Moreover, in , when considering the factors, the significance of the factor is dependent on serum creatinine.

In 2017 a team of researchers used 14 factors to predict CKD and gained 0.991 accuracy with a multi-phase decision resolution. They discarded low-value scenarios and trained the neural network and the asset retrieval model which provided 0.975 and 0.960 accuracy in total respectively. The interaction of the selected attributes varies in frequency [0.2 to 0.8]. Considering the therapeutic point of view, high blood pressure can cause CKD and CKD can cause high BP.

In 2015 Lambodar J. and Narendra Ku. K. tested 8 machine learning models using the WEKA data mining tool.[9] High Receiver Function (ROC) and accuracy are provided by Naive Bayes, Multi-layer Perception and J48 algorithms as ROC 1 and accuracy of 0.950, 0.9975 and 0.99 respectively. In the said function, Kappa Statistics is used to determine the strength of the argument and provides a maximum of 0.9947 perceptron for multiple components, 0.9786 as the top sequence of the decision table and J48 algorithms.

Considering the related function based on the UCI CKD data set [7], it was noted that the reasons for the majority were less accurate misalignment of missing values and method of selection of attributes

III. MATERIALS AND METHODS

The scope of this paper is to make use of Machine Learning ensemble algorithms to predict chronic kidney disease. The data set used for building the model is taken from the UCI repository. The dataset involves information of 400 patients with 25 attributes including the class. The dataset consists of data collected from blood test and urine test and also some of the general information such as age, appetite. Out of the 400 patients, 250 patients were diagnosed by CKD and 150 patients were healthy. The dataset is divided into training set and testing set. The Training set is utilized to prepare the model with different Machine Learning ensemble algorithms. The hyper parameters of each of the ensemble classifiers are tuned to get the best parameters that will provide the best model for predicting the chronic kidney disease in patient. The trained model is then used on the testing dataset. The model is assessed based on the performance of each model in terms of accuracy, sensitivity, specificity, precision, F-score, ROCAUC and Mathew Correlation Coefficient.

A. Missing Values

CKD dataset available in the UCI Repository is raw and needs some data preprocessing techniques before applying it to the model. The CKD dataset consist of missing values in many of the features. Figure 1 shows the count of missing values in some of the attribute. All the missing values are replaced by mean for numerical attributes and by mode for categorical attributes.

B. Feature Scaling

The dataset consists of attributes having different range. Such data cannot be applied to the machine learning model. Therefore, it requires rescaling which ensures that all the features fall under the same scale. Minmax scaling technique is used to scale the attributes in the range 0 and 1.

C. Training and Testing Dataset

The dataset is split into training dataset of 70% which includes 280 patient details and 30% testing dataset which includes 120 patient details. The 70% of the training set is further split into training set and validation set using the cross-validation technique A 10-fold cross-validation is used, which splits the training set into 10 folds. In each fold, one cluster is held as validation data and the remaining nine groups are used to train the model. For each fold the evaluation score is retained. Finally, the mean of all the evaluation score for the 10-fold cross validation is calculated.

D. Classifiers

Once the preliminary data processing has been completed the next step is to train the algorithm for a separate set of machine learning using the training and model database in the database to test and evaluate the model performance.

Bagging

Bootstrap Aggregation ordinarily known as Bagging is a sort of ensemble algorithm that arbitrarily picks some instances from the training set with substitution. In Bagging, bootstrap samples are obtained from the training dataset collection and the classifier is set up with every model. The outcome from each classifier is

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|

consolidated, and the final result is obtained from the process of majority voting. Examination shows that bagging can be used to upgrade the overall performance of a weak classifier preferably.

AdaBoost

Adaptive Boosting commonly referred to as Ada Boost is another stage of integration. A common problem with machine learning systems can be reduced using Ada Boost. Ada Boost works by selecting basic class dividers and enhances its performance by identifying non-differentiated cases in training repositories in a repetitive manner. Age of equal weight is given for all training samples and the weakest category is selected. The cycle is repeated n times, each time using the basic phase in a training set with updated weights. In the latest model, the proposed approach combines the outcomes of each weak category either by a majority vote or in the middle [7].

