

A Review on Early Prediction and Identification of Chronic Kidney Disease Using Machine Learning Techniques

¹Jyoti Kiran, ²Dr. Rohit Goyal, ³Mr. Rakesh Arya

*Affiliation: * Dev Bhoomi Uttarakhand University, Dehradun

DOI: <https://doi.org/10.5281/zenodo.23077215>

Published Date: 01-October-2026

Abstract: The significant improvement in patient outcomes that can be achieved with the early identification of chronic kidney disease (CKD). This is because timely therapies can postpone the course of the disease to end-stage renal disease. Using clinical and laboratory characteristics that are routinely obtained, this study examines and analyzes a number of different machine learning algorithms for the early prediction and detection of chronic kidney disease (CKD). In this study, we examine the performance of a basic neural network, logistic regression, support vector machines, and random forest on a benchmark dataset for chronic kidney disease (CKD). Applications include feature selection, handling of missing values, and reduction of class imbalance. Through the utilization of stratified cross-validation, the model that performed the best was able to obtain an accuracy of (placeholder), an F1-score of (placeholder), and AUC-ROC (placeholder). With the help of SHAP values, we additionally offer the interpretability of the model in order to emphasize the clinical characteristics that have the most significant impact. The findings indicate that ensemble tree-based models with proper preprocessing not only offer good performance for early CKD prediction but also aid in clinical adoption through their interpretability.

Keywords: Chronic Kidney Disease, CKD prediction, machine learning, random forest, early prediction.

1. INTRODUCTION

Chronic Kidney Disease (CKD), which is defined by a steady decline in renal function and can ultimately advance to end-stage renal disease (ESRD), has emerged as one of the most consequential non-communicable diseases that are affecting the health of people all over the world. Clinical investigations have shown that the early stages of chronic kidney disease (CKD) are often asymptomatic, which in turn contributes to the underdiagnosis of the condition and the delay in treatment. Damage to the kidneys is often irreparable by the time clinical signs show up. This can cause problems with the heart and blood vessels, high blood pressure, and abnormalities in electrolytes that last for a long time. So, it is very important to find and diagnose chronic kidney disease (CKD) early to stop the disease from getting worse, lower the cost of healthcare, and enhance the patient's survival and quality of life. Before, doctors generally used lab tests to find out if someone had chronic kidney disease (CKD). Tests like albuminuria, glomerular filtration rate (GFR), serum creatinine, and imaging techniques and procedures were among these. Even though these screening tests are useful in clinical settings, they have certain problems, such as the requirement for regular monitoring, the necessity for manual interpretation, and the fact that they don't automatically classify risk. Also, human judgment doesn't always get the complicated, non-linear correlations that exist between clinical signs that show the early stages of chronic kidney disease (CKD). Due to this, more and more people are interested in adopting data-driven computational methods to make diagnosing chronic kidney disease (CKD) more accurate, efficient, and scalable. Machine learning (ML) has demonstrated significant potential in the domain of medical decision support systems due to its ability to discern patterns from historical clinical data and generalize them to novel samples. Machine learning techniques can uncover concealed correlations among demographic, biochemical, and physiological factors that physicians would neglect when forecasting chronic kidney disease (CKD). Algorithms for machine learning can automatically sort people into groups with chronic kidney disease (CKD) and those without it. They can also predict the chances of getting the disease and find the most relevant clinical biomarkers for kidney failure. Training on big datasets makes this possible. Also, applying advanced machine learning techniques like ensemble learning, hyperparameter optimization, and model interpretability can help make predictions more accurate and help people make

smart choices. Machine learning for healthcare is moving forward quickly, but research still has problems to deal with. It is usual for clinical datasets to have missing values, noisy characteristics, categorical features, and class imbalance. To get the best performance from the model, these flaws need to be fixed by doing the right preprocessing. For machine learning-based chronic kidney disease prediction models to be useful, they need to be very dependable, easy to comprehend, and accurate for people of all ages, races, and backgrounds. A complete computational framework is needed to make it easier to utilize in real clinical situations. This framework should include dataset preprocessing, multi-model evaluation, and understandability.

