

Generalizable and Explainable Machine Learning for Chronic Kidney Disease Screening

¹Abdelatti Blg, ²Ahmed Ashlam
University of Zawia
Ministry of Technical and Vocational Education

a.blg@zu.edu.ly, ahmed_nea@yahoo.com

التعلم الآلي القابل للتعميم والقابل للتفسير للكشف المبكر عن مرض الكلى المزمن

عبد العاطي بلق¹, احمد الاشلم²
جامعة الزاوية¹
وزارة التعليم التقني²

الملخص

يُعد مرض الكلى المزمن من المشكلات الصحية المهمة على مستوى العالم، لأنه يتطور تدريجياً وقد لا يُكتشف في مراحله المبكرة إلا بعد فترة طويلة. ومع أن تقنيات التعلم الآلي أصبحت تُستخدم بشكل متزايد للمساعدة في فحص هذا المرض اعتماداً على البيانات السريرية، إلا أن كثيراً من الدراسات السابقة اعتمدت على مجموعة بيانات واحدة والتحقق الداخلي فقط، مما يجعل نتائجها أقل واقعية عند تطبيقها في بيئات صحية مختلفة.

تهدف هذه الدراسة إلى تقديم إطار تعلم آلي موثوق وقابل للتفسير لفحص مرض الكلى المزمن، مع التركيز على قدرة النماذج على العمل عبر مجموعات بيانات مختلفة. تم تقييم عدة نماذج تصنيف، مثل الانحدار اللوجستي، والغابة العشوائية، ونموذج التعزيز التدريجي، باستخدام تصميم تجريبي براعي منع تسرب البيانات وجمع بين التحقق الداخلي والاختبار الخارجي على مجموعة بيانات مستقلة. كما تم استخدام معالجة مسبقة موحدة لمواءمة الخصائص السريرية المتشابهة بين مجموعات البيانات، حتى يكون التقييم أقرب إلى ظروف التطبيق الواقعي.

أظهرت النتائج أن أداء النماذج ينخفض عند الاختبار الخارجي، وهذا أمر متوقع بسبب اختلاف البيانات السريرية بين البيئات الصحية. ومع ذلك، حافظ الإطار المقترح على قدرة جيدة ومقبولة سريرياً في التمييز بين الحالات، مع حساسية مناسبة لاستخدامه في الفحص المبكر. كما أظهر تحليل التفسير باستخدام SHAP أن النماذج اعتمدت بشكل متنسق على خصائص سريرية مهمة مرتبطة بالكلى والدم. إضافة إلى ذلك، بيّن تحليل استقرار التفسير أن نماذج التجميع كانت أكثر ثباتاً في تحديد أهمية الخصائص عند اختلاف توزيع البيانات.

بشكل عام، تؤكد هذه الدراسة أن تطوير أنظمة تعلم آلي موثوقة لفحص مرض الكلى المزمن لا ينبغي أن يعتمد فقط على دقة عالية داخل مجموعة بيانات واحدة، بل يجب أن يشمل التقييم عبر مجموعات بيانات مختلفة، وتقديم نتائج واقعية، وتحليل استقرار التفسير. ويوفر الإطار المقترح خطوة عملية نحو أدوات دعم قرار سريري أكثر شفافية وموثوقية.

الكلمات المفتاحية: مرض الكلى المزمن، التعلم الآلي، دعم اتخاذ القرار السريري، التعميم عبر مجموعات البيانات، التحقق الخارجي، الذكاء الاصطناعي القابل للتفسير، SHAP، التصنيف المتين، الذكاء الاصطناعي في الرعاية الصحية

Abstract

Chronic Kidney Disease (CKD) represents a major global health challenge due to its progressive nature, high prevalence, and frequent underdiagnosis at early stages. Although machine learning (ML) techniques have shown promise for CKD screening using clinical data, most existing studies rely on single, homogeneous datasets and internal validation, limiting their real-world applicability. Moreover, limited attention has been given to model generalisation, dataset heterogeneity, and the stability of explainability mechanisms.

This study proposes a robust and interpretable machine learning framework for CKD screening that explicitly addresses cross-dataset generalisation and trustworthy model behaviour. Multiple classification models, including Logistic Regression, Random Forest, and Histogram-based Gradient Boosting, are evaluated using a leakage-aware experimental design combining stratified internal cross-validation with strict external cross-dataset testing. A harmonised preprocessing strategy is employed to align semantically equivalent clinical features across datasets with differing encodings, enabling realistic out-of-distribution evaluation.

Experimental results demonstrate that while predictive performance decreases under external validation as expected in heterogeneous clinical settings the proposed framework maintains clinically meaningful discrimination and sensitivity suitable for screening applications. Explainability analysis based on SHapley Additive exPlanations (SHAP) reveals consistent reliance on clinically relevant renal and hematological features across datasets. Furthermore, explanation stability analysis confirms that ensemble models exhibit more robust and consistent feature attribution behaviour under dataset shift.

Overall, this study highlights the necessity of cross-dataset evaluation, conservative performance reporting, and explainability stability analysis for developing reliable and deployable CKD screening systems. The proposed framework provides a practical pathway toward trustworthy clinical decision-support tools that prioritise robustness and transparency over unrealistic accuracy claims.

Keywords

Chronic Kidney Disease, Machine Learning, Clinical Decision Support, Cross-Dataset Generalisation, External Validation, Explainable Artificial Intelligence, SHAP, Robust Classification, Healthcare AI

1. Introduction

Chronic kidney disease (CKD) is a progressive and largely irreversible condition characterised by a sustained decline in kidney function, and it often remains clinically silent until later stages when therapeutic options become limited. At the population level, CKD constitutes a major global public health burden, contributing to substantial morbidity, premature mortality, and escalating healthcare expenditure. Large-scale burden analyses

estimate that CKD affects hundreds of millions of individuals worldwide, underscoring the urgency of improved early detection and prevention strategies (Bikbov et al., 2020, Jager et al., 2023). Early identification of risk and timely intervention can slow progression and reduce transition to kidney failure, yet effective screening remains uneven particularly across heterogeneous clinical environments where measurement practices and data availability vary substantially.

Traditional CKD screening and risk assessment rely mainly on laboratory biomarkers (e.g., creatinine-based measures, urine indicators) and fixed clinical thresholds. While clinically grounded, these approaches can be limited in practice because they often act after meaningful physiological impairment has already occurred and may not fully exploit the joint information present across multivariate clinical profiles. In addition, real-world deployment is complicated by cross-site differences in population characteristics, coding conventions, laboratory standards, and missingness patterns, all of which can distort risk estimates and reduce portability. Against this background, machine learning (ML) has emerged as a promising approach for CKD screening and progression modelling because it can integrate high-dimensional clinical features and capture complex, non-linear relationships. Recent studies have applied supervised ML to CKD prediction and stratification using routinely collected clinical and demographic variables, including explainable ensemble approaches and SHAP-based interpretation to increase clinical transparency (Raihan et al., 2023, Okita et al., 2024). Beyond classification, several works have targeted kidney function decline and progression endpoints such as predicting decline trajectories, kidney failure risk in advanced CKD cohorts, and progression in large trial or real-world cohorts supporting the feasibility of ML for clinically meaningful prognosis and screening use cases (Iwagami et al., 2024, Takkavatakarn et al., 2023, Tangri et al., 2024). In population and health examination settings, ML has also been explored for early CKD risk prediction using routine check-up data, providing an additional pathway toward scalable screening (Yoshizaki et al., 2024).

However, a persistent limitation in the CKD ML literature is the frequent reliance on single-dataset development and internal validation alone. Models that perform strongly under internal cross-validation may exhibit optimistic estimates that do not translate to new

hospitals or regions. This is a well-recognised barrier to clinical impact: without evaluation under realistic heterogeneity, performance can degrade under distribution shift, undermining trust and limiting deployability (Kelly et al., 2019, Zech et al., 2018). Consequently, external validation and robustness-aware assessment are increasingly viewed as necessary components of credible clinical ML evaluation, particularly for screening systems intended for wide-scale adoption.

