Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Cost-Sensitive and Explainable Evaluation Framework for Chronic Kidney Disease Prediction Models

Kayla DaCosta

kdacosta2024@fau.edu

Florida Atlantic University https://orcid.org/0009-0004-5998-3117

Research Article

Keywords: Chronic Kidney Disease, Machine Learning Models, Ensemble Models, Black Box Models, Threshold Optimization, Cost -Sensitive Analysis, Interpretability, SHAP, Feature Importance, Clinical Decision Support

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10244248/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

A Cost-Sensitive and Explainable Evaluation Framework for Chronic Kidney Disease Prediction Models (2026)

Kayla DaCosta

Employment Details:

Assistant Director of Data Analytics at Florida Atlantic University

Contact Details

Phone: [redacted]

Email: dacostakayla@yahoo.com

ORCID - ID: 0009-0004-5998-3117

Abstract - **This paper evaluates Logistic Regression, Random Forest and Extreme Gradient Boosting models for CKD prediction using a framework that incorporates threshold optimization, cost-sensitive analysis and interpretability techniques to assess clinical deployment suitability. The cost-sensitive analysis framework assigns greater penalties to false negative predictions than to false positive predictions, to stimulate the clinical consequences of missed diagnosis. The models eliminated or reduced false negative predictions at their optimal thresholds, hence lowering the overall cost. Furthermore, the interpretability analysis consistently identified predictors including serum creatinine, glomerular filtration rate (GFR) and protein in urine across all models indicating consistency in feature importance across traditional and ensemble models contributing toward improving confidence in black box machine learning models. The research findings demonstrate that the proof of concept extending model evaluation beyond predictive performance to include threshold optimization, cost-sensitive analysis and utilizing explainability techniques provides a more clinically relevant framework for assessing the suitability of machine learning models as clinical decision support systems.**

Index Terms - Chronic Kidney Disease, Machine Learning Models, Ensemble Models, Black Box Models, Threshold Optimization, Cost -Sensitive Analysis, Interpretability, SHAP, Feature Importance, Clinical Decision Support

I. Introduction

Chronic Kidney disease affects around 35.5 million adults in the United States and is a leading cause of death in the USA [4]. Worldwide, CKD affects 10-16% of adults and can lead to end-stage renal disease [9]. One of the primary functions of the kidney is the filtering of waste products from the blood and the regulation of fluid balance through a series of physiological processes and partly controlled by anti-diuretic hormone. The kidney also plays a critical role in controlling blood pressure and stimulating red blood cell production.

CKD occurs when the kidney's ability to filter the blood deteriorates resulting in excess fluid and waste remaining in the blood [4]. There are five (5) stages of kidney disease which are determined by the glomerular filtration rate (eGFR) measured in clinical blood labs. Stages 1 and 2 are characterized by an eGFR of 60 or higher with mild damage but an operational kidney while stages 3 and 4 present an eGFR between 15-59 and a moderately damaged kidney with deteriorating functional capabilities [1]. These stages may present little to no symptoms where symptoms gradually appear as function declines. It is also noted that a temporary decline in eGFR is possible often due to shock or some medications and returns to normal with fluids. At stage 5, eGFR is less than 15 and the kidney may have failed requiring the patient to be on dialysis and/or receive a kidney transplant [1]. Patients may remain at lower stages without progressing through early diagnosis, intervention and treatment strategies.

Albuminuria, assessed through urine samples, may also indicate kidney disease, where lower levels indicate normal kidney functions and elevated levels indicate risk for kidney failure [1]. Individuals with preexisting conditions such as diabetes and hypertension are at risk for developing CKD while African American, Hispanic, native or Asian American populations are more likely to develop CKD [1]. However, this does not limit the risk factors for CKD as other conditions such as blood disorders or lifestyle factors can result in declining kidney functions over time.

Measures such as the reduction in the dietary intake of sugar and salt and maintaining adequate hydration, while engaging in regular physical activities are precautionary measures to reduce the risk of developing CKD or reducing the risk of disease progression. Treatment at early stages of CKD includes medications to lower blood pressure and cholesterol levels and control blood glucose levels. Consequently, the use of ML models for CKD prediction may support earlier identification and intervention to improve disease management and potentially reduce the risk of progression.

Supervised ML models serve as a mechanism to predict or aid in the diagnosis of diseases including CKD. One of the greatest challenges, however, in the implementation in clinical practice, is the lack of interpretability as compared to traditional statistical models. Although numerous research projects have examined the predictive performance of several models, fewer have jointly evaluated interpretability and cost sensitive measures suitable to healthcare diagnosis. The challenge is reflected in the difficulty of identifying trustworthy models that can support clinicians, expand the medical treatment process and promote more patient-centered care.