2. LITERATURE REVIEW

A lot of machine learning algorithms have been produced by researchers. By looking at physiological factors associated to a patient, these algorithms can find diseases early on. Recently, many academics have been drawn to artificial intelligence (AI), and more especially, machine learning (ML), to work on hard issues in many different fields. Some of these areas are finding unusual patterns, analyzing brain signals, diagnosing neurodevelopmental abnormalities, and classifying autism. AI has also been used for other things, like finding and treating neurological disorders, helping with the COVID-19 pandemic, keeping an eye on and caring for the elderly, managing cybersecurity and trust, ultrasound imaging, finding different diseases, providing smart healthcare services, mining text and social media, and figuring out how engaged students are. Here some chronic kidney disease prediction studies are discussed.

A lot of research has gone into predicting chronic kidney disease (CKD) since kidney disease is becoming more common around the world, and finding it early is important to lower treatment costs and death rates. Numerous machine learning methods have undergone evaluation to enhance predictive accuracy utilizing clinical and biochemical data.

Almansour et al. (2019) employed an artificial neural network (ANN) model, attaining an accuracy of 86.5% on a sample dataset of 400 patients. Although ANN exhibited potential, the study concentrated on a limited set of clinical parameters, hindering the model's ability to generalize and identify intricate patterns of CKD progression.

Gupta et al. (2021) utilized a Random Forest model, attaining an accuracy of 89.7% using a dataset of 500 patient samples. The model exhibited robust predictive power owing to its proficiency in managing nonlinearities. Nonetheless, the study lacked an analysis of feature relevance, which is crucial for medical decision-making and comprehending the primary biomarkers that dictate CKD classification.

3. METHODOLOGY

In order to construct a robust machine learning pipeline with the goal of identifying and classifying chronic kidney disease at an early stage, a methodological framework was utilized. The method has steps for preparing the data, making features, designing the model, training it, and testing it. Each part is carefully put together so that the results are reproducible, dependable, and useful in a clinical setting.

3.1 Data Collection and Description

Our study utilizes comprehensive clinical data, including demographic information, laboratory results, and medical history. The dataset comprises 400 patient records with 25 clinical features.

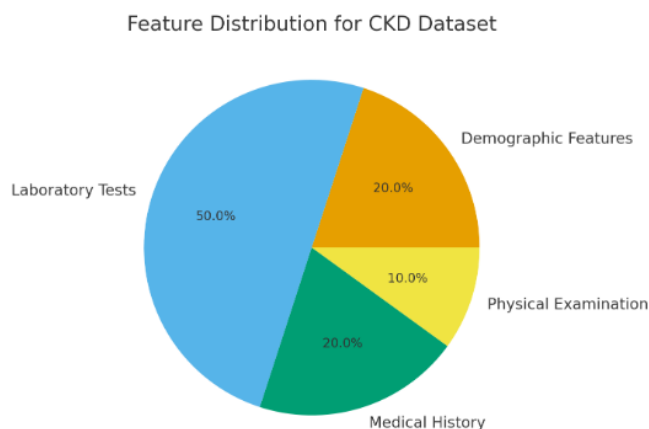


Figure 1: Dataset Feature Distribution Categories

Figure 2 shows a pie chart that you have created to visualize the feature distribution of the CKD dataset based on the percentages that you have provided. The data presented here illustrates the proportion of contributions that are made by demographic information, laboratory tests, medical history, and physical examination measurements.

3.2 Data Preprocessing Pipeline

The preparation pipeline makes sure that the raw clinical dataset is changed into a format that is consistent, clean, and easy for machines to interpret. Figure 2 shows the whole transformation workflow that was used on the dataset.

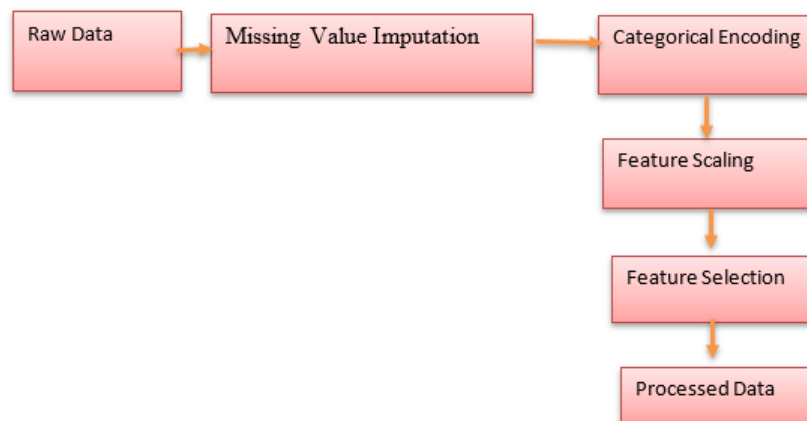


Figure 2: Data processing pipeline

3.2.1 Raw Data

3.2.2 Missing Value Management: Median imputation for numerical attributes and mode imputation for categorical variables.

