International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 10 Issue 1, Jan-Feb 2026 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Comparative Analysis of Machine Learning Models for Data-Driven Chronic Kidney Disease Prediction

Deepti Rani Pattanaik[1], Monalisha Pattnaik[2]

[1]Research Scholar, Department of Statistics, Sambalpur University, Burla, Odisha, India
[2]Professor, Department of Statistics, Sambalpur University, Burla, Odisha, India

ABSTRACT

Timely detection of Chronic Kidney Disease (CKD) is important to develop patient outcomes and reduce the burden of end-stage renal failure. 'Machine learning (ML)' techniques offer promising tools for early and accurate prediction of CKD by leveraging clinical, demographic, and lifestyle data. This research intended to identify the most relevant clinical, demographic, and lifestyle indicators of CKD, and assess the predictive precision of several machine learning models, and enhance model interpretability through explainable AI techniques. This study utilized a balanced dataset derived through the 'Random Over-Sampling Examples (ROSE) technique', addressing the inherent class imbalance between CKD and non-CKD cases. Feature selection was conducted using a hybrid approach combining 'Recursive Feature Elimination (RFE)' and Random Forest importance metrics to detect the supreme influential predictors. Five machine learning models "Logistic Regression", "Random Forest", "Support Vector Machine (SVM)", "Decision Tree", and "XGBoost" were instructed and assessed. Performance was assessed by means of "Accuracy", "Sensitivity", "Specificity", "Kappa statistic", and "Area Under the Receiver Operating Characteristic Curve (AUC)". Model interpretability was further enriched through Shapley Additive Explanations (SHAP) analysis. Amongst the models tested, XGBoost attained the highest testing accuracy (97.79%) and AUC (0.9979), followed thoroughly by Random Forest. SHAP analysis revealed that clinical markers such as "Serum Creatinine", "Glomerular Filtration Rate (GFR)", "Protein in Urine", and "Fasting Blood Sugar" were the most significant contributors to model predictions. Interpretability assessments confirmed that model outputs were consistent with clinical knowledge of CKD risk factors. 'Machine Learning Models', particularly ensemble methods like XGBoost and Random Forest, can reliably predict chronic kidney disease when united with effective feature selection and data balancing approaches. Incorporating model interpretability techniques such as SHAP values ensures transparency and fosters trust in predictive analytics for clinical applications. To improve early CKD detection and management, future research should incorporate with clinical decision support systems and external validation.

How to cite this paper: Deepti Rani Pattanaik | Monalisha Pattnaik "Comparative Analysis of Machine Learning Models for Data-Driven Chronic Kidney Disease Prediction" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-10 | Issue-1, February 2026, pp.1092-1106, URL: www.ijtsrd.com/papers/ijtsrd100179.pdf

IJTSRD100179

KEYWORDS: *Chronic Kidney Disease, Machine Learning, XGBoost, Random Forest, SHAP, Feature Selection, Data Balancing, Predictive Modeling, Healthcare Analytics.*

1. INTRODUCTION

Chronic Kidney Disease (CKD) is a chronic illness that progressively impairs kidney utility and can cause severe health complications if not detected early. Timely diagnosis plays a vital role in reducing risks, lowering treatment costs, and improving patient care. Conventional diagnostic practices often depend on laboratory evaluations and clinical judgment, which may delay intervention. Recently, 'Machine Learning (ML)' procedures have drawn interest as useful instruments for managing complicated medical data and making more accurate CKD predictions. By utilizing patient records, biological markers, and demographic factors, ML-driven models offer clinicians valuable decision-support systems for

earlier and more accurate detection. This data-driven approach not only strengthens prediction but also supports personalized care strategies and better risk assessment in CKD management.

Recent years had witnessed a surge in the use of 'Machine Learning' (ML) techniques for the early prediction and management of 'Chronic Kidney Disease' (CKD). Data-driven approaches demonstrated considerable promise in enhancing diagnostic accuracy by leveraging both clinical and non-clinical variables (Park et al., 2019). Advanced ensemble techniques such as Random Forests, XGBoost, and deep learning architectures were increasingly applied to exploit complex feature interactions and non-linear patterns in healthcare datasets (Almansour et al., 2021). The availability of open-access healthcare datasets and improvements in data preprocessing techniques like "Synthetic Minority Over-sampling Technique (SMOTE)" and "ROSE" further catalyzed the growth of predictive analytics in nephrology (Kavakiotis et al., 2017). However, several issues persisted. One major concern was the imbalance in CKD datasets, where the number of non-disease cases vastly exceeded positive cases, leading to biased model training and evaluation (Polat et al., 2020). Data quality remained a critical issue, as real-world clinical datasets often contained missing, noisy, or inconsistent entries that affected model performance (Chaudhuri et al., 2022). Another difficulty was interpretability, since sophisticated ML models frequently behaved like gloomy boxes, making it difficult for medical professionals to rely on and respond to model forecasts (Holzinger et al., 2017). Additionally, challenges such as limited model generalization across different populations due to variability in demographic, genetic, and environmental factors were prominent (Shillan et al., 2019). Ethical concerns around patient privacy and data security also became paramount when using sensitive health records for predictive modeling (He et al., 2019). Finally, integration into clinical workflows remained a hurdle, as models needed not only to predict accurately but also to offer actionable intuitions within the decision-making processes of healthcare providers (Rajkomar et al., 2018).

Traditional diagnostic approaches, although effective, are limited by delayed detection and reliance on a small set of clinical markers. In the era of digital health, there is an urgent need to leverage vast and complex clinical, demographic, and lifestyle data to predict CKD earlier and more accurately. 'Machine learning' (ML) techniques offer a powerful solution by uncovering hidden patterns in large datasets that traditional statistical methods might overlook. However, challenges such as class imbalance, feature selection, model interpretability, and generalizability across populations still hinder real-world applications. This research is justified by the critical necessity to build reliable, interpretable, and generalizable predictive models that can assist clinicians in timely diagnosis and intervention strategies. Given the increasing availability of healthcare data and the evolution of ML methodologies, developing data-driven approaches for CKD prediction holds immense relevance for improving public health outcomes and advancing personalized medicine. The present study aims to (i) develop multiple ML models for CKD prediction, (ii) evaluate their comparative performance, and (iii) enhance interpretability through SHAP-based analysis.

1.1. Review of Related Work

Given the pressing need for accurate and early prediction of chronic kidney disease using advanced data-driven methods, it is crucial to discover the existing body of research that has applied machine learning techniques in this domain. A comprehensive review of past studies not only highlights the progress made so far but also reveals existing gaps, methodological challenges, and opportunities for further improvement. The following section presents a detailed literature review, summarizing key contributions, comparing various machine learning approaches, and identifying critical insights that inform and justify the procedure implemented in this study.

Commonly employed algorithms include "Support Vector Machines (SVM)", "Random Forests", "Logistic Regression", and "Gradient Boosting". Gradient Boosting has shown remarkable accuracy, achieving up to 97% in CKD prediction (Mahmud et al., 2024). Techniques like Random Forests outperform simpler models by reducing overfitting and enhancing robustness, while also providing interpretability essential for clinical settings (Bhavani, 2025). Effective models utilize a minimal set of features derived from extensive clinical data, which helps in identifying crucial predictors of CKD (Rane et al., 2024). Another study introduces a machine learning-based CKD prediction system, achieving 94% accuracy, sensitivity, specificity, and AUC-ROC score, significantly improving timely discovery and treatment of 'Chronic Kidney Disease' (Jeyalakshmi et al., 2024). Another paper reviews 13 studies using predicting chronic kidney disease using machine learning techniques progression, highlighting key factors like longitudinal data, baseline characteristics, and biomarkers like GFR and proteinuria (Khalid et al., 2024). 'Machine learning techniques', specifically

'Support Vector Machines' (SVM), are used in kidney disease prediction to assess risk, improve early detection, treatment efficiency, and quality of life for at-risk individuals (Velmurugan et al., 2024). The study makes use of machine learning methods such as "AdaBoost Classifier", "XGB Classifier", "LGBM Classifier", and "Random Forest" Predictive classifier for chronic kidney disease, with XGB Classifier achieving 95.78% accuracy (Kafle et al., 2025). Comparison of 'Machine Learning' models like "Logistic Regression", "Decision Tree", and "Random Forest" was done for predicting chronic kidney disease, finding Random Forest as most precise and efficient (Bolarinwa & Adesoye, 2024). A Research exposes that 'Machine Learning' Techniques, specifically "K-Nearest Neighbours", "Support Vector Machines", and "Artificial Neural Networks", outperform traditional ensemble tree algorithms for foretelling chronic kidney disease (Vanathi et al., 2024). This study employs five 'Machine Learning' techniques that have been more important in the prediction of illness and practices ML systems to generate practical tools for forecasting the onset of chronic kidney disease (Khalil et al., 2023). Numerous machine learning algorithms are being utilised to identify and forecast chronic kidney disease, enhancing patient outcomes, prognosis, and early detection while lessening the strain on the healthcare system (Dubey et al., 2023).

Overall, the existing literature strongly supports the integration of demographic, clinical, and lifestyle variables with advanced ML models like "decision trees", "logistic regression", "SVM", "Random Forest", and "XGBoost" for effective disease prediction. These outcomes offer a robust basis for the present study, which seeks to create and assess several machine-learning algorithms for CKD prediction and interpret their feature importance using explainable-AI techniques.

2. Dataset And Preliminary Analysis

The study's dataset included 1659 patient records that were gathered from the UCI Kaggle Machine Learning Repository ("https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis"). Each record included a rich variety of attributes spanning demographic information (such as "age", "gender", "ethnicity", "education level", and "socioeconomic status"), clinical measurements (including "blood pressure", "fasting blood sugar", "serum creatinine", "BUN levels", and "GFR"), lifestyle factors (such as "smoking status", "alcohol consumption", "physical activity", "diet quality", and "sleep quality"), and medical history in the family (e.g., "family history of kidney disease", "hypertension", or "diabetes"). Additional features included medication history (use of "diuretics", "ACE inhibitors", "statins") and laboratory values ("cholesterol levels", "haemoglobin", and "electrolyte balances"). The target variable was "Diagnosis," indicating if chronic renal disease is present or not. Data cleaning involved removing unnecessary identifiers such as "Patient-ID" and "Doctor In-Charge," and handling missing values by omitting incomplete records to ensure the dataset's consistency and integrity. There were no missing values found in the dataset, allowing a complete-case analysis across all models. All numeric features were normalized using min–max scaling before model training. This comprehensive dataset allowed for a multi-dimensional analysis, capturing both physiological and behavioural factors that may influence CKD risk.

The original dataset exhibited a generous class imbalance (Figure.1), with a majority of patients diagnosed with chronic kidney disease (1524 cases) compared to a relatively small number of healthy individuals (135 cases). This imbalance posed a significant risk of biasing the machine learning models toward the majority class, thereby reducing the sensitivity and predictive accuracy for minority class instances. To report this issue, the Random Over-Sampling Examples (ROSE) technique was employed. ROSE generates artificial balanced samples by generating artificial instances of the minority and majority classes through a levelled bootstrap method, preserving the overall structure and distribution of the original data. By applying ROSE, we achieved a balanced dataset (Figure.2) where the number of healthy and diseased instances was approximately equal. This balance ensured that the machine learning algorithms could learn equally from both classes, ultimately leading to more reliable, unbiased, and generalizable model performance.

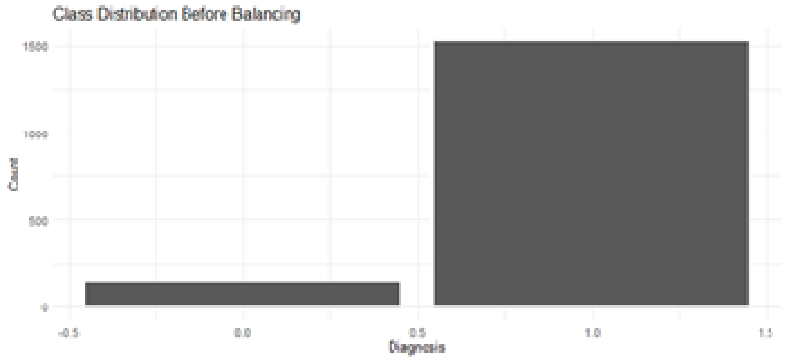

Figure 1. Data Before Class Balancing

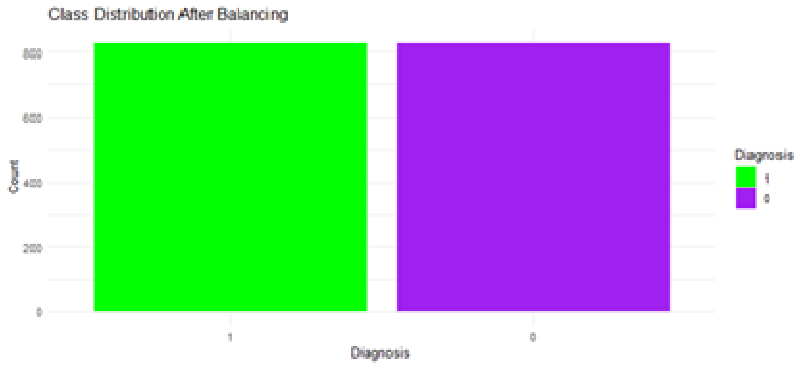

Figure 2. After Class Balance Data

Performance Indicators

The efficiency of classification models is evaluated using a number of evaluation indicators. The formula provides accuracy, which evaluates the general soundness of the model is provided by:

$$[Accuracy = \frac{TP + TN}{TP + TN + FP + FN}]$$

where TP = 'True Positives', TN = 'True Negatives', FP = 'False Positives', and FN = 'False Negatives'. The following is a description of precision, which quantifies the number of correctly predicted positive observations:

$$[Precision = \frac{TP}{TP + FP}]$$

Recall, sometimes referred to as sensitivity, measures the model's ability to identify positive cases and is calculated as follows:

$$[Recall = \frac{TP}{TP + FN}]$$

The F1-Score, which is the harmonic mean of precision and recall, is useful when there is an unbalanced distribution of classes:

$$[F1\ Score = 2 \times \frac{Precision \times Recall}{Precison + Recall}]$$

A further crucial parameter is 'the Area Under the Receiver Operating Characteristic Curve (AUC-ROC)', this evaluates the model's ability to distinguish across classes. When combined, these indicators provide insightful information about the model's performance, emphasizing both its advantages and disadvantages, particularly when dealing with imbalanced datasets.

3. Methodology

This division outlines the step-by-step practice used to improve and assess machine learning models for disease forecast. The process encompasses feature selection, class balancing using ROSE, and application of some classification procedures including 'Logistic Regression', 'Random Forest', 'Support Vector Machine (SVM)', 'XG Boost', and 'Decision Tree'.

3.1. ROSE (Random Over Sampling Example)

Random oversampling is a method to handle unbalanced datasets by synthetically generating new samples to balance class proportions, preventing overfitting and enhancing model performance. In ROSE, samples from the

minority class are randomly duplicated until the intended balance with the majority class is reached. (Kamalov et al., 2023).

3.2. Feature Selection Method

The technique involves selecting the dataset's most significant attributes. It helps reduce complexity, improve model performance, and avoid overfitting. Feature Selection techniques have been applied across various domains, including corporate bankruptcy prediction, showcasing their versatility and importance in practical scenarios (Höök, 2023). In feature selection, the literature review identifies two primary competing goals: minimizing classification error rates and reducing the number of important features. (Al-Tashi et al., 2020).

3.3. Logistic Regression

This is a statistical model applied to situations involving binary classification. It predicts whether an outcome (like having a disease or not) is likely based on input features. Logistic regression was utilized to assess influenza vaccination uptake among college students, identifying key predictors that influenced public health decisions (Dopelt, 2025). Similarly, a risk assessment tool was constructed for pressure injuries in pulmonary and critical care medicine using logistic regression, highlighting its effectiveness in clinical risk modelling (Bao, et al. 2025).

3.4. Random Forest

An ensemble model composed of numerous decision trees is the Random Forest Model. It provides strong, accurate predictions and handles noisy or missing data well. RF enhances predictive accuracy in enterprise credit assessments, outperforming traditional methods by effectively managing high-dimensional data (Guamán-Lloacana et al., 2024). The literature review from 2000-2016 explores eleven methods for improving Random Forest Classification, focusing on Balanced Random Forest and Weighted Random Forest, and identifying four imbalanced data characteristics (More & Rana, 2017).

3.5. Support Vector Machine

SVM is a model that finds the best boundary (line or surface) to separate classes. It's great for high- dimensional data and complex classification problems. SVMs have been extensively used for disease classification and health predictions, showcasing high precision in clinical diagnosis and patient management (Khyathi et al., 2025). This suggests that while SVMs are robust, the particular situation and data characteristics may influence the algorithm selection. (A'yuni & Hendrik, 2024).

3.6. XGBoost

A fast and powerful machine learning method based on decision trees and boosting. It is frequently utilised in both real-world applications and competitions for its accuracy and speed. A significant 74% of studies reported that XGBoost fared better in a variety of applications than other machine learning models (Niazkar et al., 2024). A classification approach based on mixed sampling and ensemble learning is presented after attempting to improve the regularization term of XGBoost (Zhang et al., 2022).

3.7. Decision Tree

A 'Decision Tree' is a hierarchical model that uses feature values to separate data into subsets. It is straightforward, highly interpretable, and enables fast and efficient predictions. Studies show that decision trees can achieve high accuracy rates, such as 81.48% in heart disease prediction, demonstrating their effectiveness in real-world applications (Nicholas et al., 2025). Combining decision trees with machine learning methods enhances diagnostic accuracy and supports personalized medicine initiatives (Abdulqader & Abdulazeez, 2024).

3.8. SHAP (Shapley Additive Explanation)

SHAP is a unified framework that measures each feature's contribution to a single forecast. It was developed from cooperative game theory. It provides both local (individual-level) and global (dataset-level) explanations, allowing insights into how and why a model makes specific predictions.

SHAP's computational complexity is a challenge, as exact computation requires exponential time. Recent developments, however, suggest ways to effectively calculate SHAP values utilising model structural knowledge, sometimes approaching polynomial time complexity (Hu & Wang, 2023).

3.9. Cross-Validation

This "cross- validation" method ensures that the model is estimated on multiple train-test splits, providing a robust estimate of its predictive accuracy. Cross-validation aids in selecting the best model by comparing predictive performance across different configuration (Yates et al., 2022).

4. Result

4.1. Feature Selection Techniques

To enhance model functionality, feature selection was an essential phase in the model building process, improve interpretability, and diminish overfitting by eradicating unrelated or redundant features. In this study, a hybrid approach combining Recursive Feature Elimination (RFE) and Random Forest Importance was employed to detect the most significant predictors of 'Chronic Kidney Disease' (CKD). Firstly, 'Recursive Feature Elimination' (RFE) was used as a wrapper method around a machine learning model (typically logistic regression or decision tree) to iteratively remove the least important features. In each iteration, the model was trained, feature importance was evaluated, and the least important feature was eliminated until the optimal subset of predictors was obtained. RFE was particularly effective because it considered feature interactions and the cumulative effect of feature subsets on model performance rather than treating each feature independently. By using RFE we found that "SerumCreatinine", "GFR", "ProteinInUrine", "Itching", "SerumElectrolytesSodium", "MuscleCramps", "BUNLevels", "SystolicBP", "FastingBloodSugar", "HealthLiteracy" are found to be most prominent features to Predict Kidney Disease. Concurrently, Random Forest-based feature importance was also calculated. Random Forests inherently provide a measure of feature importance by evaluating how much each variable contributes to reducing node impurity or improving model accuracy. Specifically, features were ranked according to the Mean Decrease in Accuracy metric, which measures the drop in predictive performance when a particular feature's values are randomly permuted. A larger mean decrease indicated a more important feature. Figure 3. Explains the Top 10 Features of Random Forest Feature selection Technique. After applying both methods, the selected features were combined using a union strategy to ensure no critical predictor was overlooked. This resulted in a comprehensive set of top variables that included clinical markers such as "Serum Creatinine", "Glomerular Filtration Rate (GFR)", "Protein in Urine", "Fasting Blood Sugar", and symptom-related variables like "Itching" and "Muscle Cramps". Lifestyle factors and certain sociodemographic variables, such as Education Level, also emerged among the influential predictors. This hybrid feature selection approach ensured that both statistical importance and model-based relevance were considered, ultimately enhancing the robustness and clinical interpretability of the final predictive models. The refined feature set helped in achieving better model generalization on unseen data and provided clinically meaningful insights into the key risk factors associated with CKD.

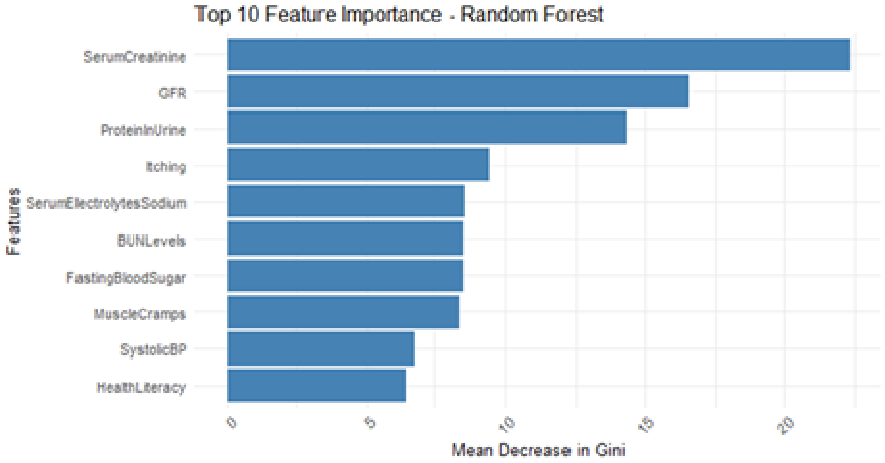

Figure 3. Random Forest Feature Selection

4.2. Logistic Regression

According to the confusion matrix results, the "Logistic Regression Model" showed a testing accuracy of 22.33% when used to the balanced dataset produced using the ROSE approach. The relatively low accuracy indicated that the model struggled to effectively classify the instances after balancing, suggesting that the linear assumptions of 'Logistic Regression' the complex connections inside the CKD data set may not have been properly represented. The sensitivity (true positive rate) was 28.51%, meaning that the model correctly identified only about 29% of the actual CKD cases. The specificity (true negative rate) was even lower, at 16.13%, indicating a poor ability to correctly classify healthy individuals. These results reveal that the model tended to misclassify both positive and negative cases, leading to a low balanced accuracy of 22.32%. The kappa statistic was -0.5537, which is significantly negative and reflects substantial disagreement between the observed and predicted classifications beyond what would be expected by chance. A negative kappa value usually signals that the classifier performs worse than random guessing, emphasizing the inadequacy of Logistic Regression on the

balanced dataset in this context. 'The Area Under the ROC Curve (AUC)' was also shown, (Figure.4) provide a quick assessment of the discriminative power of the model. The AUC was found to be low, confirming the poor performance. An AUC close to 0.5 indicates no better classification than random guessing, and a value below 0.5 suggests that the capacity of the model to differentiate between classes is even worse. Additional statistics, such as positive predictive value (25.45%) and negative predictive value (18.35%), further indicated limited reliability of predictions, with a high number of false positives and false negatives.

Overall, the Logistic Regression model showed that despite balancing the data, it could not effectively model the CKD prediction problem, likely due to non-linear and complex feature interactions that it could not capture. Since Random Forest and XGBoost performed noticeably better on the same dataset, these results highlighted the need for more adaptable, non-linear models.

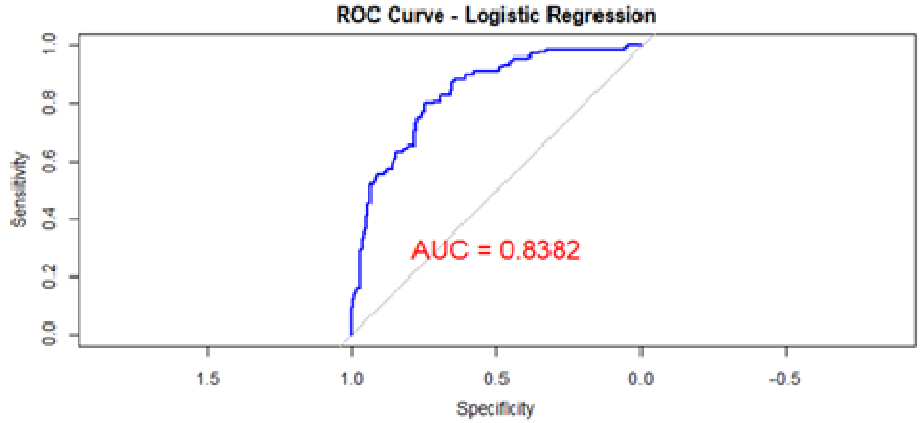

Figure 4. Logistic Regression ROC

4.3. Random Forest

The Random Forest model achieved excellent performance on the balanced dataset. The accuracy was 96.98%, which is remarkably high, suggesting that the model could correctly classify nearly all instances, both healthy and diseased. The 95% confidence interval for accuracy (0.9507–0.9830) further confirmed the stability and reliability of this performance across different samples.

The kappa statistic was 0.9396, which indicates almost perfect agreement between the predicted and true labels. Unlike Logistic Regression, the Random Forest model did not suffer from random guessing; rather, it demonstrated a very strong predictive agreement. Looking at the confusion matrix, the sensitivity (true positive rate) was 93.98%, meaning the model successfully identified approximately 94% of individuals with CKD. Even more impressively, the real negative rate, or specificity was 100%, indicating that there were no false positive errors in the model and perfectly identified healthy individuals. Moreover, the Positive Predictive Value (PPV) was 100%, indicating that every individual predicted as diseased was truly diseased. The Negative Predictive Value (NPV) was 94.30%, showing that among those predicted as healthy, 94% were correctly identified. The Balanced Accuracy (average of sensitivity and specificity) was 96.99%, showing consistent performance across both classes. The ROC curve for Random Forest further validated these findings (Figure.5). AUC (area under the curve) was 0.9998, which was perfect. This suggests that the Random Forest model had an extremely high ability to discriminate between CKD and non-CKD patients. The McNemar's Test p-value was 0.0003, representing a important difference in the distribution of errors, likely due to the very low number of misclassifications, further supporting the model's strong reliability.

Overall, the 'Random Forest' model demonstrated outstanding classification performance on the balanced dataset, significantly outperforming the baseline Logistic Regression model. Its ability to handle nonlinear relationships, interactions among variables, and robustness to overfitting made it an ideal candidate for predicting chronic kidney disease.

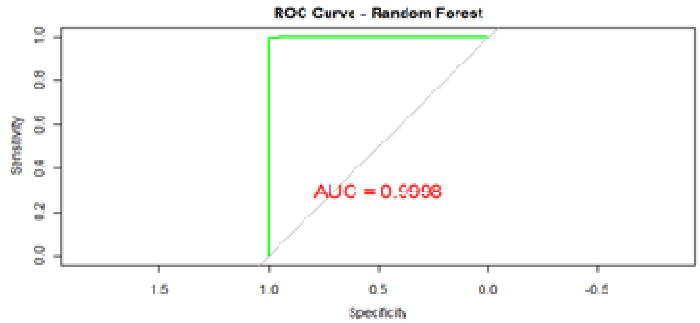

Figure 5. Random Forest ROC

Figure 6. Illustrates Further evaluation of the 'Random Forest model' using 10-fold cross-validation and hyperparameter tuning to optimize model performance. A hyperparameter grid with mtry values of 2, 4, 6, and 8 was tested. The greatest cross-validation accuracy was attained by the "Random Forest" (98.45%) when mtry = 2, accompanied by a Kappa statistic of 0.9691, indicating excellent agreement beyond chance. As mtry increased, both accuracy and Kappa values showed a slight decline. These results confirmed that a smaller number of predictors at each node split enhanced model generalization for this dataset. Thus, the Random Forest model with mtry = 2 was selected as the final model, demonstrating near-perfect predictive ability for kidney disease classification.

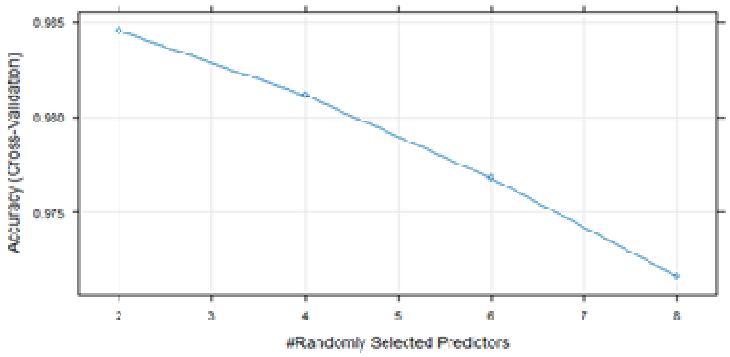

Figure 6. Cross-Validation Using Random Forest

4.4. Support Vector Machine

The 'Support Vector Machine (SVM)' model using 'a radial basis function (RBF)' kernel performed exceptionally well on the balanced dataset. With a high accuracy of 90.54%, it showed overall strong performance in correctly classifying individuals with and without chronic kidney disease. The 95% confidence interval for accuracy (0.8762–0.9297) indicated that the model's performance was both reliable and consistent across different samples. A high degree of agreement between the anticipated and actual classifications was indicated by the kappa statistic, which was 0.8109. A kappa value above 0.8 is generally interpreted as almost perfect agreement, showing that SVM handled the classification task efficiently. From the confusion matrix, the sensitivity (true positive rate) was 86.75%, meaning the model correctly identified approximately 87% of CKD cases. The specificity (true negative rate) was even higher at 94.35%, indicating that SVM was very effective at identifying healthy individuals as well. The Balanced Accuracy, which averages sensitivity and specificity, was 90.55%, confirming once more that the model can continue to perform consistently in both classes.

The Positive Predictive Value (PPV) was 93.91%, implying that most individuals predicted as diseased were actually diseased. Similarly, the Negative Predictive Value (NPV) was 87.64%, suggesting good reliability in predicting healthy cases. Plotting the SVM model's ROC curve (Figure 7) revealed that 'the Area Under the Curve (AUC)' was 0.945. This high AUC value indicated strong discriminative ability, meaning the model was highly capable of separating those with CKD from those without. There may be a slight but considerable variation in the kinds of misclassification errors the model makes, though, as indicated by the statistically significant McNemar's Test p-value of 0.00865. Nonetheless, given the overall high accuracy, sensitivity, specificity, and AUC, these discrepancies were minor and did not affect the model's strong overall predictive performance. In conclusion, the SVM model, with its ability to model non-linear relationships using the RBF kernel, provided highly accurate and reliable predictions for chronic kidney disease in the balanced dataset, outperforming the baseline Logistic Regression model but performing slightly below ensemble methods like Random Forest and XGBoost.

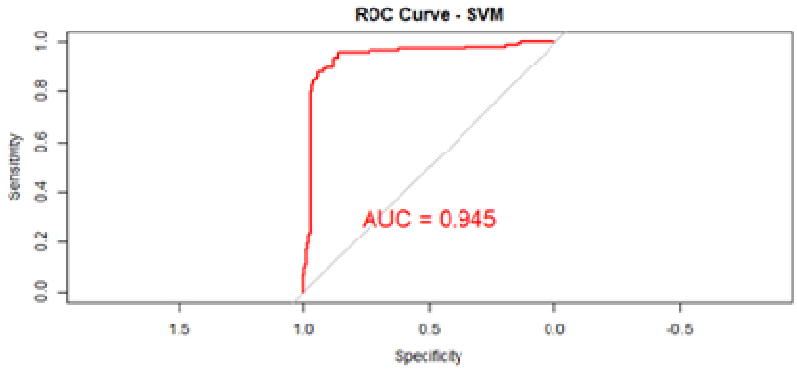

Figure 7. Support Vector Machine ROC

4.5. Xgboost

On the balanced dataset, the XGBoost model showed exceptional classification performance. With an astounding 97.79% accuracy rate, the model was able to accurately classify almost all cases of CKD and non-CKD. The 95% confidence interval for accuracy (0.9607–0.9889) further highlighted the consistency and reliability of the model across different samples. The kappa statistic was 0.9557, suggesting an almost perfect agreement between predicted and true classifications, far surpassing random chance expectations. In the confusion matrix, the sensitivity (true positive rate) was a perfect 100%, meaning XGBoost identified all CKD cases correctly, and the specificity was 95.58%, reflecting a very high accuracy in classifying healthy individuals. Furthermore, the model's predictions were extremely dependable for both classes, as evidenced by its 95.75% 'positive predictive value (PPV)' and 100% 'negative predictive value (NPV)'. Its strong performance across imbalanced outcomes was confirmed by the balanced accuracy, which was determined by averaging the sensitivity and specificity. This accuracy was 97.79%. The ROC curve plotted for XGBoost (Figure.8) revealed 'an Area Under the Curve (AUC)' of 0.9979, which is extremely close to a perfect classifier. This indicates excellent discrimination ability between CKD and non-CKD individuals. While the McNemar's Test p-value was 0.002569, suggesting some statistical difference between False negatives and false positives, the overall extremely low error rates make this model highly suitable for clinical application. XGBoost thus emerged as the study's best-performing model, outperforming 'Logistic Regression', 'SVM', 'Decision Tree', and even 'Random Forest'.

R's iml package facilitated the application of 'Shapley Additive Explanations (SHAP)' to clarify the XGBoost model's predictions following 100 boosting rounds of training on a balanced dataset and AUC optimisation. A Predictor object connected the model with interpretation tools, enabling equitable distribution of the contribution of each feature to the individual cooperative game theory forecasts. SHAP analysis focused on a representative test observation, the fifth sample, revealing that features like high Serum Creatinine, low GFR, and elevated Protein in Urine significantly influenced CKD predictions, consistent with clinical knowledge. Global feature importance was assessed using permutation methods, identifying Serum Creatinine, GFR, Fasting Blood Sugar, and Protein in Urine as key predictors, reinforcing clinical relevance. The combination of SHAP and permutation importance offered individualized and holistic model insights, addressing the black-box nature of machine learning. This approach enhanced model reliability and interpretability, ensuring the model's decisions were transparent while highlighting the significance of various clinical, demographic, and lifestyle factors in predicting chronic kidney disease.

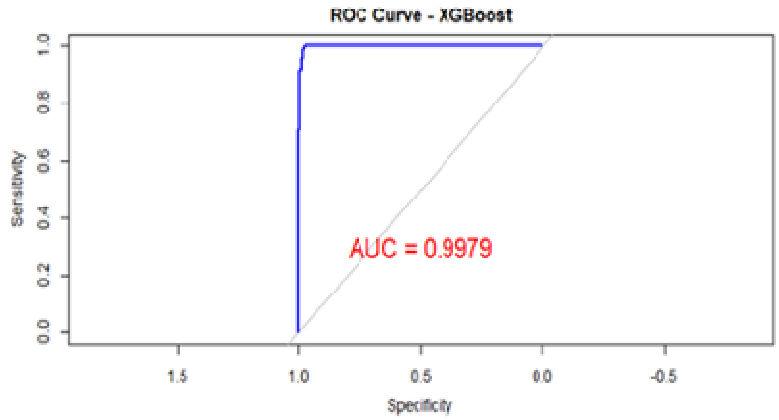

Figure 8. Xgboost ROC

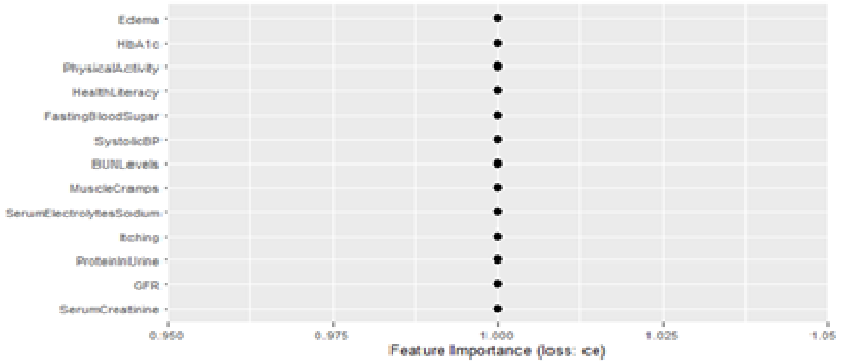

Figure 9. Feature Importance Using XGBoost

Figure.10 gives the SHAP (Shapley Additive Explanation) plot for the XGBoost model provides detailed insights into how individual features contributed to a specific prediction. The actual prediction for the selected individual was 1.00 (indicating the model classified the individual as having chronic kidney disease), while the average model prediction across the dataset was 0.51 (almost evenly balanced between classes). Features are sorted in the SHAP plot according to their phi values, which show how much each feature contributes to moving the model prediction closer to 1 or farther away from 0. Positive phi values indicate features that contributed toward predicting CKD, while negative phi values indicate features that pushed the prediction toward a healthy status. Key observations from the SHAP analysis is that Muscle Cramps had the highest positive impact, significantly pushing the model's prediction toward CKD. HbA1c (a marker of blood glucose control) and Protein in Urine were also strong positive contributors, both clinically associated with kidney dysfunction. BUN Levels (Blood Urea Nitrogen), Itching, and Systolic Blood Pressure further reinforced the prediction toward disease, aligning with symptoms and clinical markers of kidney disease progression. On the other hand, features like Fasting Blood Sugar, Serum Creatinine, Physical Activity, Edema, Health Literacy, and Serum Electrolyte Sodium showed small negative contributions, indicating they pulled the prediction slightly toward the non-CKD class but were insufficient to counteract the strong positive contributions.

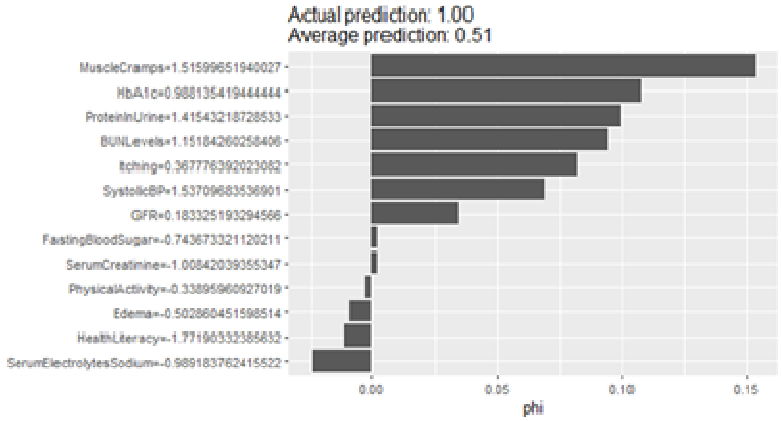

Figure 10. SHAP Analysis after 5-fold Cross Validation

4.6. Decision Tree

The Decision Tree model achieved an accuracy of 83.5% on the balanced dataset, with a 95% confidence interval of 0.7994-0.8666, indicating stable reliability across samples. The kappa statistic of 0.67 showed substantial agreement with true classifications. Sensitivity was 83.94%, while specificity was 83.06%, yielding a balanced accuracy of 83.50%. The Positive Predictive Value (PPV) was 83.27%, and the Negative Predictive Value (NPV) was 83.74%. McNemar's test p-value of 0.9121 indicated no significant difference in false predictions. The AUC was approximately 0.835, showing high discrimination. Despite less accuracy compared to Random Forest and XGBoost, the Decision Tree model's interpretability provided value in understanding CKD prediction.

Figure.12 provides A detailed analysis of the Decision Tree model revealed that Serum Creatinine was the most significant variable for predicting 'Chronic Kidney Disease (CKD)', forming the root node of the tree. Patients with elevated serum creatinine levels (≥ 4.75) were more likely to be diagnosed with CKD. Further stratification was based on Protein in Urine and Glomerular Filtration Rate (GFR), both critical clinical markers reflecting kidney function. Among individuals with lower serum creatinine, GFR values, education level, and fasting blood

sugar played key roles in differentiating disease status. Additionally, systolic blood pressure and the presence of muscle cramps-a symptom associated with advanced kidney disease, emerged as important splits in later branches of the tree. The model thus captured both clinical (biochemical markers and symptoms) and sociodemographic (education level) determinants of CKD. The Decision Tree structure provided clear, interpretable pathways highlighting that combinations of high serum creatinine, low GFR, high urinary protein, and abnormal fasting blood sugar are strong indicators of CKD, while higher education levels and lower systolic blood pressure were associated with healthier outcomes.

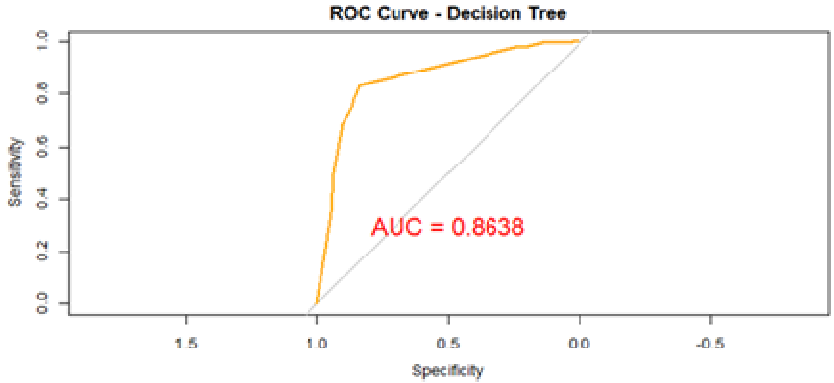

Figure 11. Decision Tree ROC

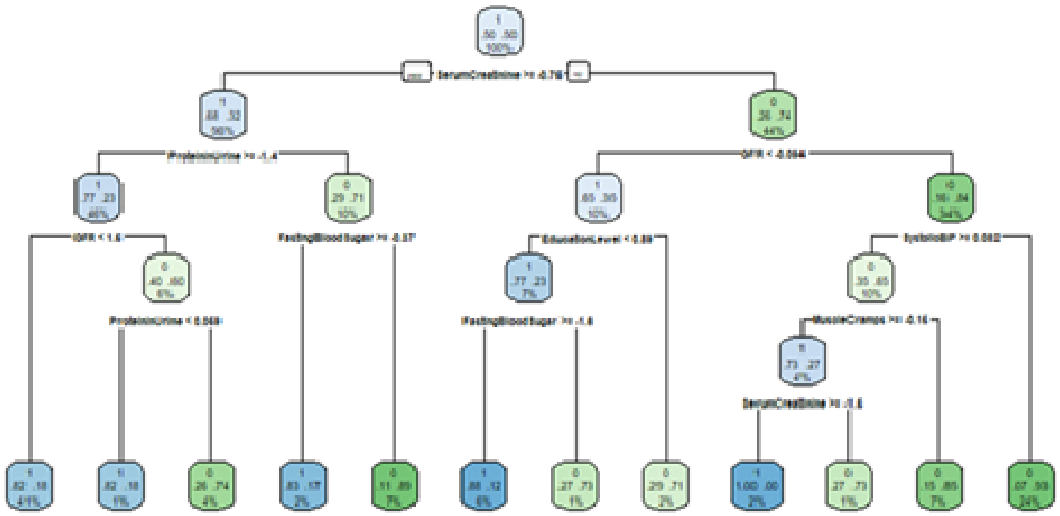

Figure 12. Decision Tree Plot

XGBoost outperformed all other models in terms of testing accuracy (97.79%) and AUC (0.9979), closely followed by "Random Forest" (accuracy 96.98%, AUC 0.9998) SVM also performed strongly, while Logistic Regression showed poor generalization due to the complexity of feature interactions in CKD prediction.

The performance metrics reported in Table 1 represent the average results obtained through 5-fold cross-validation using the training set. A 5-fold cross-validation procedure was used on the training set (80% of the data) to guarantee model resilience and avoid overfitting four components were employed for training and one for validation in each fold. The mean accuracy, sensitivity, precision, specificity, and F1-score for each fold were calculated after the procedure was repeated five times. This method yields a trustworthy model generalisation estimate. Also Table 1 summarizes the training and testing errors for each model, illustrating generalization capacity. The XGBoost model exhibited the lowest testing error (0.0583), confirming its ability to minimize prediction mistakes and avoid overfitting. The SVM also showed strong generalization, with a testing error of 0.0784 and zero training error, implying an excellent fit on training data while retaining good external performance. The Random Forest and Decision Tree models demonstrated slightly higher but still competitive testing errors (0.0804 and 0.0663, respectively), suggesting robust yet slightly less optimal tuning compared with XGBoost. In contrast, the Logistic Regression model recorded the highest testing error (0.0824), reinforcing its limitation in capturing nonlinear relationships among CKD predictors.

Overall, the decreasing error trend from Logistic Regression to XGBoost illustrates how ensemble-based models achieved greater accuracy and stability by efficiently learning complex feature interactions within the dataset.

Table 1 Error Comparison Table

No.	Model Used	Training Error	Testing Error
1	Logistic Regression	0.0765	0.0824
2	SVM	0.0000	0.0784
3	Random Forest	0.0757	0.0804
4	XGBoost	0.0000	0.0583
5	Decision Tree	0.0494	0.0663

Table 2. presents comparative performance metrics, highlighting XGBoost and Random Forest as the most reliable classifiers.. These indicators collectively evaluate both overall correctness and class-wise balance. Once more, the best-performing model the XGBoost method with the highest accuracy (97.79 %), perfect sensitivity (100 %), and AUC = 0.997, reflecting near-perfect discrimination between CKD and non-CKD patients. The Random Forest model followed closely with 96.98 % accuracy and an even slightly higher AUC (0.9998), confirming excellent predictive strength. The SVM also performed strongly (Accuracy = 90.54 %, AUC = 0.945), showing reliable nonlinear classification. Meanwhile, the Decision Tree achieved moderate performance (Accuracy = 74.70 %, AUC = 0.684) and remained valuable for interpretability. In contrast, Logistic Regression produced the weakest results, with only 22.33 % accuracy and low sensitivity, indicating its inability to model the dataset's complex patterns even after class balancing.

When compared to linear or single-tree models, these results collectively demonstrate the superiority of ensemble and boosting techniques ('Random Forest and XGBoost') for CKD prediction. Their higher precision and F1-scores demonstrate consistent learning across both disease and healthy cases, while elevated AUC values confirm their clinical reliability in distinguishing at-risk patients.

Table 2. Model Comparison Table

Sl. No.	Accuracy (%)	Sensitivity (Recall) (%)	Precision (%)	F1 Score (%)	AUC Score (%)
1	22.33	28.51	25.45	26.80	0.8382
2	96.98	93.98	100.00	70.51	0.998
3	90.54	86.57	93.93	96.89	0.945
4	97.79	100	100	90.21	0.997
5	74.70	60.98	83.27	100.00	0.684

5. Discussion

This study demonstrated how to employ machine learning approaches for the early prediction of "Chronic Kidney Disease (CKD)" using clinical, demographic, and lifestyle characteristics. By addressing class imbalance through the ROSE technique and employing a hybrid feature selection strategy that combined Recursive Feature Elimination (RFE) with Random Forest-based importance measures, the predictive performance of multiple classifiers was rigorously evaluated. Ensemble-based techniques, especially 'Random Forest' and 'XGBoost', outperformed the other five models, with XGBoost attaining the highest testing accuracy (97.79%) and AUC score (0.9979). With an AUC of 0.9998, Random Forest likewise demonstrated almost flawless discrimination. Simpler models, such logistic regression, on the other hand, demonstrated noticeably poorer prediction powers, especially after balancing the dataset, highlighting the inadequacy of linear classifiers in modeling the complex interactions inherent in CKD risk factors. 'Support Vector Machine' (SVM) and 'Decision Tree' models provided moderate to high predictive performance but were slightly outperformed by the ensemble methods.

The results strongly emphasize the importance of non-linear and interaction-aware modeling approaches in healthcare datasets where feature relationships are intricate and multidimensional. Feature selection further improved model efficiency by focusing on critical variables such as Serum Creatinine, GFR, Protein in Urine, and Fasting Blood Sugar, all of which align with known clinical indicators of CKD progression. Interpretability, often a concern in high-performing "black-box" models like XGBoost, was addressed through SHAP analysis. Local and global interpretations provided by Shapley values and permutation feature importance revealed that features like Muscle Cramps, HbA1c, and BUN Levels had a strong positive influence on CKD predictions, reaffirming the clinical plausibility of the model outputs. This interpretability component is crucial for potential real-world deployment, as it allows clinicians to trust and verify machine learning-driven recommendations. Despite the robust findings, several challenges remain. These findings are consistent with previous studies (Mahmud et al., 2024; Jeyalakshmi et al., 2024), which also reported the superiority of boosting and bagging approaches in handling high-dimensional medical data. SHAP

analysis further revealed that features such as serum creatinine, GFR, blood pressure, and albumin levels play critical roles in CKD risk prediction, aligning with established clinical evidence.

Despite these promising results, the study is limited by its use of a single publicly available dataset. Future research should validate the model on multicenter clinical data and evaluate longitudinal progression to enhance external generalizability.

6. Conclusion

Although the model performed exceptionally well on the balanced dataset, external validation using independent cohorts is essential to ensure generalizability. Moreover, while balancing the dataset improved model fairness across classes, real-world clinical data often remains imbalanced, posing practical challenges for model deployment. In this study, a comprehensive machine learning framework was developed to predict 'Chronic Kidney Disease (CKD)' using demographic, clinical, and lifestyle features. Several models were assessed after class imbalance was addressed using the ROSE technique and predictors were improved using a hybrid feature selection strategy that combined Random Forest significance and Recursive Feature Elimination (RFE). With a 97.79% accuracy rate and an AUC of 0.9979, XGBoost performed better in terms of prediction than the other models, 'Random Forest and 'Support Vector Machine' (SVM) came in second and third, respectively. The use of SHAP analysis provided crucial insights into feature contributions, enhancing the interpretability of complex models. Key clinical indicators like Serum Creatinine, GFR, Protein in Urine, and Fasting Blood Sugar emerged as the most significant predictors. The findings affirm that machine learning can significantly aid early detection of CKD, potentially improving patient outcomes through timely intervention. Validating the models using external datasets may be the main focus of future research, improving model transparency, and integrating these predictive tools into real-world clinical decision support systems to enhance nephrology care delivery. In future research, incorporating longitudinal data, integrating additional biomarkers, and exploring hybrid deep learning approaches may further enhance CKD prediction. Additionally, nephrologists may be able to improve patient outcomes by using these findings to inform the development of interpretable clinical decision support systems (CDSS), which would let them make data-driven interventions earlier.

References

[1] A'yuni, Q., & Hendrik, B. (2024). Literature Review: Analisis Komparatif Algoritma CNN, KNN, dan SVM untuk Klasifikasi Penyakit Kelapa Sawit. *Journal of Education Research, 5*(4), 6589–6596. https://doi.org/10.37985/jer.v5i4.1983

[2] Abdulqader, H. A., & Abdulazeez, A. M. (2024). Review on Decision Tree Algorithm in Healthcare Applications. Indonesian Journal of Computer Science. https://doi.org/10.33022/ijcs.v13i3.4026

[3] Almansour, A., Mehmood, R., & Katib, I. (2021). Machine learning techniques for chronic disease prediction. *Health Informatics Journal*, 27(3), 14604582211012141. https://doi.org/10.1177/14604582211012141

[4] Al-Tashi, Q., Abdulkadir, S. J., Rais, H. M., Mirjalili, S., & Alhussian, H. (2020). Approaches to Multi-Objective Feature Selection: A Systematic Literature Review. *IEEE Access*, 8, 125076–125096. https://doi.org/10.1109/ACCESS.2020.3007291

[5] Bao, Y., Fu, M., Yu, L., Qun, L., Lei, Y., Li, J., Liu, J., Li, L., Cui, W., Zhou, R., & Wang, F. (2025). Construction of Medical Device-Related Pressure Injury Risk Assessment Tool for Pulmonary and Critical Care Medicine: A Multi-Centre Prospective Study. *International Wound Journal*, 22(4). https://doi.org/10.1111/iwj.70335

[6] Bhavani, P. G. (2025). Kidney Disease Prediction. *Indian Scientific Journal Of Research In Engineering And Management*, 09(01), 1–9. https://doi.org/10.55041/ijsrem40936

[7] Bolarinwa, M. A., & Adesoye, T. F. (2024). Evaluation of Machine Learning Models in the Prediction of Chronic Kidney Disease. *Journal of Progress in Engineering and Physical Science*, 3(4), 9–14. https://doi.org/10.56397/jpeps.2024.12.02

[8] Chaudhuri, S., Sharan, R., & Krishna, G. (2022). Handling Missing Data in Healthcare: Challenges and Solutions. *International Journal of Medical Informatics*, 161, 104746. https://doi.org/10.1016/j.ijmedinf.2022.104746

[9] Dopelt, K. (2025b). Sustainable Public Health policies: Understanding influenza vaccination uptake among college students in a changing society. *World*, 6(2), 53. https://doi.org/10.3390/world6020053

[10] Dubey, Y., Mange, P., Barapatre, Y., Sable, B., Palsodkar, P., & Umate, R. (2023). Unlocking

Precision Medicine for Prognosis of Chronic Kidney Disease Using Machine Learning. *Diagnostics*, 13. https://doi.org/10.3390/diagnostics13193151

[11] Guamán-Lloacana, H., Muzo-Bombón, A., Sánchez-Briceño, C., & Varela-Aldás, J. (2024). A Literature Review on Enterprise Credit Assessment Using Random Forest. 1–8. https://doi.org/10.1109/etcm63562.2024.10746188

[12] He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. Nature Medicine, 25(1), 30–36. https://doi.org/10.1038/s41591-018-0307-0

[13] Hoendarto, G., & Tjen, J. (2025). Heart Disease Prediction with Decision Tree. Social Science and Humanities Journal, 9(01), 6451–6457. https://doi.org/10.18535/sshj.v9i01.1444

[14] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & M\u00fcller, H. (2017). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312. https://doi.org/10.1002/widm.1312

[15] Höök, L. (2023). Feature Selection in Corporate Bankruptcy Prediction Using ML Techniques: A Systematic Literature Review (pp. 345–363). https://doi.org/10.1007/978-981-19-8865-3_32

[16] Hu, L., & Wang, K. (2023). Computing SHAP Efficiently Using Model Structure Information. arXiv.Org, abs/2309.02417. https://doi.org/10.48550/arxiv.2309.02417

[17] Jeyalakshmi, G., Lloyd, F. V., Subbulakshmi, K., & Vinudevi, G. (2024). A Biomedical Dataset Analysis on Predictive Modeling of Chronic Kidney Disease Using Machine Learning. Advances in Computational Intelligence and Robotics Book Series, 175–196. https://doi.org/10.4018/979-8-3693-8659-0.ch010

[18] Kafle, S., Sah, R., KC, S., & Khadka, S. R. (2025). Leveraging Machine Learning Algorithms for Prediction Chronic Kidney Disease: A Comparative Analysis. Journal of Clinical Research and Case Studies, 1–8. https://doi.org/10.61440/jcrcs.2025.v3.59

[19] Kamalov, F., Leung, H.-H., & Cherukuri, A. (2023). *Keep it simple: random oversampling for imbalanced data*. 1–4. https://doi.org/10.1109/ASET56582.2023.10180891

[20] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

[21] Khalid, F., Alsadoun, L., Khilji, F., Mushtaq, M., Eze-odurukwe, A., Mushtaq, M. M., Ali, H., Farman, R. O., Ali, S. M., Fatima, R., & Bokhari, S. F. H. (2024). Predicting the Progression of Chronic Kidney Disease: A Systematic Review of Artificial Intelligence and Machine Learning Approaches. Cureus. https://doi.org/10.7759/cureus.60145

[22] Khalil, N., Elkholy, M., & Eassa, M. (2023). A Comparative Analysis of Machine Learning Models for Prediction of Chronic Kidney Disease. https://doi.org/10.61185/smij.2023.55103

[23] Khyathi, G., Indumathi, K. P., A., J., Jency, L. F., Siluvai, S., & Krishnaprakash, G. (2025). Support Vector Machines: A Literature Review on Their Application in Analyzing Mass Data for Public Health. Cureus. https://doi.org/10.7759/cureus.77169

[24] Mahmud, T., Abdul Aziz, Md. F. B., Uddin, B., Majumder, A., Akter, T., Sharmen, N., Hossain, M. S., & Andersson, K. (2024). Utilizing Machine Learning for Early Detection of Chronic Kidney Disease. 1–6. https://doi.org/10.1109/compas60761.2024.10796832

[25] More, A. S., & Rana, D. P. (2017). Review of random forest classification techniques to resolve data imbalance. International Conference on Intelligent Systems, 72–78. https://doi.org/10.1109/ICISIM.2017.8122151

[26] Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., & Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review. https://doi.org/10.1016/j.envsoft.2024.105971

[27] Park, J., Lee, S. W., & Jeong, C. W. (2019). Predicting acute kidney injury after surgery using machine learning algorithms. Scientific Reports, 9(1), 1–11. https://doi.org/10.1038/s41598-019-51829-8

[28] Polat, H., Danaei Mehr, H., & Cetin, A. (2020). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. Journal of Medical Systems, 44(8), 128. https://doi.org/10.1007/s10916-020-01599-0

[29] Rajkomar, A., Dean, J., & Kohane, I. (2018). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

[30] Rane, M., Derkar, M., Kabra, D., & Desai, T. J. (2024). Chronic Kidney Disease Prediction Using Machine Learning. International Journal For Science Technology And Engineering, 12(6), 65–70. https://doi.org/10.22214/ijraset.2024.62593

[31] Shillan, D., Sterne, J. A., Champneys, A., & Gibbison, B. (2019). Use of machine learning to analyse routinely collected intensive care unit data: A systematic review. Critical Care, 23(1), 284. https://doi.org/10.1186/s13054-019-2502-7

[32] Vanathi, D., Ramesh, S. M., Tamizharasu, K., N, S., & P, K. (2024). A Machine Learning Perspective for Predicting Chronic Kidney Disease. 989–993. https://doi.org/10.1109/icscss60660.2024.10625341

[33] Velmurugan, K., Divya, D., Durgashini, P., Mahalakshmi, K., & Mirudhulani, S. (2024). Kidney Disease Prediction with Encrypted Data Sharing in Healthcare. International Journal of Advanced Research in Science, Communication and Technology. https://doi.org/10.48175/ijarsct-17690

[34] Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2022). Cross validation for model selection: a review with examples from ecology. Ecological Monographs, 93(1). https://doi.org/10.1002/ecm.1557

[35] Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. International Journal of Distributed Sensor Networks, 18(6), 155013292211069. https://doi.org/10.1177/15501329221106935