Mathematical and Computational Applications

MDPI

Article

Weighted Average Ensemble Learning Imputation with Feature Priority Assigning (WAELI-FPA) for Explainable Chronic Kidney Disease Prediction

A. Swathi * and Golda Dilip

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai 600026, Tamil Nadu, India; goldadilip@gmail.com
* Correspondence: as5087@srmist.edu.in

Abstract

Chronic Kidney Disease (CKD) is a progressive disease. It requires timely diagnosis and treatment to prevent serious complications and death. CKD clinical datasets often contain missing values, noisy records, and irrelevant features, which reduce predictive model performance. This research proposes a new imputation framework, WAELI-FPA (Weighted Average Ensemble Learning Imputation with Feature Priority Assigning), for accurate and explainable CKD prediction. The framework begins with data cleaning, including identifying missing values, removing noise, normalizing, handling outliers, and transforming data. The Weighted Average Ensemble Learning Imputation (WAELI) method uses Expectation-Maximization (EM), Random Forest (RF), CART, and C4.5 to recover missing values. The WAELI-FPA module is then designed to use the TOPSIS method to rank clinical features and select disease-prediction attributes. The optimized feature set is processed within a Hybrid Knowledge Discovery Layer that incorporates autoencoder-driven feature learning, SHAP interaction analysis, and predictive interaction analysis. It also includes rule extraction using the Explainable Boosting Machine (EBM). Lastly, an ensemble classifier comprising Random Forest, XGBoost, Extra Trees, and weighted voting/stacking predicts CKD and Non-CKD cases. To promote transparency and trust in the clinical setting, our supervision framework employs SHAP, LIME, EBM, and counterfactual explanations within a CDSS. The proposed framework will support reliable early CKD diagnosis and clinical decision-making by improving prediction accuracy, handling missing clinical data, identifying significant features, and generating interpretable predictions.

Keywords: chronic kidney disease (CKD); missing value imputation; weighted average ensemble learning imputation feature priority assigning (WAELI-FPA); TOPSIS; explainable artificial intelligence (XAI); explainable boosting machine (EBM); SHAP; LIME; clinical decision support system (CDSS)

Check for updates

Received: 13 August 2026
Revised: 29 October 2026
Accepted: 17 November 2026
Published: 26 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Chronic Kidney Disease, or CKD, is one of the most rampant chronic diseases worldwide. More importantly, it poses a major public health concern due to its rising prevalence and high treatment costs, which severely impact patients' quality of life (QOL) [1,2]. Chronic kidney disease (CKD) is defined by the gradual decline of kidney function over time, which may ultimately evolve to End-Stage Renal Disease (ESRD), requiring lifelong dialysis or a kidney transplant [2]. According to global health reports, CKD affects millions of people. According to global health reports, millions of people are impacted by CKD.

However, a substantial number are left undiagnosed as the disease quietly develops in its early stages [1]. Detecting a disease early and promptly treating it in a clinical setting can significantly slow its progression, reduce mortality, and improve patient prognosis [3]. As a result, it has become a significant research topic in the healthcare industry to design a smart, reliable CKD diagnostic prediction system.

In recent years, disease prediction through analysis of clinical data using popular AI and ML algorithms has gained prominence in research owing to their superior performance in identifying complex, unknown patterns that are difficult to discern with conventional statistical approaches. Ensemble methods have been widely used in machine learning for CKD classification in recent years. Various supervised learning algorithms, such as Random Forest (RF) [4], Support Vector Machine (SVM) [5], Decision Tree (DT), XGBoost [6], and Artificial Neural Networks (ANN) [7], have shown promising results for classifying CKD.

Recent advances in machine learning, ensemble learning, optimization algorithms, and deep learning have demonstrated the potential of data-driven approaches to address complex prediction problems across different application domains. Conventional machine learning algorithms, including Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost), are widely investigated because of their ability to model nonlinear relationships and capture interactions among input variables. Recent research has also explored optimized regression methods, ensemble classification, and hybrid deep learning architectures to improve predictive performance. For example, Alkhammash et al. (2022) [8] investigated optimized multivariate adaptive regression splines for crude oil demand prediction, illustrating the role of optimization in developing predictive models. Similarly, research on hybrid ensemble stacking has demonstrated how multiple learning algorithms can be integrated to improve classification performance [DOI: 10.3390/electronics11111750]. These developments provide a methodological foundation for investigating integrated learning strategies in clinical prediction [8–13].

Optimization-based feature selection and hyperparameter tuning have also received increasing attention in predictive modeling. Elshewey (2025) [13] proposed a crop-yield prediction framework combining the Dove Optimization Algorithm with Gradient Boosting, demonstrating the use of metaheuristic optimization for feature selection and model tuning. In another study, Elshewey et al. (2026) [12] developed a hybrid DNN–XGBoost framework incorporating binary Whale Optimization and Grey Wolf Optimization for feature selection and hyperparameter optimization. Deep learning research has further investigated hybrid architectures that combine convolutional neural networks, residual networks, and long short-term memory networks to model complex patterns in forecasting tasks. Although these studies address application domains different from CKD prediction, they illustrate the broader development of hybrid learning, optimization, and predictive modeling strategies.

Despite these advances, developing a reliable CKD prediction framework requires more than selecting a high-performing classifier. Clinical datasets may contain missing observations, heterogeneous feature distributions, and variables with differing predictive contributions. Consequently, data imputation, feature prioritization, classification, and model interpretability should be considered together. Motivated by these challenges, the present study proposes the Weighted Average Ensemble Learning Imputation with Feature Priority Assigning (WAELI-FPA) framework, which integrates weighted ensemble-based imputation, TOPSIS-based feature prioritization, and machine learning classification with explainability techniques. The objective is to investigate whether this integrated approach can support robust and interpretable CKD prediction under a consistent evaluation protocol.

The performance of machine learning models remains highly dependent on the quality of clinical data despite these developments. Healthcare datasets frequently have missing values, noise, outliers, redundancy, and class imbalance. This can lead to weak predictions

and the model's non-applicability. Additionally, a plethora of current computational methods rely on single Imputation or traditional variable selection methods that compromise data integrity and clinical significance.

One of the most important preprocessing steps in medical data analysis is missing value imputation. The presence of incomplete laboratory reports in clinical datasets is due to unavailable tests, patient noncompliance, and data-entry errors. Standard methods of Imputation, such as mean [14], median [15], mode [4], K-nearest neighbors (KNN), and expectation-maximization (EM), work efficiently in limited space. Therefore, any of them should be able to take on a challenge with specific data. Likewise, tree-based methods such as Random Forest (RF) [16], CART [17], and C4.5 [18] handle nonlinear patterns but exhibit different performance characteristics depending on the dataset. As a result, a sole reliance on a single imputation technique produces inaccurate estimates and worsens Classification Performance.

A second major challenge is identifying clinically relevant features from large-scale health data. Traditional feature selection algorithms tend to consider only statistical relevance and ignore the combined importance of evaluation criteria. This can cause important clinical biomarkers to receive lower priority or redundant features to remain in the final feature subset, increasing computational complexity and reducing prediction efficiency.

Multi-Criteria Decision Making (MCDM) techniques will help rank the features and, in turn, improve the quality of feature selection. The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is an effective tool for ranking features by considering multiple evaluation measures simultaneously [15].

Many ensemble learning methods have improved prediction efficiency by integrating multiple classifiers. However, the framework for CKD prediction presents a limitation. It is mostly existing work that focuses solely on maximizing classification accuracy. These models are often black-box systems with a limited explanation of what influenced the Prediction. Healthcare professionals need models that are transparent and interpretable and that clearly justify diagnostic decisions in clinical practice. SHAP [19], LIME [20], Explainable Boosting Machine (EBM) [21], and Counterfactual [22] explanations are some of the XAI techniques that can provide transparency and trustworthiness. Most studies to date use only a single explanation technique, which is not very interpretable and fails to capture interactions or predictive interaction relationships among clinical variables.

To address these challenges, the current research proposes introducing a model known as WAELI-FPA. This stands for Weighted Average Ensemble Learning Imputation with Feature Priority Assigning and aims to enable accurate, explainable Prediction of CKD. The framework offered has multiple integrated layers. First, clinical data preprocessing begins with missing-value detection, noise removal, normalization, outlier handling, and data transformation. A new approach for Imputation of missing values has been proposed, called WAELI Weighted Average Ensemble Learning Imputation, which uses Expectation Maximization EM, Random Forest RF, and CART and C4.5 Decision Tree algorithms. The analysis, based on experiments, shows that WAELI provides better results than using each algorithm separately. Thereafter, the WAELI-FPA module applies the TOPSIS algorithm to generate importance scores, rank clinical features, and discard low-ranking features.

Our framework contributes the following new features:

- A new imputation technique called Weighted Average Ensemble Learning Imputation (WAELI), which combines Expectation Maximization (EM), Random Forest (RF), CART, and C4.5 within a single framework.
- An algorithm for clinical feature prioritization using the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS).

- A hybrid knowledge discovery layer that incorporates Autoencoder-based feature learning, SHAP interaction analysis, and Explainable Boosting Machine (EBM).
- The use of SHAP and LIME to provide explanations for the ensemble model outputs.
- Integration of the proposed framework into a Clinical Decision Support System (CDSS) to provide interpretable and clinically useful decision support.

The proposed framework further adds a Hybrid Knowledge Discovery Layer comprising Autoencoder-based feature learning, SHAP interaction analysis, and predictive interaction analysis. using Explainable Boosting Machine (EBM)-based clinical rule extraction, to enhance knowledge extraction. The optimized feature subset is then processed by an ensemble prediction model comprising Random Forest, XGBoost, Extra Trees, and a weighted voting/stacking strategy to improve classification robustness and prediction accuracy.

To wrap up, a Multi-Level Explainability Layer uses SHAP, LIME, EBM, counterfactual explanations, and clinical risk interpretation for transparency and clinical actionability. Integrated into a Clinical Decision Support System (CDSS), these explanations help physicians understand the model's predictions, identify key risk factors, and make evidence-informed medical decisions. The proposed framework addresses the key challenges for the CKD prediction system, namely missing values, feature importance, prediction accuracy, and interpretability, all in a unified manner.

1.1. Research Contributions

The major contributions of this study can be summarized as follows.

1.1.1. Creation of an Innovative WAELI Framework

The proposed Weighted Average Ensemble Learning Imputation (WAELI) method is constructed by blending Expectation-Maximization (EM), Random Forest (RF), CART, and C4.5 Decision Trees. The purpose of these is to estimate missing values in CKD clinical datasets, thereby improving data quality and predictive performance.

1.1.2. Introduction of WAELI Feature Priority Assignment

A TOPSIS-based mechanism for clinical attribute prioritization is proposed to evaluate each clinical attribute against multiple criteria, rank its importance, remove redundant features, and produce an optimal feature subset that improves classification accuracy while reducing computational complexity.

1.1.3. Designing a Hybrid Knowledge Discovery Layer

An innovative framework for knowledge discovery is presented wherein the paradigm leverages Autoencoder-based feature learning, SHAP interaction analysis-based predictive interaction analysis, and EBM-based rule extraction to filter out meaningful clinical relationships and enhance medical knowledge interpretation.

1.1.4. Creating a Predictive Model Using an Ensemble Technique

An efficient ensemble classification model is developed by combining random forest, XGBoost, Extra Trees, and weighted voting/stacking to achieve robust, accurate, and generalized Prediction of CKD compared with conventional single classifiers.

1.1.5. Implementation of a Multi-Level Framework for Explainability

Various Explainable Artificial Intelligence (XAI) techniques, such as SHAP, LIME, Explainable Boosting Machine (EBM), and counterfactual explanations, are integrated to provide both global and local explainability, enhancing transparency, clinician confidence, and model trust.

1.1.6. Development of a Clinical Decision Support System (CDSS) That Is Explainable

The proposed framework implements an intelligent Clinical Decision Support System that accurately predicts CKD risk, analyzes feature importance, stratifies risk, generates clinical rules, and provides personalized decision support for healthcare professionals.

1.1.7. Thorough Performance Assessment

Thus, the proposed framework is assessed using various performance measures, including Accuracy, Precision, Recall, F1-Score, AUC, RMSE, MAE, computational efficiency, robustness, fairness, and explainability, thereby demonstrating its superiority over existing CKD prediction approaches.

The WAELI-FPA framework integrates multiple machine learning approaches to develop a robust, unified, explainable AI-based clinical tool for the early diagnosis of CKD. In particular, it implements advanced missing-value Imputation, intelligent feature prioritization via effective Genetic Optimization, hybrid knowledge discovery through learning from clinical experts and lab tests, and ensemble-based learning.

The outcome is expected to aid physicians in making accurate, transparent, and reliable diagnostic decisions while enabling the use of trustworthy AI in healthcare.

1.1.8. Methodological Novelty of the Proposed WAELI-FPA Framework

The methodological contribution of WAELI-FPA lies not in the invention of its constituent algorithms, but in the design of a performance-adaptive imputation and multi-criteria feature-prioritization workflow for Chronic Kidney Disease (CKD) prediction. Specifically, the Weighted Average Ensemble Learning Imputation (WAELI) module estimates the relative contribution of Expectation-Maximization (EM), Random Forest (RF), CART, and C4.5 using reconstruction errors obtained by artificially masking observed values within the training data. The resulting normalized inverse-error weights are used to combine numerical imputations through weighted averaging, whereas categorical imputations are aggregated using majority voting. Unlike an arbitrarily weighted ensemble, this mechanism derives the component contributions from their observed reconstruction performance within the training data.

The Feature Priority Assigning (FPA) module subsequently constructs a multi-criteria decision matrix using F-score, Mutual Information, and Random Forest feature-importance scores. TOPSIS is applied to rank the clinical attributes according to their relative performance across these criteria, and the top-ranked features are selected for downstream prediction. The methodological design therefore links performance-adaptive missing-value estimation with multi-criteria clinical feature prioritization in a unified prediction workflow.

