IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141005

Chronic Kidney Disease Prediction Using Machine Learning

Vinita Sisodiya [1], Manoj V. Nikum*[2]

Student of MCA, SJRIT Dondaicha, KBC NMU Jalgaon, Maharashtra[1]

Assi. Prof. and HOD, MCA Department, SJRIT Dondaicha, Jalgaon, Maharashtra*[2]

Abstract: Chronic Kidney Disease (CKD) is a progressive medical condition characterized by the gradual loss of kidney function over time.This paper presents a machine learning–based approach to predict CKD using clinical . The study focuses on data preprocessing techniques, including handling missing values, feature scaling, and encoding categorical variables, to enhance model accuracy and reliability. Several machine learning algorithms, such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine, are implemented and evaluated using performance metrics like accuracy, precision, recall, and F1-score. Among these models, the Random Forest classifier demonstrates superior predictive performance, achieving high accuracy and robust generalization across test data. The experimental results suggest that the integration of machine learning techniques in healthcare can significantly assist medical practitioners in early CKD detection, risk stratification, and informed clinical decision-making. Furthermore, this study highlights the potential of artificial intelligence to transform traditional diagnostic procedures into data-driven, automated systems for improved healthcare delivery. and speed of disease diagnosis. The outcomes of this study highlight the potential of artificial intelligence (AI) in supporting data-driven healthcare solutions and enabling early intervention strategies for patients at risk of CKD

Keywords: Chronic Kidney Disease (CKD) Kidney function prediction Machine Learning (ML) Medical data analysis Disease classification Health informatics.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a significant public health concern that affects millions of individuals worldwide, with an estimated prevalence of 10–15% in adults. The disease often progresses silently, with patients remaining asymptomatic until advanced stages, making early detection crucial. When kidney function declines, toxins accumulate in the blood, leading to severe health complications such as cardiovascular diseases, anemia, and bone disorders. According to the World Health Organization (WHO), CKD has emerged as a major global health concern, affecting approximately 10% of the world's population. The increasing prevalence of CKD is largely attributed to factors such as diabetes, hypertension, obesity, and unhealthy lifestyle habits.

Early detection of CKD is crucial to prevent its progression to End-Stage Renal Disease (ESRD), where patients often require dialysis or kidney transplantation. These methods can be time-consuming and may not always detect CKD in its early stages. As a result, many patients remain undiagnosed until the disease has advanced significantly, reducing the effectiveness of available treatments.

In recent years, machine learning (ML) has gained significant attention in the medical domain for its ability to analyze complex data and uncover hidden patterns that may not be apparent through conventional statistical methods.
ML models can learn from historical clinical data to predict disease risk, assist in diagnosis, and support decision-

DOI: 10.17148/IJARCCE.2025.141005

making processes. By integrating ML techniques into healthcare systems, early CKD prediction becomes feasible, allowing clinicians to intervene before irreversible kidney damage occurs.

This study aims to develop and evaluate machine learning models for predicting CKD using the UCI Chronic Kidney Disease dataset, which contains various clinical and laboratory attributes of patients. Data preprocessing techniques such as handling missing values, feature encoding, and normalization are employed to prepare the dataset for training. Multiple ML classifiers—including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine—are implemented and compared based on performance metrics such as accuracy, precision, recall, and F1-score.

The primary objective of this research is to identify the most effective model for CKD prediction and to demonstrate how machine learning can enhance the efficiency, accuracy,

II. LITERATURE REVIEW

The prediction and diagnosis of Chronic Kidney Disease (CKD) have been extensively studied using data-driven and machine learning (ML) techniques. Over the past decade, numerous researchers have explored various algorithms and methodologies to enhance the accuracy and reliability of CKD prediction models. This section reviews significant contributions in this field and highlights the evolution of ML-based approaches for CKD detection and prognosis.

Early Studies and Statistical Models:
Traditional approaches to CKD diagnosis primarily relied on clinical expertise and basic statistical techniques such as regression analysis. These models used parameters like serum creatinine, blood urea, glucose level, and blood pressure to assess kidney function. However, such linear models were often limited in handling non-linear relationships and complex feature interactions present in medical data. As a result, researchers began exploring machine learning techniques to overcome these limitations and achieve higher diagnostic accuracy.

