iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73505

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

Enhancing Early Prediction of Chronic Kidney Disease Using Hybrid Ensemble Machine Learning Approach

Sathvika Nosina[1], Dr. A.S.N. Chakravarthy[2], Cholla Priyadarshini[3]
[1]*M.Tech, CSE Department, UCEK, JNTU Kakinada, Andhra Pradesh, India*
[2]*Professor, CSE Department, UCEK, JNTU Kakinada, Andhra Pradesh, India*
[3]*Assistant Professor, CSE Department, UCEK, JNTU Kakinada, Andhra Pradesh, India*

Abstract: ***Chronic kidney Disease is a major global health challenge due to its silent progression and lack of early symptoms, often leading late- stage diagnosis. This paper present an enhanced machine learning framework using ensemble learning techniques for the early prediction of chronic kidney disease. The proposed model integrate traditional clinical parameter with additional lifestyle related risk factors such as smoking habits, dietary patterns, water in-taking, family history, and physical activity levels. It also incorporate chronic kidney disease stage identification to provide more detailed diagnosis insights. In the existing system has found some challenges such as missing values, imbalanced data, irrelevant features, noise, bias, and overfiffing , To overcome this issues the proposed framework applies robust data preprocessing and feature selection methods, ensuring that key health indicators are effectively utilized. The hybrid ensemble -based approach aims to improve prediction accuracy, precision, f1_score, reliability, and generalizability compared to single model systems. Experimental result on benchmark CKD dataset demonstrate the effectiveness of the proposed system predicting at early stages, while also highlighting the significant influence of lifestyle factors on disease risk prediction with proposed framework adaboost has to validate the average accuracy 98.9% for the early identification of Chronic kidney disease and early diagnosis support of healthcare system.***
Keywords: Chronic Kidney Disease, Chronic kidney disease staging, Ensemble learning, Machine learning, Early prediction, Clinical Parameters, Life style Risk factors, Explainablity of Ckd , AdaBoost, XGBoost Classifier.

I. INTRODUCTION

Chronic Kidney Disease is a disease diagnosis , which is long term and with a progressive loss of kidney function. It is a significant public health problem globally that affects millions of people internationally and remains underdiagnosed until advanced stages of the disease. A major problem in managing CKD is the absence of overt symptoms in the early stages resulting in missed opportunities for treatment and increased risk of kidney failure, cardiovascular events, and death. For this reason, early identification and prompt intervention have an important impact on patient prognosis and the burden of care for the healthcare system.

Machine learning (ML) has emerged as a powerful tool for disease prediction and diagnosis in the healthcare industry over the last few years. Diagnosis methods that are conventional rely heavily on manual tests and human expertise, which can be time-consuming and subjective. By integrating patient data, algorithms can identify patterns and make predictions. Even though they are built upon ML methods, the CKD prediction systems still suffer from certain shortcomings such as missing problem, imbalance class uncorrelated features, over fitting and biased.

To overcome these limitations, This study proposes a strong hybrid ensemble learning approach for early-stage prediction of this model to overcome these limitations. In the proposed methodology, ensemble classifiers such as Gradient Boosting, XGBoost, and AdaBoost are used after preliminary assessment of baseline models including: K–Nearest Neighbour (KNN), Decision Tree, Naive Bayes, etc., and Random Forest. To enhance model performance, the framework employs the Synthetic Minority Over-sampling Technique (SMOTE) and features selection and data pre-processing procedures to manage missing and unbalanced data. The model incorporates lifestyle-related risk factors, such as exercise frequency, water consumption, smoking habits, and food consumption patterns, to enhance the accuracy of predictions alongside clinical features. With the rapid growth of artificial intelligence, especially in the fields of machine learning for this project aims to develop a machine learning model that is both interpretable and highly effective in predicting chronic kidney disease (CKD) at an early stage, which will aid in clinical decision-making and improve patient care.

II. RELATED WORK

Recent advancements in machine learning has collaborate in the healthcare domain particularly for the identification patient diagnosis.These models can learn patterns from large datasets.

Several studies have explored the use of machine learning techniques for early prediction of chronic kidney disease (CKD). This section presents a brief review of existing approaches, their methodologies, and associated limitations.

