American Journal of Sciences and Engineering Research

E-ISSN– 2348–703X, Volume 9, Issue 5, 2026

Interpretable Machine Learning for Chronic Kidney Disease Classification: A Leakage-Aware Comparison of Random Forest and XGBoost with SHAP

Zeynep Kucukakcali[1,*], Ipek Balikci Cicek [2]

[1,2] *Inonu University Faculty of Medicine, Department of Biostatistics and Medical Informatics, 44280, Malatya, Turkey*

** Corresponding author: Zeynep Kucukakcali, e-mail: zeynp.tunc@inonu.edu.tr*

Abstract:

Background: *Chronic kidney disease (CKD) is a major global health burden that is frequently asymptomatic in its early stages, making automated screening from routine clinical and laboratory data attractive. However, when the CKD label is derived from the glomerular filtration rate (GFR), including GFR as a predictor causes target leakage and artificially perfect performance. We aimed to develop and compare leakage-aware machine-learning models for CKD detection.*

Methods: *A publicly available kidney-function dataset of 5,000 patient records (26.3% CKD) was used. GFR was excluded from the feature set to avoid target leakage, and nine clinical and laboratory variables were retained. Random forest and XGBoost classifiers were trained using an 80:20 stratified split. Class imbalance was addressed with SMOTE applied only within the training and cross-validation folds. Models were assessed by five-fold stratified cross-validation and on a held-out test set with bootstrap 95% confidence intervals (1,000 replicates); discrimination, calibration, and interpretation via SHapley Additive exPlanations (SHAP) were examined.*

Results: *Both models discriminated CKD almost perfectly. XGBoost was marginally superior, with a test accuracy of 0.997 (95% CI 0.993–1.000), ROC-AUC of 0.9997, sensitivity of 1.000, specificity of 0.996, and the lower Brier score (0.003), versus an accuracy of 0.995 for random forest. Both models were well calibrated, and SHAP identified serum creatinine, blood urea nitrogen, and urinary protein as the most influential predictors—consistent with established markers of kidney function.*

Conclusions: *A leakage-aware, calibrated, and interpretable modeling workflow yielded highly accurate CKD classification, with XGBoost slightly outperforming random forest. External and prospective validation on real-world clinical data is required before clinical use.*

Keywords: *chronic kidney disease; machine learning; XGBoost; random forest; SHAP; class imbalance; calibration; data leakage.*

I. Introduction

Chronic kidney disease (CKD) is a progressive condition characterized by the gradual loss of kidney function and has become a major global public health problem. Recent estimates indicate that CKD affects roughly 9% of the world's population—on the order of 700 million people—and that its prevalence has continued to rise over the past three decades, driven largely by population growth, ageing, and the increasing burden of diabetes and hypertension (1). Untreated CKD can progress to end-stage renal disease requiring

 Received-31-08-2026 Accepted- 12-09-2026

dialysis or transplantation, and it substantially increases the risk of cardiovascular disease and premature death (1, 2).

A central difficulty in CKD management is that the disease is frequently asymptomatic in its early stages, so that many patients remain undiagnosed until kidney function is already markedly impaired (2). Diagnosis and staging rely primarily on the glomerular filtration rate (GFR) together with routine laboratory markers such as serum creatinine, blood urea nitrogen (BUN), and proteinuria (2). Because these variables are collected during ordinary clinical care, they provide a natural basis for automated screening tools that could flag high-risk individuals earlier and support timely intervention.

Machine learning has emerged as a promising approach for this task, and tree-based ensemble methods in particular have shown strong and consistent performance on tabular clinical data. Random forests aggregate many decorrelated decision trees to reduce variance and improve generalization (3), whereas extreme gradient boosting (XGBoost) builds trees sequentially with regularization to achieve high accuracy on structured data (4). Several studies have applied these algorithms to CKD prediction with encouraging results: Ghosh and Khandoker reported that XGBoost outperformed other classifiers with an area under the curve of about 0.97 (5); Halder et al. developed a multi-classifier CKD prediction pipeline deployed as a web application (6); and ensemble models combining random forest and boosting have been shown to predict renal function decline with good discrimination and external validity (7). Explainable-AI approaches have further been used to make such models more transparent to clinicians (8).

