10.48047/jocaaa.2024.33.08.81

Predictive Analytics for Chronic Disease Management: A Machine Learning Approach to Early Intervention and Personalised Treatment

Rabi Sankar Mondal
MSc. in Business Analytics
University of New Haven, CT, USA
rmond1@unh.newhaven.edu
rabi.s.mondal@gmail.com
0009-0006-0136-9354
https://scholar.google.com/citations?user=oCbym6sAAAAJ&hl=en
Co-Author:
Md Nazmul Alam Bhuiyan
MBA in Data Analytics
University of New Haven, CT, USA
mbhui2@unh.newhaven.edu
bhuiyan.unh@gmail.com
0009-0003-2830-3396

ABSTRACT

The progressive condition, Chronic Kidney Disease (CKD), exists without symptoms yet causes major health system strain throughout the world. Foretelling the development of end-stage renal failure requires early detection in addition to a quick medical response. A machine learning-based method for chronic kidney disease prediction was developed using clinical records obtained from Kaggle. A total of 400 patient records with 25 features, which include biochemical markers together with demographic indicators, make up the dataset. A three-step standardised pipeline for data preparation consisted of filling empty values and converting labels into values before applying feature normalisation techniques to analyse the data. Visual tools in Exploratory Data Analysis showed serum creatinine, haemoglobin, and blood urea strongly contributed to discriminative patterns when observing data trends. Random Forest and XGBoost models received training through an 80–20 stratified dataset split method. The accuracy and AUC result of 96.25% with 0.9987 and F1-scores above 0.94 applied to both classes emerged in both models. ROC curves confirmed the robustness of the system along with confusion matrices. The Random Forest algorithm delivered better interpretability, but XGBoost achieved the most effective computational performance. Such research establishes that interpretable, effective models can be built using minimal clinical information. The deployed framework operates smoothly in clinical environments to identify CKD early during patient screenings. The model needs validation through multiple centres and implementation of SHAP explainability tools for future development.

Keywords: *Chronic Kidney Disease, Machine Learning, Random Forest, XGBoost, Predictive Analytics, Healthcare AI, Ensemble Models, Early Detection, Clinical Decision Support, SHAP.*

1. INTRODUCTION

The prevalence of chronic diseases has become a major worldwide healthcare issue because these diseases require most hospital admissions and long-term care services, along with high healthcare expenses. The International Society of Nephrology reports that CKD affects 850 million people worldwide, one of the most significant healthcare concerns. CKD progresses extensively without apparent warning, which can involve end-stage renal failure demanding dialysis or a kidney transplant to survive [1]. Several developing nations have restricted access to such medical treatments, making prevention and early disease detection a critical need. The economic pressure on healthcare systems from CKD appears through multiple channels, including advanced intervention expenses combined with hospitalisation expenses and management costs of hypertension and diabetes. The rising numbers of lifestyle risk factors worldwide create an immediate need for affordable diagnostic systems to detect early kidney disease instances and create specialised treatment sequences for patients [2].

Many healthcare professionals overlook the advantages of early diagnosis while consistently missing CKD diagnoses in their patients. The main obstacle to detecting early-stage kidney damage is its tendency to stay silent through standard clinical procedures because patients experience no signs of disease. The current diagnosis methods using serum creatinine tests with eGFR calculations and urinalysis assessment fail to detect subtle kidney function changes in people with no apparent kidney symptoms [3]. These diagnostic methods require periodic laboratory tests and patient adherence, but both measures are inconsistent in typical clinical settings. Several patients detect their condition too late because diagnostic delays extend their disease progression to advanced phases before diagnosis. The delay of clinical decision-making in nephrology through reactive approaches diminishes the

10.48047/jocaaa.2024.33.08.81

effectiveness of initial interventions for slowing disease progression [4].

The use of rule-based diagnostic systems, which serve as fundamental elements of traditional medical practice, demonstrates certain restrictions when diagnosing the complex, non-linear condition known as CKD. Medical systems with predefined thresholds operate without recognising variations in disease expression patterns among population groups [5]. Rule-based diagnostic systems apply universal modelling since they fail to account for the substantial effects which factors like age and ethnicity, along with gender and comorbidities, exert on clinical expressions and biomarker boundaries. Such systems cannot adopt adaptivity during data learning from large-scale real-world datasets since they remain confined by human logic that omits predictive data patterns. The traditional methods fail to effectively exploit EHRs, wearable health devices, and population-level health data to reach their full potential [6].

By applying machine learning (ML), researchers can transform CKD management through the applications of real-time decision support and predictive analytics. ML algorithms find patterns in various-dimensional datasets to detect relationships that standard statistics cannot identify. These mathematical systems use multiple patient data types, such as test results, demographic characteristics, diseases, and lifestyle parameters, to generate risk estimates for individual patient needs [7]. The most important attribute of Machine Learning models lies in their adaptability because the models will accept updates with new data to maintain their value for dynamic healthcare settings. ML-based decision support systems enable clinical practitioners to detect high-risk patients better early through risk prediction, facilitating treatment and intervention planning that suits distinct patient risk factors. Machine learning links precision medicine and clinical use through its application by providing scalable solutions to healthcare systems experiencing high pressures [8].

This research project aims to establish and test machine learning algorithms that forecast Chronic Kidney Disease among patients by assessing their clinical along demographic characteristics. Two supervised learning algorithms, Random Forest and XGBoost, undergo evaluation for their ability to classify patients by assessing 25 features found in a public Kaggle dataset, which separates patients into CKD-positive and CKD-negative classes. A detailed pre-processing step and exploratory data analysis phase led the study to understand feature distributions and their correlations. The modelling process includes split-based validation through stratified splits, ensuring fair performance across the two classes. This paper evaluates model performance through standard classification metrics, including accuracy combined with precision, recall, F1-score and a measurement called area under the ROC curve (AUC). AUC data will be presented with ROC plots and confusion matrices. The main value of this research stems from showing strong accuracy levels and generalisation capabilities while providing a deployable, interpretable system that medical facilities can easily use. The research evidence demonstrates how machine learning models deliver over 96% accuracy and almost perfect AUC scores in predicting CKD, thus validating the integration of predictive analytics into preventive nephrology while highlighting the power of AI for chronic disease management.

