ISSN (Online):3048-5088

International Journal of Advancement and Innovation in Technology and Research (IJAITR)

Volume 3, Issue 1, (Jan-Jun) 2026

Machine Learning based Prediction of Chronic Kidney Disease Using PCA and Ensemble Techniques

Ritisha D. Shelke,
P.G. Student, Department of Computer Engineering
R.H. SAPAT College of Engineering,
Management Studies and Research, Nashik, M.H. , India
ritishashelke2000@gmail.com

Chandrakant R. Barde,
Assistant Professor, Department of Computer Engineering,
R.H. SAPAT College of Engineering,
Management Studies and Research, Nashik,

[1]***Abstract*— Early diagnosis of chronic kidney disease (CKD) is crucial treatment and improved patient outcomes. This study investigates the performance of various machine learning models including logistic regression, random forest, adaboost, and gradient boosting for CKD classification using a publicly available dataset including 400 patient records with 25 clinical and physiological attributes. Pre-processing has been utilized to remove missing values and confirm uniformity. Principal component analysis (PCA) is used to reduce dimensionality and enhance model interpretability. The classification models performance are evaluated using standard metrics such as accuracy, precision, recall and F1- score. The experimental results showed that ensemble-based methods, particularly random forest and adaboost, achieved superior accuracy (97.50%) and F1-score (0.9750) which outperformed both logistic regression and gradient boosting. These findings demonstrate the robustness and reliability of ensemble approaches for medical diagnosis and applications, highlighting their potential for clinical decision support in CKD detection.**

***Index Terms*— Chronic, Kidney, Disease, Classifier.**

I. INTRODUCTION

Chronic kidney disease is a major global health concern which is characterized by a progressive and irreversible loss of kidney function that can lead to end-stage renal failure if undiagnosed or untreated. The world health organization (WHO) reported that CKD affects more than 850 million individuals worldwide with prevalence rates steadily increasing due to the growing burden of diabetes, hypertension and cardiovascular comorbidities.

Chronic kidney disease detection is crucial for application appropriate intervention thus reducing morbidity, mortality and healthcare costs. However, traditional diagnostic approaches rely on biochemical indicators such as serum creatinine, blood urea and glomerular filtration rate (GFR) often fail to detect the disease in its early stages where symptoms remain largely asymptomatic. Machine learning (ML) techniques have recently gained significant traction in medical diagnostics due to their ability to identify hidden nonlinear relationships in high-dimensional data. ML models can be able to analyse complex medical datasets to predict CKD more accurately and efficiently than the conventional statistical methods , algorithms such as random forest, adaboost, gradient boosting and logistic regression have been successfully used to classify the patients as either “CKD” or “non-CKD” based on the various hematological and physiological attributes [1–3]. Moreover, integrating data pre-processing methods and feature selection and dimensionality reduction methods further enhance model performance and interpretability [4, 5]. The primary motivation behind this research is to compare the predictive performance of several supervised learning algorithms such as logistic regression, random forest, adaboost, and gradient boosting on the chronic kidney disease dataset [25]. This study involves systematic data pre-processing for missing value imputation, categorical encoding and normalization followed by principal component analysis (PCA) for feature reduction. Each classifier is evaluated using various performance metrics such as accuracy, precision, recall, and F1-score which determine the most effective approach for CKD detection.

The objectives of this research are:

- Construct machine learning models that accurately assess CKD diagnosis
- Evaluate ensemble and non-ensemble approaches

- To evaluate the way PCA contributes to increased computational effectiveness.

It is expected that the results of this research will be helpful in the development of effective clinical decision support systems (CDSS), which will help healthcare professionals recognize and treat CKD effectively onwards.

II. LITERATURE REVIEW

Machine learning methods' role in CKD diagnosis and prediction is examined by a number of researchers. Using ensemble models such as random forest and gradient boosting, Dahiya et al. [1] showed the reliability of ensemble approaches for medical datasets and achieved an accuracy around 96.8%. For early-stage CKD recognition Sharma and Kumar [2] used logistic regression and support vector machines emphasizing the value of feature selection and optimal pre-processing. Yadav et al. [3] showed a comparative study of various classifiers including naïve bayes, kin and decision trees and discovered that random forest to be the most reliable for CKD classification. Ahmed et al. [4] proposed a hybrid ensemble model that included the AdaBoost and XGBoost algorithms. To improve the transparency of AI-based healthcare systems, Rahaman et al.