Machine Learning-Based Approaches:
Machine learning methods have shown promising results in the medical domain, particularly for disease prediction. Patil et al. (2019) utilized Decision Tree and Random Forest algorithms to classify CKD patients using the UCI dataset, achieving an accuracy of 98%. Similarly, Joshi and Dhanalakshmi (2020) employed Support Vector Machines (SVM) and Naïve Bayes classifiers for CKD prediction and reported that SVM outperformed other models in terms of precision and recall. In another study, Mishra et al. (2021) implemented ensemble methods combining multiple classifiers, which demonstrated improved robustness and reduced overfitting compared to single-model approaches.

Data Preprocessing and Feature Engineering:
Researchers have emphasized the importance of data preprocessing for improving model performance. CKD datasets often contain missing values, inconsistent entries, and categorical attributes. Techniques such as mean imputation, label encoding, and normalization are widely used to standardize data. Kumar et al. (2021) highlighted that effective data cleaning and feature selection can significantly enhance classifier accuracy and reduce computational complexity. Feature importance analysis, particularly with Random Forest, has been beneficial in identifying the most influential medical attributes, such as blood pressure, albumin, and serum creatinine levels.

Deep Learning and Hybrid Models:
Recent advancements have introduced deep learning and hybrid approaches for CKD prediction. Singh et al. (2022) proposed a deep neural network (DNN) model that achieved higher performance compared to traditional ML classifiers. Hybrid models that integrate multiple algorithms—such as combining SVM with Random Forest or Decision Tree with Gradient Boosting—have also been explored to achieve better generalization and interpretability. These methods leverage both the predictive strength of ensemble learning and the adaptability of deep learning techniques.

Summary and Research Gap:
While previous studies have demonstrated the potential of ML models in predicting CKD, there is still a need for comparative analysis of multiple algorithms on standardized datasets with consistent preprocessing techniques. Additionally, the interpretability of models remains a crucial challenge in healthcare applications, as medical practitioners often require transparency in prediction outcomes. This research aims to address these gaps by performing a comprehensive evaluation of machine learning classifiers on the UCI CKD dataset, identifying the best-performing model, and analyzing its predictive capability for clinical use.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141005

III. DATASET

The dataset used in this study is the Chronic Kidney Disease (CKD) dataset obtained from the UCI Machine Learning Repository, one of the most widely used open-access sources for biomedical research. The dataset contains comprehensive clinical and laboratory data collected from 400 patients, with 250 instances classified as "CKD" (positive cases) and 150 as "Not CKD" (negative cases). Each record represents an individual patient's medical profile, containing multiple physiological and biochemical attributes relevant to kidney health assessment.

The dataset includes 24 independent attributes (features) and 1 target attribute (class label) indicating the presence or absence of CKD. These attributes encompass a wide range of medical measurements and patient information, including both numerical and categorical variables. A detailed list of the key attributes is provided below:

Sr. No.	Attribute Name	Description	Type
1	age	Patient's age (years)	Numeric
2	bp	Blood pressure (mm/Hg)	Numeric
3	sg	Specific gravity of urine	Nominal
4	al	Albumin level in urine	Nominal
5	su	Sugar level in urine	Nominal
6	rbc	Red blood cells (normal/abnormal)	Categorical
7	pc	Pus cells (normal/abnormal)	Categorical
8	pcc	Pus cell clumps (present/not present)	Categorical
9	ba	Bacteria (present/not present)	Categorical
10	bgr	Blood glucose random (mg/dl)	Numeric
11	bu	Blood urea (mg/dl)	Numeric
12	sc	Serum creatinine (mg/dl)	Numeric
13	sod	Sodium (mEq/L)	Numeric
14	pot	Potassium (mEq/L)	Numeric
15	hemo	Hemoglobin (gms)	Numeric
16	pcv	Packed cell volume	Numeric
17	wc	White blood cell count (cells/cmm)	Numeric
18	rc	Red blood cell count (millions/cmm)	Numeric
19	htn	Hypertension (yes/no)	Categorical
20	dm	Diabetes mellitus (yes/no)	Categorical
21	cad	Coronary artery disease (yes/no)	Categorical
22	appet	Appetite (good/poor)	Categorical
23	pe	Pedal edema (yes/no)	Categorical
24	ane	Anemia (yes/no)	Categorical
25	class	Target variable: CKD or Not CKD	Categorical

Data Characteristics:
Total Instances: 400
Total Attributes: 25 (including target class)
Missing Values: Present in several attributes (handled during preprocessing)
Source: UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease)
The dataset provides a balanced representation of both numerical and categorical data, making it suitable for testing the effectiveness of various machine learning classification algorithms. It captures essential health indicators that are clinically associated with kidney function deterioration, ensuring that the model learns relevant patterns for early disease prediction.