2. LITERATURE REVIEW

2.1 Overview of ML/AI in Healthcare Diagnostics

Healthcare diagnostics now function differently since healthcare providers have integrated machine learning (ML) and artificial intelligence (AI) in their disease detection and patient management strategies. The existing traditional diagnostic techniques function well in structured environments by using fixed thresholds alongside rule-based systems that demonstrate restricted adaptability capabilities when performing assessments. ML focuses on data-based techniques which locate and learn complex interconnected patterns within extensive mixed-type datasets [9]. Supervised learning models provide outstanding results in diagnosing diseases and planning treatments through their application in medical imaging, combined with genomic sequencing and electronic health records analysis. Supervised learning algorithms are key tools in classification systems because they are currently used in identifying cancer and predicting cardiovascular conditions. Medical experts can harness ML as a valuable clinical workflow instrument for its ability to recognise patterns automatically, reduce delays in diagnosis, and help determine risk levels among patients. The advances in explainable AI (XAI) have improved the model interpretability, which makes these systems more acceptable to healthcare professionals [10].

2.2 Prior ML Applications in CKD Detection

The application of ML focuses on Chronic Kidney Disease (CKD), which progresses silently since the condition manifests without noticeable symptoms. Identifying early-stage CKD using conventional methods is difficult because the disease remains asymptomatic during this period;

10.48047/jocaaa.2024.33.08.81

thus, researchers employ machine learning to extract delicate biomarkers of renal dysfunction from demographic and laboratory information [11]. Various studies have embraced publicly available datasets found in UCI and Kaggle CKD datasets to develop models identifying patients with CKD. Medical studies employ blood pressure, urea, serum creatinine, and haemoglobin measurements combined with hypertension and diabetes status. A goal exists to create prediction systems which help primary care doctors identify high-risk patients for additional clinical evaluations [12]. The study by Tripathi et al. (2020) presented a hybrid ML model which successfully detected CKD patients with a minimum accuracy rate of over 90%, thereby showing promise for implementing these methods in real clinical settings. These initial efforts provide a reliable starting point but operate only in experimental laboratories without meeting necessary deployment requirements for clinical practice [13].

2.3 Models Used in Previous CKD Studies (Logistic, NB, SVM)

The classification of CKD utilises multiple ML algorithms with Logistic Regression (LR), Naive Bayes (NB) and Support Vector Machines (SVM) appearing most frequently. Logistic Regression provides the basis for classification tasks because it remains straightforward and simple to understand. While assuming feature-to-log-odds relationships as linear, the model remains useful as a benchmarking tool [14]. The limitations of LR occur when non-linear patterns or intricate interactions exist in biomedical datasets, which are common in such contexts. The Naive Bayes classification method benefits from exceptional operational speed because it uses Bayes' theorem and predictor independence assumptions. The predictive ability of these models becomes limited through violations of independence assumptions present in clinical datasets. SVMs receive substantial praise because they excel at processing high-dimensional data and creating solid class segregation borders [15]. According to Kaur and Arora (2019), SVM modelling techniques exceeded 90% accuracy when analysing data related to CKD. The kernel adjustment requirement and vulnerable parameter selection cause SVMs to face difficulties with large-scale deployments and a clear understanding of results. The existing models deliver essential observations, though they display limited capabilities for combined patient groups and real-time diagnostic environments [16].

2.4 Limitations in Past Approaches

Previous machine learning applications for CKD prediction revealed several ongoing restrictions despite their encouraging outcome. Several research findings display poor interpretability due to using obscure models like ensemble methods alongside neural networks. Healthcare professionals tend to distrust medical models that cannot reveal their decision-making processes because clinical settings need transparent decision-making systems. Data related to CKD suffers from a typical class distribution imbalance due to non-CKD instances being more frequent than CKD cases in the samples [17]. The improper management of class imbalance causes models to over-predict the majority class, thus diminishing their sensitivity and reducing clinical effectiveness. Researchers implement three data sampling methods, including oversampling, sampling, and synthetic data generation with SMOTE, to handle the class imbalance problem, but their application and reporting practices are not standardised [18]. Small-scale validation constitutes a major difficulty among researchers. Numerous studies work with datasets which contain less than 500 records compounded by the lack of external validation which creates doubts about their general applicability. Model performance during internal testing proves acceptable, although it does not translate to effective accuracy with new data distributions. Data quality and feature relationships, alongside potential model biases, become difficult to understand because insufficient visual exploratory analysis is performed beforehand. The reliability and effectiveness of models deteriorate when they rely on data that researchers fail to understand or misinterpret correctly [19].

2.5 Novelty of This Work: Combining Interpretability, Visual EDA, and Robust Classifiers

The research presents a complete machine learning procedure for detecting CKD, which delivers interpretability, strong visualisation methods, and dependable model performance. The methodology performs extensive data analysis operations to understand CKD predictors through exploratory data analysis (EDA) methods, incorporating boxplots with violin plots, KDE plots, pair plots and correlation heatmaps [20]. The graphical displays help designers of models as well as provide transparent insights showing the distinctions between CKD and non-CKD patient groups through major variables, including haemoglobin and serum creatinine. This research utilises two sophisticated supervised classifiers, Random Forest and XGBoost, because of their strong performance, precision, and functionality to assess feature significance. The ensemble capabilities, built-in interpretability of feature

10.48047/jocaaa.2024.33.08.81

importance scores, and strength and speed in XGBoost drive its selection as a model, while Random Forest benefits from these attributes and its ensemble and interpretability capacity. Stratified train-test splits validate the predictive models, while the evaluation process depends on accuracy, precision, recall, F1-score and AUC measurement.