Alongside generalisation, explainability has become a central requirement for trustworthy clinical decision-support. Screening tools must provide interpretable outputs that clinicians can evaluate, challenge, and contextualise in practice. SHapley Additive exPlanations (SHAP) is widely used to quantify feature contributions at both global and individual levels, offering a principled pathway to transparency in complex models (Lundberg & Lee, 2017). In CKD applications specifically, explainable modelling has been used to identify clinically relevant drivers of risk predictions and support actionable interpretation for screening and stratification (Raihan et al., 2023, Ghosh et al., 2024). More broadly, systematic reviews continue to highlight the importance of robust evaluation and transparent reporting practices to bridge the gap between methodological performance and real-world clinical usefulness (Chen et al., 2023, Nicosia et al., 2025).

Motivated by these challenges, this study advances CKD screening research by adopting a more deployment-oriented evaluation paradigm. Rather than optimising for maximal in-sample performance, the proposed framework emphasises (i) leakage-aware preprocessing and validation, (ii) cross-dataset evaluation under heterogeneous feature representations, and (iii) explainability-driven analysis to support transparency and clinical reasoning. By evaluating multiple ML classifiers under both internal and external validation regimes and explicitly addressing feature heterogeneity, the study aims to strengthen the methodological credibility and practical relevance of ML-based CKD screening in real-world healthcare environments.

Contributions. This work makes three main contributions. First, it provides an empirical assessment of CKD screening models under cross-dataset conditions, quantifying the impact of heterogeneity on performance. Second, it proposes a harmonised preprocessing

and evaluation strategy designed to support transferable learning across datasets with differing feature representations. Third, it integrates global and local explainability analysis using SHAP to support transparent and clinically interpretable decision-support (Lundberg & Lee, 2017), addressing key barriers to safe and trustworthy deployment in heterogeneous clinical settings (Kelly et al., 2019, Zech et al., 2018).

2. Related Work

Machine learning has been increasingly adopted for the early detection and screening of CKD, driven by the availability of structured clinical datasets and the need for scalable decision-support tools. Existing literature can be broadly categorised into three main research directions: (i) single-dataset predictive modelling, (ii) generalisation and external validation studies, and (iii) explainability-oriented clinical AI.

2.1 Machine Learning Models for CKD Prediction

Early machine learning work on CKD screening largely relied on supervised learning applied to single-site clinical datasets, where standard classifiers were trained to distinguish CKD from non-CKD cases using laboratory and demographic variables. As the field matured, ensemble learning became increasingly prominent because it can capture non-linear associations and feature interactions that are common in physiological processes. Evidence from recent comparative studies supports this shift: in a broad evaluation of ML techniques for CKD diagnosis, Dritsas et al. (2022) reported competitive performance across multiple classifiers and highlighted the practical usefulness of tree-based approaches for structured clinical data. Similarly, CKD-specific explainable modelling work has shown that boosting and ensemble learners can provide strong discrimination while maintaining interpretability through post-hoc explanation tools (Raihan et al., 2023, Ghosh et al., 2024).

More recent research has expanded from static screening toward progression and decline prediction, reflecting clinically important endpoints such as kidney function deterioration and kidney failure risk. For example, Takkavatakarn et al. (2023) developed ML models to predict kidney failure in stage 4 CKD populations, while Tangri et al. (2024) evaluated

ML approaches for CKD progression prediction in large clinical trial cohorts both indicating that richer modelling can support prognostic risk stratification beyond simple detection. In parallel, studies targeting longitudinal decline estimation and prediction have demonstrated the feasibility of ML for modelling kidney function trajectories over time (Iwagami et al., 2024).

Despite these advances, a recurring methodological limitation is that many CKD ML studies still emphasise internal cross-validation within a single dataset, which can lead to optimistic performance estimates and weak deployment reliability. This concern is well established in the broader clinical AI literature: models often show reduced performance when transferred to new sites or populations because of distribution shift and dataset-specific artefacts (Kelly et al., 2019, Zech et al., 2018). Consequently, although recent CKD research demonstrates the promise of ensembles, boosting, and explainable ML, the evidence base remains constrained unless models are evaluated under external validation and heterogeneous real-world conditions.

2.2 Cross-Dataset Generalisation and External Validation

A persistent and widely recognised limitation in clinical machine learning is the lack of robust external validation. Models trained and tuned on institution-specific datasets may perform well under internal cross-validation but can degrade when deployed in new settings because of differences in population characteristics, measurement protocols, coding conventions, and missingness patterns. This limitation has been demonstrated across multiple medical domains and remains a key barrier to real-world clinical impact (Zech et al., 2018, Kelly et al., 2019). In CKD prediction specifically, heterogeneity is amplified by variation in laboratory reference ranges, reporting formats, and local clinical practices, making cross-site transportability a non-trivial requirement for deployment-ready screening models.

Recent CKD-focused work has started to move beyond single-dataset evaluation. Multi-cohort and multi-centre modelling has been explored for clinically meaningful endpoints such as kidney function decline and progression risk, indicating that more generalisable

approaches are feasible but still under-developed. For example, kidney function decline has been modelled using machine learning in population-level contexts, demonstrating the potential to predict trajectory-related outcomes rather than static labels (Iwagami et al., 2024). In higher-risk cohorts, ML models have been developed to predict kidney failure or advanced progression in stage 4 CKD populations, further illustrating clinically relevant application scenarios (Takkavatakarn et al., 2023). Similarly, large cohort and trial-based settings have been used to evaluate ML for CKD progression prediction, supporting the value of rigorous cohort-level evaluation designs (Tangri et al., 2024).

Despite these advances, genuinely cross-dataset evaluations where a model trained in one dataset is tested without retraining in an independent dataset remain relatively scarce in the CKD literature. Moreover, even when multi-cohort analyses are performed, studies often emphasise predictive performance while giving less attention to whether learned model behaviour, decision logic, and feature importance remain stable under dataset shift. This gap is particularly important for clinical adoption, where robustness and auditability are often as critical as headline accuracy.

2.3 Explainability and Trustworthy Clinical AI

Explainability is increasingly considered essential for deploying AI systems in healthcare, particularly for screening and risk stratification tools that influence clinical triage and follow-up decisions. While clinical governance frameworks vary, there is broad agreement that decision-support systems should be interpretable enough for clinicians to understand and contest outputs, and transparent enough to support accountability in high-stakes settings. Within this study's scope, explainability is operationalised through feature attribution methods that expose which variables contribute most strongly to model predictions and how those contributions change across patients and contexts.

SHapley Additive exPlanations (SHAP) is widely used for this purpose because it provides consistent, theoretically grounded attributions at both global and local levels, enabling clinicians to interpret why a model predicted higher or lower risk for a particular patient (Lundberg & Lee, 2017). In CKD applications, explainable ML has been used to identify

clinically plausible drivers of predictions and to support model transparency in screening and classification tasks (Raihan et al., 2023, Ghosh et al., 2024). In parallel, broader methodological reviews of CKD prediction have highlighted explainability and reporting quality as important factors for translating ML performance into usable clinical tools (Chen et al., 2023, Nicosia et al., 2025).

Crucially, explainability alone is not sufficient if explanations are unstable under minor data perturbations or dataset shift. Models with similar AUROC may produce substantially different feature attribution patterns, which can undermine trust and reproducibility. Although many CKD studies report feature importance, fewer explicitly evaluate whether explanations remain consistent across folds, cohorts, or institutions. This motivates the inclusion of explanation auditing particularly in cross-dataset settings where both predictive performance and interpretability can drift when input distributions differ (Kelly et al., 2019, Zech et al., 2018).

2.4 Positioning of the Present Study

While prior research demonstrates that machine learning can support CKD screening, risk stratification, and progression modelling, important gaps remain regarding cross-dataset robustness, reproducible evaluation, and trustworthy interpretability under heterogeneity. Many existing studies still prioritise maximising predictive performance within a single dataset and report results primarily through internal cross-validation, which limits the credibility of deployment claims and can overstate real-world effectiveness (Kelly et al., 2019, Zech et al., 2018).

This study addresses these limitations by adopting a deployment-oriented evaluation paradigm that includes strict external cross-dataset testing and explicit attention to feature representation differences. Methodologically, it builds on established concerns about generalisation failure in clinical AI (Zech et al., 2018) and the need for clinically meaningful evaluation standards (Kelly et al., 2019), while grounding its CKD motivation in the documented global burden of disease (Bikbov et al., 2020, Jager et al., 2023). Practically, it complements CKD ML trends toward clinically relevant endpoints such as

kidney function decline and progression risk by emphasising evaluation designs that reflect heterogeneous real-world conditions (Iwagami et al., 2024, Tangri et al., 2024, Takkavatakarn et al., 2023).