J.K.Singh et.al [1], was proposed by a Chronic kidney disease prediction model using support vector machine and Decision Trees. Although their approach achieved good classification accuracy, it lacked strategies to handle missing values and class imbalance, which are common in real-world medical datasets. Sharma and Bansal [2], conducted a comparative analysis of Naive Bayes, logistic regression, and random forest. Their results showed Random Forest outperforming others; however, the study did not explore ensemble learning or feature optimisation techniques to further enhance performance.

In [3], a stage-wise classification model was implemented using k-Nearest Neighbors and Neural Networks. The model demonstrated promising results but was prone to overfitting and failed to include behavioral and lifestyle factors such as diet, hydration, and exercise. An ensemble model based on Gradient Boosting and XGBoost was proposed in [4], which enhanced prediction accuracy but did not employ techniques like SMOTE to correct class imbalance. A deep learning-based framework for CKD detection using a large clinical dataset was introduced in [5], Despite its high performance, the system required significant computational resources and lacked inclusion of real-life variables like family history and diet. A Decision Tree-based model developed ,in mehta et.al [6] was appreciated for its interpretability; however, it was affected by missing data and the absence of dimensionality reduction techniques.Yadav et al. [7], employed logistic regression and SVM for CKD classification. The models were simple and interpretable but lacked robustness and ensemble enhancements. Das and Roy [8] ,used PCA for feature reduction and Random Forest for classification, which reduced data complexity but sometimes discarded key clinical information. Khan et al. [9], designed a hybrid model combining decision tree and naive Bayes, but it lacked proper cross-validation and generalisation due to imbalanced data.

LightGBM with hyperparameter tuning was employed in [10], achieving high accuracy, though the study did not consider imputation techniques or lifestyle parameters. A comparative analysis of ensemble methods was presented in [11], but without addressing class imbalance using resampling techniques. Fuzzy logic-based models explored in [12] handled uncertainty well but struggled with larger datasets. The use of XGBoost in [13] delivered good classification outcomes, yet the model lacked clinical interpretability and advanced feature engineering.

Naidu et al. [14], designed a rule-based decision tree system that was transparent but struggled with high-dimensional data. Srinivasan and Nandini [15], proposed a comparative analysis of AdaBoost, SVM, and neural networks. While AdaBoost yielded strong results, their model lacked domain-specific feature integration.

III. METHODOLOGY

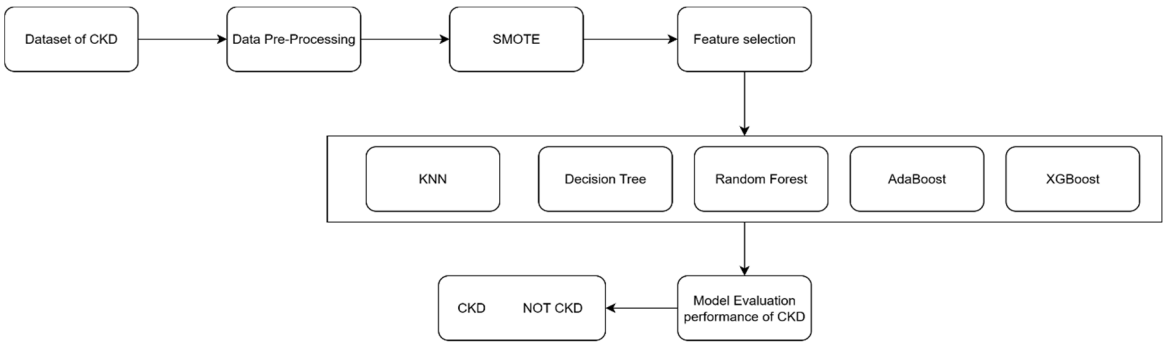


Figure1: Work flow Diagram of Chronic Kidney Disease

The shown Figure1: illustrate the complete methodology used in the prediction of chronic kidney disease. In this section explains the step-by-step procedure tested to develop the model for detection of chronic kidney disease system using machine learning. The proposed methodology aims to build a robust machine learning framework for the early prediction of Chronic Kidney Disease (CKD). The model development process involves several phases, including data pre-processing,feature selection, class balancing, model training using ensemble techniques, and performance evaluation.