This research will compare three supervised ML models namely Logistic Regression (LR), Random Forest (RF) and XG Boost (XGB). Unlike many

previous studies that focus primarily on predictive accuracy, this research jointly evaluates and compares models based on interpretability, feature contribution and cost sensitive threshold optimization to assess the practical suitability of the models for deployment to predict CKD.

The objective of the research is to identify which model presents a balance of predictive performance, feature importance, interpretability and healthcare cost considerations. Through the incorporation of a cost sensitive evaluation, feature importance analysis and explainable techniques, the study aims to provide clinicians with a more practical perspective on predictions and contributes to the discussion of black box model deployment as part of the medical decision framework.

Logistic Regression models are considered traditional statistic models and are selected due to their technique of matching data from the cause variables to predict the probability of an event [20]. This makes it suitable for healthcare predictions since normal assessments are conducted using variables, the model must operate similarly to be suitable for such critical operations. The ensemble models incorporate the use of multiple decision trees. One such, the random forest is considered a semi-interpretable model and is often used in medical predictions due to its high accuracy [13]. XG Boost is considered to perform well, is robust and well balanced, making it suitable for healthcare predictions [11].

II. Literature Review

CKD presents with little to no symptoms making early detection very challenging. Furthermore, without regular medical check-ups, the disease may progress to later stages without intervention making it difficult to treat leading to kidney failure. Interpretable models provide a logical understanding of the predictions. Logistic regression provides for interpretability through its explanation of each predictor variable [10]. It further matches data from the cause variables to predict the probability of an event [20]. The interpretability capability supports practitioner clinical knowledge and reduces the risk of generalization [7]. Given that CKD does not currently have a defined treatment or management plan but rather depends on the underlying issues or clinical values that require management, the interpretability aids in the treatment plans. The regression model utilizes a weighted approach where predictor variables are assigned a weighted coefficient [7]. These coefficients may be positive or negative providing further details on its correlation in the prediction. The interpretable model demonstrates which variables are more impactful and whether higher or lower values are more beneficial [8].

However, its simplistic approach in prediction underscores the relatively complex nature of data variables that must be assessed to arrive at an accurate prediction. One such issue is inputting highly correlated variables (multicollinearity) into the model that reduces accuracy by distributing the effect amongst the variables [19]. Furthermore, it restricts the number of variables and follows a more uniform approach in its prediction [19]. Within healthcare, this creates limitations in the prediction capabilities, especially when the logistic regression model fails to capture more complex nonlinear relationships that exist among variables. While several studies have employed the use of Logistic models for CKD prediction due to the underlying structure used for the prediction and largely due to its interpretability, the model performance reduces with increasing complexity compared to ML models. Given the diverse variable assessments in clinical diagnosis, this model is seen to be most effective for simpler cases but in the healthcare setting, more robust and flexible models are required. ML Algorithms can handle high dimensional data, large numbers of predictor variables and has the capability to handle intricate connections [2].

[12] alludes to the tradeoff between performance and interpretability. Decision tree-based models such as Random Forest sit between the traditional statistical models, with somewhat explainable predictions and more complex ensemble models such as XG Boost with elevated accuracy and increased capability of handling multiple complex variables. Although these models provide a greater degree of interpretation, their performance may be lower in comparison to complex boosting models [12].

Ensemble Tree Classifier and Boosting algorithms expand the prediction capabilities proving to be more adaptable and flexible in disease prediction. Ensemble trees are stable and robust with varying dataset sizes and assigns weights to decision trees which are continued to the final model [15]. Furthermore, the models reduce generalization errors and boost accuracy by integrating multiple ML algorithms [14]. This addresses the complexity of clinical data and reduces the need to restrict variables. Ensemble models contrast to traditional statistical models in their advanced capabilities to process multidimensional information and identify nonlinear relationships [11]. Ensemble boosting techniques such as XGBoost have proven to have the highest accuracy in predicting CKD across several studies.

This is useful in expanding models to incorporate clinical, demographic, comorbidities and other types of data that are deemed related to disease prediction. This enhances professionals' capacity to devise holistic treatment plans, early intervention plans and more unique patient tailored management plans. Multiple studies have demonstrated improved predictive performance capability when using ensemble models, particularly XGBoost in CKD prediction. The studies also emphasize the increased capability of XGBoost in handling more variables than statistical models. As a result, this boosting technique will be used as a comparator in this research.

