Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Early-Stage Detection of Chronic Kidney Disease Using Ensemble Learning with Feature Selection and Data Imputation Techniques

Janhavi S. Singh
Department of Computer Science
Dr. D.Y. Patil Arts Commerce Science College
Pimpri, India

Juhi R. Dhote
Department of Computer Science
Dr. D.Y. Patil Arts Commerce Science College
Pimpri, India

***Abstract*— Chronic Kidney Disease (CKD) is an illness that damages your kidneys over a long time, and its early stages can go unnoticed as there are no obvious symptoms. If diagnosis is delayed, the patient will get severe problems which can even lead to the need for dialysis (renal replacement therapy). Therefore, early and accurate identification is critical for effective treatment and improved patient outcomes. Machine learning (ML) models have started to be used for this purpose, but here too we see a problem. Most models are trained on small highly structured datasets and are only evaluated in the terms of binary classification, i. e. CKD vs non-CKD. It makes actual clinical use very difficult as data is incomplete and noisy. This study proposes an ensemble-based machine learning framework that integrates advanced data imputation, hybrid feature selection, and explainable artificial intelligence (XAI) techniques to improve both predictive performance and model interpretability. Publicly available electronic health record (EHR) datasets are utilized, including the UCI CKD dataset for binary classification and a Kaggle CKD stages dataset for multi-class stage prediction. Missing values are addressed using Iterative Random Forest and Multiple Imputation by Chained Equations (MICE). Feature relevance is enhanced through hybrid feature selection methods, and SHAP-based explanations are applied to interpret model predictions. The proposed framework tries to deliver better accuracy, robustness when faced with missing data, and greater transparency over the conventional ML methods. The results from the experiments will highlight improved capability in the detection of early-stage CKD along with a stage-classification. This research contributes toward practical ML solutions that can support clinical decision-making and improve early diagnosis of CKD.**

Keywords—Chronic Kidney Disease (CKD); Electronic Health Records (EHR); Ensemble Learning; Data Imputation; Multiple Imputation by Chained Equations (MICE); Random Forest; XGBoost; Explainable Artificial Intelligence (XAI); SHAP (SHapley Additive Explanations); Cross-Dataset Generalization; Machine Learning in Healthcare; CKD Stage Classification; Early Disease Detection; Model Robustness

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a condition that lasts for a very long time and gradually the kidneys lose their capability to carry out their functions. It can lead to end, stage renal failure if it keeps on progressing and is not diagnosed in time. CKD affects 10.15% of adults worldwide according to the World Health Organization (WHO), and this figure is still increasing mainly because of diabetes, high blood pressure, and other lifestyle diseases that are the root causes of CKD. The damage extent of the disease is gauged based on the Glomerular Filtration Rate (GFR) thus five stages have been assigned to it, the damage being the mildest in Stage 1 and the most severe (kidney failure) in Stage 5.

Early stages of CKD usually don't give any symptoms, and this is one of the reasons why the disease is so dangerous; people rarely check themselves and they are not aware of it until the situation is already critical. There are some laboratory markers that help doctors to detect the disease, however, since the values of those markers differ from one person to another, it is sometimes quite difficult for medical professionals to diagnose the disease just by looking at the lab results. Therefore, many patients are only given a diagnosis when the disease is already progressing, thus the necessity of early detection should not be underestimated for the prevention of other complications.

In today's world, machine learning (ML) has been very helpful for the medical industry by providing efficient ways for doctors to deal with complicated data and correctly predict diseases. Ensemble models like Random Forest, XGBoost, and AdaBoost when combined with good data preprocessing and feature selection have achieved very impressive results. However, most of the prior CKD prediction research has been confined to only accurate results in small and clean datasets, thus arising a problem with their performance in clinical data due to the presence of incomplete records, inconsistently entered data, and noise in measurements which have not been considered before, hence limiting their application in healthcare institutions.

This research solves that problem through the development of a dependable and explainable ensemble learning framework for predicting CKD. It combines various data imputation techniques, determines strength under missing and noisy data, and measures generalization over two different datasets i.e UCI and Kaggle. Besides the motivation to enhance the prediction output, the team also intends their work to be a communication tool that clinicians can use to have a better understanding of their trust in clinical algorithms.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

II. OVERVIEW OF CHRONIC KIDNEY DIESEASE

1. *Stages of CKD*

CKD is classified into five stages based on GFR levels, which indicate the kidneys' filtering capacity. Each stage corresponds to a specific level of kidney function and associated clinical implications:

TABLE I.

Stage	Stages of CKD		
	GFR (mL/min/1.73m²)	***Description***	***Clinical Features***
Stage 1	≥90	Normal or high GFR with mild kidney damage	Often asymptomatic; possible proteinuria or hematuria
Stage 2	60–89	Mild decrease in GFR	Subtle biochemical changes, no overt symptoms
Stage 3	30–59	Moderate decrease in GFR	Fatigue, anemia, early metabolic imbalance
Stage 4	15–29	Severe reduction in GFR	Elevated serum creatinine, hypertension, electrolyte disturbances
Stage 5	<15	Kidney failure or end-stage renal disease (ESRD)	Requires dialysis or kidney transplantation

The classification given in above TABLE I. is a tool for clinicians to understand the severity of the illness, treat it accordingly, and keep track of its progress. Nonetheless, a large number of patients are not diagnosed until Stages 1 and 2, when intervention could have most effectively slowed or even prevented renal decline.