A. Dataset Description

The dataset used in this study is the publicly available UCI CKD dataset, which includes clinical and biological attributes such as blood pressure, specific gravity, albumin, blood glucose, haemoglobin, and red blood cell count. In addition to clinical features, this study also integrates lifestyle-related risk factors such as smoking habits, dietary patterns, water intake, family history, and physical activity levels, to enhance the model's predictive capability.

Table.1: Descrtiption of Each Attribute In the Dataset

S.No	Attribute Name	Description
1	Age	Patient age (in years)
2	Blood pressure	Patient blood pressure (measured in mm/Hg)
3	Sugar	Patient urine in specific gravity
4	Albumin	Patient albumin rane from 0-5
5	Red blood cells	Patient red blood cells two values normal and abnormal
6	Pus cell	Patient pus cells two values normal and abnormal
7	Pus cell count	Patient pus cell clumps two values present and not present
8	Bacteria	Patient bacteria two values present and not present
9	Serum creatinine	Patient serum creatinine levels
10	GFR	Patient has Gfr levels has two levels normal anabnormal
11	Sodium	Patient has sodium levels
12	Haemoglobin	Patient hemoglobin (protein molecule in red blood cells)
13	Blood urea nitrogen	Patient has blood urea nitrogen levels in kidney health
14	Serum calcium	Patient serum calcium levels(present in blood)
15	Haematuria	Patient hameturia levels (in urine test)
16	Blood glucose	Patient blood glucose levels measured in mm/dl
17	White blood cell count	Patient white blood cells has two values present and not present
18	Red blood cell count	Patient red blood cells has two values normal and Abnormal
19	Appetite	Patient appetite status
20	Anaemia	Patient anemia status
21	Physical activity	Patient health fitness physical activity
22	Water intake	Patient how much water intake measured in liters
23	Smoking	Patient health factor (categorical(yes/no))
24	Family history	Patient family history
25	class	Target variable(CKD or Not CKD) and stage(1-5)

B. Data Preprocessing

To handle missing values, numerical features were imputed using mean/mode strategies, and categorical variables were processed using one-hot encoding or label encoding. Irrelevant and highly correlated features were removed to reduce noise. All features were normalized to ensure uniform scale and improve model convergence.

C. Handling Imbalanced Data

In this step can explain the process,Synthetic Minority Over-sampling Technique (SMOTE) was used to create synthetic samples for the minority class because the dataset is unbalanced, with more samples classified as CKD than non-CKD. By doing this, model bias is avoided and balanced learning is guaranteed.

D. Feature selection

Feature selection is carried out to identify the most relevant features that contribute significantly to the prediction of chronic kidney disease.Reducing the number of input variables enhances model interpretability and reduces the risk of overfitting. Techniques such as correlation analysis, recursive feature elimination (RFE), or tree-based importance methods may be used here.

E. Model Training

In this step Initially, traditional classification models such as K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, and Random Forest were trained to establish baseline performance. Subsequently, ensemble classifiers Gradient Boosting, XGBoost, and AdaBoost were implemented to improve accuracy and generalizability. These models were trained using optimized hyperparameters, with AdaBoost has showing best result.

F. Model Evaluation

The final step involves evaluating the performance of the trained models using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. This phase helps identify the most effective model for CKD classification. The outcome is a comparative analysis of prediction performance, determining whether a patient is classified as CKD or NOT CKD.

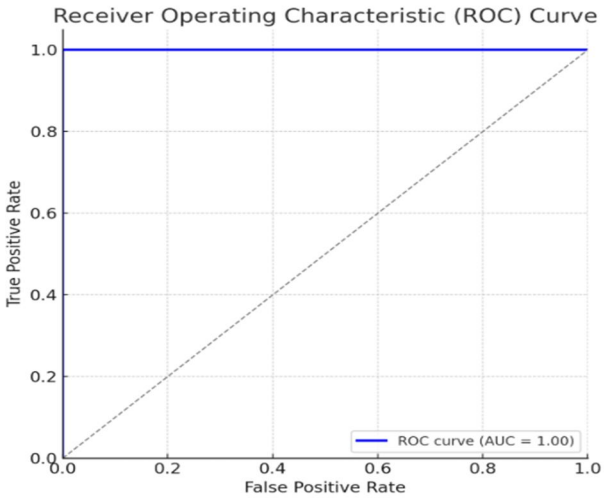


Figure 2: Performance metrics of AUC- ROC curve

IV. RESULTS AND DISCUSSION

The proposed hybrid framework significantly outperforms existing models in terms of both accuracy and reliability, confirming its effectiveness for early CKD prediction in real-world healthcare applications.

