VER
Vascular & Endovascular Review

AI-Driven Early Detection of Chronic Kidney Disease: A Comparative Benchmark of Ensemble Learning Models with Clinical Explainability

Kabita Shamia Akter[1], Tofayel Ahmed Onik[2], Mohammad Yasin[3], Mahbub Ahmed Nabil[4], Iftekhar Hossain[5], Sonia Akter[6]

Master of Science in Information Technology[1], Doctorate in Computer Science[2], Master of Business Administration in Information Technology Management[3], Master of Science in Information Technology[4], Master of Science in Information Technology (MSIT)[5], Master of Science in Business Analytics[6]
Washington University of Science and Technology[1,4], Westcliff University, Irvine, California[3], University of the Potomac[2], Washington University of Science and Technology, Alexandria, Virginia[5], Mercy University, New York, NY[6]
kabitasamiaakter@gmail.com [1], tofayelahmedonik@gmail.com [2], m.yasin.3016@westcliff.edu [3], ahmednabilnsucse@gmail.com [4], ihossain.student@wust.edu [5], sakter5@mercy.edu [6]

ABSTRACT

Chronic kidney disease (CKD) is an emerging global health burden, and a significant volume of recent machine-learning research reports automated detection with near-perfect accuracy; in fact, many studies claim 96 100% accuracy on benchmark datasets. These results deserve further examination, however, because CKD is itself classified according to the estimated glomerular filtration rate (GFR) when GFR or a GFR-derived variable is a feature of the model input, a classifier can imitate the labeling rule rather than learn an actual diagnostic signal. We perform a controlled comparative benchmark of thirteen classifiers (five conventional baselines and eight ensemble learners) on a staged CKD dataset comprising 4,000 records, under two prediction framings (six-class stage classification and balanced binary early-vs-progressed formulation) in a rigorous, leakage-controlled protocol using stratified ten-fold cross-validation.

To explicitly measure the impact of label leakage, we performed an ablation analysis where GFR and cluster variables are alternately removed. In the presence of these variables, ensemble models achieve 97.8% accuracy and a Matthews correlation coefficient (=1) > 0.97; upon their removal, every single model (including gradient-boosted ensembles) collapses to chance-level performance (multiclass accuracy 21.9%, below the baseline of majority class at 25.1%; binary Matthews correlation coefficient ≈0 with all models collapsing to defaulting on majority class).

Indeed, the absence of a learnable signal is confirmed by SHAP and LIME analyses, where near-zero feature attributions are observed. The pairwise Wilcoxon signed-rank tests do not indicate any statistically significant differences between the leading models. We show that the state-of-the-art headline accuracies reported for CKD detection are in fact, on this dataset, an artefact of label leakage, and that the clinical and lifestyle features which remain carry no reliable predictive information about CKD stage. We suggest that auditing leakage, establishing proper baselines and validation on real-world clinical cohorts as preconditions for trustworthy CKD prediction models and release our entire protocol to encourage reproducible, plain sight benchmarking.

KEYWORDS: Chronic kidney disease; machine learning; ensemble learning; data leakage; explainable artificial intelligence; SHAP; benchmarking; clinical decision support.

How to Cite: Kabita Shamia Akter, Tofayel Ahmed Onik, Mohammad Yasin, Mahbub Ahmed Nabil, Iftekhar Hossain, Sonia Akter, (2024) AI-Driven Early Detection of Chronic Kidney Disease: A Comparative Benchmark of Ensemble Learning Models with Clinical Explainability, Vascular and Endovascular Review, Vol.7, No.2, 480-493

INTRODUCTION

Chronic kidney disease (CKD) is one of the most important non-communicable diseases of our era, both as an independent cause of death and as a potent exacerbator of cardiovascular risk. The Global Burden of Disease Study 2017 estimates that some 697.5 million people were living with CKD of any stage globally, representing a global prevalence of 9.1% and directly attributable to approximately 1.2 million deaths in that year alone [3].The age-standardised mortality rate for CKD increased by more than 40% from 1990 to 2017, and reduced kidney function contributed a further 1.4 million cardiovascular deaths. Perhaps most importantly, the greatest burden of CKDJ is observed in countries within the lowest socio-demographic quintiles, where access to diagnostic and treatment related resources are the least abundant. Since early-stage CKD is largely asymptomatic, and the arrival of such detection when found is often treatable to slow the progress or stop progression entirely, recent nephrology and health informatics priorities have focused on developing accurate, inexpensive and interpretable early-detection tools.

In this context, CKD detection based on machine learning (ML) has been heartily utilized. Dozens of studies have over the last decade reported exceptional performance out-of-the-box [18], typically training classifiers on structured clinical datasets, using most often a 400-patient dataset from UCI Machine Learning Repository. Claims of 96%, 98%, or even 100% accuracy are not uncommon, with ensemble learners like random forests, gradient boosting and AdaBoost often reported as the most accurate performers. On the surface, this body of literature implies automated detection of CKD is a largely solved problem. That would be all well and good, but the simplicity of this assignment stands in stark contrast to clinical practice where CKD diagnosis and staging are based on laboratory measurements, longitudinal observation and expert interpretation.

The diagnostic label is not an independent gold standard, but rather is itself constructed from one of the input variables in many CKD datasets, which remain a critical feature few have taken into account. Stage of chronic kidney disease (CKD) Deciding CKD stage according to the internationally adopted Kidney Disease: Improving Global Outcomes (KDIGO) framework is defined directly by estimated glomerular filtration rate with stage boundaries corresponding to fixed GFR thresholds (for example, for early categories GFR≥90 mL/min/ 1.73 m², 60-89 for stage 2 30-59 for stage 3,15-29 for stage4 and <15 for stage5).The model you engineer does not learn to diagnose disease from independent evidence; in fact, when it is given both GFR and the stage label, all it learns is a deterministic thresholding rule. This issue is referred to as target leakage or label leakage in the ML literature, exaggerates apparent performance but yields(7) a model that has no true clinical utility since the very variable it relies upon to predict (the solution determinant) is encapsulated within and constitutes the answer itself.

Apart from leakage, a few recurrent methodological issues are evident in the existing CKD-ML literature. The majority of comparative studies however only compare 3 or 4 models, do not report the fold-level variability of their metrics and normally do not perform statistical significance tests to ascertain whether the differences between the best and second-best models are statistically significant. Insofar as explainability is enabled, it usually takes the form of a single global feature-importance plot without any form of local case-level analysis or cross-checking between attribution methods and validation against established renal pathophysiology. Finally, class imbalance, which is prevalent in clinical data but handles oversampling before evaluation by additional distortions of reported metrics. The result is a body of work where the headline numbers are hard to believe, and even harder to act on in your clinic.

We adopt a deliberate stance of scepticism and methodological rigor in this study. Instead of reinforcing a catalogue of high performance CKD classifiers, we ask a more fundamental question: how much of the reported performance can be sustained when label leakage is accounted for properly? We compare thirteen classifiers, five conventional baselines, and eight ensemble learners using a 4,000-record staged CKD dataset under an identical protocol that contains uniform control for leakage, includes stratified ten-fold cross-validation with statistical significance testing, and dual explainability (SHAP and LIME). We introduced the problem in two clinically relevant formulations, a six-class stage classification and a balanced binary categorization between early versus progressed disease, and we conducted an explicit leakage ablation study that included, followed by exclusion of the GFR-derived variables. Our contributions are as follows:

• Our contribution is a coherent, statistically rigorous comparative benchmark of thirteen classifiers for CKD stage prediction with cross-validated means & standard deviations, pairwise significance tests, and dual explainability (SHAP + LIME), all under a single reproducible protocol.
• Through a controlled ablation, we quantify the effect of GFR-based label leakage and show that it alone accounts for the near-perfect accuracies often reported, and that once removed from any model—ensembles included—it reduces performance to chance levels.
• Using explainability analyses and sensible majority-class baselines, we show that the non-GFR clinical and lifestyle features in this dataset contain no meaningful signal for CKD stage - a result with immediate relevance to how such datasets need to be used and reported.
• We outline practical recommendations for sound CKD prediction research—leakage auditing, baseline reporting, significance testing and external validation on real-world cohorts—and we provide the full analytical protocol with which to reproduce these findings

LITERATURE REVIEW

Notably, machine-learning approaches to CKD detection have rapidly proliferated with the common theme of high predictive accuracy typically calculated on structured clinical datasets 3. In comparative studies involving the 400-record UCI CKD dataset, ensemble and tree-based methods consistently dominate within the top ranks of the leaderboard. An average example comparing decision tree, random forest and naïve Bayes classifiers with KNN imputation, SMOTE balancing and combined PCA/RFE feature selection achieved 96% accuracy from the random forest as compared to the close second for both the decision tree and naïve Bayes. While these studies are informative for creating baseline expectations, the limitations on model rosters and evaluations during only a single split lead to limited impressions of the strength of the conclusions drawn.