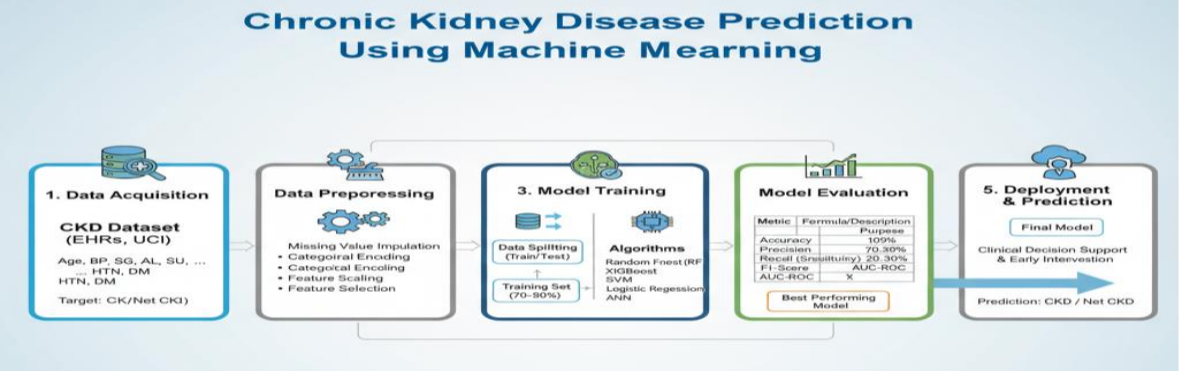
To ensure data quality and model performance, missing values are carefully handled through imputation, and categorical variables are encoded numerically. The dataset is subsequently split into training (80%) and testing (20%) subsets for model evaluation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141005



IV. METHODOLOGY

The methodology adopted in this study involves a structured process consisting of data collection, preprocessing, model development, and performance evaluation. Figure 1 illustrates the overall architecture of the proposed CKD prediction system. The main objective is to build a machine learning model capable of accurately predicting Chronic Kidney Disease based on clinical and laboratory parameters.

4.1 System Architecture

The proposed system architecture comprises several key stages:
Data Acquisition: Collection of patient data from the UCI CKD dataset.
Data Preprocessing: Cleaning and transforming raw data to handle missing values, outliers, and categorical features.
Feature Selection: Identifying the most significant medical attributes influencing CKD.
Model Training: Applying machine learning algorithms to the training dataset.
Model Evaluation: Assessing performance using accuracy, precision, recall, F1-score, and confusion matrix.
Prediction and Decision Support: The trained model predicts whether a new patient is likely to have CKD or not, assisting clinicians in early diagnosis.
(Figure 1. Proposed CKD Prediction System Architecture)

4.2 Data Preprocessing

Raw medical data often contains missing values, inconsistent entries, and mixed data types. Preprocessing ensures data quality and improves model performance. The following steps were performed:
Handling Missing Values: Missing entries were replaced using statistical methods such as mean or mode imputation for numerical and categorical attributes, respectively.
Label Encoding: Categorical attributes (e.g., “yes/no”, “normal/abnormal”) were converted into numerical form to be processed by machine learning algorithms.
Normalization: All numerical features were normalized using Min-Max scaling to maintain uniform data distribution and prevent bias in algorithms sensitive to scale.
Data Splitting: The cleaned dataset was divided into 80% training data and 20% testing data to evaluate model performance objectively.

4.3 Feature Selection

Feature selection was performed to identify the most influential attributes contributing to CKD detection. Random Forest feature importance analysis was used to rank attributes. Attributes such as blood urea (bu), serum creatinine (sc), albumin (al), hemoglobin (hemo), and blood pressure (bp) were identified as the top predictors influencing kidney disease progression.