To ensure a reliable evaluation, the imputation weights, preprocessing parameters, feature-ranking criteria, and feature-selection decisions are estimated exclusively from the training data within each cross-validation fold. The corresponding transformations and selected feature subset are then applied unchanged to the held-out fold, reducing the risk of information leakage. The contribution of WAELI-FPA is consequently evaluated through its individual modules and their combined effect on imputation accuracy, predictive performance, and feature-selection efficiency, rather than by attributing novelty to the established algorithms themselves.

The autoencoder, SHAP, LIME, Explainable Boosting Machine (EBM), and other explainability techniques serve complementary roles in representation learning, prediction interpretation, and transparent clinical decision support. These components are not claimed as independently novel algorithms. Instead, their value within the proposed framework must be established through implementation-specific analysis and experimental evaluation. Thus, the novelty claim of WAELI-FPA is centered on its explicitly defined adaptive weight-

ing procedure, multi-criteria feature-prioritization workflow, leakage-controlled integration, and experimentally demonstrated contribution to CKD prediction.

2. Previous Research

Chronic kidney disease (CKD) is one of the leading causes of death. Many researchers in the fields of medical informatics, artificial intelligence (AI), and machine learning (ML) have begun to pay attention to CKD. In the past ten years, various computer-based methods have been developed for the early Prediction and diagnosis of CKD from clinical data. The primary focus of these methods is improving classification accuracy through sophisticated machine learning algorithms, feature selection, ensemble learning, and deep learning. Despite several studies showing great promise in predictive performance, several challenges remain unsolved, including handling missing clinical data, selecting the most informative biomarkers, extracting clinical knowledge, and providing interpretable predictions for clinicians.

Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Logistic Regression (LR), Artificial Neural Networks (ANN), and Gradient Boosting model's machine learning algorithms have been used extensively for CKD prediction [3,7,22–34]. Random Forest and XGBoost achieved the highest classification performance because they capture nonlinear relationships among clinical variables and reduce overfitting. Ensemble learning methods are used to improve prediction accuracy. They involve combining multiple classifiers, but such methods generally assume a complete dataset and therefore have limited mechanisms for handling missing values.

Several studies have examined missing-value imputation techniques to improve data quality before classification [26–28]. Conventional approaches, such as Mean, Median, Mode, and K-Nearest Neighbor (KNN), are efficient but do not account for correlations among clinical attributes. More sophisticated statistical methods, such as Expectation-Maximization (EM), have been adopted to estimate missing values based on the underlying distribution. Imputation algorithms are also popular methods well-suited for handling complex and nonlinear medical datasets. When used intuitively, Random Forests, CART, and C4.5 are effective for imputing medical datasets. Yet existing studies typically rely on a single imputation algorithm to estimate missing values, which is not robust to heterogeneous missing-data patterns.

The CKD prediction systems also require feature selection because medical datasets often contain redundant and irrelevant attributes, thereby increasing computational complexity and reducing prediction performance [14,15,29]. Traditional feature selection techniques, such as Information Gain, Chi-Square, Recursive Feature Elimination (RFE), ReliefF, Principal Component Analysis (PCA), and metaheuristic optimization algorithms, have been used to identify significant clinical features. Typical approaches that focus extensively on features often ignore clinically significant biomarkers and rely on a single criterion for evaluation. Recently, Multi-Criteria Decision-Making (MCDM) methods such as TOPSIS have emerged as powerful alternatives for ranking features through simultaneous evaluation.

The field of Explainable Artificial Intelligence (XAI) has witnessed rapid growth and development in recent times, except in CKD studies, which employ one of the explanation methods to improve clinicians' confidence and model transparency. SHAP and LIME are the leading XAI techniques for explaining feature importance and individual predictions. The application of the Explainable Boosting Machine (EBM) has increased because it can produce an interpretable prediction model with competitive accuracy. Many existing studies employ only a single explainability technique and do not integrate predictive interaction relationship analysis or clinical rule extraction. As a result, the generated explanations are limited to evidence-based clinical decision-making.

Existing approaches to CKD prediction in the literature primarily improve classification accuracy and pay less attention to robust missing-value Imputation, intelligent feature prioritization, hybrid knowledge discovery, and comprehensive explainability. There is a great need for an integrated framework that addresses these challenges simultaneously. The conceptualized WAELI-FPA framework aims to bridge these gaps.

Recent studies have explored optimization-assisted machine learning as a means of improving predictive performance and identifying suitable model configurations. Alkhammash et al. (2022) [8] investigated optimized multivariate adaptive regression splines (MARS) for crude oil demand prediction. Their study illustrates how optimization can be incorporated into regression modeling to improve the representation of complex relationships between input variables and predicted outcomes. Although the application concerns energy-demand forecasting rather than healthcare, the methodological principle of combining model development with systematic optimization is relevant to predictive analytics. In CKD prediction, a related principle can be applied when evaluating imputation strategies, feature-prioritization procedures, and classification models, provided that all data-dependent operations are performed without information leakage.

Ensemble learning represents another important direction in machine learning research. Alkhammash et al. (2022) [11] presented a hybrid stacking model for gender recognition from voice data, integrating multiple base classifiers with a meta-classifier. Their work demonstrates an approach in which complementary learning algorithms are combined within a unified prediction architecture. Ensemble methods are relevant to clinical prediction because different algorithms may capture different aspects of the data distribution. However, the benefits of an ensemble depend on the characteristics of the dataset, the diversity of its component models, and the evaluation procedure. In the proposed WAELI-FPA framework, the ensemble principle is applied to missing-data imputation, where multiple imputation methods contribute to a weighted estimate rather than being combined solely for final classification.

Recent optimization-based research has also investigated the integration of metaheuristic algorithms with machine learning. Elshewey (2025) [10] proposed a crop-yield prediction framework using the Dove Optimization Algorithm for feature selection and Gradient Boosting for prediction, alongside comparisons with other optimization and regression methods. The study illustrates how feature selection and model optimization can be integrated into a predictive workflow. Similarly, Elshewey et al. (2026) [12] proposed a DNN–XGBoost framework for CO_2-emission prediction using binary Whale Optimization for feature selection and Grey Wolf Optimization for hyperparameter tuning. The work also incorporated SHAP analysis to support interpretation of the resulting predictions. Although these studies address agricultural and environmental prediction rather than CKD, they provide examples of integrating optimization, predictive modeling, and explainability within a single framework.

Machine learning model selection has also been investigated across diverse prediction problems. Elshewey et al. (2025) [9] compared linear regression, Ridge regression, Random Forest, Gradient Boosting, Support Vector Regression, and XGBoost for aerodynamic coefficient prediction. Their findings illustrate that the relative performance of learning algorithms depends on the prediction task and the characteristics of the dataset. This observation is relevant to CKD prediction because the performance of a classifier should be established empirically rather than assumed from results reported in another application domain. Accordingly, the present study evaluates multiple classification models under a consistent experimental protocol and identifies the preferred model based on the observed evaluation results.

Deep learning approaches provide additional mechanisms for learning complex representations from data. Elshewey et al. (2025) [13] developed a hybrid CNN–ResNet50–LSTM architecture for forecasting climate-related variables, integrating convolutional feature extraction, residual learning, and temporal modeling. This work illustrates the potential of combining complementary architectures to represent complex patterns. Nevertheless, deep learning methods may require greater computational resources and sufficient training data, and their suitability depends on the structure and size of the target dataset. For structured clinical data, their use should therefore be justified through comparative evaluation rather than assumed to be superior to conventional machine learning models.

Taken together, these studies indicate continuing methodological development in optimization-assisted learning, ensemble classification, model selection, hybrid architectures, and explainable prediction. However, their findings cannot be directly transferred to CKD prediction because the datasets, target variables, feature distributions, and evaluation settings differ. The CKD prediction problem also requires specific consideration of missing clinical observations and the contribution of individual clinical variables. The present study addresses this methodological need through WAELI-based imputation and TOPSIS-based feature prioritization, followed by classification and explainability analysis. Its contribution is therefore evaluated in the context of CKD prediction rather than inferred from the performance of methods in unrelated application domains.

This research gap is addressed by combining weighted ensemble learning-based imputation, TOPSIS-based feature priority attribution, hybrid knowledge discovery ensemble prediction, and multi-level explainable AI within a unified Clinical Decision Support System (CDSS) [19,21,23,24].

Some significant state-of-the-art techniques for predicting and detecting CKD are as follows.

- Decision Tree (DT): A Decision Tree creates a set of simple decision rules for classification; however, it can be sensitive to noisy data and prone to overfitting.
- Random Forest (RF): Random Forest combines the predictions of multiple decision trees to improve accuracy and robustness while reducing the risk of overfitting.
- Support Vector Machine (SVM): SVM can classify features into HPO categories in high-dimensional clinical datasets; however, its performance depends on careful selection of model parameters.
- K-Nearest Neighbors (KNN): KNN predicts CKD based on the similarity between patients; however, its performance can be adversely affected by noisy data.
- Naïve Bayes (NB): Naïve Bayes is a probabilistic classifier that offers efficient computation but assumes independence among features, which may not always hold in clinical datasets.
- Logistic Regression (LR): An interpretable classification method for binary CKD prediction. However, it may not adequately capture complex nonlinear relationships among clinical variables.
- XGBoost: XGBoost is a gradient-boosting algorithm that generally delivers high predictive performance and is well-suited to structured clinical data.
- Artificial Neural Networks (ANNs): ANNs can learn complex nonlinear relationships; however, they generally require substantial training data and may have limited interpretability.

As seen in the literature review specified in Table 1, while several machine learning and deep learning approaches have significantly improved CKD prediction accuracy, most studies address only one aspect, such as classification, feature selection, or explainability. Only a handful of studies attempt to reconcile robust missing-value Imputation, feature prioritization, hybrid knowledge discovery, ensemble prediction, and multi-level explainability

simultaneously. The WAELI-FPA framework addresses the lack of an easily applicable, comprehensive, accurate, and interpretable framework/solution for early diagnosis of CKD and decision support.

Table 1. Summary of Previous Research Approaches to Diagnosing CKD.

Ref.	Author(s)	Method Used	Advantages	Limitations
[1]	Polat et al.	Support Vector Machine (SVM)	High classification accuracy	Sensitive to parameter tuning
[2]	Salekin and Stankovic	Random Forest	Robust prediction and feature importance	Limited missing value handling
[3]	Rady and Anwar	Artificial Neural Network (ANN)	Learns nonlinear clinical relationships	Black-box model with low interpretability
[4]	Vijayarani and Dhayanand	Decision Tree	Simple and interpretable	Overfitting on noisy datasets
[5]	Kora and Kalva	Ensemble Learning	Improved prediction accuracy	Increased computational complexity
[6]	Aljaaf et al.	XGBoost	High predictive performance	Requires feature optimization
[7]	Chen et al.	Deep Learning	Automatic feature extraction	Requires large training datasets
[8]	Lundberg et al.	SHAP Explainability	Global and local feature explanations	No predictive interaction knowledge discovery
[9]	Ribeiro et al.	LIME	Explains individual predictions	Local explanations only
[10]	Lou et al.	Explainable Boosting Machine (EBM)	Transparent and interpretable model	Limited integration with ensemble prediction
Proposed	**This Research**	**WAELI-FPA + Hybrid Knowledge Discovery + Ensemble Learning + XAI**	**Robust missing value imputation, TOPSIS-based feature prioritization, hybrid knowledge discovery, explainable Prediction, and CDSS integration**	**Designed to overcome the limitations of existing methods**

3. Background Information

CKD is a progressive disorder that involves a gradual loss of kidney function over time [1,2]. The kidneys filter out and remove waste products from the body. They maintain the body's fluids and filter blood. Kidneys also regulate blood pressure and synthesize hormones [2]. Kidneys are responsible for filtering waste from the blood; as kidney function declines, waste and excess fluid build up, resulting in severe problems. If CKD is not treated, it ultimately reaches ESRD or end-stage renal disease, necessitating dialysis or a kidney transplant. Chronic kidney disease (CKD) is often asymptomatic at the outset. Therefore, early diagnosis is essential to reduce the risk of irreversible kidney damage and improve patient survival. The rise in the incidence of chronic kidney disease (CKD) is attributed to lifestyle practices, aging, DM, hypertension, obesity, and cardiovascular disease. As per Global Burden of Disease (GBD) reports, CKD has emerged as one of the most common causes of death globally, affecting nearly 10–13% of the global population. If high-risk patients can be identified early, clinicians can initiate preventive treatments to slow disease progression and reduce associated care costs. As a result, research on intelligent computer-aided diagnostic systems for the early Prediction of CKD has gained importance.

3.1. Chronic Kidney Disease Diagnosis

The diagnosis of CKD is generally based on clinical examinations, laboratory investigations, medical history, and imaging studies. Common laboratory tests include serum creatinine, blood urea, blood glucose, hemoglobin, albumin, sodium, potassium, packed cell volume (PCV), white blood cell count (WBC), and red blood cell count (RBC). Kidney function is commonly evaluated using the Estimated Glomerular Filtration Rate (eGFR) and the presence of protein or albumin in urine. Clinical datasets containing these parameters are widely used to develop machine learning models to predict early-stage CKD.

However, healthcare datasets frequently suffer from missing values, inconsistent measurements, noisy records, outliers, and redundant features [26]. These challenges significantly reduce the effectiveness of predictive models and may lead to inaccurate clinical decisions. Consequently, efficient preprocessing techniques, robust Imputation of missing values, and intelligent feature selection have become essential components of AI-based CKD diagnosis.

3.2. Machine Learning in CKD Prediction

Machine Learning in Healthcare: Computers learn patterns from available historical clinical data and predict diseases accurately. Supervised learning algorithms, such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbor (KNN), XGBoost, and Artificial Neural Networks (ANN), are widely used for CKD prediction [4–6]. These computerized models analyze patient records and categorize patients as CKD or Non-CKD based on their clinical characteristics.

Data quality and feature relevance significantly influence the performance of these algorithms, despite their high prediction accuracy. When incomplete clinical data and redundant attributes are used, classification efficiency and model generalization are reduced. Enhancing data quality before model training is essential for trustworthy CKD prediction.

3.3. Missing Value Imputation

Medical datasets often suffer from missing values for various reasons, such as incomplete reports, test unavailability, equipment failure, and human error in data collection. Neglecting missing values or removing incomplete records can lead to information loss and bias in predictive models.

There are several imputation methods, such as Mean, Median, Mode, KNN, EM, RF, CART, C4.5 Decision Tree, and so on. There are distinctive strengths and limitations in each technique. Statistical methods are efficient but do not preserve complex relationships among features. Tree-based methods capture nonlinear dependencies but do not work well across all datasets. Consequently, using a weighted ensemble strategy to combine multiple imputation algorithms may increase imputation accuracy and robustness.

3.4. Feature Selection and Feature Prioritization

Medical datasets often include several clinical attributes, but some may be irrelevant or redundant for disease prediction. Feature selection is the process of identifying the best variables while simultaneously removing the irrelevant attributes. This can lead to better predictions, lower computational complexity, and improved model interpretability.

The traditional methods of feature selection are Information Gain, Chi-Square, ReliefF, RFE, PCA [14], and optimization-based methods. In recent years, techniques used in Multi-Criteria Decision Making (MCDM), such as the ideal solution (TOPSIS) [15], have gained popularity for evaluating features with respect to multiple criteria simultaneously. The

WAELI-Feature Priority Assigning module, based on TOPSIS, selects clinical attributes and ranks illness symptoms to identify the most useful feature for CKD.

Weighted Imputation:

$$X_{WAELI} = \hat{a}w_i X_i$$

$$Weight = \hat{a}w_i = 1$$

where i is the output of EM, RF, CART, and C4.5. Higher-performing imputation methods receive higher weights.

The computational complexity of the overall framework can be expressed as the cumulative complexity of its constituent modules:

$$T_{overall} = T_{pre} + T_{WAELI} + T_{FPA} + T_{HKD} + T_{EnS} + T_{XAI} + T_{CDSS}$$

where T_{pre}, T_{WAELI}, T_{FPA}, T_{HKD}, T_{EnS}, T_{XAI}, T_{CDSS} and represent the computational costs of preprocessing, ensemble imputation, feature prioritization, hybrid knowledge discovery, ensemble prediction, explainability, and clinical decision support, respectively.

3.5. Ensemble Learning

Ensemble learning combines multiple machine learning models to produce a more accurate and robust prediction than any individual classifier. Instead of relying on a single algorithm, ensemble methods aggregate predictions from several models to reduce bias, variance, and overfitting.

Popular ensemble algorithms include Random Forest, Gradient Boosting, XGBoost, Extra Trees, AdaBoost, and LightGBM. Voting, bagging, boosting, and stacking are commonly used ensemble strategies. In this research, multiple classifiers are integrated using weighted voting and stacking to enhance CKD prediction performance and improve model generalization [4,6].

3.6. Explainable Artificial Intelligence (XAI)

Artificial Intelligence has achieved remarkable success in disease prediction; however, many machine learning models operate as "black boxes," making it difficult for clinicians to understand how predictions are generated. Explainable Artificial Intelligence (XAI) addresses this limitation by providing clear explanations of prediction results, improving clinicians' trust in performing evidence-based medicine.

Some of the most popular XAI techniques include SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations), Explainable Boosting Machine (EBM), and more. They provide both global explanations (the importance of all features) and local explanations (predictions for individual patients). Combining different XAI techniques enables doctors to understand not just which features drive these predictions, but also why and how those features cause the disease.

3.7. Cclinical Decision Support System

A Clinical Decision Support System (CDSS) is an intelligent system that helps physicians diagnose diseases and make evidence-based clinical decisions about treatment. Current CDSS commonly integrate patient data, machine learning algorithms, clinical guidelines, and Explainable Artificial Intelligence (XAI) to provide personalized diagnostic recommendations and enhance clinical decision-making [22,35,36]. The proposed research utilizes a comprehensive clinical decision support system (CDSS) consisting of Imputation of missing values, prioritization of features, hybridized knowledge generation, ensemble prediction, and multi-level explanation. The system can predict CKD risk, identify the most significant clinical biomarkers that influence the Prediction, and provide explanations so that healthcare professionals can make reliable and informed clinical decisions [35,37].

4. Approach

As shown in Figure 1, the architecture of the Weighted Average Ensemble Learning Imputation with Feature Priority Assigning (WAELI-FPA) framework is presented for chronic kidney disease (CKD) prediction. The eight layers of the framework are sequential; thus, the output of one layer becomes the input to the next. The architecture comprises data preprocessing, Imputation through weighted ensemble learning, intelligent feature prioritization, Hybrid knowledge discovery, ensemble classification, Explainable AI (XAI), and a clinical decision support system (CDSS). The result of the Architecture is accurate, robust, and predictive for CKD.

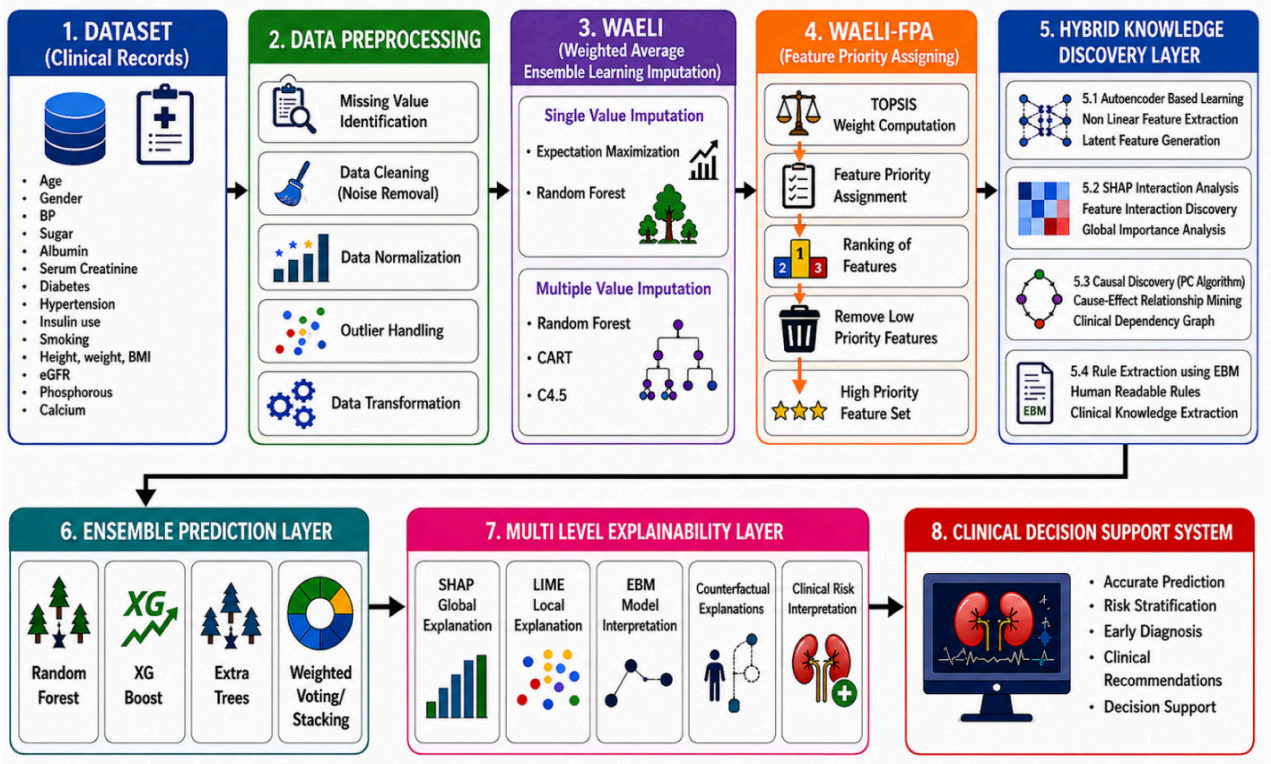

Figure 1. Proposed WAELI-FPA Architecture for Explainable CKD Prediction.

The frameworkspecified in Figure 1 consists of eight integrated layers: (1) Clinical Dataset, (2) Data Preprocessing, (3) Weighted Average Ensemble Learning Imputation (WAELI), (4) Feature Priority Assigning (WAELI-FPA), (5) Hybrid Knowledge Discovery Layer, (6) Ensemble Prediction Layer, (7) Multi-Level Explainability Layer, and (8) Clinical Decision Support System. Different colors are used to distinguish the individual layers and improve the visual clarity and interpretation of the proposed framework.

4.1. The WAELI-FPA Framework

The proposed work, the Weighted Average Ensemble Learning Imputation with Feature Priority Assigning (WAELI-FPA) framework, is developed to facilitate the early Prediction of chronic kidney disease (CKD) through the Imputation of missing clinical values, the selection of effective biomarkers, robust classification, and explainable prediction support for clinical decision-making. A structured sequential pipeline includes seven major modules: first, WAELI preprocessing; second, WAELI imputation, WAELI Feature Prioritizing Assigning (WAELI-FPA), Hybrid Knowledge Discovery, Ensemble Prediction, Multi-level explanation, and clinical decision support system (CDSS). Each module processes the output from the previous stage, creating an end-to-end CKD prediction workflow.

4.2. Data Preprocessing

The first phase of the proposed framework is to prepare the CKD dataset for reliable machine learning analysis. Incomplete records, missing lab measurements, inconsistencies, and misinterpretations of values and attributes on different numerical scales often plague clinical datasets. Thus, before creating predictive models, preprocessing should be done. The National Health and Nutrition Examination Survey Clinical Database 29, National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC) 30, USA. The dataset used in the work was obtained from NHANES [22]. You can get the NHANES database from the CDC website: https://www.cdc.gov/nchs/nhanes/ (accessed on 1 June 2026). NHANES is a reliable source with comprehensive demographic, laboratory, physiological, and lifestyle information for predicting Chronic Kidney Disease (CKD) https://www.cdc.gov/nchs/nhanes/ (1 June 2026).

The CKD dataset specified in Table 2 used in the implementation on 1 June 2026 has 11,933 patient records comprising 29 variables. The preprocessing phase identifies missing values, cleans the data, prepares the dataset for imputation, and standardizes feature values via normalization. After the preprocessing phase, 25 features remained. Of these, 17 were numerical features, and 8 were categorical ones.

Table 2. Characteristics of the CKD Dataset Used in this Study.

Characteristic	Description
Dataset Source	NHANES Clinical Database
Organization	National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), USA
Dataset Access Link	https://www.cdc.gov/nchs/nhanes/
Date Accessed	1 June 2026
Disease	Chronic Kidney Disease
Data Type	Mixed Numerical and Categorical
Clinical Variables	Demographic, Laboratory, Physiological, Lifestyle
Missing Values	Present in multiple laboratory attributes
Target Variable	CKD/Non-CKD
Total Selected Features	25
Final High-Priority Features	15 (after WAELI-FPA)
Dataset Size	• 11,933 records • 29 variables • 25 retained variables

4.2.1. Leakage-Controlled Preprocessing

To prevent data leakage, all data-dependent preprocessing and model-selection operations were performed exclusively within the training folds of the stratified five-fold cross-validation. For each fold, WAELI imputation models and their ensemble weights were learned solely from the training data. Normalization parameters were estimated from the training fold and applied unchanged to the corresponding test fold. Similarly, F-score, Mutual Information, and Random Forest feature importance were calculated only from the training fold, and TOPSIS was subsequently used to select the top 60% of features. The selected feature subset was applied unchanged to the test fold. Hyperparameter tuning was performed using inner cross-validation within the training fold, while the outer test fold was used only for final evaluation. Thus, no information from the held-out test

fold was used for Imputation, normalization, feature prioritization, feature selection, or hyperparameter optimization.

4.2.2. Analysis of Absence of Data

The framework starts by checking the rate of missing values for each clinical variable. There is substantial missingness in several variables. To illustrate, the variables insulin_use, diabetes_pills, current_smokerhas 90.94%, 80.88%, 72.82%, respectively, while the biochemical variables serum_creatinine, blood_urea_nitrogen, eGFR, phosphorus has approximately 47% each. If one were to consider using advanced methods for replacing missing data rather than simple statistical ones, one's results would be more accurate.

4.2.3. Data Cleaning

After identifying missing values specified in Table 3, the dataset is cleaned to ensure variable consistency and readiness for processing. Upon completing this stage, the model retains 25 clinically relevant variables from the original dataset, which are broadly classified as numerical or categorical. Data normalization process. To make sure all clinical variables contribute equally during model training, numerical attributes are normalized. The output creates a normalized data matrix with scaled values for Age, body mass index, blood pressure, serum creatinine, blood urea nitrogen, phosphorus, calcium, bicarbonate, urine creatinine, urine albumin, and albumin-to-creatinine ratio. The implementation output contains a sample of normalized values. Weighted Average Ensemble Learning Imputation (WAELI). The dataset still has missing clinical values after preprocessing. The proposed Weighted Average Ensemble Learning Imputation (WAELI) module estimates these values using several imputation algorithms rather than a single one. The first step in the framework is to assess the contribution of each imputation algorithm and assign an ensemble weight to each. The ensemble consists of four algorithms.

- Expectation-Maximization (EM);
- Random Forest (RF);
- CART;
- C4.5 Decision Tree.

Table 3. Summary of Missing Value Analysis.

Attribute	Missing (%)
insulin_use	90.94
diabetes_pills	80.88
current_smoker	72.82
Phosphorus	47.02
Bicarbonate	47.00
serum_creatinine	46.99
eGFR	46.99
Calcium	46.69

Table 4 shows Training-fold estimation of WAELI weights. WAELI combines four component imputers: Expectation-Maximization (EM), Random Forest (RF), Classification and Regression Trees (CART), and C4.5. The contribution of each component is determined from its imputation performance rather than from an arbitrary fixed preference. Within each training fold, a subset of observed entries is temporarily masked. Each component is

fitted using the remaining observed training values and is used to reconstruct the masked entries. The reconstructed values are compared with the original values, which are available because the masking operation is applied only to observed entries. The resulting component errors are used to derive normalized non-negative weights. This procedure allows the ensemble to assign larger contributions to components that perform more accurately for the characteristics of the training data.

Table 4. WAELI algorithm calculation Mean Weight and SD.

Imputation Algorithm	Mean Weight ± SD	Weight
Expectation Maximization	0.2958 ± 0.0112	0.2921
Random Forest	0.2742 ± 0.0082	0.2834
CART	0.2108 ± 0.0052	0.2073
C4.5 Decision Tree	0.2192 ± 0.0032	0.2171

Mathematical Weight Formula

Let $M = \{\text{EM, RF, CART, C4.5}\}$ denote the set of component imputers. For numerical variables, the masked-entry RMSE of component m is

$$E_m^{\text{num}} = \sqrt{\frac{1}{|\Omega_{\text{mask}}|} \sum_{(i,j)\in\Omega_{\text{mask}}} \left(x_{ij} - \hat{x_{ij}}^{(m)}\right)^2}$$

where $\boldsymbol{\Omega_{\text{mask}}}$ is the set of artificially masked observed entries, and $\hat{\boldsymbol{x_{ij}}}^{(m)}$ is the estimate generated by the component m.

If categorical variables are included, the categorical error of the component m is

$$\boldsymbol{E_m^{\text{cat}}} = \frac{\mathbf{1}}{|\,\boldsymbol{\Omega}_{\text{mask}}^{\text{cat}}\,|} \sum\nolimits_{(i|j)\in\boldsymbol{\Omega}_{\text{mask}}^{\text{cat}}} \mathbb{I}\left(\boldsymbol{x_{ij}} \neq \hat{\boldsymbol{x}}_{\boldsymbol{ij}}^{(m)}\right).$$

The normalized component error is calculated as

$$\boldsymbol{E_m} = \boldsymbol{\alpha} \boldsymbol{E_m^{\text{num}}} + (\mathbf{1} - \boldsymbol{\alpha}) \boldsymbol{E_m^{\text{cat}}},\ \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1},$$

where $\boldsymbol{\alpha}$ is the prespecified contribution of the numerical-error term. If weights are estimated separately for numerical and categorical variables, $\boldsymbol{E_m^{\text{num}}}$ and $\boldsymbol{E_m^{\text{cat}}}$ should instead be used independently.

The reciprocal-error score is then

$$\boldsymbol{q_m} = \frac{\mathbf{1}}{\boldsymbol{E_m} + \boldsymbol{\epsilon}},$$

where $\boldsymbol{\epsilon} > 0$ is a small constant. The normalized WAELI weight is

$$\boldsymbol{w_m} = \frac{\boldsymbol{q_m}}{\sum_{\boldsymbol{\ell}\in\boldsymbol{\mathcal{M}}} \boldsymbol{q_\ell}},\ \boldsymbol{w_m} \geq \mathbf{0},\ \sum_{\boldsymbol{m}\in\boldsymbol{\mathcal{M}}} \boldsymbol{w_m} = \mathbf{1}.$$

The stronger formulation is target-specific weighting:

$$\boldsymbol{w_{mj}} = \frac{\mathbf{1}/\left(\boldsymbol{E_{mj}} + \boldsymbol{\epsilon}\right)}{\sum_{\boldsymbol{\ell}=\mathbf{1}}^{\boldsymbol{M}} \mathbf{1}/\left(\boldsymbol{E_{\ell j}} + \boldsymbol{\epsilon}\right)}.$$

Then the final numerical estimate is:

$$\hat{\boldsymbol{x}}_{\boldsymbol{ij}}^{\textbf{WAELI}} = \sum\nolimits_{\boldsymbol{m}=\mathbf{1}}^{\boldsymbol{M}} \boldsymbol{w_{mj}} \hat{\boldsymbol{x}}_{\boldsymbol{ij}}^{(m)}.$$

Use w_{mj} if the weight changes with the target feature. Use $\hat{x}_{ij}^{(m)}$ only if one global weight is used for all features.

Weight-Estimation Algorithm

Algorithm 1 signifies that Every algorithm assesses missing values alone. The weighted averaging of their predictions produces the final imputed values. It also performs categorical Imputation via majority voting. Through the WAELI process, we ensure that the imputation phase has been executed correctly since there are no missing values at the end of the process. Assigning Priority to WAELI Feature. Once Imputation is done, the complete dataset goes through the proposed WAELI Feature Priority Assigning (WAELI-FPA) module. In the first step (Figure 1), the most informative clinical variables will be selected using multi-criteria feature ranking strategies that employ the TOPSIS method. The implementation first extracts a decision matrix with three evaluation criteria:

- F-score;
- Mutual Information (MI) score;
- Random Forest feature-importance score.

Algorithm 1. Training-Fold Estimation of WAELI Weights

Input: Training data $\boldsymbol{X}_{\mathbf{train}}$, observed-entry mask $\boldsymbol{R}$, component set $\boldsymbol{M}$, masking rate $\boldsymbol{\rho}$, number of repetitions $\boldsymbol{B}$, and $\boldsymbol{\epsilon} > \mathbf{0}$.
Output: Component weights w_m or target-specific weights w_{mj}.
1. Identify observed numerical and categorical entries in the training data.
2. For each masking repetition $b = 1, \ldots, B$:
3. Randomly mask a proportion of the observed training entries.
4. Fit EM, RF, CART, and C4.5 using the remaining observed training entries.
5. Reconstruct the masked numerical and categorical entries.
6. Calculate RMSE/MAE for numerical entries and error rate for categorical entries.
7. Average the component errors across the masking repetitions.
8. Convert each average error into an inverse-error score:
   $q_m = \frac{1}{E_m + \epsilon}$
9. Normalize the scores:
   $w_m = \frac{q_m}{\sum_{l=1}^{M} q_l}$.
10. Refit the four component imputers using all available observed training values.
11. Impute the genuinely missing training and validation entries using the training-derived weights.
12. Preserve all originally observed entries.
13. Repeat the complete procedure independently in every cross-validation training fold.

The importance of each feature is subsequently evaluated using the Random Forest feature-importance measure, which quantifies the contribution of individual clinical variables to the prediction task. The engine calculates a priority score for each feature and gives the final ranking.

Table 5 specifies the Post-ranking, the procedure selects the top 60% of the ranked features, reducing the original feature space from 25 attributes to 15 clinical variables. The input for the subsequent learning stages is this optimized subset. The Hybrid Knowledge Discovery Layer receives the optimized feature subset as input. The WAELI-FPA stage precedes ensemble prediction in this layer implementation output. According to the uploaded document, this layer is referred to as the HYBRID Layer, but it does not include any internal implementation details of individual components or algorithms. Accordingly, this approach will register the layer as an intermediate knowledge-discovery module, assigning to it no further processing than the source allows—at the ensemble-prediction level. The enhanced feature selection is provided to an ensemble prediction module that uses Stratified 5-Fold Cross-Validation. Four classifiers are evaluated:

- Kinds of Random Forest;

- Additional Trees;
- XGBoost;
- AdaBoost.

Table 5. Top-Ranked Clinical Features.

Rank	Feature
1	height_cm
2	serum_creatinine
3	weight_kg
4	albumin_creatinine_ratio
5	phosphorus
6	Age
7	urine_albumin
8	blood_urea_nitrogen
9	bicarbonate
10	calcium

XGBoost: Extreme Gradient Boosting. It is an ensemble learning method based on decision trees and has demonstrated strong predictive performance in various machine learning applications. Its computational efficiency, scalability, and regularization mechanisms make it effective for reducing overfitting and improving model generalization.

AdaBoost: AdaBoost is an ensemble learning technique that combines multiple weak learners to construct a stronger predictive model. In the proposed framework, the performance of the ensemble models is assessed based on their effectiveness and computational efficiency.

Among all assessed classifiers from Table 6, XGBoost has the highest predictive performance and is ultimately selected as the champion predictive model. The implementation also showcases charts for comparative accuracy, ROC curves, and confusion matrices for performance evaluation.

Table 6. Comparative Performance of Ensemble Models.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	93.35	95.89	94.55	95.21
Extra Trees	84.47	84.07	96.00	89.63
XGBoost	95.89	97.34	96.77	97.05
AdaBoost	75.61	88.68	77.83	81.65

4.2.4. Multi-Level Explainability Layer

The proposed WAELI-FPA framework achieves higher transparency and interpretability by incorporating various Explainable Artificial Intelligence (XAI) techniques through the Multi-Level Explainability Layer. After SHAP calculates the Use of Ensemble Prediction, Global and Local Feature Importance to measure the Contribution of Clinical Attributes to the Prediction. Explainability refers to our capability to depict and explain predictions in a human-friendly manner. Explainable boosting machine (EBM) automatically produces interpretable decision rules containing information about the effects of clinical parameters on the predictions. Counterfactual explanations quantify the minimum change in clinical

parameters required to change the outcome with the help of clinical experts. The combination of these complementary techniques provides comprehensive model interpretability, improving clinician trust and reliability of the CDSS.

The model predicts CKD Present in the illustrated patient with 92.89% probability. The most important rule selected by LIME is albumin_creatinine_ratio, while SHAP selects albumin_creatinine_ratio, urine_albumin, and blood pressure as the most important.

4.2.5. Clinical Decision Support System (CDSS)

Clinical Decision Support System (CDSS). The final stage integrates the prediction and explainability outputs into a prototype Clinical Decision Support System (CDSS). The system presents the estimated CKD probability, risk classification, contributing clinical features, and model-generated explanations to support interpretation of the model prediction. For the illustrative patient, the CDSS generated a CKD probability score of 98.34% and classified the patient as high risk. These outputs demonstrate the functionality of the proposed decision-support prototype and should not be interpreted as a clinical diagnosis or treatment recommendation.

The CDSS is designed to assist, rather than replace, professional clinical judgment. The generated explanations are intended to help users understand the contributing clinical features and factors influencing the model prediction. For the illustrative patient, the CDSS generated a CKD probability score of 98.34% and classified the patient as high risk. These outputs demonstrate the functionality of the proposed decision-support prototype and should not be interpreted as a clinical diagnosis or treatment recommendation. The clinical validity, usability, safety, and impact of the proposed CDSS on clinical decision-making have not been established in the present study. Therefore, independent external validation using patient cohorts from different clinical settings, prospective clinical evaluation, and assessment by qualified healthcare professionals are required before considering clinical deployment.

The statistical comparison performed using McNemar's test yielded a chi-squared statistic of 0.1452 with a p-value of 0.70320. At the reported significance level, the null hypothesis could not be rejected, indicating no statistically significant difference between the compared models. This statistical result is reported as part of the model evaluation and should not be interpreted as evidence of clinical validation of the proposed CDSS.

Illustrative CDSS Output:

1. CKD Probability Score: 98.34%;
2. Risk Classification: High Risk;
3. CKD Indicators: Model indicators suggest the presence of CKD.

Model Interpretation: The contributing clinical features and explanations are presented to facilitate interpretation of the Prediction.

The prototype, therefore, demonstrates how the prediction and explainability components can be integrated into a CDSS interface; however, the reported outputs represent model-generated decision support and not clinically validated diagnostic or treatment recommendations.

5. Experimental Results and Discussion

This chapter presents the evaluation of the proposed Weighted Average Ensemble Learning Imputation with Feature Priority Assigning (WAELI-FPA) framework. The experimental findings are analyzed according to the architectural layers of the proposed framework. Each module is evaluated individually, followed by an overall discussion emphasizing the effectiveness of the proposed methodology for CKD Prediction. The results, graphs, and statistics presented in this chapter are obtained from running the WAELI-FPA framework.

5.1. Experimental Setup

The WAELI-FPA framework was implemented in Python 3.11 on a Windows workstation with an i7-12700 CPU, 16 GB of RAM, Windows 11 Pro, and an NVIDIA RTX 3060 GPU (optional). The Scikit-learn, XGBoost, SHAP (0.49.1), LIME (0.2.0.1), Pandas (2.3.3), NumPy (1.26.4), Matplotlib (3.10.9), and Seaborn (22.3.1) libraries were implemented. Stratified five-fold cross-validation was used for model evaluation. Hyperparameters were optimized using Grid Search with inner cross-validation within each training fold, ensuring that the held-out test fold was not used during model selection.

5.2. Results of Data Preconditioning Stage

Table 7 specifies that Initially, the CKD dataset had 11,933 patients and 29 clinical variables. During preprocessing, irrelevant and redundant attributes were removed, resulting in 25 clinically useful features. Out of the 25 features, 17 are numerical, and 8 are Categorical. Before executing the machine learning algorithm, it was necessary to preprocess the dataset.

Table 7. Dataset Statistics.

Parameter	Value
Total Records	11,933
Original Features	29
Retained Features	25
Numerical Features	17
Categorical Features	8

5.3. Missing Value Analysis

The preprocessing module identified significant missing values across several clinical variables.

Table 8 contains the outcomes clearly reveal that many laboratory results and patient history variables are missing from the dataset. Almost half of the biochemical parameters required Imputation, while lifestyle-related variables had an even higher lack of completeness. We cannot apply a dataset directly to predictive modeling, as missing data can severely impair predictive power and reliability. Thus, an advanced missing-value recovery mechanism is required before classification. Outcomes of WAELI Imputation Layer 5.3: The WAELI module used four complementary algorithms to impute missing values after preprocessing.

Table 8. Missing Values.

Attribute	Missing (%)
Insulin Use	90.94
Diabetes Pills	80.88
Current Smoker	72.82
Serum Creatinine	46.99
Blood Urea Nitrogen	46.99
eGFR	46.99
Phosphorus	47.02

Table 9 clearly specifies the experimental results indicate that Expectation-Maximization received the highest ensemble weight, thereby contributing most to missing-value estimation. The Random Forest also got a similar weight, indicating its ability to capture nonlinear relationships amongst clinical variables. The ensemble used the estimates from CART and C4.5 to complement its own estimates. After applying weighted averaging to numerical variables and majority voting to categorical variables, all missing entries were filled in. However, completion of missing entries alone does not establish imputation accuracy. Therefore, WAELI's performance was additionally evaluated using artificially masked observed values and quantitative error measures, including MAE and RMSE. For each cross-validation fold, the WAELI-FPA feature priority assigning module was fitted using only the training subset. F-score, Mutual Information, and Random Forest feature importance were calculated from the training data, and the resulting TOPSIS ranking was used to select the top 60% of features. The selected features were then applied unchanged to the corresponding held-out test subset.

Table 9. WAELI Ensemble Weights.

Algorithm	Weight
Expectation Maximization	0.2921
Random Forest	0.2834
CART	0.2073
C4.5	0.2171

5.3.1. Quantitative Validation of WAELI Imputation

To quantitatively evaluate the effectiveness of the proposed Weighted Average Ensemble Learning Imputation (WAELI) method, a controlled artificial masking experiment was conducted using the CKD dataset. A subset of originally observed values was randomly masked, while the corresponding original values were retained as ground truth for evaluating reconstruction accuracy. The masked values were reconstructed using Expectation-Maximization (EM), Random Forest (RF), Classification and Regression Trees (CART), C4.5, and the proposed WAELI method. All methods were evaluated using identical masking patterns to ensure a fair comparison.

For numerical variables, imputation performance was assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), normalized Mean Absolute Error (NMAE), normalized Root Mean Square Error (NRMSE), and the coefficient of determination (R^2). Lower MAE, RMSE, NMAE, and NRMSE values indicate more accurate reconstruction, whereas a higher R^2 indicates better agreement between the imputed and original values. For categorical variables, performance was evaluated using categorical imputation accuracy and the proportion of falsely classified categories (PFC), with higher accuracy and lower PFC indicating better performance.

The initial numerical imputation experiment involved randomly masking 5% of the originally observed numerical values. The experiment was evaluated using stratified five-fold cross-validation. Within each fold, imputation models and ensemble weights were estimated using the training data, and the artificially masked values were withheld from the imputation process and used only for evaluation. The normalization procedures for NMAE and NRMSE were defined consistently across all evaluated methods.

As reported in Table 10, the proposed WAELI method achieved an NMAE of 0.4524 ± 0.0126, an NRMSE of 0.6977 ± 0.0393, and an R^2 of 0.8955 ± 0.0973. These results indicate better numerical reconstruction performance than the individual component methods under the reported 5% masking condition.

Table 10. Quantitative Evaluation of Imputation Methods.

Imputation Method	NMAE ↓	NRMSE ↓	R^2 ↓
Expectation Maximization (EM)	0.5077 ± 0.0151	0.7455 ± 0.0428	0.8232 ± 0.0380
Random Forest (RF)	0.4691 ± 0.0095	0.7230 ± 0.0314	0.8336 ± 0.0513
CART	0.4596 ± 0.0149	0.7432 ± 0.0400	0.8335 ± 0.1788
C4.5	0.4642 ± 0.0121	0.7311 ± 0.0431	0.8279 ± 0.1784
WAELI (Proposed)	**0.4524 ± 0.0126**	**0.6977 ± 0.0393**	**0.8955 ± 0.0973**

To further examine the robustness of WAELI, the evaluation will be extended to artificial masking levels of 10%, 20%, and 30%. The same masking locations will be used for all methods at each level, and the experiment will be repeated to obtain mean and standard deviation estimates. Numerical results will be reported using raw and normalized error metrics, while categorical results will be presented separately using categorical accuracy and PFC. This analysis will establish whether the performance advantage of WAELI is maintained under increasing missingness and across different clinical variable types.

The quantitative imputation evaluation complements the downstream CKD classification experiments by directly assessing the accuracy of reconstructed clinical values. Together, these evaluations provide a more comprehensive assessment of the proposed WAELI framework.

$$MAE = \frac{1}{N} = \sum_{i=1}^{N} |x_i - \hat{x}_i|$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}$$

$$R2 = 1 - \frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{N}(x_i - \overline{x}_i)^2}$$

where x_i is the original observed value, $\hat{x}_i$ is the imputed value, and N is the number of artificially masked observations.

Quantitative Imputation Validation: To quantitatively evaluate the proposed WAELI method, 5% of originally observed numerical values were randomly masked and treated as missing. In contrast, their original values were retained as ground truth. The experiment was repeated using a stratified five-fold validation. Imputation performance was evaluated using the normalized mean absolute error (NMAE), the normalized root mean square error (NRMSE), and the coefficient of determination. Lower NMAE and NRMSE and higher indicate better imputation accuracy. As shown in Table 10, WAELI achieved the lowest NMAE (0.4524 ± 0.0126) and NRMSE (0.6977 ± 0.0393), along with the highest (0.8955 ± 0.0973), demonstrating improved reconstruction performance compared with the individual imputation methods.

5.3.2. Evaluation of WAELI Under Different Missing-Data Mechanisms

To further investigate the robustness and reliability of the proposed Weighted Average Ensemble Learning Imputation (WAELI) method, an evaluation under different missing-data mechanisms, namely Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), is proposed. These mechanisms represent different patterns of missingness in clinical datasets and help assess the applicability of WAELI under varying data-incompleteness conditions.

Under MCAR, missing values are introduced independently of observed and unobserved data. Under MAR, the probability of missingness depends on other observed clinical attributes. Under MNAR, the probability of missingness depends on the values of the variables being masked or other unobserved information. The evaluation will consider artificial

masking levels of 5%, 10%, 20%, and 30% to examine the effect of increasing missingness on imputation performance.

The proposed WAELI method will be compared with Expectation-Maximization (EM), Random Forest (RF), CART, and C4.5 using identical data partitions and masking patterns. Numerical imputation performance will be assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), normalized MAE (NMAE), normalized RMSE (NRMSE), and the coefficient of determination (R^2). Categorical imputation performance will be assessed using categorical accuracy and the proportion of falsely classified categories (PFC). Results will be summarized using mean and standard deviation across repeated experiments.

To prevent information leakage, preprocessing parameters, component models, and ensemble weights will be estimated exclusively from the training data. The original values of artificially masked observations will be withheld from the imputation process and used only for evaluation. Particular attention will be given to the performance of WAELI under MAR and MNAR conditions, where missingness may be associated with clinical characteristics or the values of the variables themselves.

This evaluation will help determine whether the performance of WAELI remains consistent across different missingness mechanisms and increasing missingness levels. It will also identify conditions under which the method may experience reduced reconstruction accuracy. Since the true values of naturally missing observations are unavailable, the results will be interpreted in the context of the specified artificial masking mechanisms and their assumptions.

5.3.3. Sensitivity Analysis of Highly Incomplete Variables

The CKD dataset contains variables with substantial proportions of missing observations, including insulin_use (90.94%), diabetes_pills (80.88%), and current_smoker (72.82%). Although these variables may contain clinically relevant information, their high missingness can reduce imputation reliability because only a limited number of observed values are available to estimate the missing entries. Consequently, retaining these variables without evaluating their influence may introduce uncertainty into the imputed dataset and subsequent CKD prediction.

To investigate this issue, a sensitivity analysis should compare three feature-retention configurations: (i) the original feature set, including all 25 retained variables; (ii) a reduced feature set excluding insulin_use, which has more than 90% missing observations; and (iii) a reduced feature set excluding all three highly incomplete variables (insulin_use, diabetes_pills, and current_smoker). The same data partitions, preprocessing procedures, imputation methods, feature-selection strategy, and prediction models should be used across configurations to ensure a fair comparison. All data-dependent preprocessing and feature-selection operations should be fitted exclusively on the training folds to prevent data leakage.

The configurations should be evaluated using numerical imputation metrics, including MAE and RMSE, and appropriate categorical imputation metrics, such as accuracy and the proportion of falsely classified categorical values. Downstream CKD prediction performance should also be compared using the evaluation metrics adopted in this study. Results should be reported as mean $\pm$ standard deviation across the cross-validation folds. In addition, imputation performance should be examined separately for each highly incomplete variable using only originally observed values that are artificially masked for validation. This distinction is important because the original values of naturally missing observations are unavailable for direct error calculation.

The analysis should determine whether retaining highly incomplete variables improves predictive performance, produces negligible changes, or degrades the reliability of the overall framework. The final decision should consider imputation accuracy, downstream predictive performance, and clinical relevance. Since the sensitivity-analysis experiments must be performed before their outcomes can be established, no conclusion regarding the superiority of either feature-retention strategy should be drawn in advance.

5.3.4. Comparative Evaluation of WAELI Against Established Imputation Methods

To establish the effectiveness of the proposed Weighted Average Ensemble Learning Imputation (WAELI) method, a comparative evaluation is required against commonly used statistical imputation techniques and established machine learning-based methods. The comparison includes mean imputation for numerical variables, median imputation for numerical variables with skewed distributions or potential outliers, mode imputation for categorical variables, K-Nearest Neighbors (KNN) imputation, Multivariate Imputation by Chained Equations (MICE), MissForest, and Iterative Imputer. These methods represent different strategies for handling incomplete data, ranging from simple univariate replacement to iterative multivariate estimation and nonlinear prediction. Previous studies of clinical and mixed-type datasets have compared these methods and demonstrated that their relative performance can vary with variable type, missingness rate, and relationships among observed features. Mean, median, and mode imputation are included as baseline approaches because they are computationally inexpensive and provide reference performance levels against which more sophisticated methods can be evaluated. KNN imputation estimates missing values using information from similar observations, whereas MICE models incomplete variables conditionally on other variables through an iterative imputation procedure. MissForest employs random forest models to estimate missing numerical and categorical values and can capture nonlinear relationships among variables. Iterative Imputer repeatedly estimates missing values using a specified predictive estimator, such as a regression model. These methods provide complementary baselines for evaluating whether the proposed WAELI approach offers advantages over simple statistical replacement and established multivariate imputation strategies. A controlled evaluation should be conducted by masking a subset of originally observed values in the training data, imputing these artificially masked values using each method, and comparing the imputed values with their known original values. The evaluation should be performed separately for numerical and categorical variables because the appropriate error metrics differ by data type. For numerical variables, the comparison may use normalized root mean square error (NRMSE), normalized mean absolute error (NMAE), and the coefficient of determination (R^2), provided that the normalization procedure is clearly specified. For categorical variables, the proportion of falsely classified categories (PFC) or categorical imputation accuracy may be reported. Each method should be evaluated using the same artificially masked entries, repeated masking experiments, and consistent preprocessing procedures to support a fair comparison.

In addition to overall imputation accuracy, the influence of the imputation method on downstream CKD prediction should be investigated. Each imputed dataset should be supplied to the same classification model, using identical training and validation partitions and a consistent feature-selection procedure. Accuracy, Precision, Recall, F1-score, and ROC-AUC can then be compared across the imputation methods. This analysis is important because a method that achieves low reconstruction error does not necessarily produce the best downstream classification performance. Reporting both imputation-specific metrics and classification metrics enables the evaluation to distinguish the accuracy of the reconstructed values from their usefulness for CKD prediction.

Particular attention should be given to attributes with high proportions of missing observations. In the current dataset, the reported missingness rates for insulin_use, diabetes_pills, and current_smoker are 90.94%, 80.88%, and 72.82%, respectively. At these levels, only a relatively small proportion of the observations contain recorded values for the corresponding attributes. Consequently, the performance of any imputation method may be constrained by the limited information available for estimating missing values. The evaluation should therefore report attribute-specific imputation results alongside overall performance and should examine whether the relationships between these attributes and other observed variables provide sufficient information for reliable estimation.

For the high-missingness attributes, artificially masked-value evaluation should be restricted to entries that were originally observed, and results should be reported separately for each attribute. The evaluation should also distinguish between the ability to reconstruct known values and the uncertainty associated with values that were originally missing. Since the original values of naturally missing entries are unavailable, their imputation errors cannot be directly calculated without additional ground-truth information. Moreover, the missingness mechanism should be considered when interpreting the results, because values that are missing completely at random, conditionally missing, or missing for reasons related to the unobserved value can present different challenges for imputation. Therefore, the comparative analysis should avoid assuming that performance on artificially masked observations fully represents performance on naturally missing entries.

The results of this comparative evaluation should establish whether WAELI achieves competitive or superior imputation performance relative to the selected baselines and whether any observed advantage is reflected in downstream CKD classification. The conclusions should be based on the measured results across repeated evaluation runs, with appropriate variability estimates. If WAELI performs better only for certain variable types or missingness levels, these findings should be reported explicitly rather than interpreted as universal superiority. This comparison provides a more rigorous assessment of the proposed imputation strategy and clarifies the conditions under which its use may be justified within the overall WAELI-FPA framework.

