United International Journal of Engineering and Sciences
(UIJES – A Peer-Reviewed Journal); ISSN:2582-5887 | Impact Factor:8.075(SJIF)
Volume 5 | Special Issue 1 | 2025 Edition
National Level Conference on "Advanced Trends in Engineering Science & Technology" – Organized by RKCE

Early Detection of Chronic Kidney Disease Through Machine Learning Models

Y V N M SARMA[1] , E PRAVEEN[2]
[1,2] Assistant Professor
Department of ECE,
RK College of Engineering
Vijayawada, India

Abstract:

In today's fast-paced world, people are increasingly conscious of their health, yet due to busy schedules, they often only seek medical attention when symptoms arise. Chronic Kidney Disease (CKD), however, is particularly challenging to detect and predict, as it often presents no noticeable symptoms in its early stages, leading to potential long-term health complications. Machine learning (ML) provides a promising solution for this issue, offering powerful tools for prediction and analysis. This paper explores nine ML models, including K-nearest neighbors (KNN), support vector machines (SVM), logistic regression (LR), Naïve Bayes, Extra Trees Classifier, AdaBoost, XGBoost, and LightGBM, to identify the most effective approach for predicting CKD. Using a dataset with 14 attributes and 400 records sourced from Kaggle, we evaluate and compare the performance of these models. Among them, the LightGBM model achieved the highest accuracy, predicting CKD with an impressive 99.00% accuracy rate.

Keywords: Kidney disease, ML Model, Chronic Kidney Disease, K-nearest neighbors, LightGBM.

I. INTRODUCTION

Chronic kidney disease (CKD) is a severe public health issue globally, particularly in low- to middle-income countries where millions of people die due to lack of accessible healthcare. The kidney plays a crucial role in maintaining the body's balance by eliminating metabolic waste products from the bloodstream and emitting them in urine. CKD patients are at high risk for cardiovascular disease (CVD), diabetes, and high blood pressure.

The kidney's deterioration over an extended period is known as "chronic". Early detection and diagnosis can improve a patient's quality of life, and early treatment can slow the disease's progression. Machine learning (ML) is a modern technology that can predict and classify various diseases, including heart disease, breast cancer, kidney disease, and stroke. ML prediction algorithms can intelligently anticipate various diseases and provide early treatment at a lower cost, making it a potential approach for CKD diagnosis.

This research uses clinical datasets of 400 patients to determine kidney disease using nine ML algorithms and performs a comparative analysis of their corresponding results. The contribution of this study is to identify relevant features from the raw dataset by pre processing and then predicting Chronic Kidney Disease using ML techniques. This study would allow for prompt and accurate treatment of risk factors identified during appropriate and safe diagnosis of CKD.

The structure of this paper includes a literature review on applying ML in Kidney disease, methodology description, dataset representation, results discussion, and conclusion with potential future works.

II. LITERATURE REVIEW

Machine learning algorithms have been used in various industries to improve people's lives. S.Gopika, et al. developed a method for predicting chronic kidney disease (CKD) using cluster analysis, with the Fuzzy C algorithm showing the highest accuracy rate of 89%. Almasoud and Ward

analyzed 400 occurrences and 25 attributes from the CKD dataset, finding that gradient boosting had the greatest accuracy at 99.1%. Vijayarani and Dhayanand gathered a kidney function test dataset from medical labs, research facilities, and hospitals, using 584 occurrences, 6 attributes, and support vector machine (SVM) and artificial neural network classifier techniques (ANN). Sujata Drall, et al. employed a unique technique that combined data mining, machine learning, and classification algorithms to accurately predict a sick person's CKD state.

Deepika et al. developed a project for the prediction of chronic kidney disease using KNN and Naïve Bayes supervised machine learning algorithms, achieving 91% and 97% accuracy levels. Chiu et al. created cognitive models for categorizing CKD using neural network techniques, including generalized feed-forward neural networks (GRNN), backpropagation networks (BPN), and modular neural networks for early detection of CKD (MNN).

