I. J. Education and Management Engineering, **2025, 3, 24-39**
Published Online on June 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2025.03.03

Predictive Modeling for Chronic Kidney Disease: A Comparative Analysis of Machine Learning Techniques with Explainable AI for Clinical Transparency

Abhinav Shivhare
Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra, India
Email: abhinavshiv001@gmail.com
ORCID iD: https://orcid.org/ 0009-0002-3797-9323

A. Charan Kumari
Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra, India
Email: charankumari@dei.ac.in
ORCID iD: https://orcid.org/0000-0002-3160-1912

K. Srinivas*
Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra, India
Email: ksri12@gmail.com
ORCID iD: https://orcid.org/0009-0002-3884-6282
*Corresponding Author

Received: 04 November, 2024; Revised: 07 January, 2025; Accepted: 12 March, 2025; Published: 08 June, 2025

Abstract: Chronic Kidney Disease (CKD) is considered a leading cause of high morbidity and mortality. Therefore, it needs early detection to allow timely intervention aimed at the enhancement of the patient outcome. The current study presents a Transparent CKD ML which combines the predictive power of efficient ML methods with the eXplainable AI techniques for transparent interpretibility of the prediction. This study has conducted an in-depth performance evaluation of the predictive power of the following eight machine learning algorithms: Logistic Regression, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, CatBoost, XGBoost, and AdaBoost on the 'Chronic Kidney Disease' dataset provided by UCI Machine Learning Repository. As a further study on algorithm performance, performance measures of accuracy, precision, recall, and F1 score were calculated; it was determined that Logistic Regression, Random Forest, and AdaBoost were performing very well and achieved 100% score in all metrics. This study further combined the ML models with eXplainable AI (XAI) techniques to increase the transparency of the models. SHapley Additive exPlanations (SHAP) an XAI technique was used to provide critical insights into the causality that dictates the predictions of CKD. Thus, this combination ensures the best performance of the model, increasing the trust in AI within clinical practice. The present study, therefore, unleashes the transformational potential of AI technologies in radically renovating the management of CKD and improving patient outcomes across the world.

Index Terms: Chronic Kidney Disease, Machine Learning, Explainable AI, SHAP, Early Detection

1. Introduction

According to the National Kidney Foundation, over 10% of the population is affected by Chronic Kidney Disease (CKD), and millions of affected people die every year [1]. CKD has become a global health concern because of its high death rate. According to WHO, CKD is becoming a serious health issue that threatens various developing nations. In the early stage, CKD is treatable, but in the last stage, it leads to kidney failure [2]. CKD is defined as a progressive deterioration in the functioning of kidney over a period that causes the waste products accumulation and fluid imbalances in the body. People of different ages and ethnic backgrounds are impacted by CKD [3].

A common problem with CKD is that the disease progresses silently in its early stage. Lack of nonspecific symptoms causes a delay in the diagnosis and treatment. Delays in diagnosis lead to a higher chance of complications and worse results for the affected people. Even though chronic kidney disease (CKD) is quiet, it can cause various dangerous side effects, such as fluid imbalances, anemia, cardiovascular illness, and bone abnormalities. Therefore, detection of CKD with treatment in the starting phase is the best way to control CKD otherwise, the condition of daily dialysis or kidney transplant may arise to sustain daily living [4,5].

The estimated glomerular filtration rate (eGFR) [6] and the presence of proteinuria are crucial indicators of kidney function that aid in diagnosing chronic kidney disease (CKD). Managing CKD involves maintaining a healthy lifestyle, monitoring kidney function regularly, and controlling conditions like diabetes and hypertension.

Considering this, machine learning is becoming as an effective tool in the battle against chronic kidney disease. Using the enormous potential of clinical data sets, ML algorithms can find complex correlations and patterns between a variety of factors, including blood pressure, blood sugar, genetic information, and the onset of chronic kidney disease. With this improved understanding, the models can recognise those at higher risk of contracting the disease can be developed. Machine learning models have the potential to greatly improve patient outcomes and reduce the overall burden of CKD on global healthcare systems by facilitating early intervention [7].

Chronic Kidney Disease (CKD) presents significant diagnostic challenges due to its asymptomatic nature in its early stages, high variability in clinical parameters, and overlapping symptoms with other conditions. Early and accurate detection is critical but often hindered by inconsistent data quality, missing values, and the lack of integrated diagnostic tools. In this context, machine learning offers promising capabilities; however, the performance and reliability of different algorithms remain inconsistent across studies. This work aims to systematically compare a diverse set of machine learning algorithms on a standard CKD dataset, evaluate their predictive performance using multiple metrics, and assess their suitability for clinical deployment. By doing so, the study aims to identify robust models that can assist in early diagnosis and support healthcare professionals in managing CKD more effectively.

The main contributions of this research work include:

1. The paper analyzes the efficacy of various machine-learning algorithms for diagnosing Chronic Kidney Disease using advanced machine-learning techniques, leveraging histopathology data from the UCI Repository.
2. It also emphasizes the significance of hyperparameter tuning for optimizing model performance, highlighting the detailed approach taken to achieve the highest predictive accuracy.
3. The study emphasizes the importance of eXplainable AI (XAI) in meeting the informational needs of clinicians regarding the interpretation of AI models' predictions.
4. By demonstrating the effectiveness of machine learning in early CKD detection and by incorporating XAI techniques, the research aims to enhance the trust and acceptance of AI-based predictive models in clinical practice and also intends to integrate this advanced transparent CKD prediction model technology into healthcare, aiming to enhance patient outcomes and lessen the global CKD burden.

The remaining part of the paper is organized as: In section 2, Literature review is presented, the methodology is presented in section 3 which includes further details about the dataset, pre-processing of the dataset, model description and the evaluation metrics. Results and discussions are presented in section 4 and section 5 discusses the interpretability of the prediction model. Conclusion and the related future work are mentioned in section 6.

2. Literature Review

This section presents a brief overview of the methods adopted by various researches in the field.

Anusorn *et al.* [8] conducted a study utilizing four machine learning techniques, namely KNN, SVM, Decision Tree, and Logistic Regression (LR), to predict CKD. The study found that SVM exhibited the highest accuracy and sensitivity among the models tested. Navaneeth and Suchetha [9] proposed a non-invasive method for diagnosing Chronic Kidney Disease. In their research, a sensor module was designed to measure the urea level in saliva. CNN - SVM hybrid deep learning was developed to predict CKD using sensor data. Their model achieved a high accuracy of 96.59%.

Ebrahime *et al.* [10] presented a novel diagnostic system for CKD prediction. Four different classification algorithms were used including, SVM, Decision Tree, KNN and Random Forest. Random Forest performed well by achieving highest accuracy. They used the dataset available on the UCI machine learning repository. Janani J and Sathyaraj R [11] addressed the issue of CKD, emphasizing the significance of early detection and treatment. They used the CKD dataset and demonstrated the hybrid model combining Random Forest and AdaBoost that achieved superior accuracy.

