Early Chronic Kidney Disease Identification using Machine Learning

Peer Mohamed Appa M. A. Y[1], Gaduputi Sai Venkat[2], Udaygiri Charan Prasad[3] and Moilla Prasanth Reddy[4]
{Peer.appa@gmail.com[1], gaduputisai121@gmail.com[2], Charanprasad2004@gmail.com[3], mprasanthreddy337@gmail.com[4]}

Department of Computer Science and Engineering, Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India[1]
Department of Artificial Intelligence and Machine Learning, Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India[2, 3, 4]

Abstract. Chronic Kidney Disease (CKD) is a progressive, irreversible and disabling condition that severely affects kidney function, representing the commonest cause of end-stage renal failure. Traditional diagnostic techniques such as serum creatinine, glomerular filtration rate (GFR), and urinalysis are usually laborious, expensive, and may miss the opportunity of early disease detection. ML approaches are being developed with the goal of improving the early detection of manageable disease by discerning subtle patterns from, and between, different sources of patient data beyond what is evident from conventional methods. In this paper, we propose a predictive model using Extreme Gradient Boosting (XGBoost), a kind of tree-ensemble method, which well-performs on structured medical data. Through a clinical data set involving demographic, biochemical, and haematological parameters, we show that XGBoost outperforms the Logistic Regression, Decision Trees, Support Vector Machines, and Random Forest in terms of accuracy (95.8%). There are also performance measurements (precision, recall, F1-score, confusion matrix) which provide good proof of its efficiency. Without the ML pipeline, early prediction of CKD would be infeasible, leading to late medical intervention and poorer patient outcomes. In the future, we will improve the prediction performance by integrating the real-time monitoring features of patients and by exploring deep learning approaches.

Keywords: Chronic Kidney Disease, Machine Learning, XGBoost, Early Diagnosis, Predictive Analytics, Medical Data Processing, Support Vector Machines, Random Forest, Decision Trees.

1 Introduction

Abdel-Fattah et al., (2022) [1] Chronic Kidney Disease (CKD), a slowly progressive, irreversible disease, is affecting millions of people all over the world leading to potentially lethal consequences like end-stage renal disease (ESRD), cardiovascular diseases, and metabolic disorders. Senan et al., (2021) [2] The paucity of rapid diagnosis, and low treatment accessibility are predisposing factors for higher mortality and an additional financial burden to receive dialysis and transplant for the patient. Chan et al., (2020) [3] The WHO and the GBD Study have emphasized that CKD has become one of the most rapidly rising causes of death and disability in the world while authorities worldwide are facing a growing burden of CKD, a circumstance fuelled by aging of populations, diabetes, hypertension, and genetic issues. Yuan et al., (2020) [4] The number of 65 years of age and older individuals in 2050 will be over 1.5 billion; this shows the urgent need for early detection and early intervention strategies.

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358049

Kovesdy, (2022) [5] Estimation of GFR, serum creatinine, and urinalysis are common diagnostic procedures. Tests like these, though, will usually find CKD only after it has progressed to a more serious stage and caused more permanent damage to the kidneys. Biological variable levels, laboratory measurement attitudinal ties, and patient health diverge, which all interfere with accuracy of diagnosis. Jongbo et al., (2020) [10] The problem of early detection is aggravated in setting of resource poverty due to the unavailability of high end tests. Hence, there is an urgent demand for novel data-driven (automated) methods for the early prediction of CKD.

AI and machine learning (AI/ML) have advanced at a rapid pace and has great potential for revolutionizing healthcare by generating predictive models that uncover intricate relationships within high-dimensional data. The classical ML classifiers, such as Logistic Regression, Decision Trees, Support Vector Machines, and Random Forest have been applied for the detection of CKD, however, they tend to overfit as well as have computation complexity and lack of clinical relevance. Levey and Coresh, 2012 [6] To this end, we develop an XGBoost algorithm ensemble model in this work with good generalization ability, which is scalable and robust in learning effective representations in structured medical data. Ravani et al., (2020) [9] The model is developed based on a dataset with demographic, biochemical and haematological features, and pre-processing of missing values, normalization and balance of class weight are performed. It is evaluated on basis of accuracy, precision, recall, f1-score and also confusion matrix analysis.