Nevertheless, the ML approach to prediction presents a significant challenge especially in healthcare due to its lack of interpretability or its "black box" nature. These models are considered black box models and face heavy scrutiny and rejection as clinicians are unable to identify how the model arrived at its predictions. A primary reason for this black box issue is the integration of multiple models as mentioned earlier which creates complexity. This reduces the capability of identifying how an input variable impacts the final model, the way traditional models can. According to [12], the improved predictive accuracy can often be attributed to the increased model complexity, which consequently makes understanding the rationale behind the decisions more difficult. Explainable AI techniques are utilized to address the issue, used alongside the ensemble models, insights into feature contribution can be generated [6]. This is not to be considered as the ultimate solution but rather provide clarity to clinicians. SHAP, an explainable AI technique, is commonly used as it offers both local and global explanations. Across the studies I have reviewed, explainable AI techniques are incorporated into the research framework to provide some level of transparency of the predictions of the black box models. SHAP provides a level of interpretation of black box models by quantifying feature contributions, allowing both local explanations for sample predictions and global explanations across all predictions [17]. In this research SHAP will be used to identify feature contribution, how impactful it is and in which direction to provide insight on the model(s) generated by the ensemble models used. This approach is suitable to identify real world usage capabilities of the derived models.

However, despite the adjustments and supports provided to traditional or ML models, the misclassifications still stand as a significant and understandable barrier in their implementation in healthcare. False positives and false negatives are results of even the most trained models. According to [15], a slightly declined performance in XGBoost presented a 100% in sensitivity and 98.1% in specificity. Sensitivity accounts for the proportion of accurately identified positive cases and specificity accounts for the proportion of accurately identified negative cases. This presents a tradeoff, which reflects the need for evaluation beyond predictive accuracy into cost-sensitive analysis by considering the practical consequences of deployment. Any misclassification can cause severe outcomes within the clinical setting, however, a false negative bears greater risk to patient outcome.

Many research models are reported at the default classification threshold of 0.5. This threshold reflects the balance between sensitivity, specificity, false positives and false negatives. However, in healthcare, since the same weight is not applied to all given the clinical consequences of each, the threshold was varied to determine the most appropriate outcome. This research places significant emphasis on threshold optimization in model selection to identify outcomes that are more practical and aligned with real-world clinical trade-offs.

Despite the advances from traditional statistical models to supervised ML approaches, few studies have jointly evaluated model performance, interpretability, feature importance and cost sensitive trade-offs in CKD prediction. The primary focus remains on maximizing predictive accuracy, which is a critical point, however, this alone is insufficient for deployment in healthcare environments where the consequences of false negatives and false positives are both drastic but carry different weights. This gap reduces the practical adoption and implementation of robust ML models within clinical decision-making and emphasizes the need for a framework that creates a deployable balance.

III. Methodology

The study utilized a publicly available synthetic Chronic Kidney Disease dataset generated for educational and research purposes. The dataset was obtained from the Kaggle repository "Chronic Kidney Disease Dataset", [18] and imported into Jupyter notebook where analysis was performed using python. The preprocessing pipeline entailed imputing missing numeric with the median and mode for categorical variables. The median was selected as it provides a more stable representation of clinical measurements than the mean, which may be influenced by potential outliers or skewed

distributions. Categorical variables were converted to dummy variables for machine learning usage. Although the dataset was relatively clean, the preprocessing procedures were implemented to ensure data consistency.

The patient ID, education level amongst other variables were removed due to their unsuitability for machine learning and lack of predictive contribution to the model mainly as they are not routinely collected during diagnostic examinations. The diagnosis variable was selected as the dependent variable, where 1 represents the presence of CKD and 0 represents the absence of CKD.

Exploratory analysis was conducted to examine variable and class distributions. The analysis indicates severe class imbalance, as shown in figure 1 below, where a larger proportion of the data falls in the positive class as compared to the negative class.

Since the Logistic Regression model is sensitive to multicollinearity among predictors, a Variable Inflation Factor (VIF) analysis was conducted. Removal of variables was not based solely on high VIF values, but also on clinical relevance and with the aim of preserving critical information. Variables that may seemingly have overlapping information were grouped and the least relevant removed. For example, the variable 'Cholesterol total' would capture information from the variables 'Cholesterol LDL' and 'Cholesterol HDL' and hence was removed. The Logistic Regression model was then scaled to standardize the remaining predictors. To maintain consistency across all models for the purpose of comparison, the same predictor variables were used despite the capabilities of the ensemble models to handle multicollinearity. Although the ensemble models are capable of handling multicollinearity amongst variables, the features selected for the logistic regression model were also used for the remaining models to compare the differences in predictive performance based on model architecture rather than features.