Pankaj *et al.* [12] demonstrated CKD prediction using machine learning algorithms and a deep neural network. They used a UCI repository dataset, seven classifiers, and a feature selection technique. Linear support vector machine (LVSM) accuracy was 98.86%, and deep neural networks achieved highest accuracy, which highlighted deep learning as more promising. Nikhila [13] explored four ensemble algorithms using the chronic kidney disease dataset in his

study: AdaBoost, Random Forest, Boosting, and Bagging. He evaluated their performance using seven metrics. Random Forest and AdaBoost achieved superior performance with good accuracy, precision, sensitivity, and Mathew Correlation Coefficient scores.

Vijendra *et al.* [14] addressed the challenges in CKD detection. They introduced a deep learning approach for CKD prediction and compared its performance with different classification algorithms such as Logistic Regression, Naive Bayes, SVM, KNN and Random Forest. They used Recursive Feature Elimination (RFE) for selecting the key features. Their proposed model achieved good accuracy and outperformed the other methods.

Manal A *et al.* [15] addressed CKD prediction using a machine learning algorithm on Big Data platforms (Apache Spark). They used a hybrid approach by combining feature selection techniques like chi-squared and Relief—F with various machine learning algorithms, including Naive Bayes, logistic regression, decision trees, Support Vector Machine, Gradient Boosted Trees, and Random Forest. Their proposed hybrid model, Decision Tree, SVM and Gradient Boosting Trees classifiers, achieved good accuracy with Relief—F feature.

Ramya *et al.* [16] proposed a unique hybrid deep learning approach. In their study, missing values were eliminated, and noise was reduced from the dataset, and an enhanced capsule network method was used for features extraction. The hybrid models, BConvLSTM and DNetCNN, were used for the classification and outperformed other existing models with good accuracy. Samit and Ahsan [17] investigated machine learning algorithms for the prediction of chronic kidney diseases and reported that among all the methods XGBoost model performed well. In addition, they applied SHAP and LIME algorithms to demonstrate the influence of the features on the prediction.

3. Methodology

This section presents the details of the dataset used in this research work, preprocessing details, a short description of the experimented algorithms and the performance evaluation metrics.

3.1 Dataset

The 'Chronic Kidney Disease' dataset, which is available on UCI Machine Learning Repository, was used in this paper. This dataset was also used by several researchers [18]. The dataset includes 400 distinct cases representing patients' medical records and 25 different attributes or features that helped in the diagnosis. Out of 400 patient records, 250 patients were marked with CKD and the remaining 150 patient's records were categorized as healthy. Although the imbalance in the dataset affects accuracy, to address this, the model performance was also assessed using precision, recall, and F1 score, which are more appropriate for imbalanced datasets. Ensemble methods like Random Forest and AdaBoost were also implemented and tested in the present study as they can handle imbalance due to their sampling mechanisms. The description about different features is shown in Table 1.

Table 1. Description of Features Present in the Dataset

Features	**Description**
Age	Patient's Age
Blood Pressure	Blood pressure
Specific Gravity	Urine specific gravity
Albumin	Albumin
Sugar	Sugar
Red Blood Cells	level of RBCs in blood
Pus Cell	WBCs count
Pus Cell Clumps	Dead tissue and bacteria related infection
Bacteria	Presence of bacteria
Blood Glucose Random	Test to check glucose level
Blood Urea	Amount of waste product removed by kidney
Serum Creatinine	Creatinine amount in blood
Sodium	Amount of sodium present
Potassium	Amount of potassium
Hemoglobin	Protein inside RBCs
Packed Cell Volume	Blood cells volume in blood
White Blood Cell Count	WBC count
Red Blood Cell Count	RBC count
Hypertension	Severe hypertension
Diabetes Mellitus	Inadequate production of insulin
Coronary Artery Disease	Blockage of artery
Appetite	Hunger level
Pedal Edema	Swelling in body
Anemia	Low red blood cells count
Class	Target Variable

3.2 Preprocessing

For proper training of the machine learning model, it becomes essential to tackle the values that are missing in this dataset. Random sampling was used to impute missing values in numerical features to handle the missing value issue, while mode imputation was applied to categorical features. Random sampling preserves the original distribution of the data better than mean or median imputation, reducing the risk of introducing central tendency bias. Mode imputation for categorical variables is a simple and widely used approach, particularly when the mode represents a meaningful and frequent category. However, these methods may introduce bias by assuming that the missing values are missing at random (MAR) and, thus, may not fully capture the underlying variability or patterns.

The original dataset includes some categorical data (like appetite, pedal edema, and hypertension), which creates challenges in building the models. As machine learning models typically work with numbers or math, it becomes important to convert that categorical data into numeric data. So, a Label Encoding method was used that gives each category a unique number tag. Fig. 1 shows the number of Numerical and Categorical Features.

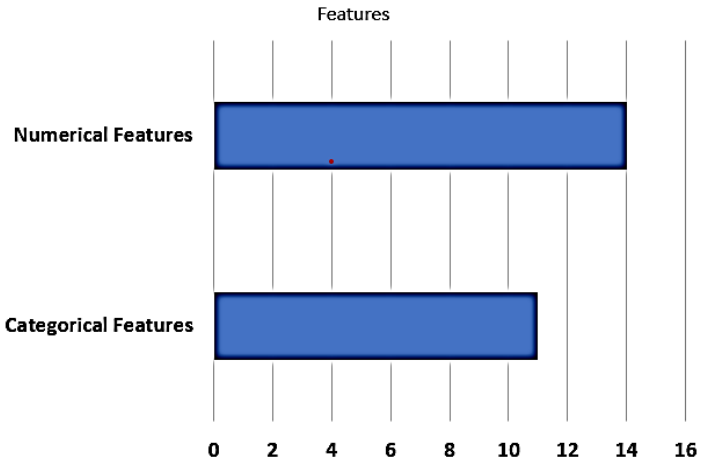

Fig. 1. Counts of Categorical and Numerical Features in Dataset

The entire dataset was separated into two subsets: one set for training purpose and another set for testing purpose. The training dataset contains 70% of the scaled data, i.e. 280 patient records were used for model's training, while the testing dataset consists of 30%, i.e. 120 patient records that were used to assess how well the trained models performed on unseen data.

3.3 Model Description

This section briefly describes the implemented machine learning algorithms. The selection of machine learning models in this study was guided by the need to represent a diverse set of learning paradigms commonly used in medical data analysis. Logistic Regression was chosen as a baseline due to its simplicity, interpretability, and strong performance in linearly separable problems. K-Nearest Neighbors and Support Vector Machine were included to explore instance-based and margin-based learning methods, respectively. Decision Tree and Random Forest represent tree-based models with Random Forest offering robustness through ensemble learning. Gradient boosting algorithms—CatBoost, XGBoost, and AdaBoost—were selected for their proven effectiveness in handling structured data and their capacity to capture complex patterns through iterative refinement. This comprehensive selection ensures a fair comparative analysis of models ranging from interpretable linear techniques to powerful ensemble methods suited for medical diagnostics.

