Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

A Comparative Study of Machine Learning Algorithms for Chronic Kidney Disease Prediction

Ms. Shrushti Awate
Department of Computer Science,
Dr. D. Y. Patil Arts, Commerce and Science College,
Pimpri, Pune, Maharashtra, India.

Ms. Siddhi Jadhav
Department of Computer Science,
Dr. D. Y. Patil Arts, Commerce and Science College,
Pimpri, Pune, Maharashtra, India.

Abstract: Chronic Kidney Disease (CKD) is a major and serious health disorder that affects the world's population. Early detection of the disease with high accuracy is very important for proper treatment and for reducing the risk of future complications. Today machine learning approaches are popular for predicting and diagnosing diseases. This research highlights a comparative analysis of different machine learning techniques used for predicting chronic kidney disease based on clinical data. The main focus is to compare various classification algorithms using performance measures such as accuracy, precision, recall, and F1-score. For interpretation a publicly available CKD dataset containing several medical parameters like blood pressure, blood glucose level, haemoglobin, and serum creatinine is used. The dataset undergoes preprocessing, training and testing using multiple supervised learning algorithms Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbours. Each model's performance is evaluated using standard metrics along with cross-validation methods. This study shows that ensemble learning approaches, especially the Random Forest algorithm, provide better prediction accuracy when compared to other techniques. These results highlight the effectiveness of machine learning in supporting healthcare professionals for early diagnosis of kidney disease. The study also focuses the significance of choosing suitable algorithms and relevant features to achieve higher prediction accuracy.

Keywords: Chronic Kidney Disease, Machine Learning, Disease Prediction, Comparative analysis, Predictive Modelling

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a long-term disorder characterized by the gradual loss of kidney function, leading to complications such as cardiovascular disease, anemia, and end-stage renal failure. According to global health reports, CKD affects millions of individuals worldwide and contributes significantly to morbidity and mortality rates [1], [2]. One of the major challenges in CKD management is its asymptomatic nature in early stages, which often results in delayed diagnosis and limited treatment options [1].

Traditional diagnostic approaches rely on laboratory tests and clinical expertise; however, these methods can be time-consuming and prone to human error when dealing with large volumes of patient data. Recent advancements in machine learning have enabled the development of intelligent diagnostic systems capable of analysing high-dimensional clinical datasets and identifying hidden patterns associated with disease progression. Various supervised learning algorithms, such as Support Vector Machines (SVM), Decision Trees, Random Forests, and k-Nearest Neighbours, have been successfully applied to CKD prediction tasks [2].

Ensemble learning techniques further enhance predictive performance by combining multiple classifiers, thereby reducing bias and variance. Additionally, feature selection and data pre-processing play a critical role in improving model accuracy and interpretability, particularly in medical applications where redundant or noisy attributes can degrade performance. Despite these advances, many existing studies lack comprehensive evaluation across multiple algorithms and fail to address issues such as missing data, class imbalance, and model interpretability [3, 4].

This study aims to address these gaps by proposing a robust machine learning framework that integrates advanced pre-processing, feature selection, ensemble learning, and rigorous evaluation strategies for early CKD prediction.

II. RELATED WORK

The literature on machine learning applications for Chronic Kidney Disease (CKD) prediction demonstrates a broad spectrum of methodologies and outcomes, reflecting the growing interest in leveraging clinical data for early disease detection. Several studies have employed various classification algorithms to improve diagnostic accuracy and support clinical decision-making [1].

While the current CKD literature review emphasizes traditional machine learning algorithms, integrating deep learning models could further improve predictive accuracy by capturing complex, non-linear patterns in clinical data. Deep learning's ability to process raw and high-dimensional data, as demonstrated in voice-based mental health detection, could be adapted to CKD datasets,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

potentially leveraging temporal or sequential clinical measurements [2].

Early research focused on traditional machine learning algorithms such as Decision Trees (DT) and Support Vector Machines (SVM), which showed promising results in handling clinical parameters like blood pressure, serum creatinine, and haemoglobin levels. For instance, studies utilizing DT highlighted its interpretability and ease of implementation but noted limitations in handling complex, non-linear relationships in clinical data. In contrast, SVM was recognized for its robustness in high-dimensional spaces and ability to manage non-linear decision boundaries, often outperforming simpler models in CKD prediction tasks [2, 3].