The data was split into training (60%), validation (20%) and testing (20%) using stratified sampling to preserve the class distribution across both datasets. This reduces the likelihood of introducing optimism bias by using the test set to decide on the optimal threshold. The training set was used for model fitting, the validation set was used for threshold optimization and cost-sensitive analysis and the test set served as the isolated data for a final evaluation. Three supervised machine learning models, Logistic Regression (LR), Random Forest (RF) and Extreme Gradient Boosting (XG Boost), were built and each accessed using statistical methods and graphical methods. Model performance was evaluated using accuracy, precision, recall, F1-score, Area Under the Receiver Operation Characteristic Curve (AUC), confusion matrices and a fivefold cross-validation AUC scores. This cross-validation assessment was performed to assess model stability and was used as a model comparator given the class imbalance.

Threshold variation was assessed to identify meaningful results applicable to the healthcare industry, which introduces a layer of model evaluation beyond the conventional metrics. This analysis considers the clinical trade-offs of false positives and false negatives. A cost assumption that false negatives usually result in greater consequences than false positives were included in the threshold analysis to stimulate clinical practices. The assumptions included assigned a value of $9000 to a false negative prediction and $5000 to false positive. Metric comparison was conducted at both the default and optimized thresholds for each model. A sensitivity analysis was also conducted, varying the price ratio.

Graphical analysis includes assessment of the ROC curve and precision-recall curve which supports the threshold analysis. Further analysis of the cost vs threshold graphs gives a visual representation of the model behavior across varying classification thresholds. The sensitivity – specificity trade-off graph demonstrates the real-world trade-off at the selected threshold compared to other thresholds including the default threshold.

Coefficient Analysis and permutation importance was conducted for the Logistic Regression model to identify key feature contributions and the direction and weight of each contribution. Interpretability analyses including permutation importance, feature importance and SHapley Additive exPlanations (SHAP) were applied to Random Forest and XGBoost models to evaluate the feature contributions of each model. The resulting features were assessed to determine their alignment with medical research and determine the extent of applicability of these models in deployment.

Portions of this paper were edited for grammatical consistency using AI. However, the manuscript has been reviewed to ensure consistency with the original idea and the author assumes full responsibility for the content.

IV. Results and Discussion

The dataset consisted of 1659 patients, of which 1524 were diagnosed with CKD and 135 were non-CKD

cases. Figure 1 illustrates the class imbalance and demonstrates the majority falling within the positive class. This insight will play a critical role in metric selection for comparison and how to evaluate performance metrics across all models.

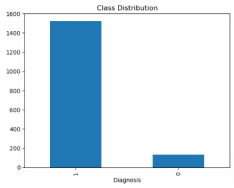


Figure 1. Class Distribution

Performance Evaluation for Each Model

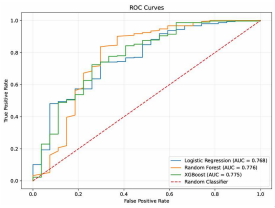


Figure 2.

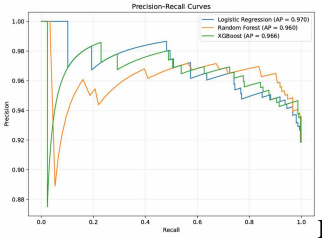


Figure 3.

Cost Functions and Sensitivity-Specificity Tradeoff Plots

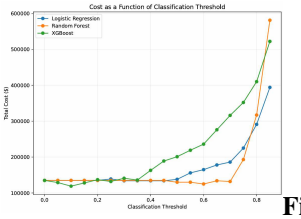


Figure 4.

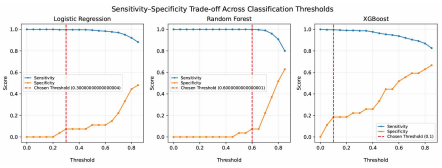


Figure 5

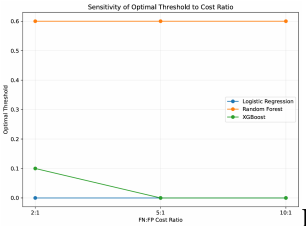


Figure 6

Logistic Regression		
Feature	Coefficient	Odds Ratio
SerumCreatinine	0.86282969	2.369857175
MuscleCramps	0.524353267	1.689365926
Itching	0.499463356	1.647836732
SystolicBP	0.446444988	1.562746715
ProteinInUrine	0.369776342	1.476650492
NauseaVomiting	-0.230355829	0.794250935
GFR	-0.809539516	0.445062963

