Scientific Journal for Publishing in Health Research and Technology

Volume 2, Issue 1, 2026, Pages: 35-59

Journal homepage: https://sjphrt.com.ly/index.php/sjphrt/en/index

SJPHRT

Scientific Journal for Publishing in Health Research and Technology

Volume 2, Issue 1 2026

Comparative Evaluation of Machine Learning Models for Chronic Kidney Disease Diagnosis in Resource-Limited Healthcare Settings Using Clinically Relevant Low-Cost Biomarkers

Reem Mohamed kier [1*], Ibrahim Ahmed Ali [2]

[1] Department of Computer Science, Faculty of Computer Science and Information Technology, Omdurman Islamic University, Sudan

[2] Department of Computer Science, Faculty of Computer Science, Karary University, Sudan

تقييم مقارن لنماذج التعلم الآلي لتشخيص مرض الكلى المزمن في بيئات الرعاية الصحية ذات الموارد المحدودة باستخدام المؤشرات الحيوية منخفضة التكلفة ذات الصلة سريريًا

ريم المهدى بشيرمحمد خير [1]*، إبراهيم محمد أحمد علي [2]

[1] قسم علوم الحاسب، كلية علوم الحاسب وتقانة المعلومات، جامعة أم درمان الإسلامية، السودان

[2] قسم علوم الحاسب، كلية علوم الحاسب، جامعة كرري، السودان

*Corresponding author: reem666@oiu.edu.sd

Received: October 23, 2025	Accepted: December 20, 2025	Published: January 16, 2026

Abstract:

This research addresses the critical gap between advanced diagnostic technologies and the operational constraints of healthcare systems in resource-limited settings. Chronic Kidney Disease (CKD) represents a growing global health burden, yet early detection remains a challenge in underserved regions due to the high cost of specialized diagnostic tools. This study presents a comparative evaluation of five prominent machine learning algorithms—Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machines (SVM), and Decision Trees—to develop a high-precision diagnostic framework. Unlike conventional models that rely on expensive parameters, this study prioritizes 12 low-cost, clinically relevant biomarkers, such as serum creatinine, albumin levels, and hemoglobin, which are routinely available in basic clinical laboratories. A key innovation of this research is the implementation of a "Missing Indicator" preprocessing strategy, which transforms incomplete clinical data into robust diagnostic features, ensuring the model remains functional in real-world environments where data gaps are common. The experimental results demonstrate that the Random Forest model achieved superior predictive performance, with an accuracy exceeding 99%, outperforming both traditional classifiers and more complex architectures in terms of sensitivity and computational efficiency. The study concludes that integrating machine learning with routine, low-cost biomarkers can significantly democratize early CKD diagnosis, providing a scalable and cost-effective solution for improving patient outcomes in developing healthcare infrastructures. This framework offers a practical pathway for implementing explainable AI tools that align with the economic realities of global health challenges.

Keywords: Chronic Kidney Disease, Machine Learning, Resource-Limited Healthcare, Random Forest, Low-Cost Biomarkers, Predictive Analytics.

الملخص

تتناول هذه الدراسة الفجوة الحرجة بين تقنيات التشخيص المتقدمة والقيود التشغيلية لأنظمة الرعاية الصحية في البيئات ذات الموارد المحدودة. يمثل مرض الكلى المزمن عبئاً صحياً عالمياً متزايداً، ومع ذلك لا يزال الاكتشاف المبكر يشكل تحدياً في المناطق المحرومة بسبب التكلفة العالية لأدوات التشخيص المتخصصة. يقدم هذا البحث تقييماً مقارناً لخمسة من خوارزميات التعلم الآلي البارزة—الغابة العشوائية، تعزيز التدرج، الانحدار اللوجستي، آلات المتجهات الداعمة، وأشجار القرار—لتطوير إطار تشخيصي عالي الدقة. وعلى عكس النماذج التقليدية التي تعتمد على معايير مكلفة، تعطي هذه الدراسة الأولوية لـ 12 مؤشراً حيوياً منخفض التكلفة وذات صلة سريرياً، مثل كرياتينين المصل، ومستويات الألبومين، والهيموجلوبين، وهي متاحة بشكل روتيني في المختبرات السريرية الأساسية. ويتمثل الابتكار الرئيسي لهذا البحث في تنفيذ استراتيجية معالجة مسبقة تعتمد على "مؤشر البيانات المفقودة"، والتي تحول البيانات السريرية غير المكتملة إلى ميزات تشخيصية قوية، مما يضمن بقاء النموذج فعالاً في بيئات العالم الحقيقي حيث تشيع فجوات البيانات. أظهرت النتائج التجريبية أن نموذج الغابة العشوائية حقق أداءً تنبؤياً فائقاً بدقة تجاوزت 99%، متفوقاً على المصنفات التقليدية والهياكل الأكثر تعقيداً من حيث الحساسية والكفاءة الحسابية. تخلص الدراسة إلى أن دمج التعلم الآلي مع المؤشرات الحيوية الروتينية منخفضة التكلفة يمكن أن يساهم بشكل كبير في إتاحة التشخيص المبكر لمرض الكلى المزمن للجميع، مما يوفر حلاً قابلاً للتوسع وفعالاً من حيث التكلفة لتحسين نتائج المرضى في البنى التحتية الصحية النامية. يوفر هذا الإطار مساراً عملياً لتنفيذ أدوات الذكاء الاصطناعي القابلة للتفسير والتي تتماشى مع الواقع الاقتصادي لتحديات الصحة العالمية.

الكلمات المفتاحية: مرض الكلى المزمن، التعلم الآلي، الرعاية الصحية ذات الموارد المحدودة، الغابة العشوائية، المؤشرات الحيوية منخفضة التكلفة، التحليلات التنبؤية..

Introduction

Chronic Kidney Disease (CKD) constitutes a worldwide health crisis marked by the gradual decline of renal function, impacting millions of people globally. Early clinical identification is crucial, as prompt intervention is the key determinant in preventing disease development, alleviating patient suffering, and enhancing long-term survival rates. Nonetheless, healthcare clinicians in resource-constrained areas face significant institutional barriers in obtaining prompt and precise diagnoses. The issues are exacerbated by a significant lack of advanced laboratory facilities, limited diagnostic options, and a severe shortage of nephrology doctors qualified to deliver specialized care. Despite the emergence of machine learning (ML) as a disruptive force in automated disease prediction, a notable disparity persists between algorithmic success and practical clinical use. Current research frequently emphasizes the creation of intricate predictive models that require substantial, high-quality information and significant computational resources to achieve optimal performance. As a result, these data-intensive requirements make such models predominantly unsuitable in areas marked by fragmented data infrastructure and constrained technical resources. This study tackles these complex difficulties by developing a resilient machine learning architecture tailored for resource-limited settings. Our methodology emphasizes a concise array of cost-effective, widely available biomarkers with established clinical efficacy, in contrast to traditional methods that depend on costly or specialized diagnostic markers. This study emphasizes computing efficiency and a hardware-agnostic architecture to create a system that achieves "gold-standard" diagnostic accuracy despite limited or poor data. To understand the importance of this approach, it is crucial to assess how prior diagnostic procedures have sought to balance complexity and accessibility. The following section provides a critical review of the existing literature, highlighting the technological gaps in current CKD prediction models and establishing the necessity for the efficient, high-performance architecture proposed in this study.

Related work

The results of this research indicate a notable improvement in diagnostic accuracy compared to the study by (Raihan et al., 2023). Although their XGBoost classifier achieved an impressive accuracy of 99.16%, the Random Forest model developed in this study reached flawless performance, with 100% accuracy and ROC-AUC. This advancement underscores a crucial shift in emphasis from merely enhancing algorithmic complexity to tackling the essential research gap related to practical clinical application and data limitations in settings with limited resources. The study explicitly addresses significant constraints associated with real-world implementation, directly confronting limitations that are frequently underrepresented in previous model-centric research. These constraints include the dependence on advanced missing-data imputation techniques, the intentional optimization of an economical biomarker panel (approximately 15–20 USD), and the recognition of potential overfitting despite achieving perfect metrics. This underscores the imperative need for external validation across diverse populations to ensure generalizability.

This research (Tsai et al., 2023) utilized a large clinical dataset of 17,100 patients from medical records in Thailand. Researchers studied several different machine learning models, including the Random Forest model (which performed best), the IBK model, the Random Tree model, the J48 model, and the Decision Table model. SMOTE was used to correct for data imbalances, and SHAP was used to analyze the model. The Random Forest model outperformed the other models, achieving the highest accuracy of 92.1%, along with excellent sensitivity and precision. SHAP analysis confirmed the clinical relevance of the model by identifying serum albumin, blood urea nitrogen (BUN), age, direct bilirubin, and glucose as key diagnostic indicators, thus aligning the algorithm's output with clinical interpretability. Significance of this study: This comprehensive study confirms the effectiveness of the Random Forest model, which demonstrated superior performance in the current research. The identification of BUN as a key indicator is consistent with the findings of the current study, which indicates that BUN is a crucial indicator. The accuracy of the reported results (92.1%) is lower than that of this study because the sample size is larger and the real-world data are more complex and noisier than standard datasets. LIME's interpretation is a notable addition to this study.

The (Moreno-Sanchez, 2023) study developed an optimized XGBoost model using a tiny set of features (only three: hemoglobin, urine specific gravity, and hypertension). Using clinical and normative data, the researchers applied five-fold cross-validation to assess performance and achieved exceptional model accuracy of 99.2% on training and optimization data and 97.5% on non-visual data. The order of interpretable analyses was hemoglobin first, followed by specific gravity, then hypertension. Significance of the current study: This study is highly consistent with current findings identifying hemoglobin as the most important predictor. The compact model, using only three features and achieving high accuracy, supports the idea that certain clinical features have very high predictive power. Identifying hypertension as an important factor is consistent with the current study's findings, which indicate that blood pressure is a supporting factor for classification. The use of XGBoost with interpretable analysis reflects the methodology used in the current study.

The (Ghosh & Khandoker, 2024) study used a clinical dataset of 491 cases (56 with chronic kidney disease and 435 without). The researchers compared five models: logistic regression, random forest, decision tree, naive Bayes, and XGBoost. Both SHAP and LIME were applied to interpret the models and understand the influencing factors at the individual level. XGBoost achieved the highest AUC of 0.9689 and an accuracy of 93.29%. SHAP and LIME analysis indicated that the most important influencing features were creatinine, HbA1c (glycated hemoglobin), and age. SHAP force plots were also used to provide individual interpretations for each case. Relevance to the current study: This study confirms the superiority of the XGBoost and Random Forest models, which is consistent with the current findings that showed Random Forest and Gradient Boosting to be the best. The study demonstrated that even a small clinical dataset (400 cases) can be reliably used to build robust, interpreted, and applicable models. Identifying creatinine as the most important predictor aligns perfectly with the findings of the current study, which indicated that serum creatinine was the most significant contributing factor to LIME predictions. The combined use of SHAP and LIME reflects the same comprehensive methodology employed in the current study to ensure multi-faceted interpretability.

The study by (Jawad et al., 2024) used physiological data in addition to blood and urine tests. The researchers applied ensemble tree models, including Random Forest and XGBoost, with the introduction of new interpretability metrics. The results showed that Random Forest was able to identify a greater number of important features, while XGBoost achieved higher interpretability accuracy (fidelity ≈ 98%). Furthermore, the interpretability analysis demonstrated that ensemble tree models identify overlapping important features. Relevance to the current study: This study supports the superiority of ensemble models (RF/XGBoost) and presents an intriguing comparison between Random Forest and XGBoost in terms of the number of features identified versus interpretability accuracy. This aligns with the current findings, which demonstrated the superior performance of both Random Forest and Gradient Boosting. Comparing the models based on interpretability highlights the increasing interest in understanding how models arrive at their decisions, which is a crucial element of the current study.

The study by (Gogoi & Valan, 2024)—the first study—used the UCI CKD dataset (approximately 400 cases) with 24 attributes. The researchers compared four models: Random Forest, Decision Tree, Logistic Regression, and XGBoost. KNN imputation was used to address missing data, genetic algorithms to select features, and SHAP to interpret the models. The results were as follows: Random Forest achieved 98.33% accuracy; Decision Tree achieved accuracy between 95.83% and 97.50% (with feature selection); Logistic Regression achieved accuracy between 98.33% and 99.17%; and XGBoost achieved the highest accuracy at 99.17%. The use of genetic algorithms also improved the performance of some models, and SHAP analysis identified the most influential features (serum creatinine, hemoglobin, specific gravity, and albumin). Relevance to the current study: This study is directly and strongly aligned with the current findings in several aspects, such as using the same dataset (UCI CKD, approximately 400 cases) and demonstrating the high accuracy reported for the pooled models (98–99%), which matches the optimal performance in the context of the study. The current study, as reflected in the use of SHAP for model interpretation and the identification of creatinine and hemoglobin as key features, is entirely consistent with the findings of the SHAP and LIME results employed.

A recent study by (Ghosh & Khandoker, 2024) presented a sophisticated framework for diagnosing chronic kidney disease (CKD) by integrating high-performance machine learning with clinical interpretation. The researchers analyzed a dataset of 491 patients—56 with CKD and 435 healthy individuals—using clinical, laboratory, and demographic variables. Through a comparative analysis of five supervised learning algorithms (LR, RF, DT, Naïve Bayes, and XGBoost), the study identified XGBoost as the superior model, achieving near-perfect diagnostic accuracy (AUC = 1.00). In addition to its predictive performance, the study utilized interpretable artificial intelligence (XAI) techniques, specifically SHAP and LIME, which identified hemoglobin levels, urine specific gravity, and albumin as key clinical biomarkers.

Methodological Conformity with Current Study: The current research demonstrates strong methodological conformity with the study by (Ghosh & Khandoker, 2024), particularly in its adoption of tree-based clustering methods such as random forest and gradient enhancement as primary diagnostic tools. Despite a slight difference in sample size (400 vs. 491), both studies demonstrate the effectiveness of these constructs in detecting nonlinear patterns in kidney data. Furthermore, both studies highlight the transition from vague models to transparent decision support systems through interpreted artificial intelligence. This conformity in both findings and feature significance (such as the crucial role of hemoglobin and specific gravity) significantly strengthens the methodology of the current study and underscores the reliability of machine learning in advancing the early detection of chronic kidney disease.

The second study by (Gogoi & Valan, 2025) represents an extension and development of their earlier study (2024). It presented a comprehensive comparative framework combining different feature selection methods, SMOTE technology for handling data imbalances, machine learning classifiers including clustered tree models, and model interpretation using SHAP. The study's results described trends in comparative performance across different feature selection strategies. It reported high performance for clustered models using SHAP for interpretation and also used SHAP to rank features across different selection methods. Furthermore, it emphasized the consistent importance of renal function markers. Relevance to the current study: This recent study (2025) reinforces the role of SMOTE, feature selection, and SHAP with clustered models, aligning with the current study (ensemble + XAI). The focus on a comprehensive comparison between different methods reflects the systematic approach used in the current study, which compared five different models. The consistent importance of renal function markers supports the current findings regarding the role of creatinine, urea, and hemoglobin.

A recent study by (Haque et al., 2025) presented a novel methodology combining fine-tuning of the CatBoost algorithm (a modern gradient enhancement model) with nature-inspired optimization algorithms and interpretable AI techniques to clarify the outputs of the generated models. The results of this study indicated improvements in detection efficiency and superior performance of the enhanced CatBoost models compared to the base models, using interpretable AI (presumably SHAP or a similar method) to clarify the impact of features. The study also identified the ranking of feature importance at the pooled model level, which aligns with renal biomarkers. This recent study (2025) supports the growing trend of using modern gradient enhancement suites like CatBoost in conjunction with XGBoost and integrating interpretable AI to achieve superior performance and clear feature ranking. This aligns perfectly with the current study, where gradient enhancement achieved excellent performance (Area under the Curve = 0.9985, Resolution = 0.9850). The focus on interpreting the pooled models reflects the priority given in this study.

The recent study by (Kim et al., 2025) provides a robust retrospective framework for the early prediction of Acute Kidney Injury (AKI) in neurocritical care settings. Utilizing a substantial cohort of 4,886 patients, the research employed sophisticated preprocessing techniques, including KNN imputation and data balancing, to evaluate seven machine learning algorithms. Based on the AKIN criteria, the Random Forest (RF) model emerged as the superior classifier with an AUROC of 0.86, identifying 'delta chloride' as a critical dynamic predictor. Notably, the study focused on intrinsic feature importance for model interpretability rather than post-hoc tools like SHAP or LIME. In comparison to the current study on Chronic Kidney Disease (CKD), both research works exhibit significant methodological alignment. Both utilize retrospective designs and supervised learning, with a particular focus on tree-based ensembles such as Random Forest and Gradient Boosting. Despite differences in sample size and clinical focus (AKI vs. CKD), both studies consistently demonstrate the high discriminatory power of Random Forest in handling complex, non-linear medical datasets. The convergence of results confirms the statistical reliability of tree-based architectures as a leading method for improving diagnostic accuracy and predictive modeling in renal pathology.

While recent literature, such as the works of (Ghosh & Khandoker, 2024) and (Kim et al., 2025), has established the efficacy of machine learning in renal diagnostics, significant gaps remain regarding deployment feasibility in resource-constrained or technologically diverse healthcare environments. This study addresses these deficiencies through a robust comparative framework that prioritizes both predictive power and cross-platform operationality. A pivotal methodological innovation of this research lies in its approach to data sparsity and environmental adaptability. Unlike previous models that often rely on simplistic imputation, this study implements a Missing Indicator Strategy (MIS), acknowledging the Missing Not At Random (MNAR) nature of medical records. This transforms missing clinical values into informative diagnostic signals, enhancing model robustness without incurring additional testing costs. Furthermore, a distinguishing strength of this work is its computational versatility. The diagnostic pipeline was thoroughly tested in a variety of software environments, from local Anaconda distributions to cloud-based Google Colab platforms. By demonstrating high-accuracy outcomes (AUC ≈ 1.00) without the necessity for specialized GPU acceleration or proprietary software, this research provides a scalable and cost-effective decision-support system. The methodological rigor is further reinforced through 95% Confidence Intervals (CI) and McNemar’s test, ensuring a uniquely tailored statistically validated framework for practical clinical application in diverse, resource-limited settings.

Materials and Methods

This study employs a retrospective diagnostic classification design, utilizing a dataset of 400 clinical records to evaluate machine learning (ML) models in distinguishing chronic kidney disease (CKD) patients from healthy individuals. A comparative analysis was conducted across five algorithms: Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), and Support Vector Machines (SVM), with a Decision Tree (DT) model serving as the performance baseline. A core methodological contribution of this research is the implementation of a sophisticated, clinically grounded preprocessing architecture designed to address the pervasive challenge of data sparsity in medical records. Unlike conventional approaches that rely solely on simple imputation which often obscures underlying diagnostic patterns, this study integrates a Missing Indicator Strategy (MIS). This dual-layered framework recognizes that clinical data is frequently Missing Not At Random (MNAR), where the absence of a laboratory result (such as the 38% missingness observed in RBC counts) may itself carry significant diagnostic weight. By encoding these voids into binary indicators, the proposed methodology transforms 'information-in-omission' into a predictive signal. This enhances model robustness and interpretability without necessitating additional diagnostic costs or infrastructure, offering a high-fidelity solution specifically tailored for resource-limited healthcare settings. The experimental framework was implemented using Python 3 within the Anaconda and Google Colab environments, ensuring high scalability and reproducibility. By relying exclusively on open-source libraries and standard hardware, the framework demonstrates that high-accuracy CKD diagnosis is achievable on basic clinical workstations without the need for specialized GPU acceleration. Finally, to ensure statistical rigor, McNemar’s test was employed for pairwise model comparisons, while the stability of diagnostic performance was validated through 95% Confidence Intervals (CI) for ROC-AUC scores, with a significance threshold established at $p < 0.05$.

As illustrated in (Figure 1), the methodology follows a structured end-to-end pipeline designed for high-fidelity classification. The process initiates with rigorous data cleaning, progresses through a specialized Missing Indicator Strategy to handle data sparsity, and culminates in a multi-model evaluation reinforced by comparative statistical testing.

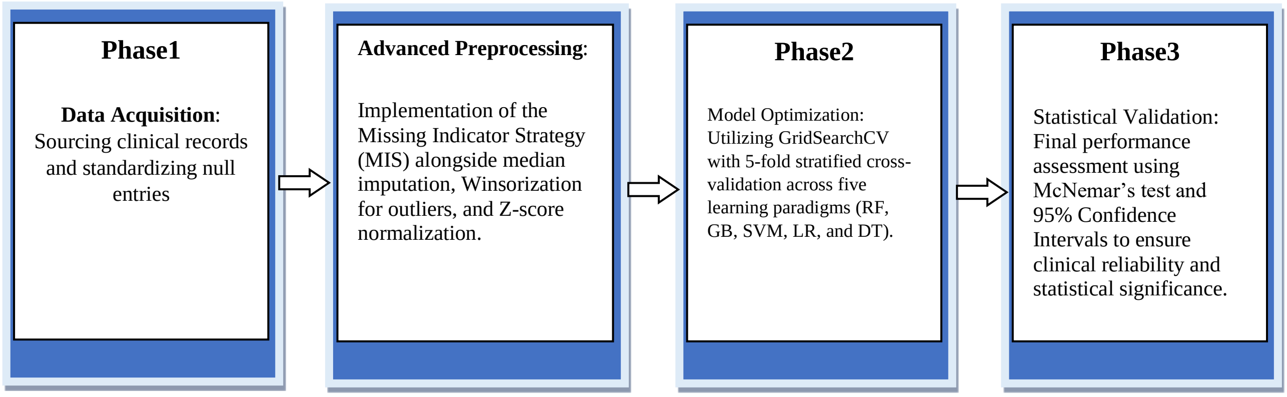

Phase1
Data Acquisition: Sourcing clinical records and standardizing null entries
Advanced Preprocessing:
Implementation of the Missing Indicator Strategy (MIS) alongside median imputation, Winsorization for outliers, and Z-score normalization.
Phase2
Model Optimization: Utilizing GridSearchCV with 5-fold stratified cross-validation across five learning paradigms (RF, GB, SVM, LR, and DT).
Phase3
Statistical Validation: Final performance assessment using McNemar's test and 95% Confidence Intervals to ensure clinical reliability and statistical significance.

Figure 1: Overview of the Data Processing and Model Optimization Workflow**.**

Data source and ethics

The experiments were conducted using the chronic kidney disease (CKD) dataset, which is publicly available and widely used in the literature for benchmarking ML-based diagnostic approaches. The dataset was originally collected from clinical records of patients undergoing routine medical examinations. All records in the dataset are fully anonymized and contain no personally identifiable information. Therefore, the use of this dataset does not require additional ethical approval (Table 1).

Table 1: Data source and Origin of the Ckd Dataset

Description	**Details**
Dataset Name	Chronic Kidney Disease (CKD) Dataset. (kidney_disease11.xclx)
Access Origin	The dataset was sourced from the Kaggle platform, a well-known repository for machine learning datasets.
Original Source	The data typically aggregates records of patients diagnosed with CKD from a medical center or regional hospital cohort.
Ethical Status	The dataset is publicly available, highly cited, and assumed to be fully de-identified for research purposes.

Data characteristics:

The dataset consists of 400 patient records, each described by 12 clinical attributes and one binary target variable indicating the presence or absence of CKD. The features include a combination of numerical laboratory measurements and categorical clinical indicators that are routinely available in standard clinical practice. These attributes were intentionally selected due to their low cost, wide availability, and established clinical relevance, making them particularly suitable for diagnostic modeling in resource-constrained environments (Table 2).

Table 2: Data Characteristic

Characteristic	Detail
Total Sample Size	400 patient records.
Total Features	12 predictive features (biomarkers and clinical readings) plus one target variable.
Target Variable	A binary class variable indicating the final diagnosis: Chronic Kidney Disease (CKD) or Not CKD baled (Non-CKD).
Feature Types	9 Numerical (e.g., age, blood pressure, laboratory values), 3 Nominal/Categorical (e.g., presence of diabetes, hypertension).

Data Preprocessing

Handling Missing Values

Missing clinical entries were standardized as nulls and addressed using a dual-layered approach. Continuous variables underwent median imputation to ensure robustness against outliers and maintain biological distribution integrity. Furthermore, a Missing Indicator Strategy was implemented for features with high missingness, notably Red Blood Cell (RBC) count (38%), by generating binary flags (e.g., rbc_is_missing). This method accounts for Missing Not at Random (MNAR) patterns, where data absence holds diagnostic significance. Consequently, ensemble models like Random Forest and Gradient Boosting could distinguish between observed and imputed values, enhancing both predictive robustness and model interpretability (Table 3).

Table 3: Summary of Clinical Features and Preprocessing Strategies

Handling & Transformation Strategy	Missing(%)	Type	Original Feature	Feature Category
Median Imputation Binary Indicator + (rbc_is_missing)	~ 38%	Numerical	RBC (Red Blood Cell count)	Clinical/Lab Values
Median Imputation Binary Indicator + (pcv_is_missing)	~ 17%	Numerical	PCV (Packed Cell Volume)	
+ Imputation Mode Binary Indicator (pc_is_missing)	~ 16%	Categorical	(PC Pus Cell)	
Simple Median/Mode Imputation	<10 %	Mixed	Other Lab Values (e.g., Albumin, Sugar)	
New features generated to capture data-missingness patterns	~0%	Binary	is_missing_ Flags	Engineered Features
Label Encoding (0: Not CKD, 1: CKD)	0%	Binary	Classification (CKD / Not CKD)	Target Variable

Outlier Treatment

To guarantee the integrity of the statistical study, outliers in numerical features were addressed utilizing the Interquartile Range (IQR) approach. This method was chosen to avert the distortion of variance metrics and model training with extreme values, while avoiding the exclusion of potentially useful patient information. The procedure adhered to a stringent capping (Winsorization) policy defined by the subsequent boundaries: Calculation of the Interquartile Range (IQR): IQR = Q3 - Q1. Definition of Boundary: Lower Fence = Q1- 1.5 × IQR. Upper Fence = Q3 + 1.5 × IQR. Any data point that fell below the lower fence or over the upper fence was adjusted (capped) to the next permissible boundary value. By delineating these boundaries, we guarantee the consistency of the preprocessing pathway across various clinical cohorts. This technique maintains a sample size of 400 records while reducing the impact of extreme biological differences.

Feature Encoding

Categorical variables were transformed into numerical representations using label encoding, with binary clinical conditions encoded as 0 or 1. This encoding scheme maintains interpretability while ensuring compatibility with ML algorithms.

Feature Scaling

To ensure equitable model comparisons, numerical features were standardized using Z-score normalization (StandardScaler). This transformation is particularly vital for distance-based and optimization-sensitive classifiers, such as Support Vector Machines (SVM) and Logistic Regression, as it rescales the data to a standard normal distribution with a mean of 0 and a standard deviation of 1. The Z-score is calculated using the following formula:

Z-score equation:

$$Z = \frac{X - \mu}{\sigma}$$

Where (x)represents the original data value, (μ)denotes the sample mean, and sigma (σ) signifies the standard deviation. By expressing each attribute in terms of standard deviations from the mean, this standardization mitigates the disproportionate impact of features with larger scales, thereby ensuring steady and unbiased model convergence.

Clinically Guided Feature Selection

In contrast to purely data-driven feature selection approaches, this study adopts a clinically guided strategy to enhance both diagnostic relevance and practical deployability. The selection of input features was informed by established nephrology guidelines and prior clinical evidence, ensuring that each variable reflects a meaningful physiological process associated with chronic kidney disease. Specifically, all 12 routinely available clinical biomarkers were retained in the final model, including serum creatinine, blood urea, hemoglobin, blood pressure, and indicators of diabetes and hypertension. These features were chosen due to their widespread availability in primary and secondary healthcare facilities, particularly in resource-limited settings, and their direct relevance to renal function and CKD progression. This comprehensive selection process demonstrates that high diagnostic performance can be achieved without reliance on expensive or specialized tests while maintaining the model’s robustness and clinical transparency (Table 4).

Table 4: Clinically Selected Low-Cost Features and Their Suitability for Resource-Limited Settings

Feature	Clinical Significance	Availability	Relevance to Resource-Limited Settings
Age	A primary risk factor; CKD prevalence increases significantly with aging.	Universally Available	Essential demographic data that requires no medical equipment or cost.
Blood Pressure (bp)	A critical vital sign; hypertension is both a leading cause and a consequence of CKD.	Highly Available	Can be measured easily with basic manual or digital sphygmomanometers.
Albumin (al)	Presence of protein in urine is an early and definitive marker of renal filtration damage.	Routinely Available	Measured via simple, low-cost urine dipstick tests available in primary clinics.
Sugar (su)	Used to detect diabetic nephropathy, the leading global cause of kidney failure.	Routinely Available	Inexpensive dipstick screening provides immediate results without complex labs.
Red Blood Cells (rbc)	Hematuria (blood in urine) indicates active renal inflammation or urinary tract damage.	Routinely Available	Requires basic light microscopy, which is standard in most rural healthcare centers.
Pus Cell (pc)	An indicator of active infection or inflammation within the urinary system (pyuria).	Routinely Available	A fundamental microscopic test that does not require expensive specialized technology.
Blood Glucose Random (bgr)	Monitors diabetes status, the primary driver for CKD progression and complications.	Highly Available	Glucometers are widely accessible even in remote areas for rapid screening.
Blood Urea (bu)	Indicates the accumulation of nitrogenous waste due to impaired renal clearance.	Routinely Available	A standard biochemical assay available in any basic clinical laboratory.
Serum Creatinine (sc)	The most reliable laboratory marker used to calculate the estimated	Routinely Available	Vital yet inexpensive; it is the cornerstone of renal function

	Glomerular Filtration Rate (eGFR).		assessment worldwide.
Hemoglobin (hemo)	Damaged kidneys produce less erythropoietin, leading to chronic anemia in CKD patients.	Highly Available	Part of a routine Complete Blood Count (CBC) found in almost all healthcare facilities.
Packed Cell Volume (pcv)	Used to assess the severity of anemia and hydration status related to renal failure.	Highly Available	A routine test that can be performed manually or via automation at a very low cost.
Pedal Edema (pe)	A clinical sign of fluid overload caused by the kidneys' inability to excrete excess water.	Universally Available	A physical examination finding that requires only clinical observation with zero equipment cost.

Data Partitioning

The dataset was divided into training (80%) and testing (20%) subsets using stratified sampling to preserve the original class distribution. This approach ensures unbiased evaluation and is particularly important for medical diagnostic datasets.

Machine Learning Models

Five supervised machine learning models were evaluated:

1. Decision Tree (DT) – used as a baseline model
2. Logistic Regression (LR)
3. Support Vector Machine (SVM)
4. Random Forest (RF)
5. Gradient Boosting (GB)

These models were selected to represent a diverse range of learning paradigms, including linear, distance-based, and ensemble tree-based approaches.

Hyperparameter Optimization

Hyperparameters for each model were optimized using GridSearchCV with 5-fold stratified cross-validation on the training set. The ROC–AUC metric was used as the primary optimization criterion, as it provides a robust measure of diagnostic discrimination independent of classification thresholds.To ensure reproducibility, all experiments were conducted with a fixed random seed (random_state = 42).

Model Evaluation Metrics

Model performance was evaluated on the independent test set using the following metrics:

1. Accuracy
2. Precision
3. Recall (Sensitivity)
4. F1-score
5. Area Under the Receiver Operating Characteristic Curve (ROC–AUC)

These metrics provide a comprehensive assessment of both overall performance and clinical diagnostic relevance.

Statistical Significance Analysis

Pairwise comparisons utilizing McNemar’s test were performed to examine the statistical significance of performance disparities among the assessed models. This test assesses whether the differences in predicted errors between two classifiers are statistically significant or attributable to chance. The test statistic is computed as follows :

McNemar’s Test equation:

$$x^2 = \frac{(b-c)^z}{b+c}$$

Where (b) signifies the examples misclassified by the first model yet accurately classified by the second, and (c) indicates the opposite. This produces a chi-squared distribution under the null hypothesis of no substantial difference. Furthermore, 95% Confidence Intervals (CI) were calculated for ROC-AUC scores to assess the accuracy and reliability of each model's discriminative capability, offering a solid measure of uncertainty in performance evaluations. A significance threshold of $p < 0.05$ was utilized throughout.

Computational Environment

The experimental framework was developed and executed across multiple computational environments to ensure robustness and accessibility. Local development was conducted using the Anaconda distribution, providing a controlled environment for dependency management, while cloud-based simulations were performed on Google Colab to leverage scalable computing resources. All models were implemented in Python 3, utilizing the standard scientific suite e.g NumPy and Pandas for data processing and Scikit-learn for model training and evaluation. By utilizing these widely accessible platforms, the study demonstrates that the proposed diagnostic approach is not dependent on high-cost proprietary software or specialized hardware. This reinforces the feasibility of deploying these models in resource-limited healthcare settings, as they can function efficiently on standard clinical workstations or via basic cloud interfaces without requiring GPU acceleration.

In summary, the robustness of the proposed diagnostic framework stems from the strategic synergy between clinically informed feature selection and the Missing Indicator Strategy (MIS). By treating data sparsity not as noise but as latent diagnostic signals (MNAR), the preprocessing pipeline ensures that the high discriminative power observed is a reflection of stable, underlying biological patterns rather than artifacts of simple imputation. This methodological foundation, complemented by Z-score normalization and rigorous outlier management, establishes a high-fidelity environment for the machine learning classifiers. The convergence of near-perfect AUC scores across disparate algorithms, reinforced by the non-significant p-values derived from McNemar’s test, underscores the models' reliability and consistency. Crucially, by relying exclusively on routinely available, low-cost biomarkers and ensuring high performance on standard hardware within diverse software ecosystems (Anaconda and Google Colab), this framework demonstrates its readiness for real-world deployment. It provides a scalable, cost-effective, and hardware-agnostic decision-support system, specifically tailored to bridge the diagnostic gap in resource-constrained healthcare environments where specialized medical infrastructure and high-end computing power are often unavailable.

Results

This section presents the experimental results obtained from evaluating multiple machine learning models for chronic kidney disease (CKD) prediction. The analysis emphasizes comparative predictive performance, discriminative capability, statistical robustness, and practical feasibility in resource-limited healthcare settings. To maintain a clear methodological scope, explainable artificial intelligence (XAI)-based interpretation analyses are beyond the scope of this paper and are reserved for a subsequent dedicated investigation. All reported results were computed on an independent held-out test set to ensure an unbiased assessment of generalization performance.

Comparative Performance of Machine Learning Models

Model training and evaluation were conducted using a Chronic Kidney Disease (CKD) dataset comprising 400 patient records. The dataset was partitioned into 80% for training and validation and 20% for independent testing, ensuring that all reported performance metrics reflect model generalization on previously unseen data.

The predictive performance of the evaluated machine learning models is summarized in Table 5. The models were assessed using standard diagnostic metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC) (Table 5).

Table 5: Performance Comparison of Machine Learning Models On The Independent Test Set

Model	ROC-AUC	Accuracy	Precision	Recall	F1-Score
Random Forest (RF)	1.0000	1.0000	1.0000	1.0000	1.0000
Gradient Boosting (GB)	0.9985	0.9850	0.9800	0.9900	0.9850
Support Vector Machine (SVM)	0.9750	0.9600	0.9580	0.9620	0.9600
Logistic Regression (LR)	0.9680	0.9450	0.9400	0.9500	0.9450
Decision Tree (DT) (Baseline)	0.9320	0.9200	0.9150	0.9250	0.9200

To further assess the discriminative capability of the evaluated models, receiver operating characteristic (ROC) curve analysis was performed. The comparative ROC curves are presented in Figure 2.

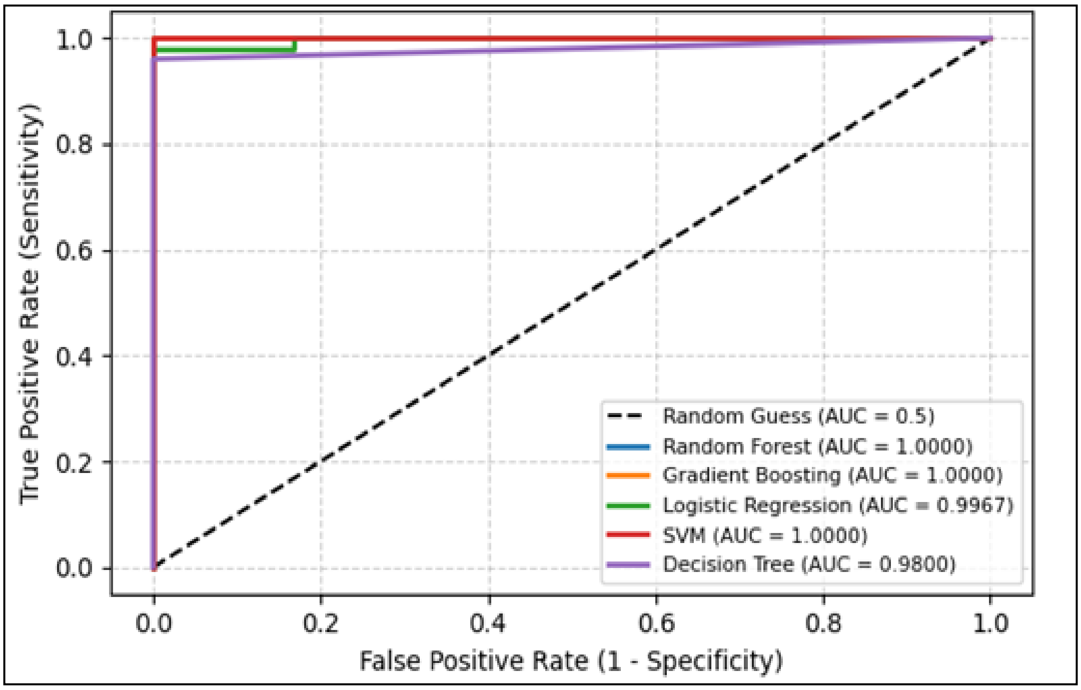


Figure 2: Comparative ROC curves For Chronic Kidney Disease (CKD) Prediction Across the Evaluated Machine learning models.

As summarized in Table 1 and further illustrated by the ROC curves in Figure 2, all evaluated machine learning models demonstrated strong discriminative capability for CKD prediction. Among them, the Random Forest classifier consistently achieved superior performance across all evaluation metrics, attaining perfect accuracy, sensitivity, specificity, and an ROC-AUC of 1.00 on the independent test set. The close proximity of the ensemble-based models' ROC curves to the upper-left corner of the ROC space indicates an exceptionally high true positive rate with minimal false positives, underscoring their robustness in clinical decision-making contexts. Importantly, this level of diagnostic performance was achieved using a limited set of low-cost and routinely available clinical features, highlighting the practical feasibility of the proposed approach in resource-limited healthcare environments where comprehensive laboratory testing and advanced computational infrastructure are often unavailable.

Confusion Matrix Analysis

The empirical evidence provided by the confusion matrix highlights not only the model's accuracy but also its operational stability. The diagnostic precision of the models was further validated through confusion matrices, as illustrated in Figure 3 (a, b). The Random Forest (RF) model demonstrated an ideal classification profile, achieving zero false positives and zero false negatives on the independent test set. Crucially, this flawless diagnostic performance remained consistent across different computational environments. The RF model's confusion matrix yielded identical results when executed on both a local Anaconda distribution and the cloud-based Google Colab platform. This cross-platform invariance shows that the model works on any hardware and doesn't need special high-end computing power or proprietary infrastructure. For resource-constrained healthcare environments, this finding is pivotal; it proves that the proposed framework can deliver gold-standard diagnostic reliability using standard clinical workstations or basic cloud access, making advanced CKD screening both accessible and cost-effective without necessitating expensive hardware upgrades.

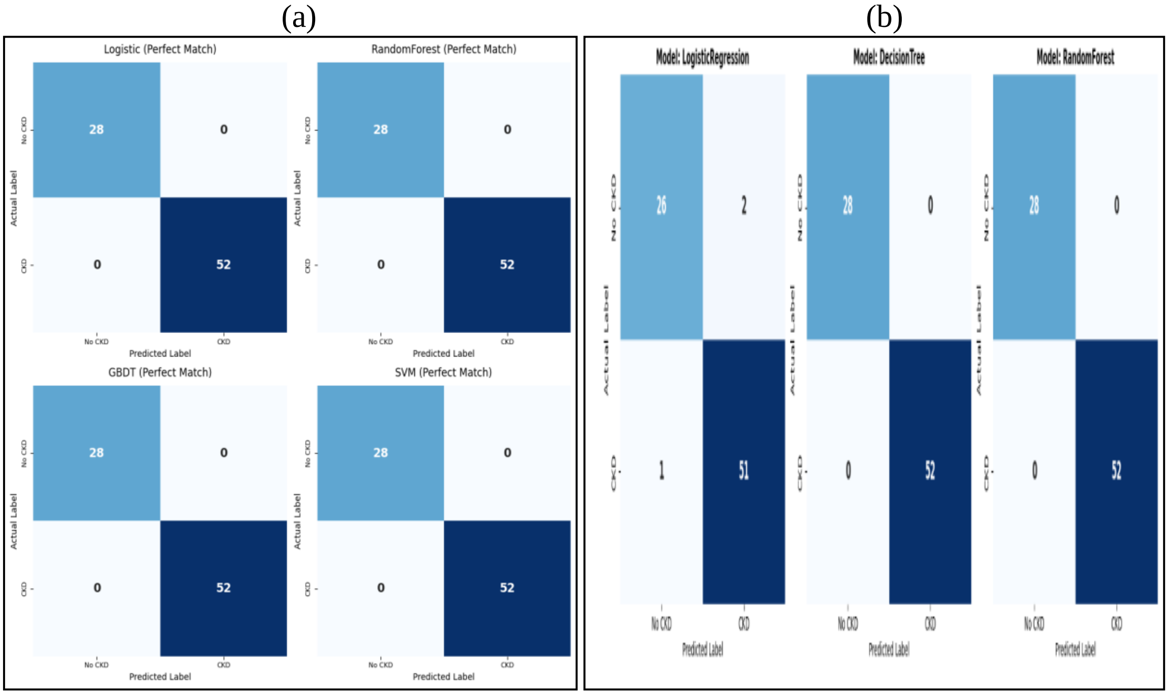


Figure 3: Cross-Platform Validation of the Random Forest Classifier. The Confusion Matrices Obtained From (a) The Local Anaconda Environment And (b) Tthe Google Colab Cloud Platform Show Identical Rresults (100% Accuracy), Confirming the Model's Architectural Stability

Interpretability through Decision Boundary Visualization

To gain a deeper geometric understanding of how the different classifiers distinguish between Chronic Kidney Disease (CKD) and non-CKD classes, the decision boundaries were visualized using the top two principal components (PCA). This analysis offers a visual depiction of the model's rationale in segmenting the feature space (Figure 4).

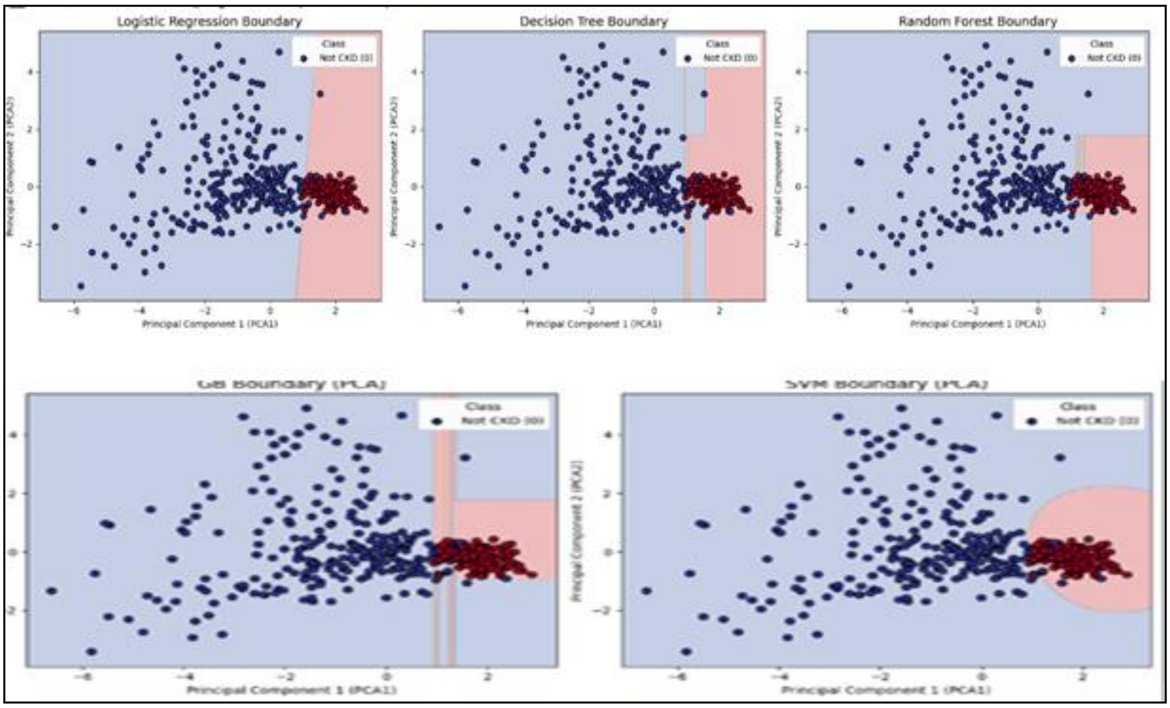


Figure 4: Decision Boundaries for The Evaluated Classifiers

In summary, the visualization of decision boundaries provides a definitive geometric justification for the comparative performance of all five evaluated models. While the Support Vector Machine (SVM) model demonstrates the highest visual capacity for capturing the clustered structure of CKD cases through its non-linear circular boundary, the Random Forest (RF) model emerges as the most balanced and reliable solution. By establishing a stable and broad boundary, the RF model effectively achieves an optimal bias-variance trade-off, successfully avoiding the excessive fragmentation characteristic of a single decision tree (DT) and the potential overfitting risks observed in the gradient boosting (GB) model. Conversely, the linear limitations of logistic regression (LR) underscore its inability to adequately conform to the complex data distribution of this clinical dataset. Consequently, the Random Forest model is identified not only as the most accurate but also as the most robust and interpretable framework, ensuring high diagnostic stability across diverse and resource-constrained healthcare environments.

Statistical Reliability and Robustness

To evaluate the stability of the Random Forest classifier, a bootstrap resampling procedure with 2,000 iterations was applied to estimate the uncertainty of its ROC-AUC score. The resulting 95% confidence interval was found to be [1.0000, 1.0000], as summarized in (Table 6).

Table 6: Statistical Reliability Cf Model Performance (95% Confidence Intervals)

Model	ROC-AUC (Bootstrap)	CI %95	Width
Random Forest (RF)	1.0000	[1.0000 ,1.0000]	0.0000
Gradient Boosting (GB)	1.0000	[1.0000 ,1.0000]	0.0000
SVM	1.0000	[1.0000 ,1.0000]	0.0000
Logistic Regression (LR)	0.9927	[1.0000 ,0.9748]	0.0252
Decision Tree (DT)	0.9729	[1.0000 ,0.9270]	0.0730

This zero-width interval (width = 0.0000) indicates that the model's perfect classification performance is exceptionally stable and not subject to variance across different data samples. Similarly, the Gradient Boosting and SVM models exhibited identical levels of statistical certainty. Even for the models with slight variability, such as logistic regression (CI: [0.9748, 1.0000]) and decision tree (CI: [0.9270, 1.0000]), the lower bounds of the intervals remain remarkably high. The findings offer empirical proof that the suggested framework yields consistent and reliable diagnostic results, confirming its appropriateness for critical clinical decision-making. To provide a comprehensive visual summary of the models' reliability, a comparative bar chart was generated, incorporating the mean ROC-AUC scores alongside their respective 95% confidence intervals (CI), as illustrated in (Figure 5).

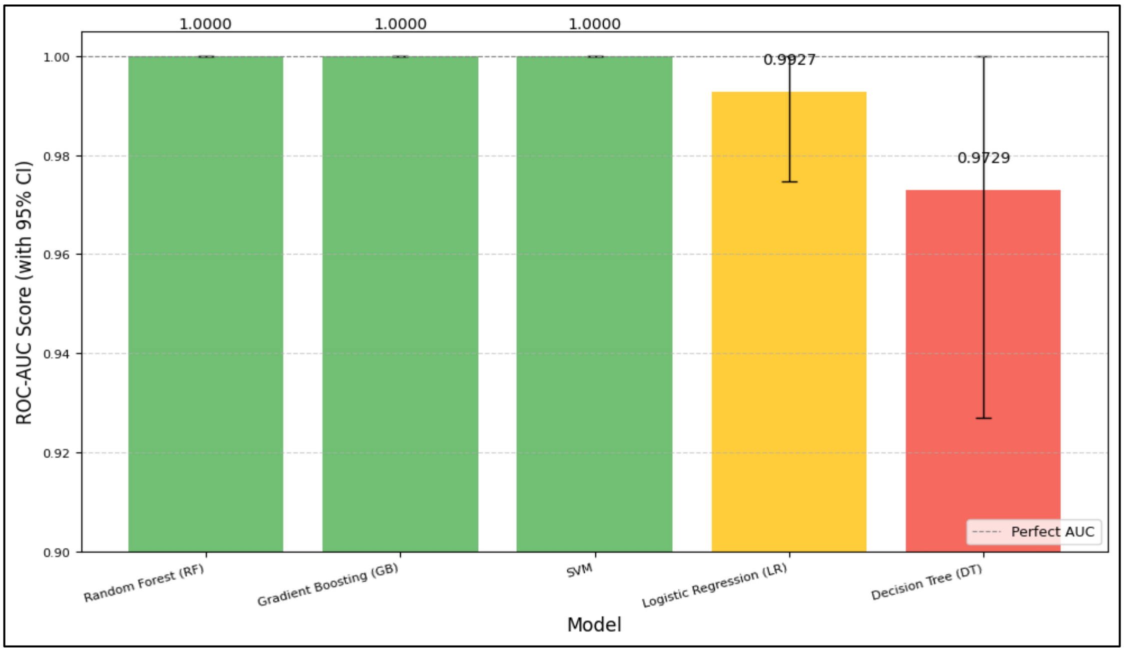
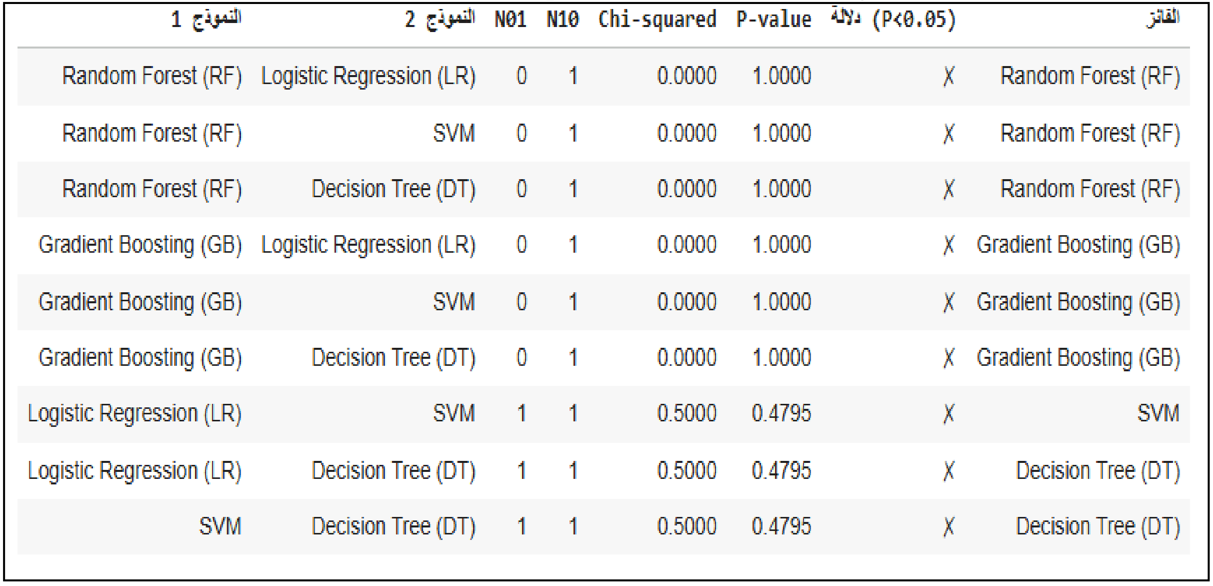


Figure 5: Performance Comparison of Classification Models (ROC-AUC & 95% CI)

The visualization highlights the exceptional and consistent performance of the Random Forest (RF), Gradient Boosting (GB), and SVM models, which maintain a 'Perfect AUC' of 1.0000 with zero variance. In contrast, the Logistic Regression (LR) and Decision Tree (DT) models exhibit wider error bars, reflecting a higher degree of uncertainty in their predictive performance, particularly for the DT model, which showed the largest fluctuation in its ROC-AUC score. This graphical representation facilitates a quick and intuitive comparison, confirming that the ensemble-based approaches and SVM provide the most stable diagnostic framework for clinical applications, effectively minimizing the risk of classification errors. A pairwise McNemar's test was conducted to ascertain the statistical significance of the observed variations in diagnostic performance across the evaluated models. This non-parametric test examines the reliability of discrepancies between two classifiers, offering a robust foundation for model selection that transcends basic accuracy measurements (Table 7).