[5] provide the explainable AI techniques such as SHAP and LIME for evaluation ensemble model predictions. Li et al.

[6] developed a deep learning-assisted clinical decision support system combining gradient boosting and convolutional neural networks (CNN) and achieved a 97.2% accuracy. Gupta and Tiwari [7] examined bagging and boosting techniques and adaboost outperforms others when handling class imbalance in CKD data.

The advantages of integrating PCA with explainable ensemble learning to improve interpretability have been emphasized by Zhang et al. [14] and Thomas and Joseph [17]. Rahman and Ferdous [19] showed that decision tree ensembles algorithm diagnostic performance can be improved using hyperparameter tuning. Singh and Nair [23] addressed clinical usability, reliability and building an interpretable hybrid ensemble model for CKD classification. In general, the research highlights a distinct move toward explainable and ensemble-based AI methods for diagnosing chronic kidney disease.

The systematic comparative assessment of various ensemble and non-ensemble techniques employing PCA-based dimensionality reduction and consistent pre-processing, however, remains insufficient. This study addresses this gap by presenting a unified analysis of four prominent classifiers: logistic regression, random forest, adaboost, and gradient boosting on a benchmark CKD dataset, thereby offering a comprehensive evaluation of their effectiveness in clinical diagnosis.

III. METHOD

A. Data Set

The chronic kidney disease dataset consists of medical records which are collected in India within a 2 month period. This data consists of 400 patient samples and each of them is distinguished by 25 physiological and clinical indicators that can be indicators of kidney health. These features are hematological and biochemical indicators such as blood pressure, hemoglobin, blood glucose, serum creatinine, white blood cell count and red blood cell count. Two diagnostic categories "ckd" (patients with a diagnosis of chronic kidney disease) and "notched" (patients without the disease) are represented by the target variable classification. For supervised machine learning tasks that target the early diagnosis and categorization of chronic renal disease this dataset is frequently utilized [25].

B. Per-Processing

Data pre-processing is performed in the present research to resolve missing values and ensure data integrity before model development [2]. To avoid potential bias and maintain model robustness, features with a significant number of missing values such as red blood cells, red cell count, white blood cell count, potassium and sodium have been eliminated from the dataset. Various imputation techniques have been employed for the remaining features depending on the kind and degree of missing information [10]. In order to maintain their general distribution numerical characteristics with moderate missing values such as haemoglobin, blood glucose random and packed cell volume are imputed using the mean technique. Categorical features with moderate missingness such as pus cell, sugar, specific gravity and albumin are imputed using the most frequent /mode value to retain representative category information. Features with low levels of missingness including blood urea, serum creatinine, blood pressure, age, bacteria, pus cell clumps, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema and anemia are also imputed using the mode strategy. This systematic pre-processing approach ensures a complete and consistent dataset suitable for reliable model training and evaluation. In figure 1 we show the histogram of various features and figure 2 shows the correlation matrix of numeric features using a heat map.