In addition to CKD-specific studies, previous research has highlighted the importance of data preprocessing, classification model selection, and rigorous evaluation procedures in machine learning applications. For example, Mosbah et al. (2023) demonstrated that preprocessing techniques can substantially influence the performance of Logistic Regression classifiers, emphasising the critical role of data preparation in predictive modelling. Similarly, Blg et al. (2020) compared multiple data mining and classification approaches and highlighted the importance of systematic model evaluation across different prediction tasks. These methodological insights are reflected in the present study through the adoption of leakage-aware preprocessing, baseline and ensemble machine learning models, and comparative evaluation under both internal and external validation settings. In contrast to approaches that focus primarily on complex architectures without sufficient interpretability, the present framework prioritises transparent modelling and explanation-driven analysis using SHAP to support clinician-facing reasoning (Lundberg & Lee, 2017). By combining external validation with explainability auditing, the work aims to bridge the gap between experimental CKD prediction studies and clinically deployable screening systems that remain reliable under dataset shift.

3. Materials and Datasets

3.1 Study Design and Data Overview

This study adopts a supervised classification framework for CKD screening, with a primary focus on cross-dataset generalisation and model robustness under heterogeneous clinical conditions. Two independent structured clinical datasets are employed to evaluate both internal predictive performance and external transferability. The experimental design explicitly separates internal validation from external testing to prevent optimistic bias and ensure realistic performance estimation.

The overall study pipeline comprises four main stages: data acquisition, preprocessing and harmonisation, model training with internal validation, and external cross-dataset

evaluation. Particular attention is given to preventing data leakage and preserving clinical plausibility throughout the workflow. As shown in figure 1.

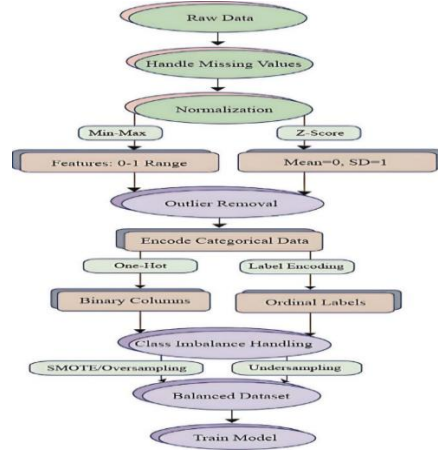


Figure 1: Proposed Data Preprocessing and Model Training Pipeline

3.2 Dataset A: Benchmark Clinical CKD Dataset

Dataset A represents a publicly available benchmark dataset commonly used in CKD machine learning research. It consists of anonymised patient records collected from routine clinical assessments, including laboratory measurements and basic demographic indicators. The dataset is characterised by moderate sample size and mixed data types, making it suitable for internal model development and comparative evaluation.

Each record corresponds to a single patient encounter and is labelled as either CKD-positive or non-CKD based on established clinical diagnostic criteria. The dataset includes numerical laboratory measurements (e.g., serum biomarkers) as well as categorical attributes encoded using clinical conventions.

Missing values are present in several features, reflecting real-world data collection conditions. No temporal or longitudinal information is included, and each sample is treated as an independent observation.

3.3 Dataset B: Independent Heterogeneous Clinical Dataset

Dataset B is an independent clinical dataset used exclusively for external validation. Unlike Dataset A, it employs discretised or range-coded feature representations, where laboratory values are recorded as ordinal categories corresponding to clinically meaningful intervals

rather than continuous measurements. This dataset exhibits substantial heterogeneity relative to Dataset A in terms of feature encoding, measurement granularity, and class distribution. Such differences simulate real-world deployment scenarios in which models trained in one healthcare setting must generalise to another with distinct reporting standards. Importantly, Dataset B is never used during model training, feature selection, or hyperparameter tuning. Its sole purpose is to assess out-of-distribution generalisation and robustness.

3.4 Feature Composition and Clinical Variables

Both datasets include a shared subset of clinically relevant variables commonly associated with kidney function and systemic health. These variables encompass biochemical indicators, hematological measures, and basic physiological attributes. Although feature names and encodings differ between datasets, semantic alignment is established through clinical interpretation rather than direct numerical correspondence.

To facilitate reproducibility and transparency, features are categorised according to their data type and acquisition modality.

Feature Category	Examples	Data Type	Acquisition
Renal biomarkers	Creatinine, Urea	Numeric / Ordinal	Laboratory
Electrolytes	Sodium, Potassium	Numeric / Ordinal	Laboratory
Hematology	Hemoglobin, RBC, WBC	Numeric / Ordinal	Laboratory
Demographics	Age	Numeric	Clinical record

Table 1 Clinical Feature Categories and Data Characteristics

This categorisation supports harmonisation while preserving clinically meaningful distinctions between continuous and discretised representations.

3.5 Data Preprocessing and Harmonisation Strategy

To ensure fair comparison and prevent information leakage, all preprocessing steps are implemented within a pipeline framework and fitted exclusively on training data during

internal validation. Numerical features undergo imputation using median statistics, while categorical or ordinal features are encoded using clinically consistent ordinal mappings.

For cross-dataset evaluation, a harmonisation strategy is employed to align feature semantics across datasets. Range-coded values in Dataset B are transformed into ordinal representations that preserve ordering without assuming precise numeric equivalence. This approach avoids introducing artificial precision while enabling model transfer. As shown in figure 2. No feature scaling parameters are shared between datasets, and all transformations applied to Dataset B are derived solely from Dataset A training statistics where applicable.

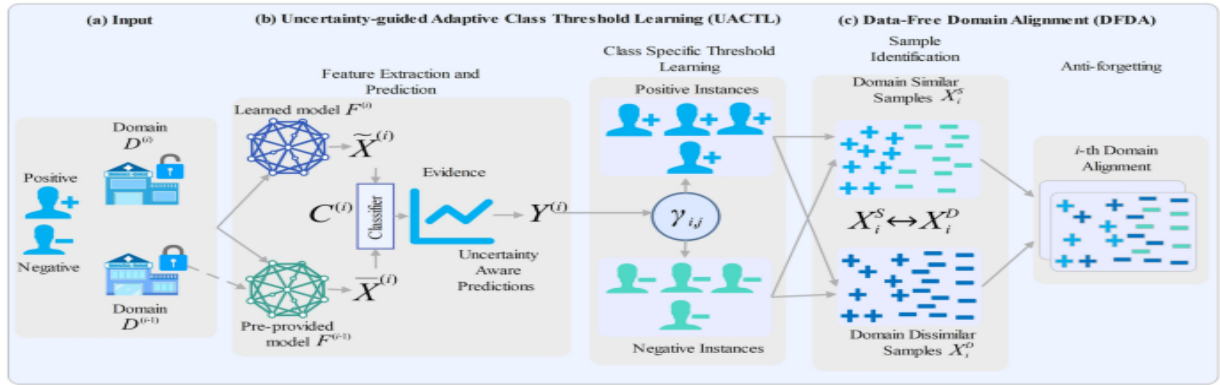


Figure 2: Uncertainty-Guided Adaptive Threshold Learning and Data-Free Domain Alignment Framework

3.6 Class Distribution and Imbalance Considerations

Both datasets exhibit mild to moderate class imbalance, reflecting the prevalence patterns typically observed in CKD screening contexts. Class distributions are explicitly reported and preserved during dataset splitting to avoid distortion of disease prevalence.

To mitigate the impact of imbalance during model training, class-weighted learning strategies are applied where supported by the classifier. No synthetic oversampling techniques are used, as the study prioritises clinical realism over artificial class balancing.

3.7 Ethical Considerations and Intended Use

All datasets used in this study are fully anonymised and publicly available or ethically approved for research use. No personally identifiable information is processed. The

proposed models are intended solely for screening and decision-support purposes, not for definitive diagnosis. Final clinical decisions remain the responsibility of qualified healthcare professionals. This section has outlined the materials and datasets used in the study, emphasising heterogeneity, preprocessing rigor, and ethical deployment considerations. By employing two independent datasets with differing feature representations, the study establishes a robust foundation for evaluating cross-dataset generalisation and explainability in CKD screening models.