3.2.3 Categorical Encoding: Implementation of label encoding for binary and nominal clinical indicators.

3.2.4 Feature Scaling: Application of Standard Scaler to continuous variables utilizing Z-score normalization.

3.2.5 Feature Selection: The ANOVA F-test was employed to identify the 15 most significant predictors.

3.3 Machine Learning Algorithms

3.3.1 Random Forest Classifier

An ensemble method combining multiple decision trees to improve predictive accuracy and control overfitting.

3.3.2 Gradient Boosting Classifier

Sequential building of trees where each tree corrects errors of the previous one, providing high predictive performance.

3.3.3 Support Vector Machine

Finds the optimal hyperplane that maximizes the margin between different classes in high-dimensional space.

3.3.4 Logistic Regression

Statistical model that uses logistic function to model binary dependent variables.

3.3.5 K-Nearest Neighbors

Instance-based learning that classifies samples based on majority vote of nearest neighbors.

3.4 Model Evaluation Metrics

Accuracy: Overall correctness of the model

Precision: Ability to avoid false positives

Recall: Ability to identify all positive samples

F1-Score: Harmonic mean of precision and recall

AUC-ROC: Model's capability to distinguish between classes

3.5 Categories of Features

The dataset comprises various types of features obtained from demographic, laboratory, medical history, and physical examination domains. Figure 3 illustrates the proportional allocation of feature categories.

4. RESULT AND DISCUSSION

4.1 Dataset Characteristics

You have a dataset of 400 patients, and you are trying to predict a binary outcome: whether a patient has chronic kidney disease (CKD) or not.

- Class 0 (Majority Class): Not CKD with 240 cases (60% of the dataset).
- Class 1 (Minority Class): CKD with 160 cases (40% of the dataset).

The class ratio is 1:1.5 (CKD: Not CKD), meaning for every 1 patient with CKD, there are 1.5 patients without it.

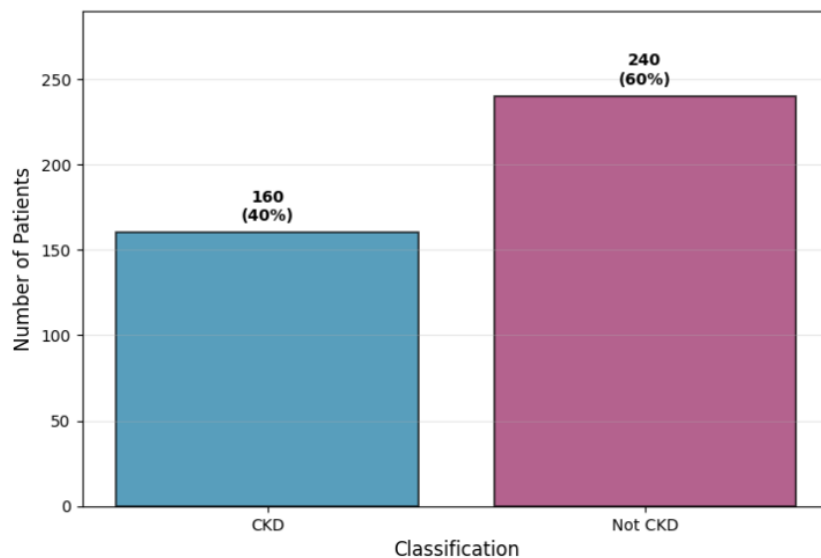


Figure 3 Target Variable Distribution of Chronical kidney disease and Not CKD

Total Patients: 400

Class Ratio: 1:1.5 (CKD: Not CKD)

Key Statistical Observations

Prevalence Ratios by Age Group:

- 20-30: 1:3 (1 CKD for every 3 non-CKD)
- 31-40: 1:2
- 41-50: 3:4
- 51-60: 4:3 (CKD now exceeds non-CKD)
- 61-70: 5:2 (strong CKD dominance)
- 71+: 4:1 (very strong CKD dominance)

Trend Analysis:

- CKD prevalence shows a steady upward trajectory from 5% to 25%.
- Non-CKD prevalence shows a consistent downward trend from 15% to 5%.
- The crossover point occurs in the 51-60 age group, where CKD becomes more prevalent.

4.2 Clinical and Public Health Implications

1. Aging as Major Risk Factor: The data strongly suggests that age is a primary risk factor for CKD development.
2. Screening Recommendations:
 - Routine screening should begin by age 40-50.
 - Enhanced monitoring is warranted for patients over 50.
 - The high-risk category clearly includes patients 60+ years old.
3. Prevention Focus:
 - Younger populations (20-40) represent the best opportunity for preventive interventions.
 - Middle-aged adults (41-60) should receive targeted risk factor management.
4. Resource Allocation: Healthcare resources for CKD management should be prioritized for the 50+ age groups where prevalence is highest.

Limitations and Considerations

- The data shows percentages but not absolute numbers.
- No information provided about comorbidities or other risk factors.
- The "71+" category may include very elderly patients with different risk profiles.

4.3 Additional Statistical Analysis

Prevalence Ratio (CKD: Non-CKD) by Age Group

20-30: 0.33:1

31-40: 0.50:1

41-50: 0.75:1

51-60: 1.33:1

61-70: 2.50:1

71+: 4.00:1

Peak CKD Prevalence: 25% in age group 61-70.

5. CONCLUSION AND FUTURE SCOPE

In this paper the research priorities include early prediction and detection of chronic renal disease, as it progresses slowly, and late-stage diagnosis has a significant impact on health and the economy. This research highlights the growing importance of machine learning (ML) in addressing diagnostic challenges. Logistic regression, support vector machines, random forests, gradient boosting techniques, and artificial neural networks can detect CKD by observing complex patterns in demographics and metabolism. Tree-based ensemble methods and gradient boosting frameworks can better handle high-dimensional clinical data than linear classifiers because they can capture non-linear interactions between features. This study highlights major challenges that still need to be addressed before ML-based CKD prediction models can be used in the clinic, even if progress is made. Small dataset sizes, class imbalances, missing values, noise in healthcare records, and a wide range of features make it difficult to apply models to other conditions. Interpretability issues of advanced ML methods raise concerns about clinical openness and accountability. Recent studies have shown that understandable AI

methods like SHAP and LIME make AI easier to understand, although they have not yet been used in routine diagnostics. Another problem is the lack of large, standardized, long-term CKD datasets from different geographic and population groups. Further research should focus on ML-based frameworks that are therapeutically scalable, interpretable, and ethically implementable. Hospitals, research institutions, and data science groups need to collaborate to develop large CKD datasets that maintain people's privacy and include lab results, imaging data, lifestyle factors, and long-term follow-up records. Deep learning and hybrid methods are helpful for discovering and keeping an eye on diseases early on. Architectures that include EHRs, medical imaging, and data from physiological sensors show promise. Personalized prediction models using federated learning and transfer learning can also overcome dataset limitations and protect patient privacy. ML models in real-time clinical decision support systems must be validated, calibrated, and evaluated for fairness across varied populations to avoid bias. To reduce clinically important false negatives in CKD diagnosis, future research should include cost-sensitive learning. The use of machine learning in CKD screening and risk stratification could revolutionize preventive nephrology. Reliable, early, and patient-centric CKD prediction systems will emerge when data availability, algorithmic transparency, and clinical application strategies improve.

