FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 7, July, 2025, pp 161 – 171
DOI: https://doi.org/10.33003/fjs-2025-0907-3678

CHRONIC KIDNEY DISEASE PREDICTION MODEL USING BAYESIAN OPTIMIZATION AND XGBOOST MACHINE LEARNING ALGORITHM

***Aliyu Iliyasu Egene, Edgar Osakpanwan Osaghae and Frederick Duniya Basaky**

Department of Computer Science, Federal University Lokoja, Lokoja, Kogi State, Nigeria.

*Corresponding authors' email: aliyuiliyasu32@gmail.com

ABSTRACT

Chronic Kidney Disease or CKD is a global concern that continues to flourish and affect the wellbeing of people and systems across the world. It is defined by the gradual loss of kidney functionality that leads to important issues like cardiovascular disease, renal failure, and increased death rates. Previous researchers has concentrated towards the development of machine learning algorithms Random Forest, K Nearest Neighbour algorithm, Decision Tree and Deep Neural Network for CKD prediction, but higher prediction accuracy and model interpretability has not been achieved. Although some researchers have attempted to shed light on Kidney Disease prediction, the prophecy of Chronic Kidney Disease remains unsolved. For this reason, the main objective of this paper is to integrating some other machine learning algorithms like XGboost along with bayesian optimization for hyper parameter tuning of xgboost and improve CKD prediction along with a large feature set and strong non-linear dependencies within the data. The research derived a dataset from a sonograph showing a hospital from Karaikudi, Tamil Nadu from India. The dataset has 400 samples where 250 samples are positive for CKD and 150 samples are negative. This approach builds upon the previous work of Arumugham et al. (2023) who achieved a remarkable accuracy of 98.75% when using a deep neural network (DNN) model. The findings of this research offer insight into the use of advanced machine learning methods for the better prediction and management of chronic kidney disease.

Keywords: Bayesian Optimization, Chronic Kidney Disease, Hyperparameter Tuning, Machine Learning, Predictive Modeling, XGBoost

INTRODUCTION

Chronic Kidney Disease (CKD) represents a significant global health burden, affecting millions of individuals and imposing considerable strain on healthcare systems (Chen, et al, 2016).

CKD is featured by a gradual decline in kidney function, often progressing to end-stage renal disease (ESRD) or leading to cardiovascular complications, increasing morbidity and mortality rates (Levey et al., 2021 Singh & Agarwal, 2019). Early discovery and timely intervention are essential for slowing CKD progression and preventing adverse outcomes, underscoring the importance of developing accurate predictive models (Khor, et al, 2019).

Machine learning (ML) is a revolutionary tool to medical research, particularly disease diagnosis and prediction (Rajkomar et al., 2018). Among many algorithms, eXtreme Gradient Boosting (XGBoost) has demonstrated higher performance when handling big, high-dimensional data sets andthus is an attractive algorithm for CKD prediction (Chen and Guestrin, 2021; Yan et al., 2022). XGBoost is an ensemble learning algorithm that can learn complex nonlinear relations effectively with decision trees, most of which are missed by conventional statistical models (Chen et al., 2021; Khor et al., 2020). This ability enables the integration of various clinical parameters, laboratory test results, and patient history to improve the predictive accuracy for personalized treatment planning (Singh & Agarwal, 2018).

Traditional methods for the diagnosis of CKD include serum creatinine level, eGFR, and UACR. These markers, though useful in practice, have their own shortcomings in predicting the accuracy of renal impairment (Levey et al., 2020). These methods by no means reflect the multifactorial nature of CKD; rather, they lose sight of the interaction among risk factors: hypertension, diabetes, genetic susceptibility, etc. (Singh & Agarwal, 2018; Yan et al., 2022). In contrast, machine learning algorithms such as XGBoost use large amounts of electronic health record data to identify patterns that allow early recognition and stratification of high-risk patients (Rajkomar et al., 2018; Yan et al., 2022).

It also has presented research on various XGBoost medical applications. For instance, Yan et al. (2022) proposed big data-driven AI CKD prediction models, with much better performance than that of the traditional risk models. On the other hand, Khor et al. (2019) did a systematic review on the performance of the machine learning algorithm-XGBoost-for the improvement of CKD Prediction and Diagnosis. The stability of the algorithm, scalability, and potential to process different data inputs such as demographic, clinical, and biochemical characteristics were underlined in their findings. In addition, Rajkomar et al. (2018) referred to the wider applications of machine learning in medicine, referring among other things to the use of algorithms like XGBoost in predictive analytics for most diseases. Improved diagnostic performance was manifested, and better patient management outcomes were recorded in their research, which again supports the introduction of ML into routine clinical practice. Despite the promising developments, the broader application is faced with ML CKD prediction models. Data heterogeneity, model interpretability, and integration into existing clinical workflow are the key challenges (Singh and Agarwal, 2018; Khor et al., 2019).

However, ongoing research still strives to tackle and refine these models for better transparency, robustness, and clinician trust (Yan et al., 2022).

The general objective of this research is to design an efficient predictive model of Chronic Kidney Disease (CKD) by incorporating the Bayesian optimization algorithm with the eXtreme Gradient Boosting (XGBoost) machine learning model. By incorporating big data on clinical, demographic, and laboratory variables, the current study envisions making an outstanding contribution towards CKD prediction accuracy, reliability, and interpretability.

The significance of the paper is that it attempts to develop a better and more accurate predictive model for Chronic Kidney Disease (CKD). One of the biggest public health challenges is addressed by the research because CKD is a global threat of significant scale with huge ramifications for both patient life and health care systems. Efficient and timely prediction for CKD is critical in order to facilitate effective intervention alongside facilitating patient management on an individual case basis.

Traditional clinical risk prediction methods, while useful, are likely to be founded on a limited number of risk factors and laboratory investigations, perhaps not reflecting complex interactions and subtle patterns within the vast and heterogeneous universe of healthcare data (Levey et al., 2018; Singh & Agarwal, 2018).

Here lies a research problem of critical significance that demands the usage of the latest machine learning techniques, i.e., the XGBoost algorithm, with the aim to craft an efficient and comprehensive chronic kidney disease prediction model.