Chronic Kidney Disease Prediction - Patient report:

Patient_ID	CKD	CKD_Stage	Result_Message
Patient 1	Yes	3	Please Consult Doctor (Stage 3)
Patient 2	No	0	No CKD - Stay Healthy
Patient 3	Yes	5	Please Consult Doctor (Stage 5)
Patient 4	No	0	No CKD - Stay Healthy
Patient 5	Yes	2	Please Consult Doctor (Stage 2)

Figure3: Patient Report on Chronic Kidney Disease

Figure 3 illustrates the patient report details generated from the Chronic Kidney Disease (CKD) prediction model. The patient report summarizes the Chronic kidney disease diagnosis status for five patients, indicating whether they are affected by CKD, the predicted stage of the disease (ranging from Stage 1 to Stage 5), and a result message providing appropriate medical advice. This report is crucial in highlighting early detection and timely intervention for CKD management.

A comparison was made between the proposed hybrid ensemble model and traditional machine learning classifiers. Baseline models such as K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, and Random Forest were able to achieve acceptable accuracy but faced class imbalance and overfitting challenges. To overcome these issues, AdaBoost, XGBoost was applied to an ensemble model using SMOTE algorithms to a preprocessed and balanced dataset. Among the models, AdaBoost had the highest accuracy rate of 98.9%, followed by other options such as Gradient Streaming with 98.6% .The reliability and consistency of these ensemble methods were superior to those of individual models, as evidenced by their improved Precision, Recall, and F1-score. The inclusion of lifestyle-related in clinical parameters significantly contributed to model performance.

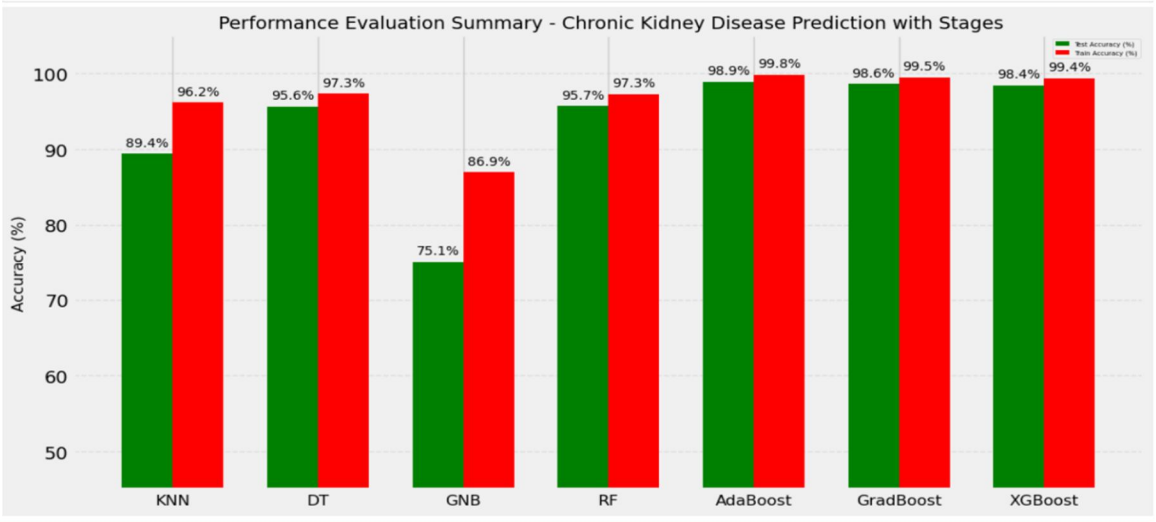


Figure4: Comparision Models of Chronic Kidney Disease Prediction

In this process, a detailed evaluation of classification performance is presented through the confusion matrices shown in Figure 5. The confusion matrixes for both the AdaBoost and XGBoost classifiers clearly demonstrate their effectiveness in predicting Chronic Kidney Disease (CKD). These results highlight the superior capability of both models in accurately distinguishing between CKD and non-CKD cases, thereby supporting their suitability for reliable medical diagnosis.