Random Forest

Random Forest helps in clinical applications for better accuracy by combining a group of weak classifiers like Decision Tree. It produces N number of Decision trees by using randomly picked attributes as their information. In Random Forest, the bias is not changed, but the number of trees increases. The outcomes from all the trees can be picked by casting a vote or averaging [7].

Gradient Boosting

As opposed to Random Forest, this model continuously creates decision trees using gradient decent to minimize the loss function. A final forecast is made using a heavy dominant part vote of the whole decision trees. Gradient boosting invalidates the over-fitting issue and manages the bias.

Performance Metrics

In order to estimate the performance of chronic kidney disease model using machine learning some of the performance metrics are utilized from the confusion table. Table 1 shows the Confusion matrix with Accuracy, Positive predicted value and negative predicted value.

Table I Confusion matrix for CKD

Confusion Matrix		**Actual Values**		**Accuracy**= (TP+TN)/(TP+FP+FN+TN)	
		CKD =1	No CKD= 0		
Observed Values	CKD =1	TP	FP	Positive Predictive Value	TP/(TP+FP)
	No CKD =0	FN	TN	Negative Predictive value	TN/(FN+TN)

True Positive (TP) = Samples correctly predicted as having CKD.

False Positive (FP) = Samples falsely predicted as having CKD.

False Negative (FN) = Samples Falsely predicted as not having CKD.

True Negative (TN) = Samples correctly predicted as not having CKD.

IV. RESULTS AND DISCUSSION

In this study, the data used was the Chronic Kidney Disease Dataset located at UCI. The database contains 400 records of which 250 records have CKD and 150 records do not have CKD. Non-database values are managed using the numerical value system and the category pricing mode. Coded labels using the encoding

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|

method. The Min-Max normalizing technique was used to measure all attributes in scales 0 and 1.The previously processed database was divided into a training and assessment data set. The training set has 280 records and the test set has 120 records. Four algorithms such as Bagging, Gradient Descent, Random Forest and Ada Boost were used. The performance of each model was evaluated using different metrics such as accuracy, recall, accuracy, accuracy, f1-score, MCC and ROC-AUC curve. Table II shows the confusion matrix of the various integration algorithms in the experimental data.

Confusion Matrix of Ensemble Classifiers

		Predicted Value				
		BAGGING			ADABOOST	
		No CKD	CKD		NO CKD	CKD
Actual value	NO CKD	36	1		37	0
	CKD	0	83		0	83
	NO CKD	37	0		37	0
	CKD	0	83		2	81
RANDOM FOREST					GRADIENT BOOSTING	

As shown in table II, bagging classifier wrongly classified 1 patient as having NoCKD. Gradient Boosting Classifier falsely classified 2 patients as having CKD. AdaBoost and Random Forest perfectly classified all the patients. Table III shows the performance of the classifier based on Accuracy, Sensitivity, Specificity and Precision. Figure 2 shows the performance of various classifiers.

Table III Accuracy, Sensitivity, Specificity and Precision of Various Ensemble Classifiers

	Accuracy	Sensitivity	Specificity	Precision
Bagging	0.991666	1	0.988095	0.972973
AdaBoost	1	1	1	1
Gradient Boosting	0.983333	0.9759	1	1
Random Forest	1	1	1	1

Table IV Performance of Ensemble Classifiers based on F1-Score, AUC and MCC

	F1- Score	AUC	MCC
Bagging	99.4	98.6	98.05
AdaBoost	100	100	100
Gradient Boosting	98.78	98.8	96.22
Random Forest	100	100	100

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|

The proposed methodology consists of 3 key steps: Data preprocessing, models training and model selection (Fig. 2).