Bello et al., (2017) [7] Our own contributions to AI based health care analytics have been sought to expand the frontier by building a more generalized CKD prediction model, and this at the same time enhances diagnostic efficiency and reliability, eliminates medical errors, and evidence-based clinical decisions. Liu et al., (2021) [8] In future works, we are planning to integrate the real-time patient monitoring gadgets that are developed on the same model of IoT, deep learning for robust feature selection, and XAI for better interpretability and trust in the health care applications.

2 Methodology

The proposed Machine Learning (ML) based technique for early CKD detection is designed in the five steps: data gathering, data pre-processing, model building, model testing, and applying the model. Each step that lead to a robust, efficient and reliable predictive system that could give the aid to doctors in an early CKD diagnosis is important. The principal goal of this study is to employ the Extremely Gradient Boosting (XGBoost) classifier that demonstrated an excellent performance to handle structured medical data. The following section details the methodical approach that is adopted in the model development and enhancement.

2.1 Data Collection

We have used the popular UCI machine learning repository [community07recommender] to conduct our research. This dataset is particularly created to predict Chronic Kidney Disease (CKD), consisted on 400 observations and 25 clinical features that are crucial for a CKD early diagnose and predict. These characteristics are demographic, biochemical, haematological and urinary variables needed to understand the physiologically and pathologically significant

features of the disease. The data has both numerical and categorical features, which need to be pre-processed in order to train the model.

Several clinical variables that are thought to be important in diagnosing CKD are included in the dataset. We control for demographic factors such as age and gender, and for variation in renal function that are age- and sex-related, as the prevalence of CKD is higher in older DKD participants and gender differences may influence renal function. Among these haematological and biochemical parameters, blood urea, serum creatinine, sodium, potassium, haemoglobin, PCV and WBC count are important for kidney health. High serum (1.5 mg/dL) and blood urea (40 mg/dL) levels define the impairment in kidney function and are frequently used to calculate the GFR and to classification the degree of kidney disease. Albumin level, sugar concentration pull particle count and red blood cell (RBC) count are measurements found in urinalysis which is essential for detecting proteinuria (presence of excess protein in urine), a risk factor for CKD. Moreover, blood pressure (systolic and diastolic readings) has an important effect on CKD progression, because hypertension is both a causative and consequent factor for CKD.

Apart from these, diabetes mellitus and hypertension are two of the leading risk factors for CKD, making blood glucose random (BGR) readings and hypertension status valuable indicators in predicting disease progression. Studies indicate that patients with diabetes have a 2-4 times higher risk of developing CKD due to chronic hyperglycaemia-induced kidney damage. Similarly, persistent hypertension (> 140/90 mmHg) accelerates nephron loss, contributing to CKD development.

Due to the fact that medical data tends to be incomplete, inconsistent, and missing, the data pre-processing was a must before feeding it to the machine learning model for training. Handling of missing values on important variables such as blood pressure, haemoglobin, and glucose was conducted through statistical imputation to maintain data integrity. Categorical columns were normalized, encoded to ensure feature representation consistency. Such pre-processing is crucial for better data quality as well as models and for more accurate prediction of CKD.

2.2 Data Pre-processing

Data pre-processing is one of the most important stages of machine learning required for maintaining the data consistency, relevance of the features and the performance of the model. For clinical datasets, pre-processing is even more important as the data may contain missing values, noisy records, different scales and class imbalances. The pre-processing pipeline included four major stages: the missing value treatment, data transformation, feature selection and class imbalance resolution.

2.2.1 Handling Missing Values

Medical dataset usually contains missing values which can be replaced by the multivariate imputation by chained equation (MICE) approach for incomplete record of patient like human errors or not performed lab test. When data is missing, estimates can be biased and inaccurate, and the resulting model is less trustworthy. When data were missing in the different datasets imputation was applied according to the type of variable:

Mean Imputation: The missing values in continuous numerical features (e.g., blood urea, serum creatinine, haemoglobin, etc) were imputed using the mean of the respective feature. Using this hews to be a good value to keep good statistical properties without distorting the distribution too much in general.

Mode Imputation: Categorical attributes such as sensor type, diabetes status, hypertension and pus cell clumps etc. were imputed with mode. This ensures that, in place the missing categorical missing values are filled in with the values which have higher probability to occur.