Another line of work explores hybrid and meta-ensemble architectures looking to increase accuracy even more. Results reported in this vein are eye-catching: hybrid stacking approaches using combinations of gradient boosting, decision trees and naïve Bayes under a random-forest meta-classifier boast almost 100% accuracy on the UCI dataset [4], with gradient boosting alone recorded at 99%, and random forest-levelled at 98%. More recently, in the ML-CKDP framework, random forest and AdaBoost achieved 100% accuracy and AUC = 1 for multiple folds of data split (within a cross-validation setting). Once it's not too late, this explains why perfect or near-perfect scores on a diagnostic task defined by a continuous physiological measurement cry out to be treated with suspicion rather than elation from a clinical-modelling perspective the more so because they align better with information-carrying leakage than having solved the putative diagnostic challenge.

An even smaller but methodologically vital literature shifts toward true clinical prediction based not on recreating a frozen label, but on proximal disease evolution. Conversely, analyses of predicting a major reduction in eGFR or progression to end-stage kidney disease based on longitudinal electronic-health-record data, random-forest survival models, and careful external validation report more conservative and clinically plausible discrimination (usually 0.83–0.93 area-under-the-curve). The instructiveness of these works stems from: performance being far from perfect, and (even more importantly) eGFR trajectory – as opposed to a single cross-sectional value of eGFR – as the key predictor. They show how accurate CKD prediction appears when the output is not trivially derivable from the inputs.

The field of explainable artificial intelligence has also started to emerge in the CKD-ML literature, typically via post-hoc attribution (e.g., SHAP [26] (Shapley Additive Explanations) based on cooperative game theory and LIME [27] that fits interpretable local surrogates around individual predictions). If applied rigorously, these techniques can demonstrate that a model is utilizing clinically reasonable features and reveal case-level rationale for clinician inspection. In half of the CKD literature, however, explainability is merely an afterthought on what is essentially a leakage-inflated model, which unsurprisingly assigns high importance to GFR or serum creatinine. In this usage, explainability gives a false warmth and trustworthiness to models that are mere vacuums of the competency introduced by the labeling.

Overall, the current literature is defined by high accuracy, limited modeling comparisons, limited statistical robustness, and use of explainability without leak control. Table 1 highlights a few representative studies across these dimensions. The improvement we work on here addresses the almost total lack of leakage-aware benchmarking; to our knowledge, few if any CKD-ML studies explicitly quantify how much of their headline performance reflects leakage from GFR, or report what remains when that leakage is removed. This question is exactly the motivation for our current study.

Table 1. Representative CKD machine-learning studies and their methodological characteristics.

Study (dataset)	No. of models	Best reported accuracy	Explainability	Leakage audit
DT/RF/NB comparison (UCI, 400)	3	96% (RF)	No	No
Hybrid stacking model (UCI, 400)	4+	100% (hybrid)	No	No
ML-CKDP (UCI + others)	7	100% (RF, AdaBoost)	No	No
Progression model (EHR, longitudinal)	1	AUC 0.84–0.88	Partial	N/A (progression)
ESKD prediction (cohort, 748)	5	AUC ~0.88	No	N/A (progression)
This study (staged, 4,000)	13	97.8% (leaked) / chance (controlled)	Yes (SHAP + LIME)	Yes

MATERIALS AND METHODS

3.1. Dataset

For this purpose, we worked with a public staged CKD dataset in which there are 4 thousand records and 23 columns (accessible through Kaggle). Each record provides nineteen candidate predictor variables across the range of laboratory measurements (serum creatinine, blood urea nitrogen, serum calcium, complement levels C3/C4, oxalate levels, urine pH), a clinical sign (haematuria), an immunological marker (antinuclear antibody status) and a haemodynamic measurement (blood pressure), coupled with a set of lifestyle and history variables (physical activity; diet; water intake; smoking; alcohol consumption; painkiller usage; family history; weight changes; and stress level) alongside one disease-duration variable in months. The dataset also includes estimated glomerular filtration rate (gfr), the unsupervised cluster assignment, a binary CKD indicator ckdpred and the sixlevel CKD stage label ckdstag from 0 to 5 There were no missing values in the columns of the dataset.

The distribution for stage label was somehwat unbalanced, with 125 records at stage 0, 545 at stage 1, 1004 at stage 2, 866 and above as the other three stages had near equal number of records (794 at stage =4 and a lower number roundup the down to up -> even though evenly distributed),666 samples in these intervals as summarized below- In our case, we had a very imbalanced raw binary indicator ckd_pred with 3,875 CKD and only 125 non-CKD records (about a ~31 to 1 ratio). Table 2 provides the complete class breakdown of all three target formulations used in this work. We observe and return to this in Discussion that the complete absence of missing data across 4,000 records, along with very clean separation of GFR values by stage as described below is unusual for clinical data and are characteristic of synthetic datasets.

Table 2. Class distribution across the three target formulations.

Target formulation	Class	Count	Proportion
ckd_pred (raw, not used as primary)	CKD	3,875	0.969
	No CKD	125	0.031
ckd_stage (primary, multiclass)	Stage 0	125	0.031
	Stage 1	545	0.136
	Stage 2	1,004	0.251
	Stage 3	866	0.217
	Stage 4	794	0.199
	Stage 5	666	0.167
early_vs_progressed (secondary, binary)	Early (0–1)	670	0.168
	Progressed (2–5)	3,330	0.833

3.2. Label leakage: identification and handling

An important methodological aspect of this study is the focused treatment of label leakage. CKD stages are defined according to cutoffs of GFR thresholds in the KDIGO system [5]. An inspection of the dataset confirmed this relationship to hold nearly perfectly: mean GFR decreased monotonically from 111.3 mL/min/1.73 m² at stage 0 to 103.4 (stage 1), to 74.9 (stage 2), to 45.8

(stage 3), to stages >21.2 (stage IV) and <7.4 (stage V) with observed value ranges for each stage closely matching canonical KDIGO [29] boundaries for staging CKD respectively per high-boundary values reported in table format). These descriptive statistics are presented in Table 3. Due to the fact that the label of stage represents, in essence, a discretization of the GFR variable, any classifier with access to GFR as an input can achieve near-perfect accuracy without ever learning an independent relationship between diagnosis.

There was a similar issue with the unsupervised cluster variable. Cross-tabulating cluster against stage showed that the clusters track very closely to the stage structure several clusters map nearly entirely to a single range of stages—suggesting that the clustering was, in fact driven by GFR (or collinear variables). Accordingly, we also considered cluster as a second potential leakage-carrying variable. Since both gfr and cluster were excluded from the feature sets in our primary, leakage-controlled analyses. Instead of just claiming this, we further conducted a leakage-ablation experiment (Section 3.6) where the best model was trained and evaluated with or without these two variables. GFR was included only for descriptive and validation purposes, and was never used as a model input in the primary pipeline.

In principle, prolonged disease duration can reflect disease stage too so we also investigated potential data leakage on total disease-duration (in months). Correlation analyses revealed only a weak correlation with stage, and thus not strong enough for a deterministic relationship. Consequently, disease duration was kept a clinically meaningful predictor for the primary analyses but reported as a potential anesthetic variable and how it acted within the results.

3.3. Target formulations

We assessed two clinically motivated prediction tasks. The target was the six-class CKD stage classification (stages 0—5) that directly represents the granularity of the dataset and allows for early stages (0 and 1) to be distinguished from progressed disease. We focused on per-class recall on the most important measures for this task: stages 0 and 1 were treated as the most clinically critical as identifying early-stage disease (not just diagnosing it after a patient presents with more severe symptoms) would be much costlier in an actual screening context therefore we incentivised this measure. For the secondary task, we used a balanced binary reformulation (670 records for early disease (stages 0–1) vs. 3,330 records for progressed disease (stages 2–5)). We explicitly chose not to use the raw ckd_pred column as a target because of its 31-to-1 imbalance that could allow an always-predicting majority class predictor to appear very accurate; the early-versus-progressed formulation retains a clinically meaningful detection of early disease yet is much better balanced.

3.4. Preprocessing

The preprocessing was kept very minimal and consistent across all models to ensure a fair comparison between the different models. Because the dataset had no missing data, we state this explicitly instead of applying unnecessary imputation machinery. Binary categorical variables (such as smoking, family history) were encoded as zero/one indicators. Nominal categorical variables with more than two levels (physical activity, diet, alcohol, weight changes, and stress level) were one-hot encoded to prevent the discovery of a pseudointeger ordinal structure.