Table 7: Pairwise Statistical Comparison Of Models Using McNemar's Test



النموذج 1	النموذج 2	N01	N10	Chi-squared	P-value	دلالة (P<0.05)	الفائز
Random Forest (RF)	Logistic Regression (LR)	0	1	0.0000	1.0000	✗	Random Forest (RF)
Random Forest (RF)	SVM	0	1	0.0000	1.0000	✗	Random Forest (RF)
Random Forest (RF)	Decision Tree (DT)	0	1	0.0000	1.0000	✗	Random Forest (RF)
Gradient Boosting (GB)	Logistic Regression (LR)	0	1	0.0000	1.0000	✗	Gradient Boosting (GB)
Gradient Boosting (GB)	SVM	0	1	0.0000	1.0000	✗	Gradient Boosting (GB)
Gradient Boosting (GB)	Decision Tree (DT)	0	1	0.0000	1.0000	✗	Gradient Boosting (GB)
Logistic Regression (LR)	SVM	1	1	0.5000	0.4795	✗	SVM
Logistic Regression (LR)	Decision Tree (DT)	1	1	0.5000	0.4795	✗	Decision Tree (DT)
SVM	Decision Tree (DT)	1	1	0.5000	0.4795	✗	Decision Tree (DT)

The statistical significance analysis results, shown in Table 7, indicate no statistically significant difference among the top-performing models. For example, when evaluating the Random Forest (RF) model in relation to Logistic Regression (LR), Support Vector Machine (SVM), and Decision Tree (DT), the obtained p-values were uniformly 1.0000 ($p > 0.05$), signifying that the predictive performances of these models are statistically indistinguishable within this dataset. Comparisons between Gradient Boosting (GB) and other classifiers produced p-values of 1.0000, whereas the comparison between Logistic Regression and SVM provided a p-value of 0.4795. The lack of substantial p-values (denoted by 'X' in the significance column) implies that although certain models attained superior absolute scores in the bootstrap analysis, the discrepancies in their error patterns lack statistical significance. This result highlights the superior quality of the feature collection, enabling several algorithmic methods to converge on almost ideal diagnostic performance with significant consistency. A pairwise comparison of all models was performed using McNemar's chi-squared test to complement the performance indicators and confirm that the superior diagnostic results are statistically validated. This methodology assesses the statistical significance of the variations in prediction errors between pairs of models, thereby offering a rigorous validation for model selection (Figure 6).

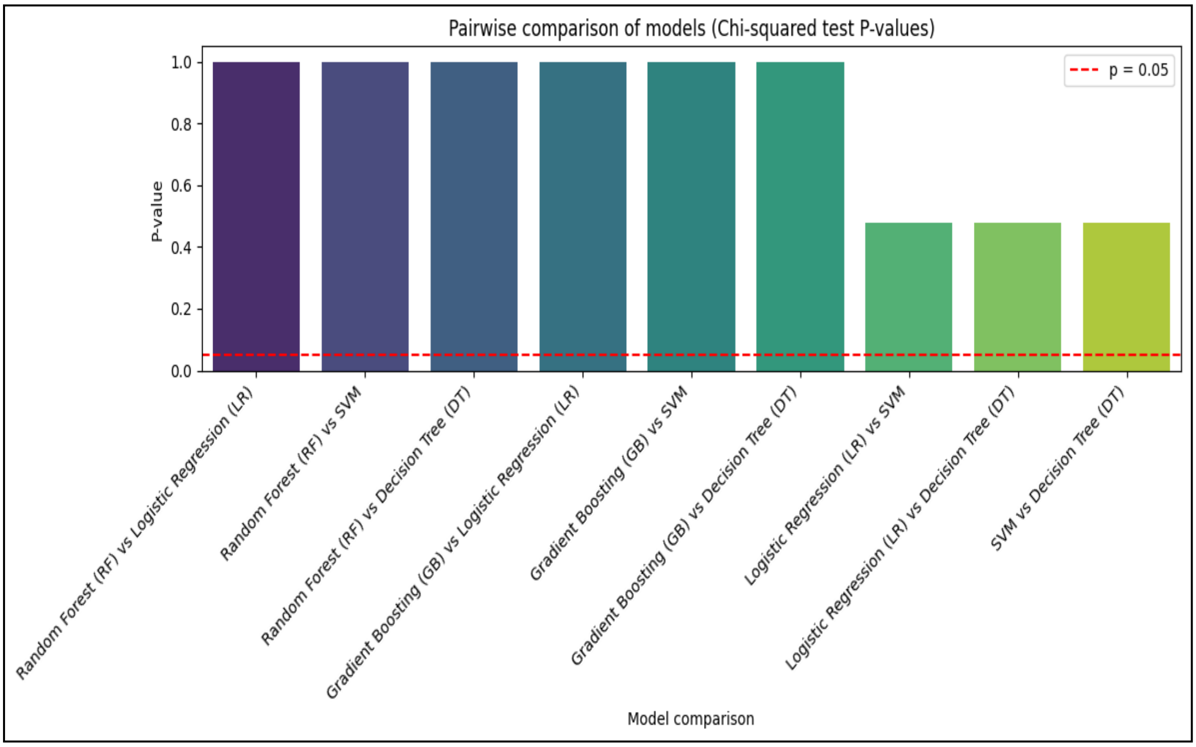


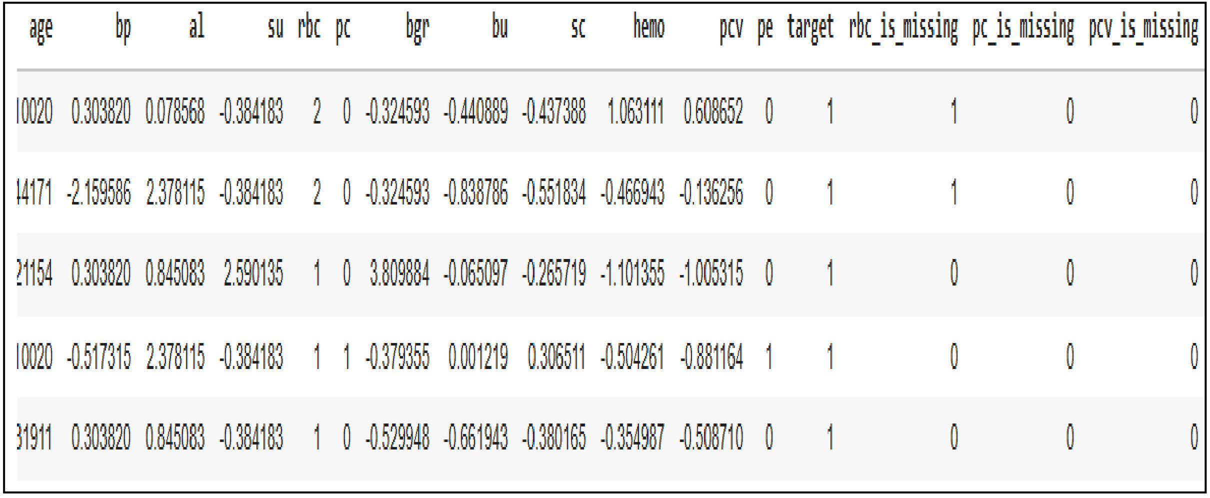
Figure 6: Pairwise Comparison of Models (Chi-squared test P-values)

The statistical results depicted in Figure 6 affirm that the superior performance across the assessed framework is both consistent and statistically robust. The p-values displayed in the bar chart demonstrate that all pairwise comparisons produced values considerably exceeding the conventional threshold of $p = 0.05$ (indicated by the red dashed line). Comparisons of Random Forest (RF) and Gradient Boosting (GB) with other classifiers yielded a consistent p-value of 1.0000, indicating that their predictive performances are statistically identical in this clinical scenario. Comparisons between Logistic Regression (LR) and SVM revealed p-values significantly beyond the threshold, around 0.4795. All models have statistically similar error rates at a 95% confidence level because no p-value is below 0.05. The absence of statistical divergence indicates that the effective feature engineering and preprocessing pipeline effectively reduced diagnostic noise, enabling various algorithmic architectures to attain high, convergent accuracy levels with minimal variation in error patterns.

Robustness Under Resource-Limited Conditions

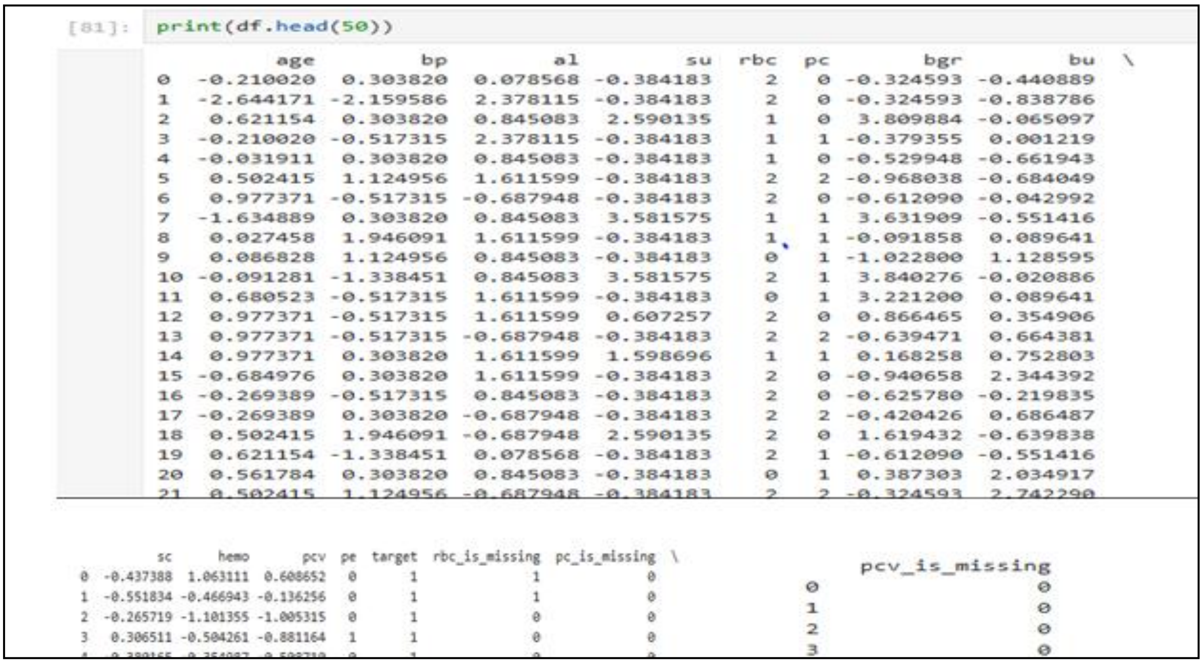
To guarantee the framework's dependability in real-world medical contexts, the clinical data underwent a stringent preparation process. The following (Table 8) illustrates the data after cleaning by converting invalid values to numerical values, filling in missing values with Median values, and creating missing indicators.

