European Journal of Molecular & Clinical Medicine

ISSN 2515-8260 Volume 10, Issue 01, 2023

Original Research Article

Utilizing Machine Learning Algorithms For Kidney Disease Prognosis

Mr. Satish Dekka[1*], Dr.K. Narasimha Raju[2], Dr.D. ManendraSai[3], Ms M. Pallavi[4]

[1*]Associate Professor, Department of Computer Science and Engineering, Lendi Institute of Engineering and Technology (A), Vizianagaram, JNTUK-AP

[2]Associate Professor, Department of Computer Science and Engineering, Gayatri Vidya Parishad College of Engineering [GVPCE], Andhra University-AP

[3]Associate Professor, Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women. JNTUK-AP

[4]Asistant Professor, Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women. JNTUK-AP

***Corresponding Author:** Mr. Satish Dekka

[*]Associate Professor, Department of Computer Science and Engineering, Lendi Institute of Engineering and Technology (A), Vizianagaram, JNTUK-AP

ABSTRACT

Chronic kidney disease (CKD) is a major problem on the healthcare system because of its high increasing prevalence and poor morbidity. Artificial Intelligence peculating its role in every field of research including healthcare and diagnosis of diseases. Recently, machine learning approaches are applied to raise consciousness about key health hazards including chronic kidney disease (CKD). When kidneys are damaged, they are unable to perform their normal role of filtering blood. Therefore, it is tough to anticipate, recognize, and prevent such a sickness, which may result in long-term health repercussions. Machine learning methods aid in more precise forecasting to tackle this problem at an early stage. With the increase of technology aids, it makes an ambiguity on which algorithm to apply for prediction of CKD. To address these issues several machine learning algorithms such as Logistic Regression, Naive Bayes, and Decision Tree are applied. Experiments are conducted utilizing the rich set of data in a MATLAB environment. Logistic Regression shows potential for reducing mortality from chronic kidney disease by enhancing prognosis and diagnostic accuracy at an early stage.

Keywords: Chronic Kidney Disease, Logistic Regression, Naive Bayes, Decision Tree

1.INTRODUCTION

Due to its high incidence, high risk of developing into end-stage renal disease, and severe morbidity, CKD places a significant cost on the healthcare system. It is progressively evolving into a global health issue. The main causes of this condition are low water intake and unhealthy eating habits. A person only lives an average of 18 days without kidneys, hence dialysis and kidney transplantation are required. Accurate approaches for CKD early diagnostic prediction are essential. Machine learning is a method that enables artificial intelligence (AI) systems to automatically learn from their experiences and improve with time. Machine learning is focused on building computer programmes that can access data and use it to learn for themselves. The current work offers a method for determining CKD status using clinical data that comprises feature extraction and data processing. Three distinct models were trained in this study for precise prediction using a range of physiological variables and ML techniques such logistic regression (LR), decision tree (DT)

classification, and k-nearest neighbour (KNN).

2. LITERARTURE SURVEY

The investigation of chronic kidney disease and its detection is making sluggish progress. Chronic renal disease (CRD), often known as CKD, has steadily gained importance, according to past studies[1]. According to the work [2], chronic kidney disease (CKD) is one of the most important health problems due to its rising prevalence. The authors tested how effectively machine learning systems could forecast chronic kidney disease using the lowest collection of information. Several statistical tests have been run to remove redundant characteristics, including the ANOVA test, Pearson's correlation, and Cramer's V test.

The performance research of renal disease over enormous data using machine learning [3] found that disorders like diabetes, glomerulonephritis, or high blood pressure are what cause the kidneys to be in poor health. These problems could creep up on you slowly over a long time, frequently without any symptoms. Renal failure may eventually set in, requiring dialysis or a kidney transplant to extend life. Therefore, with early discovery and treatment, many repercussions can be avoided or postponed.

