Performance Analysis of Different Machine Learning Models for Early Detection of CKD

[1]Reni MR,[2]Anu Maria Joykutty

[1]PG Student,[2]Assistant Professor,
[1]Department of Computer Science and Engineering,
[1]Rajagiri School of Engineering and Technology, Cochin, Kerala, India.

Abstract : Pair of kidneys are vital organs for proper functioning of the human body. Failure in effective functioning can lead to CKD (Chronic Kidney Disease). CKD is a global health problem in which the kidneys decrease their shape and lose their ability to function, which makes the human body sick in the long run. Since there are no obvious symptoms during the early stages of CKD, patients often fail to detect the disease which leads to the final stage of CKD. Thus dialysis or kidney transplantation is the only solution to this. Early detection of CKD will help patients to take timely treatment and thereby decreases the progression of this disease.

Machine learning models can achieve this goal due to their fast and accurate recognition performance. Those who are with Chronic Kidney Disease are not aware that the medical tests they take for other purposes sometimes contain useful information about CKD disease. This data is taken for developing the model. There are many machine learning models that exist, the aim of this study is to find the performance of each model and also aims to identify important features for creating the model without much complication.

Index Terms **– Feature Selection; Machine Learning; Chronic Kidney Disease (CKD); KNN Imputation; Early Detection.**

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a global health problem affecting approximately 10% of the world's population [1],[2]. There are no symptoms during the early stages of CKD, hence it is detected at the end stage where dialysis or kidney transplantation is required urgently. In addition, CKD has high morbidity and mortality, with a global impact on the human body [3]. It can induce the occurrence of cardiovascular disease [4], [5]. CKD is a progressive and irreversible pathologic syndrome [6]. Hence, the prediction and diagnosis of CKD in its early stages is quite essential, it may enable patients to receive timely treatment to decrease the progression of the disease. On the other hand several attributes of medical tests taken for other purposes contain useful information related to CKD. To effectively use these attributes the importance of these attributes with respect to CKD should be studied in detail.

Machine learning is actively being used today. ML algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. ML models can effectively achieve this goal due to their fast and accurate recognition performance. This study involves developing six different machine learning algorithms and comparing the performance of each model.

II. LITERATURE SURVEY

In [7] authors summarized the risk factors as obesity, hypertension, diabetes mellitus, cigarette smoking, established cardiovascular disease, age being greater than 60 years, aboriginal and Torres strait Islander peoples, Maori and pacific peoples, family history of stage 5 CKD or hereditary kidney disease in a first or second degree relative and severe socioeconomic disadvantage. So detection of CKD at an early stage is important to minimize the effects of disease.

Chronic Kidney Disease, also called chronic renal disease is a condition in which kidneys reduce in size and are not able to function properly. Thus it could cause problems with waste and excess fluid accumulation in the body, potentially leading to complications [8]. To identify CKD, specific urine tests and blood tests should be taken [9]. So it is not possible to detect at an early stage without any test, since there will not be any symptoms during the early stages of CKD. On the other hand several attributes or features of blood tests or medical tests taken for other purposes may contain some traces of CKD. To effectively detect CKD, it is needed to study each attribute and its importance in detail. Machine learning models effectively help to detect CKD at an early stage using medical dataset. In the medical field, machine learning has already been used to detect human body status [10], analyze the relevant factors of the disease [11] and diagnose various diseases.

The study on [12] suggests neural network technique to identify CKD, The problem in this method is , if some attributes are selected to train the model then for testing every subject should contain all those selected attributes irrespective of their relative importance. The method in [8] proposed a method which aims to identify the important features that contribute more to CKD. So that when people take medical tests for any other purposes may give traces or clues for CKD. Then they may proceed to take proper tests to confirm CKD and further they can take proper meditations.

III. PROBLEM DEFINITION

To detect the presence of chronic kidney disease at an early stage through integrated machine learning based classification model.

IV. METHODOLOGY

The motive of this paper is to compare different machine learning models and analyze the model which gives high and low accuracy. The architecture diagram of the proposed model is shown in Figure 4.1.

