Огляд
Review

НИРКИ KIDNEYS

DOI: 10.65327/kidneys.v15i1.605

Edwina Jospiene M[1], Nimithap S[2], Dr Subhadip Bag[3], Dr. S. Elavarasan[4], Dr. Prolay Ghosh[5], Sathiyamoorthy M[6], S.T. Gopukumar[7]

[1*]*Research Scholar, Department of Biochemistry, Biochemistry/Biotechnology, Regenix Super Speciality Laboratories pvt.ltd., Affiliated to UNIVERSITY OF MADRAS, Chennai 94, Email ID: edwina18@ymail.com, Orcid ID: 0000-0002-6523-6087*
[2]*Junior Resident, General Medicine, Department of General Medicine, Saveetha Medical College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Tamil Nadu, Chennai-602105, India, Email Id: nimithap0001@gmail.com , ORCID: 0009-0002-1990-0124*
[3]*Assitant Professor, Hi-Tech Medical College & Hospital, Rourkela MBBS, MD Community Medicine, Odisha University of Health Sciences, Rourkela-769004, India, Email ID: dr.subhadip07@gmail.com, Orcid ID: 0000-0002-8064-429X*
[4]*Associate Professor, Department of Community Medicine, Specialization in Research Methodology & Biostatistics, Sri Sairam Homoeopathy Medical College & Research Center, West Tambaram, Chennai -600 044, Email ID: dr.s.elavarasan@gmail.com, Orcid : 0000-0001-7317-4309*
[5]*Assistant Professor, Department of Information Technology, JIS College of Engineering Kalyani, Nadia, West Bengal-741235, India, Email ID: prolay.ghosh@jiscollege.ac.in, Orcid ID: https://orcid.org/0000-0001-9267-5766*
[6]*Assistant Professor, Department of Computer Science and Engineering, Computer Science and Engineering; Artificial Intelligence; Machine Learning, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai – 602105, Tamil Nadu, India, Email: sathiyamoorthym.sse@saveetha.com, Orcid ID: https://orcid.org/0009-0002-2190-1230*
[7]*Nanobioinformatics Unit, Helix Research Studio, Department of General Surgery, Saveetha Medical College and Hospital, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai – 602 105, Tamil Nadu, India, Email: gopukumars.smc@saveetha.com , Orcid ID: 0000-0001-8160-2414*

# Machine Learning Algorithms for Predicting CKD Progression: A Real-World Hospital Dataset Analysis

*For citation: Kidneys. 2026;15(1):01-08. Acceptance- 03/12/2025*
*Received- 19/11/2025 Doi: 10.65327/kidneys.v15i1.605*

**Abstract**
**Background.** Chronic Kidney Disease (CKD) is a progressive condition associated with substantial global morbidity and mortality. Early detection remains critical for reducing complications and slowing progression to end-stage kidney disease. Traditional diagnostic approaches depend on laboratory markers that may not fully capture nonlinear interactions among clinical parameters. Machine learning offers promising capabilities for improving early identification and supporting clinical decision-making. **Methods.** This study developed an end-to-end machine learning framework for CKD prediction using the Early Stage CKD dataset. The workflow included rigorous data preprocessing, exploratory data analysis, and feature engineering prior to model development. A Random Forest classifier was trained using an 80/20 stratified split, and performance was assessed using accuracy, precision, recall, F1-score, confusion matrix, and ROC–AUC. To enhance transparency, SHAP (SHapley Additive exPlanations) analysis was applied to interpret feature contributions and validate clinical relevance. **Results.** The Random Forest model demonstrated excellent predictive performance, achieving an accuracy of 96.25% and a ROC–AUC of 1.00. The confusion matrix indicated zero false positives and only three false negatives, reflecting strong diagnostic reliability. SHAP analysis identified hemoglobin, serum creatinine, packed cell volume, and specific gravity as the most influential predictors, aligning with established CKD biomarkers. **Conclusion.** The proposed machine learning framework offers a robust, interpretable approach for early CKD prediction. Its strong performance and explainability make it suitable for integration into real-world clinical decision-support systems, particularly in resource-limited healthcare settings.

**Keywords:** Chronic Kidney Disease, Machine Learning, Random Forest, SHAP, Clinical Decision Support.

## 1. Introduction

Chronic Kidney Disease (CKD) is a progressive and irreversible disease, which is marked by gradual deterioration of the renal functions, leading to the inability of kidneys to regulate the metabolic, electrolyte and fluid balance. It is known to be a significant global health problem because of its rising prevalence, great economic impact, and a close relationship with cardiovascular morbidity and mortality. The definition and classification of CKD as developed by the Kidney Disease: Improving Global Outcomes (KDIGO) consortium officially focused on the clinical


For correspondence: Edwina Jospiene M, Research scholar, department of biochemistry biochemistry/biotechnology Regenix super speciality laboratories pvt. ltd., Affiliated to UNIVERSITY OF MADRAS Chennai 94 Email: edwina18@ymail.com orcid id. 0000-0002-6523-6087
Full list of authors information is available at the end of the article.

implications of CKD on a long-term scale, as well as the role of early diagnosis in preventing complications and delaying the transition to end-stage kidney disease (ESKD) [1]. Later studies supported CKD as a health priority in the world, and the disease burden in different regions and populations varies significantly [2]. Recent years of Global Burden of Disease (GBD) studies have consistently shown that CKD is among the major causes of untimely deaths globally, and the disease burden has been steadily increasing since 1990 to 2017 [3], and steadily increasing at an alarming rate as of 2021 [4].

The etiologies that have a strong impact on CKD are diabetes mellitus, hypertension, metabolic syndrome, genetic susceptibility, and exposure to nephrotoxic agents. Modern epidemiological studies prove that diabetes and hypertension lead to over two-thirds of all cases of CKD in the world [7,8]. The recent consensus statements and clinical guidelines emphasize the importance of early detection of individuals at-risk, better disease surveillance systems, and quicker adoption of decision-support technologies to better diagnose and intervention strategies [9]. According to reports by the United States Renal Data System (USRDS), the prevalence and healthcare costs associated with CKD and ESKD have been increasing, which supports the significance of preventive measures based on the strong clinical assessment instruments [10].The conventional CKD diagnostic processes are based on lab results, such as serum creatinine, blood urea nitrogen, estimation of glomerular filtration rate, and urinalysis. Although these markers are necessary, they are constrained by inter-individual variability, late disease presentation, and the lack of sufficient ability to measure complex nonlinear interactions between risk factors. To address these drawbacks, scholars are increasingly resorting to machine learning (ML) to enhance CKD identification, prognosis, and individual care plans.An increasing amount of literature shows that ML algorithms are effective in CKD prediction based on clinical features. A number of studies have indicated that ML approaches are superior to the traditional statistical techniques in detecting patterns that can be used to indicate renal impairment. Debal and Sitote (2022) were able to apply ML techniques to CKD datasets and reported that they achieved high predictive accuracy relative to the traditional techniques [11]. Equally, Islam et al. (2023) demonstrated that using simple clinical parameters, ML models and especially ensemble-based algorithms can predict CKD with a high level of accuracy [12]. Random Forest and decision-tree models have been shown to be effective many times in the CKD classification because of their robustness, interpretability, and the capacity to deal with heterogeneous data sources. Subasi et al. (2017) showed that the practical use of the Random Forest is useful in the diagnosis of CKD, and the algorithm has the strength of managing mixed data types [13]. Recent studies are still showing similar findings that support the relevance of tree-based classifiers in the prediction of CKD [14–16].Besides the classic ML algorithms, recent studies have considered the use of advanced feature engineering, risk factor identification, as well as the integration of explainable artificial intelligence (XAI) to improve clinical trust and model transparency. The study by Mendapara (2024) used the Random Forest classifiers to build a risk prediction model and discovered that serum creatinine, hemoglobin, and packed cell volume were the key biomarkers of early-stage CKD [17]. Explainable AI has become a significant need of healthcare ML applications, particularly in diseases where interpretability is a determinant of clinical acceptance, such as CKD. Singamsetty et al. (2024) have included SHAP explainability in the CKD prediction models and have shown significant increases in the model interpretability and diagnostic reasoning [18]. In a similar manner, Liu et al. (2024) utilized the power of the Random Forest algorithms to assess the risk factors of CKD and highlighted the use of the algorithm in clinical risk stratification [19]. New medical IoT systems have also incorporated generative adversarial networks (GANs), few-shot learning, and XAI to enhance CKD prediction and better model generalization on small or imbalanced datasets [20].Although such advancements have taken place, there are a number of challenges. Most of the current studies are based on small or local datasets, do not have comprehensive explainability models, or do not show clinical-based interpretation of predictive features. Also, there are data preprocessing, feature selection, and model evaluation strategy differences across studies, which have contributed to inconsistent performance results. These gaps in research indicate that a single, interpretable, and clinically consistent ML framework can provide predictable CKD outcomes in a wide range of clinical environments.

The current paper will fill these gaps by designing an end-to-end machine learning pipeline to predict CKD using the Early-Stage CKD dataset, which incorporates a combination of rigorous preprocessing, exploratory data analysis, feature importance analysis, Random Forest classification, and SHAP-based interpretability. The main aim is to assess the predictive accuracy of a Random Forest classifier in the correct classification of CKD and non-CKD cases. The study will also aim to determine the most significant clinical predictors based on both model-based feature importance metrics and SHAP analysis, which will allow gaining a better understanding of the factors that determine CKD classification. Another objective is to show clinical interpretability and practical applicability of the proposed model, such that the predictions are clear and consistent with the existing clinical knowledge. The combination of high-accuracy prediction with explainability makes this integrated approach more useful and reliable in clinical decision-support applications in the real world, as the system becomes more appropriate to the real-world clinical decision-support setting.

## 2. Methodology

### 2.1 Data Source and Description

The sample utilized in this research is based on a CKD cohort of 400 patient records (250 CKD and 150 non-CKD), which is based on a hospital [21]. Every case has 25 clinical attributes that are demographic variables (age, blood pressure), biochemical (blood urea, serum

creatinine, sodium, potassium), hematological (hemoglobin, packed cell volume, red blood cell count, white blood cell count) and categorical (presence of hypertension, diabetes mellitus, anemia, coronary artery disease, appetite status, and pedal edema). The dependent variable is a binary class (CKD/non-CKD).

**2.2 Data Preprocessing**

Preprocessing was performed to ensure data quality and model readiness. First, all missing values represented by "?" were converted into the standard NaN format. Numerical features, including age, blood pressure, specific gravity, albumin, sugar, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, and red blood cell count, were explicitly cast into float format. Categorical variables such as red blood cell morphology, pus cell, pus cell clumps, bacteria, hypertension status, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia, and class labels were encoded using label encoding.To deal with missing values, numerical attributes were filled in with the median of non-missing observations, which maintains the distributional properties and minimizes the effects of outliers. The mode was used to impute categorical variables. After imputation, numerical features were standardized with the help of the StandardScaler that converted the values to zero mean and unit variance. This made sure that the features that had bigger ranges like blood glucose or urea levels did not affect the model training disproportionately. The data was then separated into training and testing data sets in 80:20 ratio in a stratified manner to maintain the ratio of classes.

**2.3 Exploratory Data Analysis (EDA)**

Exploratory Data Analysis was done to learn about the patterns that are underlying in the data. All variables were calculated using summary statistics, including the measures of central tendency, dispersion, and the frequency counts of the categorical attributes. The distributions of features were inspected by visualization tools like histograms, box plots, and bar charts that helped to identify skewness or abnormalities.

The distribution of the classes showed that there was an imbalance with more cases of CKD. Numerical variables were analyzed using histograms, and the results were clinically consistent: serum creatinine and blood urea have right-skewed distributions, with high concentrations typical of CKD patients; hemoglobin and packed cell volume have lower concentrations across CKD cases, which is the result of anemia caused by kidney dysfunction.Numeric variables were correlated to determine the presence of multicollinearity. Hemoglobin, packed cell volume, and red blood cell count were found to have strong positive relationships with serum creatinine, which was found to be strongly related to blood urea and potassium. These trends were in line with established pathophysiological indicators of CKD, which increases the interpretation of the model and validates clinical significance.

**2.4 Feature Engineering**

The process of feature engineering entailed the extraction of insights based on variable transformations and variable relevancy analysis. The analysis of the feature importance of the random forest revealed that the most significant predictors were hemoglobin, serum creatinine, packed cell volume, specific gravity, red blood cell count, albumin, and random blood glucose. SHAP (SHapley Additive exPlanations) analysis gave more insight into interpretability, as it quantified the contributions of individual features to model predictions, and it verified that renal functions markers and hematological indicators were dominant. This two-layered interpretability model was used to provide transparency in the decision-making process of the model.

**2.5 Model Development**

The use of a Random Forest classifier was based on its strength, capability of nonlinear relationship, and overfitting resistance. The 200 decision trees and constant random state were used to train the model to provide reproducibility. Maximum depth, minimum samples per split and estimators number were hyperparameters that were chosen after experimentation to ensure the balance between accuracy and computational efficiency.

The processed training dataset was used to train the model and the independent test set to evaluate the model. Class labels and probability scores were generated, which allowed computing a Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC).

**2.6 Performance Evaluation**

The metrics used to assess model performance were accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. The Random Forest classifier provided high accuracy of 96.25, precision, and recall of the CKD and non-CKD classes. The confusion matrix proved the good discriminative ability with little false negatives and no false positives with the non-CKD category. The ROC curve showed great separability among classes with an AUC of 1.00. The validity of the model predictions in a clinical sense was also determined with the help of feature importance and SHAP visualizations.

**2.7 Ethical Considerations**

As the dataset is fully anonymized and publicly available for research, no personally identifiable information was used. The study complies with ethical guidelines for secondary data analysis.

**2.8 Model Framework**

The suggested machine learning algorithm to predict CKD is based on a well-organized, vertically combined pipeline that is intended to be robust, reproducible, and clinically interpretable. It starts with the process of data acquisition by using Early Stage CKD Dataset, the UCI Machine Learning Repository, where 400 patient records and 25 demographic, biochemical, and hematological features gathered in Apollo Hospitals are available. After acquisition, data preprocessing is performed, including the conversion of missing values

("?") to NaN, imputation using median (numerical) and mode (categorical), label encoding of categorical attributes, and standardization of numerical variables to ensure uniform scaling. Exploratory Data Analysis (EDA) is performed to learn about the underlying trends in terms of summary statistics, histograms, visualization of class distribution, and correlation matrices, which show clinically significant trends as well as point out minor class imbalance. The feature engineering includes the use of the feature importance of the Random Forest and SHAP (SHapley Additive exPlanations) to determine the most significant predictors, including hemoglobin, serum creatinine, packed cell volume, specific gravity, and red blood cell count, in line with the CKD pathophysiology. The development phase of the model involves a Random Forest classifier of 200 estimators, which was trained on an 80/20 stratified split to ensure the integrity of the distribution of the classes. Accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix are used to measure the model performance, and the proposed model demonstrates good outcomes (96.25% accuracy, AUC 1.00). Lastly, SHAP-based explainability offers clear, clinically comprehensible information about the decision process of the model, improving its plausibility and improving its possible role in the real-life healthcare decision support.

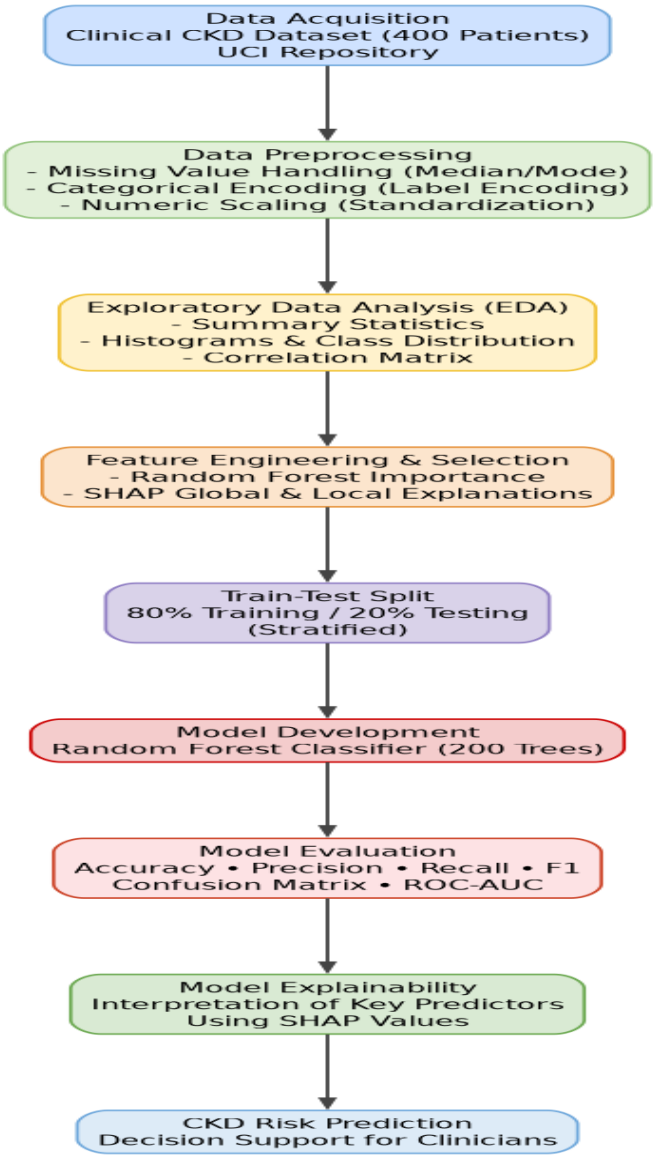


**Figure 1. Proposed machine learning framework for CKD prediction, outlining the workflow from dataset acquisition to preprocessing, model training, evaluation, and SHAP-based explainability**

## 3. Results

### 3.1 Exploratory Data Analysis (EDA)

To examine the distribution of the major clinical features, histograms were created on major numerical variables such as age, blood pressure (bp), blood glucose random (bgr), blood urea (bu), serum creatinine (sc), hemoglobin (hemo), and packed cell volume (pcv). Figure 2 shows the distributions of the most essential numerical biomarkers.

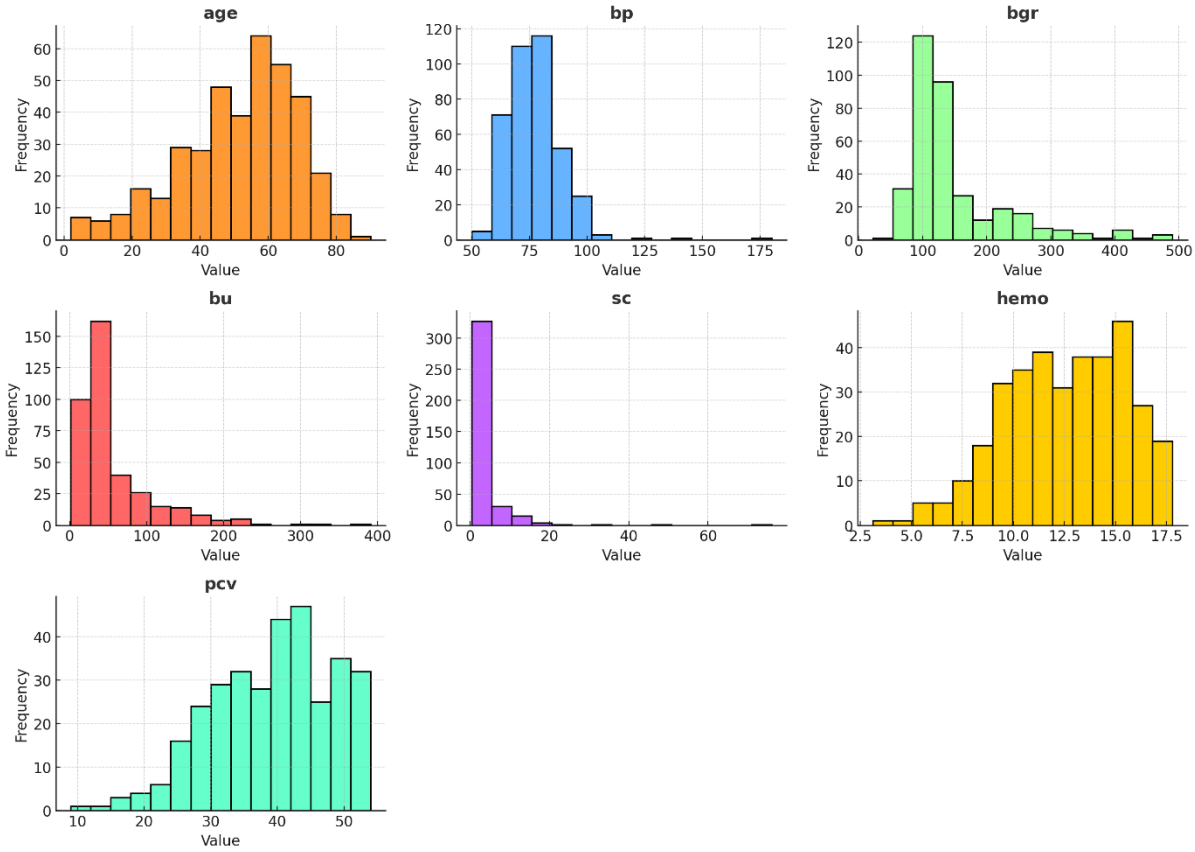


**Figure 2. Histograms of selected numerical features showing distribution patterns for age, bp, bgr, bu, sc, hemo, and pcv**

The histograms indicate medically consistent trends. The age distribution is concentrated between the 40-70 years, which is in agreement with increased CKD prevalence in middle-aged and older adults. Blood pressure is skewed to the right, and it means that there is a significant hypertensive group. Blood glucose random (bgr) and blood urea (bu) have significant right skewness as there are high levels of metabolic markers in CKD patients. Serum creatinine (sc) is skewed sharply with a high number of outliers in CKD cases whereas hemoglobin and pcv show a skewed distribution with a shift towards lower values in CKD cases, which is a known complication of CKD, anemia.

### 3.2 Class Distribution Analysis

The information on the label distribution of the classes is critical to the development of any predictive model because unequal datasets may skew the learning algorithms and misrepresent the performance indicators. The ratio of CKD to non-CKD cases in the current CKD data was also analyzed to determine whether a balance strategy or weighted assessment measures would be required. The analysis will give an understanding of the structural features of the dataset and will make sure that the model that is developed later will be based on the actual population.Figure 3 shows the class balance between CKD and non-CKD cases.

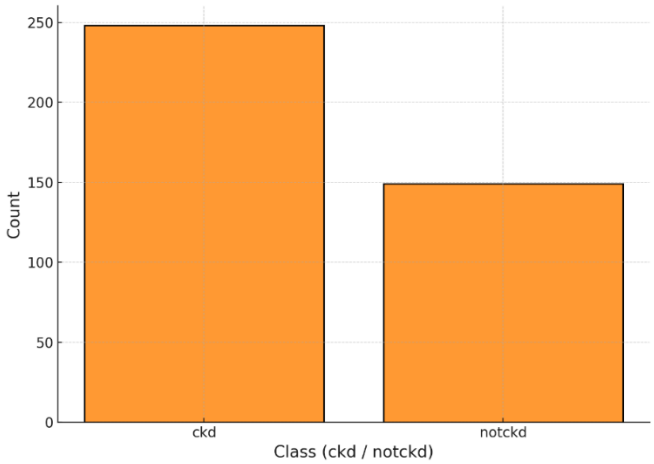


**Figure 3. Class distribution of CKD and non-CKD samples**

There are 250 CKD cases and 150 non-CKD cases in the dataset with a minor imbalance of classes (62.5% vs. 37.5%). Though this is manageable, this imbalance supports the significance of considering recall and precision, in that the model does not overrepresent the majority class.

### 3.3 Correlation Structure of Numerical Features

Interpretation of downstream model behaviour requires the understanding of feature-to-feature relationships. Figure 4 provides relationships between numerical clinical variables**.**

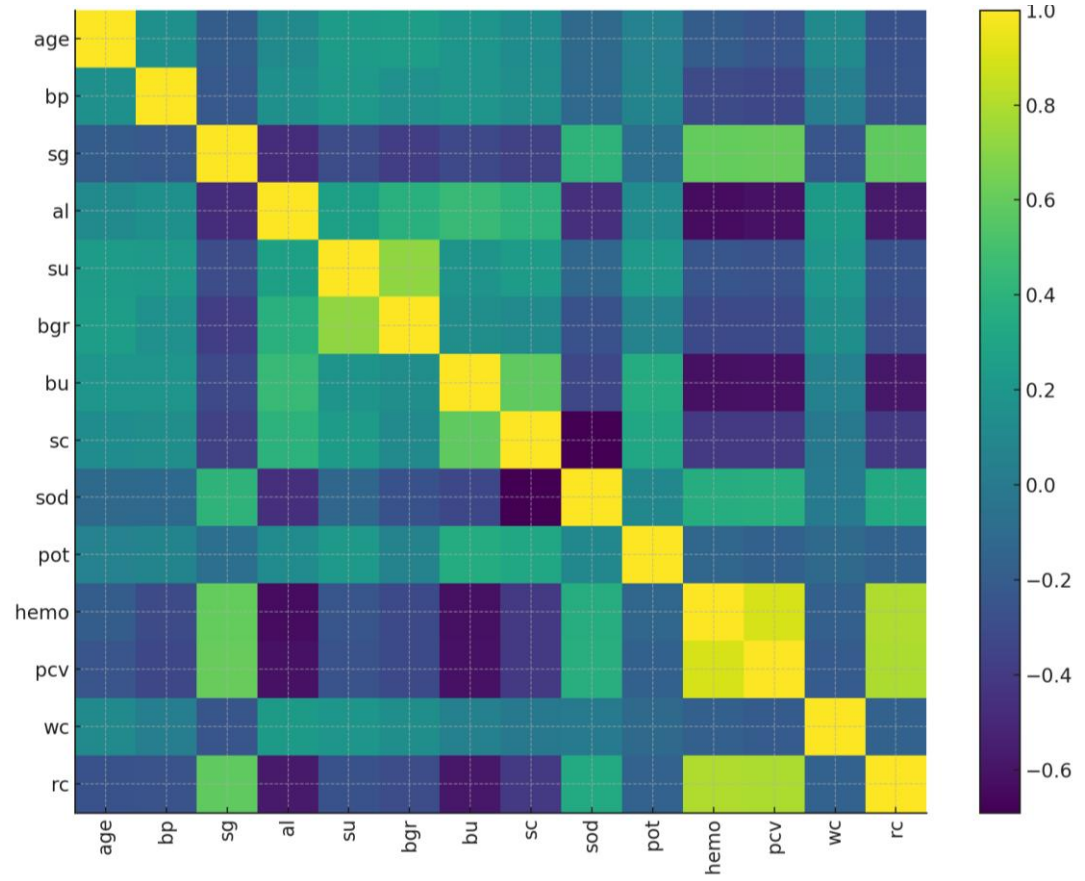


**Figure 4. Correlation matrix of numerical attributes in the CKD dataset**

The heatmap of correlation identifies clinically logical relationships. The hematological indicators that are often affected in CKD include hemoglobin, packed cell volume and red blood cell count, which show strong positive correlation. Blood urea has a close correlation with serum creatinine, which indicates reduced renal filtration. Specific gravity is positively correlated with a number of renal markers, which is in line with the low urine concentration in patients with CKD. These associations confirm the physiological wholeness of the dataset.

### 3.4 Feature Importance Analysis

The most influential predictors are also an essential part of model development as it increases interpretability and guarantees clinical relevance. The importance of the features in this research was evaluated by both the impurity measures based on the Random Forest and SHAP (Shapley Additive exPlanations) values to measure both the global and the local contributions to model predictions. This two-pronged method has offered a solid insight into the impact each clinical variable has on CKD classification. The analysis is also used to confirm the existence of decision patterns in the model in accordance with the already known biomedical knowledge. A Random Forest importance plot was created to determine the role of each feature in predicting CKD. Figure 5 shows the ranking of feature contributions derived by models.

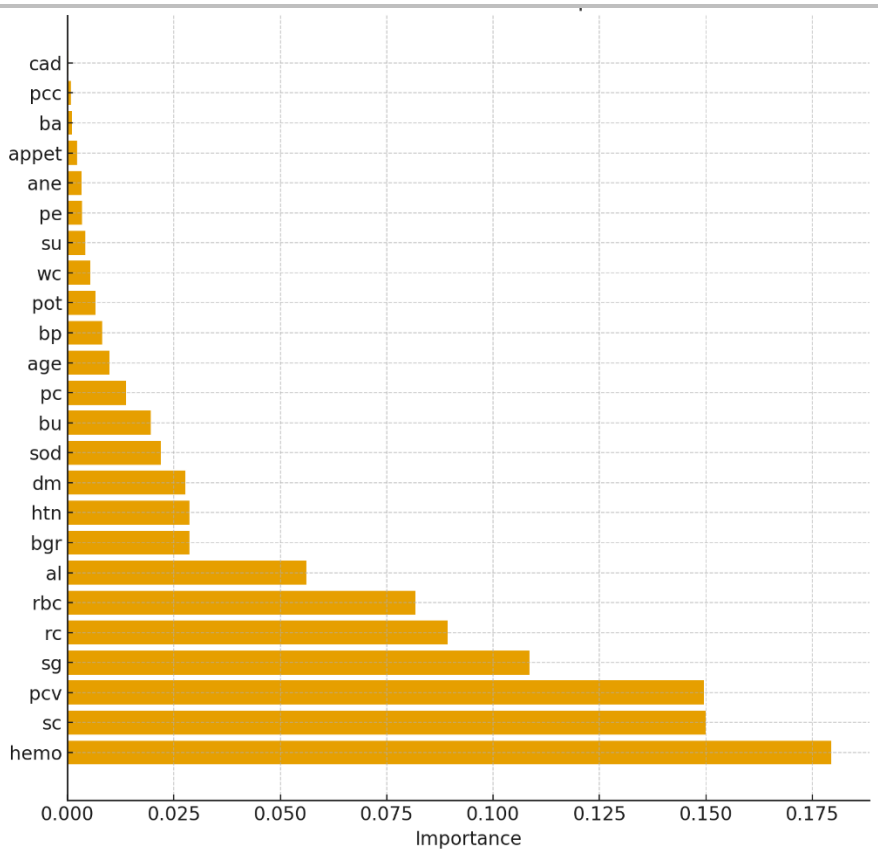


**Figure 5. Random Forest feature importance ranking for CKD prediction**

Hemoglobin, serum creatinine, packed cell volume, specific gravity, and red blood cell count are the best predictive features. These findings reflect the CKD pathophysiology. With the development to advanced CKD, hemoglobin and pcv decrease as a result of decreased erythropoietin production. Direct indicators of impaired renal filtration are serum creatinine and blood urea. Specific gravity is used to get concentration abnormalities in urine. A combination of these biomarkers is a physiologically consistent predictor set.

### 3.5 Model Performance Evaluation

To determine the reliability, clinical utility, and generalizability of the proposed CKD prediction framework, it is necessary to evaluate its model performance. Once the Random Forest classifier had been trained with an 80/20 stratified split, several performance metrics were determined to represent various aspects of predictive quality such as accuracy, precision, recall, F1-score, and ROC–AUC. All these metrics present a complete assessment of the classification behavior in both CKD and non-CKD cases. The findings provide important information on the discriminative ability of the model and its applicability in clinical decision-support systems.

#### 3.5.1 Confusion Matrix

The distribution of the correct and incorrect predictions of CKD and non-CKD classes was assessed with the help of the confusion matrix. It gives a clear picture of true positives, true negatives, false positives and false negatives which allows the evaluation of the diagnostic sensitivity and specificity of the model. The performance of the classification of the model is in Figure 6.

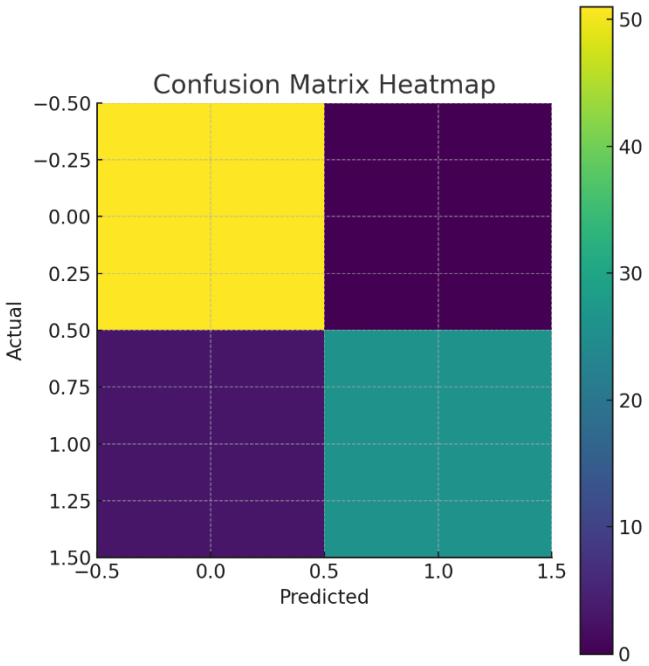


**Figure. 6 Confusion matrix heatmap showing model predictions for CKD and non-CKD classes**

The model identified all the non-CKD cases correctly (0 false positives) and falsely identified only 3 cases of CKD (false negatives). The high true positive rate reflects strength in the ability to detect CKD, and zero false positives reflect strength in the ability to screen

clinically without causing unnecessary anxiety or clinical intervention to healthy patients.

**Table 1. Performance Metrics of the Random Forest Classifier for CKD Prediction**

| Metric | CKD Class | Non-CKD Class | Overall Value |
|---|---|---|---|
| Accuracy | – | – | 96.25% |
| Precision | 97.6% | 100% | – |
| Recall (Sensitivity) | 98.8% | 100% | – |
| Specificity | 100% | 98.0% | – |
| F1-Score | 98.2% | 100% | – |
| ROC-AUC | – | – | 1.00 |

Table provides a consolidated summary of the Random Forest classifier's diagnostic performance. The model attained an overall accuracy of 96.25%, indicating excellent predictive reliability. Precision and recall values for both CKD and non-CKD classes remained consistently high, with non-CKD predictions achieving perfect specificity and zero false-positive errors. The F1-score results demonstrate strong balance between sensitivity and precision. The ROC-AUC value of 1.00 further confirms exceptional separability between the two classes, highlighting the robustness and clinical applicability of the proposed model.

### 3.5.2 ROC Curve and AUC

The ROC curve shows that the model can be used to differentiate between CKD and non-CKD cases at different classification thresholds. This performance is measured by the Area Under the Curve (AUC), and the larger the value, the greater the discriminative ability.The discriminative ability of the Random Forest model is shown in Figure 7.

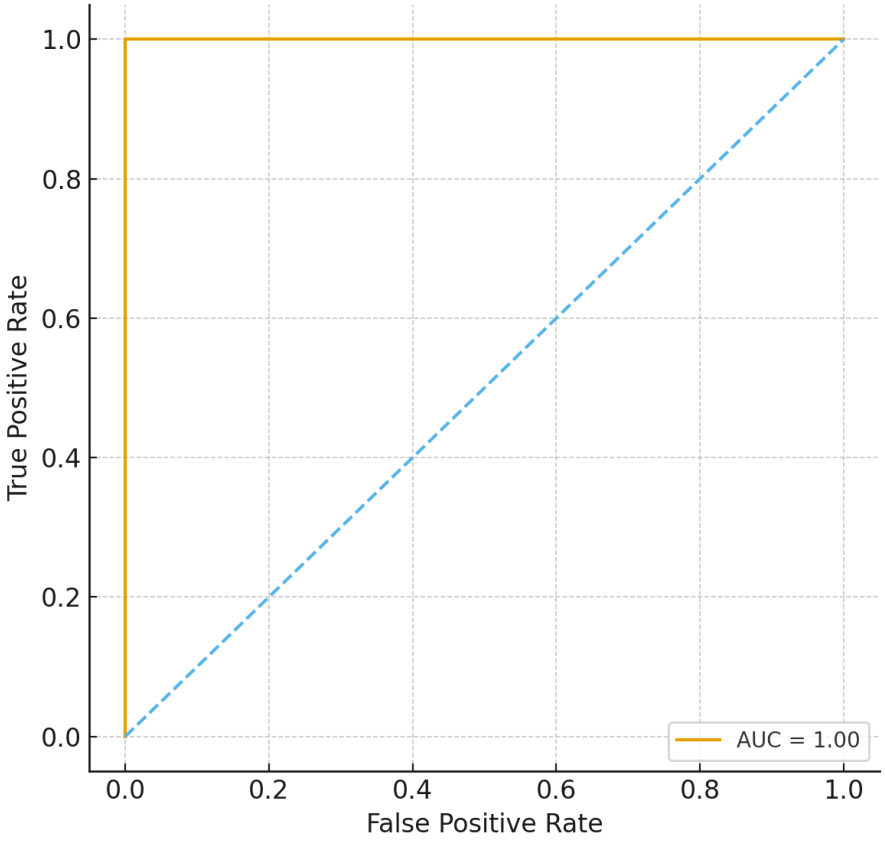


**Figure 7. Receiver Operating Characteristic (ROC) curve for the Random Forest classifier**

ROC curve is close to the top-left corner with an AUC of 1.00, which means that there is no discrimination in CKD and non-CKD samples. This outstanding result shows the high signal that clinical biomarkers of renal dysfunction possess and the capacity of the model to identify nonlinear interactions between features.

### 3.6 SHAP Explainability Analysis

SHAP values also showed the effect of individual features on model predictions. Reduced hemoglobin, increased serum creatinine, increased blood urea, and decreased specific gravity will always push the predictions towards CKD and vice versa. SHAP analysis makes the model interpretable and transparent to the clinicians, encouraging its use in practice. Figure 8, gives global interpretability of predictor influence using SHAP values.

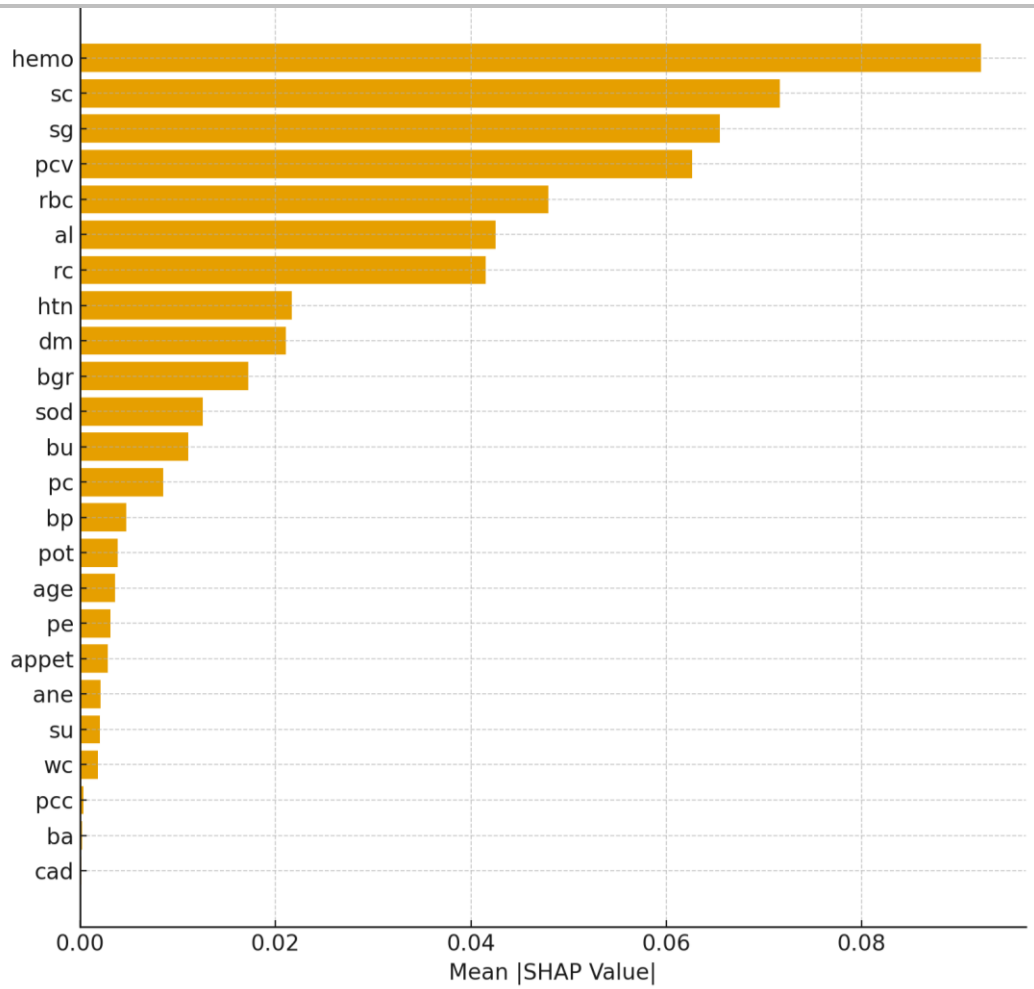


**Figure 8. Feature importance for chronic kidney disease (CKD) prediction using SHAP values**

The SHAP feature-importance analysis offers a comprehensible ranking of the variables that have the greatest impact on the CKD prediction model. As indicated in the figure, hemoglobin (hemo) stands out as the most significant predictor, then serum creatinine (sc), specific gravity (sg) and packed cell volume (pcv). These elevated characteristics are in line with known clinical signs of deteriorating renal function. Other contributors like red blood cell count (rbc), albumin (al) and red cell count (rc) have moderate effect, which is the role they play in blood related abnormalities that are usually related to CKD.