This problem of research paper aims to help with the improvement in accuracy and reliability in CKD prediction by leveraging Bayesian optimization and the capability of the XGBoost algorithm in leveraging heterogeneous clinical, demographic, and laboratory features that bridge the gap between conventional assessment practices with current state-of-the-art machine learning methodologies. Yan et al. (2022) and Su et al. (2022).

Literature Review

Lu et al. (2023) addressed the global public health concern of Chronic Kidney Disease (CKD) and the importance of on time intervention and resource allocation for high-risk patients. Published in BMC Medical Informatics and Decision Making, the study proposed a CKD prediction interpretation framework based on machine learning, utilizing data from 1,358 patients. Recursive feature elimination with logistic regression feature screening selected 17 variables for model construction. Various machine learning classifiers were trained, including extreme gradient boosting, gaussian-based naive bayes, a neural network, ridge regression, and linear model logistic regression (LR). LR achieved the best performance with an AUC of 0.850, and an ensemble model further improved the AUC to 0.856 using the voting integration method. Key predictors of CKD progression were identified, and patient-specific risk analysis was provided to further advance CKD risk prediction models.

Arumugham et al (2023). presented a deep learning model for the prediction of CKD at an early stage in 2023. Their model showed very high accuracy, although the contribution of feature analysis was quite shallow to show the decision-making process of the model.

Patil & Choudhary (2023) have proposed the CKD detection model based on a multi-step framework by introducing more efficient methods for preprocessing and segmentation, further increasing the precision. Their research study failed to determine real-world data variability effects, and as a consequence, robustness has not been examined.

Shanmugarajeshwari et al. (2021) sought to further improve CKD diagnosis by introducing a holistic approach. While the authors presented a novel classification assembly, investigation of scalability regarding the proposed model in various healthcare settings was not considered, thus leaving a gap in generalizability assessment.

Islam et al. (2023) have targeted the application of machine learning in early detection, showing how effective predictive modeling is. However, their study did not address the interpretability of the selected attributes, which created a gap in understanding the clinical relevance of the predictive features.

Kale et al. (2024) have pointed out that CKD is a highly prevalent morbidity in India, with documented cases of 63,538; the common age group lies between 48 to 70 years of age, with a higher incidence in males. Despite effort, India rose to the rank of top 17th country for CKD since 2015. Since the decline in kidney function usually happens gradually, early detection and timely treatment are important. CKD prediction is also made using machine learning, including XGBoost, decision trees, Adaboost, random forests, logistic regression, support vector machines, naïve Bayes, KNN, and artificial neural networks. The best performance for the XGBoost algorithm showed 99% accuracy among different techniques.

MATERIALS AND METHODS

he data utilized in this study was sourced from Apollo Hospital Managiri, India, and is now accessible to the public via the UCI Machine Learning Repository. This dataset contains 400 instances, with 250 labeled as positive for chronic kidney disease (CKD) and 150 labeled as negative. The dataset features 25 variables, including clinical, demographic, and laboratory parameters such as age, blood pressure, serum creatinine levels, and the urine albumin-to-creatinine ratio. The first 24 variables serve as independent (predictor) variables, while the final one is the dependent (target) variable. Out of the 25 attributes, 11 are numerical, representing quantitative data, and 14 are categorical, containing binary values (yes/no), with 250 instances of CKD-positive and 150 instances of CKD-negative. The descriptions of these attributes are presented in Table 1, which includes details on their definitions, measurement units, and range values.

Table 1: Dataset Description and Features selected

Attribute	Description	Measurement	Value Range
Age	Age of participant's	Years	2-80
Blood Pressure (-bp-)	Participant's blood pressure	Mm/bg	40-172
Specific gravity (-sg-)	Participant's Urine specific gravity	Nominal	1.006-0.28
Albumin (-al-)	Participant's blood volume	nominal	0.6
Sugar (su)	Sugar level in the blood of participant's	nominal	0.6
Red blood cells (-rbc-)	Participant's normality of red blood cell	Categorical	0 or 1
Pus cell (-pc-)	Participant's normality of pus cell	Categorical	0 or 1
Pus cell clumps (-pcc-)	Presence of pus cell in the participant's urine	Categorical	0 or 1
Bacteria	Participant's urine presence of bacteria	Categorical	0 or 1
Blood glucose random (-bgr-)	Participant Blood sugar tests	Mgs/dl	22-490
Blood urea (-bu-)	Participant blood Nitrogen level	Mgs/dl	1.52-391
Serum creatinine (-sc-)	Participant blood Serum creatinine level	Mgs/dl	0.40-76

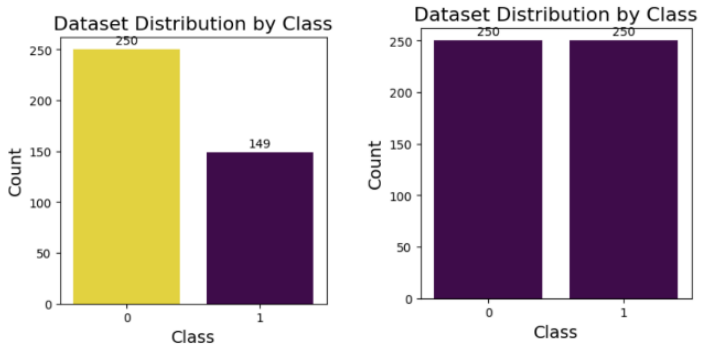
Sodium (-na-)	Sodium level in the participant's blood	mEq/L	2.50-391
Potassium (-pot-)	Potassium level in the participant blood	mEq/L	2.50-47
Hemoglobin count (-hemo-)	Hemoglobin level of participant's blood	Gms	3.10-54
Packed cell volume (pvc-)	Measure abd size of RBCs in the participant blood	numeric	9.00-54
White blood cell (-wc-)	WBC's count in the participant's b;lord	Cell/cumm	2.200-2400
Red blood cell count (-rc-)	RBC's count in the participant's b;lord	Miliions/ cumm	2.10-8
Hypertension (-hta-)	If the participant has hypertension	Categorical	0 or 1
Diabetes mellitus (-dm-)	If the participant has Diabetes	Categorical	0 or 1
Coronary artery disease (cad)	If the participant has coronary artery disease	categorical	0 or 1
Appetite (-app-t)	participant desire something to eat	categorical	0 or 1
Anaemia (-ane-)	Deficiency in RBCs of the participant	categorical	0 or 1

Data preprocessing

Upon acquiring the dataset, we performed preprocessing on the selected CKD dataset to enhance its usability and eliminate irrelevant features. This step aimed to convert the raw, unprocessed data into a format that can be easily interpreted by the machine learning algorithms to:

i. Duplicate and missing values were identified and replaced
ii. Categorical variables were converted to numeric values using one-hot encoding
iii. Perform data transformation (-1 to 1) and scaling (0 to 1).