A. Data Preprocessing: Missing Value Handling

In this the data processing was performed in 2 steps. First, attributes with more than 20% data with missing values were filtered (see Table 1). Therefore, a set of factors, (red blood cells, sodium, potassium, white blood cell count, red blood cell count) were not included in the analysis. The second step in pre-processing the data was to manage the non-data values remaining.

In the pre-processing step, shortages should be addressed based on their distribution in order to achieve optimal accuracy. In this work, in order to confirm the non-existent values, a Little's MCAR test was performed. The potential bias for missing data depends on the method that makes the data non-existent. The analytical methods used to adjust the deficit are evaluated using the chi-square test of MCAR with multivariate quantitative data. It checks if there is a significant difference between the patterns of the missing value.

TESTS FOR MEASURING MULTIPLE ATTRIBUTES AND MISSING VALUE PERCENTAGE

Attribute	Missing Percentage	Test to Obtain
Class	0.00 %	
Appetite	0.25%	Doctor's Inspection
Pedal Edema	0.25 %	Doctor's Inspection
Anemia	0.25 %	FBC
Hypertension	0.50 %	Doctor's Inspection
Diabetes Mellitus	0.50%	FBC
Coronary Artery Disease	0.50 %	Doctor's Inspection
Pus Cell Clumps	1.00 %	UFR
Bacteria	1.00 %	UFR
Age	2.25%	Doctor's Inspection
Blood pressure	3.00 %	Doctor's Inspection
Serum creatinine	4.25 %	SERUM CREATININE
Blood Urea	4.75 %	BLOOD UREA
Blood Glucose Random	11.00 %	RBS
Albumin	11.50 %	UFR
Specific Gravity	11.75 %	UFR
Sugar	12.25 %	UFR
Hemoglobin	13.00 %	FBC
Pus Cell	16.25 %	UFR
Packed Cell Volume	17.50 %	FBC
Sodium	21.75 %	SERUM ELECTROIDES
Potassium	22.00 %	SERUM ELECTROIDES
White Blood Cell Count	26.25%	FBC
Red Blood Cell Count	32.50 %	FBC
Red Blood Cells	38.00 %	UFR

Note: Text in *italics* corresponds to attributes excluded from the analysis

Based on Little's MCAR test results, as shown in Table II, the missing value is considered to be completely random because the "p" value is equal to zero. Ideally, consistent inclusion of missing values can reduce the accuracy and bias of the prognostic process given a CKD condition that is better than a negative CKD condition. This condition may be due to other related features.

Considering the flaw in defined in the related exercise, this exercise used the K-nearest neighbour algorithm to fill in the missing values. Using thealgorithm, the actual distribution of locations was stored as an integer using multiple scales (hyperparameter algorithms), providing small measurements and small changes in standard deviation.

B. Data Preprocessing: Feature Selection

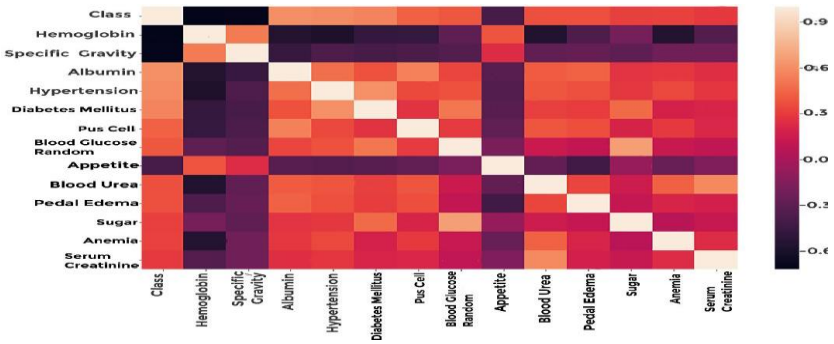
The specified values of warmth map of correlations of attributes to the category label (Fig. 3) show that haemoglobin, relative density, albumin, hypertension and DM have the very best correlations (more than 0.5). Then the secondary attributes pus cell, blood sugar random, appetite, blood urea, pedal edema, sugar, anaemia and serum creatinine are the attributes which have correlations of over 0.3.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