3.3.1 Logistic Regression

It is a popular supervised machine learning technique which is majorly applied in classification tasks. It predicts the probability, that the given instance is a part of a specific class. It makes use of a sigmoid or logistic function that converts the linear combination of input features using weights into a probability between 1 and 0. Mathematically, the logistic function is defined as given in equation 1.

$$\sigma(z) = \frac{1}{1+ e^{-z}} \tag{1}$$

where z represents the linear combination of weights and input features.

If the probability is more significant than 0.5 (threshold value), the instance is likely to belong to class 1 (positive class); otherwise, the instance belongs to class 0 (negative class).

3.3.2 K-Nearest Neighbors (KNN)

KNN is a simple ML algorithm that usually applied to problems with both regression and classification. In this algorithm, the distance of input instances from other instances in the training dataset is computed using Euclidean distance. Mathematically, the Euclidean distance formula between two points is as shown in equation 2.

$$d(m,n) = \sqrt{\sum_{i=1}^{k}(m_i - n_i)^2} \quad (2)$$

where m and n are variables for input data.

Further, the computed distance is used to select k closest neighbors. The unseen instances are identified according to the K's value. This algorithm is termed a lazy learning algorithm as it is non-parametric and saves the training data points instead of learning from them. This algorithm is majorly useful when we need to classify objects based on different attributes by performing a pattern recognition task.

3.3.3 Support Vector Machine

SVM is commonly favored for classification tasks. It constructs an optimal hyperplane in an N-dimensional space to segregate instances into classes. The hyperplane is determined by the most significant margin between classes, calculated using support vectors (data points) nearest to the hyperplane.

3.3.4 Decision Tree

Decision tree is a tree-structured algorithm where the leaf node indicates the class label or result, the internal nodes represent features and the branches indicate decision rules. The decision tree has two kinds of nodes: a leaf node having fewer branches and a decision node that have more branches. Until the leaf node is reached, the data is recursively divided into subsets based on the attribute's value to construct the tree. It is preferably used as a base learner in ensemble methods like random forest and AdaBoost to improve their generalization performances.

3.3.5 Random Forest

Random Forest can be applied to both classification as well as regression problems. It uses multiple decision trees, and each decision tree is trained using a different random subset of the training dataset and considers only a random selection of features. This technique combines the prediction of each tree and makes the outcome either by majority vote in case of classification problems or average in case of regression tasks. A few trees give higher accuracy and prevents the problem of overfitting of model and maintains the robustness.

3.3.6 CatBoost

CatBoost, or Categorical Boosting, is an ensemble learning algorithm in machine learning that does not require any encoding techniques or methods which helps in converting categorical data into numerical data present in the dataset. Multiple decision trees are iteratively trained on the training dataset to fix the errors made by the previous ones and then combined to create an ensemble model. To improve model generalization and prevent overfitting, CatBoost uses advanced regularization techniques like L2 regularization.

3.3.7 XGBoost

XGBoost, or Extreme Gradient Boosting, is a popular ensemble learning method that combines weak models to produce strong predictions. It sequentially builds multiple decision trees to correct errors made by prior models using Gradient Boosting to guide the learning process, particularly on large datasets due to features like parallelization.

3.3.8 AdaBoost

AdaBoost is another ensemble technique that combines base learners into stronger ones by iteratively training weak learners, typically decision trees, on subsets of training data with varying weights assigned to data points based on classification accuracy. It enhances weak learner performance by focusing on misclassified instances and combining results to prevent overfitting.

3.4 Evaluation Metrics

Performance metrics are quantitative measurements used to evaluate how well machine learning models predict or classify data based on a dataset. These metrics provide insight into how accurately models work in real-life scenarios by providing insights into several performance metrics such as:

3.4.1 Accuracy

Accuracy as defined in equation 3, describes how often the model's prediction comes true out of the total predictions it makes.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

3.4.2 Precision

Precision of the model focuses on the positive predictions that the model makes. It shows how often the model predict positive instances without mislabeling negative as positive as shown in equation 4.

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

3.4.3 Recall

Recall as defined in equation 5, measures how well the model is at finding every positive instance present in the dataset. It indicates whether the model is able to catch all the positives without missing any.

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

3.4.4 F1 score

It is a harmonic mean of precision and recall which provides a balance between precision and recall when the classes in the dataset are imbalanced, as indicated in equation 6. It ranges between 0 to 1. A higher F1 – score indicates that model performed well in terms of both precision and recall.

$$F1\ Score = 2 * \frac{\text{Recall*Precision}}{\text{Recall+Precision}} \tag{6}$$

where,

True Positive (TP) - Records that were rightly predicted with CKD.
True Negative (TN) - Records that were rightly predicted with Not CKD.
False Positive (FP) - Records that were mistakenly predicted with CKD.
False Negative (FN) - Records that were mistakenly predicted with Not CKD.

Although standard evaluation metrics such as accuracy, precision, recall, and F1 score provide a quantitative measure of model performance, they have limitations when applied in clinical settings. For instance, high accuracy can be misleading in imbalanced datasets where the model may correctly predict the majority class but fail to detect high-risk patients in the minority class. In CKD diagnosis, false negatives (i.e., failing to detect a CKD patient) are more critical than false positives, as delayed intervention can worsen patient outcomes. Therefore, recall becomes particularly important in minimising missed diagnoses, while precision ensures that positive predictions are reliable. F1 score offers a balance, but its clinical relevance still depends on the healthcare context.

4. Results and Discussion

This section presents the results obtained by the implemented algorithms. This research work was performed on the chronic kidney disease dataset from the UCI Repository, containing records of 400 patients with 25 features. The dataset was pre-processed by correcting missing values using random sampling for numerical data and mode imputation for categorical data. Label encoding transformed categorical data into numerical format, and the dataset was split into 70:30 ratios for testing and training. Eight algorithms were evaluated for CKD prediction, with hyperparameters tuned for optimal performance. The performance of model was evaluated using four evaluation metrics such as accuracy, Recall, Precision, and F1 score.

Table 2. Classification Report of Logistic Regression

Class	Precision	Recall	F1 score	Support
Not CKD	1.00	1.00	1.00	44
CKD	1.00	1.00	1.00	76

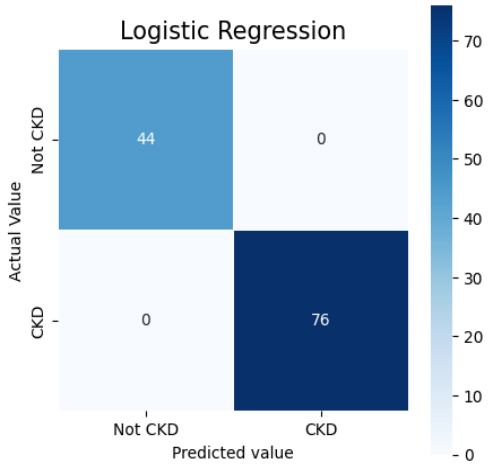

Fig. 2. Confusion Matrix of Logistic Regression

Table 2 shows the score of 100% for Logistic Regression in terms of recall, precision and F1 score for both the classes. All the records were correctly classified for both Not CKD and CKD classes which is shown in Fig. 2.