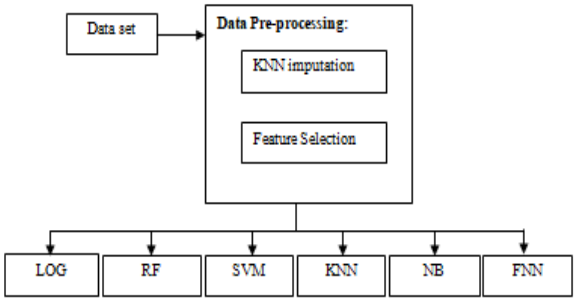

Figure 4.1. System Architecture

4.1. Description of Dataset

The CKD data set used in this study was collected from Kaggle which is also available in UCI machine learning repository. The data set contained 400 samples including both ckd and notckd classes.

This CKD dataset contains 24 predictive variables or attributes and a categorical response variable. The 24 predictive attributes constitute 11 numerical attributes and 13 categorical attributes. The categorical response variable has 2 values as ckd (sample with CKD) and notckd (sample without CKD). In the 400 samples, 250 samples belong to the category of ckd, whereas 150 samples belong to the category of notckd. The data set contains some amount of missing values. The details of each attribute in the dataset are listed in Table 4.1.

Variables	Explain	Class
age	Age	Numerical
bp	Blood Pressure	Numerical
sg	Specific Gravity	Nominal
al	Albumin	Nominal
su	Sugar	Nominal
rbc	Red Blood Cells	Nominal
pc	Pus cell	Nominal
pcc	Pus cell clumps	Nominal
ba	Bacteria	Nominal
bgr	Blood glucose random	Numerical
bu	Blood urea	Numerical
sc	Serum Creatinine	Numerical
sod	Sodium	Numerical
pot	Potassium	Numerical
hemo	Hemoglobin	Numerical
pcv	Packed cell volume	Numerical
wbcc	White blood cell count	Numerical
rbcc	Red blood cell count	Numerical
htn	Hypertension	Nominal
dm	Diabetes Mellitus	Nominal
cad	Coronary artery disease	Nominal
appet	Appet	Nominal
pe	Pedal edema	Nominal
ane	Anemia	Nominal
class	Class	Nominal

Table 4.1: Details of variables in CKD

4.2. Data Pre-processing

Each categorical variable in the dataset was encoded to numerical type to facilitate the processing in a computer. Normal and abnormal were coded as 1 and 0 respectively. Similarly the values present and notpresent as 1 and 0 respectively. The values yes and no were coded as 1 and 0, respectively. good and poor were coded as 1 and 0, respectively. Although the original data description defines three variables sg, al and su as categorical types, the values of these three variables are still numeric based, thus these variables were treated as numeric variables.

```
dataset.isnull().sum()
```

age	9
bp	12
sg	47
al	46
su	49
rbc	152
pc	65
pcc	4
ba	4
bgr	44
bu	19
sc	17
sod	87
pot	88
hemo	52
pcv	71
wbcc	106
rbcc	131
htn	2
dm	2
cad	2
appet	1
pe	1
ane	1
class	0

Figure 4.2. Count of null values in each variable

The data set contains some amount of missing data that is shown in Figure 4.2. These missing values are filled using KNN imputation. The missing values in the original CKD data set were processed and filled first after encoding the categorical variables. For each sample with missing values, KNN imputation was used, which selects the K full samples with the shortest Euclidean distance. The missing values for numerical variables are filled using the median of the corresponding variable in the K full sample. The missing values in category variables are filled with the category with the highest frequency in the corresponding variable in K full samples. People with similar physical conditions should have similar physiological measurements, which is why the system based on a KNN is used to fill in the missing values for physiological measurements. When selecting the median of corresponding variables in K complete samples, K is ideally chosen as an odd number because when the values of the numeric variables in the K complete samples are sorted by numerical value, the middle number is naturally the median. K should be chosen in a way that is neither too big nor too small. An unusually high K value could cause the inconspicuous mode to be overlooked, which could be crucial. An unusually small K value, on the other hand, creates noise, and the irregular data has a significant impact on the filling of missing values. As a result, the values of K in this work are 3, 5, 7, 9, and 11. Five full CKD data sets were created as a result.