Random Forest	
Feature	Importance
SerumCreatinine	0.098335416
ProteinInUrine	0.068591068
GFR	0.064050831
SerumElectrolytesSodium	0.041577976
Itching	0.039905337
MuscleCramps	0.039843986
SystolicBP	0.035088117

XGBoost	
Feature	Importance
FamilyHistoryKidneyDisease	0.08239482
SerumCreatinine	0.06939356
ProteinInUrine	0.05221297
GFR	0.047772914
UrinaryTractInfections	0.041798994
PreviousAcuteKidneyInjury	0.041048206
SleepQuality	0.03775814

Figure 7. Feature Importance for Each Model

Logistic Regression		
Feature	Importance Mean	Importance Std
SerumCreatinine	0.085808541	0.034780294
GFR	0.054867436	0.023060492
MuscleCramps	0.030394657	0.018288461
SystolicBP	0.020170006	0.015154894
ProteinInUrine	0.013041894	0.011826202
BMI	0.01037037	0.009656352

Random Forest		
Feature	Importance Mean	Importance Std
SerumCreatinine	0.168516485	0.042961742
GFR	0.058273629	0.016328533
ProteinInUrine	0.037128112	0.021623514
MuscleCramps	0.019156041	0.012047433
NauseaVomiting	0.010076908	0.005362824
HemoglobinLevels	0.007563246	0.003837955

XGBoost		
Feature	Importance Mean	Importance Std
SerumCreatinine	0.09907569	0.029323931
GFR	0.06729407	0.022902626
ProteinInUrine	0.019319976	0.013194282
SystolicBP	0.013444847	0.014357507
MuscleCramps	0.012260676	0.012440305
NauseaVomiting	0.009686298	0.005835664

Figure 8. Permutation Importance for Each Model

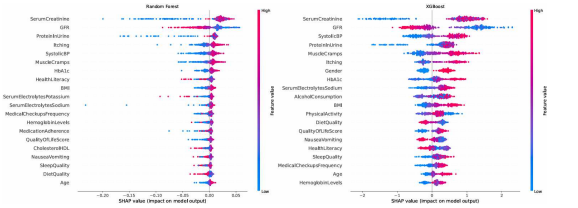


Figure 9. SHAP Summary Plots

The performance of three models, Logistic Regression (LF), Random Forest (RF) and XG Boost (XGB) were evaluated using standard classification metrics, confusion matrices, cost-sensitive analysis and interpretability techniques. The applicability of the models was found to vary mainly based on threshold behavior, clinical decision costs, interpretability and feature importance. A majority-class classifier was included as a no-information baseline due to the severe class imbalance in the dataset.

Table 1 presents the metrics at the default threshold of 0.5 and at the respective optimal thresholds for each model. At the threshold of 0.5, all models achieved between 91% to 92% accuracy with approximately 1% increase at the optimal threshold indicating strong model performance. However, given the strong class imbalance with predominantly positive CKD cases, accuracy bears less weighting in model performance classification as it is likely inflated and may overstate model effectiveness in diagnosis. According to [3], class imbalance results in noninformative accuracy. Consequently, metrics including recall, precision and F1-score will serve as reference points for model performance evaluation. At both the default and optimal thresholds, the recall value remained substantially high across all models demonstrating their ability to correctly identify CKD cases. Furthermore, precision lies between 92% - 94% across all models, indicating relatively few

unnecessary positive classifications. The F1 score ranges around 96%, signaling that all models maintain a strong balance between precision and recall. These metrics are more suitable for evaluating performance given the imbalance. However, the majority classifier achieved similar values in these metrics, demonstrating that these metrics are insufficient in evaluating performance under severe class imbalance as its capability of distinguishing non-CKD cases is weak. This is further demonstrated in the balanced accuracy where random forest achieved 50% while Logistic regression and XGBoost achieved 60% and 66% respectively. The higher balanced accuracy of XGBoost indicates that it provided a better balance between identifying positive and negative CKD cases suggesting stronger performance under the imbalanced classification setting. Figure 3 demonstrates that the average precision for all models was approximately 0.97 indicating high precision across the recall range as the graph demonstrates. The precision -recall curve is useful for model assessment in the presence of class imbalance.

Given that the objective of the models is to determine applicability in supporting early disease detection, a cost-sensitive framework was developed to extend analysis beyond standard metrics. A higher cost was assigned to false negatives ($9000) compared to false positives ($5000) reflecting the greater risk of missing a positive diagnosis. The reason for this assumption is alluded to the common clinical decision-making process that a false negative may delay treatment, contributing to disease progression and complications whereas a false positive is likely followed with additional screening or diagnostic testing. The assigned costs were not intended to represent actual healthcare costs nor any specified quantitative relationship between the two types of misclassifications but rather as a demonstration of the unequal consequences of misclassifications and to examine how this affects threshold selection and to compare model behaviour at selected thresholds.