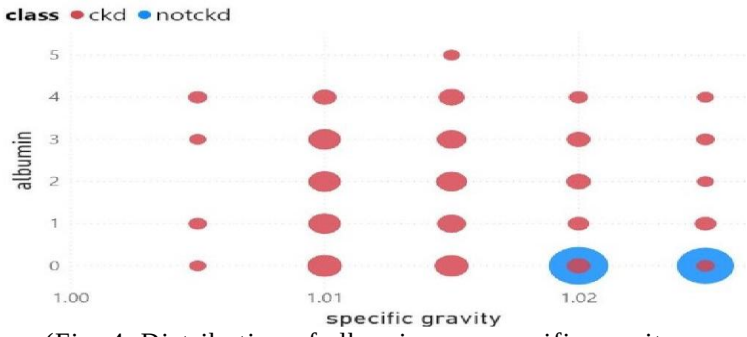
|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|



(Fig. 3. Heat map of the co-relation of attributes to the class variable.)

The specific gravity and albumin are only 5 sets of values in each (Fig. 4). When organized competitively, their numbers form a distinct group with the negative characteristics of CKD.



(Fig. 4. Distribution of albumin over specific gravity.
Note: The count of each type represented by the size of the circle.)

The amount of albumin is estimated on the basis of a test for protein in urine. Many tests should be done to confirm the condition over several weeks. Generally, the hemoglobin level can decrease due to three reasons, namely decreased red blood cell production, increased red blood cell destruction and blood loss. Healthy kidney produces a hormone called erythropoietin (EPO).A hormone is a chemical produced by the body and released into the blood to help regulate particular body functions. EPO prompts the bone marrow to make RBC which then carry oxygen throughout the body. When kidneys are damaged, they do not make enough EPO. As a result, the bone marrow makes fewer red blood cells, causing anemia but before it causes anemia (which happens the two classes: positive and negative (Fig. 5) [2].

TABLE III
PERCENTAGE CHANGE OF STATISTICS OF ATTRIBUTES AFTER FILLING MISSING VALUES

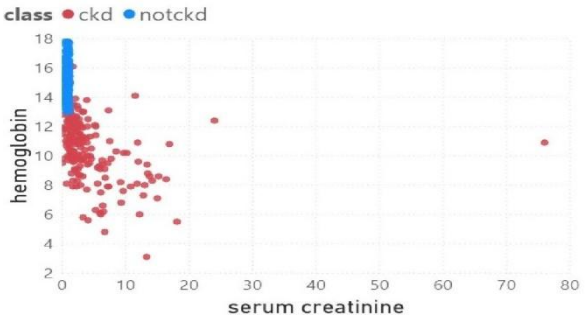
	Hb	Specific Gravity	Albumin	Hypertension	DM	Pus Cell	Blood Glucose Random	Appetite	BU	Pedal Edema	Sugar	Anemia	SC
count	13.00	11.75	11.50	0.50	0.50	16.25	11.00	0.25	4.75	0.25	12.25	0.25	4.25
mean	0.66	0.01	-1.39	-0.34	-0.36	-2.22	-0.14	0.04	-0.22	-0.10	-4.53	-0.15	-0.43
std	-6.32	-6.15	-5.95	-0.19	-0.19	-8.94	-5.86	-0.12	-2.43	-0.12	-6.30	-0.12	-2.16
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	5.29	0.49	0.00	0.00	0.00	0.00	1.98	0.00	0.00	0.00	0.00	0.00	0.00
50	2.62	-0.12	100	0.00	0.00	0.00	3.97	0.00	4.55	0.00	0.00	0.00	7.14
75	-2.56	0.00	0.00	0.00	0.00	100	-2.56	0.00	-3.00	0.00	100	0.00	0.89
max	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