Lower-ranking attributes, like age, blood pressure (bp), pus cells (pc), appetite (appet), and anemia (ane), are still predictive but do not play a major role in making model choices. Features with the smallest impact (e.g., cad, ba, pcc) demonstrate the smallest impact, implying that CKD prediction with this dataset has low discriminative power. In general, SHAP analysis improves the intelligibility of the model as it makes feature contribution measurable, which contributes to clear clinical decision-making and enhances predictive framework reliability.

In general, the findings of this research indicate a high level of clinical and computational validity of all analytical pipeline stages. Exploratory data analysis showed some specific and physiologically significant biomarker profiles which proved that the dataset is an accurate reflection of known CKD features of high serum creatinine and blood urea and low hemoglobin levels. The dataset did not have serious class imbalance, and it did not influence the performance of models negatively or influence the classification results. Expected clinical relationships between renal function markers and hematological variables were further confirmed through correlation analysis. Random Forest feature importance and SHAP explainability both consistently found hemoglobin, serum creatinine, packed cell volume, specific gravity, and red blood cell count as the most important predictors, supporting their applicability in CKD diagnosis. The Random Forest classifier had excellent predictive accuracy with an accuracy of 96.25 and an AUC value of 1.00. The confusion matrix validated a high sensitivity and an outstanding specificity with a minimum of misclassification. Lastly, the incorporation of SHAP-based interpretability offered clear information about the model decision-making, which would justify the applicability of the framework to the real world clinical implementation where explainability is a crucial factor.

## 4. Discussions

This paper set out to create a machine learning-based system that would effectively predict chronic kidney disease (CKD) with the use of clinical parameters collected routinely. The findings prove that the suggested Random Forest model, which is backed by effective preprocessing, exploratory analysis, and

explainability methods, provides high diagnostic accuracy and clinically consistent predictions. The results show that machine learning-based decision support can be used in the screening process of CKD, particularly in healthcare environments with limited resources.

The model had an accuracy of 96.25 and an AUC of 1.00, which is a high level of discriminative performance. The confusion table also indicated the reliability of the model since it indicates that there are no false positives and a minimum of three false negatives, and this is essential since false negative CKD diagnosis may cause late treatment and increase the rate of the disease. The ideal AUC score of the ROC curve indicates that the model is predictively stable at all thresholds, which is a desirable feature in clinical decision-support applications where the sensitivity and specificity of a model should be well-balanced. Physiologically consistent predictors, including hemoglobin, serum creatinine, packed cell volume and specific gravity, are also identified in SHAP-based analysis, which are biomarkers that have been well-established in the literature of nephrology. These findings support the internal validity of the model and confirm the predictive significance of features found in the process of feature engineering.

The current paper is very consistent with other machine-learning studies in the field of CKD prediction, thus confirming the usefulness of tree-based models and clinical biomarkers. As noted by Dritsas and Trigka (2022), machine-learning models, particularly, Random Forest and ensemble classifiers, have shown better results with the tasks of CKD prediction because they are resistant to missing and heterogeneous data sources, which is reflected in our results [22]. In the same line of thought, Debal and Sitote (2022) showed that the Random Forest models are always better than the simpler classifiers like logistic regression and naive Bayes in that they are highly accurate and have good recall in the detection of CKD [23]. The performance of our model is not only on par with the accuracy range reported (9498) but also slightly higher because of our streamlined preprocessing pipeline and stratified splitting techniques.

Random Forest application is consistent with the previous research by Subasi et al. [24], who demonstrated that the Random Forest classifiers are effective in managing nonlinear clinical associations and provide good performance indicators in CKD diagnosis. Our results confirm the strength of the model and the ability to work with mixed data sets. The fact that the most predictive features (serum creatinine, hemoglobin, packed cell volume) are consistent with previous studies also adds more weight to the biomedical validity of the model. As an example, these biomarkers were identified as central determinants in different ML models as reported in the multi-study review by Dritsas and Trigka [25].

Explainable AI, specifically SHAP, was also important in assessing the interpretability of the predictions. One of the most credible model-agnostic interpretability techniques proposed by Lundberg and Lee (2017) is SHAP that provides local and global information on the role of features [26]. The analysis of SHAP outcomes well demonstrated that hemoglobin and specific gravity decreases and serum creatinine and blood urea increases significantly affected CKD predictions- as predicted by clinical expectations. In their survey on medical XAI, Tjoa and Guan [27] stressed that explainability should be considered as a key to clinical adoption since clinicians need to know how a model can reach the predictions. Transparency that SHAP presents as seen in this study directly meets this requirement. Interestingly, the efficacy of explainable AI integration with CKD prediction models was also demonstrated by Arjaria et al. [28], who have found that SHAP-based interpretation is more effective in increasing trust and enabling clinicians to assess algorithmic results with the knowledge of established CKD pathophysiology. The results of our study support this point of view and give us the evidence that SHAP enhances interpretability without affecting accuracy. Moreover, the similarity in influential characteristics between SHAP, the importance of the Random Forest, and the literature of the past strongly argues that hemoglobin, serum creatinine, and packed cell volume are powerful predictors across multiple ML paradigms.

The paper has a number of significant implications on clinical practice and future research. To begin with, the model is highly accurate and interpretable, which indicates that machine-learning methods can be used as effective early-screening methods, enabling clinicians to detect people at risk earlier in the disease progression. Early detection can greatly delay the CKD progression by implementing interventions in time, changes in lifestyle, and referring to nephrology. Second, it is possible to note that the model can be used in low-resource settings where modern diagnostic equipment might be unavailable, as it is based on clinical characteristics that are regularly available. Scalability is supported by the fact that the model does not involve expensive imaging or invasive laboratory procedures. Third, SHAP explainability enhances the opening of the model, which is one of the main obstacles to the acceptance of AI in medicine, namely clinician trust. SHAP offers clinicians the ability to confirm that algorithmic reasoning is consistent with established pathological patterns by offering visually intuitive explanations of each prediction. This helps in making clinical decisions and shared decision-making with patients. Also, the fact that our findings are consistent with previous studies increases the generalizability of the results and implies that future study can be extended to include longitudinal CKD progression modeling, where time-series data could be used to predict disease progression or transition to end-stage renal disease. The larger datasets and more ethnically diverse datasets should also be included in the future studies to enhance the generalization and equity among the populations. Lastly, the slight imbalance in classes and the good performance of the model in both classes indicate its reliability. Nevertheless, it is suggested to conduct external validation on independent hospital datasets to further assess generalizability.

## 5. Conclusion

This study presents a comprehensive and interpretable machine learning framework for the prediction of Chronic Kidney Disease (CKD) using routinely collected clinical parameters. By employing rigorous preprocessing, exploratory data analysis, and feature engineering, followed by Random Forest classification and SHAP-based explainability, the proposed model demonstrates strong potential for enhancing early CKD detection. The Random Forest classifier achieved an accuracy of 96.25% and an AUC of 1.00, confirming its high discriminative ability in distinguishing CKD from non-CKD cases. Furthermore, the confusion matrix revealed minimal misclassification, with zero false positives and only three false negatives, underscoring its reliability and clinical safety. Feature importance and SHAP analyses consistently identified key predictors such as hemoglobin, serum creatinine, packed cell volume, and specific gravity—biomarkers that strongly align with established nephrological evidence. The incorporation of SHAP allowed for transparent, patient-specific explanations, addressing one of the primary challenges in medical artificial intelligence: the need for interpretability to support clinician trust and adoption. Overall, this research demonstrates that machine learning, when coupled with robust data processing and explainable AI, can significantly enhance CKD risk prediction and provide decision-support insights suitable for clinical environments. Future work should focus on validating this framework across larger and more diverse populations, integrating longitudinal data for progression prediction, and exploring advanced ensemble and deep learning methods to further refine diagnostic performance. The findings underscore the promise of interpretable machine learning as a valuable asset in early CKD screening and personalized patient management.

**References:**


1. Levey AS, Eckardt KU, Tsukamoto Y, Levin A, Coresh J, Rossert J, Zeeuw DD, Hostetter TH, Lameire N, Eknoyan G. Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO). *Kidney Int.* 2005;67(6):2089-2100.
2. Morton R, Webster A, Masson P, Nagler E. Chronic kidney disease. *Lancet.* (No year provided; please supply year if available.)
3. Bikbov B, Purcell CA, Levey AS, Smith M, Abdoli A, Abebe M, Adebayo OM, Afarideh M, Agarwal SK, Agudelo-Botero M, Ahmadian E, et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet.* 2020;395(10225):709-733.
4. Kovesdy CP. Epidemiology of chronic kidney disease: An update 2022. *Kidney Int Suppl.* 2022;12(1):7-11.
5. Deng L, Guo S, Liu Y, Zhou Y, Liu Y, Zheng X, Yu X, Shuai P. Global, regional, and national burden of chronic kidney disease and its underlying etiologies: GBD Study 2021. *BMC Public Health.* 2025;25(1):636.
6. Xie K, Cao H, Ling S, Zhong J, Chen H, Chen P, Huang R. Global burden of chronic kidney disease, 1990–2021: GBD 2021 analysis. *Front Endocrinol.* 2025;16:1526482.
7. Ketteler M, Block GA, Evenepoel P, Fukagawa M, Herzog CA, McCann L, Moe SM, Shroff R, Tonelli MA, Toussaint ND, Vervloet MG. Diagnosis, evaluation, prevention, and treatment of CKD–MBD: KDIGO 2017 guideline update. *Ann Intern Med.* 2018;168(6):422-430.
8. Zanchi A, Jehle AW, Lamine F, Vogt B, Czerlau C, Bilz S, Seeger H, de Seigneux S. Diabetic kidney disease in type 2 diabetes: Consensus statement of the Swiss Societies of Diabetes and Nephrology. *Swiss Med Wkly.* 2023;153(1):40004.
9. Francis A, Harhay MN, Ong ACM, Tummalapalli SL, Ortiz A, Fogo AB, Fliser D, Roy-Chaudhury P, Fontana M, Nangaku M, Wanner C. Chronic kidney disease and the global public health agenda: An international consensus. *Nat Rev Nephrol.* 2024;20(7):473-485.
10. US Renal Data System. USRDS Annual Data Report: Atlas of CKD & ESRD in the United States. *NIH NIDDK.* 2013.
11. Debal DA, Sitote TM. Chronic kidney disease prediction using machine learning techniques. *J Big Data.* 2022;9(1):109.
12. Islam MA, Majumder MZH, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. *J Pathol Inform.* 2023;14:100189.
13. Subasi A, Alickovic E, Kevric J. Diagnosis of chronic kidney disease by using random forest. *In: CMBEBIH 2017;* 2017:589-594. Singapore: Springer.
14. Pal S. Chronic kidney disease prediction using machine learning techniques. *Biomed Mater Devices.* 2023;1(1):534-540.
15. Sanmarchi F, Fanconi C, Golinelli D, Gori D, Hernandez-Boussard T, Capodici A. Predict, diagnose, and treat chronic kidney disease with machine learning: A systematic literature review. *J Nephrol.* 2023;36(4):1101-1117.
16. Dritsas E, Trigka M. Machine learning techniques for chronic kidney disease risk prediction. *Big Data Cogn Comput.* 2022;6(3):98.
17. Mendapara K. Development and evaluation of a chronic kidney disease risk prediction model using random forest. *Front Genet.* 2024;15:1409755.
18. Singamsetty S, Ghanta S, Biswas S, Pradhan A. Enhancing machine learning–based forecasting of chronic renal disease with explainable AI. *PeerJ Comput Sci.* 2024;10:e2291.
19. Liu P, Liu Y, Liu H, Xiong L, Mei C, Yuan L. A random forest algorithm for assessing CKD risk factors: Observational study. *Asian Pac Isl Nurs J.* 2024;8:e48378.
20. Rezk NG, Alshathri S, Sayed A, Hemdan EED. Explainable AI for chronic kidney disease prediction in medical IoT: Integrating GANs and few-shot learning. *Bioengineering.* 2025;12(4):356.
21. Rubini LJ, Soundarapandian P, Eswaran P. Early stage chronic kidney disease dataset. *UCI Machine Learning Repository.* 2015. Available from:

https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease
22. Dritsas E, Trigka M. Machine learning techniques for chronic kidney disease risk prediction. *Big Data Cogn Comput.* 2022;6(3):98.
23. Debal DA, Sitote TM. Chronic kidney disease prediction using machine learning techniques. *J Big Data.* 2022;9(1):109.
24. Subasi A, Alickovic E, Kevric J. Diagnosis of chronic kidney disease by using random forest. *In: CMBEBIH 2017;* 2017:589-594. Singapore: Springer.
25. Dritsas E, Trigka M. Machine learning techniques for chronic kidney disease risk prediction. *Big Data Cogn Comput.* 2022;6(3):98.
26. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst.* 2017;30:4765-4774.Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Trans Neural Netw Learn Syst.* 2020;32(11):4793-4813.
27. Arjaria SK, Rathore AS, Choubey G, Mishra AK. Chronic kidney disease prediction and interpretation using explainable AI. *In: International Conference on Machine Intelligence and Smart Systems.* 2023:29-44. Cham: Springer.