The cost analysis was used to identify the optimal threshold that minimized the overall cost under the proposed framework. The suitable threshold varied across all models. Various cost-sensitive approaches seek to assign different penalties to false positive and false negative outcomes to identify acceptable trade-offs and minimize cost [16]. Under the primary cost structure, in table 1, both the logistic regression and random forest models achieved zero false negative cases at their respective optimal thresholds but had higher minimum costs than XGBoost. This illustrates the trade-off between reducing false negatives and increasing false positive classifications. Furthermore, these results demonstrate that minimizing false negative cases did not necessarily achieve minimum costs which translates into healthcare decisions where eliminating false negative cases does not imply clinical usefulness.

Nevertheless, applying the optimal thresholds demonstrates changes in classification behaviour to support model assessment suggesting that model deployment should not solely rely on default thresholds for clinical practicability. However, due to the severe imbalance and the cost assumptions that do not necessarily reflect realistic operations, alternative cost structures were assigned using the ratios 2:1, 5:1 and 10:1 to evaluate threshold changes demonstrated in figure 6. Random forest selected the same threshold across all tested cost ratios and the original cost assumption indicating relative stability across the evaluated assumptions. However, this perceived stability could be an effect of the imbalance. Logistic regression demonstrates that it is most cost effective to classify all patients as CKD rather than to miss any by lowering the threshold to 0.0 across all tested ratios. The model behaviour across the evaluated ratios is equivalent to the majority class-baseline. For XGBoost, at the lowest ratio of 2:1 it maintains the threshold similar to the original cost assumption which has a relatively similar cost ratio but falls to 0 for the remaining ratios. This finding demonstrates that increasing the false negative penalty does not necessarily improve the clinical model based on the tested framework. Across all models, imbalance impacts clinical feasibility of the resulting classification behaviour but demonstrates that assigning significantly higher false negative penalties does not improve the clinical usefulness of a model based on the tested structures.

In comparing model discrimination, all three models demonstrated similar discriminating capabilities with AUC values of approximately 77-78%. This indicates moderate class separation capabilities across thresholds. However, given the wide 95% confidence intervals, there is considerable uncertainty in the estimated performance. Given the severe class imbalance, the cross-validation AUC provides an estimate of performance to support AUC reporting. XGBoost maintained a high performance in CV-AUC in comparison to the other models, suggesting comparatively stronger discrimination across the cross-validation folds despite the imbalance. This consistency is particularly important in disease prediction given the variability amongst clinical profiles and the need for robust model performance across this diversity.

The sensitivity - specificity trade-off plots at the optimal threshold in figure 5, showed a similar pattern across all three models, with high sensitivity of approximately 1 accompanied by low specificity of approximately 0.1. However, the threshold optimization process changed classification performance across all models suggesting that desirable classification performance is not always achieved at the default threshold. Therefore, model thresholds should be selected based on the alignment with deployment criteria.

To improve interpretability, permutation importance based on AUC was applied. This analysis identified the features that remained influential to the model despite feature reshuffling. Serum Creatinine was consistently selected by all models indicating strong contribution to prediction and consistency across all model behaviours. This indicates that ensemble and traditional models identify similarly important features supporting interpretability of black box models. Additional features consistently identified as influential to AUC are GFR, protein in urine, systolic BP and muscle cramps. Hypertension is a defined risk factor of CKD [10]. This medical evidence provides clinical support for the model findings of systolic BP impacting AUC. Given that some variables were removed after VIF identified highly correlated variables that logistic regression models do not handle well, the same remaining feature set was used for the remaining models to support these feature comparisons.

A similar overlap in variable selection is observed across the models using the feature importance analysis. This indicates that these variables strongly contribute to the prediction of CKD within each model. The alignment of these variables amongst traditional and ensemble models serves as useful evidence for model consistency and promotes feature importance as a tool for interpretability. Feature selection may vary slightly such as XGBoost selecting family history as a strong positive influencer or Logistic Regression identifying itching as one of the strong positive influencers highlighting the difference in model priorities and use of variables in prediction.