The models' performance visualisation includes ROC curves and confusion matrices, which enable a direct understanding of sensitivity-specificity relationships and classification precision. The ending model iteration reaches an accuracy of 96.25% and approaches 0.9987 AUC scores, which outpace the results found in most previous literature studies. This work adds interpretability value through a detailed, documented pipeline, explaining a step-by-step process researchers can readily apply and modify in clinical environments. This study establishes a method to link academic laboratory work with real-world uses by showing that visual model development with ensemble systems leads to reliable diagnostic systems evaluating chronic kidney disease.

3. METHODOLOGY

3.1 Proposed Methodology Framework

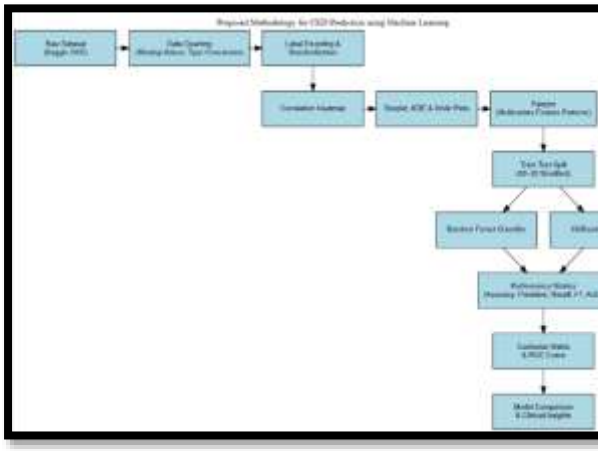


Figure 1: Proposed Methodology Diagram

Figure 1 illustrates a sequential approach to predicting CKD, which starts by structuring raw data and then continues with thorough analysis before splitting into Random Forest and XGBoost training phases. Random Forest and XGBoost training paths originate from the methodology outline, which measures their performance with identical assessment criteria. The real-world application receives guidance from clinical insights after visual comparisons during the methodology completion.

3.2 Dataset Description

The study utilises 400 records and 25 clinical features from the Chronic Kidney Disease (CKD) dataset available on Kaggle, focusing on renal function diagnostics. This data includes patient demographic attributes, blood pressure measurements, and hematology and biochemical analysis parameters such as hemoglobin and packed cell volume and blood urea and serum creatinine tests. The dataset has a binary target variable distinguishing patients from those diagnosed with CKD or those without this diagnosis. The patient records show a moderate class imbalance because they include two hundred and fifty CKD-positive cases and one hundred and fifty CKD-negative cases. Evaluation methods and stratified splitting become necessary to prevent the model from learning from becoming biased because of this data imbalance.

3.3 Data Pre-processing-processing

The machine learning-ready state of the dataset needed multiple data transformation methods. Missing values were identified in the essential numeric fields, including pcv, wc and rc. Mean imputation provided the solution for filling in missing numerical feature values:

$$x_i^{(\text{imputed})} = \frac{1}{n}\sum_{j=1}^{n} x_j \text{ where } x_j \in \text{non missing values of column x}$$

The pandas.to_numeric function with errors='coerce' option transformed pcv and wc columns into floats by replacing invalid data with NaN values. LabelEncoder applied a numerical value to each category type, including htn, dm, and applet. Conversion of the dataset leads to a numerical matrix that does not increase its dimension by applying one-hot encoding methods.

The dataset received pre-processing for XGBoost models through an application of standardisation that involved the following calculations:

$$z = \frac{x - \mu}{\sigma}$$

This equation applies the standard deviation of each feature (σ) and its corresponding mean value (μ). StandardScaler implemented through Scikit-learn performed the transformation. The procedure resulted in a ready-to-use model development dataset that normalised numerical data.

3.4 Exploratory Data Analysis (EDA)

10.48047/jocaaa.2024.33.08.81

Exploratory Data Analysis revealed important patterns in feature characteristics and relationships between classes. A class distribution plot showed that the data existed at a moderate level of imbalance [21]. A Pearson coefficient analysis of the data through correlation heatmaps showed that disease status had negative relationships with haemoglobin measurements and positive associations with serum creatinine levels.

The research utilised visual plots, including boxplots, KDE plots, and violin plots, to examine distribution patterns of vital variables between CKD subjects and non-CKD participants. The research showed that CKD patients demonstrated distinctively lower haemoglobin results. The selected features, age and haemoglobin, serum creatinine and blood urea and sodium, displayed quantifiable differences between the disease classes and indicated significant correlations.

The visual data representations proved the existence of valuable patterns while facilitating the selection of training features from the dataset.

3.5 Model Development

Random Forest and XGBoost ensemble models were chosen to predict CKD because they maintain good performance and reliability in medical decision tasks. Random Forest represents a bootstrap-aggregated ensemble model that includes decision trees, which follow this definition:

$$\hat{y} = \text{majority\_vote}\{h_1(x), h_2(x), \ldots, h_k(x)\}$$

where $h_1(x)$ is the prediction from the i^{th} decision tree trained on a random subset of the data.

XGBoost applies **gradient boosting**:

$$F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x)$$

F_t is the boosted ensemble at stage t, h_t is the new decision tree, and η is the learning rate.

The data split for training and testing occurred at 80% training to 20% testing, while stratified sampling ensured proportionate levels of each class between both parts. The parameters used default settings with random_state set to 42 as a standard for replication. The training phase used Scikit-learn for Random Forest protocols and the XGBoost Python API.

Evaluation metrics included:

- **Accuracy**:

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

- **Precision**:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity)**:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score**:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Area Under Curve (AUC)**: The Area Under Curve (AUC) metric originates from the Receiver Operating Characteristic (ROC) analysis between true positive rate and false positive rate measurements across different threshold values [22].

The models performed exceptionally by reaching 96.25% accuracy and AUC scores that virtually matched 0.9987, showing their superior discriminatory characteristics. The classification errors were depicted through confusion matrices. The models performed well in error detection by producing a few incorrect results.