Despite these advances, developing a reliable and trustworthy CKD classifier involves several methodological challenges. Clinical datasets are often imbalanced, with far fewer diseased than healthy cases, which can bias models toward the majority class; the Synthetic Minority Over-sampling Technique (SMOTE) and its variants are widely used to mitigate this problem (9). In addition, the opaque nature of ensemble models limits their clinical acceptance, motivating the use of post-hoc interpretation methods such as SHapley Additive exPlanations (SHAP) to quantify how each feature contributes to a prediction (10). A further, less frequently addressed issue is the risk of target leakage: when the outcome label is itself derived from one of the input variables, including that variable as a predictor produces artificially perfect performance that does not generalize.

In this study, we developed and compared random forest and XGBoost classifiers for CKD detection using a publicly available kidney-function health dataset. Because the CKD label in this dataset is defined from GFR, we deliberately excluded GFR from the predictor set to avoid target leakage and trained the models on the remaining clinical and laboratory variables. Class imbalance was addressed with SMOTE applied only within the training and cross-validation folds, model performance was assessed using stratified cross-validation together with bootstrap confidence intervals on a held-out test set, and calibration was examined alongside discrimination. Finally, the best-performing model was interpreted using SHAP to identify the most influential predictors and to verify that the model's behavior is consistent with established clinical knowledge of kidney function.

II. Materials and Methods

Dataset

The study used a dataset of 5,000 patient records comprising ten predictor variables and a binary chronic kidney disease (CKD) outcome (CKD_Status). The predictors consisted of eight continuous laboratory and clinical variables—serum creatinine, blood urea nitrogen (BUN), urine output, age, urinary protein, and daily water intake—together with two binary comorbidity indicators (diabetes and hypertension) and one categorical variable representing medication class (ACE inhibitor, ARB, diuretic, or no medication). In the raw data, empty cells in the medication field did not denote missing values but rather patients receiving no pharmacological treatment; these entries were therefore recoded as a dedicated "No medication" category prior to modeling.

Outcome definition and exclusion of GFR

The CKD label was derived directly from the estimated glomerular filtration rate (GFR). Because GFR is the quantity used to define the outcome, retaining it as a predictor would constitute target leakage and produce circular, over-optimistic performance estimates. GFR was therefore excluded from the feature set, and all models were trained on the remaining nine predictors. The resulting cohort showed a moderate class imbalance, comprising 3,685 non-CKD cases (73.7%) and 1,315 CKD cases (26.3%).

Study design and data partitioning

The dataset was divided into training and test sets using an 80:20 stratified partition (random seed = 42), preserving the outcome prevalence in both subsets (n = 4,000 for training and n = 1,000 for testing). The test set was held out and left untouched until the final evaluation.

Preprocessing and class balancing

All preprocessing steps were fitted within a single pipeline to prevent information leakage from the test set. Continuous variables were imputed using the median, and the categorical medication variable was ordinal-encoded, with previously unseen categories mapped to a reserved code. To address the class imbalance, the Synthetic Minority Over-sampling Technique for mixed numeric–categorical data (SMOTENC) was applied exclusively within the training data and cross-validation folds and was never applied to the test set, ensuring that resampling could not inflate the reported test performance.

Models and hyperparameters

Two ensemble classifiers were compared. The Random Forest model used 300 trees, a maximum depth of 8, and a minimum of 2 samples per leaf. The XGBoost model used 250 trees, a maximum depth of 3, a learning rate of 0.05, subsample and column-subsample ratios of 0.9, and L2 regularization ($\lambda = 3$). A fixed random seed was used for both models to ensure reproducibility.

Cross-validation and evaluation

Model stability was assessed by five-fold stratified cross-validation on the training set, using accuracy as the scoring metric. Final performance was evaluated on the held-out test set using accuracy, area under the receiver operating characteristic curve (ROC-AUC), area under the precision–recall curve (PR-AUC), sensitivity, specificity, positive and negative predictive values (PPV and NPV), F1 score, the Matthews correlation coefficient (MCC), and the Brier score for calibration. Ninety-five percent confidence intervals (CIs) were obtained by non-parametric bootstrap resampling of the test predictions (1,000 replicates). Discrimination and calibration were further examined using ROC curves, precision–recall curves, and reliability (calibration) curves based on ten quantile-defined bins.

Model interpretation

The best-performing model by test accuracy was interpreted using SHapley Additive exPlanations (SHAP). A tree-based SHAP explainer was applied to a random sample of test observations, and mean absolute SHAP values were used to rank feature importance and to characterize the direction of each feature's effect on the predicted CKD risk.

Software

All analyses were performed in Python using the scikit-learn, imbalanced-learn, XGBoost, SHAP, and Matplotlib libraries.

III. Results

Overall performance. Both models were highly stable across the five training folds, with mean cross-validated accuracies of 0.9955 (SD = 0.0017) for XGBoost and 0.9930 (SD = 0.0034) for Random Forest. On the untouched test set, both classifiers separated CKD from non-CKD cases almost perfectly. Both identified every CKD case (sensitivity = 1.000) with a negative predictive value of 1.000, meaning no true CKD cases were missed. XGBoost slightly outperformed Random Forest in accuracy (0.997 versus 0.995), specificity (0.996 versus 0.993), and PPV (0.989 versus 0.981), and produced the higher F1 score (0.994 versus 0.991) and MCC (0.992 versus 0.987). Complete metrics with bootstrap 95% confidence intervals are presented in Table 1.

Table 1. Test-set performance of the Random Forest and XGBoost models, with 5-fold cross-validated accuracy and bootstrap 95% confidence intervals (1,000 replicates).

Metric	Random Forest	XGBoost
5-fold CV accuracy (mean ± SD)	0.9930 ± 0.0034	0.9955 ± 0.0017
Accuracy	0.995 (0.990–0.999)	0.997 (0.993–1.000)
ROC-AUC	0.9998 (0.999–1.000)	0.9997 (0.999–1.000)
PR-AUC	0.9993 (0.998–1.000)	0.9992 (0.997–1.000)
Sensitivity	1.000 (1.000–1.000)	1.000 (1.000–1.000)
Specificity	0.993 (0.987–0.999)	0.996 (0.991–1.000)
PPV	0.981 (0.964–0.996)	0.989 (0.974–1.000)
NPV	1.000 (1.000–1.000)	1.000 (1.000–1.000)
F1 score	0.991 (0.982–0.998)	0.994 (0.987–1.000)
MCC	0.987 (0.976–0.997)	0.992 (0.982–1.000)
Brier score	0.0053 (0.003–0.009)	0.0033 (0.001–0.007)

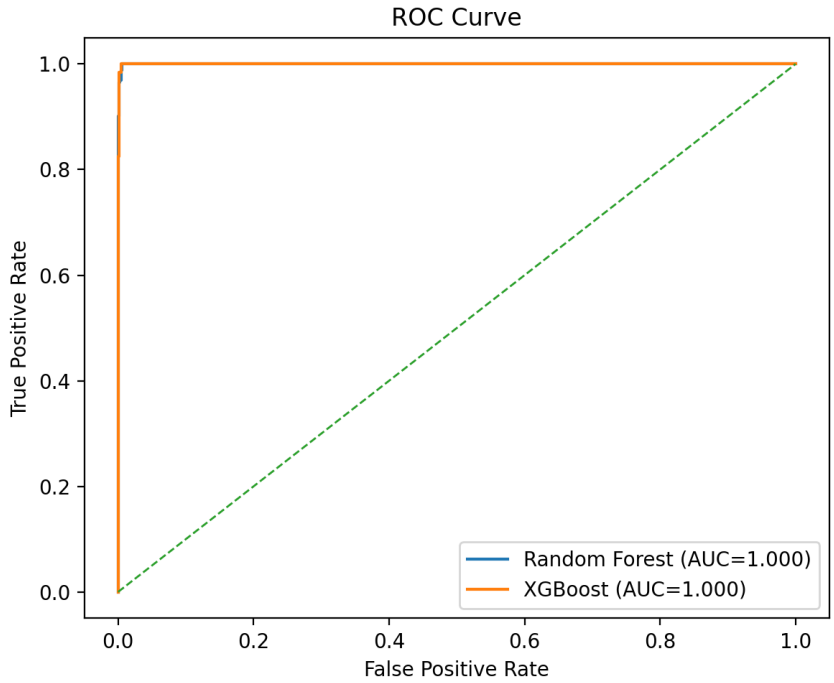

Figure 1. Receiver operating characteristic (ROC) curves for the Random Forest and XGBoost models on the test set.

The ROC curves for both models rose almost vertically toward the upper-left corner and then ran along the top of the plot, corresponding to an area under the curve of 1.000 (to three decimal places) for each model. This indicates near-perfect ranking of CKD versus non-CKD cases: across essentially all decision thresholds, the models achieved a very high true-positive rate while keeping the false-positive rate close to zero. Discrimination was therefore preserved despite the moderate class imbalance.