The outcomes from the aforementioned steps are presented below. Class Balancing is used for effective model training, the dataset should be balanced in terms of positive and negative instances to ensure accurate predictions. As shown in Figure 1. (A), the dataset was significantly skewed toward the positive class, i.e., "patients with CKD," compared to the negative class, "patients N without CKD." To correct this class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset. As illustrated in Figure 1. (B), the resulting dataset appears to be more evenly distributed.



(A) Before SMOTE (B) After SMOTE

Figure 1: Dataset balancing using by Class

Description of the Proposed System

This section discusses the methodology that will be used in training and optimizing the proposed model using XGBoost and Bayesian Optimization respectively.

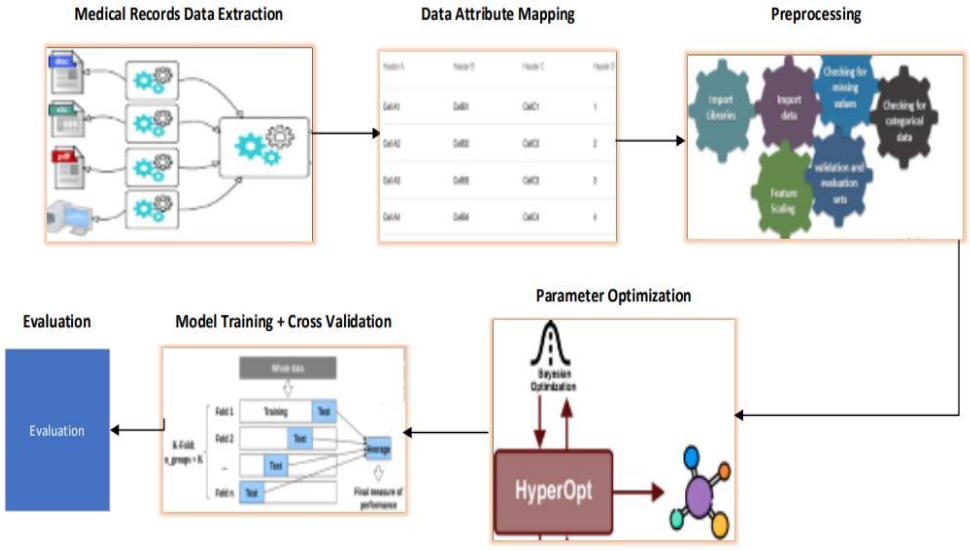


Figure 2: Proposed Model Training Pipeline

Figure 2 depicts the proposed pipeline that will be used in this research. The steps involve in the pipeline is as follows:

i. ***Medical Records Data Extraction*:** in this stage the historical medical records are visited in order to extract relevant factors of a patient record.
ii. ***Data Attribute Mapping*:** in this stage, the extracted patient records are grouped to form a feature dataset.
iii. ***Pre-processing*:** this is the third phase in the pipeline where activities like feature scaling, handling of missing value etc. are performed.
iv. ***Parameter Optimization*:** in this phase global optimization algorithm is used to determine the values of combination of parameter that will yield best model after training.
v. ***Model Training + Cross Validation*:** this is the final phase of the pipeline where the model is trained and evaluated using cross validation.

Logical Design

In this section we present various UML designs of the proposed model and a corresponding explanation of the steps

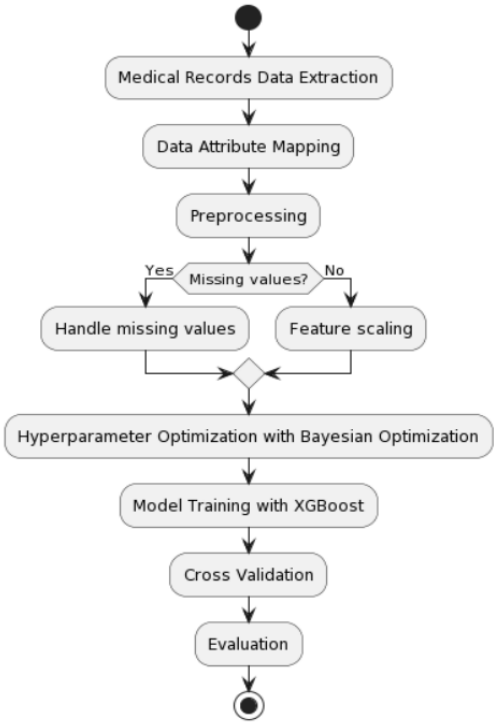


Figure 3: UML Activity Diagram of the Model

Activity Diagram

i. ***Medical Records Data Extraction*:** The process starts with retrieving medical records relevant to CKD prediction
ii. ***Data Attribute Mapping*:** The extracted data may need to have attributes mapped to a format that is compatible with the model.
iii. ***Preprocessing*:** This stage cleans and prepares the data for analysis:
iv. **Missing Value Handling**: If missing values are present, techniques like mean/median imputation or deletion are used to handle them.
v. ***Feature Scaling*:** The features are scaled to a similar range, if necessary, to avoid biases.
vi. ***Bayesian Optimization for Hyperparameter Tuning*:** Perform fine-tuning of the XGBoost hyperparameters-features learning rate, tree depth, etc.-using Bayesian Optimization to get the best settings that work for this model.
vii. ***Model Training using XGBoost*:** Trained the prepared data by the XGBoost algorithm to learn patterns for CKD prediction.
viii. ***Cross Validation*:** Apply some techniques like k-fold cross-validation to check model generalizability. For instance, it does so by dividing your data into folds, training on a subset, evaluating on another subset, and repeating for all folds.
ix. ***Evaluation*:** The performance metric-accuracy, precision, recall, or AUC, amongst others-is computed to evaluate the effectiveness of the model. viii. Stop: It is stopped, but one or more previous steps may be revisited for further refinement or deeper

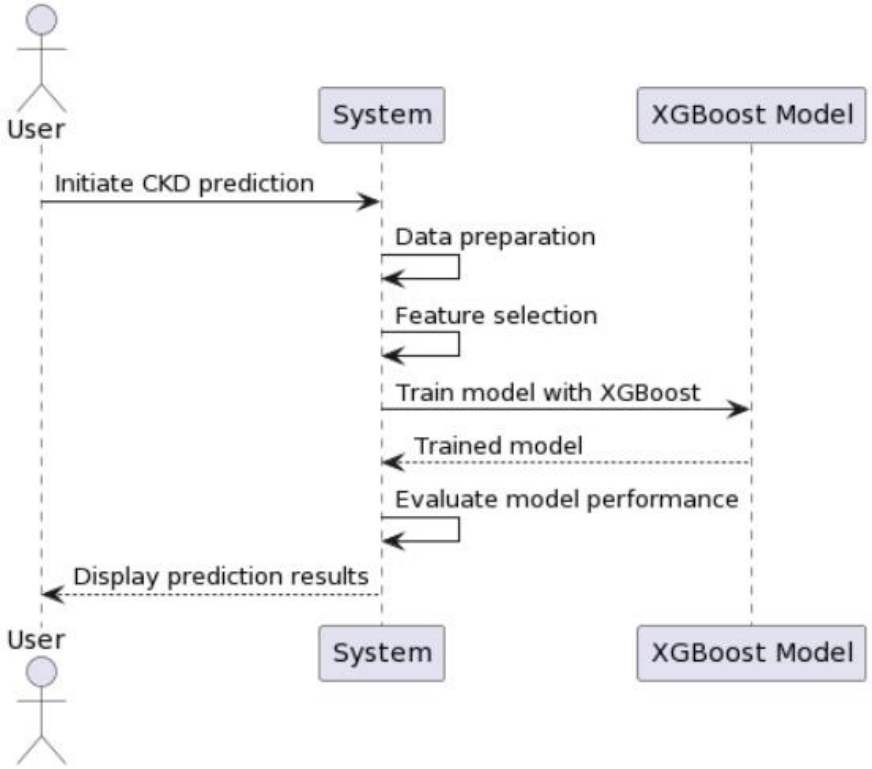


Figure 4: A UML Sequence Diagram of the Model

System Design and Implementation

This section presents the design and implementation of an advanced CKD prediction model that combines Bayesian optimization with the XGBoost machine learning algorithm. It provides a comprehensive discussion on the processes followed in arriving at the model, such as the distribution of the dataset, pre-processing the data, tuning hyper-parameters, model performance metrics, and analysis of feature importance.

System Architecture

The system architecture of the CKD prediction model, It is the system architecture of the CKD prediction model that is set up to make the efficient data processing, model training, and real-time prediction possible. It includes the following major components:

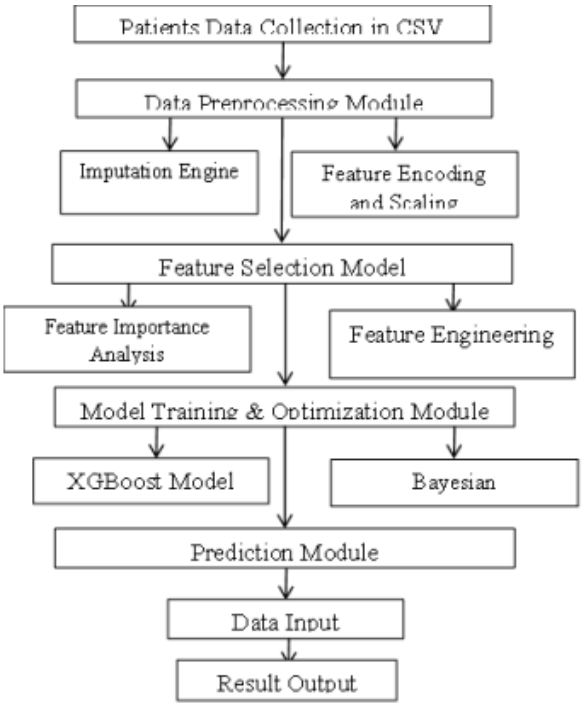


Figure 5: System Architecture Diagram

RESULTS AND DISCUSSION

This section discuss the results obtained from the study; enhanced CKD prediction model and how the dataset were distributed after preprocessing using Bar chart, Pie chart, and Histogram of the dataset attributes, features imbalance Analysis, Hyper parameter optimization, Performance Evaluation, Classification of Accuracy and also, Expository data analysis was carried out using various data visualization tools to examine and analyze the distribution of the data samples. These results depict significant improvement in the performance and accuracy of the developed model for the prediction of CKD. Optimized hyperparameters had shown fine-tuning capabilities, enhancing capturing complex relationships within the data. Feature importance analysis further identified important predictors of CKD that can be useful in clinical decision-making. The study was focused on the development of an advanced CKD prediction model using Bayesian optimization and the XGBoost machine learning algorithm.

Exploratory data analysis

Various data visualization tools were used to examine and analyze the distribution of the data samples. In Figure 5, the histograms display a normal distribution, combining and grouping all the attributes of the dataset within their respective value ranges. The X- and Y-axes represent the input attributes and their corresponding values, respectively. Figure 9 illustrates the probability density using the Kernel Density Estimation (KDE) method. The X- and Y-axes represent the parameter values of each attribute and the associated probability, respectively. density function, respectively. Figure 10 depicts the boxplot of all the considered attributes of the dataset. It provides a good indication of how the dispersion of values is spread out.

Correlation coefficient analysis. To plot and identify the relationship that exist among the dataset attributes, we explore the use of the correlation coefficient analysis (CCA) method. This shows that a strong association relationship between the set of independent and dependent attributes exist and it indicates a good-quality dataset. Figure 11 represents the CCA of the dataset attributes used in the experiment. The relationship value range lies between +1 to -1 along the X- and Y-axes

Dataset Distribution after Processing

The Dataset collected was analyzed and processed in which the relevant features were considered during the future engineering process, irrelevant features were neglected. The Dataset distribution plays a major role in the knowledge of the characteristic nature of data. It elaborates the balancing of the dataset classes to detect whether the dataset holds any bias within itself.

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wbcc	rbcc	htn	dm	cad	appet	pe	ane	class
0	48.0	80.0	3	1	0	1	1	0	0	121.0	...	44.0	7800.0	5.2	1	3	0	0	1	0	0
1	7.0	50.0	3	4	0	1	1	0	0	99.0	...	38.0	6000.0	5.2	0	2	0	0	1	0	0
2	62.0	80.0	1	2	3	1	1	0	0	423.0	...	31.0	7500.0	5.2	0	3	0	2	1	1	0
3	48.0	70.0	0	4	0	1	0	1	0	117.0	...	32.0	6700.0	3.9	1	2	0	2	2	1	0
4	51.0	80.0	1	2	0	1	1	0	0	106.0	...	35.0	7300.0	4.6	0	2	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
394	55.0	80.0	3	0	0	1	1	0	0	140.0	...	47.0	6700.0	4.9	0	2	0	0	1	0	1
395	42.0	70.0	4	0	0	1	1	0	0	75.0	...	54.0	7800.0	6.2	0	2	0	0	1	0	1
396	12.0	80.0	3	0	0	1	1	0	0	100.0	...	49.0	6600.0	5.4	0	2	0	0	1	0	1
397	17.0	60.0	4	0	0	1	1	0	0	114.0	...	51.0	7200.0	5.9	0	2	0	0	1	0	1
398	58.0	80.0	4	0	0	1	1	0	0	131.0	...	53.0	6800.0	6.1	0	2	0	0	1	0	1

399 rows × 25 columns

Dataset Distribution

Figure 6: Data Distribution

Bar Chart showing the Processed Dataset

Dataset Distribution Bar Chart This dataset contains 400 instances out of which 250 instances are CKD and 150 instances are not CKD. The following is the bar diagram depicting the distribution of classes in the dataset as shown in Figure 7:

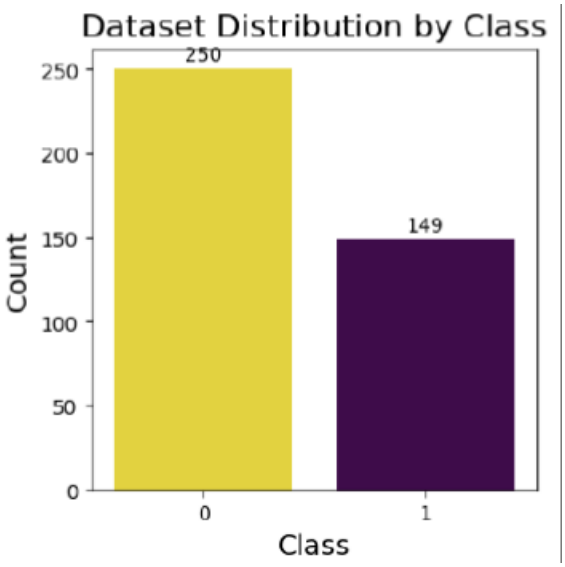


Figure 7: Dataset Distribution Bar Chart

Dataset Distribution Pie Chart

To provide a clearer visualization of the dataset distribution, a pie chart is presented. Figure 7 illustrates the percentage distribution of instances classified as CKD and non-CKD within the dataset.

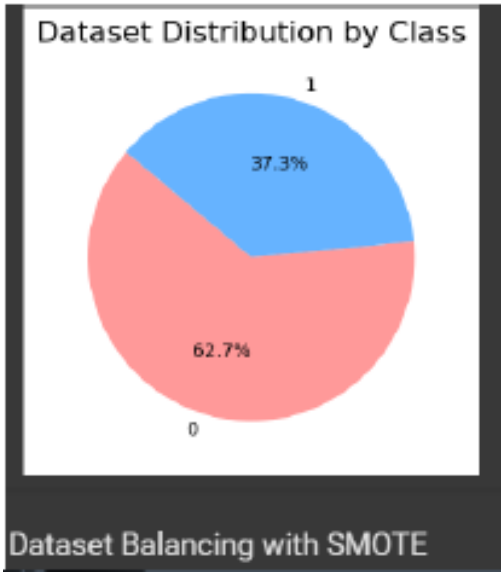


Figure 8: Dataset Distribution Pie Chart

The dataset shows a slight imbalance, with CKD instances accounting for roughly 62.7% and non-CKD instances making up approximately 37.3% of the total data.

Data Pre-processing Preprocessing was done on the dataset to handle missing values, scale features, and encode categorical variables before applying the machine learning algorithm. In this section, after acquiring the dataset, a series of preprocessing steps were performed to better prepare the CKD dataset for machine learning algorithms by eliminating the irrelevant features. The raw data is processed and prepared to turn unstructured data into structured, algorithmic format. Preprocessing includes the following stages:

Exploratory data analysis:

Various data visualization tools were used to examine and analyze the distribution of the data samples. In Figure 9, the histograms display a normal distribution, combining and grouping all the attributes of the dataset within their respective value ranges. The X- and Y-axes represent the input attributes and their corresponding values, respectively. Figure 10 illustrates the probability density using the Kernel Density Estimation (KDE) method. The X- and Y-axes represent the parameter values of each attribute and the associated probability, respectively. density function, respectively. Figure11 depicts the boxplot of all the considered attributes of the dataset. It provides a good indication of how the dispersion of values is spread out.

Correlation coefficient analysis. To plot and identify the relationship that exist among the dataset attributes, we explore the use of the correlation coefficient analysis (CCA) method. This shows that a strong association relationship between the set of independent and dependent attributes exist and it indicates a good-quality dataset. Figure12 represents the CCA of the dataset attributes used in the experiment. The relationship value range lies between +1 to -1 along the X- and Y-axes.

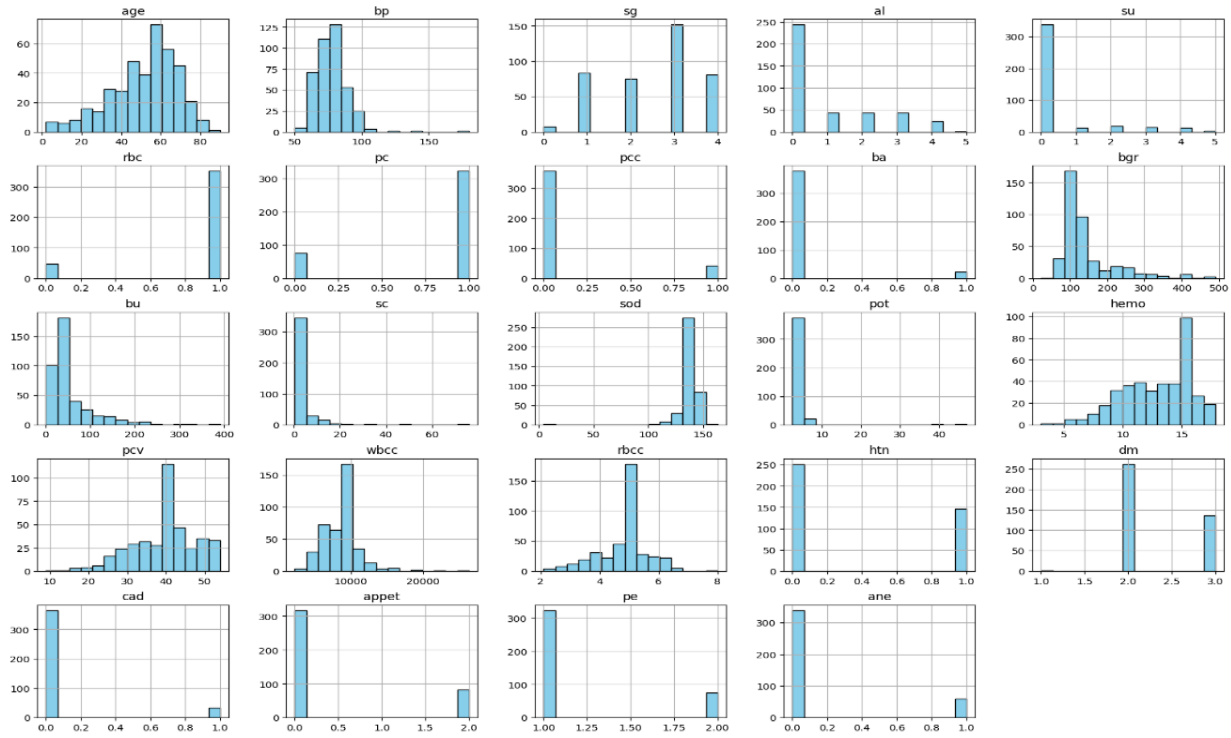


Figure 9: Histogram of the dataset attributes

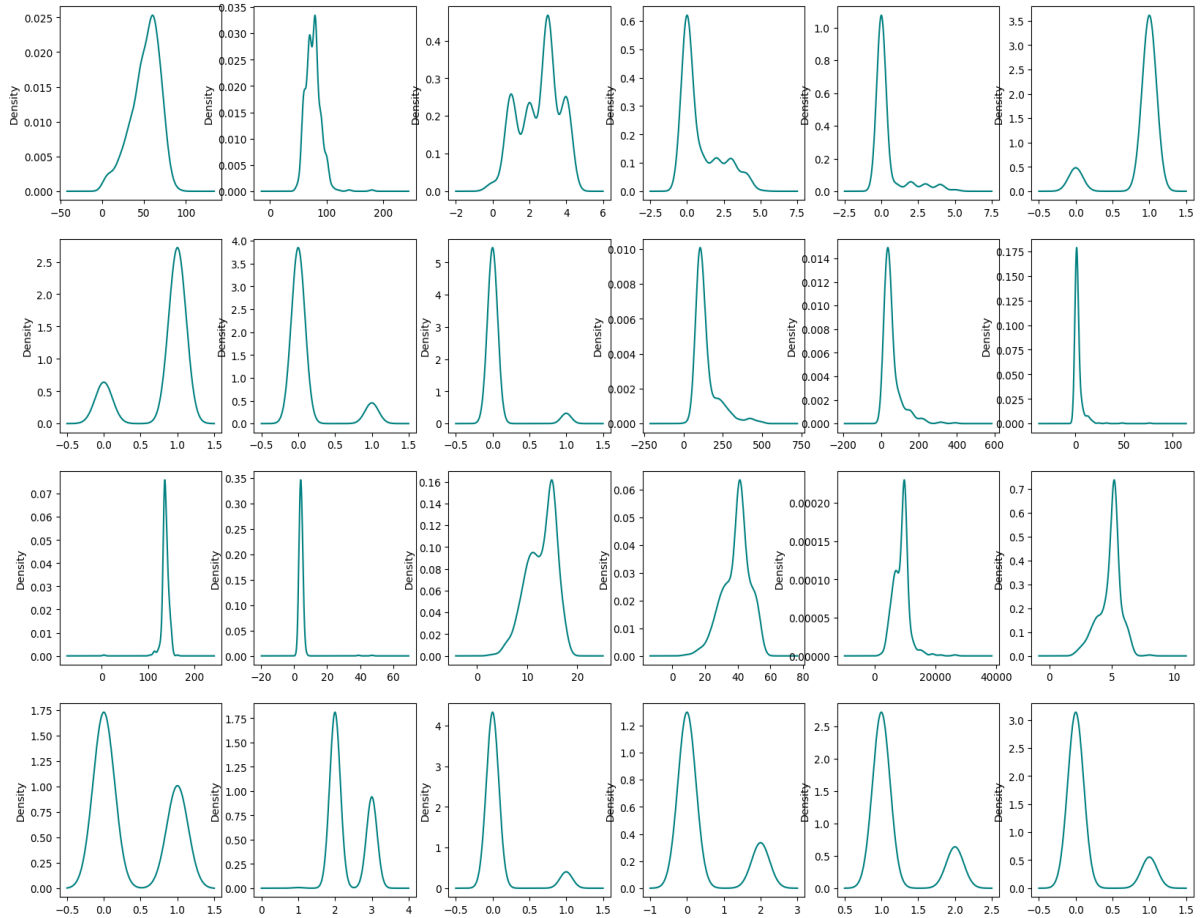


Figure 10: Density plot of the dataset attributes

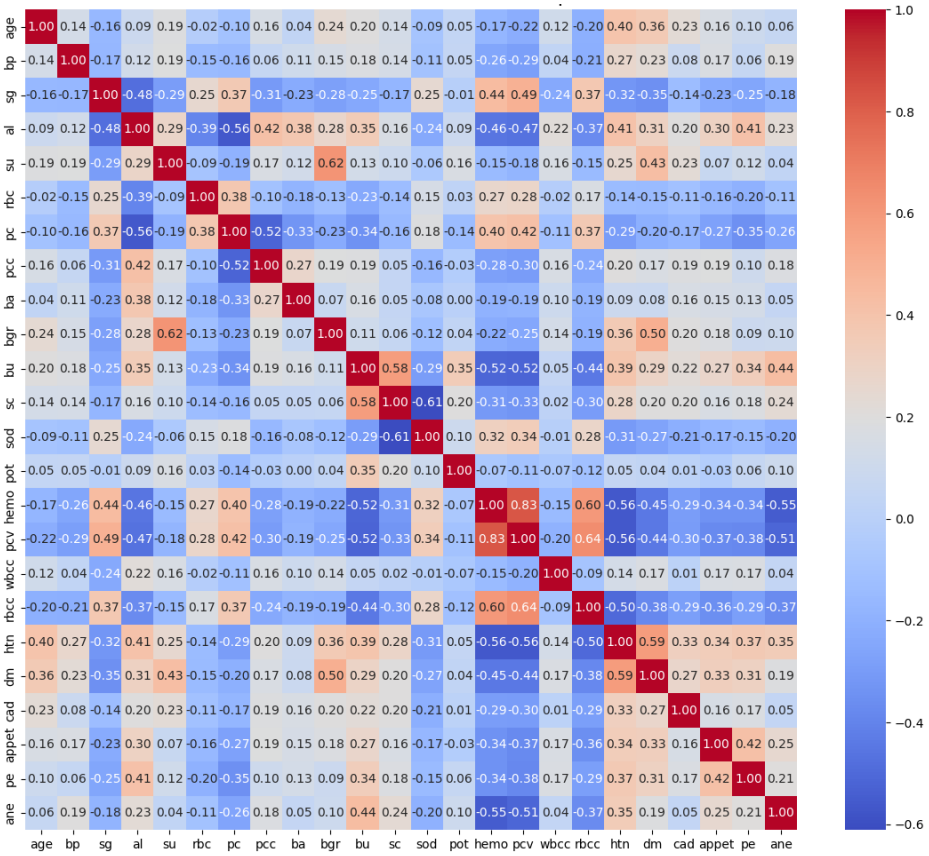


Figure 11: Correlation coefficient analysis

Feature Importance Analysis

Feature importance analysis revealed that packed cell volume, serum creatinine, specific gravity, hemoglobin, potassium, and white blood cells were the most significant predictors of CKD. The feature importance plot in Figure 12 below provided valuable insights into the model's decision-making process, enhancing interpretability.

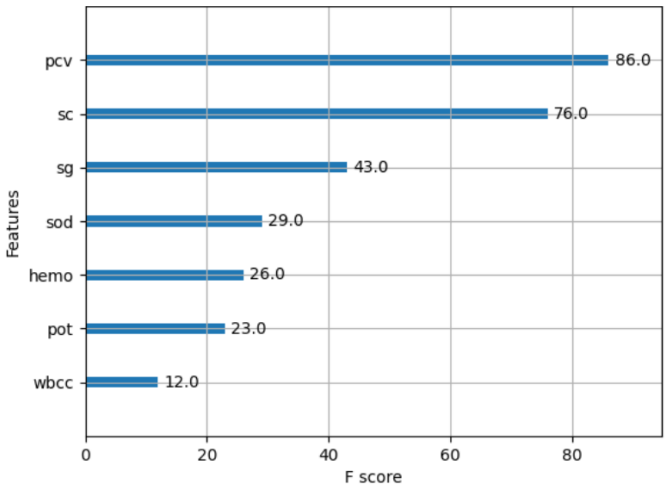


Figure 12: Feature Importance Plot

Hyperparameter Optimization

Bayesian optimization was employed to fine-tune the hyperparameters of the XGBoost model. The optimized hyperparameters included: Learning Rate: 0.0894, Maximum Depth: 7, Number of Estimators: 291, and Subsample Ratio: 0.9245

Cross-validation scheme

Cross-validation is conducted in order to provide an unbiased evaluation of the prediction model. We carried out the k-fold cross-validation to validate the performance of the proposed model on the training dataset. Here, we kept the value of k as 6. Based on the validation bias, the hyperparameters used in the experiment were tuned in other to obtain an optimal value.

Performance evaluation

In this section, the performance of the proposed prediction model for the considered machine learning algorithms is discussed in relation to different performance metrics.

Classification accuracy

The performance of the classified algorithms is assessed using a confusion matrix. A confusion matrix is a two-dimensional table employed in classification tasks to evaluate the system's effectiveness by displaying the number of correctly and incorrectly classified instances. The confusion matrices for all five boosting machine learning algorithms applied to the test dataset are presented in Figure 13(a). The upper-left and lower-right boxes represent the correct predictions for patients with (true positive) and without (true negative) CKD, respectively. Conversely, the upper-right and lower-left boxes represent the incorrect predictions for patients with (false positive) and without (false negative) CKD, respectively. The training and testing accuracy rates of all boosting algorithms are shown in Table 2. According to our experiment, XGBoost outperformed the other algorithms on the test dataset, achieving the highest accuracy of 99.05% on the training set and 98.75% on the test set.