The combined ROC curve analysis clearly showed strong agreement between Random Forest and XGBoost performance metrics because they proved to be highly effective classification methods.

A performance comparison table presented exhaustive data about the two models' accuracy, precision, recall, F1-score and AUC values. Interpretability levels were better for the Random Forest model, but XGBoost provided improved recall metrics.

4. RESULT AND DISCUSSION

4.1 Exploratory Data Analysis

10.48047/jocaaa.2024.33.08.81

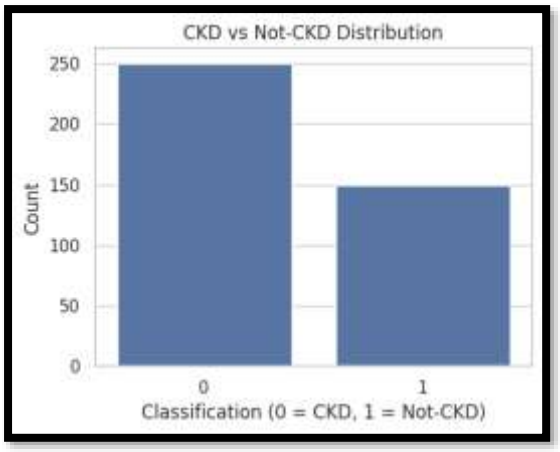


Figure 2: CKD vs Not-CKD Distribution

Fig 2 presents a moderate class distribution that appears in the dataset. Among the patients, 250 were diagnosed with CKD (label 0), whereas 150 patients did not receive a CKD diagnosis (label 1). Stratified sampling must be used along with accuracy and additional performance metrics to ensure unbiased model evaluation when classifying data collections.

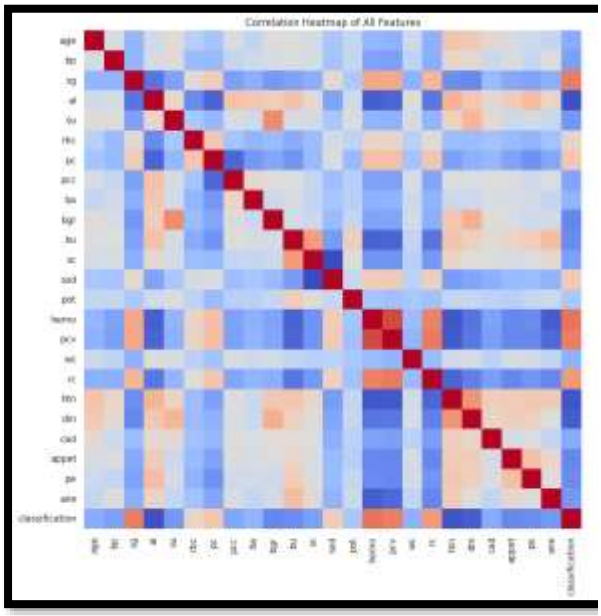


Figure 3: Correlation Heatmap

Fig 3 reveals inter-feature relationships and associations with CKD classification. Features like serum creatinine (sc), haemoglobin (hemo), and blood urea (bu) show strong correlations with CKD status. Negative correlations between haemoglobin and CKD highlight its diagnostic value. The matrix detects unimportant and predictive variables that should be considered for modelling purposes.

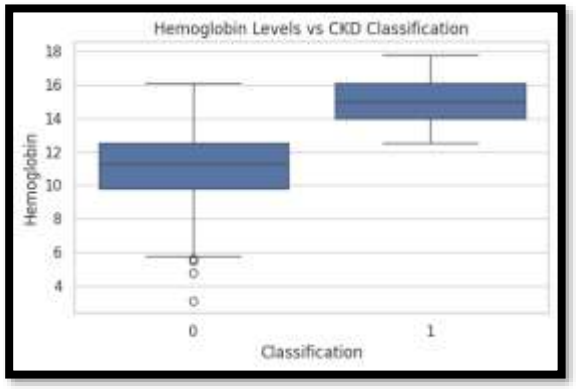


Figure 4: Hemoglobin Levels vs CKD Classification (Boxplot)

Fig 4 demonstrates the dispersion of haemoglobin levels between patients with and without CKD. The clinical observation of common anaemia in CKD patients receives support through the lower haemoglobin levels observed in patients with class 0 CKD status. The results establish that haemoglobin demonstrates the powerful discriminatory potential for classification purposes.

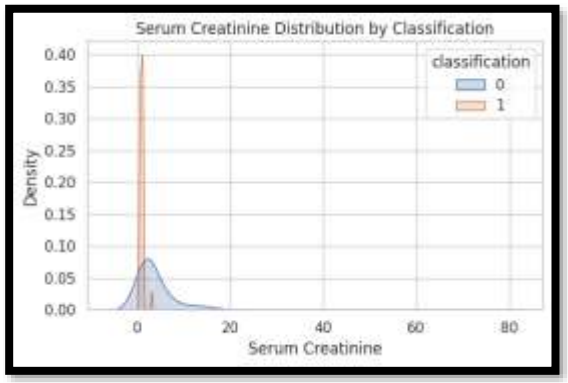


Figure 5: Serum Creatinine Distribution by Classification (KDE Plot)

Fig 5 demonstrates that serum creatinine levels stand much higher in patients with CKD than those without CKD. Patients in the CKD class group display wider variability in their serum creatinine levels compared to non-CKD patients. The diagnostic marker of CKD features clearly through creatinine because it shows decreased filtration capability of the kidneys.