Table 3. Classification Report of KNN

Class	Precision	Recall	F1 score	Support
Not CKD	0.63	0.73	0.67	44
CKD	0.83	0.75	0.79	76

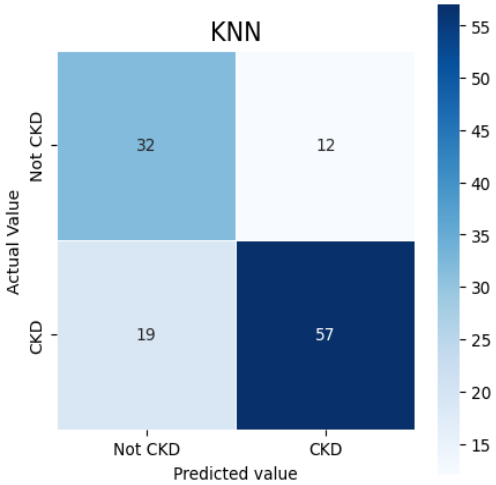

Fig. 3. Confusion Matrix of KNN

Classification report is presented in Table 3 which shows that KNN has achieved precision of 63% and 83%, recall of 73% and 75% and F1 score of 67% and 79% for Not CKD and CKD classes respectively. Fig. 3 shows KNN algorithm falsely classified 19 patients with CKD and 12 patients with Not CKD.

Table 4 Classification Report of SVM

Class	Precision	Recall	F1 score	Support
Not CKD	1.00	0.98	0.99	44
CKD	0.99	1.00	0.99	76

SVM algorithm has achieved good results as shown in Table 4. Recall score is 98% and F1 score is 99% for Not CKD class, and precision and F1 score is 99% for CKD class. Fig. 4 shows that SVM falsely predicted 1 patient as having Not CKD and rest of the instances were perfectly classified.

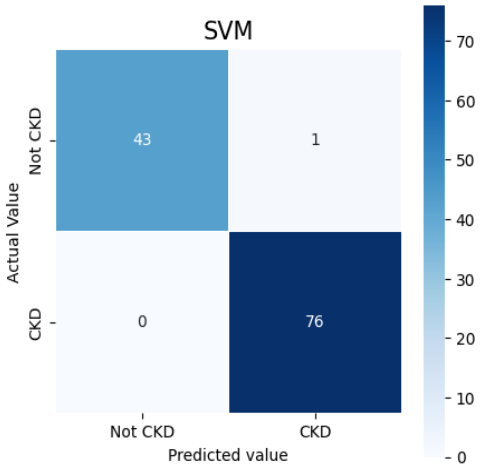

Fig. 4. Confusion Matrix of SVM

Table 5. Classification Report of Decision Tree

Class	Precision	Recall	F1 score	Support
Not CKD	0.95	0.95	0.95	44
CKD	0.97	0.97	0.97	76

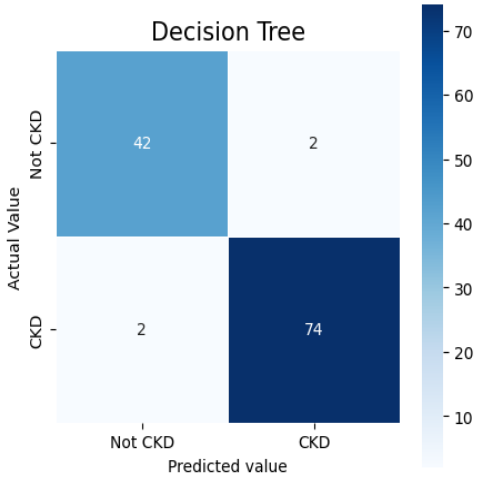

Fig. 5. Confusion Matrix of Decision Tree

Decision Tree algorithm's performance is shown in Table 5, where precision is 95%, recall and F1 score are also 95% for Not CKD class and 97% for CKD class. Confusion matrix is represented in Fig. 5, which shows Decision Tree algorithm falsely misclassified 2 patient as having Not CKD and 2 patient having CKD.

Table 6. Classification Report of Random Forest

Class	Precision	Recall	F1 score	Support
Not CKD	1.00	1.00	1.00	44
CKD	1.00	1.00	1.00	76

Table 6 shows that Random Forest achieved a perfect result with recall, precision and F1 score of 100%. Random Forest algorithm accurately classified all the instances correctly for both the classes as shown in Fig. 6.

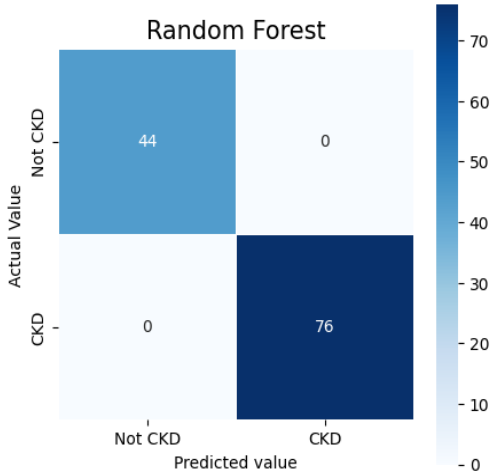

Fig. 6. Confusion Matrix of Random Forest

Table 7. Classification Report of CatBoost

Class	Precision	Recall	F1 score	Support
Not CKD	1.00	0.98	0.99	44
CKD	0.99	1.00	0.99	76

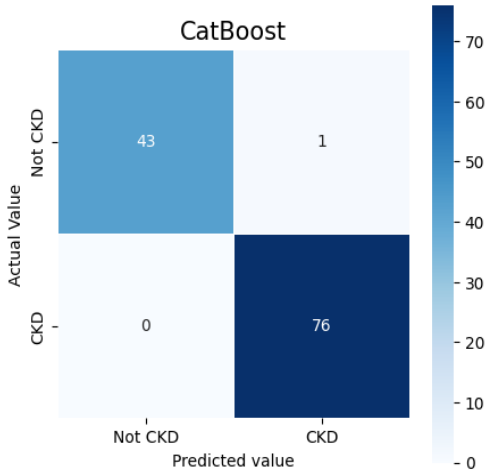

Fig. 7. Confusion Matrix of CatBoost

CatBoost algorithm has achieved good scores as shown in Table 7. Recall score is 98%, F1 score is 99% for Not CKD class, and precision and F1 score is 99% for CKD class. Fig. 7 shows that CatBoost falsely predicted 1 patient as having Not CKD and rest of the instances were perfectly classified.