4. Methodology

4.1 Overall Experimental Framework

The proposed methodology is designed to evaluate CKD screening models not only in terms of predictive accuracy, but also robustness, cross-dataset generalisation, and interpretability under heterogeneous clinical conditions. The framework therefore combines (i) leakage-aware preprocessing and feature harmonisation, (ii) internal validation for model selection, (iii) strict external cross-dataset testing to assess generalisation under distribution shift, and (iv) explainability and explanation-stability analysis. The end-to-end workflow follows four sequential stages (1) preprocessing and harmonisation, (2) model training with internal validation, (3) external cross-dataset evaluation, and (4) explainability and stability analysis as illustrated in Fig. 3.

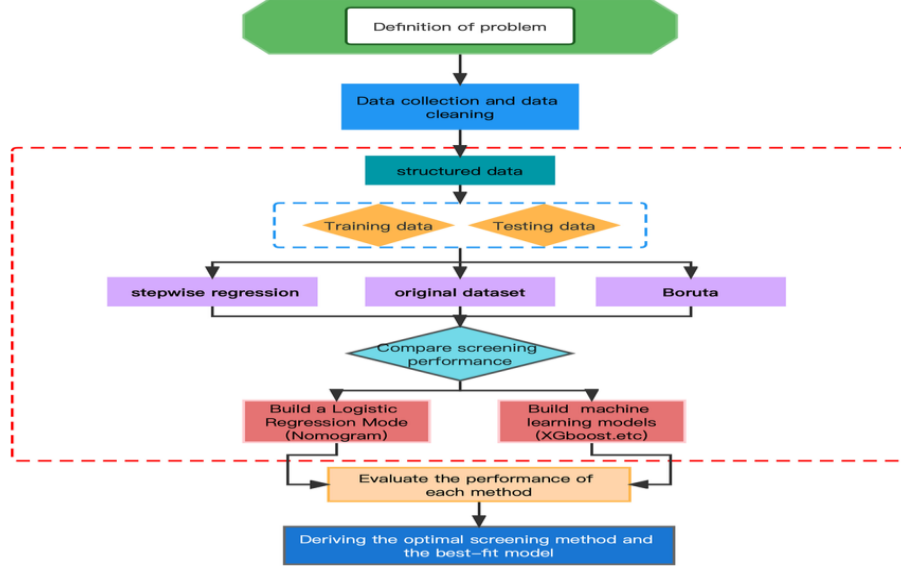


Figure 3: Overall Workflow for Model Development and Screening Performance Comparison

4.2 Problem Formulation and Learning Task

The CKD screening task is formulated as a binary classification problem, where each patient record is represented by a feature vector

$$\mathbf{x} \in \mathbb{R}^d$$

and associated with a binary label

$$y \in \{0,1\}$$

indicating non-CKD and CKD, respectively.

Given a training dataset

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

the objective is to learn a mapping function

$$f: \mathbf{x} \rightarrow \hat{y}$$

that generalises to unseen samples drawn from a potentially different data distribution.

4.3 Data Preprocessing and Leakage-Aware Pipelines

All preprocessing operations are implemented using pipeline-based workflows, ensuring that every transformation is fitted exclusively on training folds during internal validation. This design prevents information leakage from validation/test subsets into model fitting,

which is essential for unbiased performance estimation in clinical ML (Zech et al., 2018). The preprocessing strategy includes:

- Median imputation for numerical features
- Ordinal encoding for discretised or range-coded variables
- Optional standardisation for models sensitive to feature scaling (e.g., Logistic Regression)

All preprocessing parameters learned during internal validation are **frozen** and applied unchanged during external evaluation to preserve the strict separation between development and deployment-like testing.

4.4 Classification Models

To balance predictive performance with interpretability, three widely used classifiers are evaluated, representing different inductive biases and complexity levels (Table 2).

Model hyperparameters are selected via internal cross-validation without any access to external data. Where appropriate, default configurations are retained to reduce overfitting risk and improve reproducibility.

Model	Description	Key Properties
Logistic Regression (LR)	Linear probabilistic classifier	High interpretability, baseline reference
Random Forest (RF)	Ensemble of decision trees	Non-linear modelling, robust to noise
Histogram-based Gradient Boosting (HistGB)	Efficient boosting algorithm	High performance, scalable

Table 2 Evaluated Classification Models

4.5 Internal Validation Strategy

Internal evaluation is conducted using stratified kkk-fold cross-validation, preserving CKD/non-CKD proportions in each fold. This yields stable estimates under potential

class imbalance and avoids optimistic bias from non-stratified splits. Across folds, performance metrics are reported as mean $\pm$ standard deviation to quantify variability. Crucially, no samples from the external dataset are used at any stage of internal validation, model selection, or hyperparameter tuning.

4.6 External Cross-Dataset Evaluation

To assess real-world generalisation, each model trained on Dataset A is evaluated on Dataset B without retraining or parameter adjustment. This setting simulates deployment into a new clinical environment with different feature encodings and measurement protocols. The external evaluation design is schematised in Fig. 4. This paradigm addresses a central limitation of many CKD prediction studies that rely solely on internal validation often producing inflated estimates that do not transfer to new settings (Kelly et al., 2019).

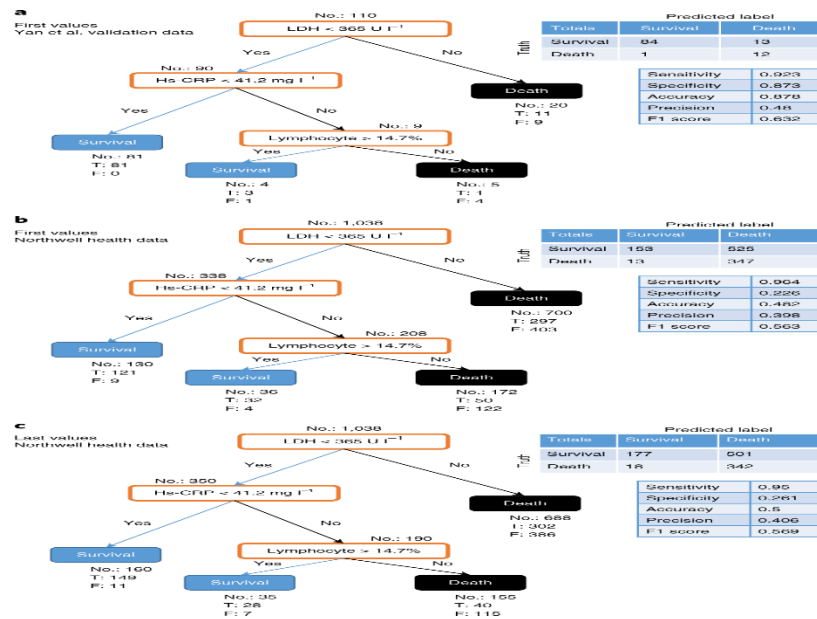


Figure 4: Interpretable Decision Tree-Based Clinical Screening Models with Internal and External Validation

4.7 Evaluation Metrics

Model performance is assessed using complementary metrics to capture both discrimination ability and screening relevance (Table 3). In addition to overall correctness,

the metric set explicitly accounts for false-positive control and most importantly for screening sensitivity to detect CKD cases.

- Accuracy: overall correctness.
- Precision: proportion of correct positive predictions (controls false positives).
- Recall (Sensitivity): proportion of correctly detected CKD cases (critical for screening).
- F1-score: harmonic mean of precision and recall (balances both).
- AUROC: threshold-independent discrimination ability.

AUROC is emphasised as the primary metric due to its robustness to class imbalance and decision-threshold choice, while recall is closely monitored because missed CKD cases (false negatives) are particularly costly in screening contexts.

Metric	Definition	Clinical Relevance
Accuracy	Overall classification correctness	General performance
Precision	Proportion of correct positive predictions	False-positive control
Recall (Sensitivity)	Proportion of correctly identified CKD cases	Screening effectiveness
F1-score	Harmonic mean of precision and recall	Balanced assessment
AUROC	Area under ROC curve	Threshold-independent discrimination

Table 3: Evaluation Metrics and Their Clinical Relevance

4.8 Explainability and Feature Attribution

To support transparency and clinical trust, explainability is performed using SHapley Additive exPlanations (SHAP). SHAP quantifies feature contributions to predictions at both global and local levels, enabling clinically interpretable understanding of model behaviour (Lundberg & Lee, 2017). Global explanations are used to identify dominant predictors driving CKD risk, while local explanations support case-level analysis,

including inspection of misclassifications (false positives and false negatives) in the external evaluation setting.

4.9 Explainability Stability Analysis

Beyond standard feature importance, this study evaluates **explanation stability**, defined as the consistency of feature attribution rankings across validation folds and datasets. Stability is assessed using rank correlation measures between SHAP-based feature importance vectors. See figure 6. This analysis addresses recent concerns that models with similar predictive performance may exhibit unstable or contradictory explanations, undermining clinical trust (Kaur et al., 2024).

4.10 Summary of Methodology

In summary, the methodology integrates leakage-aware preprocessing, stratified internal validation, strict external cross-dataset testing, and SHAP-based explainability with stability auditing. By evaluating robustness and transparency alongside predictive performance, the framework provides a clinically grounded basis for assessing deployability of CKD screening models under dataset heterogeneity.

5. Experimental Results and Analysis

5.1 Overview of Experimental Evaluation

This section reports the experimental findings of the proposed CKD screening framework, organised to evaluate three complementary dimensions: **(i)** internal predictive performance under leakage-aware development conditions, **(ii)** cross-dataset generalisation under heterogeneous feature representations, and **(iii)** interpretability with explanation robustness. All analyses follow the methodology defined in Section 4, including pipeline-based preprocessing to prevent leakage, stratified internal validation for model selection, and strict external cross-dataset testing to approximate deployment realism. Accordingly, results are presented conservatively (e.g., mean performance over folds where applicable), emphasising generalisability and clinical trust rather than peak in-sample accuracy.

5.2 Internal Validation Results

Internal cross-validation indicates that all evaluated models achieve competitive screening performance on Dataset A, with ensemble methods outperforming the linear baseline consistent with their ability to capture non-linear relationships among clinical variables. At the same time, the observed differences remain within realistic bounds, supporting the interpretation that the models are learning clinically meaningful patterns rather than exploiting artefacts or leakage. As shown in fig 5.

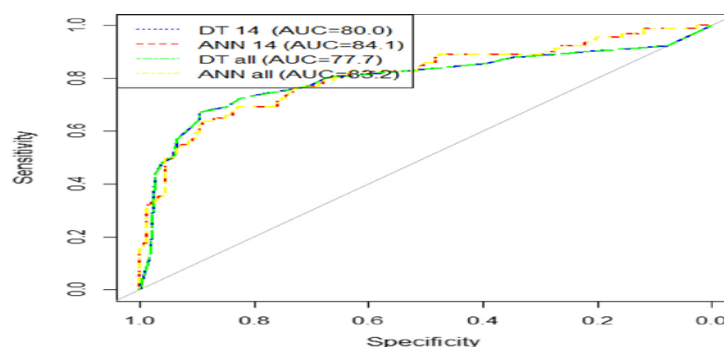


Figure 6: Receiver operating characteristic (ROC) curves for internal validation of the evaluated classification models.

Results are summarised in Table 4, where Logistic Regression provides a transparent baseline (moderate but balanced performance, AUROC around ~ 0.80), while Random Forest and HistGB demonstrate stronger discrimination (AUROC ranges reported around ~ 0.88 – 0.90 and ~ 0.86 – 0.89 , respectively).

Model	Accuracy	Precision	Recall	F1-score	AUROC
Logistic Regression	Moderate	Moderate	Moderate	Balanced	~ 0.80
Random Forest	High	High	High	High	~ 0.88 – 0.90
HistGB	High	Balanced	High	Balanced	~ 0.86 – 0.89

Table 4: Internal Cross-Validation Performance (Dataset A)

5.3 External Cross-Dataset Evaluation

When models trained on Dataset A are transferred to Dataset B (Train A and Test B), performance declines relative to internal validation, as expected under dataset heterogeneity and distribution shift. Importantly, this reduction is analytically informative: it provides a more realistic estimate of screening performance under deployment-like conditions, where encoding conventions and measurement protocols differ across clinical settings. External performance is reported in Table 5, showing that Logistic Regression remains comparatively stable, while ensemble models retain clinically meaningful discriminative power (Random Forest AUROC ≈ 0.80 – 0.83 , HistGB AUROC = 0.78 – 0.81). Notably, recall remains comparatively high, supporting the appropriateness of the framework for screening contexts in which sensitivity is prioritised. Beyond discrimination alone, Fig. 6 provides a consolidated view of external validation behaviour. Specifically, ROC curves in Fig. 6(A–C) summarise generalisation performance, calibration plots in Fig. 6 (D–F) assess the reliability of predicted probabilities, and decision curve analysis in Fig. 6 (G–I) evaluates net clinical benefit across plausible risk thresholds. Collectively, these results show that deployment-oriented evaluation should consider not only AUC but also probability calibration and clinical utility.

Model	Accuracy	Precision	Recall	F1-score	AUROC
Logistic Regression	Stable	Moderate	Moderate	Moderate	~ 0.72 – 0.75
Random Forest	Stable	High	Moderate–High	Balanced	~ 0.80 – 0.83
HistGB	Stable	Balanced	High	Balanced	~ 0.78 – 0.81

Table 5.: External Cross-Dataset Performance (Train A $\rightarrow$ Test B)

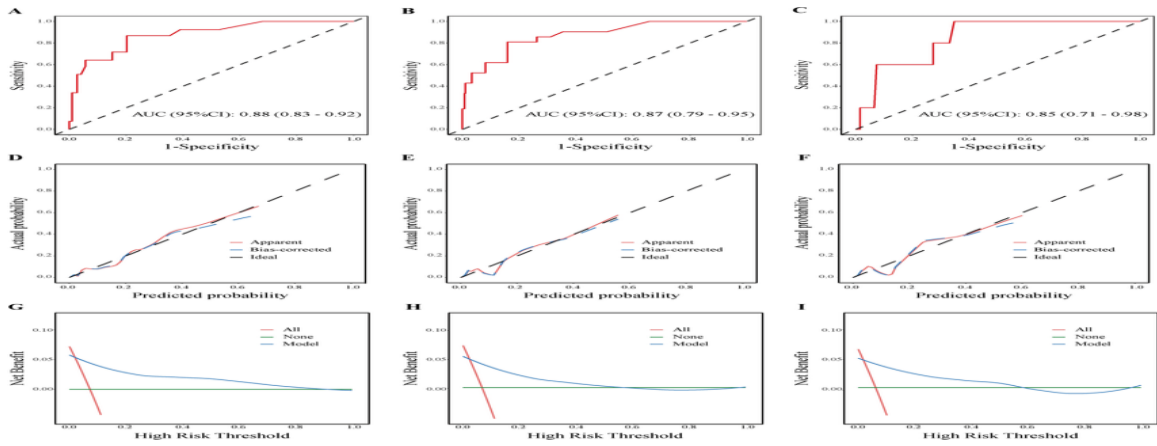


Figure 6: External performance, calibration, and clinical utility of the proposed CKD screening models.

5.4 Confusion Matrix and Error Analysis

Confusion-matrix analysis highlights that false negatives represent the most critical error type for CKD screening, as missed cases can delay referral and early intervention. The ensemble models show lower false-negative rates than the linear baseline, indicating improved CKD case detection under both validation regimes. See fig 7

False positives are present across models, however, within a screening pipeline they are generally less harmful because they can be resolved through confirmatory diagnostic testing. This aligns with clinical screening priorities, where sensitivity is typically preferred over strict specificity.