This endeavour aims to reduce diagnosis times and improve diagnosis accuracy by using classification algorithms based on several parameters. Chronic kidney disease prognosis Early utilizing predictive analytics in conjunction with machine learning Chronic kidney disease is a hazardous condition that develops over time and is either caused by renal pathology or by decreased kidney function [4]. Early detection and effective treatment may be able to stop or delay the progression of this chronic condition before dialysis or a kidney transplant are the patient's sole options for survival. This work examines the capability of various machine-learning approaches for the early prediction of chronic renal illness. The forecasting of chronic renal disease is one of the most important issues in health care analytics [5]. The most exciting and challenging activity in daily life is prediction in the field of medicine utilizing machine learning algorithms like the Decision tree algorithm and Nave Bayes algorithm. The performance of the aforementioned models is compared with one another in order to select the best classifier for predicting chronic renal illness for the given dataset.

The risk of developing chronic kidney disease is a significant cause of mortality and disability. CKD can be predicted more correctly by looking at the data of patients [6]. Poonam Sinha et al. reported chronic kidney disease (CKD), commonly known as chronic renal illness. Chronic renal disease [7] refers to illnesses that impair your kidneys and reduce their ability to keep you healthy. Data mining and machine learning can be used to predict chronic renal illness by entering the symptoms. The authors employed data mining and machine learning techniques to predict outcomes [8]. The techniques utilized to forecast the disease are KNN and SVM Ensemble. They used an ensemble technique for better machine learning accuracy.

Chronic kidney disease, according to Enes Celik et al., is a long-term disorder that damages the kidneys and reduces their capacity to carry out their normal tasks [9]. This disorder is suspected when urinary albumin excretion increases for longer than three months or when kidney function is noticeably impaired. Some of the issues that can arise from chronic kidney disease include high blood pressure, anaemia, bone disease, and cardiovascular disease. In this study, we investigated the critical parameters for early diagnosis, timely initiation of patient care, avoidance of illness complications, and disease prognosis. Lambodar Jena et al[10] .'s work concentrated on using feature selection and classification machine learning algorithms to predict the risk of chronic diseases. The biological dataset on chronic kidney disease is taken into account for the analysis of the classification model. The prediction and performance analysis of chronic renal disease using machine learning techniques is a promising approach [11]. [12].

ISSN 2515-8260 Volume 10, Issue 01, 2023

3.METHODOLOGY

Machine learning is a method that enables artificial intelligence (AI) systems to automatically learn from their experiences and improve with time. Machine learning is focused on building computer programmes that can access data and use it to learn for themselves. During the modelling step, three machine learning algorithms have been applied to the dataset to see how effectively they can detect CKD. Decision tree, naïve bayes, and logistic regression are these algorithms (LR). The system architecture is shown in Figure 2, and the algorithm is described in the section that follows. The models are created, trained on regular data, and then tested. The organisation of the data used for training and testing is governed by the benchmark ratio of 80:20