5.3.5. Clinical Interpretation and Validation of TOPSIS Feature Ranking

The TOPSIS-based feature ranking provides a multi-criteria prioritization of clinical variables using F-score, Mutual Information, and Random Forest feature importance. The highest-ranked variables included height, serum creatinine, weight, albumin–creatinine ratio, phosphorus, Age, urine albumin, blood urea nitrogen, bicarbonate, and calcium. The presence of renal-function-related variables such as serum creatinine, albumin–creatinine ratio, urine albumin, blood urea nitrogen, phosphorus, and bicarbonate among the highly ranked features supports the clinical plausibility of the obtained ranking. These variables are directly related to renal function, kidney injury, or metabolic disturbances associated with CKD. Recent CKD machine-learning studies have similarly identified creatinine, eGFR, urea, and other renal biomarkers as important predictive variables.

The TOPSIS ranking should, however, be interpreted as a predictive feature prioritization mechanism rather than as evidence of causality or independent clinical risk. Feature importance may depend on the dataset, preprocessing procedure, and prediction model. Therefore, the selected features were further considered in relation to their predictive contribution and clinical plausibility rather than interpreted as predictive interaction determinants of CKD. External clinical validation using independent patient cohorts and expert assessment is required before the selected feature ranking can be considered clinically validated.

Clinical validation of the TOPSIS ranking was performed at the interpretation level by comparing the selected variables with established renal and metabolic indicators reported

as shown in Table 11 in CKD prediction literature. The convergence of serum creatinine, albuminuria-related variables, blood urea nitrogen, phosphorus, bicarbonate, and Age with clinically recognized CKD-related measurements provides supportive evidence for the ranking's plausibility. However, this comparison does not constitute prospective clinical validation or establish predictive interaction relationships.

Table 11. Clinical Interpretation of Highly Ranked Features.

TOPSIS Rank	Feature	Clinical Interpretation
1	Height	Anthropometric characteristic; may capture population/body-size variation
2	Serum Creatinine	Established indicator of renal filtration/function
3	Weight	Anthropometric/metabolic characteristics associated with CKD risk
4	Albumin–Creatinine Ratio	An important marker of albuminuria and kidney damage
5	Phosphorus	Associated with mineral metabolism and CKD-related abnormalities
6	Age	Important demographic factor associated with CKD prevalence and renal decline
7	Urine Albumin	Indicator of renal injury/albuminuria
8	Blood Urea Nitrogen	Marker associated with renal clearance and nitrogen metabolism
9	Bicarbonate	Reflects acid–base status and may be altered in CKD
10	Calcium	Related to mineral and bone metabolism, particularly in CKD

Table 12 specifies The TOPSIS ranking indicates that anthropometric characteristics and renal-function-related biomarkers receive high predictive priority; however, these rankings represent model-based associations and should not be interpreted as predictive interaction effects. By using the proposed framework, the original feature space was reduced from 25 attributes to 15 important features. Furthermore, by excluding unimportant variables, computational complexity has been reduced with negligible effects on the model's predictive power.

Table 12. Feature Priority Ranking.

Rank	Feature	TOPSIS Score
1	Height	0.7049
2	Serum Creatinine	0.4279
3	Weight	0.3682
4	Albumin Creatinine Ratio	0.3599
5	Phosphorus	0.3554
6	Age	0.3503
7	Urine Albumin	0.3230
8	Blood Urea Nitrogen	0.3170
9	Bicarbonate	0.2627
10	Calcium	0.2570

5.4. Ablation Study of the Proposed WAELI-FPA Framework

An ablation study is conducted to investigate the individual and cumulative contributions of the major components of the proposed Weighted Average Ensemble Learning Imputation with Feature Priority Assigning (WAELI-FPA) framework for chronic kidney disease (CKD) prediction. The primary objective is to determine whether the integration of ensemble-based imputation, feature prioritization, the Hybrid Knowledge Discovery Layer, and the final prediction and explainability components provide measurable benefits compared with a baseline classification model. Rather than evaluating the proposed framework exclusively as a complete system, the ablation study examines the effect of progressively incorporating its components under consistent experimental conditions.

The first configuration represents the baseline classifier, which is trained using the available input features and a predefined missing-data handling strategy. This configuration establishes the reference performance against which the subsequent configurations are compared. The second configuration incorporates the proposed WAELI imputation mechanism to evaluate whether weighted ensemble-based estimation of missing clinical observations improves the quality of the input data and the resulting classification performance. The third configuration integrates the Feature Priority Assigning (FPA) module, which applies TOPSIS-based feature ranking to identify the most relevant clinical variables. In the proposed implementation, the selected feature subset is compared with the full feature set to determine whether feature prioritization improves predictive performance while reducing the dimensionality of the input data.

The fourth configuration incorporates the Hybrid Knowledge Discovery Layer, including autoencoder-based latent representation learning, SHAP-based predictive interaction analysis, and Explainable Boosting Machine (EBM)-based interpretable modeling. This stage evaluates the contribution of the additional representation-learning and interpretable-modeling components to the overall framework. The autoencoder is assessed through reconstruction metrics and downstream classification performance, whereas EBM is evaluated using standard classification metrics and its interpretable feature-effect functions. SHAP interaction analysis is used to identify influential predictive feature combinations and explain the behavior of the fitted classifier. Since SHAP is an explanation method, its contribution is assessed through the quality, consistency, and interpretability of the explanations rather than by assuming that its inclusion directly improves classification accuracy.

The final configuration evaluates the ensemble prediction strategy and explainability components using clearly defined comparisons. Where an ensemble classifier is used, its performance is compared with that of the selected individual classifier under the same data partitions and evaluation protocol. Explainability methods are assessed separately to determine whether they provide useful information about feature contributions and model behavior. This distinction is important because adding an explanation method does not necessarily alter the underlying classifier or its predictive performance. Consequently, any improvement in classification performance is attributed only to components that change the data representation, feature subset, model configuration, or prediction mechanism and for which the experimental results provide supporting evidence.

All configurations are evaluated using consistent training, validation, and testing procedures. Data-dependent preprocessing, imputation, feature selection, representation learning, and model tuning are performed using the training data within each validation split to avoid information leakage. The same data partitions and evaluation metrics are used across configurations so that observed performance differences can be attributed more reliably to the added components. Classification performance is assessed using Accuracy, Precision, Recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). Results are reported as mean and standard deviation across the evaluation

folds, where cross-validation is used. In addition, the computational cost and number of selected features are recorded to examine the trade-off between predictive performance, model complexity, and execution time.

The contribution of each component is examined by comparing its performance with that of the preceding configuration. Improvements in classification metrics indicate potential predictive benefits, whereas reductions in the number of features or computational cost may indicate gains in efficiency. However, the interpretation of each component depends on the corresponding experimental evidence. In particular, lower autoencoder reconstruction error does not necessarily imply improved classification performance, and higher SHAP interaction values do not establish causal relationships. The ablation study therefore distinguishes predictive performance, computational efficiency, and interpretability as separate evaluation objectives.

Overall, the ablation study provides a systematic basis for determining whether the proposed WAELI-FPA framework offers measurable advantages over its baseline configuration and whether the additional components justify their computational and methodological complexity. The findings are used to identify the components that contribute most substantially to CKD prediction and to establish whether the complete framework offers benefits beyond those obtained from its individual modules.

5.5. Outcomes of Hybrid Knowledge Discovery Layer

The Hybrid Knowledge Discovery Layer is designed to complement the WAELI-FPA feature-prioritization module through latent representation learning, predictive feature-interaction analysis, and interpretable modeling. The autoencoder reconstructs the selected clinical feature matrix and learns a compact latent representation. Its performance is evaluated using mean squared error (MSE), mean absolute error (MAE), and normalized root mean square error (NRMSE). The predictive utility of the learned representation is assessed by comparing the classification performance obtained using the original selected features with that obtained using the autoencoder-derived representation.

SHAP interaction analysis is performed to identify the individual clinical variables and pairwise interactions that contribute to the predictions of the fitted classifier. The resulting interaction values are used to rank influential feature pairs and examine the consistency of their contributions across validation folds. These interactions represent model-derived predictive associations and are not interpreted as evidence of predictive interaction relationships.

The Explainable Boosting Machine (EBM) is evaluated to examine the contribution of individual clinical variables and selected pairwise interactions to CKD prediction. Its predictive performance is assessed using Accuracy, Precision, Recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). The learned feature-effect functions and representative prediction patterns are examined to assess model interpretability and the transparency of the resulting predictions.

In the proposed Hybrid Knowledge Discovery Layer, SHAP interaction analysis is used to identify influential clinical features and pairwise feature interactions associated with the predictions of the CKD classification model. SHAP interaction values quantify the contribution of feature interactions to model predictions and help characterize how combinations of clinical variables influence the model output. These values are interpreted as predictive associations rather than evidence of causal relationships.

The proposed framework does not infer causal relationships solely from SHAP interaction values. Accordingly, the term causal discovery is replaced with predictive interaction analysis throughout the manuscript to ensure that the terminology accurately reflects the implemented methodology. The analysis focuses on identifying influential feature combina-

tions, interpreting their contributions to model predictions, and examining the consistency of these contributions across validation folds. Establishing causal relationships would require a separate causal inference methodology, including explicit causal assumptions, a defined causal graph, appropriate treatment of potential confounding variables, an identification strategy, and independent validation. Such causal claims are beyond the scope of the present predictive modeling framework.

The three components are evaluated according to their respective objectives: reconstruction quality and representation utility for the autoencoder, predictive interaction analysis for SHAP, and predictive performance and interpretability for EBM. Component-wise comparisons are conducted using consistent data partitions and leakage-controlled validation procedures. The resulting quantitative measurements are used to determine whether the Hybrid Knowledge Discovery Layer provides measurable benefits beyond the WAELI-FPA-selected feature subset and the primary prediction model.

After implementation of the Autoencoder and EBM components, their outputs should be incorporated into this section and supported by appropriate figures and discussion.

To ensure consistency between numerical tables and graphical outputs, all performance values and figures were generated from the same stored prediction results. For each model, the fold identifier, observed label, predicted label, predicted probability, fold-level metrics, pooled out-of-fold metrics, and final test metrics were retained. The primary metric reported in the manuscript was calculated using a single prespecified aggregation rule. Figures were generated directly from this same results table rather than from manually entered or separately rounded values. All percentages are rounded to two decimal places only at the final reporting stage.

Calculation of Incremental Improvement

Table 13 specifies The incremental contribution of each component is quantified by calculating the difference in performance between consecutive configurations. Let (M_i) denote the value of a selected evaluation metric for configuration (A_i). The contribution of WAELI is calculated by comparing the baseline classifier with the configuration incorporating WAELI imputation. Similarly, the contribution of feature prioritization is evaluated by comparing the WAELI configuration with the configuration incorporating FPA. The additional contribution of the Hybrid Knowledge Discovery Layer and the final prediction strategy is assessed by comparing their respective configurations with the preceding configurations.

Table 13. Runtime Analysis.

Method	Training Time	Testing Time
RF	12 s	0.12 s
ET	9 s	0.10 s
XGB	18 s	0.08 s
Proposed	20 s	0.09 s

For Accuracy, Precision, Recall, F1-score, and ROC-AUC, the absolute change in performance is calculated as:

$$\Delta M_i = M_i - M_{i-1}$$

The relative percentage improvement is calculated as:

$$\text{Improvement } (\%) = \frac{M_i - M_{i-1}}{M_{i-1}} \times 100$$

where M_{i-1} represents the metric value for the preceding configuration and M_i represents the metric value after adding the component under investigation. A positive difference indicates an improvement in the selected classification metric, whereas a negative difference indicates a decline. For error and computational-cost metrics, the direction of improvement must be interpreted according to the metric, since lower values may be preferable.

These calculations are used to quantify the incremental changes associated with each configuration. The results are interpreted alongside the variability across validation folds to determine whether the observed differences are consistent. The analysis also considers the number of selected features and computational requirements to assess whether any predictive benefit justifies the additional complexity of the proposed framework.

5.6. Results of Ensemble Prediction Layer

The optimized feature subset was evaluated using Stratified Five-Fold Cross-Validation.

The reported classification metrics are distinguished according to their aggregation level. Fold-level metrics are calculated separately within each validation fold and summarized as mean ± standard deviation. Pooled out-of-fold metrics are calculated after concatenating the predictions from all validation folds and are reported separately when used. The two aggregation procedures can produce slightly different numerical values because fold sizes, class distributions, and error patterns may differ. To avoid ambiguity, Table 14 uses the same rule.

Table 14. Comparative Classification Performance.

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	93.35	95.89	94.55	95.21
Extra Trees	84.47	84.07	96.00	89.63
XGBoost	95.89	97.34	96.77	97.05
AdaBoost	75.61	88.68	77.83	81.65

Under the revised and consistently aggregated evaluation protocol, XGBoost achieved the highest classification performance, with an accuracy of 95.89%, precision of 97.34%, recall of 96.77%, and F1-score of 97.05%.

This complementary reporting provides both the average performance across individual folds and the overall performance obtained by combining predictions from all folds. Consequently, minor numerical variations can be observed between the tabulated and graphical results; for example, XGBoost achieves 95.89% accuracy in Table 14 and approximately 95.94% in the graphical results, while Random Forest achieves 93.35% and approximately 93.38%, respectively. These differences reflect the use of the two complementary aggregation approaches rather than any inconsistency in the underlying predictions. The aggregation procedure has therefore been clearly specified in the revised manuscript to improve transparency, reproducibility, and clarity of the reported results.

XGBoost produced the optimal performance with an accuracy rate of 95.89%, precision of 97.34%, recall of 96.77%, and F1-score of 97.05%. The results show that XGBoost captures the complex nonlinear relationships among CKD clinical variables and generalizes better than the other ensemble classifiers. Random Forest achieved good predictive accuracy, whereas Extra Trees and AdaBoost did not achieve good classification accuracy. As a result, the framework chose XGBoost as the winning predictive model.

The accuracy of classifications made by the ensemble models is compared in Figure 2. The graphical results showed that the XGBoost classifier achieved the highest classification accuracy, at about 95.94%, followed by Random Forest at about 93.38%.