The CKD data set contains 25 predictive and one categorical response variables. The important variables or features are selected based on its importance or contribution. Using feature vectors or predictors, variables that are not useful for prediction or not linked to response variables can be deleted. This prevents those variables from interfering with model construction and causing the models to make accurate predictions. The method used here is the Feature importance selection method. A score is given for each attribute based on its importance or contribution. Higher the score more, important or relevant is the feature. All the least important attributes which have zero importance are dropped. To select the variables that are most meaningful to the forecast, optimal subset regression and RF is used. The model output of all possible combinations of predictors is detected, and the best combination of variables is selected. The contribution of each variable to the reduction in the Gini index is detected by RF.The higher the Gini index, the more uncertain it is to identify the samples. As a result, variables with a contribution of 0 are considered redundant. Each complete data set was subjected to the feature extraction stage. The importance of each attribute for different data sets is given below.

KNN IMPUTED 3 DATA SET :

attribute: sg Importance: 0.19
attribute: hemo Importance: 0.16
attribute: sc Importance: 0.14
attribute: al Importance: 0.11
attribute: pcv Importance: 0.09
attribute: htn Importance: 0.06
attribute: rbcc Importance: 0.05
attribute: dm Importance: 0.04
attribute: bgr Importance: 0.03
attribute: bu Importance: 0.03
attribute: bp Importance: 0.02
attribute: sod Importance: 0.02

attribute: age Importance: 0.01
attribute: su Importance: 0.01
attribute: pc Importance: 0.01
attribute: pot Importance: 0.01
attribute: wbcc Importance: 0.01
attribute: appet Importance: 0.01
attribute: pe Importance: 0.01
attribute: rbc Importance: 0.0
attribute: pcc Importance: 0.0
attribute: ba Importance: 0.0
attribute: cad Importance: 0.0
attribute: ane Importance: 0.0

KNN_IMPUTED_5 DATA SET :

attribute: sg Importance: 0.18
attribute: sc Importance: 0.16
attribute: hemo Importance: 0.14
attribute: al Importance: 0.11
attribute: pcv Importance: 0.09
attribute: htn Importance: 0.05
attribute: bgr Importance: 0.04
attribute: rbcc Importance: 0.04
attribute: dm Importance: 0.04
attribute: bu Importance: 0.03
attribute: sod Importance: 0.02
attribute: age Importance: 0.01
attribute: bp Importance: 0.01
attribute: su Importance: 0.01
attribute: pc Importance: 0.01
attribute: pot Importance: 0.01
attribute: wbcc Importance: 0.01
attribute: appet Importance: 0.01
attribute: pe Importance: 0.01
attribute: rbc Importance: 0.0
attribute: pcc Importance: 0.0
attribute: ba Importance: 0.0
attribute: cad Importance: 0.0
attribute: ane Importance: 0.0

KNN_IMPUTED_7 DATA SET :

attribute: sg Importance: 0.2
attribute: hemo Importance: 0.16
attribute: sc Importance: 0.14
attribute: al Importance: 0.11
attribute: pcv Importance: 0.09
attribute: rbcc Importance: 0.05
attribute: htn Importance: 0.05
attribute: dm Importance: 0.05
attribute: bgr Importance: 0.03
attribute: bu Importance: 0.02
attribute: sod Importance: 0.02
attribute: age Importance: 0.01
attribute: bp Importance: 0.01
attribute: su Importance: 0.01
attribute: pot Importance: 0.01
attribute: wbcc Importance: 0.01
attribute: appet Importance: 0.01
attribute: pe Importance: 0.01
attribute: rbc Importance: 0.0
attribute: pc Importance: 0.0
attribute: pcc Importance: 0.0
attribute: ba Importance: 0.0
attribute: cad Importance: 0.0
attribute: ane Importance: 0.0