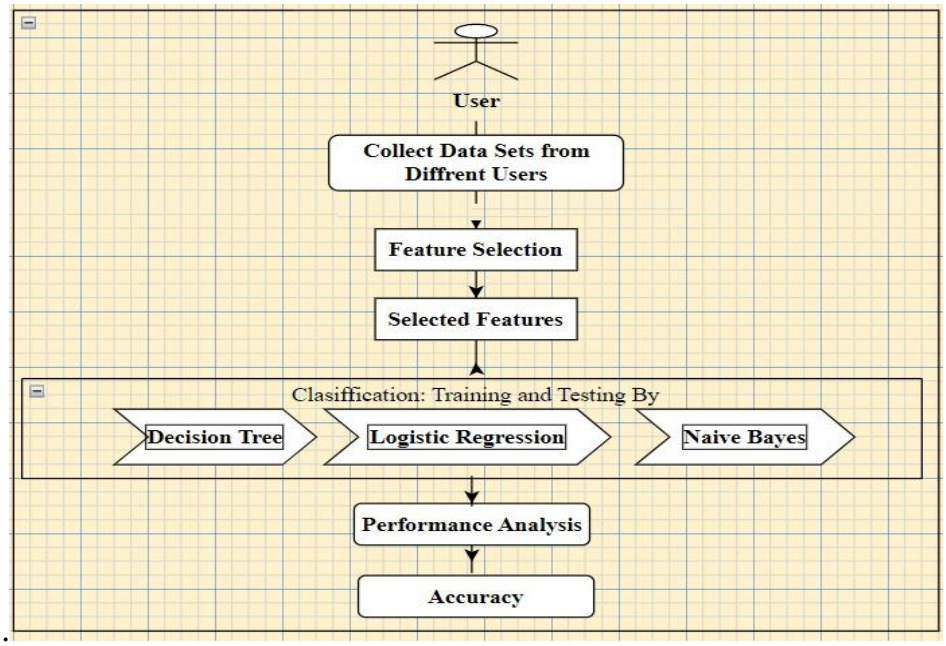

Figure 1: System Architecture

Algorithm Description
Input: Real Valued Input(x).
Output: The person is affected with CKD or not.
Step 1: Start.
Step 2: Upload the dataset.
Step 3: Extract the features
Step 4: Apply the algorithms: logistic regression (LR), Naïve bayes and Decision tree
Step 5: determine the accuracy of each algorithm
Step 6: Compare the accuracy of each algorithm with others
Step 7: Algorithm which has high accuracy is treated as best classifier.
Step 8: Stop

4. EXPERIMENTAL SETUP

MATLAB is the name of a high-performance programming language for technical computing. It combines computation, visualisation, and programming in an easy-to-use interface while describing problems and answers using conventional mathematical notation. Numerous tests are conducted, and analysis of the prediction models is done, on the dataset made public by the company "Kaggle." As shown in table 1, the data was gathered during a two-month period in India and consisted of 25 features (such as red blood cell count, white blood cell count, etc.). The aim is the classification, which can either be "ckd" or "notckd" (ckd stands for chronic kidney disease). Use machine learning techniques to determine whether a patient has chronic renal illness.

ISSN 2515-8260 Volume 10, Issue 01, 2023

Sno	Data Set Information	Attribute/Feature Details
1	age - age	Age(numerical) - age in years
2	bp - blood pressure	Blood Pressure(numerical) - bp in mm/Hg
3	sg - specific gravity	Specific Gravity(nominal) - sg - (1.005,1.010,1.015,1.020,1.025)
4	al - albumin	Albumin(nominal) - al - (0,1,2,3,4,5)
5	su - sugar	Sugar(nominal) - su - (0,1,2,3,4,5)
6	rbc - red blood cells	Red Blood Cells(nominal) - rbc - (normal,abnormal)
7	pc - pus cell	Pus Cell (nominal) - pc - (normal,abnormal)
8	pcc - pus cell clumps	Pus Cell clumps(nominal) - pcc - (present,notpresent)
9	ba - bacteria	Bacteria(nominal) - ba - (present,notpresent)
10	bgr - blood glucose random	Blood Glucose Random(numerical) - bgr in mgs/dl
11	bu - blood urea	Blood Urea(numerical) -bu in mgs/dl
12	sc - serum creatinine	Serum Creatinine(numerical) - sc in mgs/dl
13	sod - sodium	Sodium(numerical) - sod in mEq/L
14	pot - potassium	Potassium(numerical) - pot in mEq/L
15	hemo - hemoglobin	Hemoglobin(numerical) - hemo in gms
16	pcv - packed cell volume	Packed Cell Volume(numerical)
17	wc - white blood cell count	White Blood Cell Count(numerical) - wc in cells/cumm
18	rc - red blood cell count	Red Blood Cell Count(numerical) - rc in millions/cmm
19	htn - hypertension	Hypertension(nominal) - htn - (yes,no)
20	dm - diabetes mellitus	Diabetes Mellitus(nominal) - dm - (yes,no)
21	cad - coronary artery disease	Coronary Artery Disease(nominal) - cad - (yes,no)
22	appet - appetite	Appetite(nominal) - appet - (good,poor)
23	pe - pedal edema	Pedal Edema(nominal) - pe - (yes,no)
24	ane - anemia	Anemia(nominal) - ane - (yes,no)
25	class - class	Class (nominal)- class - (ckd,notckd)

