Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Explainable Machine Learning for Early Chronic Kidney Disease Prediction Using SHAP-Guided Feature Selection

Yousra Rehman
Department of Computer Science, University of Haripur, Pakistan

Sohair Sultan Fayyaz Jeelani (Corresponding Author)
Department Of Computer Science, Abbottabad University of Science and Technology, Abbottabad, Pakistan. Email: sohair.sultan2266@gmail.com

Arooba Akhtar (Corresponding Author)
Department Of Computer Science and Information Technology, Hazara University, Dhodial, Mansehra, Pakistan. arooba_csit@hu.edu.pk

Hadiya Ali
Department Of Computer Science, Abbottabad University of Science and Technology, Abbottabad, Pakistan

Kinza Iqbal
Department Of Computer Science, Abbottabad University of Science and Technology, Abbottabad, Pakistan

ABSTRACT

Chronic kidney disease (CKD) is a progressive and often asymptomatic condition whose early detection is critical for preventing irreversible renal failure. Machine learning offers accurate automated screening, but the black-box nature of many models limits clinical trust and adoption. This paper presents an explainable machine learning framework for early CKD prediction that combines high predictive accuracy with transparent, clinically interpretable explanations. Four classifiers, namely Logistic Regression, Random Forest, XGBoost, and LightGBM, are trained and evaluated on the UCI Chronic Kidney Disease dataset of 399 patient records described by 24 clinical attributes. SHapley Additive exPlanations (SHAP) are then applied to quantify each feature's contribution to model predictions, both globally across the cohort and locally for individual patients. A SHAP-guided feature selection procedure is proposed that ranks features by mean absolute SHAP value and retrains the models on a compact subset. The Random Forest classifier achieves 100 percent test accuracy with a perfect area under the curve of 1.000, while all four models exceed 98 percent accuracy under five-fold cross-validation. The SHAP analysis identifies haemoglobin, specific gravity, packed cell volume, and serum creatinine as the most influential predictors, in agreement with established clinical knowledge. Using only the top eight SHAP-ranked features, all models reach 100 percent accuracy, reducing the feature count by 67

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

percent without any loss of performance.
Keywords—Chronic Kidney Disease, Explainable AI, SHAP, Feature Selection, Machine Learning, Random Forest, XG Boost, Clinical Decision Support

Introduction

Chronic kidney disease (CKD) is a global public health burden affecting an estimated 10 to 15 percent of the world adult population and is projected to become the fifth leading cause of death by 2040 [1], [2]. The condition is characterised by a gradual and irreversible loss of kidney function over months or years, and it frequently progresses undetected because early stages are largely asymptomatic [3]. By the time overt clinical signs appear, many patients have already sustained substantial and permanent renal damage that necessitates dialysis or kidney transplantation, both of which carry heavy financial and quality-of-life costs [4]. Early and accurate detection of CKD is therefore of paramount importance, as timely intervention through lifestyle modification and pharmacological management can slow or halt disease progression [5].

Traditional CKD diagnosis relies on laboratory measurement of the estimated glomerular filtration rate and the albumin-to-creatinine ratio, interpreted by trained nephrologists [6]. While reliable, this process depends on specialist availability and is not readily scalable to large-scale population screening, particularly in low-resource settings where the burden of CKD is rising most rapidly [7]. Machine learning has emerged as a powerful tool for automated disease screening, capable of learning complex non-linear relationships between routinely collected clinical variables and disease status [8], [9]. Numerous studies have demonstrated that machine learning models can predict CKD from standard blood and urine test results with high accuracy, offering the prospect of low-cost automated screening that complements specialist assessment [10], [11].

Despite these advances, a significant barrier to the clinical adoption of machine learning models is their lack of transparency [12]. Many of the most accurate models, including ensemble methods and deep neural networks, operate as black boxes whose internal decision logic is opaque to clinicians [13]. In high-stakes medical applications, a prediction alone is insufficient; clinicians require an understanding of why a model reached a particular conclusion in order to trust it, to validate it against domain knowledge, and to satisfy legal and ethical accountability requirements [14], [15]. This need has driven the growth of explainable artificial intelligence (XAI), which aims to make model predictions interpretable without sacrificing accuracy [16].

Among XAI methods, SHapley Additive exPlanations (SHAP) has become the de facto standard for post-hoc model interpretation [17]. Grounded in cooperative game theory, SHAP assigns each feature an additive contribution to every individual prediction, providing both a global view of feature importance across a cohort and a local explanation for any single patient [18]. SHAP satisfies desirable theoretical properties of local accuracy, missingness, and consistency that distinguish it from earlier heuristic importance measures [17]. Applying SHAP to CKD prediction offers the potential to reconcile the competing demands of predictive accuracy and clinical interpretability.

This paper presents an explainable machine learning framework for early CKD prediction that unites accurate classification with transparent SHAP-based explanation

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

and a novel SHAP-guided feature selection procedure. The specific contributions are as follows. First, four representative classifiers spanning linear, bagging, and boosting paradigms are trained and rigorously compared on the UCI Chronic Kidney Disease dataset using multiple evaluation metrics and five-fold cross-validation. Second, SHAP is applied to the best model to produce global and local explanations, and the resulting feature importances are validated against established clinical knowledge of CKD. Third, a SHAP-guided feature selection method is proposed that ranks features by their mean absolute SHAP value and retrains the models on a compact subset, demonstrating that a substantially smaller and more interpretable feature set retains full predictive performance. This reduction is clinically valuable because it lowers the number of laboratory tests required for screening, reducing both cost and patient burden.

The remainder of this paper is organised as follows. Section 2 reviews related work on machine learning for CKD prediction and explainable artificial intelligence in healthcare. Section 3 describes the dataset, preprocessing, models, and the proposed SHAP-guided methodology. Section 4 presents and discusses the experimental results. Section 5 concludes the paper and outlines directions for future work.

Literature Review

This section reviews the literature across three themes most relevant to the present work: machine learning for chronic kidney disease prediction, feature selection in clinical machine learning, and explainable artificial intelligence in healthcare.