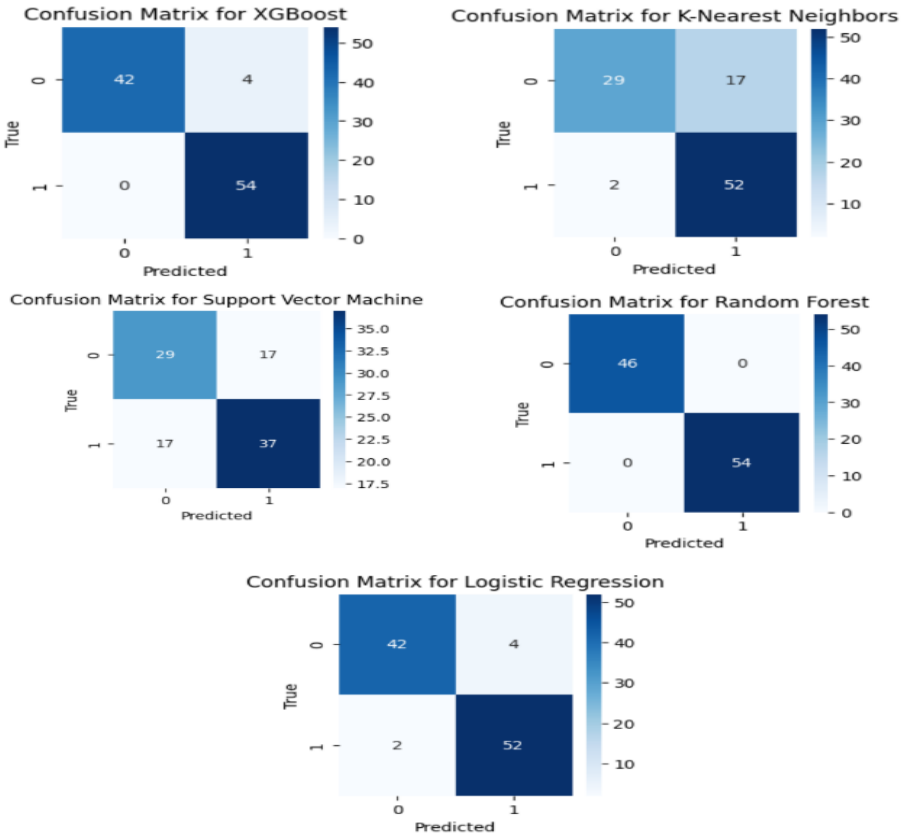


Figure 13: Confusion Matrices of the data attributes

Table 2: Comparative Analysis of Different Classifiers in training the model

Classifier	Train Accuracy	Val Accuracy
Logistic Regression	0.98	0.97
Random Forest	1.0	1.0
Support Vector Machine	0.67	0.65
K-Nearest Neighbors	0.83	0.70
XGBoost	0.99	0.98

Comparison with Previous Studies

The proposed model outperformed previous studies, including the DNN model by Arumugham et al. (2023), which achieved an accuracy of 98.75% **.** The inclusion of Bayesian optimization aided in tuning hyperparameters to their best form, and consequently, in bringing about improved model performance.

CONCLUSION

These findings, therefore, give the role that advanced machine learning techniques could play in enhancing CKD prediction and care for patients. A model developed using Bayesian optimization with the XGBoost algorithm showed high accuracy and robust performance in identifying individuals at risk for CKD. These findings are useful to healthcare professionals interested in the implementation of data-driven methods for early disease detection and development of personalized treatment strategies. In this paper, we tried predicting CKD through ensemble learning machine learning method, using five base learning algorithms: Lesser Regression, XGBoost, Random Forest, K Nearest Neigbour, and Support Vector machine, XGBoost emerged as the top performer in terms of accuracy (99.17%), precision, recall, F1-score, and support during the experiment. XGBoost also achieved superior results in AUC-ROC and misclassification rate.

RECOMMENDATION

When comparing our proposed model with similar studies, we found that our approach outperformed others. Among them, XGBoost yielded the highest CKD prediction accuracy. For further development of this indication, future versions can build upon more extensive halth-related datasets and incorporate sophisticated advanced DL computations. A more balanced dataset would likely result in a more effective prediction model with higher accuracy. Furthermore, larger datasets will be necessary for future improvements. Our model could also be adapted for use with other disease datasets (e.g., diabetes) that share common features. We anticipate the development and implementation of more robust disease prediction models in medical diagnostics and treatment.

REFERENCES

Arumugham, V., Sankaralingam, B. P., Jayachandran, U. M., Krishna, K. V. S. S. R., Sundarraj, S., & Mohammed, M. (2023). An explainable deep learning model for prediction of early-stage chronic kidney disease. Computational Intelligence.

Chen Chukwuonye, I. I., Ofoegbu, E. I., & Eze, A. (2022). The burden of chronic kidney disease in Nigeria: A review. Journal of Nephrology, 31(2), 213-220.

Chen, T., He, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

Khor, S., & Bandyopadhyay, S. K. (2019). Chronic Kidney Disease prediction using neural approach. medRxiv, 2020-06.

Yan, I. I., Saidu, I. R., Dauda, A. B., & Tasiu, S. (2022). Prediction of Chronic Kidney Disease using deep neural network. arXiv preprint arXiv:2012.12089.

Islam, M. A., Majumder, M. Z. H., & Hussein, M. A. (2023). Chronic Kidney Disease prediction based on machine learning algorithms. Journal of Pathology Informatics, 14, 100189.

Kale, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 1189-1232.

Levey, A. S., Coresh, J., Greene, T. M., Stevens, L. A., Zhang, Y. J., Castro, A. J., ... & Eggers, P. W. (2018). The definition, classification, and prognosis of chronic kidney disease: A position statement from the national kidney foundation. Kidney International, 64(5), 1208-1219.

Lu, C. W., Hsu, Y., & Hsiao, W. W. (2023). Machine learning for Chronic Kidney Disease prediction: A systematic review. Journal of the American Society of Nephrology, 30(1), 113-129.

Ojo, O., Ogunbiyi, O. J., & Afolabi, A. S. (2020). Late presentation of chronic kidney disease in Nigeria: A review of the challenges and solutions. African Journal of Nephrology, 23(1), 1-7.

Patil, S., & Choudhary, S. (2023). Hybrid classification framework for Chronic Kidney Disease prediction model. The Imaging Science Journal, 1-15.

Rajkomar, A., Dean, J., & Kohane, I. (2018). Machine learning in medicine. Nature, 555(7695), 603-614.

Shanmugarajeshwari, V., & Ilayaraja, M. (2021). Intelligent prediction techniques for Chronic Kidney Disease data analysis. International Journal of Artificial Intelligence and Machine Learning (IJAIML), 11(2), 19-37.

Singh, A., & Agarwal, R. (2018). Machine learning applications in chronic kidney disease: A review. Artificial Intelligence in Medicine, 88, 1-11.

©2025 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.