SHAP analysis was conducted for the ensemble models to identify the directional impact of the features. This analysis suggests that an increase in the levels of serum creatinine, protein in urine, systolic blood pressure alongside a decrease in GFR contributed to higher predicted CKD risk. The degree of impact, however, was stronger, particularly for Serum creatinine and GFR within the XGBoost model. Both variables emerged as strong clinical signals for the ensemble models. GFR is considered as a measure for abnormality in kidney function while serum creatinine has been tested to estimate GFR [5]. However, some variables demonstrated as strongly influential such as itching are less common and may suggest weaker clinical interpretability or suggest potential noise. However, this may prove insightful if models detect nontypical features when trained on large clinical datasets which is not established by this study.

These findings aim to contribute toward improving interpretability and trust in black box models by demonstrating that the most influential predictors in black box models correlate with those of traditional models. This suggests potential suitability, upon thorough testing and validation on clinical data, for deployment and use alongside other diagnosis screening.

Overall, these findings support the use of machine learning models as decision support tools. The employment of explainable AI techniques to reveal clinically influential predictors aim to support confidence in black box models as they align with traditional model findings and clinical knowledge. The framework showed that utilizing a cost-sensitive threshold optimization technique can improve the selection of an operable threshold that acceptably balances the misclassification trade-offs. However, cost alone cannot establish clinical usefulness under severe class imbalance where poor specificity exists. The framework demonstrates an approach or proof of concept using synthetic data that can be applied to clinical data for validation.

V. Limitations

This study presents several limitations. The dataset exhibited strong class imbalance which may impact model performance and may have influenced performance metrics despite the use of alternative suitable evaluation metrics. Additionally, the cost-sensitive approach employed the use of cost assumptions to price penalties which may not fully represent real world costs. While this enables the incorporation of a more clinically aligned thought process, the assigned values may not fully reflect real-world clinical consequences.

This dataset was synthetically generated for educational purposes, making the findings of this study limited in clinical generalizability. While the dataset reflects clinically plausible relationships among variables, it does not represent actual clinical records. Consequently, the findings should be interpreted as a methodological evaluation of the proposed framework. Furthermore, due to the limited

availability of clinical datasets for CKD with comparable feature structures, external validation was not conducted as both datasets would require alternative model comparison approaches which would limit the validity of the subsequent evaluations across models.

Conclusion

In conclusion, traditional statistical performance metrics provide a necessary foundation for model comparison. However, incorporating threshold optimization, evaluating clinical trade-offs and applying interpretability techniques provides a more comprehensive framework for identifying deployable models. The research findings demonstrate that model evaluation incorporating threshold optimization, cost-sensitive analysis and explainable techniques in addition to traditional statistical metrics, provides a more relevant framework to assess deployment suitability of machine learning models to support clinical decisions.

Further research should focus on external validation on clinical CKD dataset and the exploration of disease-specific cost functions that better reflect real-world healthcare outcomes. A similar approach may be applied to assess the applicability of this framework as a decision support tool for other critical diseases. Furthermore, future studies may explore traditional clinical risk factors and less conventional factors to expand screening capabilities to support early disease identification and timely intervention.

References

[1] American Kidney Fund. (2025, July 21). *Stages of Kidney Disease (CKD).* https://www.kidneyfund.org/all-about-kidneys/stages-kidney-disease

[2] Balikci Cicek I, Kucukakcali Z, (2023). Explainable Artificial Intelligence Method SHAP's Prediction of Risk Factors Associated with Chronic Kidney Disease Combined with Black Box Methods. *J Comm Med and Pub Health Rep* 4(10): https://doi.org/10.38207/JCMPHR/2023/NOV041004138

[3] Blagus, R., & Lusa, L. (2010). Class prediction for high-dimensional class-imbalanced data. BMC bioinformatics, 11, 523. https://doi.org/10.1186/1471-2105-11-523

[4] Centers for Disease Control and Prevention (CDC). (2024, May 15). *Chronic Kidney Disease Basics*. https://www.cdc.gov/kidney-disease/about/index.html

[5] Chen, T. K., Knicely, D. H., & Grams, M. E. (2019). Chronic Kidney Disease Diagnosis and Management: A Review. *JAMA,* 322(13), 1294–1304. https://doi.org/10.1001/jama.2019.14745

[6] Ekundayo, F. (2024). Machine learning for chronic kidney disease progression modelling: Leveraging data science to optimize patient management. *World Journal of Advanced Research and Reviews*, 24(03), 453–475. https://doi.org/10.30574/wjarr.2024.24.3.3730

[7] Gao, X., Alam, S., Shi, P., Dexter, F., & Kong, N. (2023, Jun 5). Interpretable machine learning models for hospital readmission prediction: a two-step extracted regression tree approach. *BMC Med Inform Decis Mak* 23, 104. https://doi.org/10.1186/s12911-023-02193-5