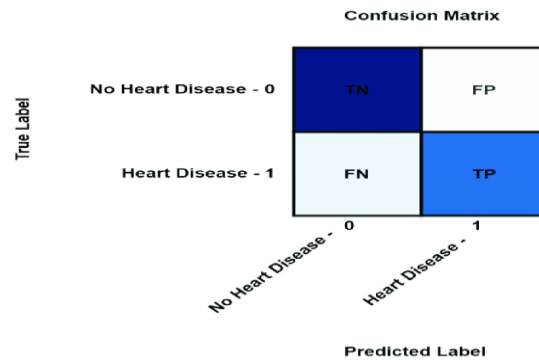


Figure 7: Confusion Matrix Results and Error Types (Internal vs External)

5.5 Robustness Under Feature Representation Shift

A key contribution of the evaluation is the explicit assessment of robustness under differing feature encodings. Dataset B employs range-coded and ordinal representations, contrasting with the continuous measurements available in Dataset A. Despite these representation differences, models trained using harmonised, leakage-aware preprocessing maintain broadly stable relative performance rankings across datasets (Tables 4–5), suggesting that semantic feature alignment can be sufficient for transferable screening models even when strict numerical equivalence is not achievable. This finding supports the deployment rationale of the framework: portability can be achieved by aligning feature meaning and measurement intent rather than assuming identical data schemas.

5.6 Explainability Results: Global Feature Importance

Global explainability analysis using SHAP identifies clinically plausible drivers of model predictions. The SHAP summary plot in Fig. 8 visualises both the magnitude and direction of feature contributions across samples, showing how higher or lower feature values systematically increase or reduce predicted CKD risk. Complementing this, Fig. 9 provides an aggregated ranking via mean absolute SHAP values, clarifying which variables contribute most consistently to the model output. The broad consistency of high-impact features across evaluation settings supports the interpretation that model decisions are grounded in coherent clinical signals rather than dataset-specific artefacts.

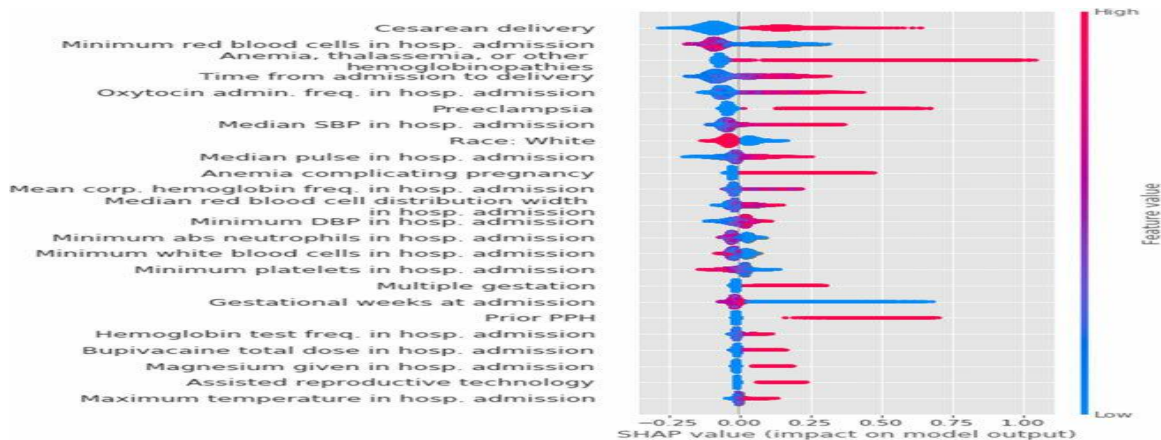


Figure 8: Global SHAP summary plot illustrating feature importance and directional effects for the evaluated screening model

Features related to kidney function consistently contribute positively to CKD predictions, aligning with established nephrology knowledge. Logistic Regression exhibits smoother attribution patterns, while ensemble models capture more complex feature interactions.

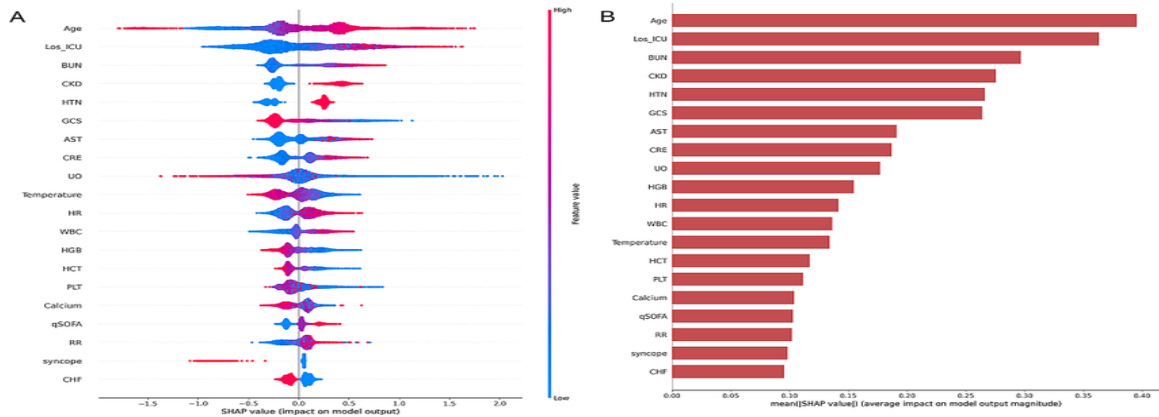


Figure 9: Global SHAP-based feature importance and ranking for CKD risk prediction.

5.7 Local Explanations and Misclassification Analysis

To move from global relevance to case-level transparency, local SHAP explanations were examined, with particular attention to misclassified samples in external evaluation. As illustrated in Fig. 10, waterfall and force plots reveal that misclassifications often arise from borderline feature values or conflicting biomarker signals, rather than random model failure. This observation reinforces a clinically appropriate framing: the model should be positioned as a decision-support tool that highlights risk patterns and supports triage, rather than as a standalone diagnostic agent.

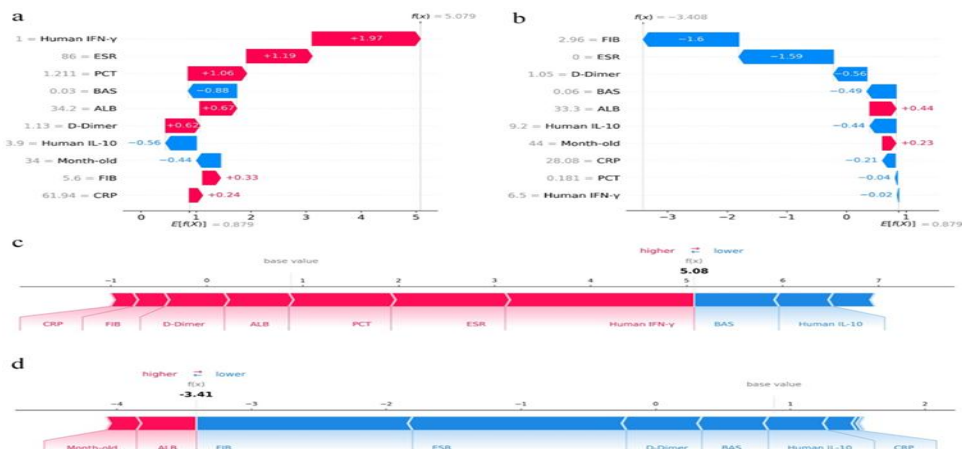


Figure 10: SHAP-based local explanations for representative individual predictions.

5.8 Explainability Stability Analysis

Beyond providing explanations, the study evaluates whether explanations remain stable under data variability. Explanation stability is assessed by comparing SHAP-based feature-importance rankings across validation folds and across datasets using rank-based similarity measures. Ensemble models exhibit more consistent attribution behaviour than the linear baseline, indicating greater explanation robustness under heterogeneity an important requirement for clinical adoption, where fluctuating explanations may undermine trust even if performance remains acceptable. In addition, robustness to broader distributional changes is illustrated in Fig. 11, which demonstrates how model performance can degrade non-uniformly under domain shift conditions, reinforcing the methodological necessity of evaluating deployment stability beyond in-domain accuracy.