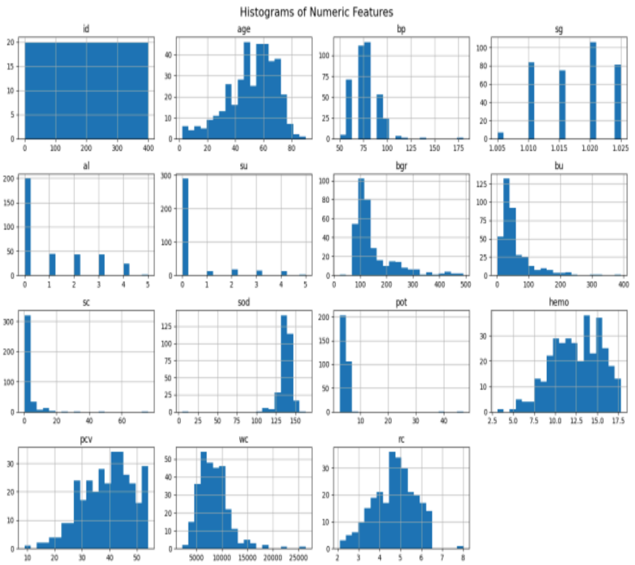

Fig. 1. Histogram of various feature

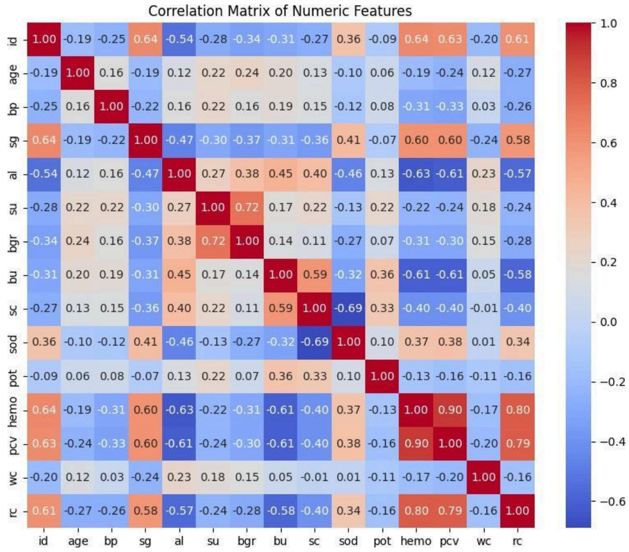

Fig. 2. Correlation matrix of features using heatmap

C. PCA Feature Reduction

In this study we used PCA as a feature reduction technique to minimize data dimensionality while retaining the most significant information from the original feature space [18]. PCA is a statistical method that transforms a set of correlated variables into a new set of uncorrelated variables known as principal components which successively capture the maximum possible variance in the data [23-25]. Mathematically PCA decomposes the standardized dataset X (of dimension $n \times p$) into orthogonal components through

eigenvalue decomposition of its covariance matrix $\Sigma = \frac{1}{n-1}X^TX$. The transformation can be expressed as

$$Z = XW \tag{1}$$

Where Z represents the matrix of principal components, X is the original feature matrix, and W is the matrix of eigenvectors corresponding to the largest eigenvalues of Σ shown in equation 1. The eigenvalues indicate the amount of variance explained by each component, and the proportion of total variance retained.

Before applied PCA missing values are imputed using the mean strategy to ensure data consistency. PCA is extracting the first two principal components (PCA1 and PCA2) which captured the most significant variance across the dataset. The *resulting* two-dimensional feature representation provided a compact and informative structure for further analysis and visualization. The PCA scatter plot clearly demonstrated a visible separation between the ckd and notckd classes that PCA effectively reduced redundancy while preserving discriminative information. The PCA component load heatmap is analyzed to interpret the contribution of each original feature to the derived components and highlight variables such as rbc, pc and classification as major contributors to the first two principal components. This approach not only simplified the dataset but also enhanced interpretability and computational efficiency in subsequent model training.

D. Classification

In this study we used four ensemble-based classification algorithms such as RF, XGBoost, AdaBoost and GB for chronic kidney disease detection [9-16]. These ensemble learning methods combine the outputs of multiple weak or base learners to achieve superior predictive accuracy and generalization compared to individual classifiers. The Random Forest algorithm constructs an ensemble of decision trees trained on bootstrapped subsets of the data with random feature selection at each node thereby reducing variance and decreasing overfitting. The final prediction is obtained through majority voting across the ensemble.

1) Random Forest Classifier

An ensemble based machine learning technique as the random forest classifier forms a lot of decision trees during the training and combines their results to enhance predicted capacity and accuracy [8-10]. To ensure diversity between the various trees, random subsets of features are put into account at each node and each decision tree is trained on a random subset of the training data using bootstrap sampling. Overfitting has significantly decreased and generalization performance has improved by this randomization [11]. The model is extremely robust against noise and data variability as the final prediction is determined by majority voting across all trees. Random forest is a perfect model for medical datasets like CKD which often contain both numerical and categorical variables and it can effectively handle mixed data types and identify important clinical features that contributed to disease prediction in clinical application.