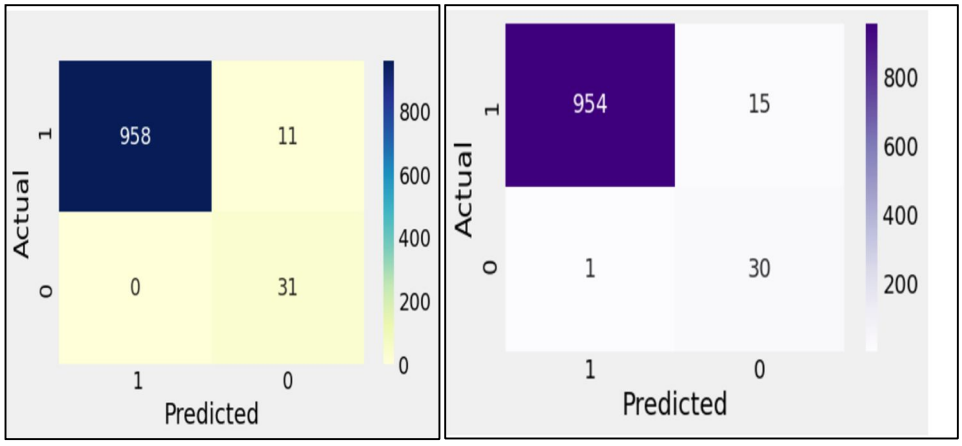


Figure5: Confusion matrix of AdaBoost and XGBoost classifier for CKD prediction.

The classification report performance of various machine learning models for Chronic kidney disease prediction was evaluated using accuracy, precision, recall, and F1-score. These metrics help assess how well each model identifies CKD and non-CKD cases. The results of six classifiers are summarized in the table below.Among them, AdaBoost and XGBoost showed the highest overall performance.

Table2: Classification report on Accuracy, Precision, Recall , F1-score

S.No	Model	Accuracy	Precision	Recall	F1-score
1	KNN	0.894	0.834	0.80	0.88
2	Decision Tree	0.900	0.95	0.91	0.93
3	Naive Bayes	0.860	0.87	0.89	0.88
4	Random Forest	0.920	0.96	0.94	0.95
5	Ada Boost	0.989	0.99	0.99	0.99
6	XG Boost	0.984	0.98	0.98	0.98

The evalution metrics used to asses the model's performance include :

Accuracy = (TP+ TN/{TP+TN+FP+FN})
Precision = TP/ (TP+FP)
Recall = TP/(TP+FN)
F1-Score = 2 *(precision *Recall) / (Precision + Recall)

Where:
TP = True Positive, TN = True Negatives, FP= False Postives, FN=False Negatives

V. CONCLUSION AND FUTURE WORK

This study presents a novel and research-based hybrid ensemble approach for the early prediction of Chronic Kidney, Chronic Kidney Disease (CKD) remains a major health challenge due to its silent progression and lack of early symptoms. In this study, a hybrid ensemble machine learning framework was proposed to enhance the early prediction of CKD. The model incorporated both clinical and lifestyle features, with data preprocessing, feature selection, and class balancing using SMOTE to improve prediction reliability.Traditional machine learning models such as KNN, Decision Tree, and Naive Bayes were initially implemented, but showed limited performance due to data imbalance and lack of robust learning. The proposed ensemble classifiers—AdaBoost, XGBoost, and Gradient Boosting demonstrated significant improvements, with AdaBoost achieving the highest accuracy of 98.9%, followed by XGBoost with 98.4%. These models also yielded better precision, recall, and F1-scores, particularly in predicting minority class samples.