10.48047/jocaaa.2024.33.08.81

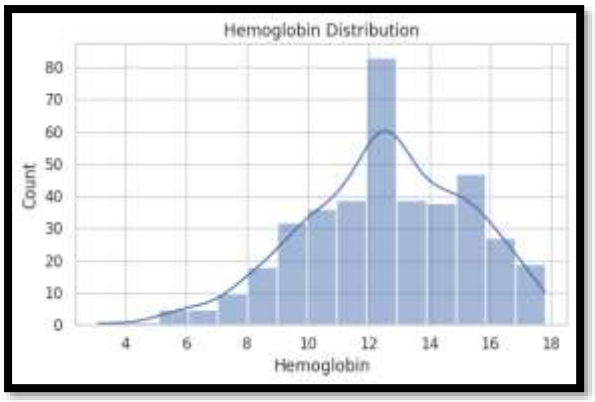


Figure 6: Hemoglobin Distribution (Histogram + KDE)

Figure 6 presents a histogram with a KDE overlay, which shows that haemoglobin values distribute mildly to the left with a central range between 12–13 g/dL. The widespread distribution of haemoglobin levels within CKD patients shows how these patients most frequently measure below this spectrum, thereby supporting the need for continuous haemoglobin tracking for diagnostic purposes and follow-up care.

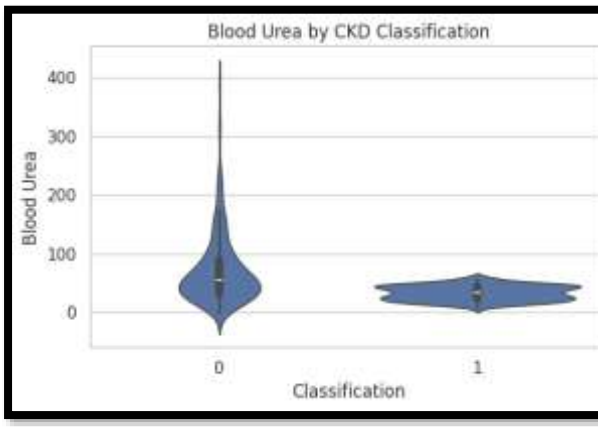


Figure 7: Blood Urea by CKD Classification (Violin Plot)

The blood urea patterns for patients with CKD show higher levels and broader ranges than those observed among patients who do not have CKD, based on Fig 7. The distribution of renal dysfunction variables extends wider within class 0 (CKD), yet shows a leaner range for healthy individuals in class 1.

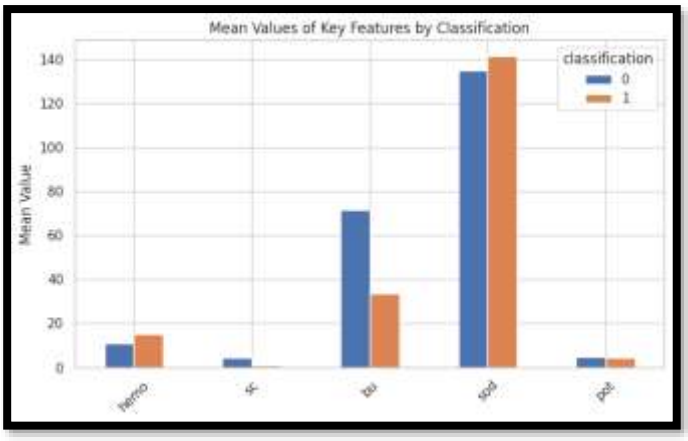


Figure 8: Mean Values of Key Features by Classification (Bar Plot)

Fig 8 compares the mean values of selected features, haemoglobin, serum creatinine, blood urea, sodium, and potassium between CKD and non-CKD classes. Clinical findings support predictive modelling, showing CKD patients have elevated creatinine and urea but decreased haemoglobin and sodium levels.

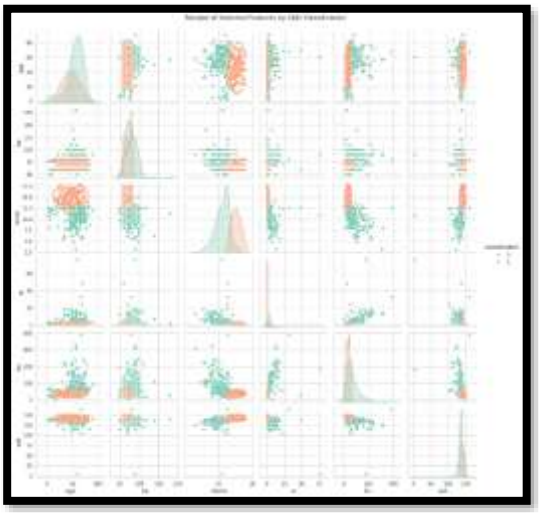
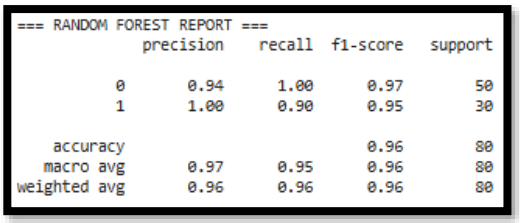


Figure 9: Pairplot

Figure 9 illustrates multivariate relationships connecting age with blood pressure, haemoglobin, serum creatinine, and blood urea measurements. The data clusters distinctly between CKD patients and those without CKD through haemoglobin and serum creatinine levels analysis. The graphical visualisation demonstrates model clustering viability because it verifies group separability, enabling accurate ensemble model classification.

4.2 Model Evaluation

10.48047/jocaaa.2024.33.08.81



=== RANDOM FOREST REPORT ===

	precision	recall	f1-score	support
0	0.94	1.00	0.97	50
1	1.00	0.90	0.95	30
accuracy			0.96	80
macro avg	0.97	0.95	0.96	80
weighted avg	0.96	0.96	0.96	80

Figure 10: Random Forest Classification Report

The Random Forest model's classification report demonstrates outstanding performance since all separate performance values sit at 0.96 macro-averaged precision-recall and F1-score (Figure 10). The model detected every case of CKD (class 0) along with perfect precision for all patients who did not have CKD (class 1). The model's overall accuracy reaches 96% despite maintaining equal identification precision between both categories. The model demonstrates strong performance across patient classes because it shows an exceptional ability to detect CKD patients, which plays a key role in clinical practice.