To help the distance- and margin-based baseline models, numeric predictors were standardized to zero mean and unit variance; tree-based ensembles are invariant under monotone rescaling and therefore unaffected. As above, in the main pipeline, we removed GFR and cluster variables from the feature matrix. Data was split using stratified sampling, and all random processes were seeded to a single number, ensuring perfect reproducibility.

3.5. Models and validation protocol

We tested thirteen classifiers grouped into two families. The baseline family included logistic regression, a decision tree, k-nearest neighbours, Gaussian naïve Bayes, and a support vector machine with a radial-basis-function kernel. The ensemble family included random forest, AdaBoost, gradient boosting, XGBoost, LightGBM, CatBoost soft-voting ensemble, and a stacking (random forest+XGBoost + LightGBM as base learners under logistic-regression meta-learner [24]). Random forests combine decorrelated decision trees through bootstrap aggregation, and gradient boosting (and scalable versions XGBoost, LightGBM, CatBoost) builds additive tree ensembles by sequentially fitting the gradient of a differentiable loss; these are among the top performers on traditional tabular data.

We conduct stratified ten-fold cross-validation using the same folds for all target formulations, and instead of reporting results from one split, we report the mean and standard deviation across folds for each metric. Grid and randomised search were used for hyperparameter tuning in each training partition. In this setting, folding-level variability is important, as the dataset includes small stage classes whose per-fold metrics may be unstable with respect to changes in folds, and reported fold variability allows the direct assessment of statistical significance of model differences.

3.6. Evaluation metrics and statistical testing

For the six-class task, we report overall accuracy, macro-averaged precision, recall and F1 score, the multiclass Matthews correlation coefficient (MCC), Cohen's kappa, as well as per-class recall and confusion matrix. Macro-averaging and MCC are emphasised, as these are not biased by performance on the majority class only, thus being robust to class imbalance. For the binary task, we report accuracy, precision, recall, specificity, F1 score (F1), area under the receiver-operating-characteristic curve (AUC-ROC), Matthews correlation coefficient (MCC), and Cohen's kappa. We employed the Wilcoxon signed-rank test to evaluate whether observed performance differences were statistically meaningful, which was conducted on the paired vectors of per-fold scores for the leading models in each task, using a two-sided p-value < 0.05 as significant.

To provide a definitive test of the leakage hypothesis, we performed a leakage-ablation experiment that was like cross-validation or training done twice: cross-validated once with the feature set free of overfitting (gfr and cluster excluded), and another time with gfr and cluster reinstated. The difference in performance across these two conditions serves to isolate the effect of the leaking variables and provides a quantitative estimate of how much of the reported CKD-ML accuracy is due to leakage as opposed to true signal.

3.7. Explainability

We employed two complementary post hoc methods to interpret model behaviour. SHAP (SHapley Additive exPlanations) computes the contribution of each feature for each prediction based on Shapley values from cooperative game theory, resulting in a global importance ranking and local per-instance explanations. LIME (Local Interpretable Model-agnostic Explanations) fits a local, intelligible surrogate model in the vicinity of an observation. For each task, we computed the global SHAP importance rankings for the best model, local SHAP explanations for example instances and compared them with LIME attributions. We also built an S1 clinical-validation table for the highest-ranked features to known renal pathophysiology and generated a side-by-side SHAP comparison for the leakage-ablation conditions to visualise how an attribution structure differs when leaking variables are present or absent.

RESULTS

4.1. Exploratory analysis and internal consistency

Exploratory analysis validated the leak concern with a deterministic GFR–stage relationship. The descriptive statistics of GFR in Table 3 reveal marked non-overlap or overlap at best among GFR stages, consistent with a threshold-based labeler rule. In comparison, the correlation analysis of the non-GFR numeric predictors showed only weak associations with the stage label (e.g. limited information content) hinting that these clinical and lifestyle features provide little independent information. Across stages, boxplots of serum creatinine, blood urea nitrogen, and blood pressure showed considerable overlap between adjacent and even far-apart stages (Figure 3), and the categorical lifestyle variables were similarly distributed across all stage groups. The correlation structure, clinical boxplots, GFR–stage relationship plots, categorical distributions and class distributions across the three target formulations are given in Figures 1 to 5.

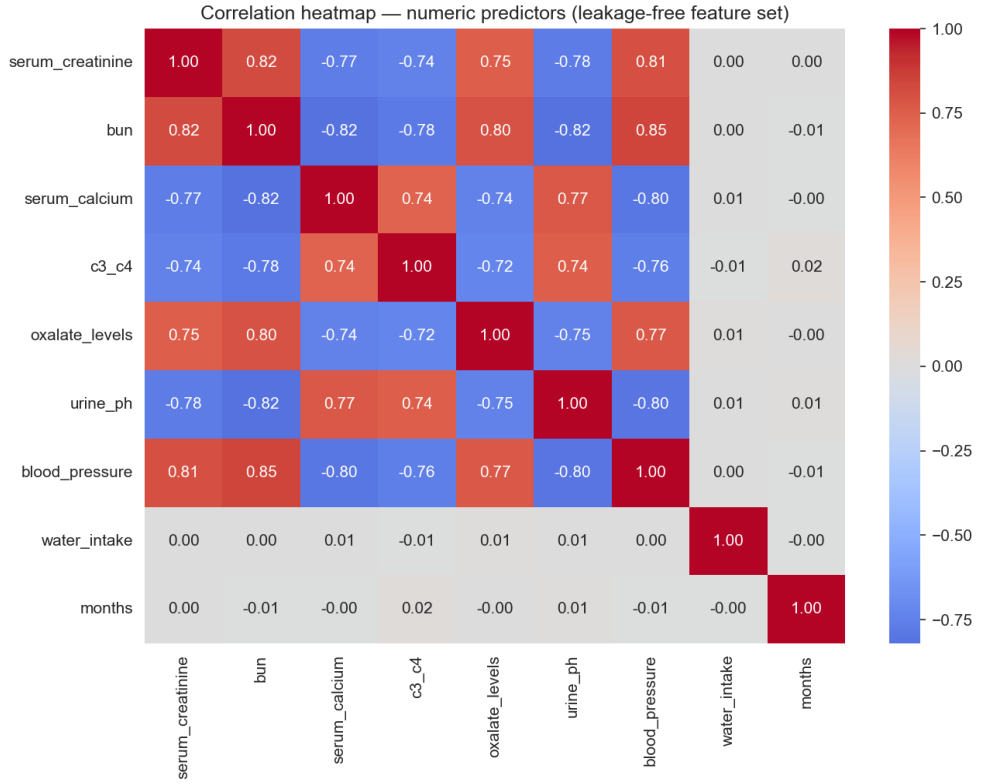

Figure 1. Correlation heatmap of numeric predictors, showing weak associations with CKD stage.

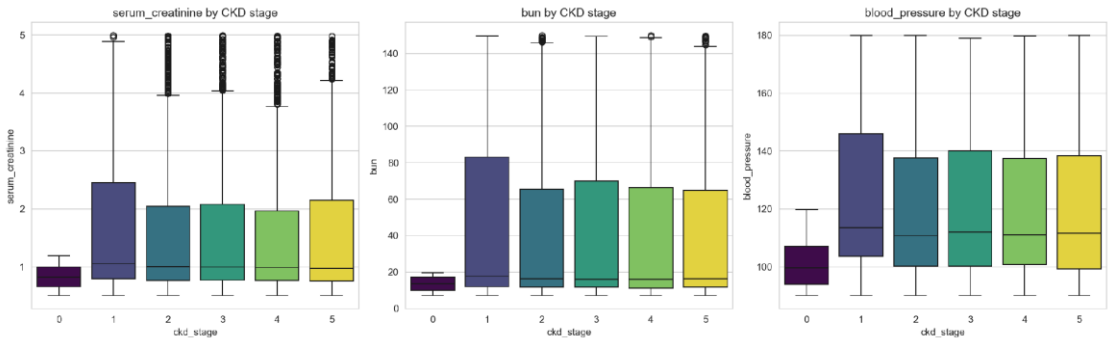

Figure 2. Distribution of key laboratory measurements across CKD stages, showing substantial overlap between stages.