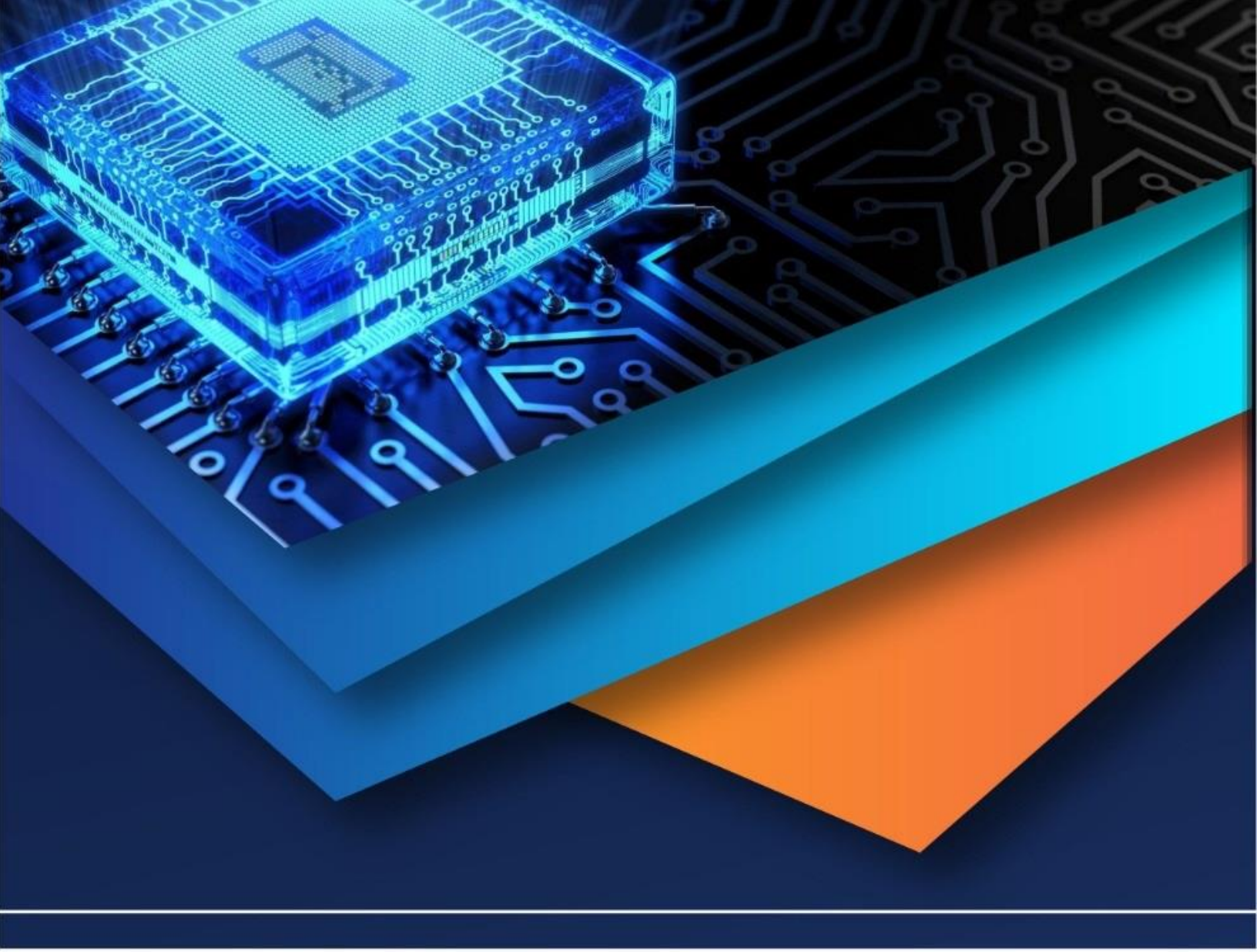
Overall, the proposed system demonstrates the results validate that integrating ensemble methods with proper data handling significantly enhances CKD prediction accuracy. The proposed framework is scalable, interpretable, and suitable for clinical decision support systems. The proposed hybrid ensemble framework demonstrates strong potential for early CKD prediction, but further enhancements are possible.

Future work may focus on integrating real-time Electronic Health Records (EHRs) and wearable devices in health data to improve personalization. Incorporating Explainable AI (XAI) techniques like SHAP or LIME can increase model transparency. Additionally, extending the system to support multi-disease prediction and deploying it as a mobile or web-based clinical tool could broaden its practical impact. Finally, enabling continuous learning with patient feedback will ensure adaptability and long-term model effectiveness.