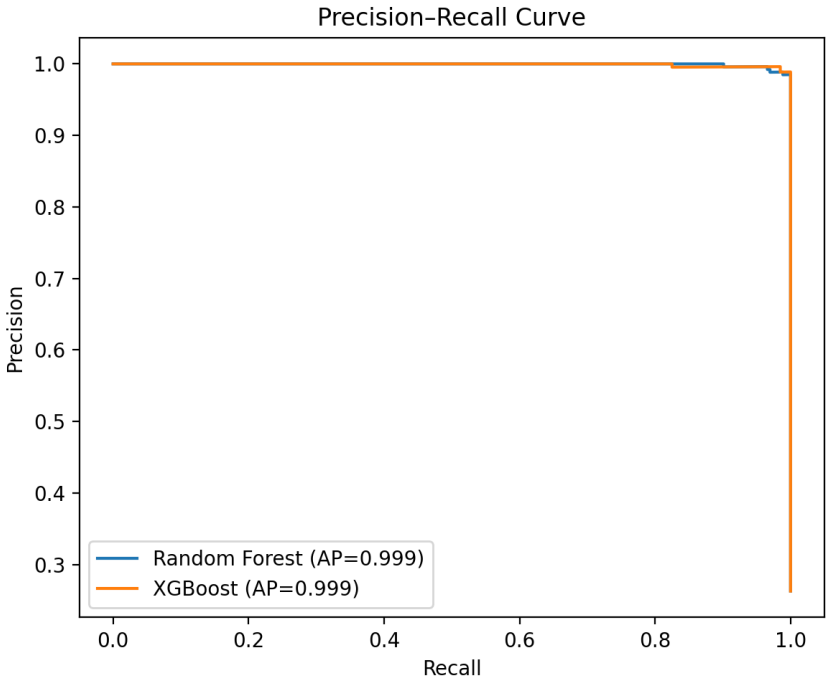

Figure 2. Precision–recall curves for the Random Forest and XGBoost models on the test set.

The precision–recall curves remained close to the upper boundary across the full recall range, with an average precision of 0.999 for both models. Because precision–recall analysis focuses on the minority (CKD) class, this result confirms that the models maintained high precision even as recall approached 1.0, and that the excellent discrimination was not an artifact of the majority class dominating the ROC analysis.

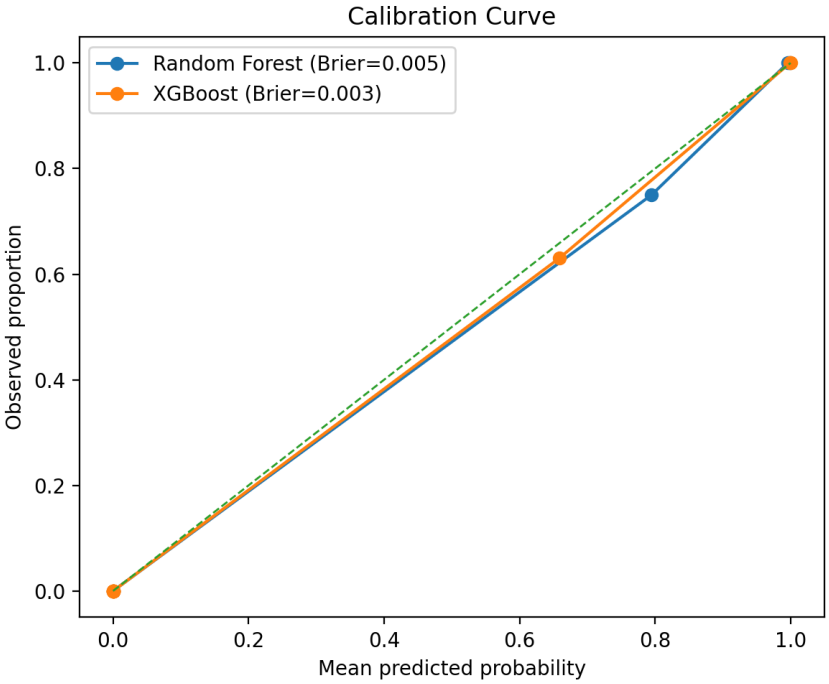

Figure 3. Calibration (reliability) curves for the Random Forest and XGBoost models, with Brier scores.