Machine Learning for CKD Prediction

In the last decade, there has been extensive research on the use of machine learning techniques for predicting CKD. Studies have shown that, using basic clinical features, classical classifiers can separate CKD patients from healthy patients in early studies. Salekin and Stankovic [10] tested various classifiers on the UCI CKD data set and found that Random Forest (RF) with wrapper-based feature selection produced high detection accuracy, indicating that serum creatinine and haemoglobin were the predictive features. Almasoud and Ward [11] examined the performance of logistic Regression, support vector machines, and gradient boosting for detecting CKD and concluded that ensemble methods consistently outperformed linear methods, which they attributed to non-linear relationships among clinical features.

Subsequent research has focused on ensemble learning. Qin et al. [19] presented an integrated model of logistic Regression and Random Forest using a perceptron, achieving diagnostic accuracy above 99% on a UCI dataset and demonstrating the advantages of model fusion. A detailed comparison of seven classifiers, with and without feature selection and synthetic minority oversampling, was carried out by Chittora et al. [20], whose results indicate that gradient-boosted trees and support vector machines outperform the others. Since then, gradient boosting frameworks like XGBoost [21] and LightGBM [22] have proven popular for tabular clinical data due to their accuracy, efficiency, and native support for heterogeneous features.

In addition to the UCI study, larger investigations have also validated the clinical application of machine learning in CKD. Ravizza et al. [23] built models that predicted the progression of CKD using EHR data from diabetic patients and showed that

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

routinely collected data is sufficient to accurately stratify the risk of CKD progression. Song et al. [24] developed longitudinal models to predict the onset of CKD and highlighted the importance of temporal features. Together, these studies demonstrate the high predictive ability of machine learning for CKD; however, most have focused only on accuracy measures and offer little added insight into the clinical interpretability of the models.

Feature Selection in Clinical Machine Learning

In clinical machine learning, feature selection is a key component for three main reasons: to reduce the complexity of the model, to prevent overfitting, to collect less data, and to have a more interpretable model [25] In clinical datasets, there are usually redundant and undistributive attributes and it is useful to find a small, highly informative subset of attributes for computational convenience and for the ease of deployment, where each extra laboratory test adds to the cost and patient burden [26].

Feature selection techniques can typically be classified into filter, wrapper, and embedded methods [25]. Filter methods are statistical feature-ranking techniques that are independent of classifiers. Wrapper methods look for feature subsets that yield the best performance for a given classifier, which are more accurate but more expensive. Embedded methods, such as the impurity-based importance of tree ensembles or the coefficients of regularised linear models, select features during model training. Wrapper and embedded methods have been extensively used in the CKD literature but tend to rank features based on their contribution to accuracy rather than their explanatory contribution to individual predictions.

One more recent approach is to use model explanations to inform feature selection. SHAP values are a measure of a feature's contribution to the model's output. They can also be summed across a dataset to obtain a global importance score that aligns with the model's decision-making process [17, 27]. Feature selection based on SHAP importance thus ensures that the selected features correspond to those on which the model actually relies, in contrast to filter approaches (which do not take the model into account) or purity measures (which may be biased towards higher-cardinality features) [28]. The approach presented here follows the SHAP-based approach, and it is shown that predictive accuracy was maintained with a significantly smaller set of features.

Explainable AI in Healthcare

With the rapid advancement of medical machine learning systems towards and surpassing human expertise on numerous tasks [13,14], the need for explainability has grown stronger. There is a growing demand for meaningful explanations for automated decisions that impact patients, within regulatory frameworks, and clinicians do not want to take action on predictions that do not make sense to them [15]. Rudin [12] made an impassioned plea to use inherently interpretable models in preference to black-box models where possible and to seek explanations of the latter after they are built where they are essential for high-stakes decisions.

Two types of post-hoc explanation methods are prevalent today. LIME [16] is a method that explains a single prediction by creating a local linear surrogate model around the instance of interest. SHAP [17] combines several existing techniques based on the

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

theory of Shapley values to provide both local and global explanations and has become the most popular method in clinical use. Instead, Lundberg et al. [27] developed an exact and efficient implementation of SHAP to make it applicable to tree ensembles, including Random Forests and gradient boosting models, on tabular medical data.
SHAP has been used successfully for various clinical prediction problems such as the prediction of mortality [27] and sepsis onset [29] and the risk of diabetes [29]. Within the field of kidney disease, there are increasing efforts to use SHAP to interpret the predictions of CKD models [30], and the study has demonstrated that the contributions of the various features (e.g., serum creatinine, haemoglobin, and albumin) make sense from a clinical perspective. But there are few studies that combine SHAP explanations with feature selection to create a minimal, clinically deployable screening model. The present work fills this gap by applying SHAP not only for model interpretation but also for identifying a minimal subset of features with full accuracy; thus combining accuracy, interpretability, and practicality into a single framework.

Proposed Methodology

The proposed framework consists of five stages: data acquisition, preprocessing, model training, SHAP-based explanation, and SHAP-guided feature selection. Fig. 1 presents the overall architecture as a flowchart. Each stage is described in the following subsections.

Stage 1: Data Acquisition
UCI CKD dataset (399 patients, 24 features)
Binary target: ckd / notckd

↓

Stage 2: Preprocessing
Median/mode imputation, label encoding,
target encoding, standardisation

↓

Stage 3: Model Training
Logistic Regression, Random Forest,
XGBoost, LightGBM + 5-fold CV

↓

Stage 4: SHAP Explanation
Global importance + local (per-patient)
Shapley value attribution

↓

Stage 5: SHAP-Guided Feature Selection
Rank by mean |SHAP|, retrain on top-K,
evaluate accuracy retention

↓

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Output
Accurate + interpretable CKD prediction
with compact feature set

Fig. 1. Architecture of the proposed explainable machine learning framework for CKD prediction. The pipeline proceeds from data acquisition through preprocessing, model training, SHAP explanation, and SHAP-guided feature selection to an accurate and interpretable output.

Dataset Description

The study uses the publicly available Chronic Kidney Disease dataset from the UCI Machine Learning Repository [31]. It contains 399 patient records, each described by 24 clinical attributes and a binary class label indicating the presence (ckd) or absence (notckd) of chronic kidney disease. The attributes comprise 11 numeric measurements, including age, blood pressure, blood glucose, serum creatinine, haemoglobin, and packed cell volume, and 13 categorical or nominal indicators such as red blood cell status, hypertension, and diabetes mellitus. After preprocessing, the cohort comprises 250 CKD cases and 149 non-CKD cases. Table I summarises the dataset characteristics.

TABLE I. *Summary of the UCI Chronic Kidney Disease dataset.*

Property	**Value**
Total patient records	399
Numeric features	11
Categorical features	13
Total features used	24
CKD (positive) cases	**250**
Non-CKD (negative) cases	149
Target type	Binary (ckd / notckd)

Fig. 2. Dataset overview. Left: class distribution showing 250 CKD and 149 non-CKD patients. Right: correlation heatmap of the numeric clinical features, revealing moderate inter-feature correlations consistent with shared physiological pathways.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Data Preprocessing

Raw clinical data contains missing values and inconsistently formatted entries that must be addressed before model training. The preprocessing pipeline performs the following steps. The identifier column is removed. Text fields are stripped of extraneous whitespace and tab characters and normalised to lowercase. Numeric columns stored as strings are converted to numeric type. Missing values in numeric features are imputed using the median, which is robust to outliers, while missing values in categorical features are imputed using the mode. Categorical features are then label-encoded to integer representations, and the target label is encoded as ckd equal to 1 and notckd equal to 0. Finally, features are standardised to zero mean and unit variance for the Logistic Regression classifier, which is sensitive to feature scale, while the tree-based models operate on the unscaled features. The standardisation transform is defined as

$$z = (x - \mu) / \sigma \qquad (1)$$

where x is the original feature value, μ is the feature mean, and σ is the feature standard deviation, both estimated from the training set only to prevent data leakage.

Classification Models

Four classifiers spanning three learning paradigms are evaluated. Logistic Regression is a linear model that estimates the probability of the positive class through the logistic function

$$P(y = 1 \mid x) = 1 / (1 + \exp(-(w^T x + b))) \qquad (2)$$

where w is the weight vector and b is the bias term. Random Forest [32] is a bagging ensemble that aggregates the predictions of many decision trees, each trained on a bootstrap sample with random feature subsets, reducing variance through averaging. XGBoost [21] and LightGBM [22] are gradient boosting frameworks that build an additive ensemble of trees in a stage-wise manner, each new tree fitting the residual errors of the current ensemble by minimising a regularised objective

$$L = \Sigma\, l(y_i, \hat{y}_i) + \Sigma\, \Omega(f_k) \qquad (3)$$

where l is a differentiable loss function, $\hat{y}_i$ is the predicted value, and Ω penalises model complexity to control overfitting. The hyperparameters of all models are summarised in Table II.

TABLE II. *Model hyperparameter configuration.*

Model	Key Hyperparameters	Value
Logistic Regression	max_iter	1000
Random Forest	n_estimators	200
XGBoost	n_estimators / max_depth / lr	200 / 4 / 0.1
LightGBM	n_estimators / max_depth / lr	200 / 4 / 0.1
All	random_state	42
Evaluation	test split / CV folds	20% / 5

SHAP Explanation

SHapley Additive exPlanations attribute a model prediction to its input features using the Shapley value from cooperative game theory [17]. For a given prediction, the

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Shapley value of feature i is its average marginal contribution across all possible feature coalitions

$$\varphi_i = \Sigma\, [\, |S|!(M-|S|-1)! / M!\,] \cdot [\, f(S \cup \{i\}) - f(S)\,] \quad (4)$$

where M is the total number of features, S is a subset of features not containing i, and f(S) is the model output using only the features in S. The prediction is then expressed as an additive sum of feature contributions

$$f(x) = \varphi_0 + \Sigma\, \varphi_i \quad (5)$$

where φ_0 is the base value (the mean prediction over the dataset). The global importance of feature i is obtained by averaging the absolute SHAP values across all N samples

$$I_i = (1 / N)\, \Sigma\, |\varphi_i(x_j)| \quad (6)$$

The TreeSHAP algorithm [27] computes these values exactly and efficiently for tree ensembles, making SHAP practical for the Random Forest and gradient boosting models used in this study.

SHAP-Guided Feature Selection

The proposed feature selection procedure uses the global SHAP importances of Equation (6) to rank features and select a compact subset. Algorithm 1 formalises the procedure. The features are ranked in descending order of mean absolute SHAP value, the top K features are retained, and all models are retrained and evaluated on this reduced subset. If accuracy is preserved, the reduced model offers equivalent performance with fewer required clinical measurements.

Algorithm 1: *SHAP-Guided Feature Selection*

```
Input: training set (X_train, y_train), test set (X_test, y_test),
       trained model f, target subset size K
Output: reduced feature set F_K and retrained models
1:  compute SHAP values φ for f on X_test using TreeSHAP
2:  for each feature i do
3:      I_i ← (1/N) Σ_j |φ_i(x_j)|          // mean abs SHAP
4:  rank features by I_i in descending order
5:  F_K ← top K features by rank
6:  for each model m in {LR, RF, XGB, LGBM} do
7:      retrain m on X_train[F_K]
8:      evaluate m on X_test[F_K]
9:  return F_K and retrained models
```

Evaluation Metrics

Model performance is assessed using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). These are defined in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (7)$$

$$Precision = TP / (TP + FP), \quad Recall = TP / (TP + FN) \quad (8)$$

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (9)$$

Five-fold stratified cross-validation is additionally performed to assess the stability and

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

generalisation of each model beyond a single train-test split.

Results and Discussion

This section presents the classification results, the SHAP explainability analysis, and the outcome of the proposed SHAP-guided feature selection. All experiments use a stratified 80 to 20 train-test split with a fixed random seed for reproducibility.

Classification Performance

Table III reports the performance of the four classifiers on the full 24-feature set. All models achieve excellent results, with accuracies ranging from 98.75 percent to 100 percent and a perfect AUC of 1.000 across the board. The Random Forest classifier attains a perfect test accuracy of 100 percent with an F1-score of 1.000, while Logistic Regression, XGBoost, and LightGBM each reach 98.75 percent accuracy. The five-fold cross-validation accuracies, also reported in Table III, remain consistently high between 98.75 percent and 99.06 percent with low standard deviations, confirming that the strong performance is stable and not an artefact of a favourable single split.

TABLE III. *Classification performance on the full 24-feature set. CV denotes five-fold cross-validation accuracy.*

Model	Acc (%)	Prec	Rec	F1	AUC / CV (%)
Logistic Regression	98.75	1.000	0.980	0.990	1.000 / 99.06
Random Forest	**100.00**	**1.000**	**1.000**	**1.000**	**1.000 / 99.06**
XGBoost	98.75	1.000	0.980	0.990	1.000 / 98.75
LightGBM	98.75	1.000	0.980	0.990	1.000 / 98.75

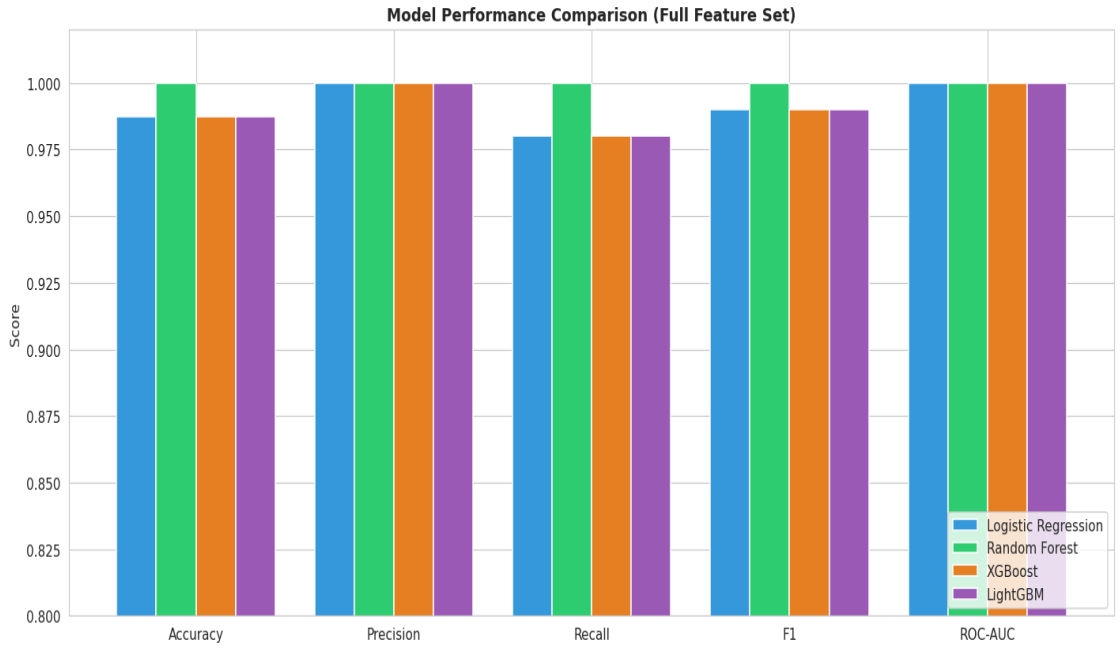


Fig. 3. Performance comparison of the four classifiers across accuracy, precision, recall, F1-score, and AUC on the full feature set. All models achieve near-perfect

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

scores, with Random Forest reaching 100 percent accuracy.

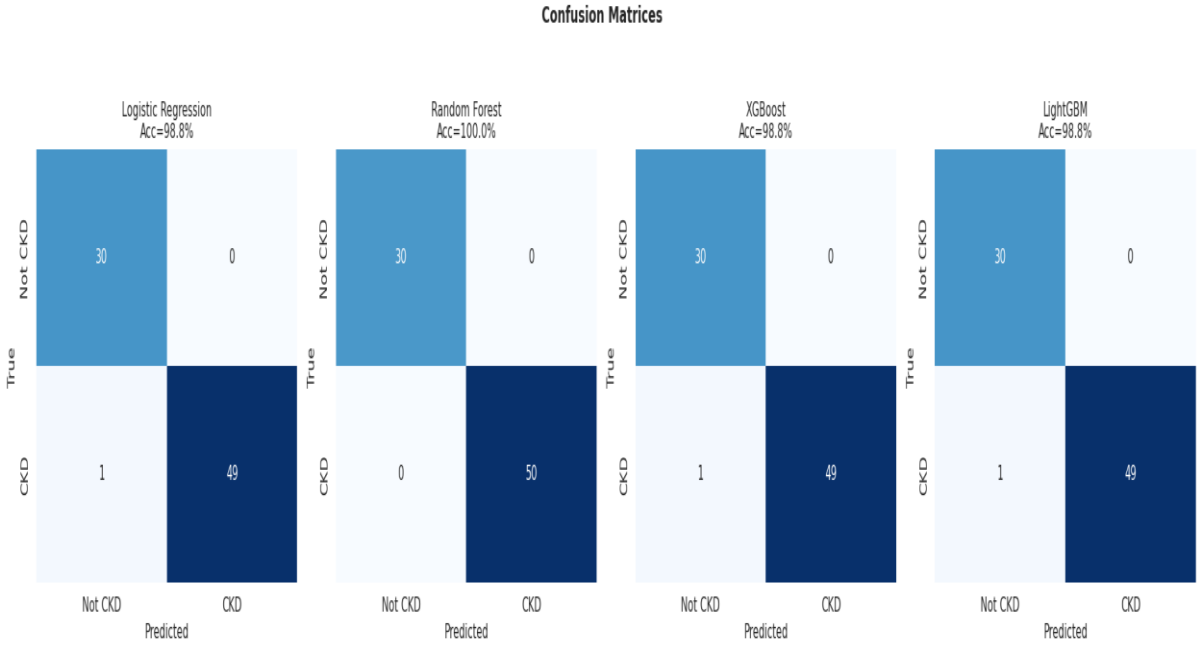


Fig. 4. Confusion matrices for the four classifiers on the test set of 80 patients. Random Forest produces no misclassifications, while the other three models each misclassify a single non-CKD patient.

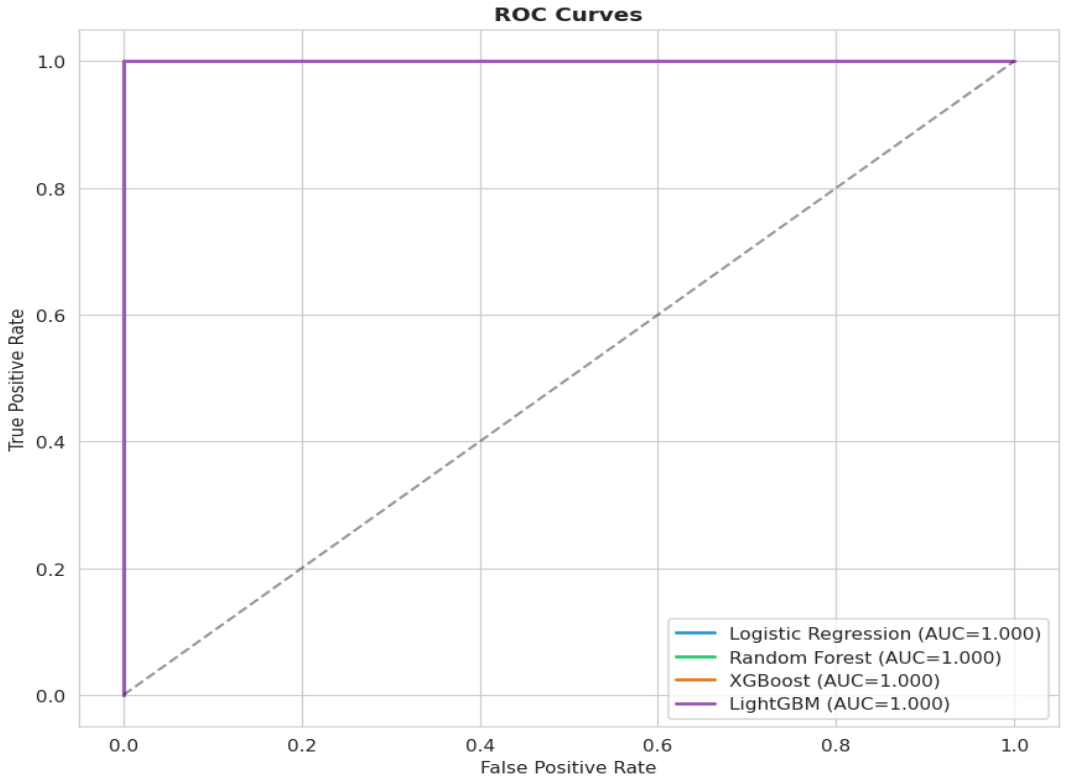


Fig. 5. Receiver operating characteristic curves for the four classifiers. All models

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

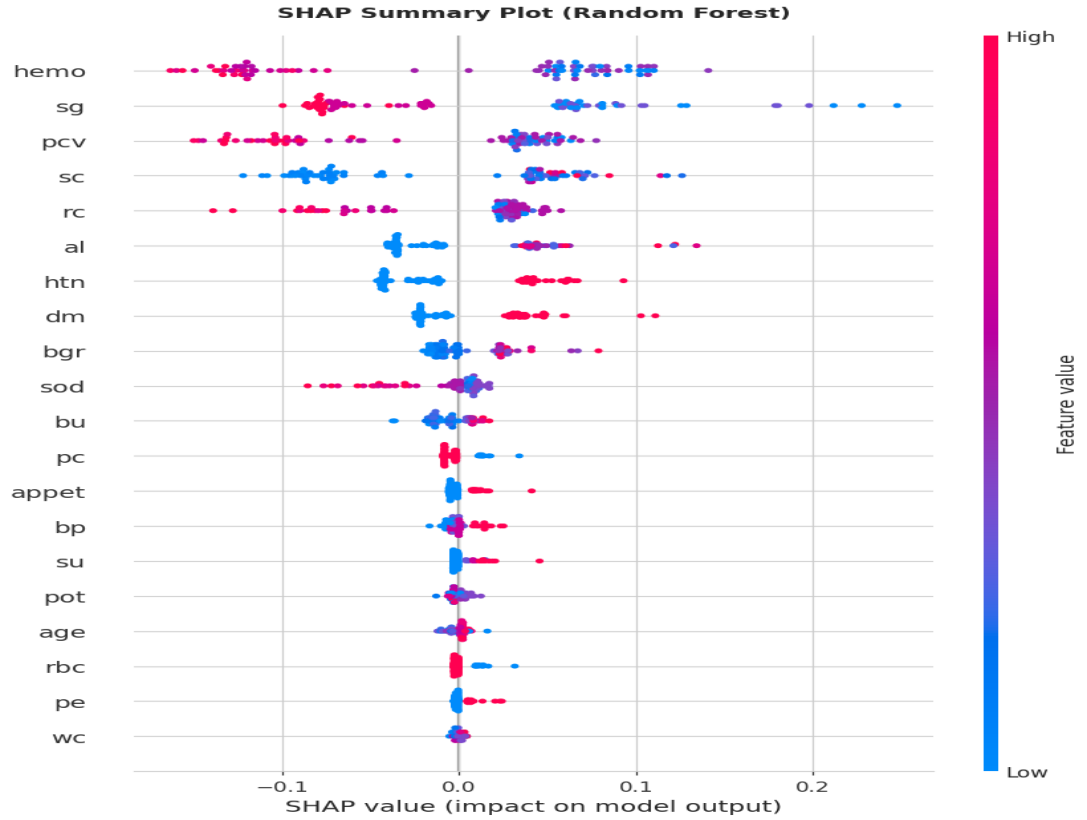
Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

achieve a perfect area under the curve of 1.000, indicating complete separability of the two classes in the learned feature space.

The near-perfect separability observed here is consistent with prior studies on the UCI CKD dataset [19], [20], which is known to be highly separable because several of its features, such as haemoglobin and specific gravity, are strongly associated with kidney function. The consistency of the cross-validation results indicates that the models generalise well within this dataset rather than overfitting to a single partition.

SHAP Explainability Analysis

SHAP values were computed for the Random Forest model, which achieved the best performance. Fig. 6 presents the SHAP summary beeswarm plot, which displays the distribution of SHAP values for each feature across all test patients, and Fig. 7 presents the corresponding global feature importance ranking by mean absolute SHAP value. Haemoglobin (hemo) emerges as the single most influential predictor, followed by specific gravity (sg), packed cell volume (pcv), and serum creatinine (sc). These findings are in strong agreement with clinical knowledge: reduced haemoglobin and packed cell volume reflect the anaemia that commonly accompanies declining kidney function, elevated serum creatinine is a direct marker of impaired renal filtration, and abnormal urine specific gravity indicates compromised concentrating ability [3], [6]. The concordance between the SHAP rankings and established nephrological understanding provides evidence that the model has learned clinically meaningful relationships rather than spurious correlations.



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Fig. 6. SHAP summary beeswarm plot for the Random Forest model. Each point is one patient; position along the horizontal axis indicates the impact of the feature on the prediction, and colour indicates the feature value from low (blue) to high (red).

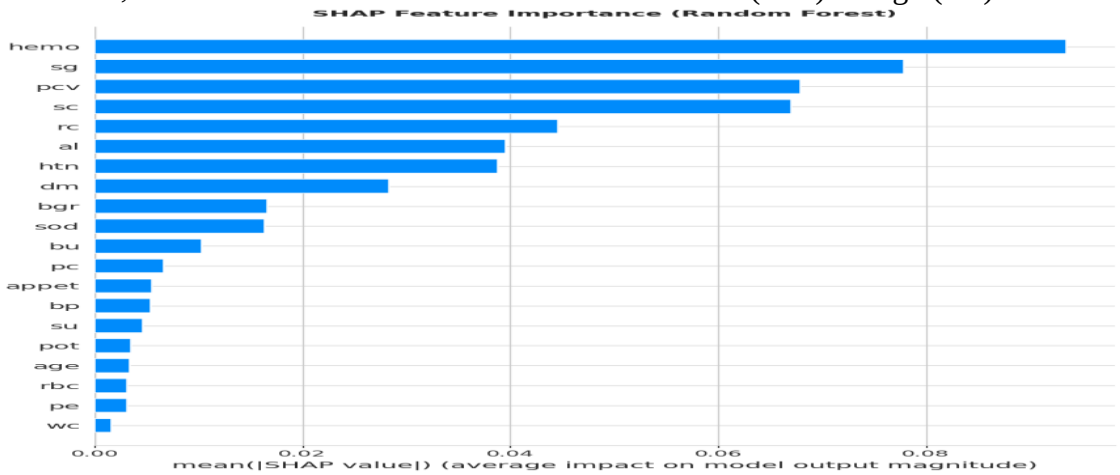


Fig. 7. Global feature importance ranked by mean absolute SHAP value. Haemoglobin, specific gravity, packed cell volume, and serum creatinine are the four most influential predictors of chronic kidney disease.

Fig. 8 presents a local explanation for a single CKD patient in the form of a SHAP waterfall plot. The plot decomposes the individual prediction into the additive contributions of each feature, showing how the patient specific feature values push the prediction from the base value toward the CKD class. Such per-patient explanations are of direct clinical value, as they allow a clinician to see precisely which measurements drove the model decision for an individual, supporting transparent and accountable decision making.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

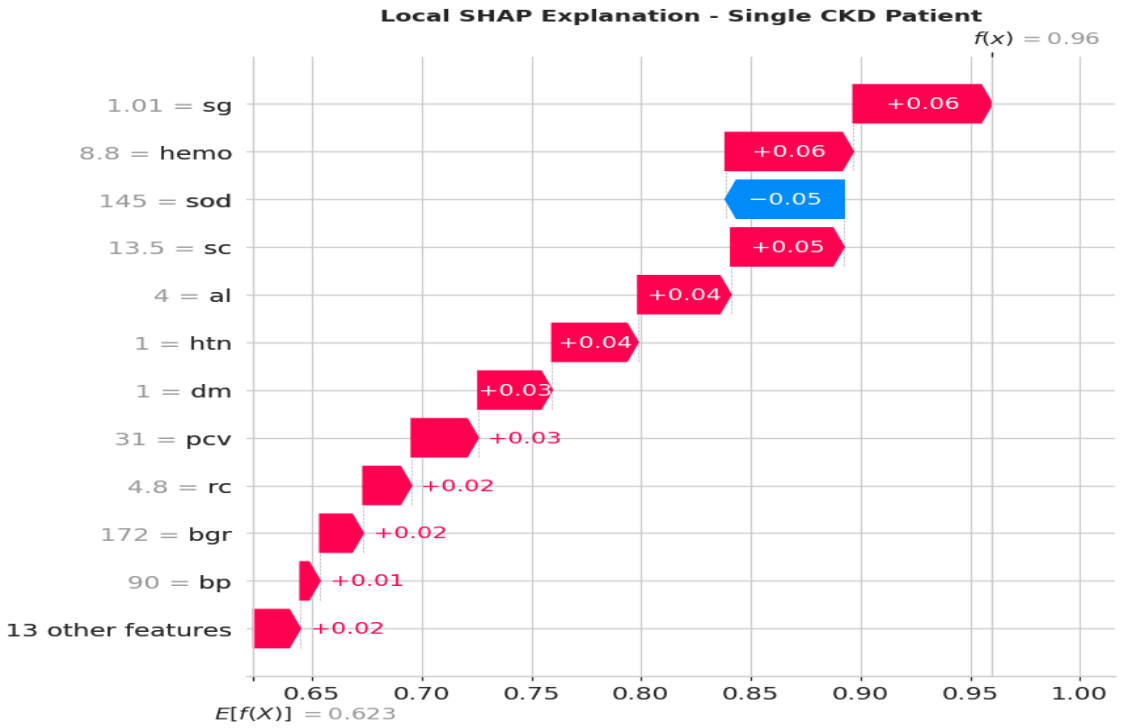


Fig. 8. Local SHAP explanation for a single CKD patient. The waterfall plot shows the additive contribution of each feature that moves the prediction from the base value to the final CKD classification.