Yashfi et al. proposed a technique for CKD risk prediction using real-time datasets from Khulna City Medical College and data from 455 patients from the UCI Machine Learning Repository. AdaBoost for ensemble learning and correlation-based feature selection (CFS) were used for feature selection, with Wibawa et al.'s system having the greatest accuracy at 98.1%. The 400 instances of the UCI machine learning repository were also used by the researchers.

Table 1. The performance of previously reported results

Authors (year)	**Dataset Collection (samples)**	**Applied Methods**	**Performance (Proposed model)**
S.Gopika, etal. [1]	Big data in healthcare created by data mining technique	Fuzzy C	89%
Almasoud and Ward [2]	400 occurrences and 25 attributes from the CKD dataset	Gradient boosting	99.1%
Vijayarani and Dhayanand [3]	Kidney function test (KFT) dataset (584 occurrences, 6 attributes)	SVM and **ANN**	87.7%
Sujata Drall, etal. [4]	UCI provided a CKD dataset with 400 occurrences and 25 attributes.	KNN and Naïve Bayes	100%
Deepika et al.[5]	24 attributes and 1 target variable	KNN and **Naïve Bayes**	97%
Chiu et al. [6]	430 clinical test outcomes for patients	BPN + GA	91.71%
Yashfi et al.[7]	455 patients from the UCI Machine Learning Repository	**RF** and ANN	97.12%
Wibawa et al. [8]	Database for machine learning at UCI.	KNN, CFS, AdaBoost	98.1% (a hybrid of KNN with Adaboost and CFS)

III. METHODOLOGY

The study used various classifiers, including XG Boost, Naive Bayes, AdaBoost, Extra Trees Classifier, Random Forest, KNN, Logistic Regression, and SVM, after data gathering. A holdout

validation process was used to train and evaluate the data set on kidney disease. The results were analyzed to identify the best approach for forecasting renal disease.

3.1 Dataset collection

The study uses a dataset from the Kaggle online domain to predict chronic kidney disease based on health records. The dataset contains 400 instances and 24 attributes, including 23 predictive attributes and 1 class attribute. Risk factors for renal illness include high blood pressure, coronary artery disease, peda edema, diabetes mellitus, specific gravity, appetite, age, anemia, sugar, albumin, red blood cells, pus cells, and more. The kidney disease class attribute is used in this process.

3.2 Dataset pre-process

The study aims to clean up patient information from public sources and remove missing values using the WEKA function "Replace Missing Values." The original dataset of 400 patients was reduced to 158 instances with 24 parameters, focusing on kidney diseases. Preprocessing included handling missing values, data cleaning, feature extraction, and transformation of categorical variables. The original dataset was reduced to 158 instances with 24 parameters.

3.3 validation process

The study used hold-out validation to test 30% of a large dataset and train 70%, calculating performance metrics like precision, recall, and F1-Score for each machine learning approach. The result analysis section provided a detailed demonstration of these metrics and output graphs, while a step-by-step flowchart depicted the overall research process. This method is ideal for large datasets.

IV. DATASET

The dataset from kaggle.com was used to predict chronic kidney disease outcomes in 242 patients with missing values. The LightGBM Classifier, with proper parameter tuning and cross-validation, achieved an accuracy of 99% and a ROC AUC of 99.9%, based on 24 health-related attributes taken from 400 patients over a 2-month period.