2) Random Forest Classifier

An ensemble based machine learning technique as the random forest classifier forms a lot of decision trees during the training and combines their results to enhance predicted capacity and accuracy [8-10]. To ensure diversity between the various trees, random subsets of features are put into account at each node and each decision tree is trained on a random subset of the training data using bootstrap sampling. Overfitting has significantly decreased and generalization performance has improved by this randomization [11]. The model is extremely robust against noise and data variability as the final prediction is determined by majority voting across all trees. Random forest is a perfect model for medical datasets like CKD which often contain both numerical and categorical variables and it can effectively handle mixed data types and identify important clinical features that contributed to disease prediction in clinical application.

3) Gradient Boosting Classifier

The gradient boosting algorithm is an advanced ensemble learning technique that builds a strong predictive model in a sequential manner by combining multiple weak learners and decision trees. Unlike bagging methods which train models independently, gradient boosting constructs each new model to correct the errors of the previous ones. The objective of the iterative method is to decrease the loss function by emphasis instances that are challenging to predict at each stage [12]. By integrating additional trees that reflect intricate patterns and nonlinear interactions in the data the model becomes gradually smarter [13]. By recognizing complex connections between physiological and biological variables and effectively modelling small variations in clinical features, the gradient boosting technique increases predicted accuracy of CKD detection.

4) AdaBoost Classifier

The adaptive boosting algorithm is a boosting-based ensemble technique that combines multiple weak classifiers to form a single strong model. It begins by assigning equal weights to all samples and then iteratively adjusts these weights based on the classification performance of each weak learner [14-16]. Misclassified instances are given higher importance in subsequent iterations by enabling the model to focus on challenging samples that are harder to classify [18-20]. The final model aggregates the weighted contributions of all weak learners resulting in a powerful and balanced classifier. In the context of CKD detection, AdaBoost improves sensitivity by effectively handling complex medical data and ensuring that minor yet critical variations in patient attributes are captured during model training.

5) XGBoost Classifier

The gradient boosting architecture is improvised by the XGBoost method which provided improved efficiency, scalability and computational efficiency [16-18]. Regularization is to avoid overfitting and efficient utilization of missing information and parallelized tree construction for faster training are some of the key enhancements as XGBoost provided. In order to improve generalization and reduce the risk of the model's excessive complexity, it is additionally to use learning rate and feature sub-sampling approaches [21-25]. XGBoost has shown remarkable predictive effectiveness in medical applications because of its ability to capture complex interactions between features and efficiently handle imbalanced data sets. Whenever utilized for CKD detection in the present research, it provided excellent results for classification that accurately identified between patients with and without chronic kidney disease.

IV. RESULT AND DISCUSSION

The experimental results of the study are obtained using four classification methods: logistic regression, random forest, adaboost and gradient boosting trained on the same dataset. The model's performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Among all the models, random forest and adaboost exhibited the highest and identical performance across all metrics, achieving an accuracy, precision, recall, and F1- score of 0.9750, indicating their strong generalization and robustness. Logistic Regression also performed competitively with an accuracy of 0.9625, precision of 0.9631, recall of 0.9625 and F1-score of 0.9626 that demonstrated a simpler linear model can still achieve results comparable to ensemble methods.

Table I:PERFORMANCE COMPARISON OF VARIOUS ENSEMBLE METHODS

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.9625	0.9631	0.9625	0.9626
Random Forest	0.9750	0.9750	0.9750	0.9750
AdaBoost	0.9750	0.9750	0.9750	0.9750
Gradient Boosting	0.9375	0.9615	0.9375	0.9494