Table 1:CKD Attributes and Functions

ISSN 2515-8260 Volume 10, Issue 01, 2023

4.1 Performance metrics

The effectiveness of a designed model is assessed using a variety of performance criteria. The TP, TN, FP, and FN values are shown in the confusion matrices.

True Positives (TP) are successfully predicted positive values, meaning that both the actual class value and the projected class value are true.

True Negatives (TN) are correctly predicted negative values, i.e., the value of the actual class is not the same as the value of the projected class.

False Positives (FP) are when the predicted class is true even though the actual class isn't.

False Negatives (FN) are when a class is actually present but not as projected.

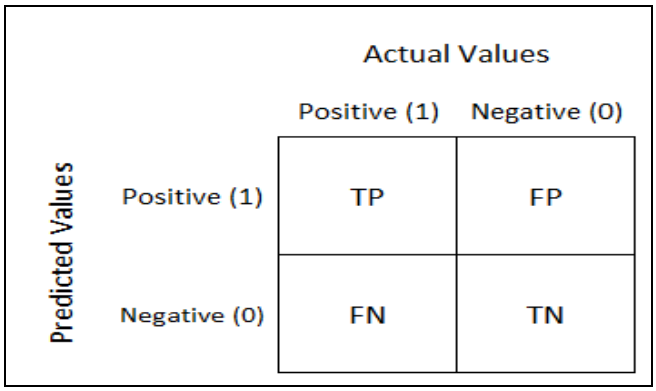

Figure 2: Confusion Matrix.

Accuracy:

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

AUC: The performance of the process is measured with the help of performance metrics like Area under Curve (AUC). The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal).

$$AUC = \frac{TP + TN}{2 \times (TP+TN+FP+FN)}$$

ROC (receiver operating characteristic) Curve: A graph of the true positive rate versus the false positive rate for various binary classification criteria. An ROC curve's form indicates how well a binary classification model can distinguish between positive and negative classes. Consider the case where a binary classification model accurately distinguishes between all the negative classes and all the positive classes.

5. RESULTS

Experiments are carried out using several machine learning models, and prediction analyses for decision tree, logistic regression, and naive bayes are provided. Three models' predictions are shown in figures 3, 4, and 5 using scatter plots. Figures 6, 7, and 8 display the three models' confusion matrix. Figures 9, 10, and 11 show three models' predictions using RoC plots. The comparison of the three models is presented in Table 2.

