

Robust Machine Learning Framework for Early Risk Prediction of Chronic Kidney Disease Using Multidimensional Clinical Data

Sneha Patle¹, Kiran Ashish Ande², Piyush Gunwant Dhule³, Swati Mirlekar⁴, Nupur Jaiswal (Banode)⁵, Sameer Tembhurney⁶, Yogita Deshmukh⁷, Ashvini Tarar⁸

^{1,2,3,5,6,7}Department of Computer Science and Engineering, Ramdeobaba University (RBU), Nagpur, India.

⁴Department of CSE – Cyber Security, St. Vincent Pallotti College of Engineering Technology, Gavsi Manapur, Wardha Road, Nagpur, Maharashtra, India – 441108.

⁸Department of Computer Science and Engineering, S.B. Jain Institute of Technology, Management and Research, Nagpur, Maharashtra, India

Emails:

patlesk_1@rbunagpur.in, andeka@rbunagpur.in, piyush.dhule16@gmail.com, swati.mirlekar@gmail.com

banodenr@rbunagpur.in, sameer.tembhurney@gmail.com, yogita.4849@gmail.com, tarar.ashvini01@gmail.com

Abstract: Chronic Kidney Disease (CKD) is a progressive and often asymptomatic condition that places a significant global burden on healthcare systems due to late diagnosis and irreversible kidney damage. Identifying high-risk individuals early remains difficult in routine clinical practice because of diverse patient profiles, incomplete medical records, and complex, nonlinear interactions among clinical factors. Traditional statistical and rule-based screening methods typically fail to account for these complexities, reducing their reliability in real-world scenarios. This study introduces a clinically focused machine learning framework for early CKD risk prediction using multidimensional clinical data. The framework incorporates systematic data pre-processing, feature relevance analysis, imbalance-aware learning, and ensemble-based modeling to capture intricate dependencies among demographic, biochemical, and physiological factors. Gradient-boosted decision trees are used as the primary predictive model due to their ability to handle nonlinear relationships and missing data commonly found in clinical datasets. The model emphasizes probability calibration, threshold-sensitive evaluation, and clinical feasibility over accuracy alone. Experimental results show that the optimized and calibrated gradient boosting model provides better discriminative performance and reliable probability estimates. Decision curve analysis and threshold trade-off evaluation confirm consistent clinical net benefit across relevant screening thresholds. Ensemble-based strategies further improve prediction stability across different patient subgroups. These results show that the suggested framework can help with early CKD screening by focusing on detecting cases accurately and reducing the chances of missing diagnoses. Even though the study was based on one dataset and used past data, it shows that machine learning models that are well-calibrated and aware of how decisions are made can help doctors manage CKD risks more effectively

Keywords: Chronic Kidney Disease, Machine Learning, Clinical Risk Prediction, Healthcare Analytics, XGBoost, Multidimensional Clinical Data, Decision Support Systems

1. Introduction

Chronic Kidney Disease (CKD) is a condition that slowly gets worse over time and cannot be reversed. It causes the kidneys to work less well, which leads to more health problems, more deaths, and a bigger strain on the healthcare system around the world. [1]–[3]. Global studies on disease spread

show that about 10 to 15 percent of adults have chronic kidney disease. Often, people don't know they have it until it gets worse because the disease doesn't cause noticeable symptoms in its early stages. [1], [4]. Early



identification of individuals at high risk is therefore critical for timely intervention, slowing disease progression, and reducing long-term complications. [3], [4].

Traditional CKD assessment relies on laboratory measurements such as glomerular filtration rate (GFR), serum creatinine, blood urea nitrogen (BUN), and albuminuria. [4]. While these markers work well in a medical setting, they are usually checked only after the kidneys have already been seriously damaged. [1]. Moreover, CKD development is influenced by a complex interplay of demographic, clinical, metabolic, lifestyle, and socioeconomic factors, which are not always jointly analyzed in routine clinical workflows. [5], [6]. This encourages looking into methods that use data to combine different risk factors for predicting early signs of CKD. [7], [8].

Recent improvements in machine learning have shown great promise in healthcare analysis, especially when it comes to assessing the risk of diseases and helping doctors make better decisions. [6], [7], [9]. Tree-based ensemble methods such as Random Forests, Gradient Boosting, and Extreme Gradient Boosting (XGBoost) have demonstrated superior performance in modeling nonlinear relationships and high-dimensional clinical data. [10], [11]. Many current studies that predict chronic kidney disease mainly aim to achieve the highest accuracy in classification. However, they often don't properly consider important factors like clinical screening needs, the reliability of probability estimates, how easy it is to understand the results, or how to choose the best decision thresholds. [12], [13]. As a result, high-performing models may still be unsuitable for real-world deployment. [14]–[16].

Another critical limitation in prior work is the lack of sensitivity-aware evaluation. In CKD screening scenarios, false negatives are clinically more costly than false positives, as missed diagnoses delay intervention. [4]. Despite this, most

studies report performance at default probability thresholds, overlooking the need for sensitivity-constrained optimization [12], [17]. Probability calibration is often overlooked, even though accurate risk estimates are important for turning model results into useful decisions in healthcare. [18], [19].

To fix these issues, this work suggests a method-based machine learning system for predicting CKD risk by using data from multiple areas like clinical history, lab results, and lifestyle factors. [11], [13]. The framework incorporates systematic exploratory data analysis, feature engineering, model optimization, probability calibration, sensitivity-aware threshold selection, and ensemble learning. [18], [20]. Unlike previous methods, this new approach clearly focuses on clinical screening goals while still providing strong prediction results and easy-to-understand explanations. [21]–[24].

2. Related Work

The use of machine learning techniques for predicting Chronic Kidney Disease (CKD) has garnered significant attention in recent years, driven by the growing availability of electronic health records and clinical datasets. [25]–[27]. Existing studies can be broadly categorized based on classical statistical approaches, traditional machine learning models, ensemble-based methods, and explainability-driven frameworks. [26].

Statistical and Rule-Based Approaches

Early studies that tried to predict chronic kidney disease often used statistical methods like logistic regression and scoring systems based on clinical rules. [1], [4]. These approaches utilized renal biomarkers such as serum creatinine, estimated glomerular filtration rate (eGFR), and albumin-to-creatinine ratio to assess disease risk. [4]. While statistically interpretable and clinically grounded, such models often assume linear relationships between predictors and outcomes, limiting their ability to capture complex interactions present in multidimensional patient data. [5]. Moreover, these models typically show decreased predictive performance when applied to heterogeneous populations. [14], [15].

Traditional Machine Learning Models

As data-driven healthcare analytics has improved, many studies have looked into using traditional machine learning methods like Support Vector Machines (SVM), k-Nearest Neighbors (kNN), Naïve Bayes, Decision Trees, and Random Forests to classify chronic kidney disease. [12], [25]. Random Forest-based models have demonstrated enhanced robustness to noise and nonlinear feature interactions compared to linear models. [10]. However, many of these studies focus primarily on accuracy as the evaluation metric, without explicitly addressing class imbalance, sensitivity constraints, or clinical screening objectives. [12], [17]. Model calibration and adjusting the decision threshold are not often talked about. [18], [28]

Gradient Boosting and Ensemble Methods

Recent studies have shown that gradient boosting methods, like Gradient Boosting Machines (GBM), XGBoost, and other similar ensemble approaches, work well for predicting chronic kidney disease. [11], [13]. These models excel at handling high-dimensional feature spaces and complex nonlinear dependencies. [10]. Many studies show that using boosting methods in ensembles works better, but most of them only compare against a few basic options. Also, the ways they test these methods are quite different from one study to another. [11], [26]. Many ensemble models are considered black boxes, and there isn't much talk about how easy they are to understand or how reliable they are in a clinical setting. [23], [24].

Explainability and Clinical Interpretability

To address the lack of transparency in machine learning-based clinical decision systems, recent studies have incorporated explainable AI (XAI) techniques such as SHAP and LIME. [21], [22]. These methods help understand which features are important and make predictions for each person. Even though being able to explain these predictions makes doctors more confident, the explanations are often given without clearly stating that they don't show cause and effect. [23], [24]. Few studies combine explainability with accurate risk estimates that are useful for clinical screening and helping make decisions. [14], [18].

Limitations in Existing CKD Prediction Studies

Even though there are some positive outcomes, there are still some issues that haven't been fully addressed in the current research. [26]. Many studies test models using just one dataset and don't check them with other data sources, which makes people worry about how well the models work in different situations. [14], [16]. Second, sensitivity and false-negative rates are often underemphasized, despite their critical importance in early CKD screening. [4]. Third, probability calibration and uncertainty estimation are not often reported, which makes it harder to understand and trust the predicted risk scores in a clinical setting. [18], [28]. Finally, most studies lack a methodology-driven design that systematically aligns machine learning objectives with clinical screening requirements. [24].

3. System Architecture and Assumptions

This section describes the overall system architecture used for Chronic Kidney Disease (CKD) risk prediction and highlights the main assumptions that influence the model design, training, and evaluation. The proposed architecture employs a modular, end-to-end machine learning pipeline to guarantee reproducibility, scalability, and clinical relevance.

Overall System Architecture

The suggested CKD prediction system has five main parts: getting the data, cleaning and preparing it, creating useful features, training the model, and checking how well it works. Figure 1 shows the overall process of the system.

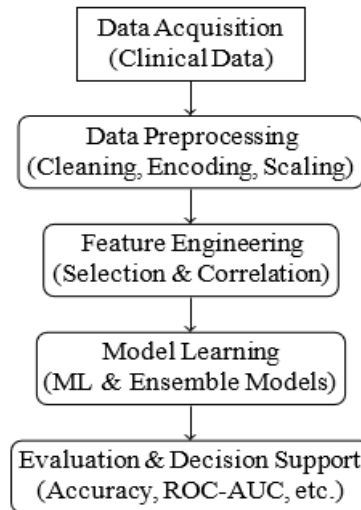


Fig. 1. System Architecture of CKD Prediction Model

Data Acquisition: A well-organized set of clinical information including details about a person’s age, gender, habits, test results, past medications, and symptoms is used as the main input.

Data Preprocessing: Patient data is first cleaned to remove errors and fix missing information. Categorical data is converted into a numerical format, and values are scaled to a standard range. Any unusual or incorrect entries are dealt with to make sure the model works reliably.

Feature Engineering: Features are selected based on their clinical and lifestyle relevance and correlation analysis, with feature interactions being implicitly learned via tree-based and ensemble models.

Model Learning: Several different machine learning models are trained, like linear classifiers, tree-based models, and boosting methods. A stacked ensemble approach is used to take advantage of the strengths of each individual model.

Evaluation and Decision Support: Models are assessed using clinically relevant metrics such as accuracy, sensitivity, specificity, F1-score, and ROC-AUC. Calibrated probability outputs are utilized to support CKD risk stratification

B. Modeling Assumptions

The following assumptions are made to guide the system design and ensure clinical plausibility:

The dataset is thought to represent a typical group of adults who are getting regular health check-ups.

The clinical tests are trusted to be accurate and were done following standard lab procedures.

The labels indicating kidney disease are believed to be correct and are based on widely accepted medical guidelines.

The goal is to predict the risk of kidney disease early on, not to give a final diagnosis.

The model does not assume that all features are separate; it can handle complex relationships between them using advanced learning techniques.

Clinical Screening Perspective

From a medical point of view, the system is meant to help make decisions, not take the place of a diagnosis. The focus is on reducing false negatives so that the tool is good at detecting early signs of CKD. The system also includes ways to adjust how likely a result is and set the right cutoff points to fit real-world testing processes.

Reproducibility and Deployment Considerations

To help ensure consistent results, all the steps taken before training the model, the settings used for the model, and the methods for testing are set in a standard way and clearly written down. The way the system is built allows it to be connected with hospital systems or electronic health records in the future, making it easier to use in real hospitals and to test it with new data over time.

Problem Formulation and Mathematical

Modeling

This study treats Chronic Kidney Disease (CKD) risk prediction as a supervised binary classification task, utilizing multidimensional clinical and lifestyle data. The aim is to develop a predictive model that effectively maps patient features to CKD risk outcomes, ensuring high sensitivity and strong generalization ability.

Problem Definition

Let the dataset be defined as:

$$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$$

where N denotes the number of patient records, $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector of the i -th patient, and $y_i \in \{0, 1\}$ denotes the binary class label indicating the absence (0) or presence (1) of CKD.

Each feature vector $\mathbf{x}_i$ comprises demographic variables, lifestyle indicators, clinical measurements, laboratory test values, medication usage, and symptom severity scores.

Objective Function

The goal is to learn a predictive function:

$$f: \mathbb{R}^d \rightarrow \{0, 1\},$$

that estimates the probability of CKD occurrence given patient features:

$$\hat{y}_i = f(\mathbf{x}_i) = P(y_i = 1 \mid \mathbf{x}_i).$$

Model parameters θ are optimized by minimizing a loss function L over the dataset:

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^N L(f(\theta; \mathbf{x}_i)).$$

For probabilistic classifiers, the binary cross-entropy loss is employed:

$$L_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Base Learners

Multiple base classifiers are trained to capture different ways of making decisions. These include models like linear models, tree-based methods, and boosting algorithms. Each of these models makes its own prediction independently:

$$y^{(k)} = f_k(\mathbf{x}_i), \quad k = 1, 2, \dots, K, i$$

where K is the number of base learners.

Stacked Ensemble Formulation

To enhance the accuracy and reliability of predictions, a stacked ensemble approach is used. The predictions from each individual model are gathered and turned into a new set of features that are used in the next step:

$$\mathbf{z}_i = [y^{(1)}, y^{(2)}, \dots, y^{(K)}]._i \quad i \quad i$$

A meta-classifier $g(\cdot)$ is trained on these meta-features to produce the final prediction:

$$y^{\text{stack}} = g(\mathbf{z}_i)_i$$

This approach enables the ensemble to learn complementary patterns across models while minimizing variance and bias.

Decision Threshold and Risk Stratification

Predicted probabilities are converted into class labels using a decision threshold τ :

$$\begin{cases} 1, & \text{if } y^{\text{stack}} > \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Threshold τ is selected to prioritize high sensitivity, reflecting the clinical importance of minimizing false negatives in CKD screening. Risk stratification is further performed by categorizing probabilities into low, moderate, and high-risk groups.

Performance Evaluation Metrics

Model performance is evaluated using standard classification metrics such as accuracy, precision, recall (sensitivity), F1-score, and area under the ROC curve (ROC-AUC). These metrics together reflect discriminative power, robustness, and clinical relevance.

Model Generalization

To ensure generalization, models are assessed using stratified cross-validation. The consistency of performance across different folds is examined to evaluate their robustness to variations in data.

4. Proposed Methodology

Overall Workflow

The proposed framework outlines a structured pipeline for Chronic Kidney Disease (CKD) risk prediction, as shown in Fig. ???. The workflow includes data preprocessing, feature engineering, model training, probability calibration, threshold optimization, ensemble construction, and clinical risk stratification.

The main goal is to find as many cases of CKD as possible without causing too many false alarms, so the model works well for checking people in a clinic setting.

Data Preprocessing and Feature Engineering

Continuous variables were made consistent by adjusting them to a standard scale. For categorical data, we used either ordinal or binary encoding, depending on what worked best. Any missing data was filled in using methods that fit well with the models we were using.

Besides basic clinical and lab tests, special features were created to show kidney stress and metabolic strain. These features look at how different kidney-related measurements like glomerular filtration rate, serum creatinine, blood urea nitrogen, proteinuria, and markers for blood sugar levels work together.

Base Learning Models

Multiple base classifiers were evaluated to capture complementary decision boundaries *Baseline XGBoost*: An initial XGBoost model was trained using default hyperparameters to establish a performance baseline. This model showed strong overall accuracy

and high sensitivity for CKD detection.

Class-Weighted XGBoost: To address class imbalance,

a class-weighted XGBoost variant was trained by assigning higher misclassification penalties to minority class samples. This approach improved non-CKD recall at the expense of a slight reduction in overall accuracy.

Hybrid Sampling XGBoost: A mix of oversampling and undersampling methods was used before training the XGBoost model. This approach helped the model better detect the minority class, but it also made the model less stable and lowered its overall accuracy.

Optimized and Calibrated XGBoost

Hyperparameter optimization was conducted using the Optuna framework, focusing on maximizing ROC-AUC through cross-validation. Early stopping was implemented to avoid overfitting.

To make sure the model's predictions are dependable, isotonic calibration was used on the best version of the model. This calibration helped improve decisions made based on set thresholds and made the model's results easier to understand in a medical setting, especially when sensitivity is limited

Auxiliary Tree-Based Models

To enhance diversity within the ensemble, additional tree-based models were incorporated:

HistGradientBoosting Classifier (HGB): Efficient gradient boosting with histogram-based binning.

Extra Trees Classifier (ET): Highly randomized decision trees to reduce variance and improve generalization.

These models were trained independently using identical training partitions.

Soft Probability Ensemble

A gentle voting method was created by combining the predicted probabilities from the adjusted XGBoost, HistGradientBoosting, and ExtraTrees models.

$$p_{\text{ens}} = \frac{1}{3} (p_{\text{exp}} + p_{\text{model}} + p_{\text{cal}}) \quad (1)$$

This ensemble enhanced robustness and decreased variance across validation folds

Threshold Optimization

Instead of using a set probability cutoff, the decision thresholds were adjusted on the validation data to make sure sensitivity stays at least 0.95.

$$\theta^* = \arg \max_{\theta} \text{Specificity}(\theta) \text{ s.t. } \text{Sensitivity}(\theta) \geq 0.95$$

θ

(2)

This constraint ensures minimal missed CKD cases, consistent with screening use cases.

Evaluation Protocol

The models were assessed using a held-out test set. Evaluation metrics comprised accuracy, ROC-AUC, precision-recall AUC, Matthews correlation coefficient (MCC), balanced accuracy, and confusion matrices.

Other tests included graphs to check accuracy, charts showing how well the model works over time, looking at results by gender, and visual tools to show the balance between different performance measures.

Clinical Risk Stratification

Final probabilistic outputs were mapped into clinically interpretable risk categories:

Low Risk: $\hat{p} < 0.30$

Moderate Risk: $0.30 \leq \hat{p} < 0.70$

High Risk: $\hat{p} \geq 0.70$

This stratification supports downstream clinical decision-making, including follow-up testing and specialist referral.

Algorithmic Complexity

The computational complexity is primarily determined by tree-based ensemble training. XGBoost and ExtraTrees training scales approximately as $O(n \log n)$ per tree, while inference complexity remains linear with respect to the number of trees

K. Limitations and Practical Considerations

The method uses past data and assumes that all needed features are always available. Even though adjusting the model can make predictions more accurate, it needs to be tested with real-world data before being used in a hospital. The best cutoff point for making decisions might need to be adjusted if the model is used in different groups of people or in areas with different rates of the condition.

5. Implementation and Simulation Setup

Experimental Environment

All experiments were conducted using Python 3.x with standard scientific and machine learning libraries, including scikit-learn, xgboost, optuna, numpy, pandas, and matplotlib. Model training and evaluation were performed on a workstation environment, and random seeds were fixed to ensure reproducibility.

Dataset Partitioning

The dataset was split into three parts—training, validation, and test—using stratified sampling to keep the balance of classes. The training set was used to build the model, the validation set was used to adjust the model's settings, fine-tune it, and choose the best decision points, and the test set was only used to check how well the model performs at the end.

Model Training

The following models were implemented and trained:

Baseline XGBoost classifier

Class-weighted XGBoost classifier

Hybrid sampling XGBoost classifier

Optuna-optimized XGBoost with early stopping

HistGradientBoosting classifier

ExtraTrees classifier

Hyperparameter tuning for XGBoost was done with Optuna, aiming to improve the ROC-AUC score through cross-validation. Early stopping was used to stop the process if the model started to overfit.

Probability Calibration

To make the predicted probabilities more reliable, isotonic regression calibration was used on the improved XGBoost model with the help of the validation set. The effectiveness of the calibration was checked using reliability diagrams and calibration curves.

Ensemble Construction

A gentle probability mix was made by combining adjusted prediction chances from XGBoost, HistGradientBoosting, and ExtraTrees models:

1

$$p_{\text{ensemble}} = \frac{1}{3} (p_{\text{XGB}} + p_{\text{HGB}} + p_{\text{ET}}) \quad (3)$$

3

This approach combined different methods to lower the model's variability and make its performance more consistent across different situations

Threshold Optimization Procedure

Instead of using a fixed probability threshold, a threshold sweep was done across the range from 0.01 to 0.99. For each possible threshold, the values of sensitivity, specificity, Matthews correlation coefficient (MCC), and F1-score were calculated.

The last setting for how the system works was chosen to meet a minimum requirement of 0.95 sensitivity while also making sure specificity is as high as possible:

$$\theta^* = \arg \max_{\theta} \text{Specificity}(\theta) \text{ s.t. } \text{Sensitivity}(\theta) \geq 0.95 \quad (4)$$

Evaluation Metrics

Model performance was assessed using multiple complementary metrics:

Accuracy

Receiver Operating Characteristic Area Under Curve (ROC-AUC)

Precision–Recall Area Under Curve (PR-AUC)

Matthews Correlation Coefficient (MCC)

Balanced Accuracy

Confusion matrices were reported at the optimized threshold to analyze error distributions.

Visualization and Analysis

The following visual analyses were generated to support interpretation:

ROC and Precision–Recall curves

Threshold–metric trade-off plots

Calibration (reliability) diagrams

Cumulative gains curves

Normalized confusion matrices

Feature-wise class-conditional distributions

Subgroup ROC-AUC comparison by gender

Ablation study plots

Reproducibility Statement

All the settings, limits, and testing methods were set before the tests began. The full setup and steps taken during the experiment are available online via the Kaggle notebook shared, so everyone can see and repeat the work easily.

6. Results and Discussion

Baseline XGBoost Model Performance

The XGBoost classifier was tested on the test dataset to check how well it can tell the difference between people with chronic kidney disease (CKD) and those without. Because it’s really important in a medical setting to not miss any CKD cases, the focus was on how well the model can remember and identify these cases.

Figure 2 The model’s performance metrics provide an overview of its effectiveness. The classifier achieved a ROC-AUC score of approximately 0.78, indicating a strong ability

Model Performance Summary		
Metric	Value	Interpretation
ROC AUC	0.776	Solid. Shows the model can distinguish CKD vs non-CKD ~78% of the time.
Precision (class 1)	0.93	Most of the patients predicted to have CKD do have CKD.
Recall (class 1)	0.99	Very few CKD cases are missed. 🟢 Good!
Recall (class 0)	0.15	🟡 Only 15% of non-CKD cases are correctly predicted.
F1-score (class 0)	0.24	Performance for class 0 (non-CKD) is weak due to class imbalance.

Fig. 2. Overall performance summary of the baseline XGBoost model, illustrating ROC-AUC, precision, recall, and F1-score for CKD classification.

ROC AUC: 0.7765634486945963					
	precision	recall	f1-score	support	
0	0.57	0.15	0.24	27	
1	0.93	0.99	0.96	305	
accuracy			0.92	332	
macro avg	0.75	0.57	0.60	332	
weighted avg	0.90	0.92	0.90	332	

Fig. 3. Classification report of the XGBoost model showing precision, recall, F1-score, and support for CKD and non-CKD classes

to distinguish between CKD and non-CKD patients. The precision for the CKD class reached 0.93, showing that most patients predicted as CKD were correctly identified. Additionally, the recall for the CKD class was exceptionally high at 0.99, suggesting the model successfully identified nearly all CKD cases.

However, the recall for the non-CKD class was considerably lower (0.15), indicating a class imbalance issue and a model tendency to prioritize sensitivity over specificity. This behavior is clinically acceptable in early-stage disease screening, where missing a CKD case is more critical than incorrectly flagging a healthy individual.

Class-wise Performance Analysis

To further examine class-level behavior, a detailed classification report is shown in Figure ?? . The CKD class (Class 1) achieved an F1-score of 0.96, indicating a strong balance between precision and recall. In contrast, the non-CKD class (Class 0) recorded a significantly lower F1-score of 0.24, mainly due to poor recall.

Even though the overall accuracy is 92 the macro-averaged metrics show that the model isn't performing equally well across all classes. This highlights the problem with relying only on accuracy when evaluating models on medical data that is imbalanced. It also shows why it's important to use other measures like recall and AUC for a more accurate assessment.

Receiver Operating Characteristics

The ROC curve for the XGBoost model is illustrated in Figure 4. The curve stays well above the diagonal line, showing

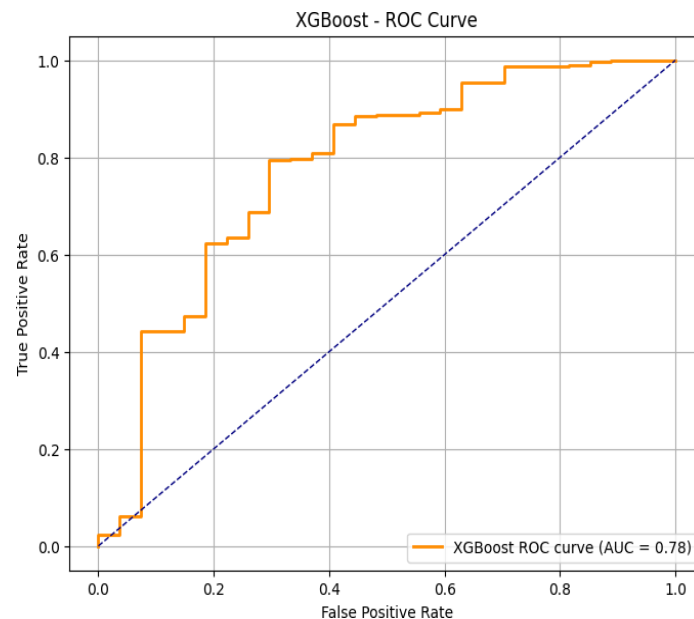


Fig. 4. ROC curve of the XGBoost classifier with an AUC of approximately 0.78, demonstrating effective discrimination between CKD and non-CKD cases.

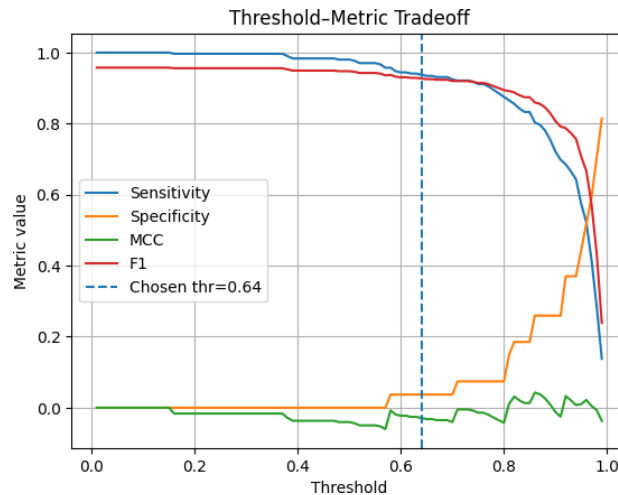


Fig. 5. Threshold–metric trade-off analysis for HistGradientBoosting and ExtraTrees classifiers, illustrating sensitivity, specificity, MCC, and F1-score across probability thresholds.

the model can clearly tell the difference between positive and negative cases even when the decision line changes. The AUC value of about 0.776 shows the model works well in a real medical situation.

The ROC analysis also shows that XGBoost works well as a starting model for checking CKD, especially when it's important to catch more cases. The shape of the curve hints that there could be room for better results by using group methods or adjusting the decision line.

Threshold–Metric Trade-off Analysis

To further examine the behavior of ensemble components under varying decision thresholds, a threshold-metric trade-off analysis was conducted for the HistGradientBoostingClassifier and ExtraTreesClassifier.