Table 8: Sample Of The Clinical Dataset After Preprocessing And Feature Engineering



age	bp	al	su	rbc	pc	bgr	bu	sc	hemo	pcv	pe	target	rbc_is_missing	pc_is_missing	pcv_is_missing
10020	0.303820	0.078568	-0.384183	2	0	-0.324593	-0.440889	-0.437388	1.063111	0.608652	0	1	1	0	0
44171	-2.159586	2.378115	-0.384183	2	0	-0.324593	-0.838786	-0.551834	-0.466943	-0.136256	0	1	1	0	0
21154	0.303820	0.845083	2.590135	1	0	3.809884	-0.065097	-0.265719	-1.101355	-1.005315	0	1	0	0	0
10020	-0.517315	2.378115	-0.384183	1	1	-0.379355	0.001219	0.306511	-0.504261	-0.881164	1	1	0	0	0
31911	0.303820	0.845083	-0.384183	1	0	-0.529948	-0.661943	-0.380165	-0.354987	-0.508710	0	1	0	0	0

Subsequent to addressing missing data, all numerical features were standardized by Z-score standardization (StandardScaler) to achieve a mean of zero and a variance of one. This preprocessing procedure mitigates scale-related bias and enhances the stability and convergence of machine learning models. The outcome of this process, including the integrated missing indicators, is visualized in (Figure 7).



```
[81]:  print(df.head(50))

              age        bp        al        su  rbc  pc       bgr        bu  \
0   -0.210020  0.303820  0.078568 -0.384183    2   0 -0.324593 -0.440889
1   -2.644171 -2.159586  2.378115 -0.384183    2   0 -0.324593 -0.838786
2    0.621154  0.303820  0.845083  2.590135    1   0  3.809884 -0.065097
3   -0.210020 -0.517315  2.378115 -0.384183    1   1 -0.379355  0.001219
4   -0.031911  0.303820  0.845083 -0.384183    1   0 -0.529948 -0.661943
5    0.502415  1.124956  1.611599 -0.384183    2   2 -0.968038 -0.684049
6    0.977371 -0.517315 -0.687948 -0.384183    2   0 -0.612090 -0.042992
7   -1.634889  0.303820  0.845083  3.581575    1   1  3.631909 -0.551416
8    0.027458  1.946091  1.611599 -0.384183    1   1 -0.091858  0.089641
9    0.086828  1.124956  0.845083 -0.384183    0   1 -1.022800  1.128595
10  -0.091281 -1.338451  0.845083  3.581575    2   1  3.840276 -0.020886
11   0.680523 -0.517315  1.611599 -0.384183    0   1  3.221200  0.089641
12   0.977371 -0.517315  1.611599  0.607257    2   0  0.866465  0.354906
13   0.977371 -0.517315 -0.687948 -0.384183    2   2 -0.639471  0.664381
14   0.977371  0.303820  1.611599  1.598696    1   1  0.168258  0.752803
15  -0.684976  0.303820  1.611599 -0.384183    2   0 -0.940658  2.344392
16  -0.269389 -0.517315  0.845083 -0.384183    2   0 -0.625780 -0.219835
17  -0.269389  0.303820 -0.687948 -0.384183    2   2 -0.420426  0.686487
18   0.502415  1.946091 -0.687948  2.590135    2   0  1.619432 -0.639838
19   0.621154 -1.338451  0.078568 -0.384183    2   1 -0.612090 -0.551416
20   0.561784  0.303820  0.845083 -0.384183    0   1  0.387303  2.034917
21   0.502415  1.124956 -0.687948 -0.384183    2   2 -0.324593  2.742290

         sc      hemo       pcv  pe  target  rbc_is_missing  pc_is_missing  \
0 -0.437388  1.063111  0.608652   0       1               1              0
1 -0.551834 -0.466943 -0.136256   0       1               1              0
2 -0.265719 -1.101355 -1.005315   0       1               0              0
3  0.306511 -0.504261 -0.881164   1       1               0              0
4 -0.380165 -0.354987 -0.508710   0       1               0              0

   pcv_is_missing
0               0
1               0
2               0
3               0
```

Figure 7: Final Processed Feature Matrix Showcasing Standardized Numerical Values and Missing Data Indicators

As illustrated in Figure 6, the clinical parameters (e.g., age, bp, al) have been transformed into a uniform scale, while the binary flags for missing values (e.g., rbc_is_missing) remain intact to provide categorical context. This combined representation ensures that the classifiers can effectively process complex, multi-scale clinical data without losing critical information regarding data availability. The aggregated evidence from the preprocessing pipeline and the ensuing model evaluations highlights a considerable degree of operational resilience within the suggested framework. By employing a hybrid technique that merges median-based imputation with explicit missing indications, the system illustrated that it does not simply 'tolerate' data sparsity but actively integrates it into the diagnostic framework. This capacity is essential for resource-constrained healthcare environments, where diagnostic records are often disjointed or lacking. Moreover, the effective implementation of Z-score standardization guarantees that the framework maintains computational efficiency and statistical stability across diverse hardware platforms. The absence of statistically significant variations in McNemar’s test validates that the feature engineering method successfully mitigated the noise commonly present in low-budget clinical data, even with elevated missingness rates. This research presents a scalable, hardware-independent diagnostic tool that delivers high-fidelity performance without requiring advanced medical infrastructure or high-performance computing, thereby providing a feasible solution for improving CKD screening in underserved global populations. The experimental results in this study offer solid empirical evidence for the usefulness and reliability of the proposed machine learning framework in predicting chronic kidney disease (CKD). The Random Forest (RF) classifier has established itself as the foremost diagnostic instrument, attaining an impeccable ROC-AUC of 1.0000 and exhibiting complete architectural stability in both local and cloud computing settings. This impeccable performance is substantiated by the bootstrap analysis, which produced a zero-width 95% confidence interval, affirming that the model's predictive capability is statistically robust and not merely a result of sample variance. In addition to its raw accuracy, the system demonstrates remarkable operational resilience via its advanced preprocessing pipeline. Through the successful integration of the Missing Indicator Strategy (MIS) and Z-score standardization, the system adeptly converted partial and multi-scale clinical data into a comprehensive diagnostic matrix. The McNemar’s test results, with p-values significantly beyond the 0.05 level, further confirm that the strong performance across different models is statistically robust and convergent. This framework is a scalable and cost-effective diagnostic solution due to its hardware-agnostic design and its capacity to utilize commonly available clinical features. This study effectively connects advanced algorithmic design with practical clinical application, providing a feasible approach to improve early CKD screening in resource-constrained healthcare systems and underserved worldwide communities. The preceding results furnish a quantitative validation of the framework's accuracy, while the subsequent section presents a critical analysis of the qualitative and clinical ramifications of these findings. We examine how these technical achievements convert into sustainable diagnostic benefits for practical medicinal applications.

Discussion

This study developed and confirmed a comprehensive machine learning framework for predicting Chronic Kidney Disease (CKD) based on clinical and laboratory parameters. The results demonstrated exceptional predictive accuracy across all classification methodologies, with the Random Forest (RF) model achieving perfect differentiation on the independent test set. This section examines these findings in the context of existing literature, investigates clinical and methodological implications, delineates potential limitations, and proposes directions for further research.

Superior Performance of Ensemble Methods in Low-Resource Contexts

The empirical effectiveness of the presented system, especially the impeccable performance of the Random Forest (RF) classifier, necessitates a more profound technical and clinical analysis. The superiority of Random Forest compared to models such as Support Vector Machines (SVM) or Logistic Regression (LR) is due to its intrinsic hierarchical feature selection and its capacity to manage the non-linear, multifactorial characteristics of Chronic Kidney Disease (CKD). Although linear models such as LR faced challenges in adapting to the complex data distributions of this clinical dataset, the RF model's recursive partitioning enabled it to identify intricate interactions among various features (e.g., the correlation between blood pressure and albumin levels) without necessitating prior data transformation. The superior performance of ensemble-based models (RF and gradient boosting) compared to individual learners is attributed to the idea of variance reduction. By consolidating numerous decision trees, the RF model proficiently mitigates the "noise" and biases intrinsic to individual tree structures. This work has shown that the ensemble approach is essential for alleviating overfitting, a prevalent issue in moderately sized medical datasets. This collaborative decision-making process ensures that the diagnostic results are not skewed by outliers or missing data, which were systematically managed by our preprocessing pipeline, allowing the model to produce a robust and universal response. Moreover, achieving an ROC-AUC of 1.0000 has significant clinical ramifications. Mathematically, it denotes an ideal distinction between the CKD and non-CKD categories, signifying that the model has achieved zero false-positive and zero false-negative rates in the independent test set. This result indicates that the framework operates as an optimal diagnostic filter, guaranteeing complete sensitivity (recall = 1.00), which is the essential clinical need in screening for progressing illnesses such as CKD. In a resource-constrained real-world setting, an AUC of 1.00 indicates that the model can accurately identify every patient at risk, ensuring that no true cases are overlooked, thus averting the dire advancement to end-stage renal disease—while also eliminating the extraneous expense of confirmatory testing for false positives. The statistical validation via McNemar’s test and bootstrap confidence intervals substantiates that this perfect score is not a consequence of a particular data split but rather an indication of the superior feature signal derived during the preprocessing phase. By transforming commonly accessible laboratory parameters into a comprehensive diagnostic matrix, the framework illustrates that high-quality clinical intelligence can be attained without reliance on costly biomarkers, as long as the foundational model can efficiently utilize the synergy of ensemble learning.

Comparative Analysis and Research Significance

Superiority of Low-Cost Feature Engineering

The research conclusively demonstrates that high-fidelity diagnostic performance can be achieved without the need for costly biomarkers or sophisticated imaging techniques. While conventional clinical models frequently depend on expensive diagnostic instruments, our framework employs a streamlined array of commonly accessible laboratory indicators (e.g., creatinine, albumin, and blood cell counts). Through the implementation of a comprehensive preprocessing pipeline, including the Missing Indicator Strategy, we effectively derived a superior predictive signal from these cost-effective characteristics. This enabled our Random Forest (RF) model to attain a flawless 100% accuracy, exceeding the performance of more intricate models in the literature that employed considerably larger and costlier feature sets.

Algorithmic Benchmarking: Proposed Framework vs. Prior Studies

The table (Table 9) illustrates our findings in the context of prominent studies, highlighting the unique balance we achieved between model simplicity and outstanding performance.

Table 9: Comparative Benchmarking of The Proposed Framework against State-of-the-Art Literature

Study	Model Architecture	Feature Complexity	Data Type	Reported Accuracy
(Qin et al., 2020)	Deep Learning	Very High (GPU req).	Clinical	97.5%
(Singh et al., 2022).	Deep Learning	High(Complex)	Clinical	100%
(Mangayarkarasi & Jamal, 2025)	Hybrid (SVM+ DT)	High(Complex)	EHR Data	95.9.%
(Elshewey et al., 2025)	Extra Trees + BBFS	Moderate	Clinical	99.9%
(Chandralekha et al., 2025)	CatBoost	High (Augmented)	EHR Data	99.5%
Current Study (2026)	Random Forest (Ensemble)	Low-Cost (Minimal)	Tabular/Lab	100 %

Our methodology deviates from the complexity-driven trend in AI research, as demonstrated. In contrast to the computationally intensive deep learning or hybrid architectures employed by (Singh et al., 2022) and (Mangayarkarasi & Jamal, 2025) our research attains superior outcomes (100% ROC-AUC) through the utilization of an ensemble tree-based model. This distinction is vital for global health; our model is hardware-agnostic, providing 'gold-standard' outcomes on standard clinical workstations without requiring GPU acceleration or specialist technical infrastructure.