K-Nearest Neighbours (KNN) Imputation: A KNN-based imputation was employed for features with more complex missing patterns. Estimates were constructed on the values of k nearest neighbouring patients with available data, and the imputed values should not change the global properties of the dataset. Through the use of these imputation methods, the data was cleaned up and prepared, preserving valuable data in place of missing values.

2.2.2 Data Transformation

Clinical data in its raw form is also mixed, in the sense of the presence of both categorical and continuous variables, which could be detrimental to the model's performance if not treated appropriately. The standardization and augmenting model efficiency were carried out through following transformations:

- Categorical Encoding: Some of the categorical features such as Albumin, Pus Cell Clumps, and Hypertension Status were transformed to a numerical representation using one hot encoding. This result influences for good the statistical thinking of non-structural learning: learning is given some degree of power in order to perform categorical processing without creating artificial orderings of the categories.
- Feature Scaling: As data includes variables with different magnitudes (e.g., serum creatinine varies from 0.4 to 15.2 mg/dL, and haemoglobin varies between 3.1 and 17.8 g/dL), we scaled all continuous layers to make them have a value between 0 and 1 by using the min-max scaling. This approach scales the values between 0 and 1 which means it will not make the attributes with large numerical values dominate the model.

Application of these transformations also organized the dataset in a way that was convenient for machine learning, and enabled numerical stability and faster convergence during the training of the model.

2.2.3 Feature Selection

Feature selection is crucial in medical machine learning tasks, to improve model efficiency, interpretability and generalization. By removing such irrelevant or redundant features, we can reduce the overfitting and thus the computing time of the model. The most significant features were chosen by two approaches:

- Recursive Feature Elimination (RFE): RFE was applied to progressively remove the least significant features and select the most important ones for CKD prediction.

- XGBoost Feature Importance Scores: XGBoost model comes with a built-in feature importance ranking mechanism with default ranking based on information gain and decision tree splits. The top 10 most significant features contributing to CKD classification are detailed below in Table 1.

Table 1. Feature Score.

Rank	Feature	Importance Score (%)
1	Serum Creatinine	**22.4**
2	Blood Urea	**17.6**
3	Haemoglobin	**15.3**
4	Packed Cell Volume (PCV)	**11.9**
5	Albumin	**10.7**
6	Blood Pressure	**8.4**
7	Blood Glucose Random	**7.2**
8	Sodium	**3.9**
9	Potassium	**2.6**
10	White Blood Cell Count	**1.8**

The results indicate that Serum Creatinine (22.4%) and Blood Urea (17.6%) are the most critical biomarkers for CKD detection, as they directly correlate with glomerular filtration rate (GFR), a key indicator of kidney function. Haemoglobin (15.3%) and Packed Cell Volume (11.9%) are also significant, as CKD often leads to anaemia due to reduced erythropoietin production. Blood Pressure (8.4%) and Blood Glucose Random (7.2%) further highlight the strong association between CKD, hypertension, and diabetes mellitus.

By selecting only, the most relevant features, the model reduces noise, improves efficiency, and enhances predictive performance.

2.2.4 Handling Class Imbalance

In medical classification tasks, imbalanced datasets pose a major challenge, as models may become biased toward the majority class, leading to poor sensitivity (recall) for detecting CKD cases. In the given dataset:

- 63% of instances were labelled as non-CKD (normal patients)
- 37% of instances were labelled as CKD (diseased patients)

Since the dataset exhibits a class imbalance ratio of approximately 1.7:1, it was necessary to apply resampling techniques to improve model generalization and ensure equal representation of both classes.

The Synthetic Minority Over-Sampling Technique (SMOTE) was used to address this issue. SMOTE works by generating synthetic instances of the minority class (CKD cases) using k-nearest neighbours. Unlike random oversampling, which simply duplicates existing samples, SMOTE synthesizes new instances based on feature-space similarities, thereby enhancing model robustness and preventing overfitting.

3 Algorithm for CKD Prediction using XGBoost

The Extreme Gradient Boosting (XGBoost) algorithm is employed in this study for Chronic Kidney Disease (CKD) classification due to its efficiency, scalability, and ability to handle structured medical data. XGBoost is an optimized gradient boosting framework, designed to provide high predictive accuracy as shown in Fig 1 while minimizing overfitting through regularization techniques. Unlike conventional decision tree-based models, XGBoost uses parallelized tree learning, weighted feature selection, and gradient optimization to enhance classification performance.