REFERENCES

- [1] M. Karthikadevi and N. Sneha, "Early Chronic Kidney Disease Detection Using Machine Learning Techniques," in *2nd International Conference on Machine Learning and Autonomous Systems, ICMLAS 2025 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 151–155. doi: 10.1109/ICMLAS64557.2025. 10967970.
- [2] P. Gogoi and J. A. Valan, "chronic kidney disease prediction using machine learning techniques: a comparative study of feature selection methods with SMOTE and SHAP," *Multiscale and Multidisciplinary Modeling, Experiments and Design*, vol. 8, no. 4, Apr. 2025, doi: 10.1007/s41939-025-00806-2.
- [3] P. M. A. M. A. Y, G. S. Venkat, U. C. Prasad, and M. P. Reddy, "Early Chronic Kidney Disease Identification using Machine Learning," in *Proceedings of the 4th International Conference on Information Technology, Civil Innovation, Science, and Management, ICITSM 2025, 28-29 April 2025, Tiruchengode, Tamil Nadu, India, Part II*, EAI, 2025. doi: 10.4108/eai.28-4-2025.2358049.
- [4] S. Umme Habiba *et al.*, "Early Prediction of Chronic Kidney Disease Using Machine Learning Algorithms with Feature Selection Techniques," in *Communications in Computer and Information Science*, Springer Science and Business Media Deutschland GmbH, 2024, pp. 224–242. doi: 10.1007/978-3-031-68639-9_14.
- [5] S. Moosavi Kashani and S. Zargar Balaye Jame, "Comparing the Performance of Machine Learning Models in Predicting the Risk of Chronic Kidney Disease," *Journal of Archives in Military Medicine*, vol. 11, no. 4, Feb. 2024, doi: 10.5812/jamm-140885.
- [6] I. Shaheen, N. Javaid, N. Alrajeh, Y. Asim, and S. Aslam, "Hi-Le and HiTCLe: Ensemble Learning Approaches for Early Diabetes Detection Using Deep Learning and Explainable Artificial Intelligence," *IEEE Access*, vol. 12, pp. 66516–66538, 2024, doi: 10.1109/ACCESS.2024.3398198.
- [7] N. Sonone and A. Daniel, "Early Prediction and Progrssion of Chronic Kidney Disease Using Machine Larning Techniques," in *Proceedings of the 2nd IEEE International Conference on Networking and Communications 2024, ICNWC 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICNWC60771.2024. 10537571.
- [8] M. Mustafizur Rahman, M. Al-Amin, and J. Hossain, "Machine learning models for chronic kidney disease diagnosis and prediction," *Biomedical Signal Processing and Control*, vol. 87, Jan. 2024, doi: 10.1016/j.bspc.2023. 105368.
- [9] S. Pal, "Chronic Kidney Disease Prediction Using Machine Learning Techniques," *Biomedical Materials and Devices*, vol. 1, no. 1, pp. 534–540, Mar. 2023, doi: 10.1007/s44174-022-00027-y.
- [10] Md. M. Hassan *et al.*, "A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patients Clinical Records," *Human-Centric Intelligent Systems*, vol. 3, no. 2, pp. 92–104, Feb. 2023, doi: 10.1007/s44230-023-00017-3.
- [11] C. Choudhary, L. S. Nagra, P. Das, J. Singh, and S. S. Jamwal, "Optimized Ensemble Machine Learning Model for Chronic Kidney Disease Prediction," in *Proceedings - 4th IEEE 2023 International Conference on Computing, Communication, and Intelligent Systems, ICCIS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 292–297. doi: 10.1109/ICCIS60361.2023.10425073.

- [12] M. Kaur and G. Kaur, "Comparative Analysis of Machine Learning Algorithms for Early Detection of Chronic Kidney Disease," in *2023 4th International Conference on Electronics and Sustainable Communication Systems, ICESC 2023 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1754–1759. doi: 10.1109/ICESC57686.2023.10193695.
- [13] F. Sanmarchi, C. Fanconi, D. Golinelli, D. Gori, T. Hernandez-Boussard, and A. Capodici, "Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review," *Journal of Nephrology*, vol. 36, no. 4. Springer Science and Business Media Deutschland GmbH, pp. 1101–1117, May 01, 2023. doi: 10.1007/s40620-023-01573-4.
- [14] Rajeshwari and H. K. Yogish, "Prediction of Chronic Kidney Disease Using Machine Learning Technique," in *2022 4th International Conference on Cognitive Computing and Information Processing, CCIP 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/CCIP57447.2022.10058678.
- [15] R. Alanazi, "Identification and Prediction of Chronic Diseases Using Machine Learning Approach," *Journal of Healthcare Engineering*, vol. 2022, 2022, doi: 10.1155/2022/2826127.
- [16] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," *Journal of Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00657-5.
- [17] M. B. Nirmala, D. K. Priyamvada, P. R. Shetty, and S. Pallavi Singh, "Chronic kidney disease prediction using machine learning techniques," in *12th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT 2021*, Grenze Scientific Society, 2021, pp. 185–190. doi: 10.30574/wjarr.2025.25.2.0384.
- [18] X. He, H. Cheng, Z. Liu, J. Yang, and T. Huang, "Comparative study on anti-lock braking control strategies based on heavy-duty multi-axle special vehicles," in *IMCEC 2021 - IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference*, Institute of Electrical and Electronics Engineers Inc., Jun. 2021, pp. 780–785. doi: 10.1109/IMCEC51613.2021.9482376.