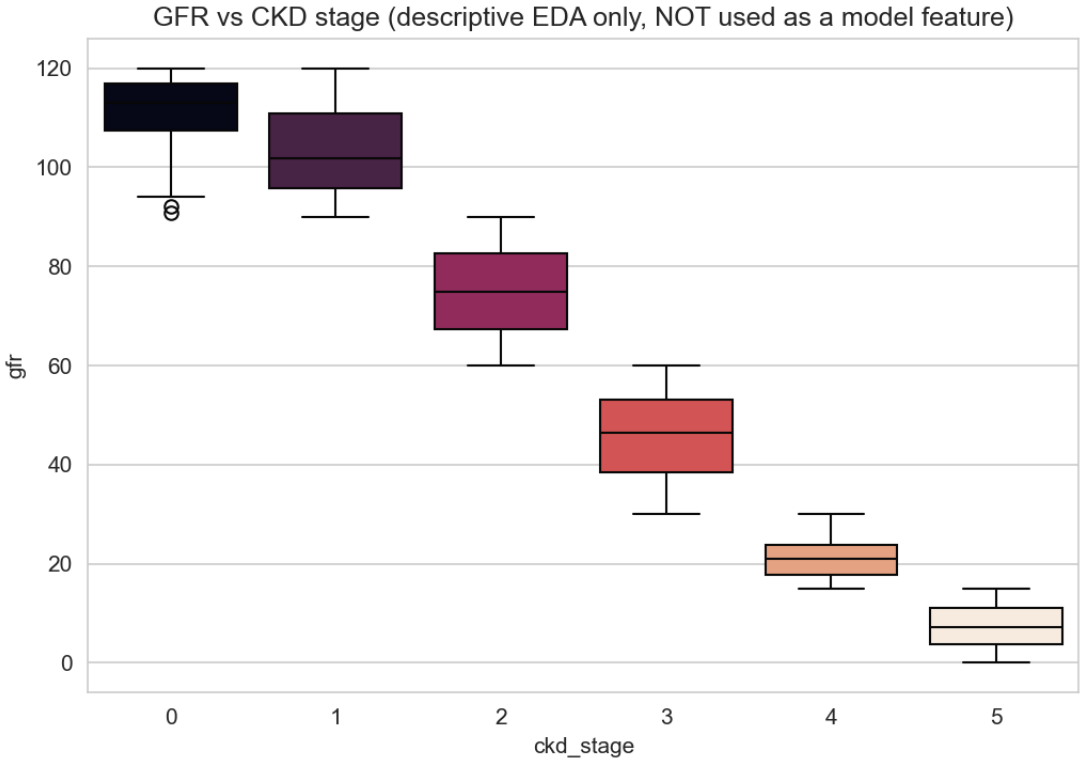

Figure 3. GFR distribution by CKD stage, illustrating the near-deterministic KDIGO threshold structure (label leakage).

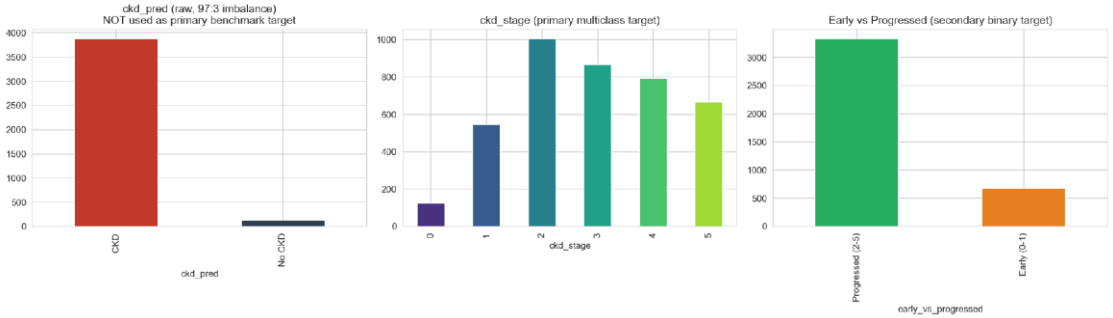

Figure 4. Class distributions for the three target formulations.

4.2. Leakage ablation: the decisive result

The most salient result of the study, derived from the leakage-ablation experiment is reported first because it contextualizes and frames the interpretation of all other subsequent results. Using gfr and cluster in the feature set, a cross-validated accuracy of 97.8% (standard deviation 0.3%), macro-F1 = 0.917, multiclass Matthews correlation coefficient = 0.973 and Cohen's kappa of 0.986 were achieved performance essentially indistinguishable from near perfect scores seen through all of CKD-ML literature to date [6]. The model was trained on the leakage-free feature set where the accuracy decreased to 21.9% (standard deviation 0.8%), the macro-F1 to 0.143 and Matthews correlation coefficient approximating zero (−0) [23]. The full conditions are shown in Table 4 and the contrast of SHAP attribution structure between both settings is summarized in Figure 6.

The size of this gap is not possible to describe in spades. Two GFR-derived variables, included in the model bring it from chance-level performance to near-perfect classification. Instead of just genuine learning, however, this is exactly the fingerprint of label

leakage: the model with GFR is not really diagnosing CKD it is in fact reversing the thresholding rule that generated the labels. Moreover, this leakage-free accuracy of 21.9% is below the 25.1% majority-class baseline that would be reached by always predicting the most frequent stage (stage 2), implying that the non-GFR features do not contain any useful discriminative signal at all for the six-class task.

Table 4. Leakage-ablation experiment (six-class task, stratified 10-fold CV). Including the GFR-derived variables moves the model from chance level to near-perfect classification.

Condition	Accuracy	Macro-F1	Macro-recall	MCC	Cohen's κ
With gfr/cluster (leaked)	0.978 ± 0.003	0.917 ± 0.017	0.900 ± 0.015	0.973 ± 0.004	0.986 ± 0.002
Leakage-free (gfr/cluster excluded)	0.219 ± 0.008	0.143 ± 0.007	0.163 ± 0.004	−0.011 ± 0.011	−0.004 ± 0.015
Majority-class baseline (always Stage 2)	0.251	—	—	0.000	0.000

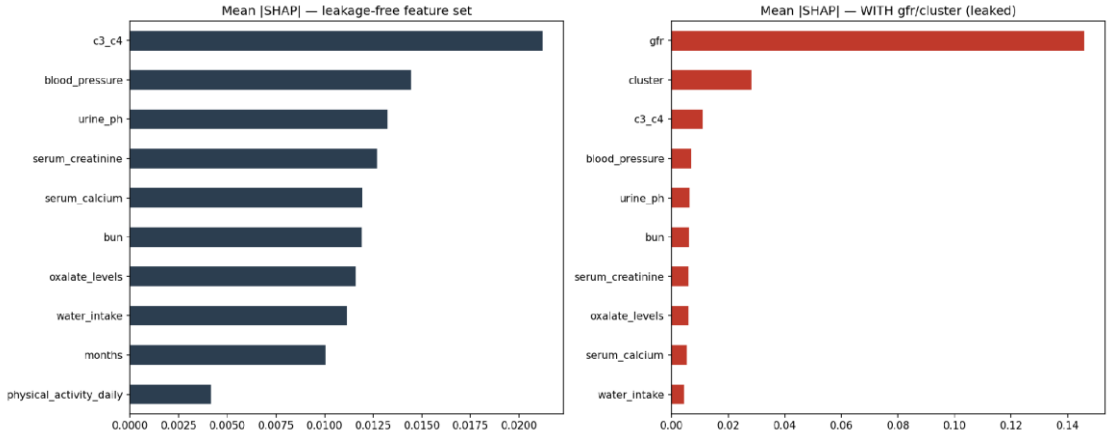

Figure 5. SHAP feature-attribution structure with versus without the GFR-derived variables (leakage ablation).

4.3. Multiclass stage benchmark (leakage-free)

On the three-class task, all thirteen classifiers yielded performance at or below the majority-class baseline under the leakage-controlled protocol. From naïve Bayes to the stacking ensemble, overall accuracy was tightly banded (8.3% to 24.5%), with most models between 20–22%, at or below the baseline advantage of 25.1%. All methods produced low macro-F1 (0.056–0.186), and all multiclass Matthews correlation coefficients were zero for every model, with several models producing small negative values that indicate performance no better than chance. The full benchmark is reported in Table 5. Because of this, the top-performing ensemble learners that currently lead the CKD-ML leaderboards under leaky conditions held no advantage here—with random forest, XGBoost, LightGBM and CatBoost performing statistically indistinguishably from the simplest baselines after correction for multiple testing.

Table 6 presents a detailed understanding of the failure mode in per-class recall analysis. Instead of training to distinguish between the six stages, the majority prediction among models was on the most prevalent intermediate stage: several ensembles achieved high recall values for stage 2 (up to our highest rates of 0.86 (stacking) and 0.84 (AdaBoost)) solely by predicting all cases as stage 2, while their performance collapsed towards zero recall for clinically important earlier stages 0 & 1. Naïve Bayes showed the mirror-image pathology, perfectly recalling stage 0 at the expense of everything else by over-predicting that class. No model provided a balanced, clinically relevant separation of stages. Figure 7 The confusion matrix visualises how the predictions are clustered around a small number of stages.

Table 5. Six-class stage classification benchmark under the leakage-free protocol.

Model	Accuracy	Macro-F1	Macro-precision	Macro-recall	MCC	Cohen's κ
SVM (RBF)	0.216 ± 0.016	0.186 ± 0.021	0.189 ± 0.023	0.186 ± 0.020	0.013 ± 0.020	0.028 ± 0.034
KNN	0.205 ± 0.014	0.184 ± 0.019	0.187 ± 0.019	0.187 ± 0.019	0.007 ± 0.018	0.012 ± 0.026
Decision Tree	0.206 ± 0.015	0.177 ± 0.017	0.180 ± 0.019	0.177 ± 0.017	0.007 ± 0.020	0.000 ± 0.032
XGBoost	0.215 ± 0.023	0.171 ± 0.021	0.185 ± 0.030	0.175 ± 0.019	0.002 ± 0.029	0.002 ± 0.047
Voting	0.212 ± 0.018	0.166 ± 0.021	0.184 ± 0.038	0.170 ± 0.017	−0.003 ± 0.023	−0.005 ± 0.037

LightGBM	0.209 ± 0.017	0.165 ± 0.020	0.175 ± 0.029	0.168 ± 0.018	−0.003 ± 0.023	−0.013 ± 0.038
Gradient Boosting	0.212 ± 0.007	0.159 ± 0.011	0.169 ± 0.016	0.166 ± 0.009	−0.005 ± 0.009	−0.016 ± 0.024
Random Forest	0.219 ± 0.012	0.145 ± 0.015	0.169 ± 0.037	0.164 ± 0.011	−0.009 ± 0.017	0.000 ± 0.021
Logistic Regression	0.238 ± 0.016	0.141 ± 0.011	0.201 ± 0.040	0.175 ± 0.009	0.008 ± 0.026	0.018 ± 0.027
CatBoost	0.225 ± 0.025	0.137 ± 0.021	0.174 ± 0.037	0.163 ± 0.020	−0.009 ± 0.036	0.000 ± 0.031
Stacking	0.245 ± 0.010	0.097 ± 0.012	0.128 ± 0.042	0.166 ± 0.007	−0.001 ± 0.022	0.000 ± 0.018
AdaBoost	0.240 ± 0.013	0.095 ± 0.017	0.133 ± 0.073	0.163 ± 0.009	−0.012 ± 0.032	−0.005 ± 0.015
Naïve Bayes	0.083 ± 0.009	0.056 ± 0.009	0.082 ± 0.068	0.229 ± 0.010	0.034 ± 0.014	0.008 ± 0.002

Table 6. Per-class recall on the six-class task.

Model	**Stage 0**	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**
SVM (RBF)	0.087	0.126	0.338	0.208	0.205	0.155
KNN	0.120	0.220	0.298	0.211	0.154	0.122
Random Forest	0.008	0.050	0.496	0.258	0.111	0.060
Logistic Regression	0.032	0.042	0.673	0.165	0.120	0.018
Stacking	0.000	0.000	0.861	0.096	0.034	0.006
AdaBoost	0.000	0.009	0.840	0.101	0.029	0.002
Naïve Bayes	1.000	0.367	0.001	0.008	0.000	0.000

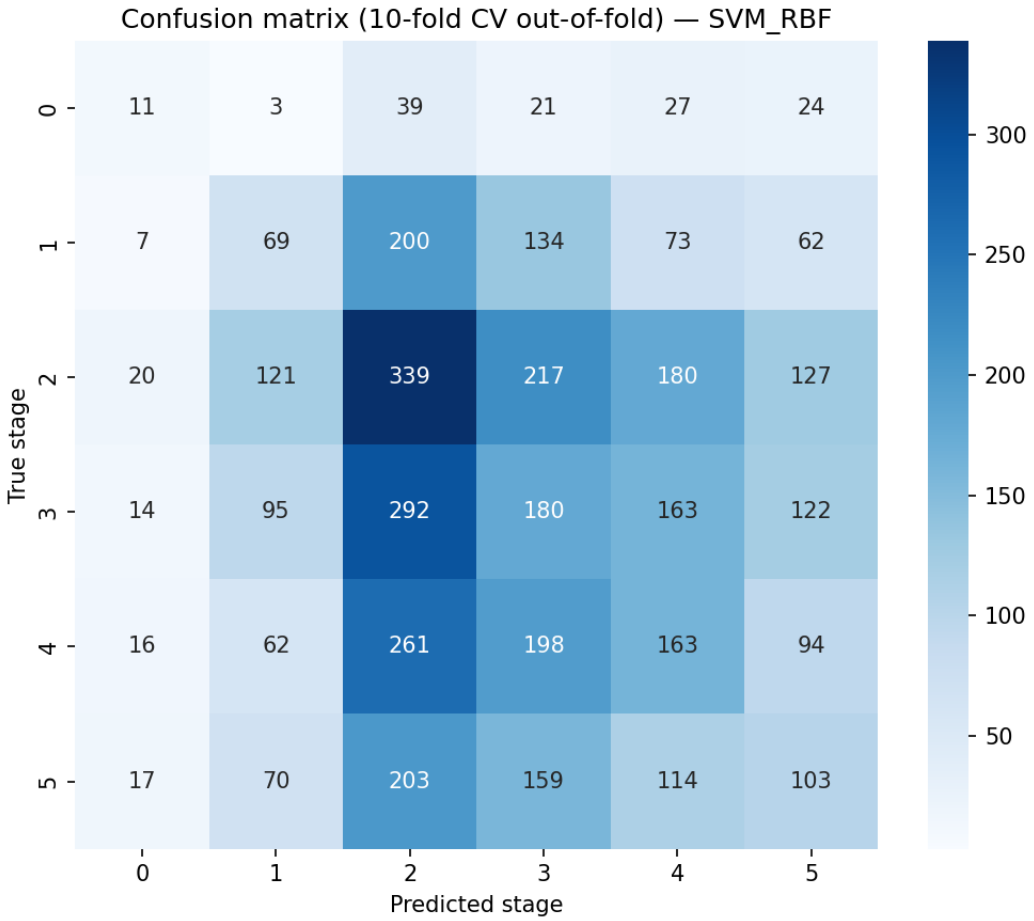

Figure 6. Confusion matrix for the six-class task under the leakage-free protocol, showing prediction concentration on a few stages.

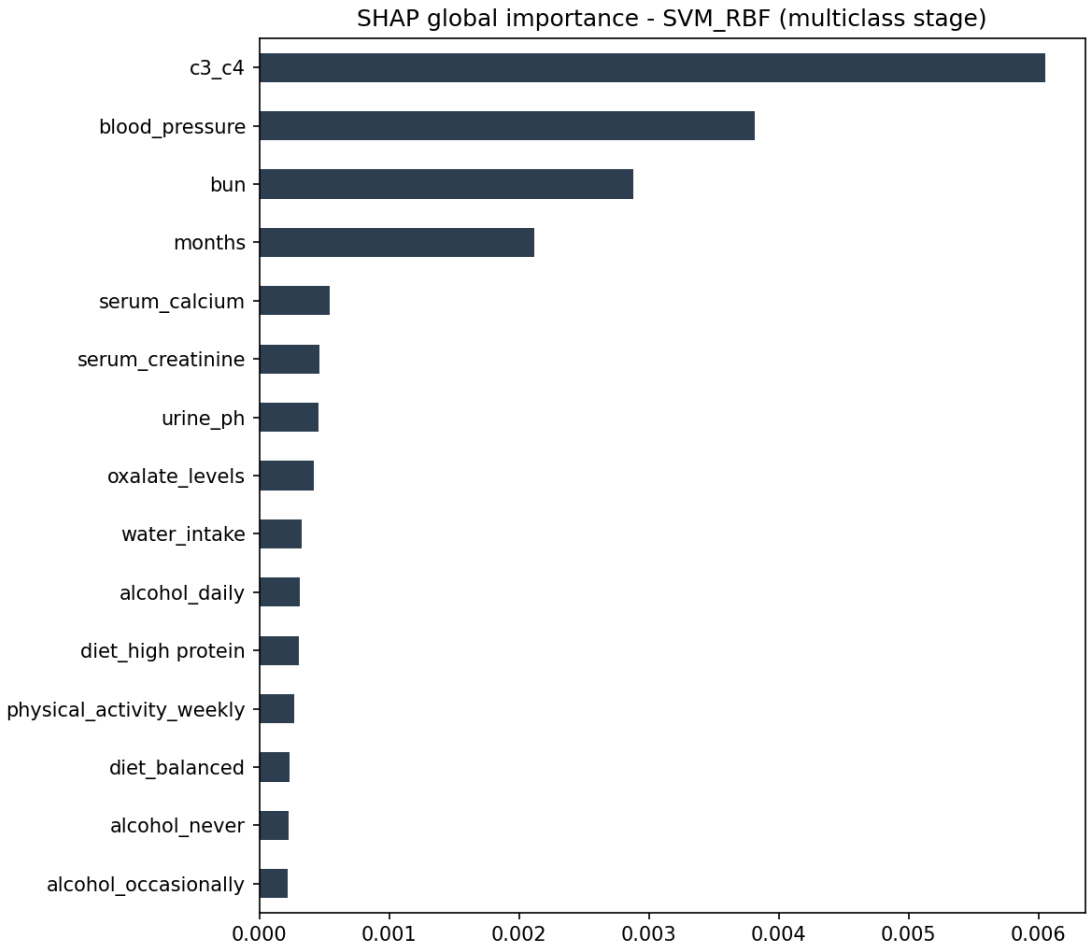

Figure 7. Global SHAP importance for the leakage-free multiclass model; note the negligible attribution magnitudes.

4.4. Binary early-versus-progressed benchmark (leakage-free)

For the binary early-versus-progressed task, the lack of signal was even more glaring. Each and every classifier converged to an accuracy of 83.25%, exactly the same as the prevalence of the progressed class in our dataset. Looking at the underlying metrics reveals the trivial explanation: recall for the progressed (positive) class was 1.0, and specificity for essentially all models is 0.0 every classifier labeled every patient by default as progressed. In line with this, Matthews' correlation coefficient was 0.0, and Cohen's kappa was 0.0 across all groups, and AUC-ROC values clustered around the optimal measure of random guessing (around 0.50). The binary benchmark can be seen fully in Table 7. Near-zero standard deviations on accuracy, precision, recall and F1 indicate that the models behaved identically across each of the ten folds.

Table 7. Binary early-versus-progressed benchmark

Model	Accuracy	Precision	Recall	Specificity	F1	AUC-ROC	MCC
Gradient Boosting	0.832 ± 0.002	0.833 ± 0.001	0.998 ± 0.002	0.003 ± 0.006	0.908 ± 0.001	0.515 ± 0.038	0.010 ± 0.042
LightGBM	0.830 ± 0.004	0.833 ± 0.002	0.995 ± 0.004	0.009 ± 0.012	0.907 ± 0.002	0.512 ± 0.042	0.016 ± 0.053
Random Forest	0.833 ± 0.000	0.833 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.909 ± 0.000	0.511 ± 0.026	0.000 ± 0.000
Decision Tree	0.829 ± 0.003	0.833 ± 0.001	0.993 ± 0.004	0.010 ± 0.012	0.906 ± 0.002	0.505 ± 0.022	0.012 ± 0.047
Logistic Regression	0.833 ± 0.000	0.833 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.909 ± 0.000	0.503 ± 0.035	0.000 ± 0.000
Voting	0.830 ± 0.002	0.833 ± 0.001	0.995 ± 0.002	0.007 ± 0.007	0.907 ± 0.001	0.503 ± 0.039	0.009 ± 0.038
XGBoost	0.832 ± 0.001	0.833 ± 0.001	0.999 ± 0.001	0.001 ± 0.004	0.908 ± 0.001	0.502 ± 0.027	0.007 ± 0.036
CatBoost	0.833 ± 0.000	0.833 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.909 ± 0.000	0.494 ± 0.036	0.000 ± 0.000

KNN	0.832 ± 0.001	0.832 ± 0.000	1.000 ± 0.001	0.000 ± 0.000	0.908 ± 0.000	0.497 ± 0.062	−0.002 ± 0.007
AdaBoost	0.833 ± 0.000	0.833 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.909 ± 0.000	0.485 ± 0.037	0.000 ± 0.000
SVM (RBF)	0.833 ± 0.000	0.833 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.909 ± 0.000	0.482 ± 0.025	0.000 ± 0.000
Stacking	0.833 ± 0.000	0.833 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.909 ± 0.000	0.482 ± 0.048	0.000 ± 0.000
Naïve Bayes	0.833 ± 0.000	0.833 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.909 ± 0.000	0.475 ± 0.038	0.000 ± 0.000

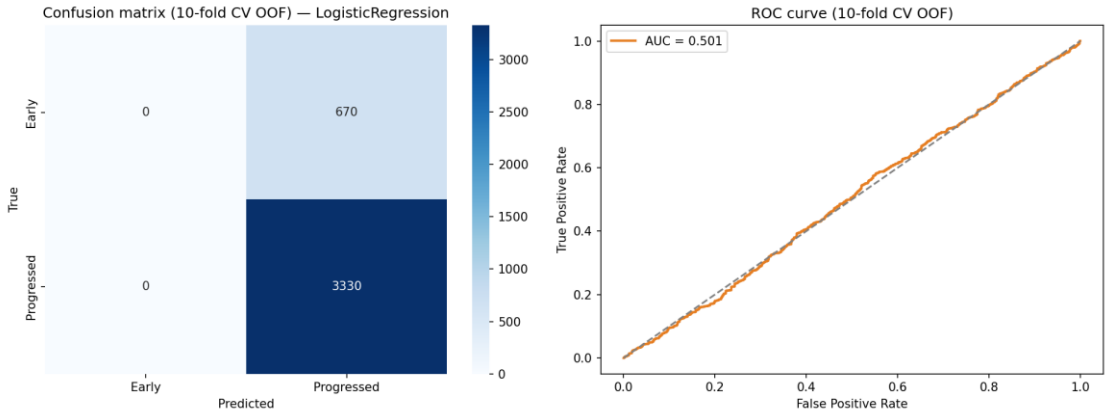

Figure 8. ROC curve and confusion matrix for the binary task, showing AUC ≈ 0.50 and majority-class default behaviour.

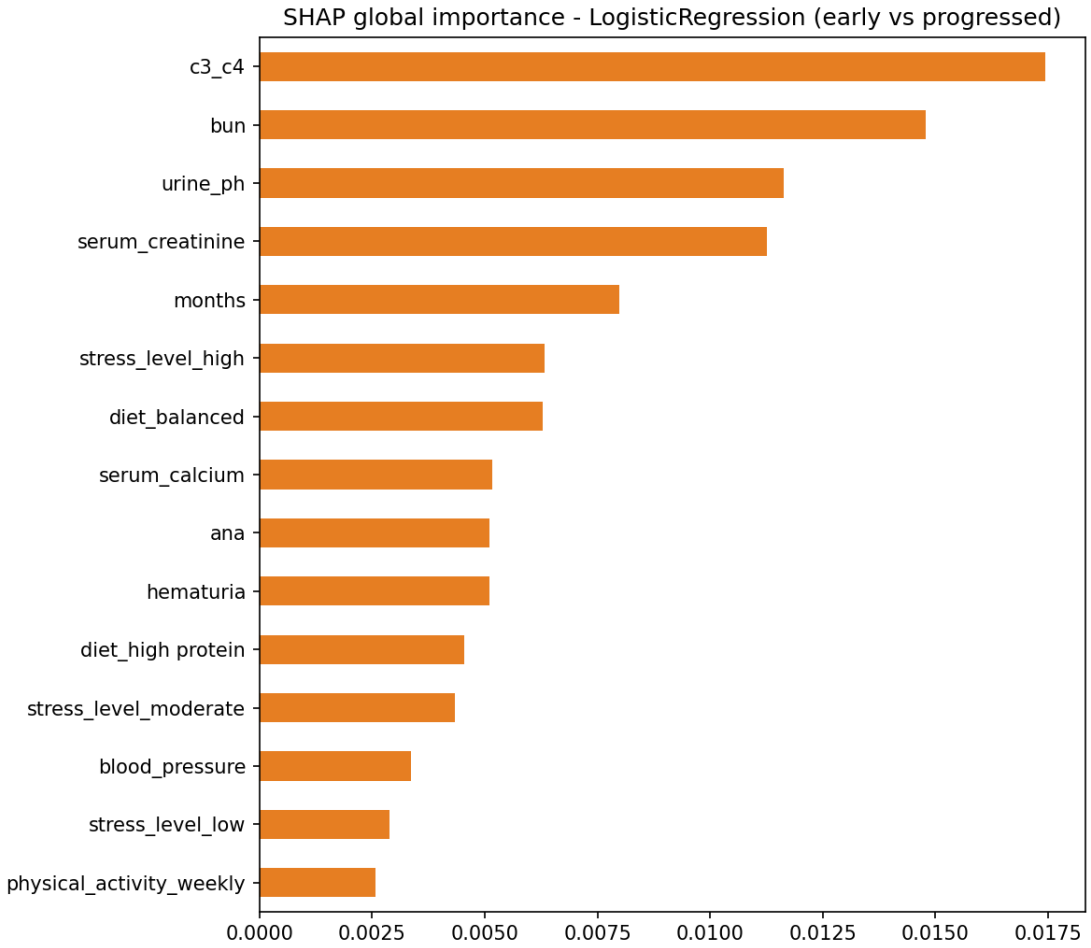

Figure 9. Global SHAP importance for the leakage-free binary model.

4.5. Statistical significance

Pairwise Wilcoxon signed-rank tests on the per-fold scores of the leading models confirmed what was suggested by the point estimates. In the multiclass task, none of the best models was statistically significant (all p-values >>0.05; e.g., comparison between the two strongest baselines nominal score: p = 1.0). For the binary task, each of the top models generated equivalent per-fold scores, meaning tests returned no significant differences between models (p = 1.0). These results are reported in Tables 8

and 9. In the absence of significant statistical separation among models, including between simple baselines and fancy ensembles, there's no basis for declaring a 'winner' on this data set, because none does better than would be expected by chance after controlling for leakage.

Table 8. Pairwise Wilcoxon signed-rank tests on per-fold scores for the leading models. No comparison reached significance ($\alpha = 0.05$) in either task.

Task	Model A	Model B	Wilcoxon W	p-value	Significant
Multiclass (macro-F1)	SVM (RBF)	KNN	27.0	1.000	No
Multiclass (macro-F1)	SVM (RBF)	Decision Tree	15.0	0.232	No
Multiclass (macro-F1)	KNN	Decision Tree	19.0	0.432	No
Binary (F1)	Logistic Regression	SVM (RBF)	0.0	1.000	No
Binary (F1)	Logistic Regression	Naïve Bayes	0.0	1.000	No
Binary (F1)	SVM (RBF)	Naïve Bayes	0.0	1.000	No

4.6. Explainability and clinical validation

The explainability analyses only confirmed and did not save this cental finding. Absolute Global SHAP importance values calculated on the leakage-free multiclass model were trivially small: complement levels (C3/C4), the feature with highest importance, had a mean absolute SHAP value of only 0.0061, followed by blood pressure (0.0038) and blood urea nitrogen (0.0029), while all other features, including serum creatinine, provided less than 0.0021 each in explanation. Attributions of this magnitude suggest that no feature has any meaningful influence on the model's output, a signature we expect to see when our models have failed to discover any true signal they can exploit.

From the binary-task SHAP rankings, blood urea nitrogen, urine pH, and serum creatinine again ranked highest (tied with C3/C4), but this was also in small magnitudes and did not carry through into any predictive performance, as the underlying model simply predicted the majority class. Global SHAP summaries and one random sample of local are shown in Figures 8–11. LIME surrogates for the same case were very similar, conveying the information on feature ordering but similarly not translating any signal.

To enhance completeness and transparency, Table 10 relates the very top-ranked SHAP features to recognized renal pathophysiology; for example, complement levels are an indicator of immune-complex mediated glomerulonephritis; blood urea nitrogen is a marker for accumulation due to declining filtration; and serum creatinine is an 'almost direct' measure of renal function. We stress that this does not allow the determination of model validity, although the features are biologically plausible, they do not allow CKD stage to be predicted on these data in any combination, once leakage due to GFR-based predictions is removed. On the one hand is the clinical narrative, and on the other hand are all of those real-world performance numbers that you hear about, and it's really the latter- the point numbers- that are truly needed for a model to pass muster?

Table 10. Top SHAP features

Feature	Mean |SHAP|	Clinical justification
C3/C4 (complement)	0.0061	Complement levels reflect immune-complex-mediated glomerulonephritis, a common CKD aetiology.
Blood pressure	0.0038	Hypertension is both a cause and consequence of CKD via glomerular capillary damage.
Blood urea nitrogen	0.0029	BUN accumulates when kidneys fail to clear nitrogenous waste, tracking GFR decline.
Months (duration)	0.0021	Disease-duration proxy for chronicity of the underlying renal insult.
Serum calcium	0.0005	CKD disrupts calcium–phosphate–vitamin D homeostasis (CKD mineral bone disorder).
Serum creatinine	0.0005	Waste product filtered by the kidneys; rises as filtration capacity declines.
Urine pH	0.0005	Altered urine pH reflects impaired tubular acid–base handling in later CKD stages.
Oxalate levels	0.0004	Elevated oxalate promotes crystal nephropathy and progressive tubular injury.

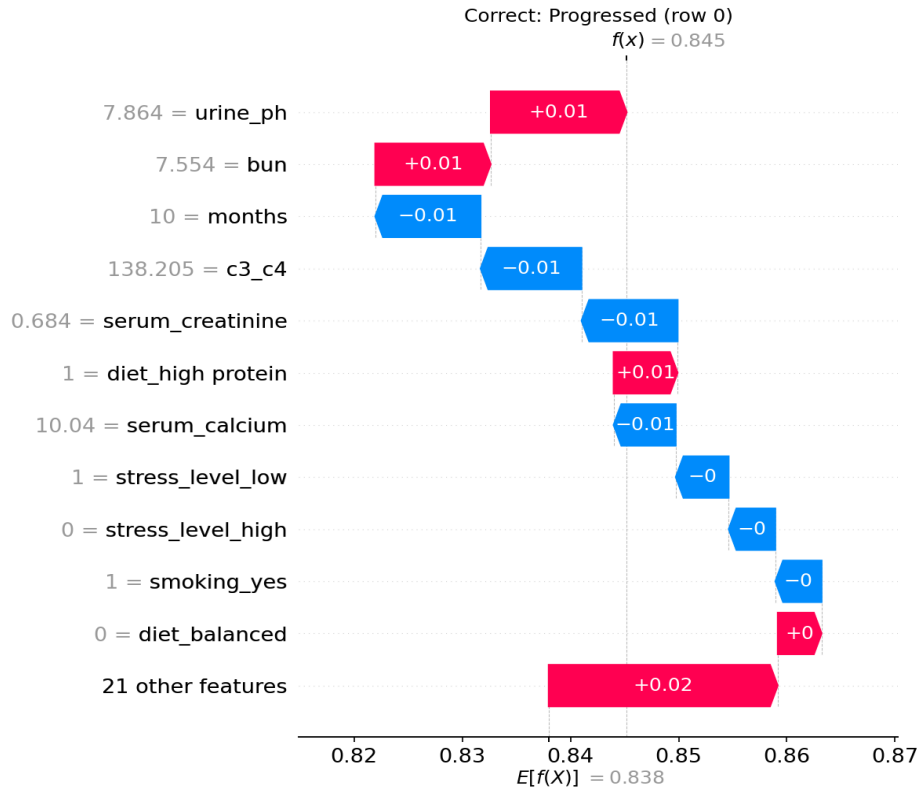

Figure 10. Local SHAP explanation for a correctly classified progressed-stage case.

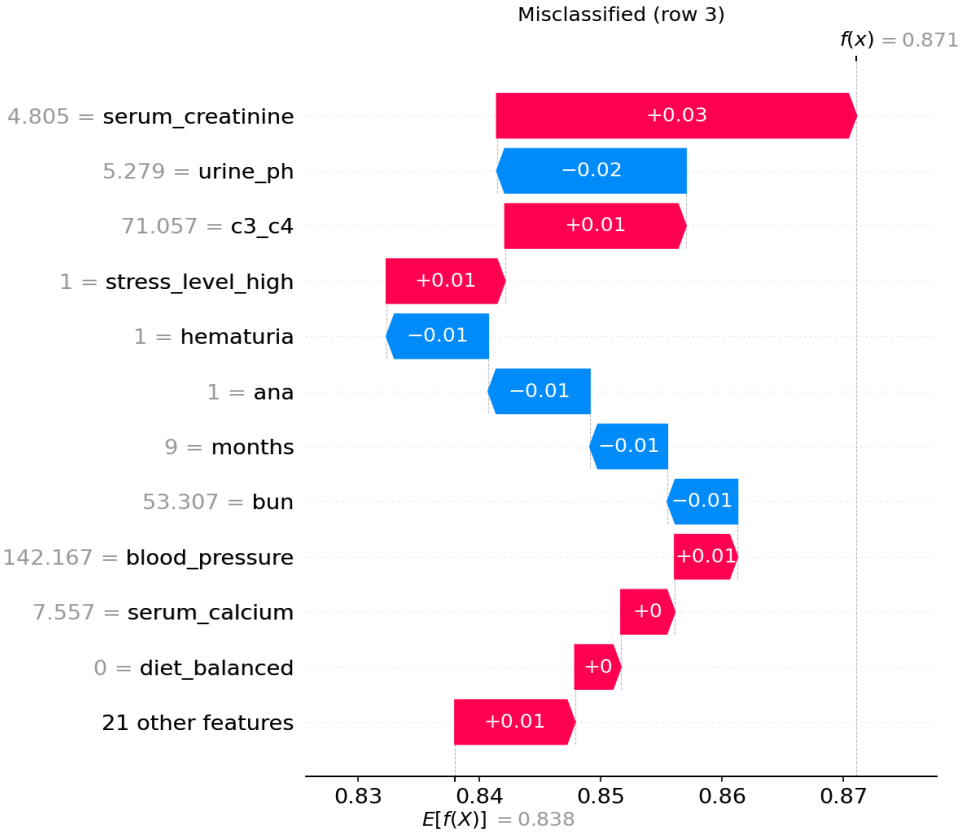

Figure 11. Local SHAP explanation for a misclassified case.