International Journal of Innovative Research in Computer and Communication Engineering

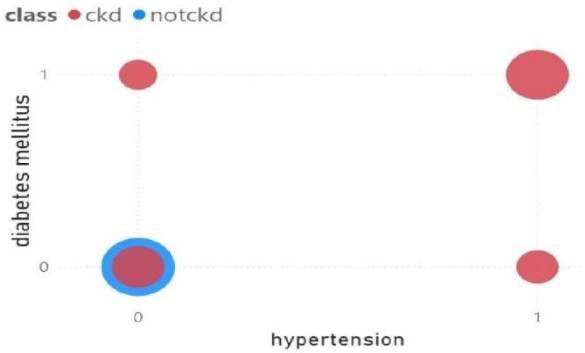
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|



(Fig.5.Distribution of hemoglobin over serum creatinine.)



(Fig. 6. Distribution of diabetes)

Creatinine is a waste product produced by muscles when a compound called creatine is broken down. Creatinine is thrown from the body through the kidneys. This test is for measuring the amount of creatinine in the blood. Creatine is part of the cycle that produces the energy needed to contract muscles. Both creatine and creatinine are produced in a certain proportion in the body. In addition to kidney problems, high-protein diets, heart failure, diabetes problems, and dehydration can also increase blood creatinine levels. The normal range of creatinine is 0.6-1.1 mg / dL for women and 0.7-1.3 mg / dL for men. The two main causes of chronic kidney disease are diabetes and hypertension (see Figure 6).

C. Model Training

In this work, 11 classification models in training were considered. They are logistic regression, kNearest Neighbors (KNN) regression, SVC with linear kernel, SVC with RBF kernel, Gaussian NB, decision tree classifier, random forest classifier, XGB classifier, extra tree classifier, Ad. a Boost classifier, and classic. Neuropil. The dataset was randomly divided into three parts: 70% training data, 15% cross-validation data, and 15% test data. The model was further optimized by hyperparameter tuning from the genetic algorithm and a grid search of the training dataset.

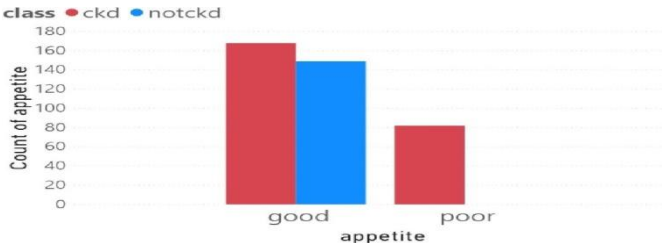


Fig. 7. The distribution of appetite.

Of the 11 algorithms above, 6 outperformed training accuracy, test accuracy, and cross-validation accuracy. These are Decision Tree Classifiers, Random Forest Classifiers, XG Boost Classifiers, Extra Tree Classifiers, Ad Boost Classifiers, and Classic Neural Networks. Implementation and evaluation was done using the Python Scikit and Keras frameworks.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|