KNN_IMPUTED_9 DATA SET :

attribute: sg Importance: 0.2
attribute: hemo Importance: 0.17
attribute: sc Importance: 0.12
attribute: pcv Importance: 0.11
attribute: al Importance: 0.1
attribute: rbcc Importance: 0.05
attribute: htn Importance: 0.05
attribute: bgr Importance: 0.04
attribute: dm Importance: 0.04
attribute: bu Importance: 0.02
attribute: sod Importance: 0.02
attribute: age Importance: 0.01
attribute: bp Importance: 0.01
attribute: su Importance: 0.01
attribute: pot Importance: 0.01
attribute: wbcc Importance: 0.01
attribute: appet Importance: 0.01
attribute: pe Importance: 0.01
attribute: rbc Importance: 0.0
attribute: pc Importance: 0.0
attribute: pcc Importance: 0.0
attribute: ba Importance: 0.0
attribute: cad Importance: 0.0
attribute: ane Importance: 0.0

KNN_IMPUTED_11 DATA SET :

attribute: hemo Importance: 0.18
attribute: sg Importance: 0.17
attribute: pcv Importance: 0.13
attribute: sc Importance: 0.11
attribute: al Importance: 0.1
attribute: rbcc Importance: 0.05
attribute: htn Importance: 0.05
attribute: bgr Importance: 0.04
attribute: dm Importance: 0.04
attribute: bu Importance: 0.03
attribute: sod Importance: 0.02
attribute: age Importance: 0.01
attribute: bp Importance: 0.01
attribute: su Importance: 0.01
attribute: pc Importance: 0.01
attribute: pot Importance: 0.01
attribute: wbcc Importance: 0.01
attribute: appet Importance: 0.01
attribute: pe Importance: 0.01
attribute: rbc Importance: 0.0
attribute: pcc Importance: 0.0
attribute: ba Importance: 0.0
attribute: cad Importance: 0.0
attribute: ane Importance: 0.0

4.3. Model Creation

The following machine learning models for diagnosing CKD were generated by using the corresponding subset of features or predictors on the entire CKD data sets.

1. LOGISTIC REGRESSION (LOG)
2. RANDOM FOREST (RF)
3. SUPPORT VECTOR MACHINE (SVM)
4. K-NEAREST NEIGHBOUR (KNN)
5. NAIVE BAYES (NB)
6. FEED FORWARD NEURAL NETWORK (FNN)

Diagnostic samples are usually distributed in a multidimensional space in disease diagnosis. Predictors are stored in this space and are used to classify data (ckd or notckd). Due to their various categories, data samples in the space are clustered in different regions. As a result, the two categories are separated by a line, and the distances between samples in the same category are smaller. The aforementioned methods are selected for disease diagnosis based on classification effectiveness.

Logistic regression is a supervised classification algorithm. The sigmoid function is used in logistic regression to model the data. The target variable in binary logistic regression has only two potential outcomes: yes or no, ckd or notckd. The weight of each predictor and a bias are obtained using LOG, which is based on linear regression. If the number of all predictor effects exceeds a certain threshold, the sample will be classified as ckd or notckd.

By randomly sampling training samples and predictors, RF produces a large number of decision trees. The aim of each decision tree is to find a boundary that maximizes the difference between ckd and notckd. The final decision is based on all of the trees' predictions in the disease diagnosis. The Random Forest algorithm has two phases. Forest formation at random, makes a guess based on the first stage's random forest classifier.

THE ALGORITHM I – LOG MODEL BUILDING

Step 1: Load the dataset that has been prepared.

Step 2: Creating a training set and a test set from the dataset

Step 3: Using the Logistic Regression() function, construct a Logistic Regression classifier object.

Step 4: Using fit(), place the model on the train set.

Step 5: Perform prediction on the test data using predict().

Step 6: Evaluate the performance of the classification model.

THE ALGORITHM II – RANDOM FOREST CREATION

Step 1: Pick “K” features at random from a total of “m” features.

Step 2: Calculate the node "d" using the best split point among the "K" features.