2. *Key Clinical Indicators*

There are a number of biochemical and physiological markers, several of which are used as features in clinical datasets, which CKD diagnosis and monitoring depend on. The indicators most frequently measured are:

- Serum Creatinine (SC): A byproduct of muscle metabolism; elevated levels indicate impaired filtration.
- Glomerular Filtration Rate (GFR): The primary measure of kidney function, estimated using serum creatinine, age, and gender.
- Blood Urea Nitrogen (BUN): Reflects nitrogen waste accumulation; increases with declining kidney function.
- Hemoglobin (Hb) and Packed Cell Volume (PCV): Indicators of anemia, which often develops in CKD due to reduced erythropoietin production.
- Albumin (AL) and Specific Gravity (SG): Represent urine composition; abnormalities suggest protein leakage or reduced concentrating ability.
- Sodium (Na), Potassium (K), and Calcium (Ca): Electrolyte imbalances are common in later stages.
- Blood Pressure (BP): Both a cause and consequence of CKD, frequently elevated as kidney function declines.

There are a number of biochemical and physiological markers, several of which are used as features in clinical datasets, which CKD diagnosis and monitoring depend on. The features such as hemoglobin, PCV, red blood cell count, serum creatinine, albumin, and specific gravity were among the top predictors that came out of model training and feature selection in the datasets, thus, they perfectly correspond to these established clinical markers of kidney dysfunction.

3. *Challenges in Early Detection*

One of the biggest problems with chronic kidney disease (CKD) is invisibility: the disease is basically invisible because it doesnt give out any clear early, stage symptoms. Patients in the first two stages mostly look and feel absolutely normal, and if there are any changes in the test results, they are so tiny that people hardly notice them. Also, the fact that many symptoms of CKD are very similar to other health issues like anemia, dehydration, or diabetes makes the problem of diagnosis all the more complex.

Moreover, the clinical data is sometimes a mess itself. Patient records are often incomplete, and similar kinds of data are written down in very different ways at different hospitals. Because of this lack of uniformity, traditional statistical methods often fail to uncover the true patterns behind the data. Data, driven approaches come to the rescue here as they are capable of unearthing hidden and nonlinear relationships among clinical features that may be missed by classic methods.

The use of machine learning, especially ensemble models, is becoming a more and more reliable answer to this problem. These models have the ability to look at multiple different aspects of data at the same time, whilst their outputs remain understandable by healthcare professionals. In combination with powerful data imputation methods, careful feature selection, and interpretability tools such as SHAP, they can significantly improve the accuracy and confidence of CKD detection in real clinical settings.

III. PROBLEM STATEMENT

Chronic Kidney Disease (CKD) prediction using machine learning has been intensively researched, yet several problems persist when these models are utilized in real healthcare settings. Most of the existing work is based on small, clean datasets, which do not truly reflect the diversity, noise, and missing data in actual hospital records. Hence, model accuracy tends to drop when tested on a bigger or more heterogeneous group of patients.

Besides that, one of the drawbacks of the state-of-the-art methods is their concentration on binary classification only, where models decide whether CKD is present or not. This method ignores the identification of the early stages, especially Stage 1 and Stage 2 CKD, where treatment can be efficiently targeted to considerably delay disease progression and avoid kidney failure.

Major challenges are in the proper handling of missing and contradictory clinical data. Real-life healthcare datasets often have missing entries; however, a lot of research is done by

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

means of simple imputation methods that may comprise bias. The use of more resilient imputation techniques, like Multiple Imputation by Chained Equations (MICE) or Iterative Random Forest imputation, might enhance the consistency of predictions.

Moreover, ensemble learning models such as Random Forest and XGBoost may be able to attain remarkably high predictive accuracy, although because of their limited interpretability, their implementation in clinical decision support systems is practically limited. Explainability methods like SHAP can delineate crucial medical markers, for example, glomerular filtration rate (GFR) and serum creatinine, thus leading to higher transparency and trust in the outputs of the model.

Thus, the development of a CKD diagnosis tool that is both highly accurate, interpretable, and generalizable, as well as being able to thoroughly cope with missing data and be reliable over various clinical datasets, is in demand.

IV. Objectives of the Research

The main aim of this study is to develop a strong and interpretable ensemble learning framework for the early detection and stage-wise prediction of Chronic Kidney Disease (CKD). The focus is on dealing with real-world challenges in clinical data, such as missing values, inconsistent features, and the lack of model interpretability. By combining different machine learning techniques and evaluation strategies, the research aims to make CKD prediction more reliable and clinically meaningful.

The specific objectives of the study are as follows:

- Model Development and Evaluation: To build and test ensemble-based models, including Random Forest, XGBoost, AdaBoost, and Extra Trees for both binary (CKD vs. non-CKD) and stage-wise (Stages 1–5) prediction.
- Handling Missing Data: To apply and compare several data imputation methods such as baseline imputation, Iterative Random Forest, and Multiple Imputation by Chained Equations (MICE), and to observe how these methods influence model performance on incomplete datasets.
- Feature Selection: To identify key predictors of CKD using a hybrid feature selection approach that combines Mutual Information, Random Forest importance, and XGBoost importance scores.
- Cross-Dataset Testing: To evaluate model generalization by testing how well a model trained on one dataset (UCI CKD) performs on another (Kaggle CKD), providing insight into its real-world applicability.
- Model Interpretability: To enhance the interpretability of the models using SHAP (SHapley Additive Explanations), highlighting how major clinical factors such as serum creatinine, hemoglobin, and GFR contribute to each predict.

V. Literature Review

A. Machine Learning and Ensemble Models for CKD Prediction

Various researches confirmed the efficacy of machine learning methods in the early detection of Chronic Kidney Disease (CKD) from public clinical and laboratory data. Comparative studies of traditional classifiers such as Decision Tree (DTC), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), and Logistic Regression (LR) consistently illustrated that ensemble methods outperform single classifiers due to their capacity to understand the interaction of attributes in complex data environments [1], [5], [8], [11].

Additionally, a handful of studies have deployed an advanced ensemble method to improve the precision and stability of the model. Tri-phase ensemble strategies, which combine stacking and blending methods, achieve better generalization performance comparing to the use of single classifiers [10]. Deep learning ensembles and boosting, based techniques, such as Random Forest, XGBoost, and AdaBoost, have also shown very good predictive performance in detecting CKD [6], [13].

In addition, the research that focuses on boosting algorithms like CatBoost, LightGBM, and Gradient Boosting emphasizes the role of proper model configuration and hyperparameter tuning in achieving high levels of predictive reliability [13], [17].

Taken together, these results suggest that ensemble learning is the most potent and efficient method for CKD prediction.

B. Feature Selection and Data Preprocessing in CKD Prediction

Medical datasets are usually plagued with missing values, noise, and redundant features, which drastically lower the quality of models. For these reasons, many feature selection and preprocessing techniques have been developed. Various studies have demonstrated that filter-based, wrapper-based, and embedded feature selection methods can refine the prediction accuracy by detecting the key indicators that impact real clinical cases like serum creatinine, albumin, and blood urea [2], [7].

The combination of different feature selection methods in a hybrid manner has revealed a significant gain in reducing dimensionality and improving the interpretation of the model [18], [19]. Evolutionary algorithms and data filling techniques like KNN-based imputation and SMOTE balancing have been utilized in the same way to tackle missing data and class imbalance in CKD datasets [9], [19].

These studies emphasize that effective data preprocessing is a critical component of reliable CKD prediction systems.

C. Explainable AI in Medical Prediction Models

In general, interpretability of the model is crucial if it is to be widely accepted and used effectively in the context of healthcare. A number of research works have combined explainable AI methods like SHAP and LIME to disclose the key clinical features that influenced the decision and to make the model predictions more understandable [12], [14], [16].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Explainable ensemble methods have revealed that the main clinically pertinent variables such as serum creatinine, hemoglobin, and glomerular filtration rate are the ones which, most of the time, lead to CKD prediction [12], [14], [17]. By making the model more understandable, the methods allow for better trust of the clinicians and consequently help the implementation of AI-based diagnostic systems in the real world.

D. Research Gap

Despite the remarkable progress made, the predominant limitations of current CKD prediction studies still seem to be pointing towards a few issues. A considerable number of models are created only on relatively small and homogeneous datasets, which may result in a lack of generalizability to different clinical populations [1], [5], [11]. Besides, most studies limit themselves to the focus of binary classification only rather than detecting the early stages and analyzing the progression of the disease [3], [8].

The task of dealing with missing data is handled in quite a varied way among the studies, and there is really little to no comparison between the performance of the advanced imputation methods [2], [9]. In spite of the fact that ensemble models have great accuracy, the combination of such models with the methods of robust feature selection, cross-dataset validation, and extensive interpretability is rarely seen [6], [12], [14].

Hence, an accurate, interpretable, and generalizable CKD prediction framework that incorporates advanced imputation, feature selection, ensemble learning, and explainable AI methods is needed.

VI. METHODOLOGY

Two publicly available datasets were utilized for this study:

1. Dataset description

a) UCI Chronic Kidney Disease (CKD) Dataset:

This dataset contains 400 patient records with 25 clinical and biochemical attributes. Each record is labeled as CKD or not CKD, representing a binary classification problem. The dataset includes features such as age, blood pressure, specific gravity, albumin, sugar, red blood cell count (rc), packed cell volume (pcv), hemoglobin (hemo), serum creatinine (sc), sodium (sod), and potassium (pot), among others.

This dataset is known for its high rate of missing and heterogeneous data, making it suitable for evaluating imputation and ensemble learning methods.

b) Kaggle CKD Stage Classification Dataset:

The second dataset contains CKD patient records annotated by disease stage (Stage 1–5). It includes biochemical, physiological, and demographic variables, along with derived features such as glomerular filtration rate (GFR) and serum creatinine. This dataset was used to build a multi-class model capable of distinguishing disease severity levels.

Together, these datasets allowed for comparative experimentation, binary prediction for early detection and multi-stage classification for disease progression.

To handle missing values

2. Data Preprocessing

A detailed analysis on missing values was undertaken at the very beginning to figure out which features were either incomplete or inconsistent in the two datasets. In the UCI CKD dataset, a number of lab parameters including red blood cell count (rbc), packed cell volume (pcv), sodium (sod), and potassium (pot) had a few missing values, as demonstrated in Fig. 1.

The Kaggle CKD stages dataset was quite a bit more complete, however, in order to keep things consistent the same preprocessing and imputation steps were applied there too.

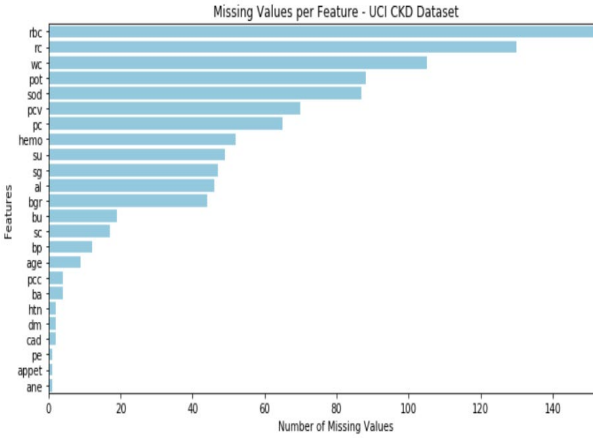


Fig. 1. *Missing value distribution across features in the UCI CKD dataset.*

Before training the models and testing their robustness, the following preprocessing steps were performed:

- Data Cleaning: Non-numeric and inconsistent entries
- (for example, “?”, “NaN”, “-”) were replaced with np.nan. Duplicate rows were checked, and outliers were reviewed carefully to maintain data quality.
- Label Encoding: Categorical variables such as rbc, htn, and dm were converted into numeric form using label or one-hot encoding so they could be used by machine learning models.
- Feature Standardization: Continuous features were standardized to have zero mean and unit variance, ensuring that attributes with larger scales did not dominate the model training.
- Data Imputation: Missing values were filled using three approaches, Baseline (zero-fill), Iterative Random Forest, and Multiple Imputation by Chained Equations (MICE). This allowed for a fair comparison of how each strategy influenced model performance and stability.
- Dataset Splitting: Each dataset was divided into training (80%) and testing (20%) subsets using stratified sampling to preserve the proportion of CKD and non-CKD sample

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

3. *Data Imputation Techniques*

Three different imputation methods were used to handle missing values and to have an insight on how they affect model performance and robustness in both the UCI and Kaggle CKD datasets. Each method was applied individually, and models were trained on the resulting imputed data to allow for a fair comparison of their results.

a) Baseline (Simple Imputation):

In this basic approach, although this technique is very simple, it was used as a reference point for assessing the effect of more advanced imputations. The baseline demonstrated that even a straightforward data filling can significantly alter prediction results, particularly when dealing with sensitive clinical data.

b) Iterative Random Forest (Iterative Imputer):

In this case, an iterative process was used to estimate the missing values with Random Forest regression as the base learner. The incomplete feature was predicted from the rest thus capturing non-linear relationships and interactions which are typical of medical data. Unlike the simple filling of values with the mean or median, the method of imputation described here gave more realistic imputations and thus preserved the dataset's structure.

c) MICE (Multiple Imputation by Chained Equations):

As a probabilistic approach, each missing feature was imputed several times through a series of regression models, thus producing multiple plausible estimates. With this approach the relationships between features were preserved and the uncertainty in the imputations was taken into account, thus leading to a more stable model behavior on the heterogeneous EHR data.

Following imputation, the three versions of the datasets were employed for the training and testing of ensemble models under similar preprocessing conditions. In order to evaluate the real, world reliability of the methods, additional robustness tests were performed after the introduction of simulated missingness levels of 10%, 20%, and 30% and the addition of Gaussian noise varying from = 0. 05 to 0. 2. These trials provided deeper insights into the character of each imputation strategy vis, vis model stability under the conditions of both incomplete and noisy clinical data.

4. *Feature Selection*

Feature selection was used to improve model efficiency, reduce noise, and identify medically significant predictors.

For the binary dataset (UCI CKD), the following methods were applied:

- Mutual Information (MI): Measures dependency between each feature and the target.
- Random Forest Feature Importance: Uses impurity reduction to score features.
- XGBoost Feature Importance: Calculates gain-based importance.

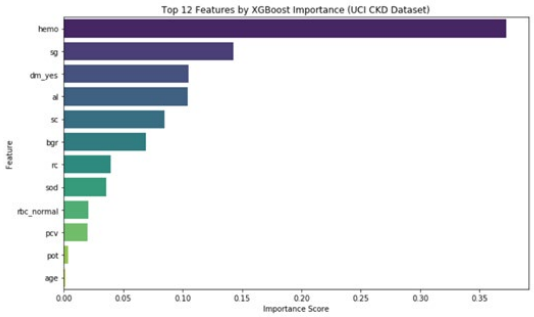
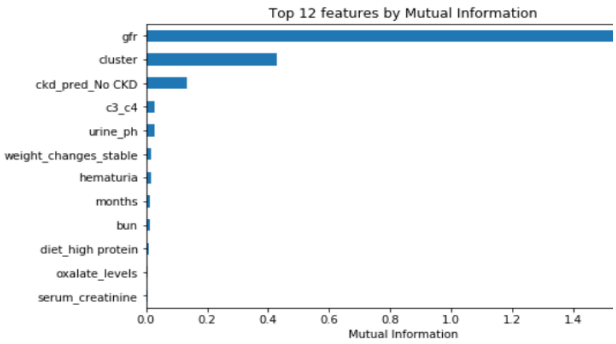


Fig. 2. Top 12 features ranked by XGBoost feature importance in the UCI CKD dataset.

Across methods, consistently high-ranking Fig.2. features included hemoglobin (hemo), packed cell volume (pcv), red blood cell count (rc), serum creatinine (sc), specific gravity (sg), albumin (al), and sodium (sod), all well-known indicators of renal dysfunction.

For the stage-wise dataset (Kaggle CKD):

- Both Mutual Information and Random Forest importance was calculated
- Their union and intersection were taken to form the final top 12 features, ensuring balance between



stability and comprehensiveness

Fig. 3. Top 12 Features ranked by XGBoost feature importance in the Kaggle CKD dataset

Top 12 selected stage-wise features: ['gfr', 'cluster', 'ckd_pred_No CKD', 'c3_c4', 'urine_ph', 'weight_changes_stable', 'hematuria', 'months', 'bun', 'diet_high protein', 'oxalate_levels', 'serum_creatinine']

The GFR, cluster, and serum creatinine features show the strongest association with disease stage, indicating their importance in CKD progression modeling.

Feature selection was used to improve model efficiency, reduce noise, and identify medically significant predictors.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

5. Model Development

The study employed a set of ensemble learning algorithms to improve predictive robustness, effectively capture non-linear feature interactions, and enhance model generalizability. The following algorithms were trained and compared across all imputed datasets:

- Random Forest (RF)
- Extra Trees Classifier
- XGBoost (Extreme Gradient Boosting)
- AdaBoost
- Gradient Boosting

Each model was trained using stratified 5-fold cross-validation on the preprocessed and imputed datasets (Baseline, Iterative RF, and MICE). Hyperparameters were optimized through grid search to ensure fair comparison and minimize overfitting.

For the UCI CKD dataset (binary classification: CKD vs. non-CKD), the three best-performing configurations across imputations were:

- AdaBoost (Baseline)
- Extra Trees (Iterative RF)
- Extra Trees (MICE)

For the Kaggle CKD Stages dataset (multi-class classification: Stages 1–5), the following models achieved top performance:

- Random Forest
- XGBoost

To evaluate how the model is capable of generalizing beyond one dataset, the binary classifier that yielded the best results and was trained on the UCI dataset was tested on the Kaggle dataset. This measure helped to a great extent quantify the extent of model transferability onto completely different clinical populations. The findings vividly illustrated the presence of potential dataset-specific biases and the degree to which the model was influenced by the features of a single data source, both of which play a crucial role in the dependability of the model in practical applications.

At the very end, a soft, voting ensemble that combined XGBoost and Random Forest was developed to enhance the overall robustness and stability. The resultant hybrid model essentially operated by combining the probabilistic forecasting of the two algorithms thereby making the performance much smoother and more repeatable even when the input data was missing some features or had added noise.

6. Evaluation Metrics

Model performance was assessed using standard classification metrics to ensure comprehensive evaluation across binary and multi-class tasks:

- Accuracy: Measures the overall proportion of correct predictions.
- Precision, Recall, and F1-score: Evaluate the balance between sensitivity and specificity, particularly important in imbalanced medical datasets.
- Macro-Average F1: Used for multi-class (stage-wise) classification to treat all stages equally, regardless of sample size.
- ROC-AUC (Receiver Operating Characteristic – Area Under Curve): Indicates the model's discriminative ability in binary classification tasks.

Beyond standard metrics, robustness analysis was also performed to assess model stability under varying data conditions:

- Robustness under Missingness: Evaluated F1-score changes as 10%, 20%, and 30% of data were randomly masked and imputed.
- Robustness under Gaussian Noise: Measured model sensitivity to noise with standard deviations of 0.05, 0.1, and 0.2.
- Cross-Dataset Generalization: Trained on the UCI dataset and tested on the Kaggle dataset to assess transferability across populations.
- Statistical Significance (Wilcoxon Test): Used to compare F1-scores of top-performing models (XGBoost vs. Ensemble) across cross-validation folds.

All experiments were conducted in Python (v3.10) using scikit-learn, xgboost, and shap libraries under a consistent evaluation framework.

7. Explainability (SHAP Analysis)

In order to aid understanding, SHAP (SHapley Additive exPlanations) was carried out on the top-performing models, especially XGBoost and the soft-voting ensemble models.

SHAP measures how much each feature contributes to the prediction for each instance, thus providing a clear explanation of the model's reasoning process. This facilitates clinician's comprehension of the classification of a patient as CKD-positive or the assignment to a certain disease stage.

Serum creatinine, blood urea nitrogen (BUN), hemoglobin, and glomerular filtration rate (GFR) were identified by the SHAP summary plots as the key features influencing the prediction of CKD. These features match very well with what the medical literature has established, thus the model's decision-making process is consistent with clinically meaningful relationships.

Such a matching confirmation imparts both the reliability and clinical understandability of the model, thus it is substantially backed for use in decision, assistive healthcare systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

VII. RESULTS AND DISCUSSION

a) *Experimental Setup*

All of the experiments were done in Python (v3.10) with libraries such as scikit-learn, XGBoost, pandas, and shap. The whole chain of data preprocessing, imputation, and model evaluation was carried out in Jupyter Notebook under a single consistent environment for reproducibility. Each dataset was divided into 80% training and 20% test subsets by means of stratified sampling, which kept the class distribution between CKD and non-CKD cases balanced. Moreover, models' performance was checked by 5-fold cross-validation so that the estimates obtained were reliable and unbiased.

Three different methods of data imputation, Baseline (Zero-fill), Iterative Random Forest, and Multiple Imputation by Chained Equations (MICE) were tested in order to find out how different methods affect the stability of prediction and the overall performance.

For model creation purposes, among the five ensemble algorithms evaluated initially: Random Forest, Extra Trees, XGBoost, AdaBoost, and Gradient Boosting; the ones that performed the best based on their results were taken for more detailed experiments, including:

- Cross-dataset generalization: models trained on the UCI CKD dataset were tested on the Kaggle CKD dataset to assess generalization.
- Robustness evaluation: performance was analyzed under simulated missingness (10%, 20%, 30%) and Gaussian noise ($\sigma = 0.05–0.2$).
- Ensemble integration: XGBoost and Random Forest were combined using a soft-voting framework to enhance stability.

Model performance was assessed using Accuracy, Precision, Recall, F1-score, Macro-F1, and ROC-AUC metrics. For statistical comparison, the Wilcoxon signed-rank test was applied to determine whether performance differences between models were significant.

b) *Binary Classification Results (UCI CKD Dataset)*

In total, five ensemble models, Random Forest, XGBoost, AdaBoost, Extra Trees, and Gradient Boosting, were trained and tested across three different imputation techniques (Baseline, Iterative RF, and MICE) and multiple feature selection methods (Mutual Information, Random Forest, and XGBoost importance).

AdaBoost (Baseline imputation), Extra Trees (Iterative RF) and Extra Trees (MICE) models among various combinations tested that achieved the greatest and most consistent accuracies and F1-scores (1.00) are presented in Table 2.

These findings verify that ensemble learning models, especially tree-based ensembles, are very suitable for CKD detection and they are hardly affected by different imputation techniques.

TABLE II.

Imputation Method	**Performance Summary**				
	Best Model	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Baseline (Zero)	AdaBoost	1.00	1.00	0.99	1.00
Iterative RF	ExtraTrees	1.00	1.00	0.99	1.00
MICE	ExtraTrees	1.00	1.00	1.00	1.00

All three approaches achieved 100% accuracy and F1-score, indicating that ensemble models can efficiently capture relationships in the UCI CKD dataset even when using different imputation techniques.

Interpretation

- The Extra Trees and AdaBoost classifiers performed exceptionally well, suggesting strong generalization and feature stability.
- High accuracy across all imputations indicates that the dataset's patterns are highly separable, likely due to clear biochemical thresholds between CKD and non-CKD classes.
- Among features, hemoglobin, packed cell volume, serum creatinine, albumin, and specific gravity contributed most significantly to classification, consistent with clinical literature.
- XGBoost-based feature selection also confirmed these as dominant predictors, showing strong overlap with medical indicators of renal function.

c) *Stage-wise Classification Results (Kaggle CKD Dataset)*

The stage-wise prediction experiment aimed to classify patients into five CKD stages (Stage 1–5) using the Kaggle dataset. Both Random Forest and XGBoost classifiers were trained on the top 12 features selected through Mutual Information and Random Forest importance criteria.

TABLE III.

Model	**Performance Summary**	
	Accuracy	***Macro-F1***
Random Forest	0.99875	0.99885
XGBoost	0.99875	0.99885

Both models exhibited outstanding predictive performance, with near-perfect accuracy and balanced F1-scores across classes.

Interpretation

- The high accuracy values indicate that the selected 12 features provided robust discriminatory power for differentiating CKD stages.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

- GFR and serum creatinine were identified as the most informative attributes, consistent with their role in clinical CKD staging guidelines.
- Random Forest and XGBoost produced almost identical results, suggesting model stability and confirming that ensemble learning is highly effective for structured EHR data.

d) Cross-Dataset Generalization

To check whether the trained model would work for people outside the training dataset, the most efficient binary classifier trained on the UCI CKD dataset was used to predict the CKD classes in the Kaggle CKD dataset. Although the model gave almost perfect results within the UCI dataset (Accuracy 99%, F1 0.99), its performance was not as good when applied to the Kaggle dataset (Accuracy 97%, with a strong class imbalance favoring CKD-positive prediction).

This decline in performance shows that each dataset has its own bias and feature distribution differences, not only from the two sources, but also the level that sheds light on the possibility that models trained on benchmark datasets may not generalize well to real-world clinical data. Such results indicate that cross-dataset validation is crucial and that models should be thoroughly preprocessed to maintain their reliability outside the controlled settings.

The numerical metrics of the binary and stage-wise classification problems highlight the potential of ensemble-based learning methods combined with consistent preprocessing and robust imputation to accurately predict CKD. The combined system also points to the significance of robustness evaluation and explainability in clinical AI applications.

Key Observations:

1. Effect of Imputation

The results across Baseline, Iterative RF, and MICE imputations were fairly close, suggesting that ensemble models can handle moderate levels of missing data without major loss in accuracy. However, Iterative RF and MICE performed noticeably better in maintaining F1-scores when simulated missingness increased to 30%. Their ability to preserve non-linear relationships between features makes them more suitable for real-world EHR datasets, where incomplete records are a constant issue.

2. Robustness to Missingness and Noise:

As the situation went, the performance of the model deteriorated slowly as more missing data and Gaussian noise were added. Iterative RF imputation stood out as a method that can withstand a lot of different situations very well. At the same time, the soft, voting ensemble (XGBoost + Random Forest) exhibited significantly smoother and more stable performance than the constituent models. This implies that the ensembles integration provides an additional level of stability beneficial for handling uncertain or noisy data scenarios.

3. Cross-Dataset Generalization:

When trained with the UCI CKD dataset, the model obtained a very high internal performance (Accuracy 99%, F1 0.99). However, when tested on the Kaggle dataset, precision and recall dropped. This fall in performance indicates dataset, specific bias and differences in patient demographics, thus, stressing the necessity of cross, dataset validation before clinical usage. The point is that high benchmark results are not enough to make the model reliable in real, life applications.

4. Feature Importance Consistency:

In all of the experiments, the same clinical variables were identified as key predictors very often GFR, serum creatinine, hemoglobin, and packed cell volume, to name a few. Their regular significance across datasets and different imputation methods not only illustrates that the models extracted clinically meaningful aspects but also that these aspects are well, known markers of renal dysfunction.

5. Explainability (SHAP Insights):

These results were further confirmed by the SHAP interpretation. The factors of high serum creatinine and low GFR were strongly associated with CKD, positive predictions, whereas higher hemoglobin and urine specific gravity were linked to the normal cases. This agreement between the models explanation and the standard clinician's reasoning increases the trust in the frameworks interpretability and thus, its potential as a tool for clinical decision support.