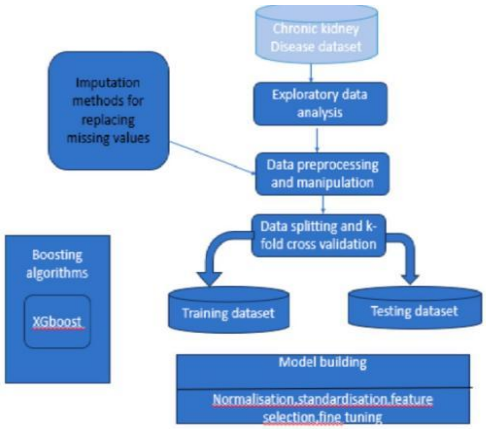

Fig. 1. CKD algorithm.

3.1 Introduction to Gradient Boosting and XGBoost

Gradient Boosting is an ensemble learning method that builds multiple weak learners (decision trees) sequentially, where each subsequent tree corrects the errors made by the previous ones. The final model aggregates these weak learners to form a strong predictive model. XGBoost enhances this process with advanced optimization techniques, including:

- Gradient-Based Tree Pruning: Reduces overfitting by eliminating unnecessary splits.
- L1 and L2 Regularization (Ridge & Lasso Penalties): Controls model complexity and prevents overfitting.
- Weighted Feature Importance: Prioritizes the most significant attributes for CKD prediction.

- Missing Value Handling: Naive detection of the pattern of missing data and then filling it very conveniently.

The main strength of XGBoost over regular machine learning models (like Random Forest and SVM) is that it can effectively deal with large, complex, unbalanced class distributed datasets, as the ones you provided (CKD datasets).

This section presents a detailed description of the experimental setting and the quality measure used to evaluate the performance of boosting approaches in the context of CKD classification. In order to obtain robust and stable predict ions, we made extensive experiments using five

This section presents a detailed explanation of experimental setup and evaluation metrics employed for the evaluation of the performance of boosting algorithms in prediction of Chronic Kidney Disease (CKD). In order to achieve accurate and reliable prediction, we have performed an array of experiments with the aid of five

various boosting algorithms and systematically investigated their predictive ability. The dataset was divided into a training test and a test and validation sample (60% and 40% proportion, respectively) in order to provide a balanced evaluation framework. The accuracy, precision, recall, F1-score, micro- and macro-weighted averages, as well as the running time (in seconds) of each algorithm were used to evaluate its performance. Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

3.2 XGBoost Model Training & Hyperparameter Tuning:

The XGBoost classifier is initialized and trained using the following hyperparameters:

- Learning Rate (η): 0.1 (controls the contribution of each tree).
- Maximum Depth of Trees: 6 (limits tree complexity to prevent overfitting).
- Number of Estimators (n_estimators): 200 (boosting rounds for improved accuracy).
- Subsample Ratio: 0.8 (controls the fraction of training data used for each boosting round).
- L1/L2 Regularization Parameters (Alpha & Lambda): Reduce overfitting by penalizing excessive complexity.

During training, each weak learner improves upon the previous learner's misclassified cases, iteratively refining the model.

Gradient-based optimization ensures that each successive tree corrects errors from prior trees, reducing the overall training loss.

3.3 Mathematical Formulation of XGBoost

XGBoost is an advanced gradient boosting technique that optimizes a given objective function using gradient descent and second-order approximations. The objective function consists of two components: a loss function measuring prediction error and a regularization term controlling model complexity.

Loss Function and Regularization The objective function for XGBoost is given as: *n*

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{k} \Omega(f_k) \quad (1)$$

Formula 1 Loss Function

Where:

- L(θ) represents the total loss function.
- l(yi ,y^ i) is the loss between the actual target yi and the predicted output y^ I.
- Ω(fk) denotes the complexity term that penalizes model complexity to prevent overfitting.
- K is the number of trees in the ensemble.

The optimization process uses gradient and Hessian calculations to approximate the function's second-order derivative for improved convergence.