Both models were well calibrated, with reliability curves lying close to the diagonal of perfect calibration, so that predicted probabilities closely matched the observed event proportions. XGBoost showed marginally

better calibration, reflected in a lower Brier score (0.003 versus 0.005). This means the predicted risks can be interpreted as meaningful probabilities rather than as bare classification labels.

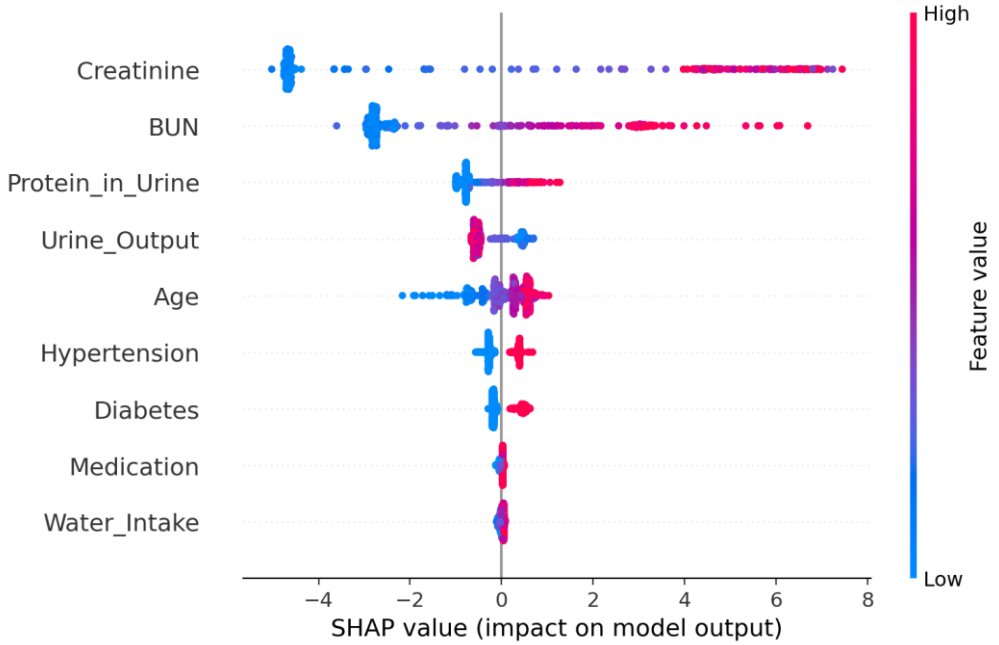

Figure 4. SHAP summary plot for the XGBoost model, showing the impact of each predictor on model output.

Because XGBoost achieved the highest test accuracy, it was selected for SHAP-based interpretation. Serum creatinine and BUN were by far the most influential predictors: higher values of both features (red points) shifted predictions strongly toward CKD, whereas lower values (blue) pushed predictions toward the non-CKD class. Urinary protein, urine output, and age formed a second tier of contributors, with elevated urinary protein and reduced urine output increasing the predicted CKD risk. Hypertension and diabetes contributed more modestly, each raising the predicted risk when present, whereas medication class and water intake had comparatively little influence on model output. These patterns are physiologically coherent with established markers of renal function and support the clinical plausibility of the model. Overall, after excluding GFR to avoid outcome leakage, both ensemble models classified CKD with excellent and comparable performance, and XGBoost was retained as the final model.

IV. Discussion

In this study, we developed and compared two tree-based ensemble classifiers—random forest and XGBoost—for the detection of chronic kidney disease (CKD) using a publicly available kidney-function dataset. Because the CKD label in this dataset is derived directly from the glomerular filtration rate (GFR), we deliberately excluded GFR from the predictor set and trained the models on the remaining clinical and laboratory variables. Both models achieved excellent discrimination on the held-out test set, correctly identifying every CKD case (sensitivity = 1.000), and XGBoost was marginally superior across accuracy (0.997), specificity, positive predictive value, F1 score, and the Matthews correlation coefficient. Both models were also well calibrated, with XGBoost showing the lower Brier score (0.003), so its predicted probabilities can be interpreted as meaningful risk estimates rather than bare labels.