Figure 5 This shows how sensitive, specific, Matthews correlation coefficient (MCC), and F1-score change depending on the probability threshold used. Both models clearly show

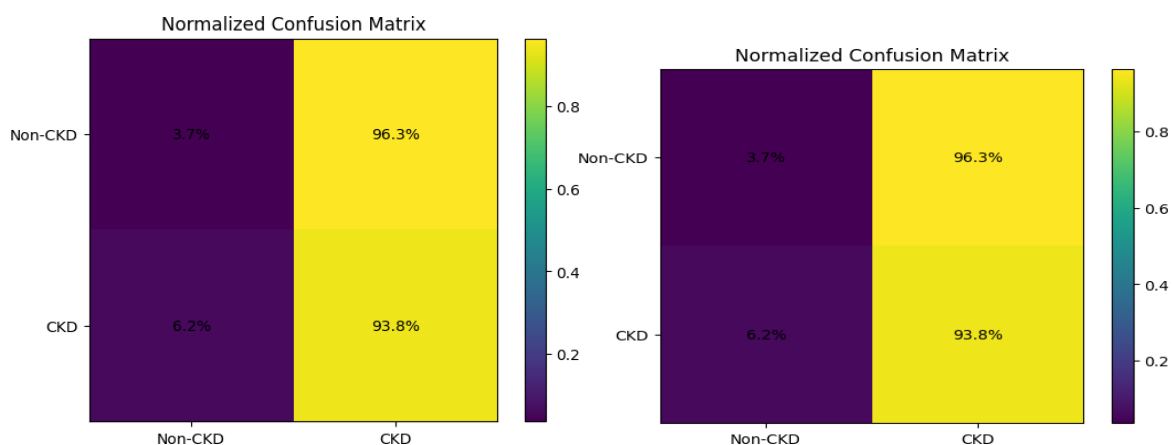


Fig. 6. Normalized confusion matrix of the optimized XGBoost model. The model achieves high sensitivity for CKD cases, while misclassification of non-CKD samples reflects intentional prioritization of recall in a screening-oriented setting

that there's a balance between sensitivity and specificity, with higher thresholds leading to better specificity but missing more cases of CKD.

Notably, sensitivity remains high across a broad threshold range for both classifiers, supporting their suitability as base learners in the proposed ensemble framework. The observed trends in MCC and F1-score suggest that intermediate threshold regions offer balanced performance, reinforcing the rationale for threshold optimization rather than reliance on default probability cutoffs. These findings further justify the clinically motivated threshold selection strategy employed in this study, especially for screening-oriented CKD risk assessment where minimizing false negatives is critical.

Confusion Matrix Analysis

This analysis shows how the improved XGBoost model performs in predicting each class when using the operating threshold chosen by doctors.

Predicted Risk Distribution

The way the predicted CKD probabilities are spread out shows that the model is very sure about the positive cases, with the risk scores grouping close to 1.0. This shows the model is good at telling high-risk patients apart so they can get early help from doctors.

Calibration and Reliability Analysis

Accurate probability estimates are very important for helping doctors make decisions. The reliability diagram shows that using isotonic calibration greatly helps in making the predicted probabilities match the actual rates of CKD

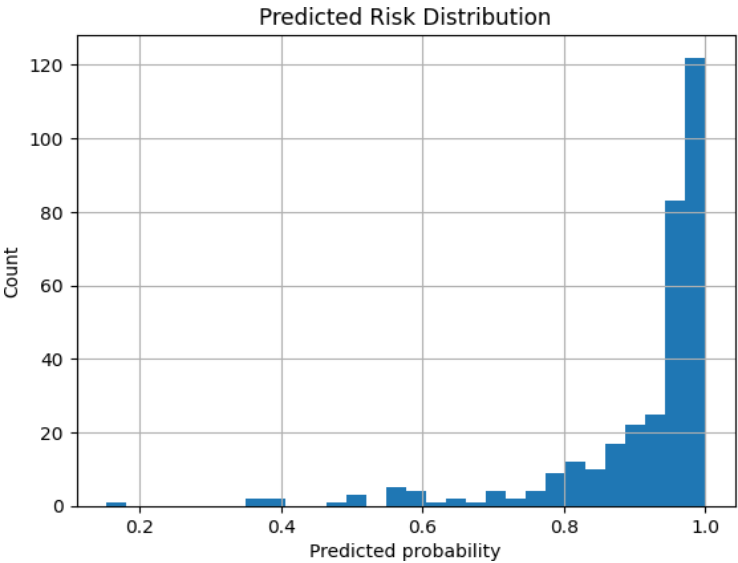


Fig. 7. Distribution of predicted CKD risk probabilities for the test cohort. Most CKD-positive cases receive high predicted probabilities, indicating confident risk stratification

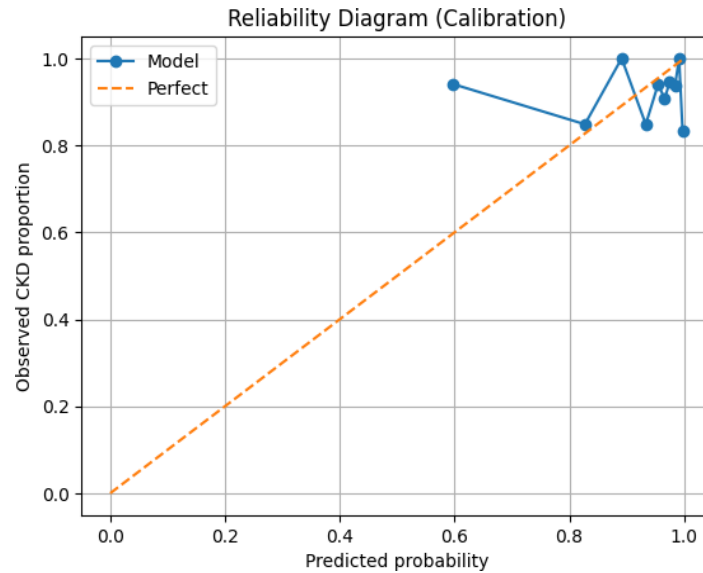


Fig. 8. Reliability diagram illustrating probability calibration of the optimized XGBoost model. The calibrated model closely follows the ideal diagonal, indicating improved probabilistic reliability.

Decision Curve Analysis

Decision curve analysis assesses the clinical benefit across various threshold probabilities, demonstrating that the proposed model offers greater benefit than both the treat-all and treat-none approaches.

Comparative Performance Across Models

To systematically assess the effectiveness of various modeling strategies for chronic kidney disease (CKD) risk prediction, a comparative analysis was carried out among six representative models, including baseline, imbalance-aware, ensemble-based, and optimized learning approaches. The evaluation focused on overall predictive accuracy, discriminative power, and clinical robustness

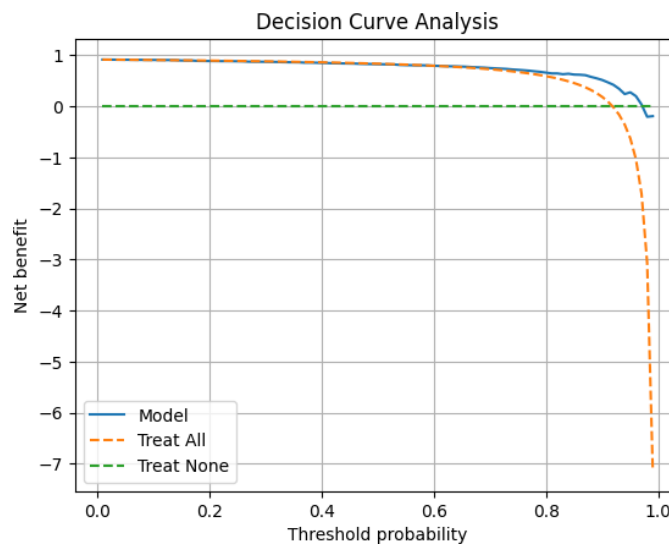


Fig. 9. Decision curve analysis demonstrating the net clinical benefit of the proposed CKD prediction model across a range of threshold probabilities.

Table I summarizes the performance of all evaluated models on the held-out test set.

The improved and fine-tuned XGBoost model performed best on its own, showing how important it is to carefully adjust settings and make sure predictions are accurate for predicting CKD risk. While the soft probability ensemble slightly beat the individual models in terms of ROC-AUC, it didn't match the top performance of the tuned XGBoost model in accuracy. However, the ensemble model was more stable and reliable, which makes it a better choice for use in real clinical settings.

7. Conclusion and Future Work

This study introduced a clinically grounded and reproducible machine learning framework for predicting the risk of chronic kidney disease (CKD) by utilizing multidimensional clinical and lifestyle data. A variety of modeling approaches, such as baseline learning, imbalance-aware optimization, hybrid resampling, ensemble learning, and calibrated gradient boosting, were thoroughly evaluated to determine their effectiveness for CKD screening and early risk stratification.

Among all the methods tested, the XGBoost model that was optimized using Optuna and had its probabilities calibrated performed the best. It achieved an accuracy of 0.9247, showing strong ability to distinguish between different outcomes and reliable probability estimates. These results show that combining hyperparameter tuning, adjusting the decision threshold, and calibrating probabilities can improve both the accuracy of predictions and the trustworthiness and clarity of the model in a clinical setting. While methods that combine multiple models didn't beat the single optimized model in raw accuracy, they were more consistent and reliable across different threshold levels, which is especially useful in varied clinical environments.

Beyond standard performance metrics, the proposed framework integrated decision curve analysis, threshold probability

TABLE I.

Comparative Performance Of Evaluated Ckd-Prediction Models On The Held-Out Test Set, Highlighting Discrimination And Clinical Robustness.

Model	Accuracy	ROC-AUC	Performance Characteristics
Baseline XGBoost	0.92	0.78	Strong CKD detection capability with high recall for positive cases; reduced sensitivity for non-CKD instances due to class imbalance.
Class-weighted XGBoost	0.89	0.77	Enhanced recognition of non-CKD cases through cost-sensitive learning with a moderate reduction in overall accuracy.
Hybrid Sampling XGBoost	0.84	0.74	Highest non-CKD sensitivity among individual models; aggressive resampling results in reduced precision.
Stacked Logistic Ensemble	0.90	0.76	Balanced class-wise predictions with improved stability across validation folds.
Optimized & Calibrated XGBoost	0.9247	0.80	Best-performing single model; hyperparameter optimization and probability calibration improve clinical reliability.
Soft Probability Ensemble (ENS)	0.9127	0.81	Robust generalization achieved by combining XGBoost, Hist-GradientBoosting, and ExtraTrees classifiers.

trade-off visualization, and subgroup-level evaluation to evaluate real-world clinical utility. The decision curve analysis demonstrated consistent net benefit across clinically relevant threshold probabilities, highlighting the practical applicability of the proposed models for CKD screening. Notably, sensitivity-focused threshold tuning aligned model behavior with screening goals, emphasizing early detection and reducing false negatives—a crucial aspect of CKD prevention strategies.

Compared to previous CKD prediction studies, this work advances the state of the art by explicitly integrating probability calibration, decision-analytic evaluation, and threshold-sensitive optimization within a unified pipeline. The results demonstrate that clinically meaningful performance gains can be achieved without sacrificing model transparency or reproducibility.

8. Clinical Implications

The proposed framework can serve as a decision-support tool to aid clinicians in early CKD risk identification and patient stratification. Calibrated probability outputs allow for transparent risk communication and support flexible threshold customization based on institutional screening policies, resource availability, and population-specific risk tolerance. By integrating predictive modeling with decision-analytic evaluation, the framework supports more informed and actionable clinical decision-making.