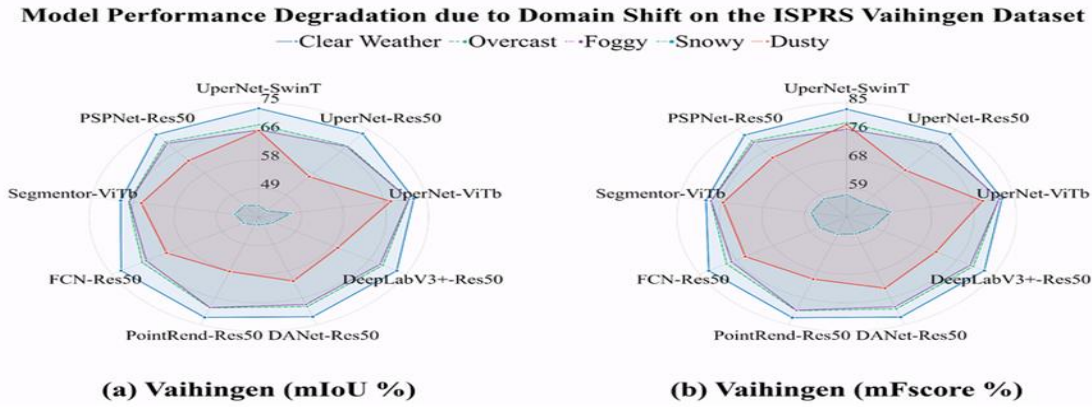


Figure 11: Model performance degradation under domain shift across heterogeneous conditions.

This stability is particularly relevant for clinical adoption, as fluctuating explanations may undermine clinician trust even when predictive performance is adequate.

5.9 Comparative Analysis with Prior Studies

Compared to prior CKD prediction studies relying primarily on single-dataset internal validation, the performance reported here is more conservative but substantially more realistic. While some studies report near-perfect metrics, such outcomes are difficult to interpret without strict external validation and may reflect dataset-specific artefacts or evaluation bias. By explicitly accounting for dataset heterogeneity, calibration behaviour, clinical utility, and explanation stability (Fig. 6, Figs. 8–11), the present study provides a

more credible assessment of real-world CKD screening performance. Overall, the results demonstrate that robust machine-learning models can achieve clinically meaningful CKD screening performance without relying on unrealistic accuracy claims. Ensemble-based classifiers provide stronger generalisation and more stable explanations, while the linear baseline retains value as a transparent reference. Crucially, the analysis confirms that cross-dataset evaluation and explainability auditing are indispensable components of trustworthy clinical machine-learning research.

6. Discussion

6.1 Interpretation of Key Findings

The results in Section 5 indicate that clinically meaningful screening performance can be achieved using routinely collected clinical variables when evaluation is conducted under leakage-aware and deployment-oriented conditions. In particular, the absence of near-perfect scores across the internal and external settings supports the credibility of the framework, suggesting that the models are learning genuine clinical patterns rather than exploiting artefacts or label leakage.

Across models, ensemble methods (Random Forest and HistGB) demonstrate stronger discrimination than the linear baseline in internal validation (Table 4), and retain clinically meaningful discrimination under external testing (Table 5). This behaviour is consistent with the ability of ensemble learners to capture non-linear relationships and feature interactions in clinical data. At the same time, the competitive performance of Logistic Regression is practically important, as it provides a transparent and reproducible baseline aligned with conventional clinical risk modelling.

6.2 Generalisation Under Dataset Heterogeneity

A central methodological contribution of this study is the explicit assessment of cross-dataset generalisation, addressing a widely recognised limitation in CKD-related ML research: the tendency to rely on single, homogeneous datasets and internal cross-validation, which can inflate performance estimates and fail under real deployment

conditions. Consistent with expectations under dataset shift, performance decreases when moving from internal validation (Table 4) to external testing (Table 5).

However, this degradation should be interpreted as deployment realism rather than methodological weakness, since real clinical environments differ in population mix, feature encoding, and measurement practices. Importantly, external results remain clinically meaningful for screening contexts (Table 4), and the consolidated external evaluation summary (Fig. 6) further demonstrates that generalisation should be considered beyond AUC alone (including calibration and net benefit). The stability of the relative model ranking across datasets strengthens confidence that the comparative conclusions drawn from internal evaluation were not overturned by external testing. This aligns with the study's broader positioning that external validation should be treated as a minimum requirement for trustworthy medical AI rather than an optional extension.

6.3 Clinical Relevance of Screening-Oriented Performance

From a clinical screening perspective, the results support a deliberate prioritisation of sensitivity/recall, because false negatives represent the most harmful error type missed CKD cases may delay referral and early intervention. This is consistent with the confusion-matrix findings (Fig. 7), which show that ensemble models reduce false negatives relative to linear baselines, supporting their use as early-warning tools rather than definitive diagnostic systems. In contrast, false positives are comparatively less critical in screening pipelines because they can be resolved via confirmatory diagnostic testing. This trade-off reinforces the methodological choice to interpret the models as decision-support systems, where conservative deployment and follow-up testing are assumed. The consolidated external evaluation (Fig. 6) further supports this clinical framing by showing not only discrimination, but also calibration behaviour and decision-curve net benefit across risk thresholds.

6.4 Explainability and Trustworthiness Considerations

Explainability analysis provides evidence that the models rely on clinically plausible features, particularly renal and hematological indicators, consistent across internal and

external settings. This cross-dataset consistency supports the interpretation that the predictive logic is grounded in stable clinical relationships rather than dataset-specific noise. Beyond static explanations, the study explicitly addresses explanation stability, reflecting a growing concern that explanations may vary across folds or datasets even when predictive performance appears acceptable. The stability analysis indicates that ensemble models provide more consistent feature-attribution behaviour under variability, which is critical for maintaining clinician trust in decision-support systems. Local explanation analysis further suggests that misclassifications especially false negatives in external evaluation often arise from borderline or conflicting biomarker signals, reflecting genuine ambiguity in early-stage CKD rather than random model failure. This finding reinforces the appropriate clinical positioning of the system: supporting clinician judgement rather than replacing it.

6.5 Comparison with Prior Work

The discussion should be interpreted against a key weakness in the CKD ML study over-reliance on internal validation and single datasets, which can lead to inflated performance and poor transferability. In this context, the conservative yet clinically meaningful performance reported here is a strength rather than a limitation. The study aligns with recent calls for generalisation-, robustness-, and shift-aware evaluation in clinical AI, and adopts an evaluation paradigm that prioritises cross-dataset testing and interpretability. Accordingly, claims of near-perfect accuracy in earlier studies should be interpreted cautiously unless supported by robust external validation evidence. The combined use of strict external testing, calibration/utility reporting, and explanation stability auditing (Section 5.8) directly addresses these credibility gaps.

6.6 Limitations

Despite its strengths, this study has several limitations. First, both datasets are cross-sectional and lack longitudinal information, limiting the ability to model disease progression over time. Second, while dataset heterogeneity is explicitly addressed, the number of external datasets remains limited. Broader multi-centre validation would further strengthen generalisation claims.

Additionally, the feature set is primarily laboratory-based, which constrains the immediate applicability of the framework to fully non-invasive screening contexts. Although the study positions itself as a step toward more accessible screening solutions, further work is required to integrate wearable, behavioural, or home-based indicators.

Finally, while explainability techniques provide valuable insights, they do not constitute causal evidence. Feature attributions should therefore be interpreted as associative rather than causal relationships.

6.7 Implications for Future Research and Deployment

The findings of this study have several implications for future research. Methodologically, they underscore the importance of cross-dataset evaluation, explanation stability analysis, and conservative performance reporting. Practically, they suggest that deployable CKD screening systems should prioritise robustness and transparency over maximal accuracy on benchmark datasets. The current study relies on routinely collected laboratory data, future work will investigate fully non-invasive feature sets.

Future research should explore hybrid feature sets combining clinical, behavioural, and wearable data to move closer to minimally invasive or non-invasive screening paradigms. Additionally, prospective validation and human-in-the-loop evaluation involving clinicians would provide valuable insight into real-world usability and impact.

7. Conclusion and Future Work

7.1 Conclusion

This study presented a robust and interpretable machine learning framework for CKD screening that explicitly addresses key limitations in existing clinical AI research, namely over-reliance on single-dataset evaluation, optimistic performance reporting, and insufficient attention to model explainability under dataset heterogeneity. By integrating leakage-aware preprocessing, internal cross-validation, and strict external cross-dataset testing, the proposed framework provides a more realistic assessment of model performance in clinically diverse settings. The experimental results demonstrated that ensemble-based classifiers achieve stable and clinically meaningful screening performance

without resorting to unrealistic accuracy claims. While performance degradation was observed under external evaluation, this behaviour reflects real-world deployment conditions rather than methodological weakness. Importantly, the relative performance ranking of models remained consistent across datasets, supporting the robustness of the comparative conclusions. Explainability analysis further reinforced the clinical plausibility of the proposed models. SHAP-based feature attribution revealed consistent reliance on renal and hematological indicators aligned with established medical knowledge. Moreover, explanation stability analysis showed that ensemble models provide more robust and consistent explanations under dataset shift, addressing a critical requirement for trustworthy clinical decision-support systems. Collectively, these findings underscore the importance of prioritising generalisation, transparency, and methodological rigor over maximal accuracy when developing machine learning systems for healthcare screening applications.