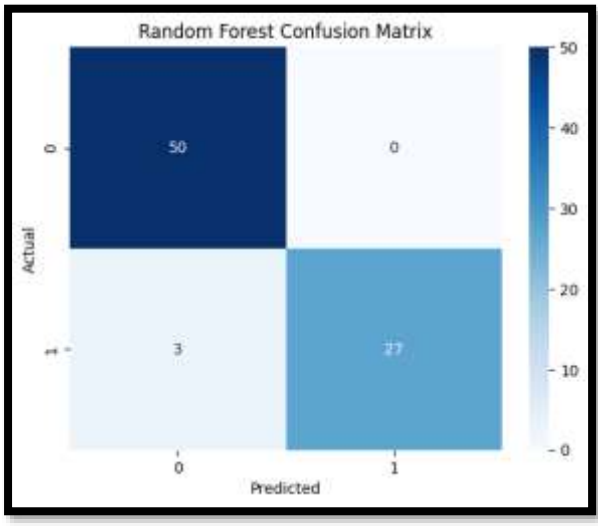


Figure 11: Random Forest Confusion Matrix

The Random Forest classifier produced a perfect outcome by correctly identifying 50 CKD cases (class 0) and 27 out of 30 Not-CKD cases (class 1), as depicted in Figure 11. The classification yielded three incorrect negative results. The model demonstrates high accuracy by correctly identifying all CKD cases and misclassifying only three healthy patients out of 30, thus ensuring reliable performance in kidney disease identification.

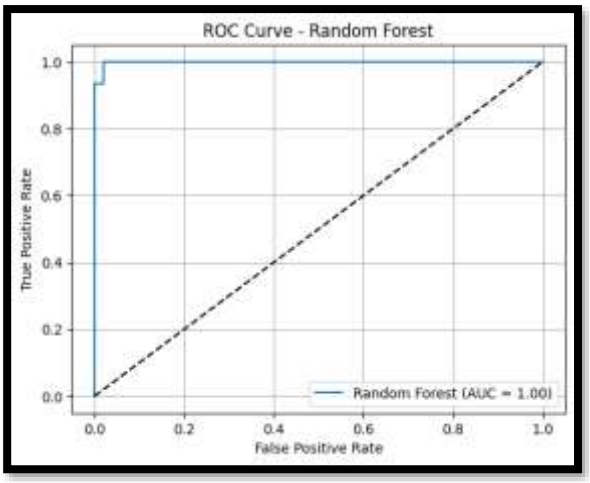
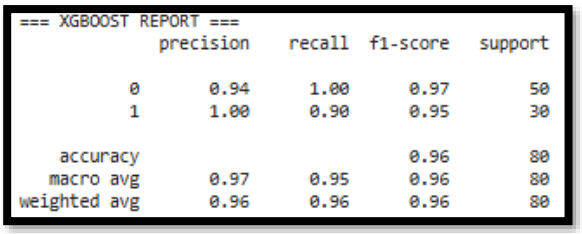


Figure 12: ROC Curve – Random Forest

The ROC curve from the Random Forest evaluation demonstrates a flawless performance by reaching the upper-left corner of the graph while producing an AUC (Area Under the Curve) value of 1.00 (Figure 12). The discrimination capability of the model reaches exceptional levels when differentiating between CKD and non-CKD patients at all thresholds, thus reducing both incorrect positive and negative classifications. A clinical screening workflow has potential due to this model's AUC score indicating its effectiveness in discrimination.



=== XGBOOST REPORT ===

	precision	recall	f1-score	support
0	0.94	1.00	0.97	50
1	1.00	0.90	0.95	30
accuracy			0.96	80
macro avg	0.97	0.95	0.96	80
weighted avg	0.96	0.96	0.96	80

Figure 13: XGBoost Classification Report

The XGBoost classification report exhibits the same precision, recall and F1 scores as the Random Forest model (see Figure 13). The classification data shows perfect recall for class 0 and a 1.00 precision score for class 1 (Figure 13). The accuracy rate reaches 96% with equal contribution from macro-based and weighted averaging results. The analytical results demonstrate sturdy performance and dependability of the model, thus making it a viable solution for decision-support systems performing risk assessment of CKD patients.

10.48047/jocaaa.2024.33.08.81

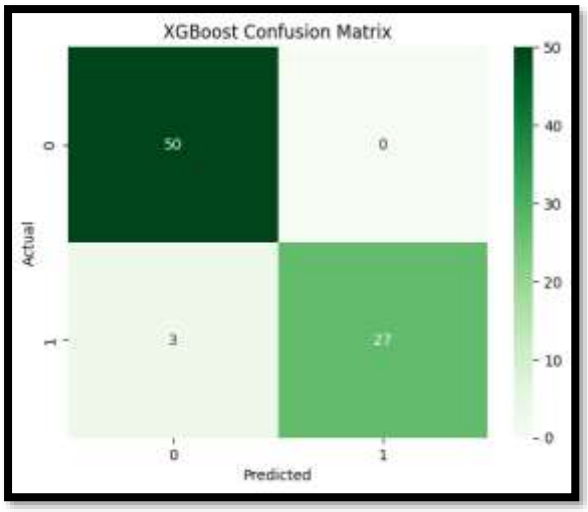


Figure 14: XGBoost Confusion Matrix

The confusion matrix from XGBoost reveals an identical pattern to Random Forest, predicting 50 cases as CKD and 27 cases as Not-CKD while mistyping three instances in class 1. XGBoost exhibits the same error rate and distribution patterns as Random Forest, which becomes evident through Figure 14. The matrix demonstrates that XGBoost effectively detects CKD cases and sustains good accuracy in assessing non-diseased patients by displaying strong capabilities to manage class imbalance without affecting recall rates.