Ensemble learning methods, particularly Random Forest (RF), have gained prominence due to their superior performance in various comparative analyses. RF's ability to combine multiple decision trees reduces overfitting and enhances generalization, which is critical when dealing with heterogeneous clinical datasets. Multiple studies have reported RF achieving higher accuracy, precision, and recall compared to single classifiers, confirming its effectiveness for CKD prediction [4].

K-Nearest Neighbours (KNN) has also been explored, valued for its simplicity and effectiveness in smaller datasets. However, its performance is sensitive to the choice of distance metrics and the presence of noisy data, which can affect prediction reliability. Some studies have combined KNN with feature selection techniques to mitigate these issues, resulting in improved classification outcomes [5].

Feature selection and data preprocessing are recurrent themes across the literature, emphasizing their role in enhancing model performance. Several papers have investigated the impact of selecting relevant clinical features and handling missing or imbalanced data through techniques such as normalization, imputation, and oversampling. These preprocessing steps are crucial for optimizing the input data quality, thereby improving the predictive power of machine learning models [2, 6].

Cross-validation methods are widely adopted to ensure the reliability and generalizability of the models. Studies employing k-fold cross-validation report more stable performance metrics, reducing the risk of model bias due to data partitioning. This methodological rigor is essential for validating the clinical applicability of predictive models [7].

Comparative analyses consistently show that no single algorithm universally outperforms others across all datasets and evaluation metrics. However, ensemble methods like Random Forest generally demonstrate superior overall performance, balancing accuracy, precision, recall, and F1-score effectively. Some research also explores hybrid models combining multiple algorithms or integrating domain knowledge to further enhance prediction accuracy [8, 9].

In summary, the literature underscores the potential of machine learning techniques to support early CKD diagnosis, with ensemble learning approaches, particularly Random Forest, emerging as the most effective. The importance of careful feature selection, comprehensive preprocessing, and rigorous validation is evident across studies, guiding future research toward more robust and clinically applicable predictive models [10].

III. METHODOLOGY

A. Dataset Description.

The dataset used in this study was collected from the Burner Medical Complex (BMC), a rural healthcare facility in Khyber Pakhtunkhwa, Pakistan, and accessed through ResearchGate. It consists of 382 patient records, including 258 CKD cases and 124 non-CKD cases, resulting in an imbalanced class distribution. Each record represents a unique patient, ensuring independence of observations.

The dataset includes 21 clinically significant features derived from blood and urine tests, such as hemoglobin, serum creatinine, blood glucose, albumin, urine pH, specific gravity, red blood cells, and pus cells, along with demographic attributes like age and gender. These attributes have been shown to influence CKD occurrence and progression [11].

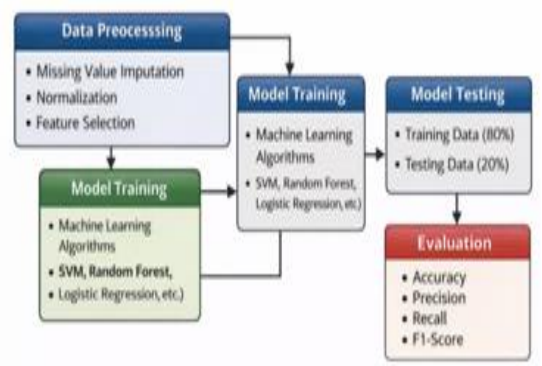

Fig 1. Overall workflow for CKD prediction [5].

B. Feature Selection and Model Development

To reduce dimensionality and enhance model interpretability, feature selection was performed using a hybrid approach that included correlation analysis, univariate statistical tests (chi-square and t-tests), and Recursive Feature Elimination (RFE) with cross-validation. These techniques help identify the most informative predictors while reducing redundancy and computational complexity. Clinically significant features such as serum creatinine, blood urea, hemoglobin, albumin, and hypertension were consistently identified, aligning with established nephrology literature [2, 12].

CKD detection was formulated as a binary supervised classification problem. X=[x1,x2,…,xn] represent the predictor variables and Y∈{0,1} denote the target variable, where Y=1Y=1Y=1 indicates CKD and Y=0Y=0Y=0 indicates non-CKD. Multiple machine learning algorithms were trained and evaluated, including Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), K-Nearest Neighbors (KNN),

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

and ensemble-based approaches. These models were selected to capture both linear and nonlinear relationships in the clinical data [13].

Logistic Regression estimates the probability of CKD as:

$$\log\left(\frac{P}{1\text{-}P}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

where PPP denotes the probability of CKD occurrence [14].