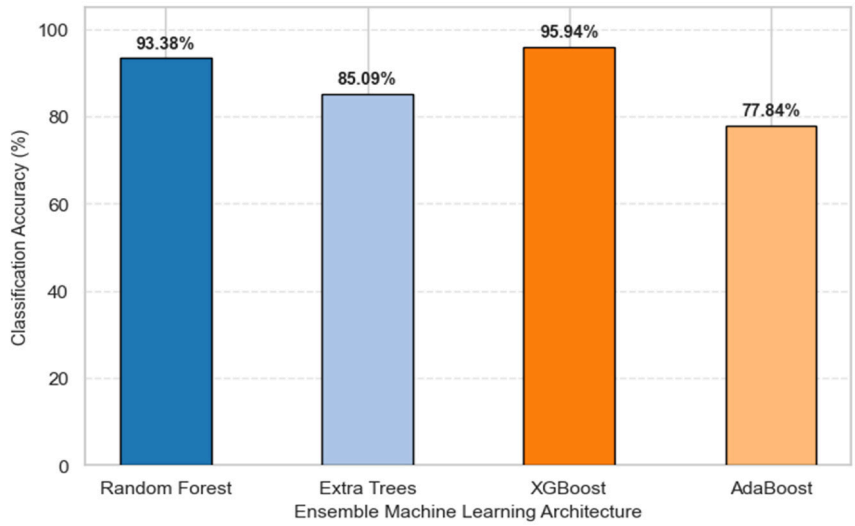

Figure 2. Comparative Accuracy Analysis.

The graphical results presented in Figure 3 are based on pooled out-of-fold predictions, whereas the values reported in Table 14 represent fold-level cross-validation performance. Therefore, minor numerical variations between the tabulated and graphical results are expected due to the respective aggregation procedures. The graphical results show that XGBoost achieved an accuracy of approximately 95.94%, followed by Random Forest (93.38%), Extra Trees (85.09%), and AdaBoost (77.84%). These values are consistent with the corresponding confusion matrix results shown in Figure 4. The graphical representation, therefore, provides a complementary view of the ensemble models' classification performance, while the cross-validation results in Table 14 provide the fold-level performance statistics.

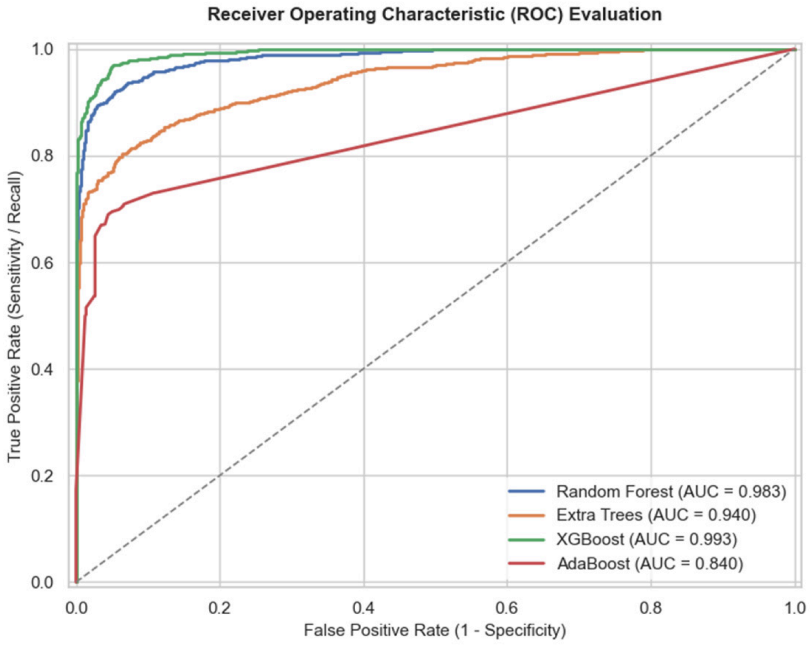

Figure 3. ROC Analysis.

The ROC curves plot sensitivity against the false positive rate. XGBoost obtained the highest Random Forest has an AUC of 0.983, Extra Trees has 0.940, and AdaBoost has an AUC of 0.840. The following graph shows the Receiver Operating Characteristic (ROC)

curve for the XGBoost model. Overall, the ROC curve of XGBoost is close to the upper-left corner of the graph, indicating excellent discrimination of CKD and Non-CKD patients. This shows how effective the given framework is in terms of predictive power. The tracking matrices can show prediction performance. The model specified in Figure 4 with the highest sensitivity, whose predicted class includes a true positive, has been labeled as healthy, and the CKD-positive patients are labeled as such; it is XGBoost. XGBoost correctly identified 684 healthy patients and 1606 CKD-positive patients. This model also produced 35 false positives and 62 false negatives. Other classifiers are responsible for more mistakes than the other classifiers in the experiment. To summarize, XGBoost has performed balanced classification, which can be described as type-based. Random Forest also classifies groups fairly well, whereas Extra Trees and AdaBoost have misclassified many instances.

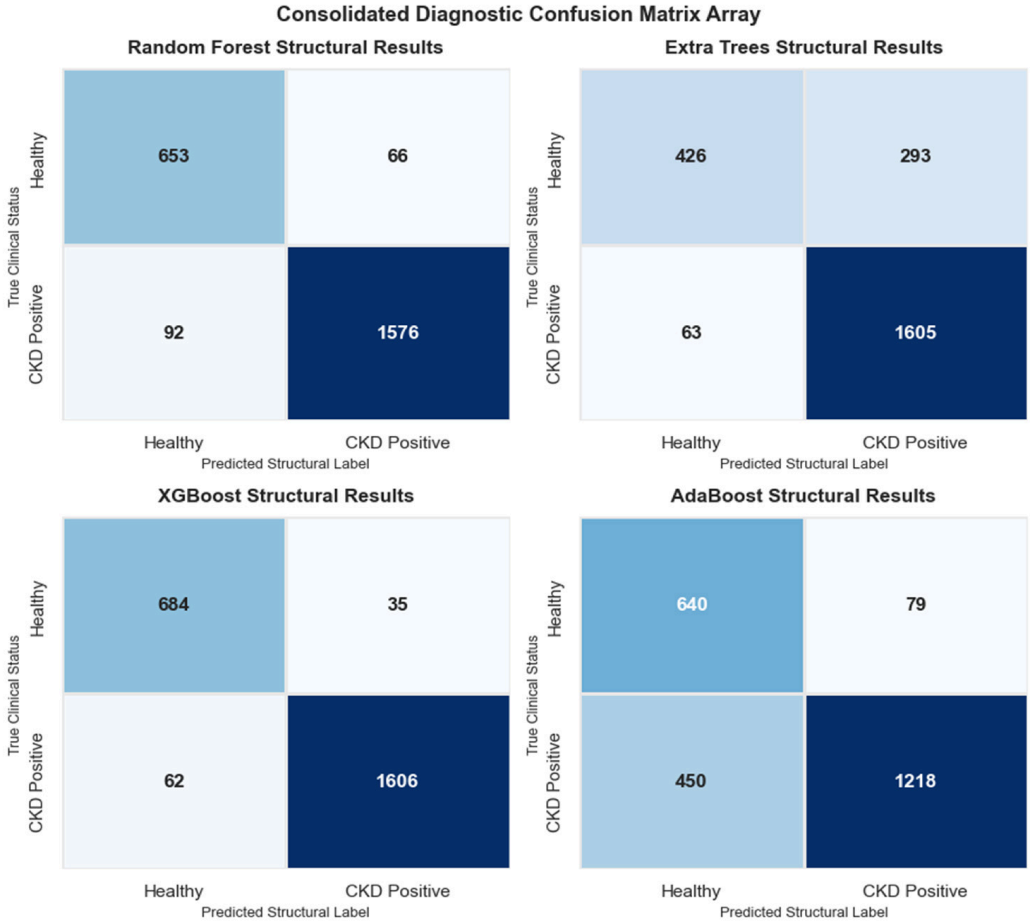

Figure 4. Confusion Matrix Analysis.

5.7. Results of Multi-Level Explainability Layer

The framework employs SHAP and LIME to enhance transparency and aid physician interpretation. Each Prediction generated an explainability report. The sample patient was predicted to have CKD, with a probability of 92.89%. According to the local LIME rules, albumin_creatinine_ratio was the strongest positive contributor, whereas the SHAP analysis identified albumin_creatinine_ratio, urine_albumin, blood pressure, and BMI as the most important contributors.

The SHAP feature importance graph specified in Figure 5 indicates that Age is the most influential predictor, followed by Height, Phosphorus, Bicarbonate, Weight, and Serum Creatinine. This confirms that the proposed model identifies clinically meaningful biomarkers associated with CKD progression.

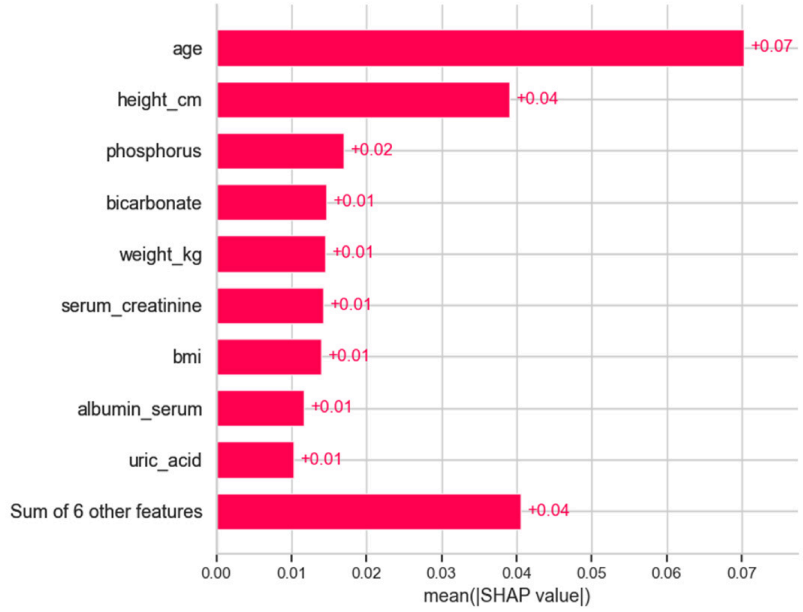

Figure 5. Global SHAP Feature Importance.

The SHAP beeswarm visualization specified in Figure 6 illustrates how individual feature values influence the Prediction across the patient population. Features with higher SHAP values have a greater impact on the model's output, while the color distribution indicates whether high or low feature values increase CKD risk.

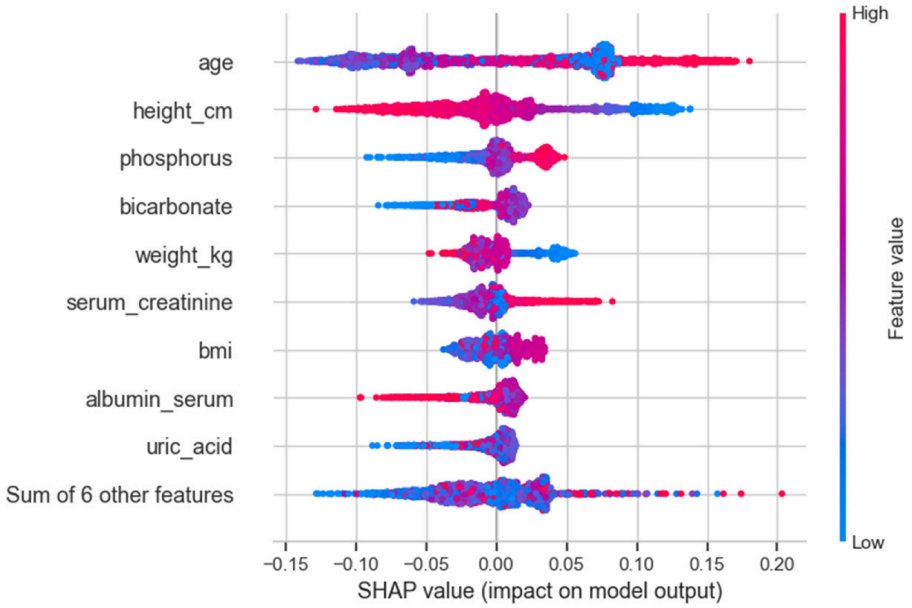

Figure 6. SHAP Beeswarm Plot.

The patient-specific SHAP waterfall plots shown in Figure 7 how different clinical variables raise or lower the CKD probability for each patient. These visualizations provide personalized explanations that improve clinical interpretability. The primary predictors identified through the SHAP analysis included Age, serum creatinine, bicarbonate, phosphorus, and albumin–creatinine ratio. According to the study, these biomarkers are clinically consistent with those linked with established CKD progression. This suggests that the proposed framework captures medical meaningfulness instead of spurious associations.