The gradient update step follows:

$$g_i = \frac{\partial L}{\partial \hat{y}_i}, h_i = \frac{\partial^2 L}{\partial \hat{y}_i{}^2} \quad (2)$$

Formula 2 gradient and Hessian calculations

Where:

- gi represents the first-order gradient term.
- hi represents the second-order Hessian term.

4 Experiment, Results and Discussion

In this section, we detail the experimental setup, evaluation metrics, and performance measures used to evaluate the boosting algorithms for CKD prediction. We experimentally conducted comprehensive evaluation of the proposed methods with five kinds of boosting algorithms and demonstrated that they predicted efficiently and robustly. The data set was randomly divided into 60% for training and 40% for testing and validation, which assures a balanced evaluation framework. The performance of algorithms was calculated and compared using as accuracy, precision, recall, F1-score, micro-averaged and macro-average accuracy and the Receiver Operating Characteristic (AUC-ROC). Fig. 2 presents the Contributing features in Prediction of CKD for all boosting algorithms.

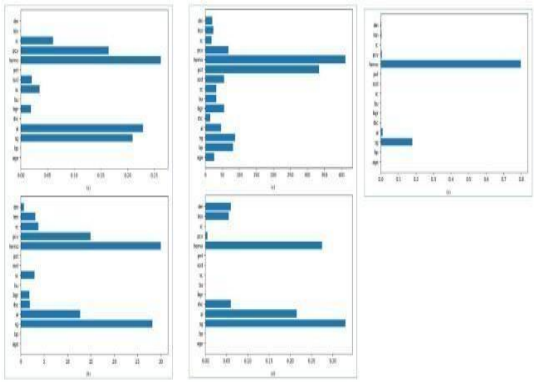

Fig. 2. Contributing features in CKD prediction for all boosting algorithms.

The listed values for each parameter for the respective algorithm were found to be the best performers in our experiment. Fig. 3 shows the Decision Tree Confusion Matrix.

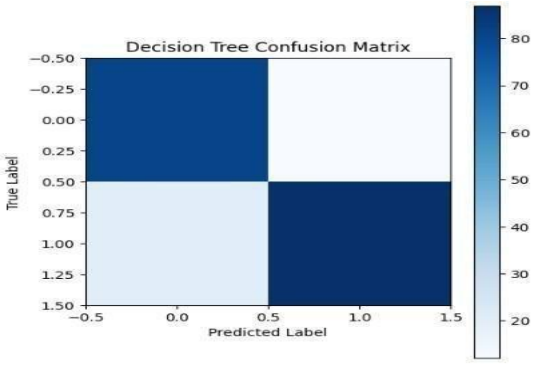

Fig. 3. Decision Tree Confusion Matrix.

The confusion matrix for a decision tree classifier. Dark squares signify correct predictions, while light squares indicate incorrect ones. This matrix helps in evaluating the classifier's performance. Fig. 4 shows the Random Forest Confusion Matrix.

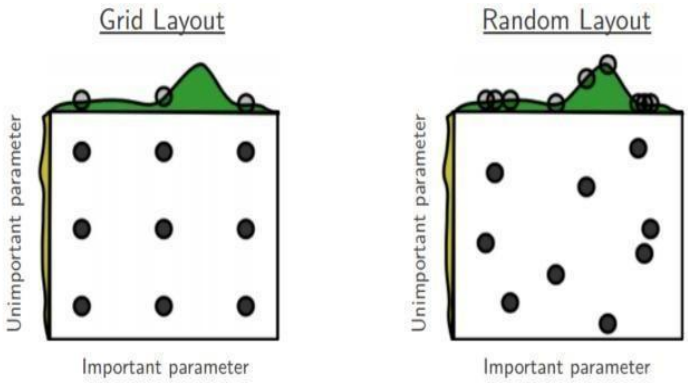

Fig. 4. Random Forest Confusion Matrix.

The confusion matrix for a Random Forest classifier. Dark squares signify higher correct prediction values, while lighter squares indicate fewer incorrect predictions. This matrix assists in evaluating the Random Forest model's performance. Fig. 5 shows the SVM Confusion Matrix.