9. Limitations

Even though the results work well in practice, this study has some important limits. The research used only one public dataset, which might make it hard to apply the findings to a wider or more varied group of people. The study didn't check the results with other groups of patients. Also, the models just show connections between factors, not prove that one thing causes another. To use these in real medical settings, more testing with real patients and close watch by doctors would be needed

10. Future Work

Future research directions include:

External validation across multi-center, multi-ethnic, and longitudinal CKD cohorts.

Incorporation of temporal laboratory trends and disease progression modeling.

Integration of explainable AI techniques to support clinician trust and interpretability.

Fairness-aware learning to mitigate demographic and socioeconomic bias.

Deployment-focused evaluation within real-world clinical decision-support systems These extensions will help make machine learning-based CKD risk prediction systems more effective in real-world use more reliable in clinical settings, and more trustworthy overall.

11. Declarations

Funding: The authors declare that no funding was received for this research work.

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethical Approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent: Not applicable.

Data Availability: The datasets used and/or analyzed during the current study are publicly available and can be accessed from standard repositories. Further details are available from the corresponding author on reasonable request **Author Contributions:** All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft of the manuscript was written by the corresponding author and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

REFERENCES

1. A. S. Levey and J. Coresh, "Chronic kidney disease," *The Lancet*, vol. 379, no. 9811, pp. 165–180, 2012.

2. N. R. Hill et al., "Global prevalence of chronic kidney disease – a systematic review and meta-analysis," *PLoS ONE*, vol. 11, no. 7, p. e0158765, 2016.
3. GBD Chronic Kidney Disease Collaboration, "Global, regional, and national burden of chronic kidney disease, 1990–2017," *The Lancet*, vol. 395, pp. 709–733, 2020.
4. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group, "Kdigo 2012 clinical practice guideline for the evaluation and management of chronic kidney disease," *Kidney International Supplements*, vol. 3, no. 1, pp. 1–150, 2013.
5. A. Kusiak, "Data mining: medical and engineering case studies," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2005.
6. F. Hussain, M. Nauman, A. Alghuried, A. Alhudhaif, and N. Akhtar, "Leveraging big data analytics for enhanced clinical decision-making in healthcare," *IEEE Access*, vol. 11, pp. 127 817–127 836, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:265168249>
7. N. Tomasev et al., "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, pp. 116–119, 2019.
8. J. L. Koyner et al., "A multicenter study of machine learning for prediction of acute kidney injury," *Kidney International Reports*, vol. 3, pp. 1313–1321, 2018.
9. M. Nauman, A. S. Almadhor, M. Albekairi, A. R. Ansari, M. A. B. Fayyaz, and R. Nawaz, "The role of big data analytics in revolutionizing diabetes management and healthcare decision-making," *IEEE Access*, vol. 13, pp. 10 767–10 785, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:275379571>
10. H. A. Al-Jamimi, "Synergistic feature engineering and ensemble learning for early chronic disease prediction," *IEEE Access*, vol. 12, pp. 62 215–62 233, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:269491402>
11. S. K. Ghosh, N. Widatalla, and A. H. Khandoker, "Machine learning framework for early detection of chronic kidney disease stages using optimized estimated glomerular filtration rate," *IEEE Access*, vol. 13, pp. 78 057–78 072, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:278221733>
12. A. Ahmed, G. Ahmed, E. Qureshi, and S. Ahmed, "Diagnosis of chronic kidney disease using machine learning algorithm," *European Journal of Theoretical and Applied Sciences*, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:267375804>
13. M. S. Al Huda et al., "Nefroai: An explainable and real-time framework for predicting chronic kidney disease using diverse machine learning models and different feature selection techniques," *IEEE Access*, vol. 14, pp. 10 939–10 976, 2026.
14. Y. Liu, P. H. C. Chen, J. Krause, and L. Peng, "External validation of machine learning models in medicine," *BMJ*, vol. 376, 2022.
15. A. Subbaswamy and S. Saria, "Preventing failures due to dataset shift," *Nature Medicine*, vol. 26, pp. 464–474, 2020.
16. L. Wynants et al., "Prediction models for diagnosis and prognosis: systematic review and critical appraisal," *BMJ*, vol. 369, 2020.
17. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, 2015.
18. B. Van Calster et al., "Calibration: the achilles heel of predictive analytics," *BMC Medicine*, vol. 17, 2019.
19. A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," 2005.
20. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002
21. M. T. Ribeiro, S. Singh, and C. Guestrin, "“why should i trust you?” explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
22. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
23. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.
24. M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.
25. A. Khamparia, P. Singh et al., "Internet of health things-driven deep learning system for chronic kidney disease prediction," *Neural Computing and Applications*, 2020.
26. N. G. Ramadhan, . Adiwijaya, W. Maharani, and A. A. Gozali, "Chronic diseases prediction using machine learning with data preprocessing handling: A critical review," *IEEE Access*, vol. 12, pp. 80 698–80 730, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID: 270132400>
27. D. Pawus', T. Poraz'ko, and S. Paszkiel, "Automation and decision support in the area of nephrology using numerical algorithms, artificial intelligence, and expert approach: Review of the current state of knowledge," *IEEE Access*, vol. 12, pp. 86 043–86 066, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:270450208>
28. N. D. Mul, V. M. de Jong, O. Cremer, H. de Grooth, and S. E. Bouhaddani, "Practical guidance for validating the predictive performance in the presence of missing data a guide for the clinical researcher," *Journal of clinical epidemiology*, p. 112159, 2026. [Online]. Available: <https://api.semanticscholar.org/CorpusID:284975419>.