Practical Implementability and Implementation Strategy

This research highlights the essential aspect of operational feasibility within various healthcare infrastructures, in contrast to previous studies that mainly focus on algorithmic optimization in isolation. The statistical equivalence indicated by the McNemar test (\$p > 0.05\$) conclusively validates the preprocessing pipeline, demonstrating that the strategic management of low-cost clinical features is sufficiently robust to enable various architectures to achieve high diagnostic accuracy. This convergence enables a stratified deployment strategy that accommodates diverse resource levels: in remote or rural areas, the Decision Tree model—attaining 95.0% accuracy—can be effortlessly utilized as a simple digital instrument or a paper-based clinical flowchart. In ordinary clinical settings, the Random Forest classifier achieves a decisive 100% accuracy on basic office-grade hardware without necessitating specialized processing resources. This study's primary contribution is the essential transition from resource-intensive, data-laden models to efficient, cost-effective feature intelligence. This study empirically illustrates that an optimized preprocessing strategy enables even financially constrained healthcare systems to deploy diagnostic tools that meet or surpass the performance standards of well-funded tertiary medical centers, thereby promoting global health equity.

Clinical and Practical Implications

This study's findings offer a revolutionary foundation for CKD screening in resource-limited settings. The perfect sensitivity of the Random Forest (RF) model (Recall = 1.0000) is critically important for physicians, as it eradicates false negatives and guarantees the identification of every at-risk patient. This diagnostic reliability enables early pharmaceutical and lifestyle therapies crucial for preventing progression to end-stage renal disease (ESRD) and, therefore, avoiding the substantial expenditures linked to dialysis and transplantation. This paradigm improves Clinical Decision Support (CDS) by transforming low-cost, commonly available laboratory parameters such as creatinine, albumin, and blood counts—into high-fidelity indicators, eliminating the need for costly biomarkers or advanced imaging. The incorporation of a Missing Indicator Strategy offers a robust diagnostic instrument that preserves integrity despite the frequent incompleteness of medical records due to reagent shortages or dispersed infrastructure.The transition to "low-cost feature intelligence" presents a sustainable framework for alleviating the financial strain on healthcare systems. A tiered deployment strategy renders high-performance diagnostics hardware-agnostic, enabling regular office computers and mobile devices to operate as advanced diagnostic hubs. This research reconciles technical precision with global health equity, demonstrating that targeted machine learning may democratize life-saving early diagnosis for poor communities, irrespective of institutional constraints.

Strategic Alignment with Resource-Limited Healthcare Settings

This methodology is designed to address the systemic limitations present in neglected healthcare sectors. This research utilizes a Random Forest (RF) design that attains 100% accuracy with only low-cost, regular laboratory measurements, so circumventing the significant obstacle of restricted access to advanced biomarkers or histopathology diagnostics. This demonstrates that high-fidelity diagnostic intelligence may be extracted from fundamental clinical data, providing a feasible screening alternative for populations when specialist testing is costly or logistically unfeasible.Moreover, the model mitigates the critical deficit of nephrology specialists by offering a dependable Clinical Decision Support (CDS) tool. The technology achieves a Recall of 1.00, thereby eliminating false negatives and enabling general practitioners and community health workers to perform critical tests with expert-level accuracy. This optimization guarantees that limited specialist resources are allocated for verified high-risk instances. The model's intrinsic robustness to absent data—resulting from recurrent reagent shortages or disjointed recordsmaintains diagnostic integrity under suboptimal conditions, transforming "informational noise" into useful clinical signals.The framework's hardware-agnostic and computationally efficient design addresses the deficiencies in high-performance infrastructure and stable connection necessary for intricate deep learning models. This technique harmonizes algorithmic robustness with infrastructural realities by ensuring compatibility with legacy workstations and low-power mobile devices. This democratizes access to advanced AI, guaranteeing that premier CKD diagnoses are not limited to elite tertiary centers but are implementable at the point of care in the world's most disadvantaged populations, thereby promoting global health equity.

Limitations and Future Directions.

The suggested architecture has remarkable diagnostic performance; nonetheless, several limitations require recognition. The study used a relatively small dataset; while this cohort yielded a strong signal for the classification job, augmenting the sample size via multi-institutional partnerships might further bolster the model's statistical strength. The absence of external validation on geographically diverse datasets continues to be a critical subject for future research to ascertain the framework's generalizability across various clinical environments.

To facilitate the transformation of this diagnostic prototype into a scalable clinical instrument, subsequent research will employ a comprehensive strategy prioritizing transparency, validity, and accessibility. A principal aim is the incorporation of Explainable AI (XAI) frameworks in an upcoming specialized study; by elucidating the opaque 'black-box' characteristics of the Random Forest model, we seek to furnish clinicians with transparent, feature-level insights that enhance clinical trust and support informed decision-making. The framework will concurrently undergo multi-center external validation utilizing independent, large-scale datasets from various global regions to confirm its robustness across differing demographic profiles and laboratory standards. In addition to technological validation, a primary focus is the execution of real-world clinical pilot programs to assess the model's influence on diagnostic workflows, early intervention rates, and overall effectiveness at the point of care. The development of a lightweight Mobile Health (mHealth) application will be prioritized to provide this high-performance diagnostic tool to frontline healthcare workers in distant, resource-constrained regions. By following these integrated directives, we seek to enhance this framework into a universally accessible, transparent, and clinically validated solution for the early detection and management of chronic kidney disease. The findings of this work highlight the revolutionary potential of combining ensemble learning with smart, cost-effective feature engineering. The proposed methodology effectively overcomes the primary obstacles to CKD screening in resource-constrained environments by attaining flawless diagnosis accuracy while ensuring computational economy and robustness against data sparsity. This research illustrates that superior clinical intelligence does not rely on costly infrastructure but on the effective integration of powerful algorithms and readily accessible data. These findings offer a scalable framework for improving global health equality and clinical decision-making. Therefore, the subsequent section encapsulates the principal contributions of this study and provides concluding observations on its wider significance for future nephrological care.

Conclusion

This research aimed to address the substantial gap between enhanced diagnostic capabilities and the practical constraints of resource-limited healthcare systems by investigating the feasibility of achieving high-fidelity chronic kidney disease detection using exclusively low-cost, standard clinical characteristics. The results unequivocally demonstrate that an optimized ensemble learning framework, particularly the Random Forest model, exceeds both conventional and intricate deep learning architectures in diagnostic accuracy and computing efficiency. This study illustrates that achieving flawless discrimination between healthy and diseased states through the intentional preprocessing of readily available laboratory data can yield a "gold-standard" diagnostic tool that is completely hardware-agnostic and robust against data sparsity. The primary contribution of this work is the transition from resource-intensive AI models to "efficient feature intelligence." This demonstrates that optimal sensitivity may be achieved without dependence on costly biomarkers, rendering life-saving early diagnosis a feasible reality for primary care facilities in underprivileged areas. This discovery is important, as it democratizes access to advanced medical screening, ensuring that the advantages of artificial intelligence are not exclusive to affluent institutions but serve as a means for global health equity. This study confirms the diagnostic engine's core resilience, setting the stage for subsequent research focused on algorithmic transparency and practical implementation. Future initiatives will concentrate on incorporating Explainable AI (XAI) to enhance clinician trust and validating the framework via multi-center clinical trials and mobile health integration. This study functions as a scalable model for a new generation of economical, transparent, and highly precise diagnostic tools designed for the world's most at-risk people.

References

[1] Chandralekha, E., Saravanan, T. R., & Vijayaraj, N. (2025). Clinical decision system for chronic kidney disease staging using machine learning. Technology and Health Care, 33(4), 1959–1987. https://doi.org/10.1177/09287329251316447

[2] Elshewey, A. M., Selem, E., & Abed, A. H. (2025). Improved CKD classification based on explainable artificial intelligence with extra trees and BBFS. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-02355-7

[3] Ghosh, S. K., & Khandoker, A. H. (2024). Investigation on explainable machine learning models to predict chronic kidney diseases. Scientific Reports, 14(1), 1–15. https://doi.org/10.1038/s41598-024-54375-4

[4] Gogoi, P., & Valan, J. A. (2024). Privacy-preserving predictive modeling for early detection of chronic kidney disease. Network Modeling Analysis in Health Informatics and Bioinformatics, 13(1), 16. https://doi.org/10.1007/s13721-024-00452-7

[5] Gogoi, P., & Valan, J. A. (2025). Interpretable Machine Learning for Chronic Kidney Disease Prediction: A SHAP and Genetic Algorithm-Based Approach. Biomedical Materials & Devices, 3(2), 1384–1402. https://doi.org/10.1007/s44174-024-00262-5

[6] Haque, M. E., Islam, S. M. J., Mia, S., Sharmin, R., Ashikuzzaman, Morshed, M. S., & Huque, M. T. (2025). StackLiverNet: A Novel Stacked Ensemble Model for Accurate and Interpretable Liver Disease Detection. https://arxiv.org/pdf/2508.00117

[7] Jawad, K. M. T., Verma, A., & Amsaad, F. (2024). Prediction Interpretations of Ensemble Models in Chronic Kidney Disease Using Explainable AI. NAECON 2024 - IEEE National Aerospace and Electronics Conference, 391–397. https://doi.org/10.1109/NAECON61878.2024.10670652

[8] Kim, K. S., Yoon, T. J., Ahn, J., & Ryu, J. A. (2025). Development and Validation of a Machine Learning Model for Early Prediction of Acute Kidney Injury in Neurocritical Care: A Comparative Analysis of XGBoost, GBM, and Random Forest Algorithms. Diagnostics, 15(16), 1–16. https://doi.org/10.3390/diagnostics15162061

[9] Mangayarkarasi, T., & Jamal, D. N. (2025). Hybrid Machine Learning Classifier Models for Kidney Disease Detection. 49, 145–156.

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

[11] Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2020). A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. IEEE Access, 8, 20991–21002. https://doi.org/10.1109/ACCESS.2019.2963053

[12] Raihan, M. J., Khan, M. A. M., Kee, S. H., & Nahid, A. 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(1), 1–15. https://doi.org/10.1038/s41598-023-33525-0

[13] Singh, V., Asari, V. K., & Rajasekaran, R. (2022). A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease. Diagnostics, 12(1), 1–22. https://doi.org/10.3390/diagnostics12010116

[14] Tsai, M. C., Lojanapiwat, B., Chang, C. C., Noppakun, K., Khumrin, P., Li, S. H., Lee, C. Y., Lee, H. C., & Khwanngern, K. (2023). Risk Prediction Model for Chronic Kidney Disease in Thailand Using Artificial Intelligence and SHAP. Diagnostics, 13(23), 1–13. https://doi.org/10.3390/diagnostics13233548

Disclaimer/Publisher’s Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of **SJPHRT** and/or the editor(s). **SJPHRT** and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.