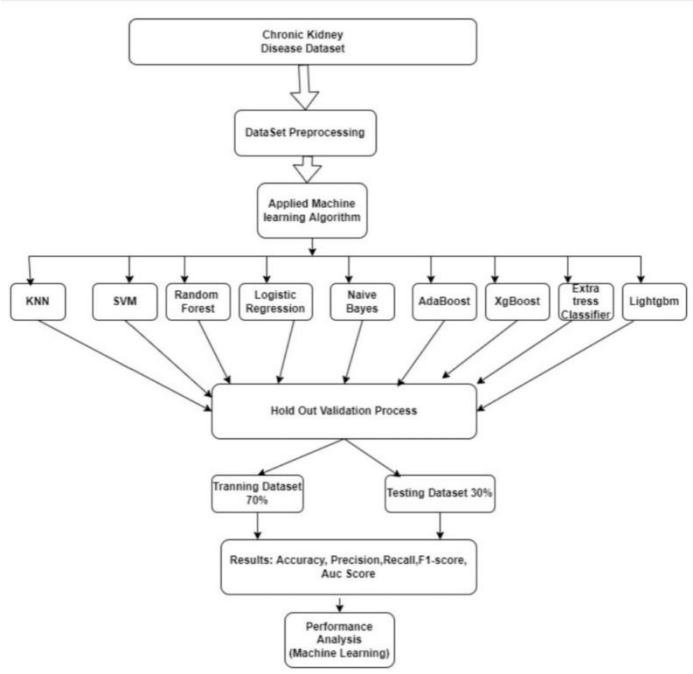

Fig 1. Proposed Methodology Overview

V. RESULTS

Table 2 compares nine machine learning algorithms based on performance metrics like recall, precision, F1 score, accuracy, and area under the curve (AUC). Accuracy alone isn't enough to evaluate a model's effectiveness. The AUC value is crucial for assessing a model's ability to distinguish between classes and assessing True Positive Rate and False Positive Rate at different thresholds along a probability curve.

Table 2. Comparison of nine machine learning algorithms.

ML Models	Accuracy (%)	Precision (%)	Recall (%)	f1- score (%)	AUROC
KNN	67.05%	74.00%	69.77%	71.51%	0.737
Random Forest	95.03%	92.00%	96.00%	98.00%	0.999
Logistic Regression	92.085%	90.33%	91.00%	94.66%	0.944
SVC	60.06%	30.89%	50.56%	75.00%	0.377
Gaussian NB	94.76%	96.00%	90.00%	95.48%	0.973
AdaBoost	96.02%	96.07%	94.54%	95.67%	0.955
XGBoost	97.012%	98.00%	94.74%	96.00%	0.999
ExtraTrees Classifier	97.23%	100%	94.33%	97.67%	0.998
LightGBM	99.00%	100%	98.89%	99.00%	0.999

The study reveals that LightGBM has the highest accuracy score of 99.00%, followed by SVM with the lowest score of 60.06%. Extra Trees Classifier scored 97.23%, while XGBoost, AdaBoost, and Random Forest scored 97.012%, 96.02%, and 95.03% respectively.

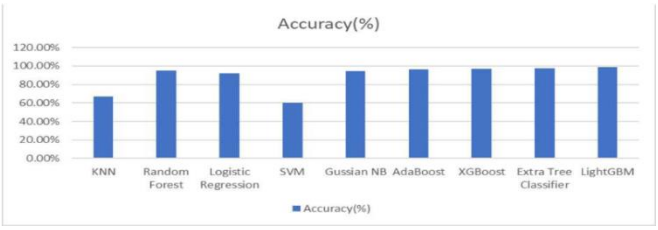

Fig 2: Accuracy analysis for predicting the Kidney Disease using machine learning models

The AUC is a measure of a model's ability to distinguish between positive and negative classes. A higher AUC indicates better results, with values ranging from 0 to 1. A score of 1 indicates perfect accuracy, while an AUC between 0.7 and 0.8 is considered acceptable. Excellent performance is between 0.8 and 0.9, and scores above 0.9 are exemplary. A hold-out validation process was used to train 70% of the dataset and test 30%. Results showed LightGBM outperforming Xgboost, Adaboost, and LightGBM in terms of accuracy, precision, and F1score.