Step 3: Split the node into child nodes using the best split.

Step 4: Repeat steps 1–3 until you cross the “l” number of nodes.

Step 5: To make a forest, repeat steps 1 through 4 for a “n” number of times to get a “n” number of trees.

THE ALGORITHM III – RANDOM FOREST PREDICTION

Step 1: Takes the test features and predicts the outcome using the rules of each randomly generated decision tree, then saves the predicted result (target)

Step 2: Estimate the number of votes for each expected target.

Step 3: Consider the random forest algorithm's final prediction as the target with the most votes.

SVM separates various types of samples by creating a decision surface in a multidimensional space that includes the samples' predictors.SVM plots data items in n-dimensional space, where n denotes the number of features. A coordinate value is assigned to each feature. It then seeks out a hyper plane that distinguishes the classes. Maximizing the distances (Margins) between the nearest data point and the hyper plane will assist us in selecting the appropriate hyper plane.

K- Nearest Neighbors (KNN) is a supervised machine learning algorithm. The performance of k-NN classification is a class membership. The object is assigned to the class that has the most members among its k closest neighbors.

THE ALGORITHM IV – K-NEAREST NEIGHBORS

Step 1: Load the desired data.

Step 2: Decide on the value of k.

Step 3: Consider Iterate from 1 to the total number of training data points to get the expected class.

3.1: Calculate the distance between each row of training data and the test data.

3.2: Sort the measured distances by distance values in ascending order.

3.3: Get the first k rows of a sorted list.

3.4: Get the most common form of these rows.

3.5: Get the expected class back.

In Naïve Bayes classifier, the conditional probabilities of the sample under the interval are calculated using the number of ckd and notckd samples in each measurement interval. It is a family of algorithms that share a common concept, namely that each pair of features being classified is independent of the others.

FNN stands for a neural network that does not contain a cycle. The following components make up feed forward neural networks: input layer, output layer, hidden layer and neuron weights.

V. RESULTS AND DISCUSSION

In this section, the performance of created models is compared with each other and is evaluated by using confusion matrix. The accuracy of each model for training and testing with different imputation values are given in Figure 5.1, Figure 5.2, Figure 5.3,Figure 5.4, Figure 5.5 and Figure 5.6.

	train_acc	test_acc	true_neg	false_pos	false_neg	true_pos
knn_imputed_3	0.993750	1.0000	28	0	0	52
knn_imputed_5	0.996875	0.9750	26	2	0	52
knn_imputed_7	1.000000	0.9875	27	1	0	52
knn_imputed_9	1.000000	0.9875	27	1	0	52
knn_imputed_11	1.000000	0.9875	27	1	0	52

Figure 5.1: Logistic Regression

	train_acc	test_acc	true_neg	false_pos	false_neg	true_pos
knn_imputed_3	100.0	98.75	27	1	0	52
knn_imputed_5	100.0	97.50	26	2	0	52
knn_imputed_7	100.0	97.50	26	2	0	52
knn_imputed_9	100.0	97.50	26	2	0	52
knn_imputed_11	100.0	97.50	26	2	0	52

Figure 5.2: Random Forest

	train_acc	test_acc	true_neg	false_pos	false_neg	true_pos
knn_imputed_3	0.990625	0.9875	28	0	1	51
knn_imputed_5	0.990625	0.9750	27	1	1	51
knn_imputed_7	0.993750	0.9750	27	1	1	51
knn_imputed_9	0.990625	0.9750	27	1	1	51
knn_imputed_11	0.993750	1.0000	28	0	0	52

Figure 5.3: Support Vector Machine

	train_acc	test_acc	true_neg	false_pos	false_neg	true_pos
knn_imputed_3	0.971875	0.9875	28	0	1	51
knn_imputed_5	0.975000	0.9750	27	1	1	51
knn_imputed_7	0.981250	0.9875	28	0	1	51
knn_imputed_9	0.971875	0.9875	28	0	1	51
knn_imputed_11	0.971875	0.9875	28	0	1	51