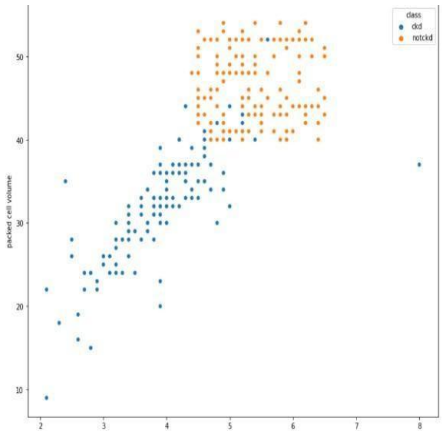

Fig. 5. SVM Confusion Matrix.

The confusion matrix for a Support Vector Machine (SVM) model. Dark squares denote higher correct prediction values, while light squares indicate fewer incorrect predictions. This matrix aids in evaluating the SVM model's performance. Fig. 6 shows the ROC Curve.

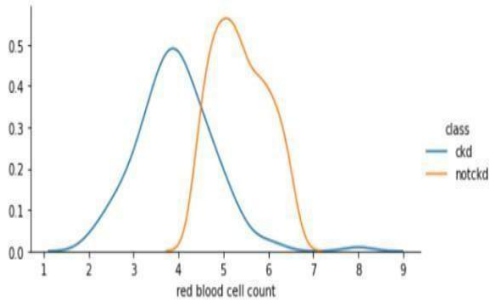

Fig. 6. ROC Curve.

ROC curve comparing multiple classification models, illustrating their true positive rate against the false positive rate. Fig. 7 shows the Heat Map.

Fig. 7. Heat Map.

Heatmap of summary statistics for heart disease-related features, including mean, standard deviation, and quartiles. Fig. 8 shows the Histograms with KDE Plots.

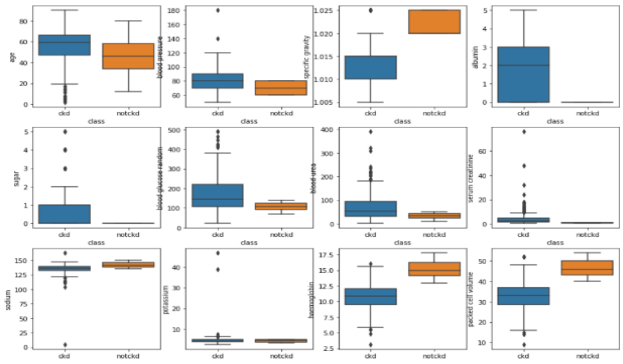

Fig. 8. Histograms with KDE Plots.

The distribution of various features in a dataset, likely related to medical parameters, using histograms with KDE plots

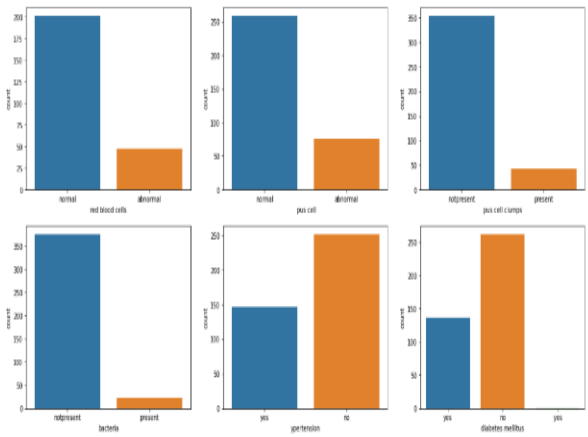

Fig. 9. Box Plots.

Box plots show the distribution of heart disease-related features, highlighting outliers in chol and old peak. Fig. 9 shows the Box Plots.

5 Conclusion

Diagnosis and control of chronic kidney disease have presented as difficult task for health workers and others in authorities. At least in part, it is possible to cope with if it can be pre-diagnosed early enough. We further applied their implementation to perform a thorough benchmark in five state of the art boosting algorithms – XGBoost, CatBoost, LightGBM, AdaBoost and Gradient Boosting – on a medical dataset from the UCI Machine Learning Repository. These models were accordingly trained via a fine-tuned pipeline that incorporate the robust pre-processing steps like multilevel imputation (mean, mode, and K-nearest neighbours (KNN)), normalization (Min-Max scaling), and standardization (Z-score), and via advanced feature selection with a recursive elimination and model-based importance scoring. The class imbalance was handled by applying the SMOTE approach to maintain fair and balanced model evaluation. AdaBoost outperformed any other model except for having a better accuracy (99.17%) compared to the rest had the best precision, recall, F1-score and AUC-ROC among all the models tested. While XGBoost performed well in prediction (a prediction accuracy of 95.8%) and feature ranking analysis revealed that serum creatinine and blood urea are the most contributing biomarkers, we observed that it was outperformed by AdaBoost in terms of total classification effectiveness. Comparative analyses carried out demonstrated that classic models such as SVM, Logistic Regression and Random Forest, while successful, were unsatisfactory at exploiting the nonlinear complex patterns in medical dataset as well as ensemble boosting models. Our results emphasize the need for efficient machine learning pipelines in clinical prediction tasks, as the earlier it receives medical attention the better the outcome for the patient.