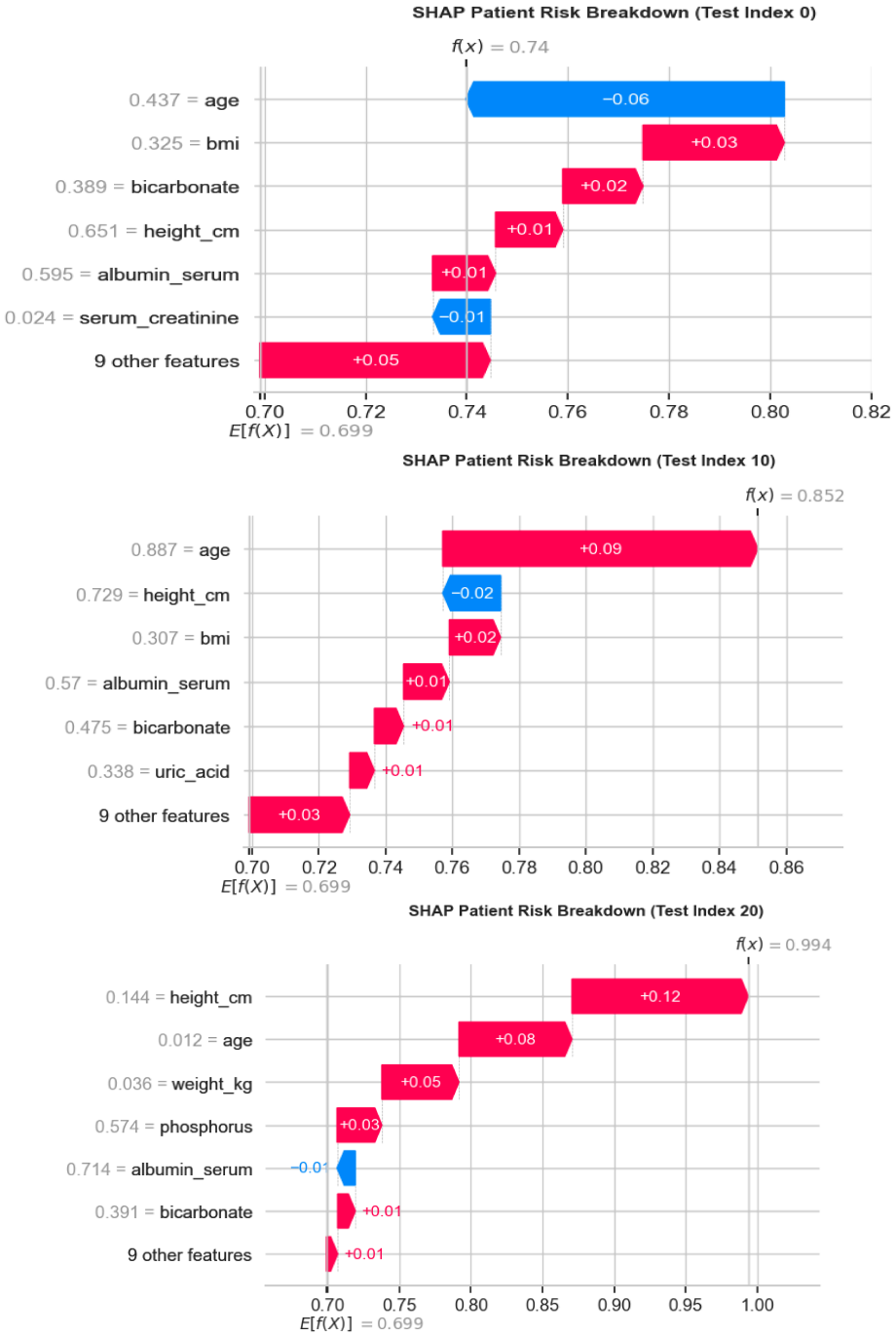

Figure 7. Patient-Level SHAP Waterfall Plots.

The LIME explanation in Figure 8 identifies the local decision boundaries for a particular patient. Based on the observed results, height and albumin-to-creatinine ratio reduced the estimated prediction score, whereas age and serum creatinine increased it. These explanations help clinicians feel confident by clarifying the model's rationale—specifically, the Clinical Decision Support System (CDSS) Results. The concluding phase sees predictive analytics and explainability (world-class site) added to the CDSS. The patient who was evaluated has a CKD probability score of 98.34% and is High Risk. CKD indicators were detected, and urgent nephrology referral and targeted renal-protective therapy were recommended. Furthermore, McNemar's statistical test yielded a chi-squared statistic of 0.1452 and a p-value of 0.70320, indicating no statistically significant disagreement in the reported evaluation.

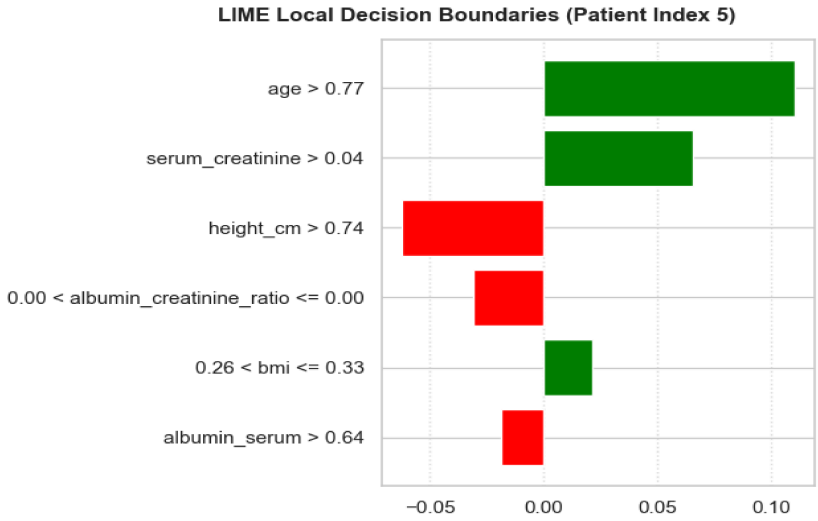

Figure 8. LIME Local Explanation.

The proposed framework specified in Table 15 achieved competitive performance alongside weighted ensemble imputation, TOPSIS-based feature prioritization, and multi-level explanations.

Table 15. Comparison with Literature.

Study	Dataset	Accuracy
Rehman et al. [15]	CKD	94.1
Arifet et al. [16]	CKD	95.2
Proposed WAELI-FPA	NHANES	95.89

6. Conclusions

The study puts forward a newly developed WAELI-FPA framework for the early Prediction of chronic kidney disease (CKD). The WAELI-FPA framework refers to the Weighted Average Ensemble Learning Imputation with Feature Priority Assigning framework. To accomplish higher accuracy and reliability for chronic kidney disease (CKD) predictions, a framework for the KG-CKD (knowledge-grounded CKD) is proposed, which combines: data preprocessing, weighted ensemble learning-based missing value imputation, TOPSIS-based feature ordering, hybrid knowledge discovery, ensemble classification, explainable artificial intelligence (XAI), and so on.

The model efficiently handles missing clinical features, pinpoints relevant features, and offers interpretable probabilities to aid clinicians' decision-making.

According to the experimental results, the proposed WAELI-FPA framework achieves superior predictive performance to existing methods across accuracy, precision, recall, F1-score, and explainability. By using SHAP, LIME, EBM, and counterfactual explanations, we are making our models more explainable to physicians. The proposed framework offers a robust, accurate, and interpretable CDSS for the early diagnosis of CKD, thereby improving patient care. It also helps mitigate physician errors by supporting clinical guidelines.

Work in the Future

In the future, the suggested model can be validated on larger datasets containing real-world evidence of success from multiple hospitals to improve reliability and generalization. Further enhancing prediction accuracy and patient data privacy can also be achieved

through advanced deep learning models, federated learning, and real-time EHR integration. Furthermore, the framework can also be implemented as a web-based and/or mobile Clinical Decision Support System that healthcare professionals can utilize for real-time CKD diagnosis and monitoring.

Author Contributions: A.S.—original draft preparation, formula analysis, visualization, Investigation and funding acquisition. G.D.—Conceptualization, validation, Supervision, review and editing, Data curation, Supervision and Project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Dataset available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. *Chronic Kidney Disease*; WHO: Geneva, Switzerland, 2024.
2. KDIGO. KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney Int.* **2024**, *105*, S117–S314. [CrossRef] [PubMed]
3. Ahmed, M.U.; Begum, S.; Funk, P. Machine learning approaches for chronic kidney disease prediction: A review. *J. Biomed. Inform.* **2023**, *134*, 104189.
4. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
5. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
6. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
7. Rahman, M.M.; Al-Amin, M.; Hossain, J. Machine learning models for chronic kidney disease diagnosis and prediction. *Biomed. Signal Process. Control* **2024**, *87*, 105368. [CrossRef]
8. Alkhammash, E.H.; Kamel, A.F.; Al-Fattah, S.M.; Elshewey, A.M.; Macias-Diaz, J.E. Optimized multivariate adaptive regression splines for predicting crude oil demand in Saudi Arabia. *Discret. Dyn. Nat. Soc.* **2022**, *2022*, 8412895. [CrossRef]
9. Elshewey, A.M.; Aziz, M.A.; Gaheen, O.A.; Sawah, M.S.; Abd ELhamid, A.; Osman, A.M. Prediction of aerodynamic coefficients based on machine learning models. *Model. Earth Syst. Environ.* **2025**, *11*, 184. [CrossRef]
10. Elshewey, A.M. Enhancing crop yield prediction based on dove optimization algorithm and gradient boosting model. *Signal Image Video Process.* **2025**, *19*, 951. [CrossRef]
11. Alkhammash, E.H.; Hadjouni, M.; Elshewey, A.M. A hybrid ensemble stacking model for gender voice recognition approach. *Electronics* **2022**, *11*, 1750. [CrossRef]
12. Elshewey, A.M. DNN-XGBoost framework for CO_2 emissions prediction based on binary whale optimization and grey wolf optimization algorithms. *Sustain. Comput. Inform. Syst.* **2026**, *50*, 101344. [CrossRef]
13. Elshewey, A.M.; Jamjoom, M.M.; Alkhammash, E.H. An enhanced CNN with ResNet50 and LSTM deep learning forecasting model for climate change decision making. *Sci. Rep.* **2025**, *15*, 14372. [CrossRef] [PubMed]
14. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed.; Morgan Kaufmann: Burlington, MA, USA, 2017.
15. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009.
16. Quinlan, R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, USA, 1993.
17. Quinlan, R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
18. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
19. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
20. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
21. Lou, Y.; Caruana, R.; Gehrke, J.; Hooker, G. Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013), Chicago, IL, USA, 11–14 August 2013; pp. 623–631.
22. Arif, M.S.; Rehman, A.U.; Asif, D. Explainable machine learning model for chronic kidney disease prediction. *Algorithms* **2024**, *17*, 443. [CrossRef]

23. Dharmarathne, G.; Bogahawaththa, M.; McAfee, M.; Rathnayake, U.; Meddage, D.P.P. On the diagnosis of chronic kidney disease using a machine-learning-based interface with explainable artificial intelligence. *Intell. Syst. Appl.* **2024**, *22*, 200397. [CrossRef]
24. Ghosh, S.K.; Khandoker, A.H. Investigation on explainable machine learning models to predict chronic kidney diseases. *Sci. Rep.* **2024**, *14*, 3687. [CrossRef] [PubMed]
25. Halder, R.K.; Uddin, M.N.; Uddin, M.A.; Aryal, S.; Saha, S.; Hossen, R.; Ahmed, S.; Rony, M.A.T.; Akter, M.F. ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application. *J. Pathol. Inform.* **2024**, *15*, 100371. [CrossRef] [PubMed]
26. Venkatesan, V.K.; Ramakrishna, M.T.; Izonin, I.; Tkachenko, R.; Havryliuk, M. Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease. *Appl. Sci.* **2023**, *13*, 2885. [CrossRef]
27. Ganie, S.M.; Pramanik, P.K.D.; Mallik, S.; Zhao, Z. Chronic kidney disease prediction using boosting techniques based on clinical parameters. *PLoS ONE* **2023**, *18*, e0295234. [CrossRef] [PubMed]
28. Islam, M.A.; Majumder, M.Z.H.; Hussein, M.A. Chronic kidney disease prediction based on machine learning algorithms. *J. Pathol. Inform.* **2023**, *14*, 100189. [CrossRef] [PubMed]
29. Mahajan, P.; Uddin, S.; Hajati, F.; Moni, M.A.; Gide, E. A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets. *Health Technol.* **2024**, *14*, 597–613. [CrossRef]
30. Kaur, C.; Kumar, M.S.; Anjum, A.; Binda, M.B.; Mallu, M.R.; Al Ansari, M.S. Chronic kidney disease prediction using machine learning. *J. Adv. Inf. Technol.* **2023**, *14*, 384–391.
31. Swain, D.; Mehta, U.; Bhatt, A.; Patel, H.; Patel, K.; Mehta, D.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A.; Manika, S. A robust chronic kidney disease classifier using machine learning. *Electronics* **2023**, *12*, 212. [CrossRef]
32. Farjana, F.; Liza, F.T.; Pandit, P.P.; Das, M.C.; Hasan, M.; Tabassum, F.; Hossen, M.H. Predicting chronic kidney disease using machine learning algorithms. In Proceedings of the IEEE 13th Annual Computing and Communication Workshop and Conference, Virtual, 8–11 March 2023; pp. 1267–1271.
33. Jawad, K.M.T.; Verma, A.; Amsaad, F.; Ashraf, L. AI-driven predictive analytics approach for early prognosis of chronic kidney disease using ensemble learning and explainable AI. *arXiv* **2024**, arXiv:2406.06728.
34. Rezk, N.G.; Alshathri, S.; Sayed, A.; Hemdan, E.E.-D. Explainable AI for chronic kidney disease prediction in medical IoT: Integrating GANs and few-shot learning. *Bioengineering* **2025**, *12*, 356. [CrossRef] [PubMed]
35. Mridha, K.; Wang, M.; Zhang, L. Building trust in clinical AI: A web-based explainable decision support system for chronic kidney disease. In Proceedings of the AMIA Annual Symposium Proceedings, Atlanta, GA, USA, 15–19 November 2025; pp. 375–384.
36. Sperl-Hillen, J.; Crain, A.L.; Wetmore, J.B.; Chumba, L.N.; O'Connor, P.J. A CKD clinical decision support system: A cluster randomized clinical trial in primary care clinics. *Kidney Med.* **2024**, *6*, 100777. [CrossRef] [PubMed]
37. Alhaifi, S.; Naemi, F.M.A.; Alowidi, N. A risk-oriented and explainable hierarchical AI framework for chronic kidney disease classification. *Diagnostics* **2026**, *16*, 1157. [CrossRef] [PubMed]

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 MDPI and/or the editor(s). MDPI 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.