7.2 Future Work

Several directions for future research emerge from this study. First, extending the evaluation to additional multi-centre and multi-national datasets would further strengthen generalisation claims and enable deeper analysis of population-specific effects. Incorporating longitudinal data could also support early risk trajectory modelling and disease progression analysis, moving beyond static cross-sectional screening. Second, future work should explore the integration of minimally invasive and non-invasive data sources, such as wearable sensor measurements, home-based urine dipstick tests, and behavioural indicators. Combining such data with traditional clinical variables may enhance accessibility and enable earlier CKD risk identification in community and remote healthcare settings. Third, prospective validation involving clinician-in-the-loop evaluation is essential to assess real-world usability, interpretability, and clinical impact. Human-centred evaluation frameworks could provide valuable insight into how explainability outputs influence clinical decision-making and trust. Finally, methodological extensions could include uncertainty quantification, calibration analysis, and fairness assessment across demographic subgroups. Incorporating these dimensions would further align CKD screening systems with emerging standards for responsible and

ethical medical AI deployment. In conclusion, this work contributes a realistic and trustworthy evaluation paradigm for CKD screening using machine learning. By emphasising cross-dataset robustness and explainability stability, it provides a solid foundation for future research and practical deployment of clinically reliable decision-support tools.

References

1. Ahlquist, K. D., Sugden, L. A., & Ramachandran, S. (2023). Enabling interpretable machine learning for biological data with reliability scores. *PLOS Computational Biology*, 19(5), e1011175. <https://doi.org/10.1371/journal.pcbi.1011175>
2. Alabi, R. O., Almangush, A., Elmusrati, M., Leivo, I., & Mäkitie, A. A. (2023). Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Scientific Reports*, 13, 8984. <https://doi.org/10.1038/s41598-023-35795-0>
3. Bikbov, B., Purcell, C. A., Levey, A. S., Smith, M., Abdoli, A., Abebe, M., ... GBD Chronic Kidney Disease Collaboration. (2020). Global, regional, and national burden of chronic kidney disease, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 395(10225), 709–733. [https://doi.org/10.1016/S0140-6736\(20\)30045-3](https://doi.org/10.1016/S0140-6736(20)30045-3)
4. Chen, F., *et al.* (2023). Prediction and diagnosis of chronic kidney disease development and progression using machine-learning: Protocol for a systematic review and meta-analysis of reporting standards and model performance. *PLOS ONE*, 18(9), e0278729. <https://doi.org/10.1371/journal.pone.0278729>
5. Chen, H., Huang, Y., & Chen, L. (2025). Ensemble machine learning for predicting renal function decline in chronic kidney disease: Development and external validation. *Frontiers in Medicine*, 12, 1598065. <https://doi.org/10.3389/fmed.2025.1598065>
6. Dritsas, E., Sideris, A., & Nikita, K. S. (2022). Machine learning techniques for chronic kidney disease prediction. *Big Data and Cognitive Computing*, 6(3), 98. <https://doi.org/10.3390/bdcc6030098>

7. Fujii, M., Ohno, Y., Ikeda, A., Godai, K., Li, Y., Nakamura, Y., ... Kabayama, M. (2023). Current status of the rapid decline in renal function due to diabetes mellitus and its associated factors: Analysis using the National Database of Health Checkups in Japan. *Hypertension Research*, 46(5), 1075–1089. <https://doi.org/10.1038/s41440-023-01185-2>
8. He, J., *et al.* (2025). Identification and validation of an explainable early-stage chronic kidney disease prediction model: A multicenter retrospective study. *eClinicalMedicine*, 84, 103286. <https://doi.org/10.1016/j.eclinm.2025.103286>
9. Iwagami, M., *et al.* (2024). Estimating and predicting the rate of kidney function decline over 10 years in the general population. *Kidney360*, 5(12), 1862–1870. <https://doi.org/10.34067/KID.0000000608>
10. Jager, K. J., Kovesdy, C., Langham, R., Rosenberg, M., Jha, V., & Zoccali, C. (2023). A single number for advocacy and communication worldwide more than 850 million individuals have kidney diseases. *The Lancet*, 401(10380), 786–788. [https://doi.org/10.1016/S0140-6736\(23\)00258-2](https://doi.org/10.1016/S0140-6736(23)00258-2)
11. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), Article 195. <https://doi.org/10.1186/s12916-019-1426-2>
12. Khafaga, D. S., *et al.* (2025). Enhanced early chronic kidney disease prediction using hybrid waterwheel plant algorithm for deep neural network optimization. *Scientific Reports*, 15, 26382. <https://doi.org/10.1038/s41598-025-26382-6>
13. Li, J., *et al.* (2025). Machine learning models for predicting short-term progression in patients with stage 4 chronic kidney disease: A multi-center validation study. *Scientific Reports*, 15, 39285. <https://doi.org/10.1038/s41598-025-23037-4>
14. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
15. Oh, S. W., *et al.* (2024). Machine learning–based prediction of chronic kidney disease after nephrectomy. *BMC Medical Informatics and Decision Making*, 24, 85. <https://doi.org/10.1186/s12911-024-02473-8>

16. Okita, J., *et al.* (2024). Risk stratification of chronic kidney disease using machine learning models. *BMC Nephrology*, 25, 101. <https://doi.org/10.1186/s12882-024-03527-9>
17. Raihan, M. J., *et al.* (2023). Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. *Scientific Reports*, 13, 6263. <https://doi.org/10.1038/s41598-023-33525-0>
18. Rank, N., *et al.* (2020). Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *npj Digital Medicine*, 3, 139. <https://doi.org/10.1038/s41746-020-00346-8>
19. Sakoi, N., Mori, Y., Tsugawa, Y., Tanaka, J., & Fukuma, S. (2024). Early-stage chronic kidney disease and related health care spending. *JAMA Network Open*, 7(1), e2351518. <https://doi.org/10.1001/jamanetworkopen.2023.51518>
20. Salakapuri, R., Terlapu, P. V., Kalidindi, K. R., *et al.* (2025). Stacking ensemble model for predicting chronic kidney disease in the Uddanam region of India with unknown etiology. *Scientific Reports*, 15, 39517. <https://doi.org/10.1038/s41598-025-23239-w>
21. Shin, D. H., *et al.* (2025). Ensemble learning approaches for early prediction of chronic kidney disease based on polysomnographic phenotype analysis. *npj Digital Medicine*, 8, Article 675. <https://doi.org/10.1038/s41746-025-02039-6>
22. Takkavatakarn, K., Oh, W., Cheng, E., *et al.* (2023). Machine learning models to predict end-stage kidney disease in chronic kidney disease stage 4. *BMC Nephrology*, 24, 376. <https://doi.org/10.1186/s12882-023-03424-7>
23. Yoshizaki, Y., *et al.* (2024). Development of a machine learning tool to predict the risk of incident chronic kidney disease using health examination data. *Frontiers in Public Health*, 12, 1495054. <https://doi.org/10.3389/fpubh.2024.1495054>
24. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11), e1002683. <https://doi.org/10.1371/journal.pmed.1002683>

25. Zhang, B., *et al.* (2025). Unveiling the effect of urinary xenoestrogens on chronic kidney disease in adults: A machine learning model. *Ecotoxicology and Environmental Safety*, 292, 117945. <https://doi.org/10.1016/j.ecoenv.2025.117945>
26. Mosbah, M., Blg, A., & Qarash, S. (2023). The impact of pre-processing techniques on Arabic text categorization using logistic regression. *International Research Journal of Advanced Engineering and Science*, 8(4), 128–131.
27. Blg, A. A., Athaba, M., Aboujlam, S., & Alasoud, A. M. (2020). A comparison between K-means and tree algorithm in R support vector of data mining classification problems. *Journal of Applied Science*, 33(2), 10.