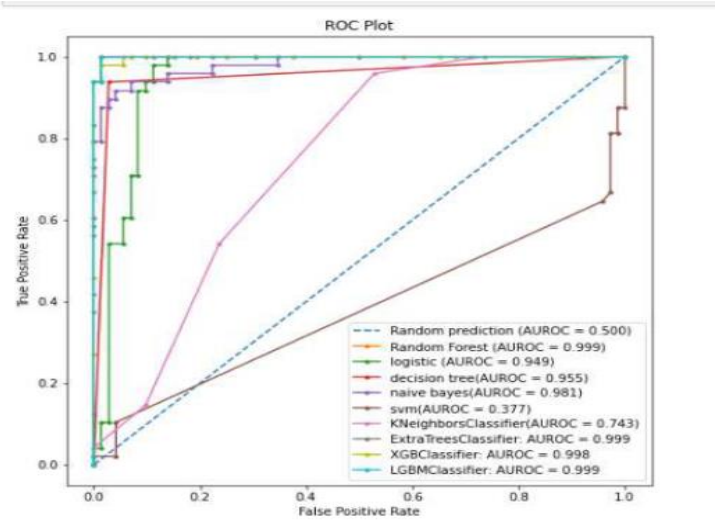

Fig 3: The ROC curve (receiver operating characteristic curve) plot.

VI. CONCLUSION

This research aims to predict Chronic Kidney Disease (CKD) early using real-world patient data. The study uses machine learning algorithms such as XGBoost, Random Forest, ExtraTrees Classifier, Naïve Bayes, Logistic Regression, SVC, AdaBoost, Light GBM, and KNN to detect CKD in patients. The results show that Light GBM outperforms other models in terms of accuracy, precision, and F1 score. These ML-based models can be used for developing resources and public health initiatives, including patient monitoring and early CKD detection. Future work aims to apply additional datasets and classification techniques like Deep Learning to improve results. The goal is to reduce the suffering of patients affected by CKD and improve patient monitoring.

VII. REFERENCES

[1] Gwozdzinski, Krzysztof, Anna Pieniazek, and Lukasz Gwozdzinski. "Reactive oxygen species and their involvement in red blood cell damage in chronic kidney disease." Oxidative medicine and cellular longevity 2021 (2021): 1-19.

[2] Saikat, Abu Saim Mohammad, Ranjit Chandra Das, and Madhab Chandra Das. "Computational Approaches for Structure-Based Molecular Characterization and Functional Annotation of the Fusion Protein of Nipah henipavirus." Chemistry Proceedings 12.1 (2022): 32.

[3] Saikat, Abu Saim Mohammad, et al. "In-Silico Approaches for Molecular Characterization and Structure-Based Functional Annotation of the Matrix Protein from Nipah henipavirus." Chemistry Proceedings 12.1 (2022): 21.

[4] R. K. Al-Ishaq, P. Kubatka, M. Brozmanova, K. Gazdikova, M. Caprnda, and D. Büsselberg, "Health implication of vitamin D on the Heidarian, Esfandiar, and Ali Nouri. "Hepatoprotective effects of silymarin against diclofenac-induced liver toxicity in male rats based on biochemical parameters and histological study." Archives of Physiology and Biochemistry 127.2 (2021): 112-118.

[5] "Preventing chronic diseases: a vital investment: WHO global report." https://apps.who.int/iris/handle/10665/43314?fbclid=IwAR0Fy2HvtoTvEI9cJLm1w8eWXwbKVs0S_UxvWTFenQ60lbjq9VfatepLCiQ (accessed Jan. 23, 2023).

[6] Bastos, Marcus Gomes, and Gianna Mastroianni Kirsztajn. "Chronic kidney disease: importance of early diagnosis, immediate referral and structured interdisciplinary approach to improve outcomes in patients not yet on dialysis." Brazilian Journal of Nephrology 33 (2011): 93-108.

[7] Roth, Jan A., et al. "Introduction to machine learning in digital healthcare epidemiology." Infection Control & Hospital Epidemiology 39.12 (2018): 1457-1462.
[8] Gopika, S., & Vanitha, M. (2017). Survey on Prediction of Kidney Disease by using Data Mining Techniques. International Journal of Advanced Research in Computer and Communication Engineering, 6(1).