Table 8. Classification Report of XGBoost

Class	Precision	Recall	F1 score	Support
Not CKD	1.00	0.95	0.98	44
CKD	0.97	1.00	0.99	76

Table 8 shows the performance of XGBoost algorithm where recall and F1- score for Not CKD class are 95% and 98% respectively and for CKD class precision is 97% and F1 score is 99%. Confusion matrix is represented in Fig. 8 which shows XGBoost algorithm falsely misclassified 2 patients as having Not CKD.

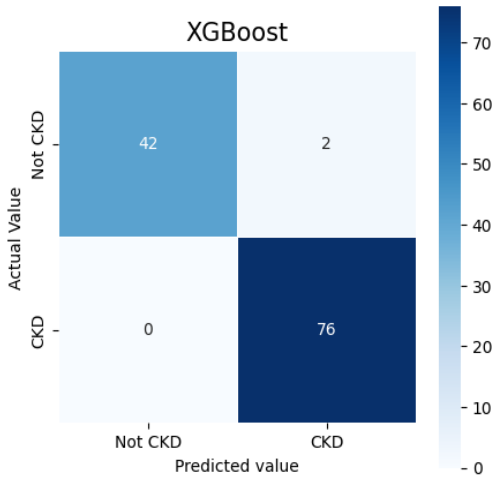

Fig. 8. Confusion Matrix of XGBoost

Table 9. Classification Report of AdaBoost

Class	Precision	Recall	F1 score	Support
Not CKD	1.00	1.00	1.00	44
CKD	1.00	1.00	1.00	76

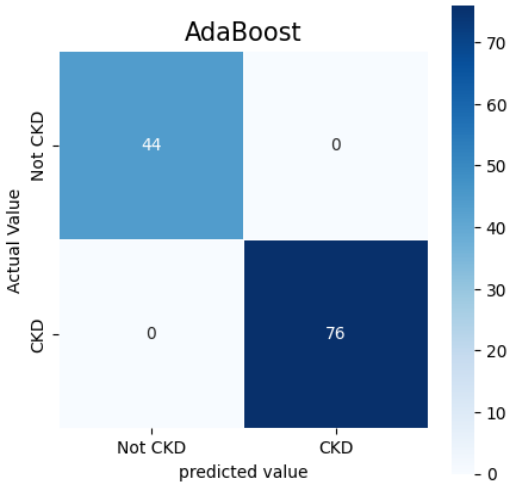

Fig. 9. Confusion Matrix of AdaBoost

AdaBoost algorithm has achieved a remarkable result of 100% accuracy, recall, precision and F1 score as shown in Table 9. AdaBoost algorithm accurately classified all the instances correctly for both the classes as shown in Fig. 9. Table 10 reports a comparative analysis of all the implemented algorithms based on various performance metrics. It is evident from Table 10 that Logistic Regression, Random Forest and AdaBoost have achieved 100% accuracy while SVM and CatBoost have 99.16% accuracy, KNN has 74.16% accuracy, Decision Tree has 96.66% accuracy and XGBoost has 98.33% accuracy. In respect of Precision, Logistic Regression, Random Forest and AdaBoost have achieved the precision of 100%, then SVM, CatBoost and XGBoost have a precision of 99%, whereas KNN and Decision Tree have a precision of 73% and 96%, respectively. In case of recall, Logistic Regression, Random Forest and AdaBoost have achieved the recall of 100%, SVM and CatBoost obtained a recall of 99% and whereas KNN, Decision Tree and XGBoost have a recall of 74%, 96% and 98% respectively. In terms of F1-score, Logistic Regression, Random Forest and AdaBoost have an F1-score of 100%, whereas SVM and CatBoost have an F1-score of 99%. whereas KNN, Decision Tree and XGBoost obtained the lowest F1-score of 73%, 96% and 98%, respectively.

Table 10. Comparative analysis of algorithms based on their performance metrics

ML Algorithms	Accuracy	Precision	Recall	F1 score
Logistic Regression	1.00	1.00	1.00	1.00
KNN	0.74	0.73	0.74	0.73
SVM	0.99	0.99	0.99	0.99
Decision Tree	0.96	0.96	0.96	0.96
Random Forest	1.00	1.00	1.00	1.00
CatBoost	0.99	0.99	0.99	0.99
XGBoost	0.98	0.99	0.98	0.98
AdaBoost	1.00	1.00	1.00	1.00

While several models in this study achieved high accuracy, clinical applications demand a deeper understanding of the types of errors made. In the context of CKD detection, false negatives—where the model fails to identify an actual CKD patient—are particularly concerning, as they can lead to delayed diagnosis and progression to severe stages without intervention. On the other hand, false positives may result in undue stress, further invasive testing, or unnecessary treatments. Therefore, reducing false negatives should be prioritized in such applications, even at the cost of slightly increasing false positives. The evaluation using recall and precision helps quantify this trade-off.

5. Explainability Analysis of the Random Forest Model

Based on the results presented in the previous section, it is apparent that Random Forest, Logistic regression and AdaBoost performed well in classifying CKD and non-CKD patients. This section discusses the interpretability of predictions obtained by Random Forest model.

5.1. Summary Plot

The summary plot serves as a visual depiction showcasing the significance of each feature within the model. It proves to be a valuable tool in comprehending the model's prediction mechanisms and pinpointing the pivotal features. Within Figure 10, the feature importance hierarchy is illustrated through the SHapley Additive exPlanations (SHAP) summary plot for the Random Forest model. Notably, the three most critical features influencing the predictive model are hemoglobin, specific gravity, and albumin. The graph delineates the SHAP value along the x-axis and the feature rankings along the y-axis. Features endowed with higher SHAP values are deemed more impactful in discerning the category and are positioned at the forefront. In Fig. 10, the features are arranged in descending order based on their SHAP scores.

5.2. Waterfall Plots

The SHAP waterfall plot provides a visual representation of how each feature contributes to the final prediction of a machine learning model for a specific instance. Each bar in the plot represents the contribution of a specific feature to the model's prediction for a given instance. Red bars are indicative of positive impacts (enhancing the prediction score), whereas blue bars signify negative impacts (reducing the prediction score). Bars of greater length wield a stronger influence on the prediction process. The function f(x) denotes the output or forecast of the model for a particular instance 'x', while E[f(x)] signifies the anticipated value of the model's output across all feasible subsets of features for the instance 'x'. This concept is frequently employed as the fundamental or benchmark for elucidating the impacts of individual features in the SHAP values, aiding in the comprehension of the extent to which each feature diverges from the model's mean prediction when producing the final forecast for the instance 'x'.