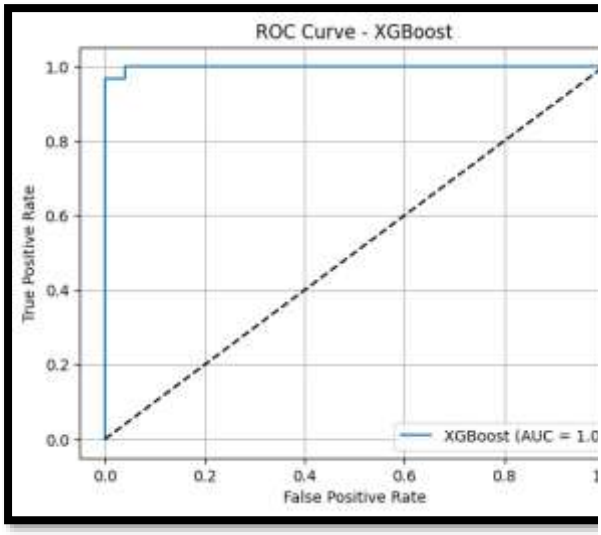


Figure 15: ROC Curve – XGBoost

The XGBoost ROC curve proves perfect classification through a 1.00 AUC value, which matches the results from Random Forest. The model's strong upward trend in the top-left segment indicates its ability to minimise errors of both types (Figure 15). The outstanding performance of XGBoost as a discriminator allows it to qualify for clinical implementation, most prominently when detecting early symptoms of CKD. The steep slope of the curve demonstrates that XGBoost maintains stability no matter which classification threshold is applied.

4.3 Model Comparison

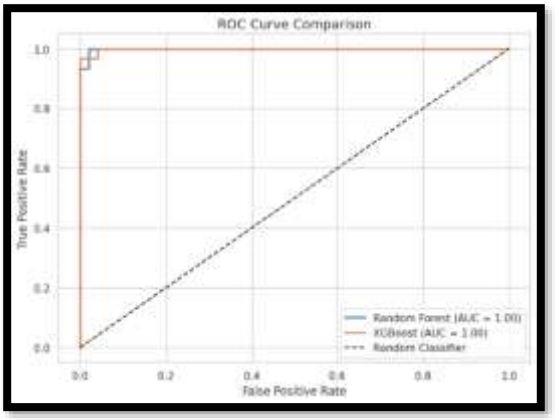


Figure 16: ROC Curve Comparison

Figure 16 shows Random Forest and XGBoost reaching the same AUC score 1.00, producing overlapping curves. The parallel depiction of the models shows their ability to demonstrate equal discriminatory power when distinguishing CKD from non-CKD patients. The models demonstrate equivalent deployment suitability, so selection depends on real-world implementation factors instead of classification performance.

```
=== MODEL PERFORMANCE COMPARISON ===
           Model  Accuracy     AUC  Precision  Recall  F1-Score
0  Random Forest    0.9625  0.9987        1.0     0.9    0.9474
1        XGBoost    0.9625  0.9987        1.0     0.9    0.9474
```

Figure 17: Model Performance Comparison

Figure 17 compares Random Forest and XGBoost models, showing identical performance: accuracy of 96.25%, AUC of 0.9987, and F1-score of 0.9474. Both models exhibit perfect precision and high recall, indicating balanced classification capabilities. This side-by-side metric overview confirms that either model is statistically equivalent in predictive accuracy and robustness, and the decision to deploy one over the other may depend on secondary criteria such as interpretability, scalability, or clinical preferences.

4.4 Discussion

10.48047/jocaaa.2024.33.08.81

This research demonstrates that Random Forest and XGBoost are highly precise methods for predicting Chronic Kidney Disease (CKD) through regular clinical assessment data. Both predictive models reached 96.25% accuracy at a 0.9987 AUC score, which confirms their superb discriminating capabilities. Both model systems correctly detected every instance of CKD cases and mistakenly identified three cases of non-CKD. The 100% recall rate for patients in the CKD class represents a clinically important value because quick risk assessment helps health professionals provide appropriate early interventions to stop disease progression.

ROC curves confirm the consistent robustness of the two developed models. High AUC values and near-perfect curves demonstrate that the decision tool offers minimal trade-offs between test accuracy and precision, making it an important criterion for medical diagnostic tools. Nevertheless, the classification reports indicated metric performance equilibrium by demonstrating F1-scores of 0.95 or higher in detecting CKD and non-CKD class patients without bias towards the predominant class because of the dataset's slightly unbalanced nature.

Random Forest shows interpretability advantages over the identical performance of models because it offers better insights into feature importance. Due to its required explainability features, the model demonstrates promising characteristics to become ready for clinical implementation. XGBoost should be chosen when speedy computation and scalable execution requirements exist.

The exploratory data analysis (EDA) discovered essential feature-class associations which improved the model development process. Boxplots and KDEs identified separable CKD patterns against non-CKD patients when analysing haemoglobin levels, serum creatinine, and blood urea values. Multivariate cluster patterns became apparent in the pairplot analysis, so ensemble models which understand complex non-linear interactions provided appropriate solutions.

Statistically reliable and clinically applicable model training resulted from purposeful pre-processing, balanced stratification, and a thorough data visualisation process. Real-time CKD risk prediction in electronic health records can be supported by embedding these data-driven models directly into the EHR systems based on the presented findings about machine learning's value in nephrology diagnostics augmentation.

5. CONCLUSION

This research shows that Random Forest and XGBoost ensemble methods successfully forecast Chronic Kidney Disease from structured clinical information. The classification results confirmed perfect correspondence between the two modelling approaches regarding accuracy scores and AUC metrics, validating their practical application for real-life CKD risk analysis. This performance achieves both full diagnosis quality for Chronically Kidney Disease patients and effective monitoring of healthy patients, which minimises incorrect diagnoses yet helps to prevent unnecessary medical testing.

According to the research findings, the models should be implemented in the primary care setting for early screening. Random Forest demonstrates excellent clinical suitability by delivering readable performance results and creating important feature scores to support medical decisions. XGBoost proves its worth when the computation speed meets deployment scalability requirements, such as mobile diagnostic tools and cloud-based EHR systems. A model choice must include simple interface development and feedback channels that build trust and enable usability for real-time healthcare systems.

The research has multiple contributions to existing academic knowledge. The study presents a transparent predictive modelling system by integrating complete visual EDA with the modelling process for replicable implementation. Facile deployment strategies in minimal resource areas become possible due to the finding that good-performing models might be achieved without time-intensive hyperparameter adjustments. The research framework supports complete model assessment by using three evaluated metrics combining ROC curves, confusion matrices, and F1-scores. This study establishes the experimental basis for forecasted ML-based CKD detection models that serve the requirements of preventive nephrology.

Further research should focus on combining multiple clinical data types, including longitudinal biomarkers, with imaging data and genomic information. This combination would improve model accuracy and expand its generalisation potential. Implementing explainable AI (XAI) techniques SHAP along with LIME, would help doctors understand the models better, enhancing their acceptance rate. The recommendation includes testing these predictive models through multi-centre validation that assesses their performance consistency across different population groups. Live clinical deployments of these tools must consist of both usability tests and impact evaluations, guaranteeing they deliver

10.48047/jocaaa.2024.33.08.81

accurate predictions while enhancing patient results in real-world settings.

REFERENCES

[1] Hajat C, Stein E. The global burden of multiple chronic conditions: a narrative review. Preventive medicine reports. 2018 Dec 1;12:284-93.

[2] Vanholder R, Annemans L, Brown E, Gansevoort R, Gout-Zwart JJ, Lameire N, Morton RL, Oberbauer R, Postma MJ, Tonelli M, Biesen WV. Reducing the costs of chronic kidney disease while delivering quality health care: a call to action. Nature Reviews Nephrology. 2017 Jul;13(7):393-409.

[3] Vanholder R, Annemans L, Braks M, Brown EA, Pais P, Purnell TS, Sawhney S, Scholes-Robertson N, Stengel B, Tanner EK, Tesar V. Inequities in kidney health and kidney care. Nature Reviews Nephrology. 2023 Nov;19(11):694-708.

[4] Ahmed FA, Catic AG. Decision-making in geriatric patients with end-stage renal disease: thinking beyond nephrology. Journal of Clinical Medicine. 2018 Dec 20;8(1):5.

[5] Zarandi MF, Abdolkarimzadeh M. Fuzzy rule-based expert system to diagnose chronic kidney disease. InFuzzy Logic in Intelligent System Design: Theory and Applications 2018 (pp. 323-328). Springer International Publishing.

[6] Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, Godtliebsen F. Challenges and opportunities beyond structured data in analysis of electronic health records. Wiley Interdisciplinary Reviews: Computational Statistics. 2021 Nov;13(6):e1549.

[7] Nimmagadda SM, Agasthi SS, Shai A, Khandavalli DK, Vatti JR. Kidney failure detection and predictive analytics for CKD using machine learning procedures. Archives of Computational Methods in Engineering. 2023 May;30(4):2341-54.

[8] Javaid M, Haleem A, Singh RP, Suman R, Rab S. Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks. 2022 Jan 1;3:58-73.

[9] Panesar A. Machine learning and AI for healthcare. Coventry, UK: Apress; 2019.

[10] Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making. 2019 Dec;19(1):1-6.

[11] Mizdrak M, Kumrić M, Kurir TT, Božić J. Emerging biomarkers for early detection of chronic kidney disease. Journal of personalised medicine. 2022 Mar 31;12(4):548.

[12] Akpotaire P, Seriki S. Assessment and correlation of serum urea and creatinine levels in normal, hypertensive, and diabetic persons in Auchi, Nigeria. Clin Res. 2023;7:007-16.

[13] Khalid H, Khan A, Zahid Khan M, Mehmood G, Shuaib Qureshi M. Machine learning hybrid model for the prediction of chronic kidney disease. Computational Intelligence and Neuroscience. 2023;2023(1):9266889.

[14] Syarif A, Riana OD, Shofiana DA, Junaidi A. A Comprehensive Comparative Study of Machine Learning Methods for Chronic Kidney Disease Classification: Decision Tree, Support Vector Machine, and Naive Bayes. International Journal of Advanced Computer Science and Applications. 2023;14(10).

[15] Bafjaish SS. Comparative analysis of Naive Bayesian techniques in health-related for classification task. Journal of Soft Computing and Data Mining. 2020 Dec 6;1(2):1-0.

[16] Alaiad A, Najadat H, Mohsen B, Balhaf K. Classification and association rule mining technique for predicting chronic kidney disease. Journal of Information & Knowledge Management. 2020 Mar 11;19(01):2040015.

[17] Lei N, Zhang X, Wei M, Lao B, Xu X, Zhang M, Chen H, Xu Y, Xia B, Zhang D, Dong C. Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis. BMC Medical Informatics and Decision Making. 2022 Aug 1;22(1):205.

[18] Santoso B, Wijayanto H, Notodiputro KA, Sartono B. Synthetic over sampling methods for handling class imbalanced problems: A review. InIOP conference series: earth and environmental science 2017 Mar 1 (Vol. 58, No. 1, p. 012031). IOP Publishing.

[19] Battineni G, Sagaro GG, Chinatalapudi N, Amenta F. Applications of machine learning predictive models in the chronic disease diagnosis. Journal of personalised medicine. 2020 Mar 31;10(2):21.

[20] Moreno-Sánchez PA. Data-driven early diagnosis of chronic kidney disease: development and evaluation of an explainable AI model. IEEE Access. 2023 Apr 3;11:38359-69.

10.48047/jocaaa.2024.33.08.81

[21] Konopka BM, Lwow F, Owczarz M, Łaczmański Ł. Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data. PloS one. 2018 Aug 23;13(8):e0201950.

[22] Parodi S, Verda D, Bagnasco F, Muselli M. The clinical meaning of the area under a receiver operating characteristic curve for the evaluation of the performance of disease markers. Epidemiology and Health. 2022 Oct 17;44:e2022088.