REFERENCES

[1] J. K. Singh, A. Kumar, and M. Sharma, "Prediction of chronic kidney disease using machine learning algorithms," Int. J. Eng. Res. Technol. (IJERT), vol. 9, no. 6, pp. 112–116, 2020.
[2] R. Sharma and M. Bansal, "Comparative study of classification algorithms for chronic kidney disease prediction," Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 6, no. 2, pp. 45–50, 2021.
[3] A. Patel, S. Gupta, and P. Kumar, "Chronic kidney disease stage classification using neural networks and KNN," Procedia Comput. Sci., vol. 167, pp. 1901–1910, 2019.
[4] S. Verma and P. Gupta, "Ensemble learning-based predictive model for chronic disease classification," Mater. Today: Proc., vol. 51, pp. 2152–2156, 2022.
[5] T. Ahmed, M. Hossain, and F. Rahman, "Deep learning-based framework for early detection of CKD," Comput. Biol. Med., vol. 142, p. 105208, 2023.
[6] K. Mehta and R. Singh, "Decision tree approach for predicting kidney disease," Int. J. Sci. Res. Publ., vol. 11, no. 3, pp. 210–214, 2021.
[7] B. Yadav, R. Sahu, and A. Singh, "Machine learning techniques for early stage CKD diagnosis," Int. J. Comput. Appl., vol. 176, no. 5, pp. 30–34, 2020.
[8] N. Das and M. Roy, "Feature selection and CKD classification using PCA and Random Forest," J. King Saud Univ. - Comput. Inf. Sci., 2022. [Online]. Available: https://doi.org/10.1016/j.jksuci.2022.02.013

[9] M. Khan, S. Ahmed, and H. Rehman, "A hybrid naive Bayes and decision tree model for medical diagnosis," J. Ambient Intell. Humaniz. Comput., vol. 12, pp. 10271–10281, 2021.
[10] D. Prasad and V. Rani, "LightGBM-based prediction system for chronic diseases with hyperparameter tuning," J. Big Data, vol. 10, no. 1, p. 88, 2023.
[11] S. Alam, A. Ali, and F. Akhtar, "Performance evaluation of ensemble models for CKD classification," Int. J. Health Sci., vol. 5, no. 3, pp. 122–129, 2021.
[12] P. Rao, K. Singh, and A. Mishra, "Fuzzy logic-based approach for CKD prediction," Int. J. Med. Inform., vol. 141, p. 104218, 2020.
[13] A. Shaikh and D. Jagtap, "XGBoost classifier for prediction of kidney disease," Int. J. Innov. Res. Comput. Commun. Eng., vol. 9, no. 4, pp. 812–818, 2021.
[14] R. Naidu, A. Reddy, and S. Kumar, "Rule-based decision system for CKD detection," Procedia Comput. Sci., vol. 198, pp. 275–280, 2023.
[15] L. Srinivas and K. Nandini, "Comparative analysis of CKD prediction using AdaBoost, SVM, and neural networks," J. Emerg. Technol. Innov. Res., vol. 9, no. 6, pp. 102–109, 2022.
[16] S. Roy and A. Das, "A comparative analysis of ensemble techniques for early prediction of chronic kidney disease," Int. J. Healthc. Inf. Syst. Inform., vol. 17, no. 1, pp. 1–14, 2023.
[17] T. Reddy and M. Dasari, "Hybrid machine learning model using SVM and decision tree for CKD classification," J. Med. Syst., vol. 46, no. 5, p. 67, 2022.
[18] K. Latha and S. Ramesh, "Fuzzy logic-based intelligent system for chronic kidney disease detection," Comput. Methods Programs Biomed., vol. 219, p. 106798, 2022.
[19] P. Bhattacharya, A. Jain, and R. Singh, "IoMT-based diagnostic framework for real-time CKD prediction," IEEE Access, vol. 11, pp. 34560–34570, 2023.
[20] M. Srivastava, D. Kumar, and R. Kaushik, "Interpretable machine learning model using SHAP for CKD prediction," Appl. Soft Comput., vol. 138, p. 110256, 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY