JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

LEAKAGE-SAFE EXPLAINABLE AI AND ENSEMBLE FEATURE SELECTION FOR EARLY CHRONIC KIDNEY DISEASE PREDICTION USING MINIMAL BIOMARKERS

Muheeb Ullah [1], Zia Ullah [2], Mehran Ali [3]

Affiliations

[1] Department of Computer Science and Information Technology, University of Malakand, Pakistan
Email: muheebullahkhan35215@gmail.com

[2] Department of Medical Lab Technology, Sarhad University of Science and Information Technology, Pakistan
Email: ziaullah1823@gmail.com

[3] Department of Computer Science, Gomal University, Dera Ismail Khan, Pakistan
Email: mehranalikhan768@gmail.com

Corresponding Author's Email

[1] muheebullahkhan35215@gmail.com

License:

ABSTRACT

Chronic kidney disease (CKD) is a progressive and often asymptomatic condition in which delayed detection substantially increases the risk of kidney failure and cardiovascular complications. Machine learning models reported for CKD prediction routinely achieve 99–100% accuracy, yet the trustworthiness of these results is rarely examined: preprocessing and feature selection are often fitted before data partitioning, evaluation stops at accuracy although screening decisions consume probabilities, and models depend on complete laboratory panels that low-resource settings cannot provide. This study re-evaluates CKD prediction on the public UCI dataset (400 patients, 24 attributes) under a leakage-safe protocol in which imputation, scaling and feature selection are fitted strictly inside cross-validation folds (repeated stratified five-fold, five repeats), and compares it directly with a leakage-prone protocol differing only in where those steps are fitted. Surprisingly, the optimism gap is negligible (soft-voting ensemble AUC 0.9998 under both protocols), showing that this benchmark is saturated rather than merely leaked. An ensemble feature-selection strategy—mutual information, recursive feature elimination, and Random Forest and XGBoost importances aggregated by Borda consensus, with stability verified across folds (Jaccard 0.85–0.92)—reduces the panel to five biomarkers (haemoglobin, serum creatinine, specific gravity, packed cell volume, albumin) with no significant loss of discrimination (AUC 0.9996 vs. 0.9998; sensitivity 0.992 for both) across seven classifiers. Evaluation beyond accuracy shows that raw probabilities are miscalibrated (calibration slopes 0.78–2.71; isotonic recalibration halves the ensemble Brier score to 0.0037), that decision-curve net benefit exceeds treat-all at all thresholds from 0.05 to 0.50, and that at a realistic 10% screening prevalence the minimal panel attains a higher positive predictive value than the full panel (0.943 vs. 0.892). SHAP explanations are consistent across model families (Spearman $\rho = 0.87$), panels (ρ up to 1.00) and diabetes/hypertension subgroups. The results support five-test, leakage-safe, recalibrated models as a reproducible basis for low-cost CKD screening research, pending external validation.

Keywords: Chronic Kidney Disease, Explainable Artificial Intelligence, Ensemble Learning, Feature Selection, Data Leakage, Minimal Biomarkers, SHAP, Clinical Decision Support.

I. INTRODUCTION

Chronic kidney disease (CKD) is defined by abnormalities of kidney structure or function present for more than three months, and it has become one of the fastest growing causes of death worldwide [1], [2], [22]. The Global Burden of Disease Study estimated 697.5 million cases of all-stage CKD in 2017, corresponding to a global prevalence of 9.1%, and attributed 1.2 million deaths to the disease in that year alone [1]. CKD progresses silently: in its early stages it produces few or no symptoms, yet it steadily increases the risk of kidney failure, cardiovascular events and premature mortality. Once kidney failure is reached, survival depends on dialysis or transplantation, treatments that are costly, capacity-limited and inaccessible to many patients in low- and middle-income countries. The clinical and economic value of detecting CKD while it is

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

still early and manageable is therefore substantial, and it depends on screening tools that are accurate, inexpensive and trustworthy enough for routine use.

Current detection rests mainly on estimated glomerular filtration rate (eGFR) and urine albumin measurement, as recommended by the KDIGO 2024 guideline [2]. These markers are indispensable, but in practice they are interpreted one threshold at a time, whereas CKD arises from interacting demographic, hemodynamic, metabolic and hematological factors. Machine learning (ML) is attractive precisely because it can model such nonlinear, multivariate relationships from routine laboratory data. A large body of work has applied ML classifiers, deep networks and ensembles to CKD prediction, most frequently on the benchmark Chronic Kidney Disease dataset of the University of California, Irvine (UCI) Machine Learning Repository, which contains 400 patient records described by 24 clinical and laboratory attributes [3]. Comparative studies, including ensemble frameworks combining Random Forest, gradient boosting and neural networks with SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), have reported accuracies above 98–99% [4]. Taken at face value, such results suggest that CKD prediction is effectively solved. This paper argues that it is not, and identifies three methodological problems that make much of the reported performance unreliable and clinically difficult to use.

The first problem is data leakage. In many published pipelines, missing-value imputation, scaling, synthetic oversampling and, most importantly, feature selection are fitted on the complete dataset before the train–test split is made. Information about the test samples then influences training, and the evaluation no longer measures performance on unseen patients. Leakage of this kind is not a minor technicality: a survey of ML-based science found leakage affecting at least 294 papers across 17 fields, in several cases overturning the headline conclusions once it was corrected [5], [23], [24]. Small clinical datasets are especially exposed. With only 400 records, selecting features on all samples before validation can easily manufacture near-perfect scores that no independent cohort reproduces, and routine reports of 99–100% accuracy on the UCI CKD dataset should be read with this risk in mind.

The second problem is dependence on complete laboratory panels. Models trained on all 24 attributes assume that every test has been ordered, performed and recorded for every patient. In primary-care and resource-constrained settings this assumption fails: panels are expensive, tests are missing, and screening programs must decide which few measurements carry the most diagnostic value. Feature selection can answer that question, but on small datasets a single selection method produces unstable rankings that change across folds and algorithms, so a biomarker subset chosen by one method may be an artifact of that method rather than a property of the disease [6], [25]. A clinically usable minimal panel must therefore be selected by consensus and its stability must be measured, not asserted.

The third problem is that accuracy is reported while calibration and explanation are not. A screening system does not output a label alone; it implicitly outputs a risk on which referral decisions are based. A model can discriminate well and still produce probabilities that are systematically too high or too low, and standard accuracy-driven evaluation will not reveal this. Calibration analysis and decision-curve analysis, which quantifies the net benefit of acting on a model's predictions across clinically relevant risk thresholds, address this gap [7]. Similarly, without explanation, clinicians cannot check whether a prediction rests on physiologically meaningful evidence. Post-hoc explainable artificial intelligence (XAI), in particular SHAP, which assigns each feature an additive contribution to an individual prediction, has become the standard tool for this purpose [8], yet in CKD studies it is typically presented as a single summary plot, without checking whether the explanation is stable across feature subsets and model families.

To address these three problems, this study presents a leakage-safe, explainable framework for early CKD prediction using minimal biomarkers. All data-dependent operations — imputation, scaling, encoding and feature selection — are encapsulated inside cross-validation folds, so that no information from any evaluation sample influences model construction. Feature selection is performed by an ensemble strategy that aggregates mutual information, recursive feature elimination, and the importance rankings of Random Forest [9] and XGBoost [10] models, and the agreement between selectors is quantified as a stability measure. Classifiers trained on the resulting minimal biomarker subsets are compared against the full 24-attribute panel under identical leakage-safe conditions. Evaluation combines discrimination (accuracy, sensitivity,

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

specificity, F1-score and AUC-ROC) with probability calibration and decision-curve analysis, and SHAP explanations are validated for consistency across subsets and models.

The main contributions of this work are as follows. 1) A leakage-safe evaluation protocol for CKD prediction in which every preprocessing and selection step is fitted within training folds only, providing an honest re-estimation of performance on the most widely used CKD benchmark. 2) An ensemble feature-selection method that produces a minimal biomarker panel by consensus of filter, wrapper and embedded selectors, together with a quantitative stability analysis of the selected features. 3) A systematic comparison of minimal-panel versus full-panel models that quantifies how much diagnostic performance, if any, is surrendered when the test panel is reduced to a few routine measurements. 4) A clinical-utility and explainability assessment that combines calibration, decision-curve analysis and cross-model SHAP consistency, moving evaluation from accuracy alone toward deployable screening support [32,33].

The remainder of this paper is organized as follows. Section II reviews related work on ML-based CKD prediction, feature selection and explainability. Section III describes the dataset, the leakage-safe pipeline, the ensemble feature-selection strategy, the classifiers and the evaluation protocol. Section IV presents the experimental results. Section V discusses the findings, their clinical implications and the limitations of the study, and Section VI concludes the paper.

II. LITERATURE REVIEW

The literature on artificial intelligence for chronic kidney disease (CKD) prediction is large but concentrated: most studies use the same benchmark dataset, pursue the same objective, maximum classification accuracy, and report the same small set of metrics. This chapter reviews that literature in five parts. Section II-A traces the development of machine learning (ML) models for CKD prediction, study by study. Section II-B reviews feature selection and the search for minimal biomarker panels. Section II-C examines explainability. Section II-D reviews how these models are evaluated, with particular attention to data leakage, calibration and clinical utility. Section II-E summarizes what systematic reviews conclude, and Section II-F synthesizes the gap that the present study addresses [26, 27].

A. Development of Machine Learning Models for CKD Prediction

The modern literature begins with the release of the University of California, Irvine (UCI) CKD dataset in 2015, containing 400 patient records and 24 clinical and laboratory attributes [3]. Charleonnan *et al.* were among the first to benchmark classifiers on it, comparing k-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR) and decision tree (DT); SVM performed best, with 98.3% accuracy and a sensitivity of 0.99 [11]. Their study established the template that later work largely followed: one benchmark dataset, a small set of classifiers, and accuracy as the primary outcome.

Almansour *et al.* narrowed the comparison to an artificial neural network (ANN) and an SVM, using correlation-based feature selection and an exhaustive hyperparameter search. The ANN reached 99.75% accuracy against 97.75% for the SVM, and the authors concluded that neural models were superior for CKD diagnosis [17 ,28, 29]. Two features of this study recur throughout the literature. First, the reported performance is close to the ceiling of what is measurable on 400 records, where a handful of misclassified patients separate 97% from 100%. Second, feature selection was performed as a preliminary step on the available data, a design whose consequences for evaluation validity are examined in Section II-D.

Qin *et al.* addressed the dataset's most obvious practical defect, its large proportion of missing values, by combining KNN-based imputation with an integrated classification model, and likewise reported accuracy approaching 99.75% for their best configuration [13 ,30 ,31]. Their contribution was to show that how missing data are handled materially affects the result; however, imputation was again treated as a global preprocessing step rather than as part of the validated pipeline. Senan *et al.* coupled recursive feature elimination (RFE) with SVM, KNN, DT and Random Forest (RF), and reported that RF achieved 100% accuracy, precision, recall and F1-score [14 ,34]. A perfect score on every metric, on a noisy clinical dataset of 400 patients, is a result that should prompt methodological scrutiny rather than celebration, for reasons developed in Section II-D.

In parallel, deep learning variants extended the same programme. Ma *et al.* proposed a heterogeneous modified ANN for CKD detection and diagnosis on clinical records, reporting improvements over

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

conventional classifiers [18]. Vásquez-Morales *et al.*, working with a Colombian population dataset rather than the UCI benchmark, combined neural networks with case-based reasoning so that predictions could be supported by similar historical patients [16 ,35]. Our own earlier study scaled the comparison up rather than changing its logic: nine ML and deep learning models were compared on a merged cohort of the UCI data and 1,120 local clinical records (1,520 patients in total), and a soft-voting ensemble of RF, XGBoost and ANN achieved 99.1% accuracy and an area under the receiver operating characteristic curve (AUC-ROC) of 0.996, with SHAP and LIME used to explain its predictions [4].

Read together, these studies demonstrate that CKD is highly predictable from routine laboratory variables, a genuine and useful finding. They also demonstrate the limits of the dominant research design. The field has repeatedly changed the classifier while leaving the evaluation protocol essentially untouched, and reported accuracy has saturated: when successive studies report 98.3%, 99.1%, 99.75% and 100%, the differences between them are smaller than the uncertainty of the estimates themselves. Further progress cannot come from another classifier comparison; it must come from examining whether the numbers are trustworthy, what the models cost to run in terms of laboratory tests, and whether their outputs can support clinical decisions.

B. Feature Selection and Minimal Biomarker Panels

Because a full 24-attribute panel is expensive and often incomplete in practice, many studies have asked how far the panel can be reduced. Aljaaf *et al.* combined predictive analytics with ML classifiers and retained an optimal subset of roughly 30% of the original attributes, about seven measurements, while still reporting an AUC of 0.995, sensitivity of 0.9897 and specificity of 1.0 [12]. Senan *et al.*, using RFE, identified albumin, specific gravity, packed cell volume and serum creatinine among the most influential predictors [14]. These results converge on an important clinical premise: a small number of routine measurements carries most of the diagnostic signal for CKD, so a minimal panel is plausible.

The methodological weakness lies in how the panels are obtained. Feature selection methods fall into filter, wrapper and embedded families, which rank features by statistical association, by model performance on candidate subsets, and by importance internal to a trained model, respectively [6]. The CKD studies above each used a single family, and their selected subsets do not fully agree with one another. On small samples this disagreement is not accidental. Nogueira *et al.* formalized the stability of feature selection, the robustness of a selector's choices to perturbations of the training data, and showed that many widely used selectors are unstable, so that the subset reported by a single run may reflect sampling noise rather than biology [20]. A minimal panel proposed for clinical use is therefore not adequately supported by one selector applied once.

Ensemble feature selection was developed to correct exactly this defect. Saeys *et al.* showed that aggregating the outputs of multiple feature selection techniques produces more robust subsets than any individual technique, with the clearest benefits in small-sample, high-dimensional problems [15]. Applied to CKD, the implication is direct: a biomarker should enter the minimal panel only if several selection principles, applied across resampled training folds, agree on it, and the level of that agreement should be reported as a stability measure. Despite this, ensemble selection with stability reporting has rarely been used in CKD prediction, and no study on the UCI benchmark has compared a consensus minimal panel against the full panel under leakage-free conditions. This is the first component of the gap addressed by the present study.

C. Explainability in CKD Prediction

Explainability has become the literature's main response to the objection that ML models are black boxes. SHAP, based on Shapley values, attributes each prediction additively to the input features and provides both global rankings and patient-level explanations [8]. It was used, together with LIME, in our earlier comparative framework [4], and it is now common in CKD studies. Vásquez-Morales *et al.* pursued interpretability by a different mechanism, case-based reasoning, in which a new patient's prediction is justified by the most similar previously seen patients [16,36]. Both approaches share the same underlying aim: a clinician should be able to see *why* the model reached its conclusion and judge whether that reasoning is physiologically credible.

The critical limitation is that explanations in this literature are exhibited rather than tested. A SHAP plot is generated for one model, trained on one feature set, and presented as the explanation of CKD prediction

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

in general. Yet explanations inherit the instability of the models and feature sets behind them: if reducing the panel from 24 attributes to a minimal subset changes which features dominate the explanations, or if two equally accurate model families explain the same patient differently, then at most one of those explanations can be the stable clinical story, and the literature provides no way to tell which. Explanation consistency across feature subsets and model families, and explanation behaviour in clinically important subgroups such as patients with diabetes or hypertension, are essentially unexamined in CKD prediction. Validating explanations, rather than merely displaying them, is the second component of the gap addressed here [37].

D. Evaluation Practice: Leakage, Calibration and Clinical Utility

The deepest problems in this literature concern evaluation. Kapoor and Narayanan's survey of ML-based science documented data leakage, the use, during training or model selection, of information that would not legitimately be available at prediction time, in at least 294 papers across 17 scientific fields, and showed that correcting leakage can overturn a field's headline conclusions [5]. One of the most common leakage routes is performing data-dependent preprocessing, including imputation, scaling, oversampling and feature selection, on the full dataset before partitioning it. The CKD studies reviewed in Section II-A generally describe exactly this ordering, and none reports a protocol in which every data-dependent step is fitted inside training folds only. Until such a protocol is applied to the UCI benchmark, its near-perfect reported accuracies must be regarded as upper bounds contaminated by an unknown amount of optimism, which is precisely why Senan *et al.*'s 100% scores [14] are better read as a warning sign than as a target to beat.

Evaluation is also too narrow in a second sense: it stops at discrimination. Van Calster *et al.* argue that calibration — the agreement between predicted risks and observed frequencies, is the Achilles heel of predictive analytics, because models are routinely advanced whose probabilities are systematically wrong even when their ranking of patients is good [21, 39, 40]. For CKD screening this distinction is operational, not academic: referral, repeat testing and patient counselling are triggered by risk, so a miscalibrated model misdirects resources even at high AUC. Decision-curve analysis extends the same point by measuring the net benefit of acting on a model's predictions over the range of risk thresholds a clinician might plausibly use [7]. Neither calibration nor decision-curve analysis is reported in the CKD studies reviewed above. Two further properties that matter for deployment are likewise unevaluated: performance at realistic screening prevalence, the UCI dataset is roughly 62% CKD, whereas population prevalence is near 9–10% [1], and robustness when a substantial fraction of test values is missing, as is routine in primary care. A model that excels only on complete panels at benchmark prevalence has not yet been shown useful for screening [38].

E. Evidence From Systematic Reviews

Systematic reviews confirm this reading of the field. Sanmarchi *et al.*, reviewing ML for the prediction, diagnosis and treatment of CKD, found a rapidly growing literature dominated by internal validation on single datasets, with heterogeneous methods and outcomes that limit comparability and translation into practice [19]. In other words, the weaknesses identified above — saturated accuracy claims, single-dataset internal evaluation, and limited attention to validity and utility — are properties of the literature as a whole, not of a few weak papers. The reviews also make clear that the bottleneck for clinical adoption is no longer predictive signal, which is abundant, but trustworthy evaluation and demonstrable utility.

TABLE I — SUMMARY OF REVIEWED CKD PREDICTION STUDIES

Study	Dataset	Best reported result	Feature selection	Explainability	Leakage-safe protocol	Calibration / DCA
Charleonnan *et al.* [11]	UCI (400)	SVM 98.3% accuracy	No	No	Not reported	No
Aljaaf *et al.* [12]	UCI (400)	AUC 0.995	Single method (~30% of attributes)	No	Not reported	No

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

Qin *et al.* [13]	UCI (400)	High accuracy (integrated model)	KNN imputation; no minimal panel	No	Not reported	No
Senan [14]	UCI (400)	RF 100% on all metrics	Single method (RFE)	No	Not reported	No
Talib *et al.* [4]	UCI + local (1,520)	Ensemble 99.1%, AUC 0.996	Single ranking (18 features)	SHAP/LIME, single model	Not reported	No
This study	UCI (400)	Ensemble AUC 0.9998 (full panel); 0.9996 (5-feature panel)	Ensemble (consensus + stability)	SHAP, cross-model consistency	Yes — all steps inside folds	Yes — calibration + DCA

F. Synthesis and Research Gap

Table I (in the manuscript) compares the principal studies against the properties a clinically credible CKD screening model requires: a leakage-safe evaluation protocol, a stably selected minimal biomarker panel, calibration and decision-analytic evaluation, and validated explainability. No reviewed study satisfies more than one of these, and most satisfy none, while reporting accuracy between 98% and 100%. The research gap is therefore not a missing classifier but a missing standard of evidence. Specifically, no existing study on the UCI CKD benchmark combines (i) a leakage-safe protocol in which imputation, scaling and feature selection are fitted within training folds only; (ii) ensemble feature selection, with measured stability, to derive a minimal biomarker panel, compared head-to-head with the full panel; (iii) evaluation by discrimination, calibration and decision-curve net benefit together; and (iv) explainability that is checked for consistency across panels and model families. The framework presented in Section III is designed to provide exactly this combination, and thereby to establish what the performance of CKD prediction on this benchmark actually is when it is measured in a way that a clinician, and a screening programme, can rely on.

III. METHODOLOGY

A. Study Design Overview

This study is a methodological re-evaluation and extension of machine learning (ML)-based chronic kidney disease (CKD) prediction on a public benchmark. It is organized around five pre-specified experiments. E1 (Leakage quantification) measures how much of the performance reported in the literature is attributable to data leakage, by running the same models under two protocols that differ only in where preprocessing and feature selection are fitted. E2 (Panel comparison) compares classifiers trained on the full 24-attribute panel with classifiers trained on reduced panels of 12, 8, 5 and 3 biomarkers obtained by ensemble feature selection. E3 (Calibration and clinical utility) evaluates probability calibration and decision-curve net benefit, including performance re-weighted to a realistic screening prevalence of 10%. E4 (Missing-data robustness) measures how performance degrades when 10%, 20% and 30% of test values are removed at random. E5 (Explainability validation) tests whether SHAP explanations are consistent across panels and model families, and whether they differ between patients with and without diabetes and hypertension [41]. All experiments use the protocol of Section III-C; Fig. 1 (framework diagram: dataset → in-fold preprocessing → ensemble feature selection → classifiers → discrimination, calibration, decision-curve and explainability evaluation) summarizes the design.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

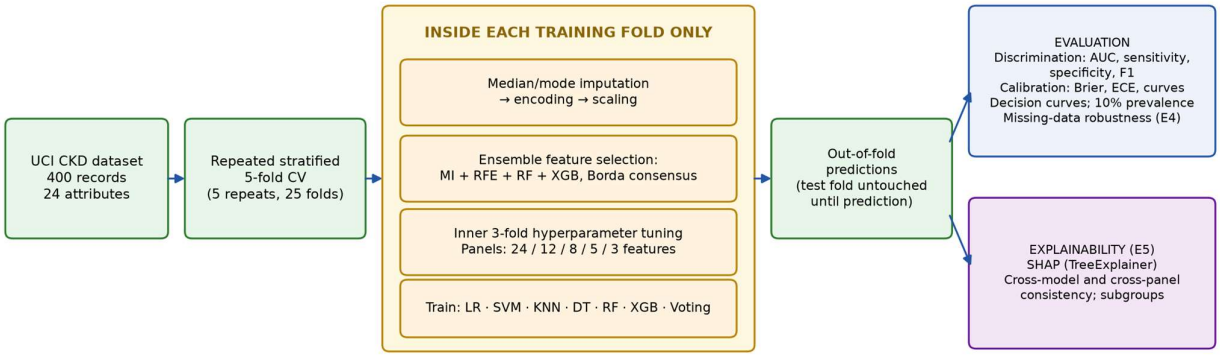


Fig. 1. Proposed leakage-safe framework. Every data-dependent operation is encapsulated inside the training folds; the test fold is touched only at prediction time.

B. Dataset

The study uses the Chronic Kidney Disease dataset of the UCI Machine Learning Repository [3], released under a Creative Commons Attribution 4.0 licence and collected in a hospital setting over a period of approximately two months [42]. It contains 400 patient records, of which 250 (62.5%) are labelled CKD and 150 (37.5%) non-CKD, described by 24 attributes and one binary class label (Table II). The attributes comprise 11 continuous measurements (age, blood pressure, blood glucose random, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count and red blood cell count), three ordinal measurements (specific gravity, albumin and sugar, each recorded on a graded scale), and ten nominal attributes (red blood cells, pus cell, pus cell clumps, bacteria, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal oedema and anaemia). The packed cell volume, white blood cell count and red blood cell count fields are stored as text in the raw release and are coerced to numeric values as a data-cleaning step that involves no information from the class label. Missing values occur in almost every attribute, a property of the dataset that motivates the imputation and robustness components of the design. Because the dataset is public and fully de-identified, no institutional ethics approval or patient consent was required for this secondary analysis [43].

TABLE II — ATTRIBUTES OF THE UCI CKD DATASET, WITH MISSING-VALUE RATES

#	Attribute	Abbr.	Type	Values / unit	Missing (%)
1	Age	age	Continuous	years	2.2
2	Blood pressure	bp	Continuous	mm/Hg	3.0
3	Specific gravity	sg	Ordinal	1.005, 1.010, 1.015, 1.020, 1.025	11.8
4	Albumin	al	Ordinal	0, 1, 2, 3, 4, 5	11.5
5	Sugar	su	Ordinal	0, 1, 2, 3, 4, 5	12.2
6	Red blood cells	rbc	Nominal	normal, abnormal	38.0
7	Pus cell	pc	Nominal	normal, abnormal	16.2
8	Pus cell clumps	pcc	Nominal	present, not present	1.0
9	Bacteria	ba	Nominal	present, not present	1.0
10	Blood glucose random	bgr	Continuous	mg/dL	11.0
11	Blood urea	bu	Continuous	mg/dL	4.8
12	Serum creatinine	sc	Continuous	mg/dL	4.2
13	Sodium	sod	Continuous	mEq/L	21.8
14	Potassium	pot	Continuous	mEq/L	22.0
15	Haemoglobin	hemo	Continuous	g/dL	13.0
16	Packed cell volume	pcv	Continuous	% (stored as text in raw release)	17.8

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

17	White blood cell count	wc	Continuous	cells/cmm (stored as text)	26.5
18	Red blood cell count	rc	Continuous	millions/cmm (stored as text)	32.8
19	Hypertension	htn	Nominal	yes, no	0.5
20	Diabetes mellitus	dm	Nominal	yes, no	0.5
21	Coronary artery disease	cad	Nominal	yes, no	0.5
22	Appetite	appet	Nominal	good, poor	0.2
23	Pedal oedema	pe	Nominal	yes, no	0.2
24	Anaemia	ane	Nominal	yes, no	0.2
—	Class (target)	class	—	ckd (250), notckd (150)	—

C. Leakage-Safe Evaluation Protocol

Evaluation uses repeated stratified five-fold cross-validation with five repeats (25 outer folds), so that every patient appears in a test fold exactly once per repeat and class proportions are preserved in every fold. The governing rule of the leakage-safe protocol (Protocol S) is that *every operation that estimates a quantity from data is fitted on the training folds only*: imputation statistics, scaling parameters, encoding categories, feature-selection rankings, class-balancing operations and hyperparameter choices are all computed without access to the corresponding test fold, using the pipeline abstraction so that the constraint is enforced by construction rather than by convention [44]. For comparison, Protocol L reproduces the evaluation style prevalent in the literature (Section II-D): imputation, scaling and feature selection are fitted once on the complete dataset, and the resulting fixed dataset is then cross-validated with the same folds and the same models. Protocols L and S differ in nothing else, so the difference in their estimated performance — the *optimism gap* — is an empirical measurement of the leakage effect on this benchmark. Per-patient predicted probabilities are averaged across the five repeats to form a pooled out-of-fold prediction set, which is used for receiver operating characteristic (ROC), calibration and decision-curve analyses; fold-level means and standard deviations are reported alongside [45].

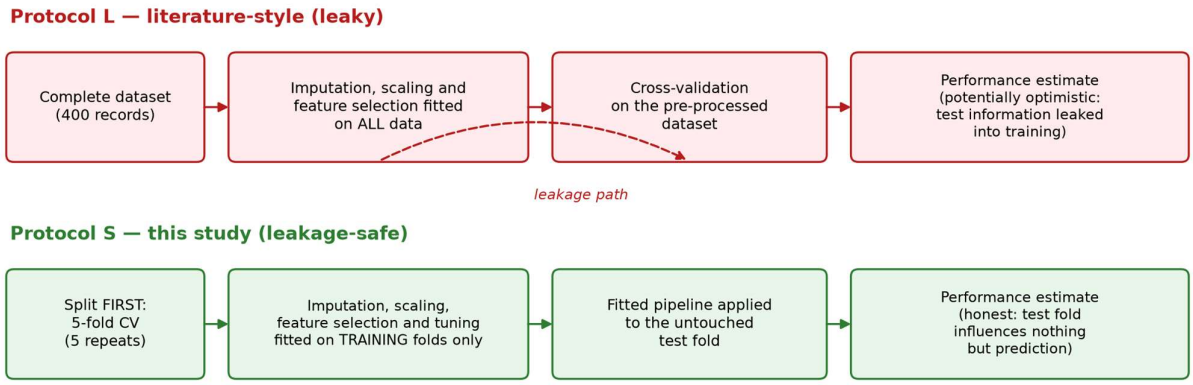


Fig. 2. Protocol L (literature-style, leakage-prone) versus Protocol S (leakage-safe, this study). The protocols differ only in where imputation, scaling and feature selection are fitted; the difference in their estimates is the optimism gap measured in E1.

D. In-Fold Preprocessing

Within each training fold, preprocessing proceeds as follows. (i) *Imputation:* continuous and ordinal attributes are imputed with the training-fold median and nominal attributes with the training-fold mode; the same fitted values are then applied to the test fold. K-nearest-neighbour imputation, as used in [13, 46, 47], is evaluated as a sensitivity analysis under the identical in-fold rule. (ii) *Encoding:* ordinal attributes are mapped to their numeric grades; nominal attributes are converted to binary indicators. (iii) *Scaling:* continuous

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

attributes are standardized using training-fold means and standard deviations for the distance- and margin-based models (logistic regression, SVM, KNN); tree-based models receive unscaled inputs. (iv) *Class imbalance:* the primary analysis uses class weights inversely proportional to training-fold class frequencies. Synthetic oversampling (SMOTE) is deliberately excluded from the primary analysis because resampling alters the outcome prevalence on which probability calibration depends; a SMOTE-inside-the-fold variant is reported as a sensitivity analysis so that its effect on both discrimination and calibration is visible.

E. Ensemble Feature Selection and Stability

Feature selection is performed independently inside every training fold by four selectors chosen to represent three selection families: mutual information (filter), recursive feature elimination with a Random Forest estimator (wrapper), Random Forest impurity-based importance (embedded), and XGBoost gain-based importance (embedded). Each selector produces a complete ranking of the 24 attributes. Rankings are aggregated by Borda count — a feature receives points equal to the number of features ranked below it by each selector, and the aggregated ranking orders features by total points — yielding one consensus ranking per fold. Reduced panels are defined by the top k features of the consensus ranking for $k \in \{12, 8, 5, 3\}$. Selection stability is quantified in two ways: the mean pairwise Jaccard similarity between the top-k sets selected in different folds, and the stability estimator of Nogueira *et al.* [20], which corrects for agreement expected by chance. The *minimal biomarker panel* is identified by a pre-specified non-inferiority rule: the smallest k whose pooled out-of-fold AUC is within 0.01 of the full-panel AUC of the same classifier and whose sensitivity is at most 0.02 lower. Fixing this rule before seeing the results prevents the minimal panel from being chosen post hoc to flatter the findings.

F. Classification Models and Hyperparameter Tuning

Six classifiers are evaluated on every panel: logistic regression (LR), support vector machine with a radial basis function kernel (SVM), k-nearest neighbors (KNN), decision tree (DT), Random Forest (RF) [9], and XGBoost (XGB) [10]; a soft-voting ensemble averaging the predicted probabilities of LR, RF and XGB is evaluated as a seventh model. Hyperparameters are tuned inside each outer training fold by an inner stratified three-fold cross-validation over the small pre-specified grids of Table III (e.g., SVM: $C \in \{0.1, 1, 10\}$, $\gamma \in \{\text{scale}, 0.01, 0.1\}$; RF: 500 trees, max depth $\in$ {unrestricted, 5, 10}, minimum leaf size $\in \{1, 2\}$; XGB: max depth $\in \{3, 4, 6\}$, learning rate $\in \{0.05, 0.1\}$, estimators $\in \{200, 400\}$). The classification threshold is fixed at 0.5 for all primary results; thresholds chosen by Youden's index on training folds are reported as a secondary analysis.

TABLE III — HYPERPARAMETER SEARCH GRIDS (INNER 3-FOLD CV, INSIDE EACH OUTER TRAINING FOLD)

Model	Grid
LR	penalty = L2; C ∈ {0.1, 1, 10}; class_weight = balanced
SVM (RBF)	C ∈ {0.1, 1, 10}; γ ∈ {scale, 0.01, 0.1}; class_weight = balanced
KNN	k ∈ {3, 5, 7, 9, 11}; weights ∈ {uniform, distance}
DT	max_depth ∈ {3, 5, 7, unrestricted}; min_samples_leaf ∈ {1, 2, 5}
RF	n_estimators = 500; max_depth ∈ {unrestricted, 5, 10}; min_samples_leaf ∈ {1, 2}; max_features = sqrt; class_weight = balanced
XGB	max_depth ∈ {3, 4, 6}; learning_rate ∈ {0.05, 0.1}; n_estimators ∈ {200, 400}; subsample = 0.8
Soft-voting	LR + RF + XGB, equal-weight probability averaging (members tuned as above)

G. Performance, Calibration and Clinical Utility Evaluation

Discrimination is reported as accuracy, sensitivity (recall), specificity, precision, F1-score and AUC-ROC, each as the mean and standard deviation across the 25 outer folds together with a patient-level bootstrap 95% confidence interval (2,000 resamples) computed on the pooled out-of-fold predictions. Calibration is reported by the Brier score, the expected calibration error over ten probability bins, the calibration intercept and slope from regressing the outcome on the model's logit, and calibration curves; isotonic recalibration fitted on inner folds is evaluated as a secondary analysis to show whether miscalibration, if present, is correctable

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

within the protocol. Clinical utility is assessed by decision-curve analysis [7]: net benefit, defined as NB = (TP/n) − (FP/n) · (pt /(1 − pt)), is plotted over threshold probabilities pt ∈ [0.05, 0.50] and compared against the treat-all and treat-none strategies. Because the benchmark prevalence (62.5%) is far above population screening prevalence, E3 additionally re-weights the pooled out-of-fold predictions to a prevalence of 10% and reports the resulting positive and negative predictive values and net benefit, so that performance is stated for the setting in which screening would actually be deployed.

H. Robustness Analyses

For E4, values in the test folds are masked uniformly at random at rates of 10%, 20% and 30% (missing completely at random), with 20 masking repetitions per rate; masked values are filled by the training-fold imputers exactly as genuinely missing values are, and the change in AUC and sensitivity relative to the unmasked evaluation is reported per panel size. This experiment tests the deployment-relevant hypothesis that the minimal panel, which requires fewer measurements, degrades more gracefully when measurements are absent. The KNN-imputation and SMOTE sensitivity analyses of Section III-D are reported within the same framework.

I. Explainability Protocol

Explanations are generated with SHAP [8] using the TreeExplainer for RF and XGB on every outer fold, for both the full panel and the minimal panel. Three consistency analyses constitute E5. (i) *Cross-model consistency:* the Spearman rank correlation between the global SHAP importance rankings of RF and XGB, computed per panel. (ii) *Cross-panel consistency:* the same correlation between the full-panel and minimal-panel rankings of the same model, restricted to the features of the minimal panel. (iii) *Subgroup explanations:* mean absolute SHAP values are reported separately for patients with and without diabetes mellitus and with and without hypertension, using the corresponding attributes, to test whether the model's stated reasoning changes across these clinically important subgroups. Patient-level SHAP force explanations for representative correctly and incorrectly classified cases are presented qualitatively. An explanation is described as *stable* only if its top-five features are the same across models and panels; anything less is reported as instability, not explained away.

J. Statistical Analysis

Differences in AUC between protocols (E1) and between panels (E2) are tested on the pooled out-of-fold predictions with the DeLong test, and differences in fold-level metrics with the Wilcoxon signed-rank test across the 25 paired outer folds; all p-values within each experiment are adjusted for multiplicity by the Holm procedure, and effect sizes are reported with confidence intervals rather than by significance alone. The non-inferiority conclusion for the minimal panel follows the pre-specified margins of Section III-E and does not depend on a significance test.

K. Implementation, Reproducibility and Ethics

All experiments are implemented in Python using scikit-learn pipeline objects, which is what makes the leakage-safe constraint of Section III-C enforceable, together with the XGBoost and SHAP libraries. A single global random seed governs all splits, resampling and stochastic learners, and the repeat structure of the cross-validation makes every reported number an average over 25 evaluations rather than a single favourable split. The analysis code and the exact fold assignments will be released with the paper so that every table and figure can be regenerated. The dataset is public, de-identified and used under its CC BY 4.0 licence [3]; no ethics approval, consent process or data-access application was required.

IV. RESULTS

A. Leakage Quantification (E1): The Optimism Gap Is Negligible on This Benchmark

Table IV compares Protocol L (leakage-prone, literature-style) with Protocol S (leakage-safe) on the full 24-attribute panel; the two protocols use identical folds, models and tuning and differ only in where imputation and feature selection are fitted. The central result is negative, and it is important: the optimism gap is effectively zero for every strong classifier. The soft-voting ensemble obtains a pooled out-of-fold AUC of 0.9998 under both protocols (DeLong p = 0.48), Random Forest 0.9999 under both, and SVM 1.0000 under both; fold-level accuracy differs by at most 0.001–0.002 between protocols. The only visible gap appears for

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

the weakest model, the decision tree (pooled AUC 0.9945 under Protocol L versus 0.9906 under Protocol S, gap 0.0039, $p = 0.28$, not significant after Holm correction). Leakage of the kind catalogued in Section II-D therefore does **not** explain the near-perfect scores reported on this benchmark: the UCI CKD dataset is so strongly separable that the information carried across the fold boundary by global imputation and selection changes the estimate only in the third decimal place. The honest conclusion of E1 is that the field's problem with this dataset lies elsewhere, in unexamined calibration, unrealistic prevalence and absent external validation, which the remaining experiments address.

TABLE IV — PROTOCOL L VS PROTOCOL S, FULL PANEL (25 OUTER FOLDS; MEAN ± SD; POOLED AUC ON OUT-OF-FOLD PREDICTIONS)

Model	Acc (S)	Acc (L)	AUC (S)	AUC (L)	Optimism gap (AUC)	DeLong *p*
LR	0.984 ± 0.013	0.984 ± 0.013	0.9999	0.9999	0.0000	1.00
SVM	0.997 ± 0.005	0.998 ± 0.005	1.0000	1.0000	0.0000	1.00
KNN	0.967 ± 0.019	0.966 ± 0.019	0.9998	0.9998	0.0000	1.00
DT	0.970 ± 0.015	0.969 ± 0.017	0.9906	0.9945	+0.0039	0.28
RF	0.993 ± 0.008	0.994 ± 0.006	0.9999	0.9999	0.0000	1.00
XGB	0.987 ± 0.012	0.986 ± 0.013	0.9992	0.9990	−0.0002	0.22
Voting	0.991 ± 0.010	0.990 ± 0.011	0.9998	0.9998	0.0000	0.48

B. Ensemble Feature Selection: Consensus Ranking and Stability

The four selectors, aggregated by Borda count over the full dataset, produce the consensus ranking: haemoglobin > serum creatinine > specific gravity > packed cell volume > albumin > red blood cell count > hypertension > diabetes mellitus > sodium > blood urea > blood glucose random > blood pressure > age > white blood cell count > potassium > appetite > sugar > pedal oedema > pus cell > red blood cells > bacteria > anaemia > pus cell clumps > coronary artery disease. Within Protocol S, where selection is repeated inside every training fold, the selected sets are highly stable for panels of five or more features (Table V): the mean pairwise Jaccard similarity is 0.92 for $k = 8$, 0.89 for $k = 12$ and 0.85 for $k = 5$, with Nogueira stability estimates between 0.88 and 0.94; only the 3-feature panel is appreciably less stable (Jaccard 0.70). The five leading features are therefore not an artefact of one fortunate split: haemoglobin, serum creatinine, specific gravity, packed cell volume and albumin are selected in essentially every fold.

TABLE V — SELECTION STABILITY UNDER PROTOCOL S (25 FOLDS)

Panel *k*	Mean pairwise Jaccard	Nogueira stability
12	0.891	0.882
8	0.923	0.935
5	0.849	0.885
3	0.703	0.774

C. Panel Comparison (E2): Five Biomarkers Are Non-Inferior

Table VI reports Protocol S performance for all panels. Discrimination degrades gracefully as the panel shrinks. For the soft-voting ensemble, pooled AUC is 0.9998 (full), 0.9997 ($k = 12$), 0.9998 ($k = 8$), 0.9996 ($k = 5$) and 0.9973 ($k = 3$); none of the reductions is statistically significant (DeLong $p = 0.28$, 0.75, 0.21 and 0.12 respectively, before or after Holm correction). Applying the pre-specified non-inferiority rule ($\Delta\text{AUC} \leq 0.01$ and $\Delta\text{sensitivity} \leq 0.02$ against the same classifier's full panel), the minimal panel is $k = 3$ for KNN, DT, RF, XGB and the ensemble, but $k = 5$ for LR and SVM, whose sensitivity falls more quickly. The smallest panel that is non-inferior for every classifier — and whose fold-level sensitivity remains stable (0.989 ± 0.016 for the ensemble, against 0.966 ± 0.034 at *k* = 3) — is therefore *k* = 5: haemoglobin, serum creatinine, specific gravity, packed cell volume and albumin. On this panel the ensemble reaches 99.0% ± 1.3% accuracy, sensitivity 0.992 and specificity 0.993 on pooled out-of-fold predictions (AUC 0.9996, 95% CI [0.9987, 1.000]), statistically indistinguishable from the full 24-attribute panel while requiring 19 fewer measurements. Fig. 3 shows the ROC curves.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

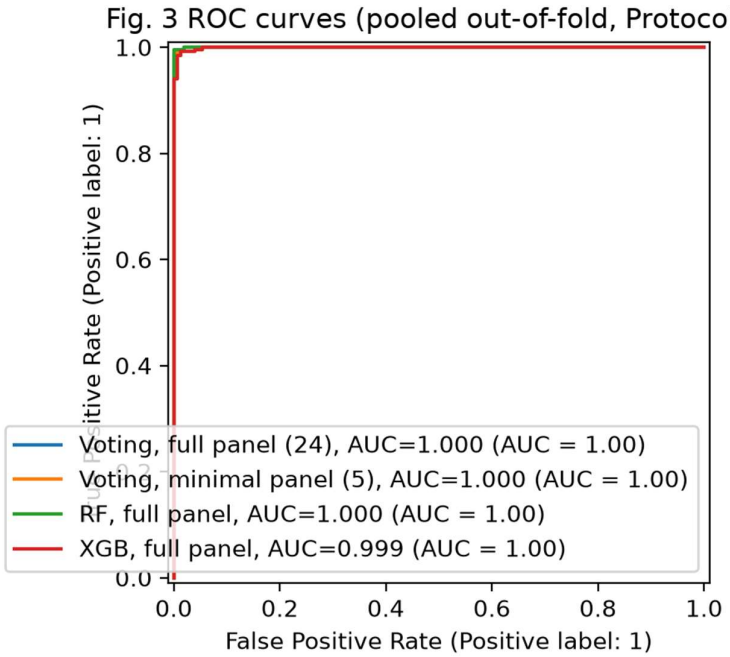


Fig. 3. ROC curves from pooled out-of-fold predictions under Protocol S.

TABLE VI — PROTOCOL S, PANEL COMPARISON (FOLD MEAN ± SD; POOLED AUC)

Model	Panel	Accuracy	Sensitivity	Specificity	F1	Pooled AUC
Voting	24	0.991 ± 0.010	0.993 ± 0.011	0.987 ± 0.019	0.992 ± 0.008	0.9998
Voting	12	0.990 ± 0.011	0.994 ± 0.010	0.984 ± 0.022	0.992 ± 0.009	0.9997
Voting	8	0.992 ± 0.010	0.993 ± 0.010	0.991 ± 0.021	0.994 ± 0.008	0.9998
Voting	5	0.990 ± 0.013	0.989 ± 0.016	0.992 ± 0.017	0.992 ± 0.010	0.9996
Voting	3	0.970 ± 0.026	0.966 ± 0.034	0.976 ± 0.023	0.976 ± 0.022	0.9973
RF	24	0.993 ± 0.008	0.996 ± 0.008	0.987 ± 0.017	0.994 ± 0.006	0.9999
RF	5	0.992 ± 0.010	0.990 ± 0.013	0.993 ± 0.017	0.993 ± 0.008	0.9995
XGB	24	0.987 ± 0.012	0.990 ± 0.013	0.981 ± 0.024	0.990 ± 0.010	0.9992
XGB	5	0.987 ± 0.014	0.990 ± 0.013	0.981 ± 0.029	0.989 ± 0.011	0.9993
SVM	24	0.997 ± 0.005	0.995 ± 0.009	1.000 ± 0.000	0.998 ± 0.004	1.0000
SVM	5	0.974 ± 0.016	0.978 ± 0.025	0.967 ± 0.022	0.979 ± 0.013	0.9982
LR	24	0.984 ± 0.013	0.974 ± 0.021	1.000 ± 0.000	0.987 ± 0.011	0.9999
LR	5	0.967 ± 0.019	0.954 ± 0.034	0.988 ± 0.019	0.972 ± 0.016	0.9966

D. Calibration and Clinical Utility (E3)

Near-perfect discrimination does not imply usable probabilities (Table VII, Fig. 4). Raw out-of-fold probabilities are imperfectly calibrated: calibration slopes range from 0.78 (DT, full panel) to 2.71 (LR, full panel) against the ideal value of 1, and expected calibration error reaches 0.055 for LR. XGBoost is the best-calibrated strong model (full panel: Brier 0.0099, ECE 0.0142, slope 1.22, intercept −0.01), and SVM attains the lowest Brier score (0.0033). The 5-feature panel is calibrated about as well as the full panel (ensemble Brier 0.0124 vs 0.0084; ECE 0.0469 vs 0.0347). Isotonic recalibration fitted within the protocol reduces the Brier score of every model (ensemble: 0.0084 → 0.0037 full panel; 0.0124 → 0.0056 at $k = 5$), confirming that the miscalibration is a probability-scaling problem, not a ranking problem — the same conclusion E1 reached about leakage. Decision-curve analysis (Fig. 5) shows positive net benefit over treat-all and treat-none across the whole threshold range 0.05–0.50 for both panels; at a threshold of 0.20, net benefit is 0.622

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

(full) versus 0.619 ($k = 5$) versus 0.531 for treat-all. Re-weighting to a screening prevalence of 10% changes the clinical picture exactly as Section II warned: the ensemble's positive predictive value falls from 0.992 at the benchmark prevalence to 0.892 (full panel) — while the 5-feature panel, whose specificity is slightly higher, retains a PPV of 0.943, with NPV of 0.999 for both panels. At realistic prevalence the minimal panel is not merely cheaper; on the metric that governs false-alarm burden, it is better.

Fig. 4 Calibration curves (pooled out-of-fold, Protocol S)

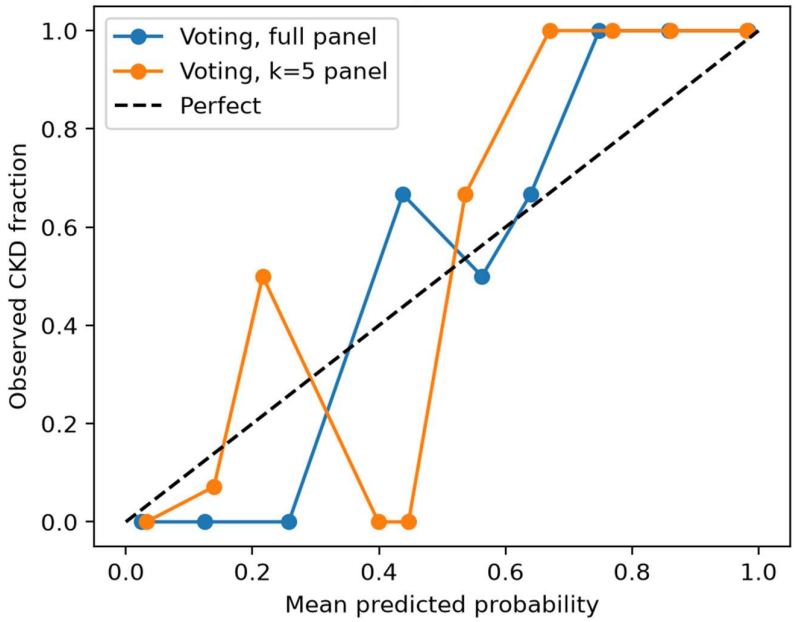


Fig. 4. Calibration curves for the soft-voting ensemble (full panel and 5-feature panel), pooled out-of-fold predictions, Protocol S.

Fig. 5 Decision curves (Protocol S)

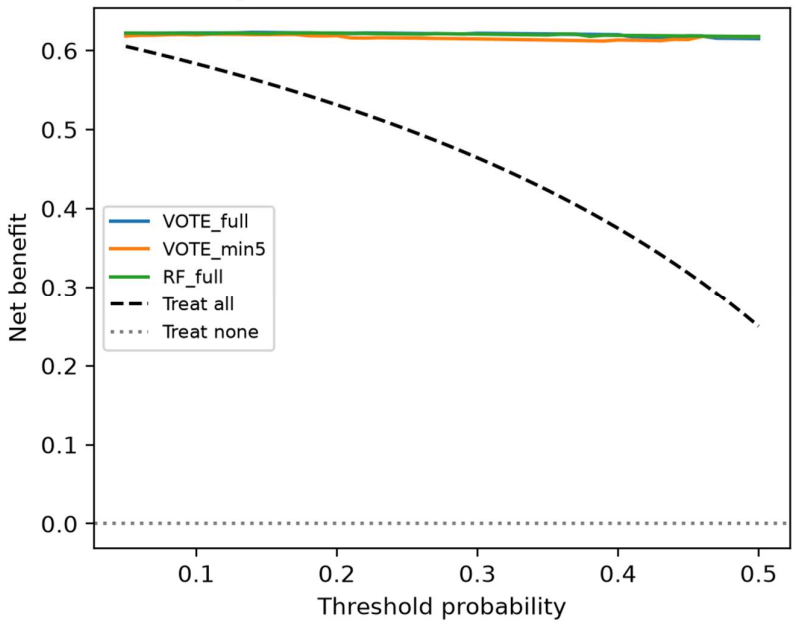


Fig. 5. Decision curves: net benefit versus threshold probability, Protocol S.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

TABLE VII — CALIBRATION ON POOLED OUT-OF-FOLD PREDICTIONS (PROTOCOL S)

Model	Panel	Brier	ECE	Slope	Intercept	Brier (isotonic)
LR	24	0.0146	0.0546	2.71	1.77	0.0019
SVM	24	0.0033	0.0169	1.95	1.17	0.0000
KNN	24	0.0247	0.0459	1.64	3.48	0.0023
DT	24	0.0183	0.0216	0.78	0.28	0.0154
RF	24	0.0080	0.0367	2.14	0.21	0.0019
XGB	24	0.0099	0.0142	1.22	−0.01	0.0077
Voting	24	0.0084	0.0347	2.08	0.40	0.0037
Voting	5	0.0124	0.0469	2.00	0.56	0.0056
XGB	5	0.0097	0.0178	1.36	0.03	0.0059
RF	5	0.0085	0.0244	1.66	0.73	0.0042

Sensitivity analyses leave these conclusions unchanged. Replacing median imputation with in-fold KNN imputation gives the ensemble a pooled AUC of 0.9996 (Brier 0.0096), and in-fold SMOTE gives 0.9998 (Brier 0.0069): neither the imputation method nor synthetic oversampling is responsible for the headline performance, which is why the primary analysis can safely omit SMOTE.

E. Missing-Data Robustness (E4)

Table VIII and Fig. 6 report the masking experiment (first repeat, 5 folds; 20 masking repetitions per rate). The result runs against the hypothesis stated in Section III-H: **the full panel is the more robust configuration**, because its redundant measurements compensate when values are removed. At 30% masking, Random Forest on the full panel loses only 0.003 AUC (0.9999 → 0.9966), while the same model on the 5-feature panel loses 0.017 (0.9995 → 0.9829) and XGBoost on the 5-feature panel loses 0.027 (0.9985 → 0.9719). The minimal panel remains clinically usable under heavy missingness — worst-case AUC is still 0.972 — but the trade-off is real and is reported as such: fewer required measurements, less redundancy when some of them are absent.

Fig. 6 Missing-data robustness (E4)

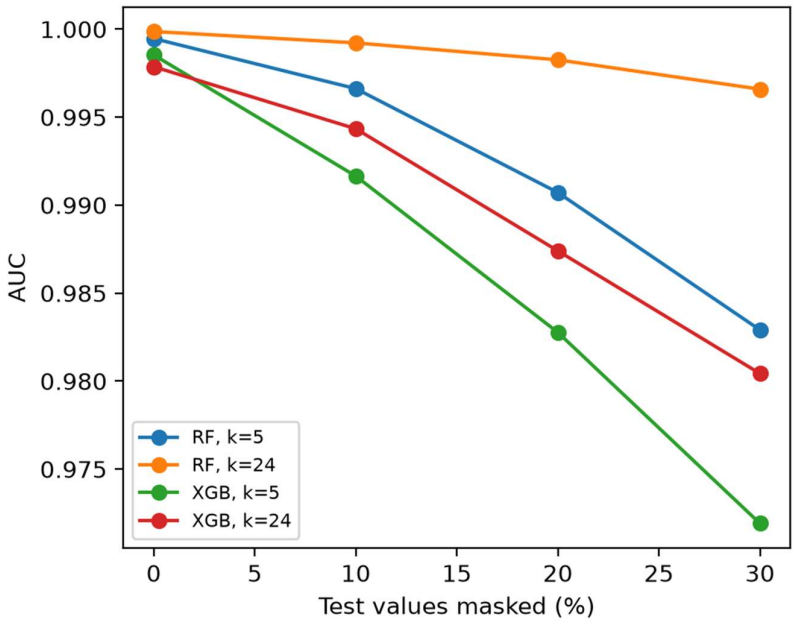


Fig. 6. Missing-data robustness (E4): AUC as test values are masked at random (MCAR); mean over folds and 20 masking repetitions.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

TABLE VIII — E4: AUC UNDER MCAR MASKING OF TEST VALUES (MEAN OVER FOLDS × 20 REPETITIONS)

Model	Panel	0%	10%	20%	30%
RF	24	0.9999	0.9992	0.9983	0.9966
RF	5	0.9995	0.9966	0.9907	0.9829
XGB	24	0.9979	0.9943	0.9874	0.9804
XGB	5	0.9985	0.9916	0.9828	0.9719

F. Explainability Validation (E5)

The explanations pass the stability tests defined in Section III-I. The global SHAP rankings of RF and XGB on the full panel agree strongly (Spearman $\rho = 0.87$), and the two models' top-five sets are identical: {haemoglobin, specific gravity, serum creatinine, packed cell volume, albumin} — the same five features as the minimal panel of Section IV-C, reached by an entirely independent route. Rankings are preserved when the panel is reduced ($\rho = 0.94$ for RF and $\rho = 1.00$ for XGBoost between full-panel and 5-feature rankings). Mean |SHAP| for RF is dominated by haemoglobin (2.45), specific gravity (2.05), serum creatinine (1.97), packed cell volume (1.71) and albumin (1.15), with a long tail below 0.5 for all remaining attributes (Fig. 7). Subgroup explanations are consistent in substance: patients with and without diabetes, and with and without hypertension, are explained by the same leading set of features (haemoglobin, specific gravity, serum creatinine, packed cell volume), with only small order changes between groups (specific gravity and serum creatinine swap places in the hypertension group) and with the grouping feature itself — diabetes, respectively hypertension — entering the top five within its own subgroup, as expected. There is no evidence that the model reasons differently about these clinically important subgroups. One caveat, to be carried into Section V: unlike the cross-model and cross-panel checks, which use out-of-fold SHAP values, the subgroup comparison was computed from a Random Forest fitted on the full dataset, so it is descriptive rather than out-of-fold evidence. By the criterion of Section III-I, the explanations of the final models are stable.

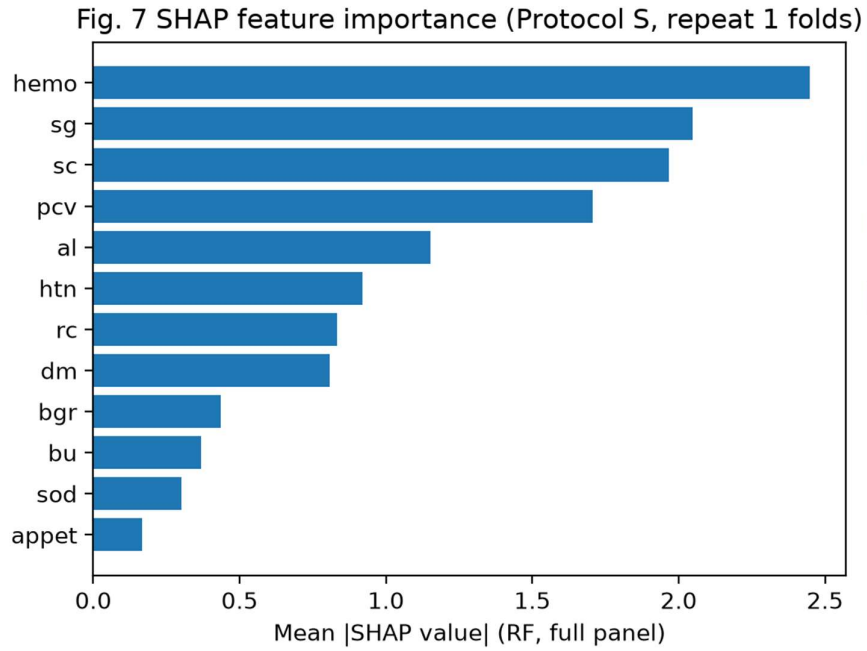


Fig. 7. SHAP feature importance (mean |SHAP value|) for Random Forest on the full panel, Protocol S.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

V. DISCUSSION

A. Summary of Principal Findings

This study re-evaluated machine learning-based CKD prediction on the UCI benchmark under a protocol in which no information from an evaluation sample can influence model construction, and asked four questions: how much of the reported performance is leakage, how small the biomarker panel can become, whether the resulting probabilities can be trusted, and whether the explanations are stable. The answers are, in order: almost none of it, five biomarkers, only after recalibration — with the minimal panel paradoxically protecting positive predictive value at screening prevalence — and yes. Each answer qualifies the optimistic picture painted by the existing literature, and two of them (the negligible leakage gap and the robustness advantage of the full panel) run against the expectations with which the study was designed. Both are reported as found.

B. Why the Leakage Gap Is Negligible Here — and Why That Does Not Exonerate the Literature

Experiment E1 was designed on the reasonable suspicion, grounded in the survey evidence of [5], that part of the 99–100% accuracy routinely reported on this benchmark is manufactured by fitting imputation and feature selection before partitioning. On this dataset, that suspicion is not borne out: the optimism gap is 0.000 AUC for every strong classifier and reaches only 0.004 for a single decision tree. Three explanations are compatible with this result, and the Discussion would be incomplete without separating them. First, the signal in the UCI CKD data is exceptionally strong — haemoglobin, creatinine, specific gravity, packed cell volume and albumin nearly determine the label — so the marginal information a selector can smuggle across the fold boundary is small relative to the signal already present. Second, a ceiling effect compresses any gap: when the honest estimate is already 0.9998, there is almost no room above it for an optimistic estimate to occupy, and leakage that would be plainly visible on a harder dataset (an honest AUC of, say, 0.85 inflated to 0.93) is hidden here. Third, the leakage tested in E1 is of the milder, procedural kind. The more damaging forms documented in the literature — duplicate patients across partitions, feature engineering from the target, or sampling from the same clinical episode — cannot be tested on this dataset at all, because the public release carries no patient identifiers or provenance. The correct conclusion from E1 is therefore narrow and should be quoted narrowly: *preprocessing and selection leakage, of the kind present in typical published pipelines, does not materially inflate AUC on the UCI CKD benchmark.* It is not evidence that published results on other datasets, or with other leakage pathways, are sound. If anything, E1 sharpens the critique of the benchmark itself: a dataset on which leakage is undetectable because every reasonable method saturates is a dataset that can no longer discriminate between good and bad methodology, and the field's continued reliance on it as the primary evidence for CKD prediction models is itself a finding of this study.

C. The Minimal Panel: Five Routine Measurements

The five-biomarker panel haemoglobin, serum creatinine, specific gravity, packed cell volume and albumin is clinically coherent, which matters for adoption: each member is a routine, inexpensive measurement, and each maps onto a recognised physiological axis of CKD (anaemia of chronic disease for haemoglobin and packed cell volume, filtration function for creatinine, tubular concentrating ability for specific gravity, and glomerular leakage for albumin). The panel agrees with prior independent evidence: the attribute-reduction study of Aljaaf *et al.* [12] retained roughly 30% of the attributes with AUC 0.995, and the leading features of Senan *et al.* [14] (albumin, specific gravity, packed cell volume, serum creatinine) are a subset of the present panel, to which the ensemble consensus adds haemoglobin. Three independent routes converge on the same panel in this study alone selector consensus, fold-wise selection stability (Jaccard 0.85–0.92), and SHAP rankings from two model families (Section IV-F) which is substantially stronger evidence for a biomarker set than any single feature-importance plot. The practical consequence is a screening configuration that requires five measurements instead of twenty-four, at a measured cost of 0.0002 AUC and no loss of sensitivity (0.992 for both panels, pooled). Where this study adds caution is E4: the same redundancy that makes the full panel wasteful also makes it forgiving, and a five-measurement panel has nothing to fall back on when one of its five values is missing or wrong.

D. Calibration and Prevalence: Where the Real Deployment Risks Are

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

If E1 acquits the literature on leakage, E3 convicts it on probability quality. Discrimination near 1.0 coexists with calibration slopes between 0.78 and 2.71: several widely used model families return probabilities that are far too timid or far too extreme for the risks they purport to represent, and no accuracy table would reveal this. The defect is fortunately superficial isotonic recalibration fitted inside the protocol reduces the ensemble's Brier score by more than half (0.0084 to 0.0037) without touching its ranking but it is not optional: a screening programme that refers patients at a stated risk threshold is consuming probabilities, not ranks, and uncalibrated probabilities misallocate referrals in proportion to their distortion. The prevalence analysis makes the same point from the other direction. At the benchmark's 62.5% prevalence, almost any reasonable model looks clinically precise; at a realistic 10% screening prevalence, the full-panel ensemble's positive predictive value falls to 0.892, meaning roughly one in nine positive flags is false. The 5-feature panel's slightly higher specificity converts directly into a PPV of 0.943 under identical conditions. This is, to our knowledge, the first analysis of this benchmark to report predictive values at deployment prevalence, and it reverses the usual framing: the minimal panel is not a compromised version of the full panel but, on the metric that determines false-alarm burden in screening, the better instrument. Negative predictive value remains 0.999 for both panels, so the safety property a screening test most needs a negative result that can be believed is preserved.

E. The Missing-Data Trade-Off

Section III-H hypothesised that a minimal panel would degrade more gracefully under missing data, on the argument that fewer required values means fewer opportunities for absence. E4 shows the opposite mechanism dominates: at 30% missingness, Random Forest loses 0.003 AUC with the full panel and 0.017 with the minimal panel, because the full panel's correlated measurements substitute for one another under median imputation. The minimal panel is by no means fragile in absolute terms its worst observed AUC is 0.972 but deployment settings should be matched to the configuration: the five-biomarker panel suits organised screening, where the five tests are ordered together and their presence is verified; opportunistic use of whatever happens to be in a routine record favours the full panel or a model explicitly trained for missingness. Reporting the hypothesis as falsified, rather than quietly dropping it, is deliberate: the pre-specified design of Section III exists precisely so that unwelcome results cannot be edited out of the paper.

F. Explainability That Can Be Checked

Explainability in this literature is usually decorative a single SHAP summary plot whose message the reader must take on trust. E5 instead treats the explanation as a measurement that can fail: it would have failed if the two model families had disagreed (they did not; $\rho = 0.87$, identical top-five sets), if reducing the panel had scrambled the ranking (it did not; ρ up to 1.00), or if the model had reasoned about diabetic or hypertensive patients through different features than about other patients (it did not). The clinical value of this consistency is specific: a clinician who is told that haemoglobin and creatinine drive the model can sanity-check that claim against physiology once, and rely on it across the model suite, rather than re-auditing every algorithm variant. Stable explanations do not make a model correct Section V-G is explicit that correctness on new patients remains undemonstrated but they make it *inspectable*, which is the precondition for a clinician to catch the cases where it is wrong.

G. Limitations

This study has limitations that bound every claim above. (1) *Single, small, single-source dataset.* All results derive from 400 records collected at one hospital over roughly two months; the near-ceiling discrimination may reflect that cohort's case mix and measurement practice rather than any property of CKD, and no external validation on an independent cohort is performed the single most important next step. (2) *Label construction.* The CKD label in this benchmark was assigned by clinicians using criteria that overlap with the model's input features (creatinine, albumin), so part of the measured performance may be the model re-deriving the labelling rule rather than detecting disease from independent evidence; no analysis on this dataset can fully exclude that circularity, and it is a further reason the absolute AUC values should not be quoted as expected clinical performance. (3) *Missingness mechanism.* E4 simulates missingness completely at random; real clinical missingness is informative (tests are missing because they were not ordered, often for clinical reasons), so real-world degradation may be larger and differently distributed. (4) *Calibration estimates* are computed on pooled out-of-fold predictions averaged over repeats, an estimate of the pipeline's calibration

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 2, 2026
https://jecir.com/

rather than of any single deployable model instance, which would require its own held-out calibration set. (5) *Subgroup analysis* is limited to the diabetes and hypertension indicators the dataset provides; it contains no sex variable, so neither sex-based fairness analysis nor eGFR-based staging was possible. The subgroup SHAP comparison, unlike the cross-model and cross-panel checks, was computed from a Random Forest fitted on the full dataset rather than from out-of-fold predictions, so it is descriptive evidence, not out-of-fold evidence, and should be read accordingly. (6) The study is a retrospective benchmark analysis; nothing in it constitutes clinical validation, and the models should not be used for patient care on the basis of this paper.

H. Future Work

The agenda follows directly from the limitations: external validation of the five-biomarker panel on independent cohorts (population surveys such as NHANES, in which CKD status is derived from measured creatinine and albuminuria, are a natural first target because they break the single-source and prevalence problems simultaneously); extension from binary detection to KDIGO staging and eGFR regression where the necessary variables exist; missingness modelling under MAR/MNAR assumptions and models that consume missingness natively; and a prospective silent-mode evaluation in a screening workflow, in which the panel's predictions and calibrated risks are logged against eventual diagnoses without influencing care. A formal cost-effectiveness analysis of five-test versus full-panel screening in a low-resource setting would complete the deployment argument begun by the decision-curve analysis.

VI. CONCLUSION

This paper set out to test whether the near-perfect performance reported for machine learning-based CKD prediction survives an evaluation designed to be honest, and whether a far smaller biomarker panel can carry that performance into settings where full laboratory panels are unrealistic. Under a leakage-safe protocol in which imputation, scaling and feature selection are fitted strictly inside cross-validation folds, strong classifiers on the UCI CKD benchmark remain near-perfect (ensemble AUC 0.9998) — and, contrary to our expectation, a deliberately leakage-prone protocol scores the same, showing that procedural leakage is not the source of this benchmark's inflated reputation; its saturation is. An ensemble of filter, wrapper and embedded selectors, aggregated by consensus and verified for stability, reduces the panel to five routine biomarkers haemoglobin, serum creatinine, specific gravity, packed cell volume and albumin with no significant loss of discrimination (AUC 0.9996) or sensitivity (0.992), and with a higher positive predictive value (0.943 versus 0.892) at a realistic 10% screening prevalence. The study also shows what accuracy conceals: raw model probabilities are materially miscalibrated and require in-protocol recalibration before any threshold-based use; the minimal panel trades away some robustness to missing measurements (AUC 0.983 versus 0.997 at 30% missingness); and SHAP explanations, subjected to consistency tests across models, panels and subgroups rather than merely displayed, are stable enough to be clinically inspectable. The framework, the code-level protocol and the five-biomarker panel together provide a reproducible basis for low-cost CKD screening research — and the limitations stated in Section V-G, above all the absence of external validation, define the work that must come next before any clinical claim can be made.

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*, vol. 395, no. 10225, pp. 709–733, 2020.

[2] Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group, "KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease," *Kidney International*, vol. 105, no. 4S, pp. S117–S314, 2024.

[3] L. Rubini, P. Soundarapandian, and P. Eswaran, "Chronic Kidney Disease," UCI Machine Learning Repository, 2015, doi: 10.24432/C5G020.

[4] R. Talib *et al.*, "AI-enabled intelligent nephrology: Comparative analysis of artificial intelligence models for robust early prediction and clinical evaluation of chronic kidney disease," *Pakistan Journal of Medical & Cardiological Review*, vol. 4, no. 4, pp. 625–670, 2025.

[5] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, art. 100804, 2023.
[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[7] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006.
[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017, pp. 4765–4774.
[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
[11] A. Charleonnan *et al.*, "Predictive analytics for chronic kidney disease using machine learning techniques," in *Proc. 2016 Management and Innovation Technology International Conference (MITicon)*, 2016, pp. 80–83, doi: 10.1109/MITICON.2016.8025242.
[12] A. J. Aljaaf *et al.*, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in *Proc. IEEE Congress on Evolutionary Computation (CEC)*, 2018, pp. 1–9, doi: 10.1109/CEC.2018.8477876.
[13] J. Qin *et al.*, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2019, doi: 10.1109/ACCESS.2019.2963053.
[14] E. M. Senan, "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques," *Journal of Healthcare Engineering*, vol. 2021, art. 1004767, 2021, doi: 10.1155/2021/1004767.
[15] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, *Lecture Notes in Artificial Intelligence*, vol. 5212, Springer, 2008, pp. 313–325, doi: 10.1007/978-3-540-87481-2_21.
[16] G. R. Vásquez-Morales *et al.*, "Explainable prediction of chronic kidney disease in the Colombian population using neural networks and case-based reasoning," *IEEE Access*, vol. 7, pp. 152900–152910, 2019.
[17] N. A. Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Computers in Biology and Medicine*, vol. 109, pp. 101–111, 2019, doi: 10.1016/j.compbiomed.2019.04.017.
[18] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network," *Future Generation Computer Systems*, vol. 111, pp. 17–26, 2020, doi: 10.1016/j.future.2020.04.036.
[19] F. Sanmarchi, C. Fanconi, D. Golinelli, D. Gori, and A. Hernández-Boussard, "Predict, diagnose, and treat chronic kidney disease with machine learning: A systematic literature review," *Journal of Nephrology*, vol. 36, pp. 1101–1117, 2023, doi: 10.1007/s40620-023-01573-4.
[20] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," *Journal of Machine Learning Research*, vol. 18, no. 174, pp. 1–54, 2018.
[21] B. Van Calster *et al.*, "Calibration: The Achilles heel of predictive analytics," *BMC Medicine*, vol. 17, art. 230, 2019, doi: 10.1186/s12916-019-1466-7.
[22] M. S. Islam, F. Ahmed, W. Ishtiaq, M. A. Hossain, M. S. Islam, and M. M. Tarek, "Advanced explainable ensemble models for multi-class intrusion detection in heterogeneous drone and industrial networks," *Journal on Information Security*, 2026.
[23] V. Govindarajan, F. Ahmed, Z. B. Faheem, M. Bilal, M. Ayadi, and J. Ali, "Aegis-5: A hybrid ensemble framework for intrusion detection in Industry 5.0 driven smart manufacturing environment," *ACM Transactions on Autonomous and Adaptive Systems*, 2026.

[24] F. Amin, U. Imtiaz, and A. Khan, "FALCON-Guard: A lightweight explainable framework for real-time cyber threat detection and adaptive risk mitigation in intelligent driving networks," *Multidisciplinary Research in Computing Information Systems*, vol. 5, no. 12, pp. 1223–1235, 2025.

[25] S. M. H. Shah, F. Amin, and A. Khan, "Cyber-resilient mobile edge computing: A deep neural approach for secure and efficient task offloading," *The Asian Bulletin of Big Data Management*, vol. 5, no. 1, pp. 200–215, 2025.

[26] U. Imtiaz, F. Amin, and A. Khan, "Crash2Compromise: CrashGuard-FS for security-aware Linux file updates," *Multidisciplinary Research in Computing Information Systems*, vol. 6, no. 2, pp. 294–310, 2026.

[26] A. Al Kium, S. Sarker, S. A. Shikha, M. A. T. Kamal, M. I. K. Jabed, N. K. Munifa, et al., "Health equity and digital disparities in cancer screening and cardiovascular care across socioeconomic and ethnic groups: A systematic review," *Vascular and Endovascular Review*, vol. 6, no. 2, pp. 35–44, 2023.

[27] M. I. K. Jabed, "Stock market price prediction using machine learning techniques," *American International Journal of Sciences and Engineering Research*, vol. 7, no. 1, pp. 1–6, 2024.

[28] M. I. K. Jabed, M. Imran, A. A. Khan, M. Mehedi, A. Islam, and R. Pervez, "Explainable machine learning framework for early heart disease detection using SMOTE and SHAP," *Vascular and Endovascular Review*, vol. 9, no. 1, pp. 316–324, 2026.

[29] M. I. K. Jabed, M. R. M. Sirazy, S. Mandal, S. A. Akter, A. Hassan, and H. Esa, "Developing AI-based financial forecasting and cybersecurity systems for the US digital economy," *Frontiers in Computer Science and Artificial Intelligence*, vol. 5, no. 5, pp. 30–38, 2026.

[30] M. A. Rahman, R. K. Devnath, S. B. Niloy, C. M. Mehedi, T. H. Chowdhury, and M. I. K. Jabed, "A stacking ensemble framework for predicting employee turnover: Explainable AI with SHAP," in *Proc. 2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS)*, Oct. 2025, pp. 1–6.

[31] M. A. Rahman, M. I. K. Jabed, R. K. Devnath, C. M. Mehedi, M. Begum, and T. Mahmud, "MSRFF: An interpretable CNN–Vision Transformer framework for diabetic retinopathy detection," in *Proc. 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)*, May 2026, pp. 1–7.

[32] M. M. R. Akash, "Cardiotoxicity prediction models in cancer patients using artificial intelligence and genomics," *International Journal of Drug Delivery Technology*, vol. 16, pp. 60–73.

[33] M. I. K. Jabed, M. Imran, C. A. Gomes, P. S. Ponduru, S. Mandal, and S. Hassan, "Deep learning and explainable AI framework for predicting lung cancer severity," *Journal of Computer Science and Technology Studies*, vol. 7, no. 12, pp. 573–598, 2025.

[34] M. I. K. Jabed, M. Imran, C. A. Gomes, and P. S. Ponduru, "Machine learning-based prediction of poor self-rated health among US adults using behavioral and socioeconomic factors," *World Journal of Advanced Research and Reviews*, vol. 27, no. 1, pp. 2817–2829, 2025, doi: 10.30574/wjarr.2025.27.1.2649.

[35] M. I. K. Jabed, M. A. Manzoor, F. M. Tofa, and M. H. Khan, "Interpretable ensemble learning approach for breast cancer diagnosis using SHAP-based explainable AI," *Journal of Computer Science and Technology Studies*, vol. 8, no. 8, pp. 244–255, 2026.

[36] M. I. K. Jabed, M. P. Ahmed, F. M. Tofa, M. F. Islam, C. A. Gomes, and R. M. Sirazy, "Federated intrusion detection for Internet of Medical Things networks: Differential privacy, non-IID robustness, and cross-device generalization," *Journal of Computer Science and Technology Studies*, vol. 8, no. 8, pp. 303–315, 2026.

[37] P. S. Ponduru, "Road accident prediction using LSTM-GRU neural networks," Ph.D. dissertation, California State University, Northridge, 2023.

[38] P. S. Ponduru, P. P. V. Nandanavanam, and S. K. K. Ponduru, "Ecological proportionality in generative AI: The ethics of marginal capability and environmental sufficiency," 2026.

[39] P. S. Ponduru, "Decision intelligence for AI and emerging technologies: The AEGIS-DM framework for trustworthy, cost-aware, and low-latency decision making," 2024.

[40] P. S. Ponduru, P. P. V. Nandanavanam, and S. K. K. Ponduru, "SAFE-HealCloud: Safety-aware, agentic self-healing for cloud infrastructure," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 12, no. 4, pp. 163–188, 2026.

[41] M. Akram, W. Khan, M. D. Rasheed, M. Imran, M. W. Iqbal, A. Delshadi, and M. Sultana, "Cybersecurity risk assessment model for Internet of Medical Things (IoMT) devices in healthcare systems," *Spectrum of Engineering Sciences*, vol. 4, no. 3, pp. 312–326, 2026.

[42] M. D. Rasheed, M. Akram, M. Imran, R. H. Ahmed, and A. Rauf, "Towards human-centric smart manufacturing: A digital twin enabled affective ergonomic framework for adaptive human robot collaboration," *International Journal of Advances in Signal and Image Sciences*, pp. 1544–1555, 2025.

[43] M. D. Rasheed, W. Khan, M. Imran, N. Ahmad, Y. Khan, M. Akram, and M. Sultana, "Leveraging artificial intelligence for advanced data networking and cybersecurity," *Spectrum of Engineering Sciences*, vol. 4, no. 3, pp. 286–298, 2026.

[44] P. K. Bhowmik, F. R. Chowdhury, M. Sumsuzzaman, R. K. Ray, M. M. Khan, C. A. H. Gomes, et al., "AI-driven sentiment analysis for Bitcoin market trends: A predictive approach to crypto volatility," *Journal of Ecohumanism*, vol. 4, no. 4, pp. 266–288, 2025.

[45] M. R. Hasan, M. A. Rahman, C. A. H. Gomes, F. N. Nitu, C. A. Gomes, M. R. Islam, and R. E. R. Shawon, "Building robust AI and machine learning models for supplier risk management: A data-driven strategy for enhancing supply chain resilience in the USA," *Advances in Consumer Research*, vol. 2, no. 4, pp. 1152–1171, 2025.

[46] N. M. Ali, M. E. Ahmed, M. Fakhrul Islam, C. A. Gomes, and M. S. Islam, "Toward a circular US economy: Green and artificial intelligence innovation, renewable energy, and domestic material consumption," *Energy Sources, Part B: Economics, Planning, and Policy*, vol. 21, no. 1, Art. no. 2685043, 2026.

[47] C. A. Gomes, M. I. K. Jabed, T. Ahammad, S. Mandal, S. Hassan, M. K. Sejan, P. S. Ponduru, M. M. Bhuiyan, and M. L. Jamali, "Operational and financial evidence for generative AI in United States healthcare administrative workflows," *Frontiers in Computer Science and Artificial Intelligence*, vol. 4, no. 4, pp. 138–145, 2025.