The illustration in Fig. 11 showcases the waterfall plot of a rightly classified non-CKD instance. Within the realm of predicting non-CKD instances, the examination through SHAP waterfall plots furnishes valuable perspectives into the determinants influencing the forecasts. Manifesting a prediction score (f(x)) of 0.965 that considerably exceeds the expected value (E[f(x)]) of 0.382, the model showcases a pronounced inclination towards categorizing instances as non-CKD. Through an assessment of the impacts of distinct features, it becomes apparent that hemoglobin, packed cell volume, and specific gravity emerge as the foremost contributors, exhibiting SHAP values of 0.11, 0.08, and 0.07, correspondingly.

The waterfall plot of a correctly predicted CKD case is depicted in Fig. 12. In the context of predicting CKD cases, the SHAP waterfall analysis reveals significant insights into the predictive factors influencing the model's decision-making process. With a prediction score (f(x)) of 0.99, far exceeding the expected value (E[f(x)]) of 0.618, the model demonstrates a high confidence in classifying instances as CKD. Upon closer examination of feature contributions, it is observed that hemoglobin, specific gravity, and albumin to be the top three influencers, with respective SHAP values of 0.07, 0.06, and 0.05.

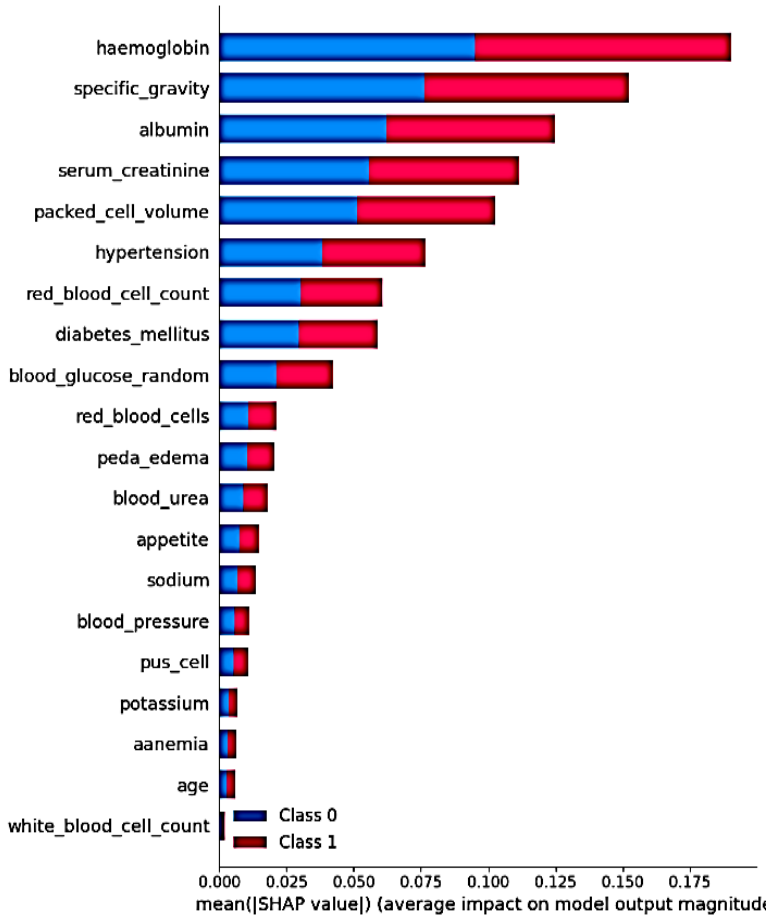

Fig. 10. Ranking of feature importance indicated by SHAP algorithm for prediction

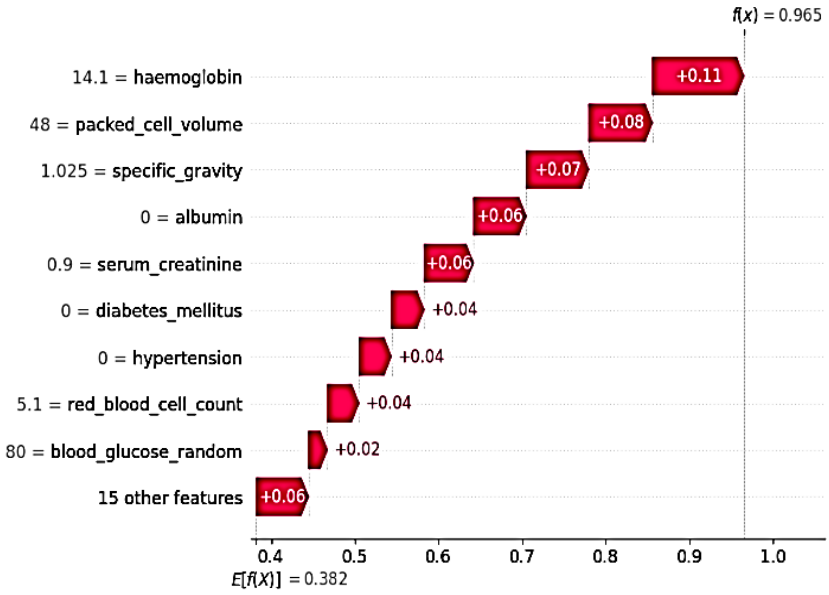

Fig. 11. Waterfall plot of a correctly classified non_CKD case

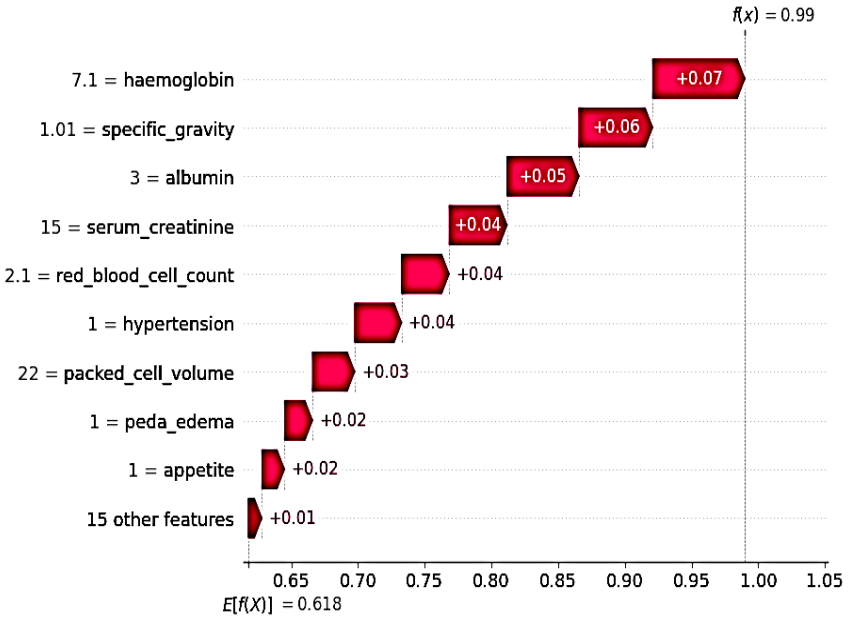

Fig. 12. Waterfall plot of a correctly classified CKD case

5.3. Force Plots

SHAP force plots are visualizations designed to provide a detailed understanding of individual predictions made by machine learning models. They display the contribution of each feature to the final prediction for a specific instance. The vertical dashed line in the center represents the base value, which is the model's average prediction across the dataset. Each horizontal bar corresponds to a feature in the dataset. The length of the bar indicates the magnitude of the feature's impact on the prediction. Positive (red) bars indicate features pushing the prediction upwards, while negative (blue) bars indicate features pushing it downwards. The final prediction for the instance is represented by the sum of the base value and the contributions of each feature. It is indicated by the arrow pointing to the final prediction value. Beneath each bar, the actual value of the feature for the instance is displayed. The force plots help in understanding the factors driving individual predictions and provide transparency into the model's decision-making process.

The force plot illustrating a properly classified non-CKD instance is presented in Fig. 13. Analysis of the plot reveals a model prediction probability value of 0.97. The base value, defined as the mean of the model's output across the training dataset, approximates 0.38. The numerical annotations on the plot's arrows correspond to the feature's value for this observation. Variations in impact are visualized along the x-axis. It is notable that specific features like hemoglobin, packed cell volume, and specific gravity exhibit favorable effects on the prediction outcome. These features demonstrate higher SHAP values, underscoring their significance in forecasting non-CKD cases. This observation suggests that the model effectively utilized these features to achieve precise predictions for non-CKD instances.

The force plot of a properly classified case of Chronic Kidney Disease (CKD) is depicted in Fig. 14. Within the SHapley Additive exPlanations (SHAP) force plot concerning the CKD instance, variables like hemoglobin, specific gravity, and albumin manifested favorable impacts on the forecast, as indicated by their positive SHAP values. This indicates that elevated levels of these biological markers were linked to an escalated probability of being categorized as having CKD. Additionally, a high predictive score (f(x)) of 0.99, notably surpassing the anticipated value (E[f(x)]) of approximately 0.62, emphasizes the model's certainty in identifying cases as CKD.

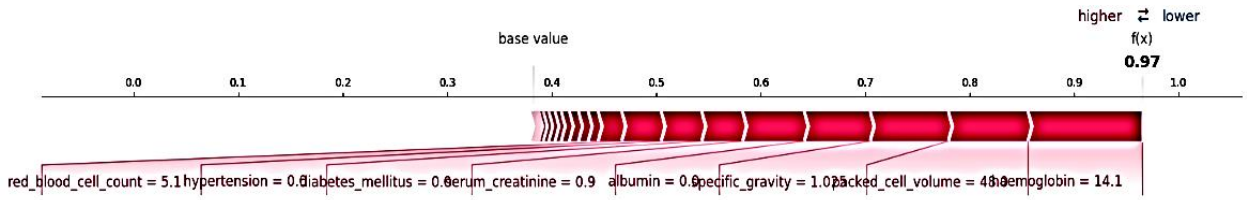

Fig. 13. Force plot of a correctly classified non_CKD case

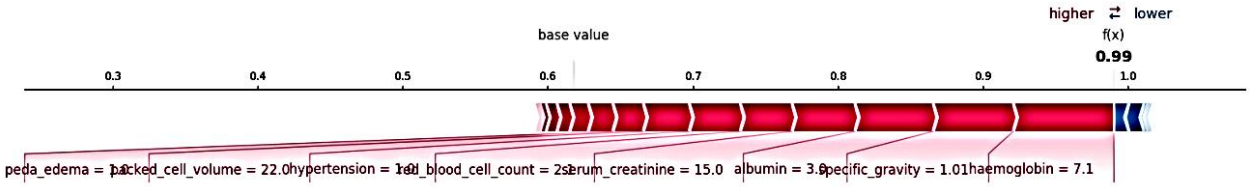

Fig. 14. Force plot of a correctly classified CKD case

The examination of SHAP force plots for the CKD instance underscores the model's capacity to utilize pertinent biomarkers for precise disease categorization. These results underscore the promise of machine learning models, reinforced by interpretable AI methods, in advancing early diagnosis and treatment of chronic kidney disease.

The use of SHAP values enhances transparency by showing the contribution of each feature to a specific prediction, enabling clinicians to understand *why* a model has classified a patient as CKD or non-CKD. This can support clinical decision-making in multiple ways. First, SHAP summary plots help identify key biomarkers (e.g., hemoglobin, specific gravity, albumin) that strongly influence the model's predictions, aligning with established clinical knowledge or uncovering new associations. Second, individual force and waterfall plots offer case-specific insights by highlighting which features pushed the prediction toward a CKD diagnosis. This can be particularly valuable in borderline cases, where physicians may use the explanation to verify model reasoning and combine it with other diagnostic information. By making AI decisions traceable, SHAP enhances clinicians' trust in model outputs and can serve as a complementary tool in diagnostic workflows.

6. Conclusion and Future Work

This research has successfully demonstrated the power of different machine learning algorithms in revolutionizing the early detection of CKD, a critical step towards mitigating its progression and enhancing patient care. Through the meticulous analysis of the dataset, incorporating advanced preprocessing techniques and a comprehensive evaluation of multiple machine learning models, this research has assessed multiple machine learning algorithms, including Logistic Regression, K– Nearest Neighbor (KNN), Support Vector Machine, Decision Tree, Random Forest, CatBoost, XGBoost and AdaBoost, and identified Logistic Regression, Random Forest and AdaBoost as the most effective tools for CKD prediction. These models, with their unparalleled performances on different evaluation metrics, stand at the forefront of this technological leap in healthcare system.

This study also incorporates SHAP (SHapley Additive exPlanations) an eXplainable AI technique to enhance the explainability of predictions made by a Random Forest model for the detection of chronic kidney disease. The visualization of summary, waterfall, and force plots, provided insights into the model's decision-making process for both CKD and non-CKD cases. The high prediction scores observed in both CKD and non-CKD cases, surpassing the expected values, underscored the model's confidence in accurately classifying instances. Furthermore, the transparent nature of SHAP plots provide clinicians and researchers with valuable insights into how the model leveraged biomarkers to make predictions, enhancing interpretability and facilitating informed decision-making in clinical practice.

These research findings contribute to the ongoing efforts to integrate cutting-edge technologies into medical diagnostics and open new avenues for future research, particularly in exploring the capabilities of ensemble methods and deep-learning architectures. As we move forward, the integration of such advanced computational techniques promises to enhance the management and diagnosis of chronic diseases significantly, ultimately resulting in better health outcomes for patients worldwide. This research demonstrates the transformative power of a transparent CKD prediction model in healthcare, marking a significant step towards the early and accurate detection of CKD.