[8] Harder, H.D. (2020, Jan. 4). Interpretable Machine Learning Models. *Towards Data Science*. https://towardsdatascience.com/interpretable-machine-learning-models-aef0c7be3fd9/

[9] Kazancioğlu R. (2013). Risk factors for chronic kidney disease: an update. *Kidney international supplements*, *3*(4), 368–371. https://doi.org/10.1038/kisup.2013.79

[10] Lavuri, N. R., Borugadda, P., Krishna, P. R., & Reddy, G. K. (2025). Developing a Logistic Regression Model for Predicting Chronic Kidney Disease. *IEEE Access.* https://doi.org/10.1109/ICOEI65986.2025.11012936

[11] Li, J., Du, X., Zhang, R., Li, X., Xu, J., Song, X., Zhao, Y., Liu, L., Xu, G., & Fan, Y. (2025). Machine learning models for predicting short-term progression in patients with stage 4 chronic kidney disease: a multi-center validation study. *Scientific reports*, *15*(1), 39285. https://doi.org/10.1038/s41598-025-23037-4

[12] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy (Basel, Switzerland)*, *23*(1), 18. https://doi.org/10.3390/e23010018

[13] Liu, P., Liu, Y., Liu, H., Xiong, L., Mei, C., & Yuan, L. (2024). A Random Forest Algorithm for Assessing Risk Factors Associated With Chronic Kidney Disease: Observational Study. *Asian/Pacific Island nursing journal*, 8, e48378. https://doi.org/10.2196/48378

[14] Mahajan, P., Uddin, S., Hajati, F., & Moni, M.A. (2023). Ensemble Learning for Disease Prediction: A Review. *Healthcare*, 11(12), 1808. https://doi.org/10.3390/ healthcare11121808

[15] Moreno- Sánchez, P.A. (2023). Data-Driven Early Diagnosis of Chronic Kidney Disease: Development and Evaluation of an Explainable AI Model. *IEEE Access, 11, 38359-38369,* https://doi.org/ 10.1109/ACCESS.2023.3264270

[16] Pes, B., & Lai, G. (2021). Cost-sensitive learning strategies for high-dimensional and imbalanced data: a comparative study. *PeerJ. Computer science*, *7*, e832. https://doi.org/10.7717/peerj-cs.832

[17] Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and translational science*, *17*(11), e70056. https://doi.org/10.1111/cts.70056

[18] Rabie El Kharoua. (2024). Chronic Kidney Disease Dataset [Dataset]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/8658224

[19] Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in clinical research*, *8*(3), 148–151. https://doi.org/10.4103/picr.PICR_87_17

[20] Singh, B., Arora, K., & Iyer, S. (2022). Chronic Kidney Disease Detection Using Machine Learning Regression Models. *The Electrochemical Society, 107*(1), 2191. https://doi.org/10.1149/10701.2191ecst

Tables

At standard Threshold - 0.5				
Metrics	Logistic Regression	Random Forest	XG Boost	Majority - class
Accuracy	0.9187	0.9187	0.9036	0.9187
Precision	0.9344	0.9187	0.9446	0.9187
Recall	0.9803	1	0.9508	1
F1 Score	0.9568	0.9576	0.9477	0.9576
Specificity	0.2222	0	0.3704	0
Balanced Accuracy	0.6013	0.5	0.6606	0.5
AUC	0.7679	0.7757	0.7750	
AUC 95% CI	0.669,0.860	0.648, 0.886	0.671, 0.871	
CV AUC	0.765	0.7943	0.8201	
At Optimal Threshold				
Metrics	Logistic Regression	Random Forest	XG Boost	
Threshold	0.3	0.6	0.1	
Accuracy	0.9247	0.9247	0.9277	
Precision	0.9242	0.9242	0.9297	
Recall	1	1	0.9967	
F1 Score	0.9606	0.9606	0.962	
Specificity	0.0741	0.0741	0.1481	
Balanced Accuracy	0.537	0.537	0.5724	

Table 1. Performance Metric Evaluation at the default threshold of 0.5 and at their respective optimal threshold

Model	Threshold	TP	TN	FP	FN	Cost
Logistic Regression	0.50	299	6	21	6	$159,000
	0.30	305	2	25	0	$125,000
Random Forest	0.50	305	0	27	0	$135,000
	0.60	305	2	25	0	$125,000
XGBoost	0.50	290	10	17	15	$220,000
	0.10	304	4	23	1	$124,000
Majority Class	0.50	305	0	27	0	

Table 2. Confusion Matrices at the default and Optimal Thresholds