Figure 5.4: K-Nearest Neighbor

	train_acc	test_acc	true_neg	false_pos	false_neg	true_pos
knn_imputed_3	0.931250	0.9875	28	0	1	51
knn_imputed_5	0.934375	0.9875	28	0	1	51
knn_imputed_7	0.937500	0.9875	28	0	1	51
knn_imputed_9	0.937500	0.9875	28	0	1	51
knn_imputed_11	0.937500	0.9875	28	0	1	51

Figure 5.5: Naive Bayes Classifier

	train_acc	test_acc	true_neg	false_pos	false_neg	true_pos
knn_imputed_3	1.0	1.0000	28	0	0	52
knn_imputed_5	1.0	1.0000	28	0	0	52
knn_imputed_7	1.0	0.9875	28	0	1	51
knn_imputed_9	1.0	1.0000	28	0	0	52
knn_imputed_11	1.0	0.9875	28	0	1	51

Figure 5.6: Feed forward Neural Network

It is evident that the accuracy of each imputation model varies. The naive bayes model has the same accuracy for all KNN imputation values. Logistic regression and feed forward neural networks outperform the other three models in terms of accuracy.

VI. CONCLUSION

This work proposed a comparative study based on the performance of six different machine learning algorithms. In terms of data imputation and sample diagnosis, the proposed CKD diagnostic approach is feasible. However, due to the limitations of the circumstances, the available data samples are relatively small, with only 400 samples in the process of establishing the model. As a consequence, the model's generalization efficiency could be reduced. It is assumed that as the size and quality of the data grows, these models and their comparative analysis will improve.

VII. FUTURE SCOPE

This initiative can be improved in the future by integrating two week models and creating a GUI for fast and easy CKD detection. In addition a large amount of more nuanced and representative data will be collected in order to train the model and enhance its generalization efficiency while also allowing it to detect disease severity.

References

[1] Z.Chenetal. Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers. Chemometr.Intell.Lab.,vol.153,pp.140-145,Apr. 2016.

[2] A. Subasi, E. Alickovic, J. Kevric. Diagnosis of chronic kidney disease by using random forest. in Proc. Int. Conf. Medical and Biological Engineering, Mar. 2017, pp. 589-594.

[3] C. Barbieri et al. A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis. Comput. Biol. Med.,vol.61,pp.56-61,Jun. 2015.

[4] V. Papademetriou et al. Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study. Am. J. Med., vol. 130, no. 12, Dec. 2017.

[5] N. R. Hill et al. Global prevalence of chronic kidney disease - A systematic review and meta- analysis. Plos One, vol. 11, no. 7, Jul. 2016.

[6] M.M.Hossainetal. Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo clinical results in kidney allografts. IEEE Trans. Ultrason. Ferr., vol. 66, no. 3, pp. 551-562, Mar. 2019.

[7] D. Johnson. (2012, July). Risk factors for early chronic kidney disease.[Online]. Available: http://www.cari.org.au/CKD/CKD%20early/Risk_Factors_Early_CK D.pdf [June 10, 2016].

[8] A.Nishanth and T. Thiruvaran. Identifying Important Attributes for Early Detection of Chronic Kidney Disease. in *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 208-216, 2018, doi: 10.1109/RBME.2017.2787480.

[9] "National Chronic Kidney Disease Fact Sheet", June. 6, 2016. [Online]. Available: Centers for Disease Control and Prevention [June. 10,2016].

[10] R. Abbas et al. Classification of foetal distress and hypoxia using machine learning approaches. in Proc. Int. Conf. Intelligent Computing, Jul. 2018, pp. 767-776.

[11] M. Mahyoub, M. Randles, T. Baker and P. Yang. Comparison analysis of machine learning algorithms to rank alzheimer's disease risk factors by importance.inProc.11thInt.Conf.DevelopmentsineSystemsEngineering, Sep. 2018.

[12] L. J. Rubini and P. Eswaran, "Generating comparative analysis of early stage prediction of Chronic Kidney Disease." International Journal Of Modern Engineering Research, vol. 50, pp. 49–55, Jul. 20