A promising direction for future research is the validation of the proposed machine learning models in real-world clinical settings. While the current study utilises the UCI CKD dataset for benchmarking, future work can focus on evaluating the models using independent, diverse datasets collected from hospitals or clinical registries. This would help assess the generalizability of the models across different patient populations and healthcare environments. Future research can explore advanced feature engineering techniques to enhance model performance. Real-world clinical data is often noisy, incomplete, and heterogeneous, significantly impacting model performance. To ensure practical applicability, future research should involve stress-testing the models with noisy or partially missing data, assessing generalizability across diverse patient populations, and validating performance using cross-institutional datasets. These steps will help determine the proposed predictive models' actual robustness and clinical readiness. Beyond predictive performance, a critical future direction involves exploring the practical deployment of these models within real-world healthcare systems. Effective implementation would require integration with electronic health record (EHR) systems, real-time data processing capabilities, and user-friendly interfaces for clinicians.

References

[1] Global Facts: About Kidney Disease. Available at https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease/. Accessed 9 May 2024

[2] Ma F, Sun T, Liu L, Jing H (2020) Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. Future Generation Computer Systems, 111:17–26. 10.1016/j.future.2020.04.036

[3] National Kidney Foundation. Chronic Kidney Disease (CKD). Available at https://www.kidney.org/atoz/content/about-chronic-kidney-disease. Accessed 9 May 2024.

[4] Kalaiselvi, K., V. J. Sara, S.B. (2022). A Hybrid Filter Wrapper Embedded-Based Feature Selection for Selecting Important Attributes and Prediction of Chronic Kidney Disease. In: Ramu, A., Chee Onn, C., Sumithra, M. (eds) International Conference on Computing, Communication, Electrical and Biomedical Systems. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-86165-0_14

[5] Henry, B.M., Lippi, G. (2020) Chronic kidney disease is associated with severe coronavirus disease 2019 (COVID-19) infection. Int Urol Nephrol 52, 1193–1194. https://doi.org/10.1007/s11255-020-02451-9

[6] Estimated Glomerular Filtration Rate (eGFR) Available at https://www.kidney.org/atoz/content/gfr. Accessed 9 May 2024.

[7] Shehab M, Abualigah L, Shambour Q, Abu-Hashem MA, Shambour MKY, Alsalibi AI, Gandomi AH (2022). Machine learning in medical applications: A review of state-of-the-art methods. Computers in Biology and Medicine 145:105458. https://doi.org/10.1016/j.compbiomed.2022.105458.

[8] Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N (2017) Predictive analytics for chronic kidney disease using machine learning techniques. Management and Innovation Technology International Conference (MITicon), DOI: 10.1109/MITICON.2016.8025242.

[9] Bhaskar N, Suchetha M: An Approach for Analysis and Prediction of CKD using Deep Learning Architecture (2020). International Conference on Communication and Electronics Systems (ICCES), DOI: 10.1109/ICCES45898.2019.9002214.

[10] Senan EM, Al-Adhaileh MH, Alsaade FW, Aldhyani THH, Alqarni AA, Alsharif N, Uddin MI, Alahmadi AH, Jadhav ME, Alzahrani MY(2021). Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of Healthcare Engineering: 1–10. https://doi.org/10.1155/2021/1004767

[11] Janani J, Sathyaraj R (2021) Diagnosing Chronic Kidney Disease Using Hybrid Machine Learning Techniques. Turkish Journal of Computer and Mathematics Education, 12: 6383–6390.

[12] Chittora P, Chaurasia S, Chakrabarti P, Kumawat G, Chakrabarti T, Leonowicz Z, Jasinski M, Jasinski L, Gono R, Jasinska E, Bolshev V (2021) Prediction of Chronic Kidney Disease - A Machine Learning Perspective. IEEE Access, DOI: 10.1109/ACCESS.2021.3053763.

[13] Nikhila: Chronic Kidney Disease Prediction using Machine Learning Ensemble Algorithm (2021) International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), DOI: 10.1109/ICCCIS51004.2021.

[14] Singh V, Asari VK, Rajasekaran R (2022) A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease. Diagnostics, 12: 116. 10.3390/diagnostics12010116

[15] Abdel-Fattah MA, Othman NA, Goher N (2022) Predicting Chronic Kidney Disease Using Hybrid Machine Learning Based on Apache Spark. Computational Intelligence and Neuroscience: 1–12. https://doi.org/10.1155/2022/9898831

[16] Latha Busi RA, Meka JS, Reddy PVGDP (2023) A Hybrid Deep Learning Technique for Feature Selection and Classification of Chronic Kidney Disease. International Journal of Intelligent Engineering and Systems,16: 638–649.

[17] Ghosh, S.K., Khandoker, A.H (2014) Investigation on explainable machine learning models to predict chronic kidney diseases. Sci Rep 14, 3687. https://doi.org/10.1038/s41598-024-54375-4

[18] Rubini, L., Soundarapandian,P., and Eswaran,P. (2015). Chronic Kidney Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C5G020.

Authors' Profiles

Abhinav Shivhare received his Bachelor of Technology in Electrical Engineering with a specialization in Computer Science from Dayalbagh Educational Institute, Dayalbagh, Agra. Currently, he is working as a Software Engineer. His research interests include Artificial Intelligence and Machine Learning, with a focus on advancing practical applications and theoretical foundations within these domains.

A. Charan Kumari received her Ph.D from Dayalbagh Educational Institute, in collaboration with Indian Institute of Technology, Delhi, India, under their MOU. She is currently working as an Assistant Professor in the Electrical Engineering Department, Dayalbagh Educational Institute (Deemed to be University) Agra, India. Her current research interests include Machine Learning, Search Based Software Engineering, Evolutionary computation and soft computing techniques. She has published papers in journals and conferences of national and international repute. She has delivered an invited talk at 43rd CREST open workshop on Hyper-heuristics at University College London, London. She is a member of IEEE, Computer Society of India (CSI) and Systems Society of India (SSI).

K. Srinivas received a B.E. in Computer Science & Technology, an M.Tech. in Engineering Systems and a Ph.D. in Electrical Engineering. He is currently working as an Associate Professor in the Electrical Engineering Department, Faculty of Engineering, Dayalbagh Educational Institute (Deemed to be University) Agra, India. His research interests include AI Machine Learning, Deep Learning, Soft Computing, and Optimization using Metaheuristic Search Techniques, Search-Based Software Engineering, Mobile Telecommunication Networks, and Systems Engineering. He is an active researcher, guiding research, publishing papers in journals of national and international repute, and being involved in R&D projects. He is a member of IEEE and a life member of the Systems Society of India (SSI).

How to cite this paper: Abhinav Shivhare, A. Charan Kumari, K. Srinivas, "Predictive Modeling for Chronic Kidney Disease: A Comparative Analysis of Machine Learning Techniques with Explainable AI for Clinical Transparency", International Journal of Education and Management Engineering (IJEME), Vol.15, No.3, pp. 24-39, 2025. DOI:10.5815/ijeme.2025.03.03