References

[1] M. A. Abdel-Fattah, N. A. Othman, and N. Goher, "Predicting chronic kidney disease using hybrid machine learning based on Apache Spark," Computational Intelligence and Neuroscience, vol. 2022, art. no. 9898831, 2022. [Online]. Available: https://doi.org/10.1155/2022/9898831

[2] E. M. Senan, M. H. Al-Adhaileh, F. W. Alsaade, T. H. H. Aldhyani, A. A. Alqarni, N. Alsharif, M. I. Uddin, A. H. Alahmadi, M. E. Jadhav, and M. Y. Alzahrani, "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques," Journal of Healthcare Engineering, vol. 2021, art. no. 1004767, 2021. [Online]. Available: https://doi.org/10.1155/2021/1004767

[3] L. Chan, A. Vaid, and G. N. Nadkarni, "Applications of machine learning methods in kidney disease: Hope or Hype?" Current Opinion in Nephrology and Hypertension, vol. 29, no. 3, pp. 319–326, 2020. [Online]. Available: https://doi.org/10.1097/MNH.0000000000000604

[4] Q. Yuan, H. Zhang, T. Deng, S. Tang, X. Yuan, W. Tang, Y. Xie, H. Ge, X. Wang, Q. Zhou, and X. Xiao, "Role of Artificial Intelligence in Kidney Disease," International Journal of Medical Sciences, vol. 17, no. 7, pp. 970–984, 2020. [Online]. Available: https://doi.org/10.7150/ijms.42078

[5] C. P. Kovesdy, "Epidemiology of chronic kidney disease: an update 2022," Kidney International Supplements, vol. 12, no. 1, pp. 7–11, 2022. [Online]. Available: https://doi.org/10.1016/j.kisu.2021.11.003

[6] A. S. Levey and J. Coresh, "Chronic kidney disease," Lancet, vol. 379, no. 9811, pp. 165–180, 2012. [Online]. Available: https://doi.org/10.1016/S0140-6736(11)60178-5

[7] A. K. Bello, A. Levin, M. Tonelli, I. G. Okpechi, J. Feehally, D. Harris, K. Jindal, B. L. Salako, A. Rateb, M. A. Osman, B. Qarni, S. Saad, M. Lunney, N. Wiebe, F. Ye, and D. W. Johnson, "Assessment of global kidney health care status," JAMA, vol. 317, no. 18, pp. 1864–1881, 2017. [Online]. Available: https://doi.org/10.1001/jama.2017.4046

[8] P. Liu, R. R. Quinn, N. N. Lam, H. Al-Wahsh, M. M. Sood, N. Tangri, M. Tonelli, and P. Ravani, "Progression and regression of chronic kidney disease by age among adults in a population-based cohort in Alberta, Canada," JAMA Network Open, vol. 4, no. 6, art. no. e2112828, 2021. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2021.12828

[9] P. Ravani, R. Quinn, M. Fiocco, P. Liu, H. Al-Wahsh, N. Lam, B. R. Hemmelgarn, B. J. Manns, M. T. James, Y. Joanette, and M. Tonelli, "Association of age with risk of kidney failure in adults with stage IV chronic kidney disease in Canada," JAMA Network Open, vol. 3, no. 9, art. no. e2017150, 2020. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2020.17150

[10] O. A. Jongbo, A. O. Adetunmbi, R. B. Ogunrinde, and B. Badeji-Ajisafe, "Development of an ensemble approach to chronic kidney disease diagnosis," Scientific African, vol. 8, Art. no. e00456, 2020. [Online]. Available: https://doi.org/10.1016/j.sciaf.2020.e00456.