TABLE IV
ACCURACIES OF EACH ALGORITHM

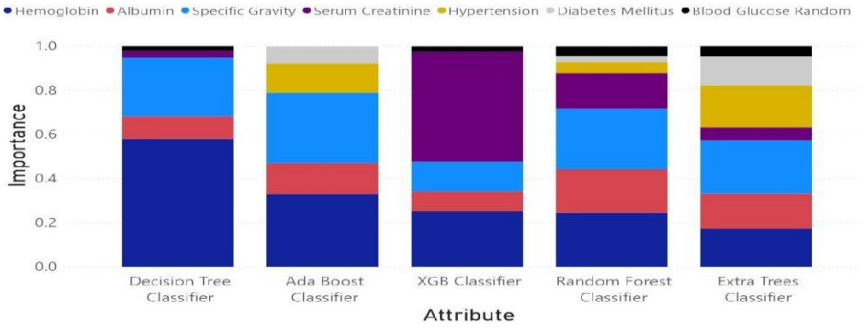
Algorithm	Training accuracy	Cross validation accuracy	Testing accuracy
Decision Tree Classifier	100.00%	100.00%	100.00%
Random Forest Classifier	100.00%	100.00%	100.00%
XGB Classifier	99.28%	100.00%	100.00%
Extra Trees Classifier	100.00%	100.00%	100.00%
Ada Boost Classifier	100.00%	100.00%	100.00%
KNN	97.85%	98.33%	98.33%
Classical Neural Network	97.81%	97.50%	97.50%
SVC Linear	97.14%	96.66%	96.66%
Logistic Regression	96.07%	96.66%	95.00%
SVC RBF	94.64%	95.00%	95.00%
Gaussian NB	95.35%	95.00%	93.33%

TABLE V
PRECISION, RECALL AND F1-SCORE OF EACH ALGORITHM

Algorithm	Precision	Recall	F1-Score
Decision Tree Classifier	1.000	1.000	1.000
Random Forest Classifier	1.000	1.000	1.000
XGB Classifier	1.000	1.000	1.000
Extra Trees Classifier	1.000	1.000	1.000
Ada Boost Classifier	1.000	1.000	1.000
KNN	1.000	0.975	0.987
Classical Neural Network	0.962	1.000	0.981
SVC Linear	1.000	0.950	0.974
Logistic Regression	1.000	0.925	0.961
SVC RBF	1.000	0.925	0.961
Gaussian NB	0.973	0.925	0.948

D. Model Evaluation and Selection

Based on the results (Tables 4, 5), an algorithm that provides high accuracy for all selected data sets. These are the categories of Decision Tree, Random Forest, XG Boost, Extra Tree, and Ada Boost . Although the above models provide 100% accuracy, it is important to identify the factors that have the most impact on each model in order to make a decision. Table 6. clearly shown in Table 6, Table 7, Figure 8. The results (Table 4, 7,8) show that the additional tree categories have a very low feature. styles near random forest dividers. The decision tree category has the highest bias.



(Fig. 8. Feature importance of each trained model.)

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|

IV. DISCUSSION

The distribution of data on the overall background of CKD is well documented, but common factors such as food allergies, anemia, and swelling of the feet are distorted in CKD. It's easy to get accurate predictions using this dataset, but in the course of a normal event, this can lead to false positives, as shown in the memory columns of Table V. Given the health values of the attributes, some attributes overlap slightly compared to other attributes due to the patient category. Model training has a significant impact on accuracy. After model training, the selected attributes show a clear class distinction without serum creatinine expression, which clearly shows that the tree structure is more accurate than other classification algorithms. This is excluded from the transfer of collected data. Finally, as shown in Table 4, some trained models are slightly biased when choosing an algorithm. Considering the causes of price fluctuations, there are many options other than CKD. Therefore, if you rely on a single attribute to make decisions, it is advisable to choose an additional tree splitter.

TABLE VI
FEATURE IMPORTANCE OF EACH ALGORITHM

Attribute	Decision Tree Classifier	Random Forest Classifier	XGB Classifier	Extra Trees Classifier	Ada Boost Classifier
Hemoglobin	0.580	0.246	0.252	0.174	0.330
Specific Gravity	0.265	0.275	0.135	0.242	0.320
Serum Creatinine	0.031	0.160	0.500	0.057	0.000
Albumin	0.103	0.196	0.089	0.158	0.140
Hypertension	0.000	0.051	0.000	0.192	0.130
Diabetes Mellitus	0.000	0.026	0.000	0.130	0.080
Blood Glucose Random	0.022	0.046	0.024	0.048	0.000

TABLE VII
STANDARD DEVIATION OF FEATURE IMPORTANCE OF ALGORITHMS

Attribute	Extra Trees Classifier	Random Forest Classifier	Ada Boost Classifier	XGB Classifier	Decision Tree Classifier
Standard Deviation	0.070684746	0.102219419	0.136224604	0.181369263	0.214295746