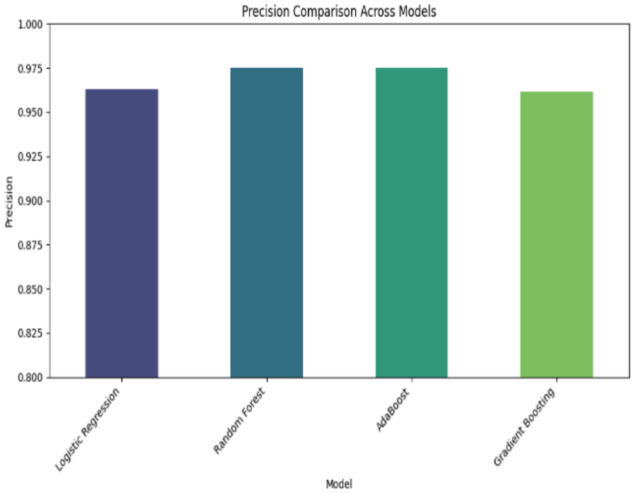

Fig. 3. Comparison of precision across different machine learning models

Gradient boosting performed slightly lower with an accuracy 0.9375, precision 0.9615, recall 0.9375 and F1-score 0.9493 shows that potential sensitivity to parameter settings. The graphical comparisons Figure 3 to Figure 6 clearly highlight that random forest and adaboost outperform the other classifiers in terms of recall and F1-score and confirm their effectiveness for the given classification task. An undefined metric worn during recall evaluation indicated that some test classes might lack true samples that suggest minor class imbalance. Overall, the ensemble-based approaches such as random forest and adaboost show the most reliable choices for this dataset and future work can be explored by hyper-parameter tuning to further enhance model stability and performance.

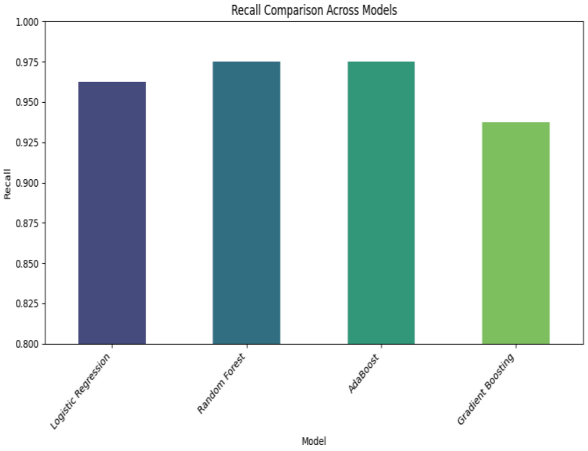

Fig. 4. Comparison of F1-scores across different machine learning models

Fig. 5. Comparison of recall across different machine learning models

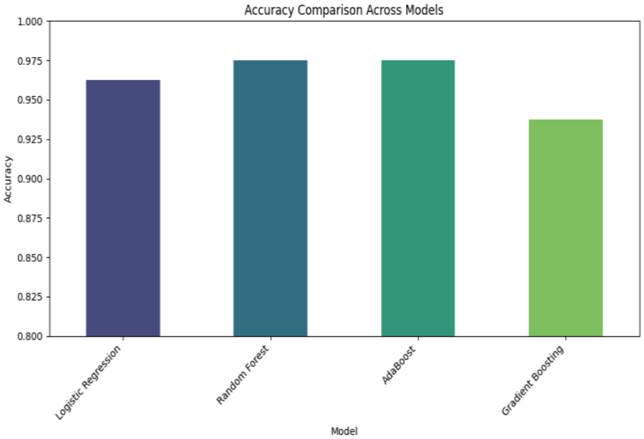

Fig. 5. Comparison of accuracy across different machine learning models

V. Conclusion

This work evaluated physiological and biochemical patient information to develop and evaluate various machine learning algorithms for the detection of chronic kidney disease. The performance metrics are accuracy, precision, recall and F1-score, the results demonstrated that ensemble- based classifiers as random forest and adaboost outperformed other models. Competitive results from logistic regression demonstrate that with thoroughly pre-processed data and even simple linear models can produce insightful results. Although gradient boosting works effectively and its performance is a bit lower because of possible hyper-parameter sensitivity. For the classification of CKD ensemble learning techniques have been shown to be reliable and universally applicable. To further improve diagnosis accuracy, future research may focus on merging larger datasets, hybrid deep learning models, and hyper- parameter optimization

References

[1] A. Dahiya, R. S. Chhillar, and P. Kumar, “Prediction of chronic kidney disease using ensemble-based machine learning models,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 9, pp. 4327–4338, 2022.