These findings are consistent with the broader literature, in which random forest and gradient boosting repeatedly emerge as top performers for CKD classification. For example, Raihan et al. reported high accuracy for an XGBoost classifier combined with SHAP-based interpretation of the contributing attributes (11). Our SHAP analysis ranked serum creatinine and blood urea nitrogen (BUN) as the two dominant predictors, followed by urinary protein and urine output. This ordering is clinically coherent: serum creatinine is the principal biochemical marker used to estimate GFR and stage kidney function (12), so its prominence—

together with BUN and proteinuria—confirms that the models learned physiologically meaningful relationships rather than spurious associations.

A distinguishing feature of our design is the explicit attention paid to target leakage. Because the outcome was defined from GFR, retaining GFR as an input would have produced artificially perfect performance that reflects the label definition rather than genuine predictive signal. Data leakage of exactly this kind has been identified as a widespread and under-recognized cause of over-optimistic, non-reproducible results across many machine-learning fields (13). By removing GFR and by applying the Synthetic Minority Over-sampling Technique only within the training and cross-validation folds—never to the test set—we guarded against both feature leakage and resampling leakage, so that the reported test metrics reflect performance on genuinely unseen data.

We also placed particular emphasis on calibration and uncertainty, aspects that are frequently neglected in applied machine-learning studies even though poorly calibrated models can be misleading in clinical decision-making (14). This is especially relevant when class imbalance is corrected, because oversampling methods such as SMOTE can distort predicted probabilities and harm calibration (15). Restricting resampling to the training folds and then confirming calibration on the untouched test set allowed us to obtain the benefits of imbalance handling while preserving reliable probability estimates, and the accompanying bootstrap confidence intervals convey the statistical uncertainty around each performance metric.

To make the models transparent, we interpreted the best-performing classifier using SHAP with a tree-based explainer, an approach specifically designed to yield fast and consistent explanations for tree ensembles and validated on medical prediction problems(16). The resulting feature-importance profile not only enhances trust but also provides a clinically auditable account of how each variable influences the predicted risk, which is an important prerequisite for the adoption of such tools in practice.

Several limitations should be acknowledged. First, the analysis relies on a single, publicly available dataset; the near-perfect discrimination observed here likely reflects a high degree of separability in these data and may not transfer to noisier, real-world electronic health records. Second, we did not perform external validation on an independent clinical cohort, which is essential before any claim of clinical utility. Third, the predictor set was relatively small and did not include some markers used in contemporary practice. Future work should therefore focus on external and prospective validation using real patient data, evaluation across more heterogeneous populations, and reporting in line with established guidance for prediction-model studies such as the TRIPOD statement (17). Despite these limitations, the present study demonstrates a leakage-aware, calibrated, and interpretable modeling workflow that distinguishes it from analyses that report perfect accuracy without addressing these methodological concerns.

V. REFERENCES

1. Bikbov B, Purcell CA, Levey AS, Smith M, Abdoli A, Abebe M, et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The lancet. 2020;395(10225):709-33.
2. Kalantar-Zadeh K, Jafar TH, Nitsch D, Neuen BL, Perkovic V. Chronic kidney disease. The lancet. 2021;398(10302):786-802.
3. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
4. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
5. Ghosh SK, Khandoker AH. Investigation on explainable machine learning models to predict chronic kidney diseases. Scientific Reports. 2024;14(1):3687.
6. Halder RK, Uddin MN, Uddin MA, Aryal S, Saha S, Hossen R, et al. ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application. Journal of Pathology Informatics. 2024;15:100371.
7. Chen H, Huang Y, Chen L. Ensemble machine learning for predicting renal function decline in chronic kidney disease: development and external validation. Frontiers in Medicine. 2025;12:1598065.

8. Arif MS, Rehman AU, Asif D. Explainable machine learning model for chronic kidney disease prediction. Algorithms. 2024;17(10):443.
9. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321-57.
10. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.
11. Raihan MJ, Khan MA, Kee SH, Nahid AA. Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. Sci Rep. 2023;13(1):6263.
12. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, Sang Y, et al. New creatinine-and cystatin C–based equations to estimate GFR without race. New England Journal of Medicine. 2021;385(19):1737-49.
13. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9).
14. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW, tests TGEd, et al. Calibration: the Achilles heel of predictive analytics. BMC medicine. 2019;17(1):230.
15. Van den Goorbergh R, Van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. Journal of the American Medical Informatics Association. 2022;29(9):1525-34.
16. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nature machine intelligence. 2020;2(1):56-67.
17. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Journal of British Surgery. 2015;102(3):148-58.