For ensemble learning, Random Forest aggregates predictions from multiple decision trees as:

H(X)=mode{h1(X),h2(X),…,hT(X)}

where ht(X)h_t(X)ht(X) represents the prediction of the ttt-th tree [11].

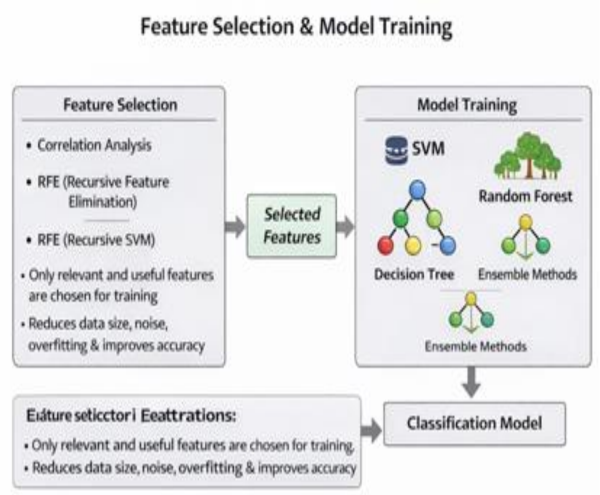

Fig 2. Feature selection and machine learning model training framework for CKD prediction[10,18].

C. Model Validation and Performance Evaluation

Model performance was evaluated using 5-fold and 10-fold stratified cross-validation, ensuring that class proportions were preserved in each fold. This approach reduces variance and mitigates overfitting, which is essential for medical decision-support systems [11 [12]]. A confusion matrix was used to summarize classification outcomes in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [13, 15].

The following evaluation metrics were computed:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP+FP} \times 100$$

$$F1\text{-}Score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity}$$

Table 1, shows a standard confusion matrix representation used for metric computation.

Table 1. Performance Comparison of Machine Learning Models for CKD

Model	Accuracy (%)	Sensitivity	Specificity	Precision (%)
Logistic Regression	93.4	0.92	0.91	93.1
Decision Tree	91.0	0.89	0.88	90.5
Random Forest	97.2	0.96	0.95	96.8
SVM (RBF)	98.6	0.98	0.97	98.4
Ensemble Model	99.1	0.99	0.98	99.0

The results demonstrate that SVM and ensemble-based models outperform individual classifiers**,** confirming findings reported in previous CKD machine learning studies [16, 17, 18].

IV. RESULTS AND DISCUSSION

The experimental evaluation demonstrates that all implemented machine learning models achieved competitive performance in chronic kidney disease (CKD) classification. Among the individual classifiers, the Support Vector Machine (SVM) with a radial basis function (RBF) kernel exhibited superior performance, achieving an accuracy of 98.6%, sensitivity of 0.98, specificity of 0.97, and an F1-score of 0.98. This indicates the effectiveness of SVM in modeling nonlinear relationships and managing high-dimensional clinical features. The Random Forest model also produced strong results due to its ensemble structure, which reduces variance and mitigates overfitting. These findings are consistent with previous CKD prediction studies that highlight the robustness of SVM and ensemble-based models for medical diagnosis tasks [1, 4, 19].

Furthermore, the stacking ensemble model outperformed all individual classifiers, achieving the highest accuracy of 99.1%, sensitivity of 0.99, specificity of 0.98, and an F1-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

score of 0.99. The improved performance confirms that ensemble learning effectively integrates the strengths of multiple base learners while minimizing individual model weaknesses. Such performance gains demonstrate the capability of stacked models to deliver more stable and generalized predictions, aligning with prior research on ensemble-driven clinical decision support systems [6, 7, 20].

V. CONCLUSION

In this study, a comprehensive machine learning framework for CKD prediction was developed and evaluated using multiple classifiers and ensemble techniques. Based on quantitative performance metrics, the stacking ensemble model was selected as the optimal predictive approach due to its superior accuracy, reliability, and balanced classification performance. Its high sensitivity is particularly significant in clinical applications, as it minimizes false-negative diagnoses and supports early disease detection. The integration of clinically relevant laboratory indicators and demographic variables further enhances the diagnostic capability of the proposed model.