V. CONCLUSION AND FUTURE SCOPE

Chronic kidney disease (CKD) is life-threatening disease that affects about 14% of the world's population, with 100% complete prediction making it possible. The first stage
people get cheaper, less risky treatment. Some qualified engineering helps reduce the number of tasks required in the prediction algorithm and the number of medical tests performed. Completing non-existent values based on the distribution and integration of other attributes in the vicinity of the K (KNN imputer) instead of a direct variable makes the prediction model more accurate than the related function performed again. Be. Website. In addition, additional tree categories and random forest dividers are the best algorithms for making CKD predictions, as they have 100% complete accuracy and minimal deviations from certain structures compared to other structures. This feature proposes a new workflow that includes pre-processing data, managing missing values, and selecting the ability to predict CKDs status as positive or negative. In addition, this work emphasizes the importance of integrating domain information with selected features when analyzing CKD-related clinical data.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005123|

Accordingly, it is worthwhile to explore the use of KN Imputer based approach to handle missing values in data sets related to multiple diseases in future. Further, more insights into CKD can be gained by adding knowledge of genomics, water consumption patterns and food types into the analysis.

Chronic Kidney Disease (CKD) is that chronic disease that affects the people in large numbers. As the symptoms of CKD are not visible in the early stages many a times the disease is only detected when it has reached an advanced stage. This may lead to failure of the kidney and hence death. Machine learning classifiers provide an efficient way to predict the disease at an early stage. Ensemble classifiers combine the predicted output of various classifiers which further enhance the performance of the model. The four-ensemble algorithm like Bagging, Random Forest, AdaBoost and Gradient Boosting were used. The performance of these classifiers was evaluated using different metrics. Based on Accuracy the AdaBoost and Random Forest performed better with 100% Accuracy. But since the dataset was slightly imbalanced accuracy cannot be the only parameter to be considered for evaluation. Based on Precision, Bagging showed 97.29% and AdaBoost, Gradient Boost and Random Forest showed 100%. The F1-Score and AUC of 100% for AdaBoost and Random Forest was better compared to Bagging and Gradient Boost. Based on the evaluation AdaBoost and Random Forest was the best classifier when compared with Bagging and Gradient Boosting.

REFERENCES

1) N. Health, "World Kidney Day 2019: Important aspects for Chronic Kidney Disease in Modern time," *Narayana Health Care*, Mar. 14, 2019.
2) https://www.narayanahealth.org/blog/world-kidney-day-2019-important-aspects-for-chronic-kidney-disease-in-modern-time/ (accessed May 12, 2020).
3) "World Kidney Day 2019: CKD is 6th deadliest disease worldwide causing 2.4 million deaths per year; here's how to reduce risk of renal ailments," *Firstpost*. https://www.firstpost.com/india/world-kidney-day-2019-ckd-is-6th-deadliest-disease-worldwide-causing-2-4-million-deaths-per-yearheres-how-to-reduce-risk-of-renal-ailments-6256331.html (accessed May 12, 2020).
4) P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Dec. 2018, pp. 1–4, doi: 10.1109/CCAA.2018.8777449.
5) J. Aljaaf*et al.*, "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, Jul. 2018, pp. 1–9, doi: 10.1109/CEC.2018.8477876.
6) E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Inform. Med. Unlocked*, vol. 15, p. 100178, Jan. 2019, doi: 10.1016/j.imu.2019.100178.
7) Charleonnan, T. Fufaung, T. Niyomwong, W.Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *2016 Management and Innovation Technology International Conference*.
8) A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in 2018 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2018, pp. 1–9.
9) P. Yildirim, "Chronic kidney disease prediction on imbalanced databy multilayer perceptron: Chronic kidney disease prediction," 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), vol. 02, pp. 193–198, 2017.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details