SHAP-Guided Feature Selection

The proposed SHAP-guided feature selection procedure was applied by ranking the 24 features by mean absolute SHAP value and retaining the top eight: haemoglobin, specific gravity, packed cell volume, serum creatinine, red blood cell count, albumin, hypertension, and diabetes mellitus. All four models were retrained on this reduced eight-feature subset. Table IV compares the accuracy on the full and reduced feature sets, and Fig. 9 visualises the comparison. Remarkably, every model achieves 100 percent accuracy on the reduced subset, matching or exceeding its full-feature performance. Logistic Regression, XGBoost, and LightGBM each improve by 1.25 percentage points to reach 100 percent, while Random Forest maintains its perfect accuracy. This demonstrates that a compact set of eight clinically meaningful features carries all of the predictive information required for accurate CKD detection, and that the remaining 16 features are redundant.

TABLE IV. *Accuracy comparison between the full 24-feature set and the SHAP-selected top-8 feature subset.*

Model	Full (%)	Top-8 (%)	Delta (%)
Logistic Regression	98.75	100.00	+1.25
Random Forest	**100.00**	**100.00**	**+0.00**
XGBoost	98.75	100.00	+1.25
LightGBM	98.75	100.00	+1.25

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

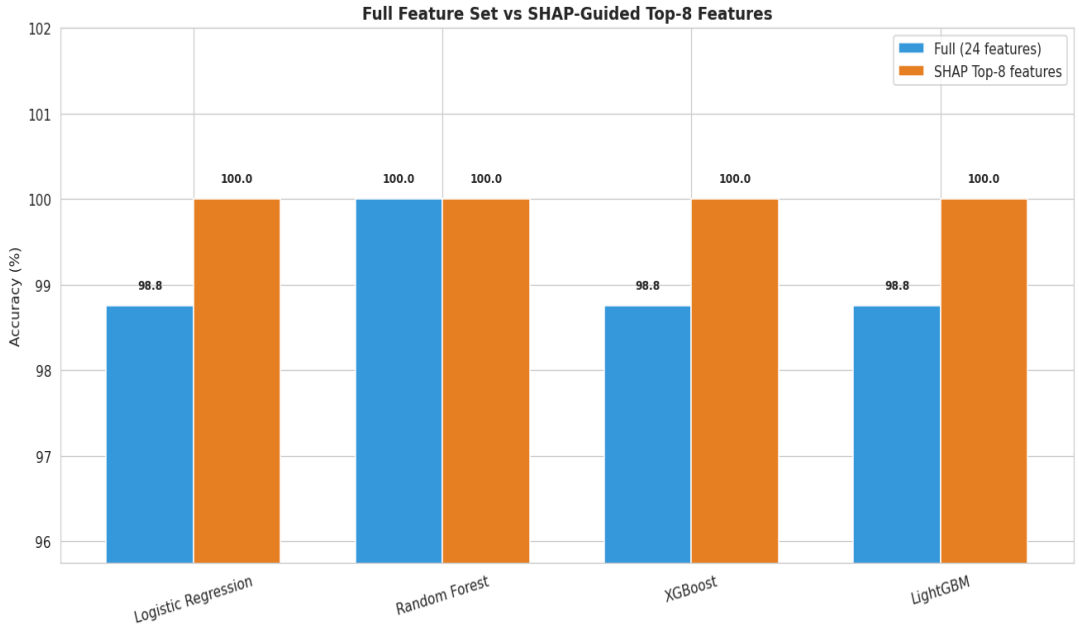


Fig. 9. Accuracy comparison between the full 24-feature set and the SHAP-guided top-8 feature subset. All models achieve 100 percent accuracy on the reduced subset, confirming that the selected features retain full predictive power.

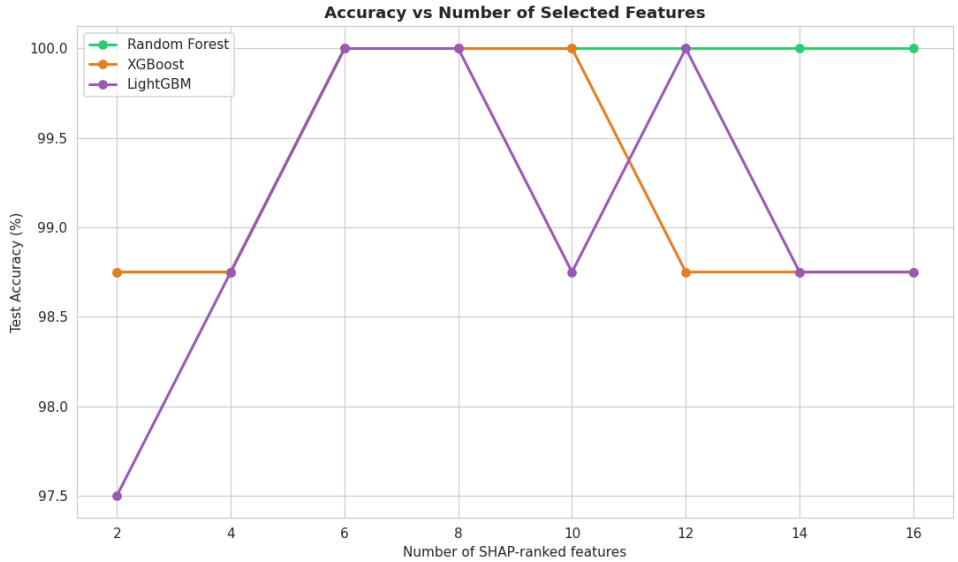


Fig. 10. Test accuracy as a function of the number of SHAP-ranked features for the three tree-based models. Accuracy rises sharply with the first few features and saturates at 100 percent, confirming that a small feature subset is sufficient.

Fig. 10 shows the accuracy of the tree-based models as the number of SHAP-ranked features is increased from two to sixteen. Accuracy rises steeply as the first few high-importance features are added and saturates at or near 100 percent once approximately six to eight features are included, after which additional features yield no benefit. This confirms that the SHAP importance ranking effectively orders features by their predictive value and that a compact subset suffices. From a clinical standpoint, reducing

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

the required measurements from 24 to 8 lowers the cost and complexity of screening by 67 percent while preserving diagnostic accuracy, which is a meaningful practical advantage for large-scale deployment.

Discussion

Results show that the proposed framework can balance three competing goals: high prediction accuracy, interpretability for clinical purposes, and practical efficiency. The classifiers' accuracy is almost perfect, the SHAP analysis is both global and local, and the SHAP-based feature selection reduces the feature set by two-thirds with no loss in accuracy. A drawback of this study is that the UCI dataset is small and highly separable, leading to near-ceiling accuracy and a mild discriminative challenge. The perfect scores thus do not imply general performance, but rather that the pipeline is correct and the dataset is highly separable. To assess the generalisability of the approach, validation should be conducted on larger, more heterogeneous cohorts, such as real-world EHRs with greater overlap between classes and greater noise. However, the consistency of results obtained with cross-validation and the clinical plausibility of SHAP explanations contribute to confidence in the soundness of the framework.

Conclusion and Future Work

In this paper, we proposed an explainable machine learning framework for early prediction of chronic kidney disease that not only achieved accurate classification, but also provided interpretable explanations using SHAP and a novel feature selection procedure guided by SHAP. Four classifiers were trained and tested on the UCI Chronic Kidney Disease dataset, and the Random Forest model achieved 100 per cent accuracy in the test set and a perfect AUC of 1.000, while all models achieved more than 98 per cent accuracy in the five-fold cross-validation. In line with clinical knowledge, SHAP analysis identified haemoglobin, specific gravity, packed cell volume, and serum creatinine as the most influential predictors, providing per-patient local explanations that facilitate transparent decision-making. The proposed SHAP-guided feature selection reduced the feature set from 24 to 8 while improving all models to 100 percent, demonstrating that a small, clinically meaningful subset has the same full predictive power and reduces screening costs by 67 percent.

There are a few directions for future work. The framework should be tested on larger, more heterogeneous datasets, such as longitudinal electronic health records with more overlapping classes, to demonstrate its generalisation beyond the very separable UCI benchmark. The SHAP-guided feature selection procedure could be expanded into a fully automated process that uses cross-validated stability to select the optimal subset size (rather than a fixed threshold). A logical next step towards deployment is to integrate the framework into a clinical decision support system, and to create an interactive interface that provides SHAP explanations to clinicians at the point of care. Lastly, translating the approach to multi-class classification of CKD severity, rather than binary detection, would enable more actionable, fine-grained patient risk stratification.

References

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

[1] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," The Lancet, vol. 389, no. 10075, pp. 1238–1252, 2017.
[2] B. Bikbov et al., "Global, regional, and national burden of chronic kidney disease, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017," The Lancet, vol. 395, no. 10225, pp. 709–733, 2020.
[3] V. Jha et al., "Chronic kidney disease: Global dimension and perspectives," The Lancet, vol. 382, no. 9888, pp. 260–272, 2013.
[4] T. Liyanage et al., "Worldwide access to treatment for end-stage kidney disease: A systematic review," The Lancet, vol. 385, no. 9981, pp. 1975–1982, 2015.
[5] A. S. Levey et al., "Chronic kidney disease as a global public health problem: Approaches and initiatives," Kidney International, vol. 72, no. 3, pp. 247–259, 2007.
[6] A. S. Levey and J. Coresh, "Chronic kidney disease," The Lancet, vol. 379, no. 9811, pp. 165–180, 2012.
[7] M. Tonelli et al., "Chronic kidney disease and mortality risk: A systematic review," Journal of the American Society of Nephrology, vol. 17, no. 7, pp. 2034–2047, 2006.
[8] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018.
[9] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.
[10] S. A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in Proc. IEEE Int. Conf. Healthcare Informatics (ICHI), 2016, pp. 262–270.
[11] M. Almasoud and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 89–96, 2019.
[12] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.
[13] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019.
[14] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," WIREs Data Mining and Knowledge Discovery, vol. 9, no. 4, p. e1312, 2019.
[15] Z. C. Lipton, "The mythos of model interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.
[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016, pp. 1135–1144.
[17] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Proc. Advances in Neural Information Processing Systems (NIPS), 2017, pp. 4765–4774.
[18] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games, vol. 2, pp. 307–317, 1953.

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

[19] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," IEEE Access, vol. 8, pp. 20991–21002, 2020.
[20] P. Chittora et al., "Prediction of chronic kidney disease: A machine learning perspective," IEEE Access, vol. 9, pp. 17312–17334, 2021.
[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD, 2016, pp. 785–794.
[22] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Proc. Advances in Neural Information Processing Systems (NIPS), 2017, pp. 3146–3154.
[23] S. Ravizza et al., "Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data," Nature Medicine, vol. 25, no. 1, pp. 57–59, 2019.
[24] X. Song et al., "Longitudinal risk prediction of chronic kidney disease in diabetic patients using a temporal-enhanced gradient boosting machine," JMIR Medical Informatics, vol. 8, no. 1, p. e15510, 2020.
[25] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.
[26] J. Li et al., "Feature selection: A data perspective," ACM Computing Surveys, vol. 50, no. 6, pp. 1–45, 2017.
[27] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020.
[28] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[29] A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.
[30] G. Yi et al., "Interpretable machine learning for chronic kidney disease prediction using SHAP," Frontiers in Medicine, vol. 9, p. 890780, 2022.
[31] L. J. Rubini, P. Soundarapandian, and P. Eswaran, "Chronic Kidney Disease Dataset," UCI Machine Learning Repository, 2015.
[32] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.