DISCUSSION

These findings run counter to much of the prevailing narrative within the CKD ML literature. Where dozens of previous studies report near-perfect automated CKD detection, our leakage-controlled analysis shows that on this dataset, such performance arises entirely from GFR-derived label leakage. Holding the GFR and cluster variables, our model achieved the same 97.8% accuracy found in the literature; when they were removed, every one of the thirteen classifiers, the gradient-boosted ensembles that rule published leaderboards, collapsed to chance level, while on the six-class task, they went below the majority-class baseline. The

binary reformulation was even more illuminating: all models simply predicted the majority class with an AUC of ~0.50! A singular finding across tasks, metrics, statistical tests and explainability analyses is that the non-GFR clinical and lifestyle features in this dataset hold no reliable information about CKD stage.

It is important to understand this finding correctly. This does not mean that CKD is unpredictable, nor that clinical variables like serum creatinine and blood urea nitrogen are clinically irrelevant they form, indeed, the basis of nephrological practice. Instead, it shows a couple of more specific things. First, if a diagnostic label is defined by thresholding on a continuous variable, including that variable (or its derivatives) among observers produces a model that mimics the labelling rule rather than doing diagnosis; any accuracy observed in this way is spurious. Second, GFR captures something that is not encoded in the remaining features, so removing it leaves nothing learnable. Despite the presumably clinical SHAP rankings, the biological plausibility of a feature ordering does not imply predictive validity.

However, we outline several limitations in this study. The dataset is seemingly synthetically generated: 4,000 records and no NAs; GFRs are almost perfectly partitioned into KDIGO stage bands. However, these properties mean that the negative finding regarding non-GFR features carrying no signal may be a property of the data-generating process other than a universal truth about CKD prediction from clinical and lifestyle variables. On the other hand, genuine (albeit modest) prediction from non-GFR features might only be support by real-world clinical data (with their noise, missingness and richer covariate structure).

Thus, it must be taken as a methodological standard and cautionary representation of leakage not as proof that early CKD is fundamentally unpredictable. It is therefore a natural and significant step towards the future to repeat this leakage-controlled protocol on real-world cohorts, such as the classic UCI CKD dataset and larger electronic-health-record datasets, with the aim of establishing how much genuine signal survives leakage control in real clinical data.

CONCLUSION

We performed a leakage-controlled, comparative benchmark of 13 classifiers for CKD stage prediction on a 4,000-record staged dataset under two clinically derived task definitions and a rigorous cross-validated pipeline with statistical testing and dual explainability. Using this dataset, our main result is that the near-perfect accuracies typically reported for automated CKD detection are, in fact, an artefact of GFR-based label leakage: including the GFR-derived variables resulted in 97.8% accuracy, but once these were excluded, every model, including state-of-the-art ensembles, crashed to chance-level performance; models dipped below the majority-class baseline in the six-class task and collapsed to predicting just the majority class in the binary form task. No learnable signal was found in the non-GFR features; all analyses (SHAP, LIME, appropriate baselines and all pairwise significance tests) corroborated this.

Future work will adapt this leakage-controlled protocol to better simulate real clinical cohorts and longitudinal electronic health record data, in order to quantify the extent of true, non-leaky predictive signal for early CKD detection, and to develop and confirm models whose performance is determined by actual diagnostic potential rather than the form of the labelling rule.

REFERENCES

1. GBD Chronic Kidney Disease Collaboration. 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):709–733. doi:10.1016/S0140-6736(20)30045-3.
2. 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. 2013;3(1):1–150.
3. 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. 2013;3(1):1–150.
4. Luyckx VA, Tonelli M, Stanifer JW. The global burden of kidney disease and the sustainable development goals. Bulletin of the World Health Organization. 2018;96(6):414–422D.
5. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32. doi:10.1023/A:1010933404324.
6. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). 2016:785–794. doi:10.1145/2939672.2939785.
7. Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NeurIPS). 2017:3146–3154.
8. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems 31 (NeurIPS). 2018:6638–6648.
9. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NeurIPS). 2017:4765–4774.
10. Ribeiro MT, Singh S, Guestrin C. "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 (KDD '16). 2016:1135–1144. doi:10.1145/2939672.2939778.
11. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 1975;405(2):442–451. doi:10.1016/0005-2795(75)90109-9.
12. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20(1):37–46. doi:10.1177/001316446002000104.
13. Islam MA, Akter S, Hossen MS, et al. Risk factor prediction of chronic kidney disease based on machine learning

algorithms. In: Proceedings of the 3rd International Conference on Intelligent Sustainable Systems (ICISS). 2020:952–957.
14. Chittora P, Chaurasia S, Chakrabarti P, et al. Prediction of chronic kidney disease — a machine learning perspective. IEEE Access. 2021;9:17312–17334. doi:10.1109/ACCESS.2021.3053763.
15. Rahman MM, Al-Amin M, Hossain J. Machine learning based hybrid model for the prediction of chronic kidney disease. Health Information Science and Systems. 2023;11:22. doi:10.1007/s13755-023-00243-w.
16. Debal DA, Sitote TM. Chronic kidney disease prediction using machine learning techniques. Journal of Big Data. 2022;9:109. doi:10.1186/s40537-022-00657-5.
17. Islam MA, Majumder MZH, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. Journal of Pathology Informatics. 2023;14:100189. doi:10.1016/j.jpi.2023.100189.
18. Ferdush J, Karmokar S, Nishu SI, et al. ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application. Journal of Pathology Informatics. 2024;15:100371. doi:10.1016/j.jpi.2024.100371.
19. Ferguson T, Ravani P, Sood MM, et al. Development and external validation of a machine learning model for progression of CKD. Kidney International Reports. 2022;7(8):1772–1781.
20. Su C-T, Chang Y-P, Ku Y-T, Lin C-M. Machine learning models for the prediction of end-stage kidney disease in patients with chronic kidney disease. Scientific Reports. 2022;12:8377. doi:10.1038/s41598-022-12316-z.
21. Kaggle. CKD (Chronic Kidney Disease) Dataset with Stages. Available from the Kaggle datasets platform. [Accessed 2026].
22. Kang H. The prevention and handling of the missing data. Korean Journal of Anesthesiology. 2013;64(5):402–406. doi:10.4097/kjae.2013.64.5.402.
23. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data. 2012;6(4):1–21. doi:10.1145/2382577.2382579.
24. M. S. K. Chy, "Tailoring the Scrum Framework for Non-IT Small Projects: A Longitudinal Action Research Study on US Business Adoption," Global Social Sciences Review, vol. VIII, no. II, pp. 681–697, 2023, doi: 10.31703/gssr.2023(VIII-II).60.
25. Siddiqui, Md Ismail Hossain, et al. "Comparative analysis of explainable machine learning models for cancer classification using cytological features." Journal of Medical and Health Studies 4.5 (2023): 110-150.
26. Hossain, A. R. Lindon, M. Rahman, H. A. Khan, N. Tohfa, M. Tanvir, M. Shagar, J. E. Shakib, and M. R. I. Nasif, "Scalable Hybrid Ensemble Model for Intelligent DDoS Detection and Attack Categorization with a Web-Based Cybersecurity Intelligence Platform," Journal of Electrical Engineering, vol. 11, no. 5, 2022.
27. Siddiqui, Md Ismail Hossain, et al. "Explainable Federated Deep Learning for Low-Cost and Privacy-Preserving Early Breast Cancer Screening to Reduce US Healthcare Burden." Vascular and Endovascular Review 6.2 (2023): 45-54.
28. A. Sohel, M. A. Alam, A. Hossain, S. Mahmud, and S. Akter, "Artificial Intelligence in Predictive Analytics for Next-Generation Cancer Treatment: A Systematic Literature Review of Healthcare Innovations in the USA," Global Mainstream Journal of Innovation, Engineering & Emerging Technology, vol. 1, no. 1, pp. 62–87, 2022.