ISSN 2515-8260 Volume 10, Issue 01, 2023

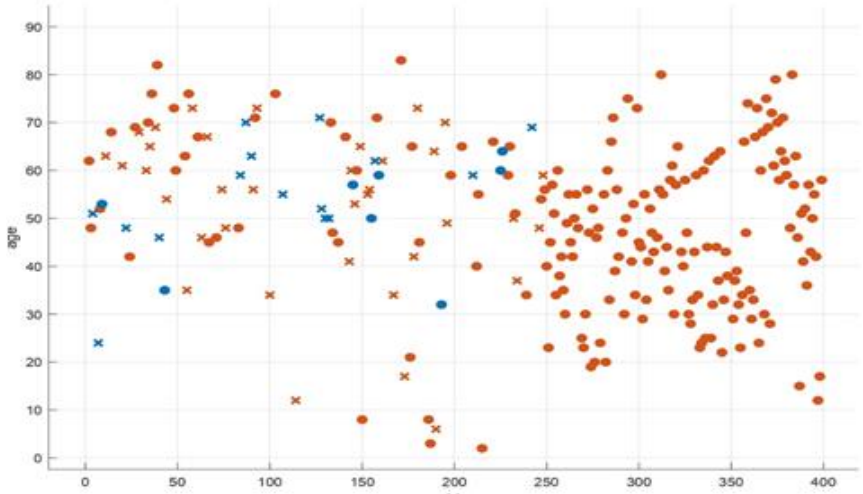

Figure 3: Running scenarios of Scatter Plot for Decision Tree

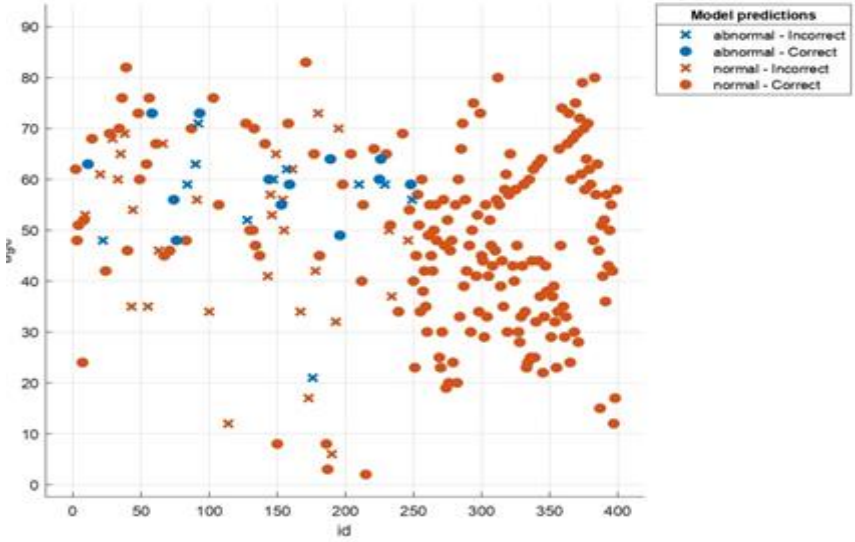

Figure 4: Running scenarios of Scatter Plot for Logistic Regression

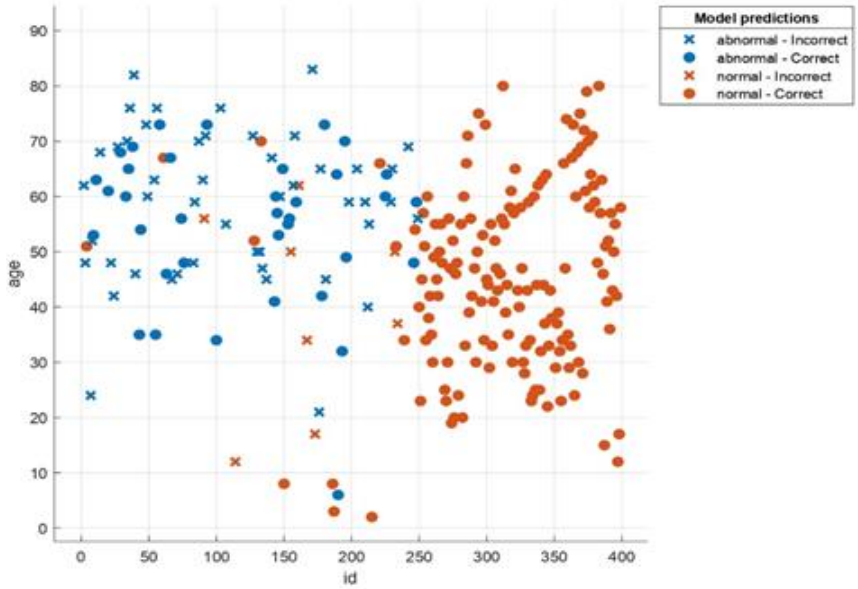

Figure 5: Running scenarios of Scatter Plot for Naive Bayes

ISSN 2515-8260 Volume 10, Issue 01, 2023

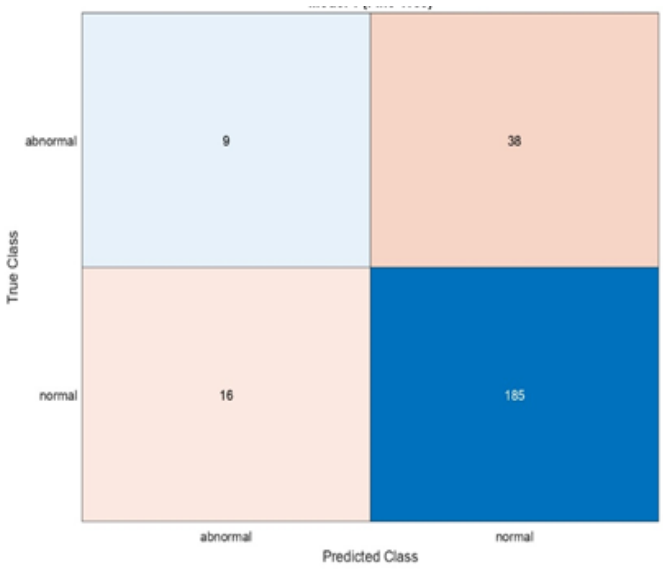

Figure 6: Running scenarios of Confusion Matrix for Decision Tree

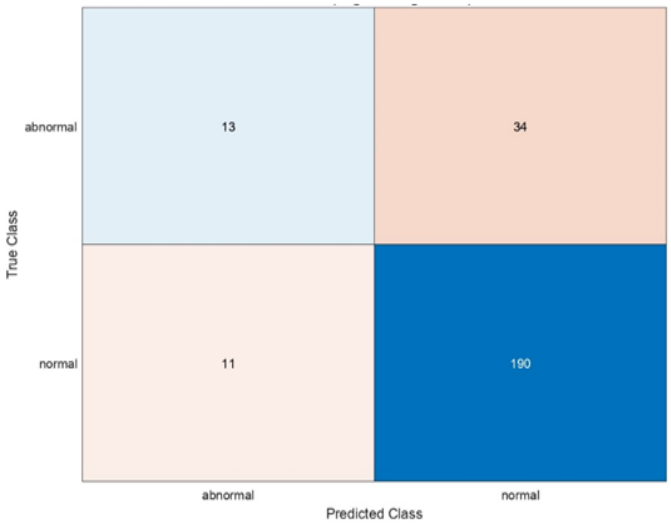

Figure 7: Running scenarios of Confusion Matrix for Logistic Regression

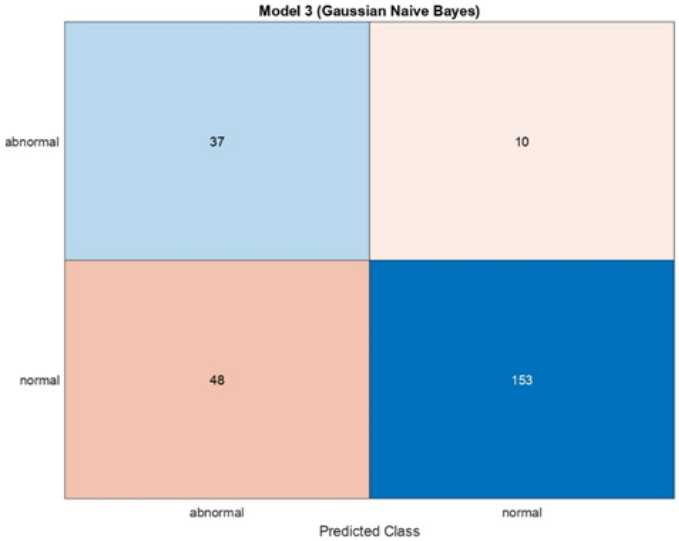

Figure 8: Running scenarios of Confusion Matrix for Naive Bayes

ISSN 2515-8260 Volume 10, Issue 01, 2023

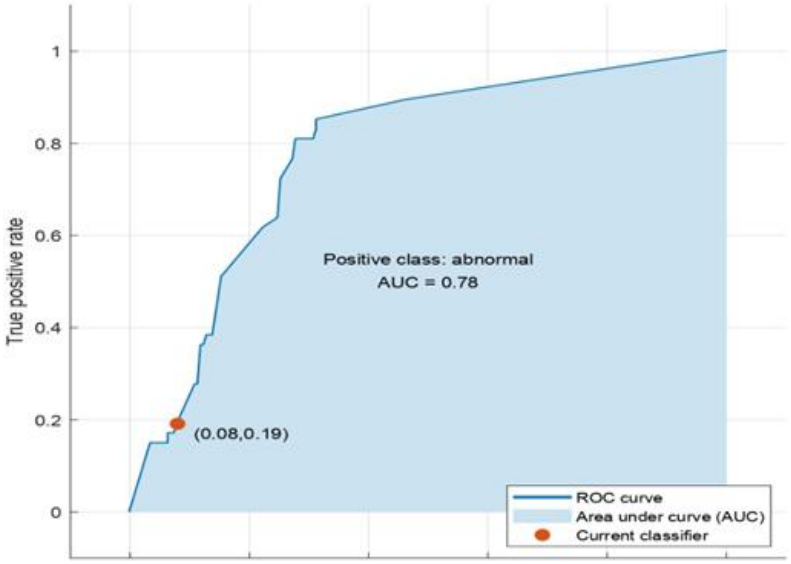

Figure 9: Running scenarios of ROC Curve for Decision Tree

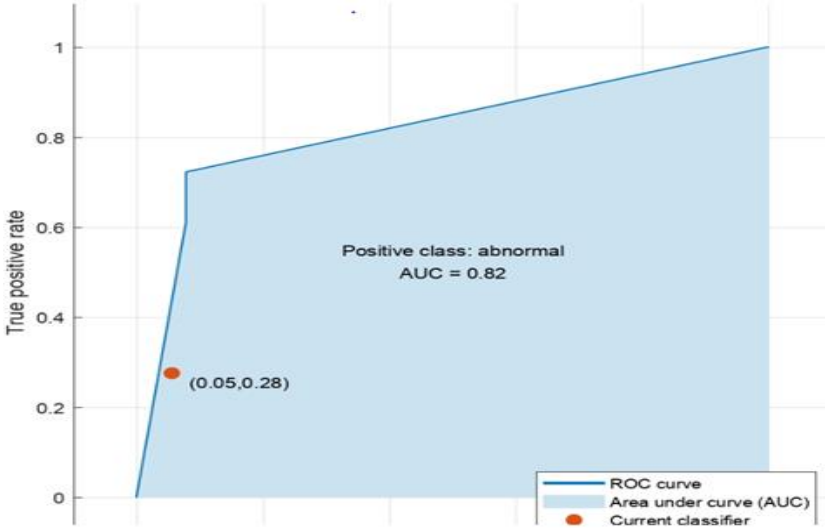

Figure 10: Running scenarios of ROC Curve of Logistic Regression

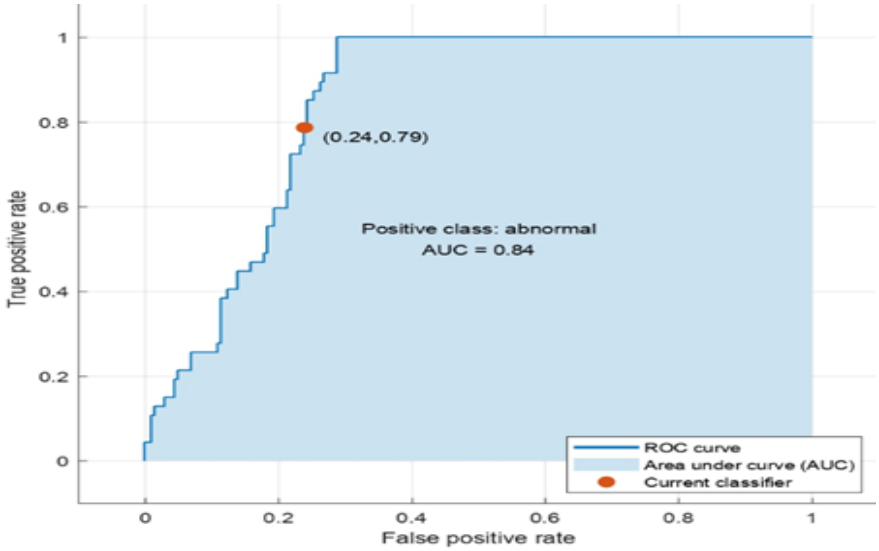