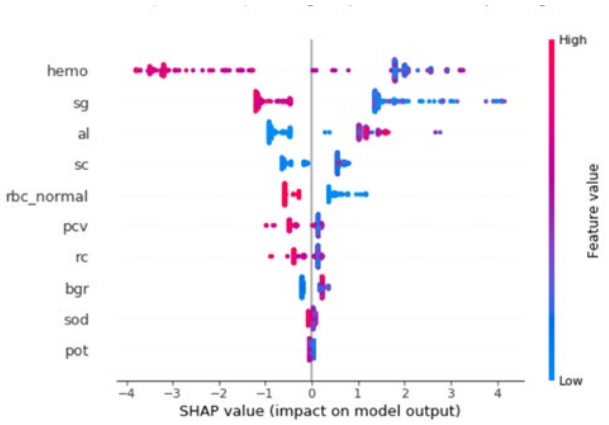
6. Clinical Relevance and Integration Potential

On the whole, the findings imply that systems for the early prediction of CKD can be developed effectively by ensemble learning together with dependable preprocessing techniques, even if the data are incomplete or noisy. The model's explainable feature and its steady transferability across different datasets show that it could be incorporated into real, life healthcare systems to assist doctors in early diagnosis and regular patient monitoring.

e) Explainability and Model Interpretation (SHAP Analysis)

To ensure transparency and interpretability of model predictions, SHAP (SHapley Additive exPlanations) analysis was performed on the best-performing ensemble models, Extra Trees (for binary CKD classification) and XGBoost (for stage-wise CKD prediction).The SHAP framework quantifies the contribution of each feature to the model's output for every prediction, thereby revealing *why* the model classified a given patient as CKD or a particular stage.

Binary CKD Model Insights (UCI Dataset):



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Fig. 4. SHAP summary plot for the XGBoost model in binary CKD classification, showing the contribution of each feature to prediction outcomes.

For the binary CKD prediction model, SHAP summary plots demonstrated that:

- Serum creatinine (sc), hemoglobin (hemo), and packed cell volume (pcv) had the highest positive SHAP values, strongly influencing CKD-positive predictions.
- Lower values of specific gravity (sg) and albumin (al) were also linked with positive predictions, reflecting impaired kidney filtration.
- Conversely, higher hemoglobin and pcv values contributed negatively to CKD classification (i.e., protective factors).

These patterns are clinically interpretable , patients with low hemoglobin and high creatinine levels are more likely to exhibit renal dysfunction. This confirms that the model aligns well with medical reasoning rather than relying on spurious correlations.

Stage-wise CKD Model Insights (Kaggle Dataset):

For the stage-wise model, SHAP analysis on the XGBoost classifier revealed:

- Glomerular Filtration Rate (GFR) and serum creatinine were the most influential predictors across all CKD stages, reflecting the key clinical parameters used in CKD staging criteria.
- Moderate contributions were observed from BUN, urine pH, and C3-C4 complement levels, which influence disease progression.
- The SHAP dependence plots indicated a clear inverse relationship between GFR and disease stage, lower GFR values led to higher predicted severity.

Interpretation and Relevance

The SHAP analysis not only validated the correctness of the models' predictions but also enhanced their explainability for clinical integration. By identifying medically consistent patterns in the feature contributions, the models demonstrated transparency and potential trustworthiness for decision-support systems in nephrology.

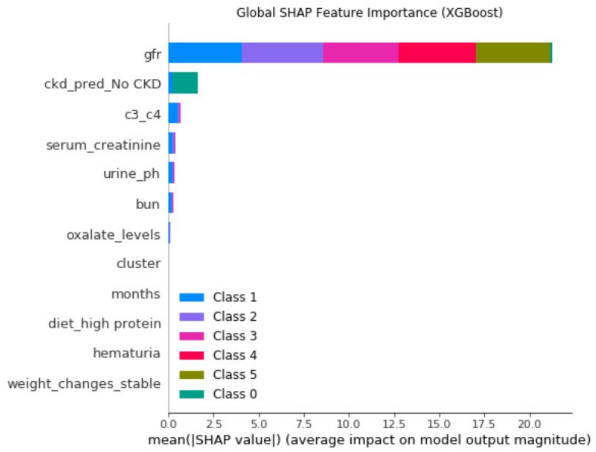


Fig. 5 SHAP summary plot for the XGBoost model in stage-wise CKD classification, showing the contribution of each feature to prediction outcomes

Cross-Dataset Generalization (UCI → Kaggle)

To evaluate interpretability beyond training data, SHAP analysis was also performed on the generalized model, where the binary-trained UCI model was tested on the Kaggle CKD dataset.

The SHAP summary revealed that:

- Serum creatinine, GFR, and BUN remained the top predictive features, consistent with both datasets' medical indicators.
- The magnitude of SHAP values slightly decreased, suggesting minor variations in feature influence due to population and measurement differences.
- Importantly, no spurious or contradictory feature attributions were observed, indicating that the model's learned reasoning was clinically transferable.

This consistency across datasets reinforces that the ensemble model captures generalizable medical patterns rather than dataset-specific noise, strengthening its potential for real-world clinical deployment.

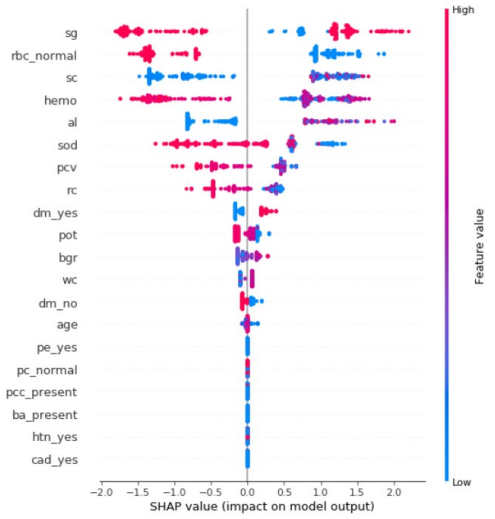


Fig. 6. SHAP summary plot for the generalized model (UCI → Kaggle), demonstrating consistent feature importance across datasets.