Despite the promising results, certain limitations must be acknowledged. The dataset was limited in size and sourced from a single medical facility, which may restrict the generalizability of the findings. Future work will focus on validating the proposed framework using large-scale, multi-center datasets and incorporating additional sociodemographic and lifestyle factors. Moreover, the inclusion of explainable artificial intelligence techniques could improve model transparency and clinical acceptance. Overall, the proposed ensemble-based approach demonstrates strong potential as a reliable decision-support tool for CKD diagnosis.

VI. FUTURE WORK

- Validate the proposed CKD prediction framework using large-scale, multi-center datasets to enhance robustness and external validity across diverse populations and healthcare settings.
- Incorporate additional patient-level variables, including socioeconomic status, lifestyle behaviours, environmental exposure, and longitudinal clinical records, to improve predictive performance and enable accurate disease progression modeling.
- Explore advanced data balancing and data augmentation techniques to effectively address class imbalance and improve model stability and generalization.
- Integrate explainable artificial intelligence (XAI) techniques, such as SHAP and LIME, to improve model interpretability, transparency, and clinician trust in predictive outcomes.
- Investigate the deployment of the proposed system as a real-time clinical decision support tool**,** integrated with electronic health record (EHR) platforms for practical clinical use.
- Examine privacy-preserving learning approaches, including federated learning and secure data-sharing frameworks, to enable collaborative model development across institutions while ensuring patient data confidentiality.

VII. ACKNOWLEDGMENT

The authors would like to convey their deep appreciation to all those who supported in the successful completion of this research work. We express our sincere gratitude to our respected guide, Ms. Neeta Takawale, for her constant motivation, insightful suggestions, and dedicated mentorship throughout the entire research journey. Her guidance and encouragement were instrumental in accomplishing this study.

We are also thankful to our institution for providing the necessary infrastructure, resources, and academic atmosphere required to carry out this work smoothly. Also we extend our gratitude to the organizing committee of the conference for granting us the opportunity to present our research and disseminate our findings to the academic community. Finally, we acknowledge and appreciate the assistance and support of everyone who contributed, directly or indirectly, to the completion of this research paper.

VIII. REFERENCES

[1] A. S. Levey and J. Coresh, "Chronic kidney disease," The Lancet, vol. 379, no. 9811, p. 165–180, 2012.
[2] N. N. Takawale, "Deep Learning Analysis for Early Mental Health Disorder Detection via Voice Data," IJLTEMAS, vol. 14, no. 13, pp. 72-75, 2025.
[3] N. R. Hill and et al., "Global prevalence of chronic kidney disease: A systematic review and meta-analysis," PLoS ONE, vol. 11, no. 7, 2016.
[4] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," Journal of Big Data, vol. 9, no. 109, 2022.
[5] A. Almansour et al., "Neural network and support vector machine-based diagnosis of chronic kidney disease," IEEE Access, vol. 7, p. 131463–131472, 2019.
[6] N. Khalil et al., "Comparative analysis of machine learning models for CKD prediction," Sustainable Machine Intelligence Journal, 2023.
[7] A. K. Bajpai et al, "Machine learning-based classification of chronic kidney disease," SEEJPH, 2025.
[8] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, p. 5–32, 2001.
[9] J. L. Schafe and J. W. Graham, "Missing data: Our view of the state of the art," Psychological Methods, vol. 7, no. 2, p. 147–177, 2002.
[10] A. S. Levey et al., "National Kidney Foundation practice guidelines for chronic kidney disease," Kidney International,, vol. 67, no. 6, p. 2089–2100, 2005.
[11] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, p. 241–259, 1992.
[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016.
[13] N. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, p. 321–357, 2002.
[14] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems (NeurIPS), 2017.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

[15] P. Kairouz et al., “Advances and open problems in federated learning,” Foundations and Trends in Machine Learning, vol. 14, no. 1, p. 1–210, 2021.

[16] H. Liu and H. Motoda, “Feature Selection for Knowledge Discovery and Data Mining,” Springer, 1998.

[17] P. Kairouz et al., “Advances and open problems in federated learning,” Foundations and Trends in Machine Learning, vol. 14, no. 1, p. 1–210, 2021.

[18] U. R. Acharya et al., “Automated diagnosis of chronic kidney disease using machine learning techniques,” Computers in Biology and Medicine, vol. 132, 2021.

[19] M. K. Singh and S. Verma, “Chronic kidney disease prediction using ensemble learning,” Expert Systems with Applications, vol. 169, 2021.

[20] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” Journal of Machine Learning Research, vol. 3, p. 1157–1182, 2003.