Figure 11: Running scenarios of ROC Curve of Naive Bayes

Comparison of machine learning techniques

Decision Tree	Logistic Regression	Naïve Bayes
Accuracy: 75.0 %	Accuracy: 80.2 %	Accuracy: 76.6 %
Training Time: 12.847sec	Training Time: 18.893 sec	Training Time: 5.0831 sec
Prediction Speed: ~ 3100 obs / sec	Prediction Speed: ~ 3100 obs / sec	Prediction Speed: ~ 5400 obs / sec

Table 2: Comparison of ML Techniques

6. CONCLUSION

changes in Healthcare professionals may make better decisions, spot trends and innovations, and increase the effectiveness of research and clinical trials with the help of an effective machine learning implementation. For the purpose of predicting the accuracy of CKD diagnosis, three machine learning algorithms are used. In order to preserve the patient's life at an early stage of chronic renal disease, logistic regression offers promising outcomes for better prediction and high accuracy. According to the study's findings, Logistic Regression outperforms Decision Trees and Naive Bayes at predicting chronic renal disease.

REFERENCES

1. Imesh Udara Ekanayake,Damayanthi Herath, “Chronic Kidney Disease Prediction Using Machine Learning Methods” in in British Machine Vision Conference (BMVC).

2. Marwa AL Masoud and Thomas E Ward, “Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors” in Moratuwa Engineering Research Conference (MERCon).

3. Ayesha, Parul Sinha,“Comparative study of attributes for a performance analysis of renal disease over big data using machine learning” in International Conference on Machine Learning (ICML)

4. Ahmed J. Aljaaf , Divya AI -Jumiley, Hussien M. Haglan, “Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics” in European Conference on Machine learning and knowledge discovery in databases(ECMLPKDD).

5. S. Sahitya Priya, Maria Sultana “Chronic Kidney Disease Prediction Using Machine Learning” in British Machine Vision Conference (BMVC).

6. Shanila Yunus Yashfi, Pranjal Shingavi, “Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms” in Data & Analytics Summit VIC.

7. Poonam Sinha, Parul Sinha, “Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM ” in GOTO Chicago 2021 – The International Software Development Conference.

8. Adeeba Azmi, Ayman S.Anwar,“Chronic Kidney Disease Prediction Using Data Mining and Machine Learning” in Ninth International Conference on Learning Representations(ICLR).

9. Enes Celik " The Diagnosis and Estimate Of Chronic Kidney Disease Using the Machine Learning Methods." in International Conference on Machine Learning, Optimization, Data Science (ICMLODS).

10. Lambodar Jena, Mohy Uddin,"Chronic Kidney Risk Prediction in Bio-Medical Data Using Machine Learning Approach" in The Genetic and Evolutionary Computation Conference(GECCO).

11. Minhaz Uddin Emon, Mahmud Imran, Rakibul Islam ,"Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches" in International Conference on Inventive Computation Technologies (ICICT)

12. Siddheshwar Tekale,Pranjal Shingavi,Sukanya Wandhekar,Ankit Chatorikar:"Prediction of Chronic Kidney Disease Using Machine Learning Algorithm" in International Conference on Machine Learning (ICML).