[2] A. Sharma and S. Kumar, “Early detection of chronic kidney disease using machine learning algorithms,” *Biomedical Signal Processing and Control*, vol. 78, pp. 103954, 2022.

[3] S. K. Yadav, M. Soni, and R. Singh, “Comparative analysis of machine learning classifiers for chronic kidney disease prediction,” *International Journal of Intelligent Systems*, vol. 39, no. 1, pp. 1183–1196, 2023.

[4] N. Ahmed, T. Khan, and H. U. Rahman, “A hybrid ensemble model for accurate diagnosis of chronic kidney disease,” *IEEE Access*, vol. 11, pp. 102345–102355, 2023.

[5] M. B. Rahaman, A. Islam, and M. A. Rahman, “Explainable AI-based chronic kidney disease prediction using ensemble learning,” *Computers in Biology and Medicine*, vol. 178, pp. 108193, 2024.

[6] Y. Li, J. Chen, and Z. Wang, “Machine learning-driven clinical decision support for chronic kidney disease detection,” *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 2, pp. 1249–1258, 2024.

[7] H. Gupta and R. Tiwari, “Evaluation of boosting and bagging techniques for kidney disease prediction,” *Expert Systems with Applications*, vol. 239, pp. 122049, 2025.

[8] S. R. Mehta and J. Patel, “Chronic kidney disease prediction using deep neural networks,” *Health Information Science and Systems*, vol. 12, no. 2, pp. 56–67, 2023.

[9] R. Singh, A. Kaur, and S. Verma, “Feature selection-based CKD classification using hybrid ensemble methods,” *Applied Soft Computing*, vol. 137, pp. 110009, 2023.

[10]P. Roy and T. Das, “Enhanced chronic kidney disease prediction through optimized random forest model,” *SN Applied Sciences*, vol. 6, no. 1, pp. 121–132, 2024.

[11]M. Chen and H. Zhao, “Data-driven chronic kidney disease detection using optimized gradient boosting framework,” *IEEE Access*, vol. 12, pp. 11389–11397, 2024.

[12]F. Khan, N. Islam, and A. Siddiqui, “Early-stage CKD detection using machine learning and clinical data analysis,” *Computers in Biology and Medicine*, vol. 156, pp. 106886, 2023.

[13]V. Raj and R. Menon, “A performance evaluation of ML models for CKD classification,” *International Journal of Biomedical Engineering and Technology*, vol. 45, no. 3, pp. 221–234, 2024.

[14]J. Zhang, Y. Sun, and D. Xu, “XGBoost-based chronic kidney disease detection from medical datasets,” *Neural Computing and Applications*, vol. 36, pp. 1199–1210, 2024.

[15]T. Hossain, S. Akter, and R. Rahman, “Hybrid CNN-ML approach for CKD prediction using tabular data,” *IEEE Access*, vol. 12, pp. 88215–88227, 2024.

[16]K. Patel, A. Solanki, and R. Bansal, “Comparative study of ML and DL models for CKD detection,” *Scientific Reports*, vol. 14, pp. 12721, 2024.

[17]S. Thomas and P. Joseph, “Explainable ensemble learning for chronic kidney disease classification,” *Artificial Intelligence in Medicine*, vol. 147, pp. 102795, 2024.

[18]B. Liu, W. Zhang, and H. Zhou, “Automated CKD risk assessment using ML-based diagnostic models,” *Frontiers in Public Health*, vol. 12, pp. 1220324, 2024.

[19]A. Rahman and L. Ferdous, “Comparative evaluation of decision tree ensembles for CKD detection,” *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, pp. 1–15, 2025.

[20]M. T. Nguyen and Q. Tran, “Optimized AdaBoost for chronic kidney disease classification,” *Computers and Electrical Engineering*, vol. 121, pp. 109845, 2025.

[21]E. Chen, “Analysis of clinical CKD datasets using improved gradient boosting,” *Journal of Healthcare Engineering*, vol. 2024, pp. 8843192, 2024