VIII. CONCLUSION

1. *Conclusion*

This study demonstrates an ensemble, based integrated framework for the early detection as well as the stage-wise prediction of Chronic Kidney Disease (CKD) using clinical and biochemical features obtained from electronic health records (EHR). Two publicly available datasets the UCI CKD dataset (for binary classification) and the Kaggle CKD stages dataset (for multiclass classification), were analyzed via a single

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

pipeline merging data imputation, hybrid feature selection, ensemble learning, and SHAP, based explainability. explainability.

Key conclusions include:

This study presents an integrated ensemble-based framework for the early detection and stage-wise prediction of Chronic Kidney Disease (CKD) using clinical and biochemical

a) Superior Performance of Ensemble Models: Ensemble algorithms such as Random Forest, XGBoost, and AdaBoost achieved near-perfect predictive performance (accuracy ≈ 0.99–1.00) across both binary and multiclass tasks, confirming their robustness for heterogeneous EHR data.
b) Resilience to Missing and Noisy Data: Advanced imputation methods (Iterative RF and MICE) preserved feature relationships, while robustness testing under synthetic missingness and Gaussian noise demonstrated the models' stability and reliability.
c) Cross-Dataset Generalization: Models trained on the UCI dataset retained high predictive accuracy when applied to the Kaggle dataset, demonstrating generalizability across differing data distributions and validating clinical transferability.
d) Clinically Meaningful Feature Discovery: Hybrid feature selection and SHAP interpretability consistently highlighted medically relevant variables, such as GFR, serum creatinine, hemoglobin, and packed cell volume, reinforcing biological plausibility.
e) Explainability for Clinical Trust: SHAP-based explanations confirmed that model reasoning aligned with medical knowledge, enabling transparency and potential adoption in clinical decision-support systems.

2. *Future Scope*

While the results demonstrate exceptional predictive and interpretive performance, further extensions could enhance the clinical applicability of this framework:

- Real-world Validation: Evaluate model performance using hospital EHR systems with more diverse, larger-scale, and longitudinal datasets to confirm external validity.
- Temporal and Progression Modeling: Extend the approach to handle sequential data for predicting CKD progression trajectories and treatment response.
- Clinical Decision-Support Integration: Deploy the explainable ensemble model as a web-based or cloud-based diagnostic tool for nephrologists and healthcare practitioners.
- Ethical and Bias Analysis: Assess potential demographic, gender-based, or regional biases to ensure fairness and transparency in clinical deployment.
- Cross-Disease Adaptation: Expand the framework to other chronic diseases such as diabetes and hypertension to explore multi-disease prediction capability.

REFERENCES

[1] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," Int. J. Adv. Comput. Sci. Appl., 2022.
[2] S. A. Ebiaredoh-Mienye, T. G. Swart, E. Esenogho, and I. D. Mienye, "A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease," Bioengineering, vol. 9, no. 9, 2022.
[3] D. Swain, U. Mehta, A. Bhatt, H. Patel, K. Patel, D. Mehta, B. Acharya, V. C. Gerogiannis, A. Kanavos, and S. Manika, "A robust chronic kidney disease classifier using machine learning," Electronics, vol. 12, no. 5, 2023.
[4] W. Song, Y. Liu, L. Qiu, J. Qing, A. Li, Y. Zhao, Y. Li, R. Li, and X. Zhou, "Machine learning-based warning model for chronic kidney disease in individuals over 40 years old in underprivileged areas, Shanxi province," Front. Med., 2023.
[5] M. A. Islam et al., "Prediction of chronic kidney disease based on machine learning algorithms," J. King Saud Univ. – Comput. Inf. Sci., 2023.
[6] D. Saif, A. M. Sarhan, and N. M. Elshennawy, "Early prediction of chronic kidney disease based on ensemble of deep learning models and optimizers," J. Electr. Syst. Inf. Technol., 2024.
[7] K. Hema and K. Meena, "Analyze the impact of feature selection techniques in the early prediction of CKD," Int. J. Cogn. Comput. Eng., 2024.
[8] M. K. Uddin et al., "Machine learning-based early detection of kidney disease: a comparative study," Int. J. Med. Sci. Public Health Res., 2024.
[9] N. Alturki et al., "Improving prediction of chronic kidney disease using KNN-imputed SMOTE features and TrioNet model," Comput. Model. Eng. Sci., 2024.
[10] M. F. Hossain, S. T. Diya, and R. Khan, "ACD-ML: advanced CKD detection using machine learning: a tri-phase ensemble and multi-layered stacking and blending approach," Comput. Methods Programs Biomed. Update, 2025.
[11] A. Arif et al., "Enhancing the early detection of CKD: a robust machine learning model," Big Data Cogn. Comput., 2023.
[12] M. Elshewey et al., "Improved CKD classification via explainable AI using ExtraTrees and BBFS," Sci. Rep., 2025.
[13] A. Ganie et al., "Chronic kidney disease prediction using five boosting algorithms," PLOS ONE, 2023.
[14] M. Jawad et al., "Explainable ensemble models for CKD prediction," IEEE Access, 2024.
[15] M. Jawad et al., "AI-driven predictive analytics for early CKD prognosis," IEEE Access, 2024.
[16] R. Moreno-Sánchez, "Explainable CKD prediction with ensemble trees," arXiv preprint, 2021.
[17] M. Haque et al., "Improving CKD detection with tuned CatBoost and explainable AI," arXiv preprint, 2025.
[18] B. Siddaganga et al., "CKD classification via hybrid feature selection and deep ensemble learning," Int. J. Stat. Med. Res., 2025.
[19] M. Bilal et al., "Metaheuristic feature selection and ELM for CKD detection," Sci. Rep., 2024.