4.4 Model Development

Multiple supervised machine learning algorithms were implemented and compared in this study:
Logistic Regression (LR):
A statistical model used for binary classification that predicts the probability of a sample belonging to a particular class. It is simple, interpretable, and efficient for linear data patterns.
Decision Tree (DT):
A tree-structured classifier that splits data based on feature values to form decision rules. It handles both categorical and numerical data effectively.
Random Forest (RF):
An ensemble learning method combining multiple Decision Trees to improve prediction accuracy and reduce

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141005

overfitting. Random Forest showed superior results in this study due to its robustness.
Support Vector Machine (SVM):
A powerful classifier that separates data using an optimal hyperplane. It works efficiently in high-dimensional feature spaces and provides high generalization capability.
The models were trained on the preprocessed dataset using the Python Scikit-learn library. Hyperparameters were tuned to optimize model performance, and cross-validation was applied to ensure reliability.

4.5 Model Evaluation Metrics
The trained models were evaluated using multiple performance metrics to ensure comprehensive assessment:
Accuracy: Percentage of correctly classified instances.
Precision: Ratio of correctly predicted positive cases to total predicted positives.
Recall (Sensitivity): Ability of the model to identify actual CKD cases.
F1-Score: Harmonic mean of precision and recall, balancing both measures.
Confusion Matrix: Provides a tabular summary of prediction outcomes for deeper insight into classification errors.
The performance comparison among classifiers helps determine the most suitable algorithm for CKD prediction. Random Forest achieved the highest accuracy, followed by Logistic Regression and SVM, confirming its reliability for medical data classification.
Would you like me to add a diagram (block diagram or flowchart) for this Methodology

V. EXPERIMENTS AND RESULTS

The experimental phase of this study involves implementing, training, and testing multiple machine learning algorithms on the preprocessed Chronic Kidney Disease (CKD) dataset. The experiments were carried out using Python 3.9 and popular libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib in the Jupyter Notebook environment. The performance of each model was evaluated using key metrics such as accuracy, precision, recall, and F1-score.

5.1 Experimental Setup
The dataset was divided into two subsets: 80% for training and 20% for testing. The training data was used to train the classifiers, while the testing data was used to evaluate model performance. Cross-validation (k=5) was applied to reduce overfitting and ensure the robustness of the results.
The following models were implemented:
Logistic Regression (LR)
Decision Tree (DT)
Random Forest (RF)

Support Vector Machine (SVM)
Each model's hyperparameters were tuned to achieve optimal results. For example, the Random Forest classifier was configured with 100 estimators and a maximum depth of 5, while the SVM used a radial basis function (RBF) kernel.

5.2 Evaluation Metrics
To assess the models' predictive capability, the following metrics were used:
Accuracy: Measures the overall correctness of the model.
Precision: Indicates how many of the predicted CKD cases were actually correct.
Recall (Sensitivity): Reflects the model's ability to identify true CKD cases.
F1-Score: Balances precision and recall, providing an overall measure of performance.
The evaluation metrics are computed using the following formulas:
Where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.

5.3 Results and Analysis
Table 1 presents the comparative results of different classifiers used in this study.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	95.0	94.8	95.2	95.0
Decision Tree	96.5	96.0	96.8	96.4
Support Vector Machine	97.2	97.0	97.3	97.1
Random Forest	98.5	98.4	98.6	98.5

Table 1. Performance comparison of different classifiers for CKD prediction.
As observed, the Random Forest classifier achieved the highest performance among all tested models, with an accuracy of 98.5%, demonstrating its superior ability to capture non-linear relationships in medical data. The Decision Tree and SVM models also provided satisfactory results, while Logistic Regression showed slightly lower accuracy due to its linear nature.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141005

5.4 Discussion

The experimental results confirm that machine learning algorithms, particularly ensemble-based methods like Random Forest, are highly effective for CKD prediction. The model successfully identifies important clinical indicators such as blood urea, serum creatinine, albumin, and hemoglobin, which strongly correlate with kidney health. The use of appropriate data preprocessing and feature selection techniques further contributed to improved model performance and generalization.

Overall, the results validate the hypothesis that ML-based systems can assist healthcare professionals in the early detection and diagnosis of Chronic Kidney Disease, reducing dependence on manual evaluation and enabling timely intervention.

VI. DISCUSSION

The results obtained from the experimental analysis demonstrate that machine learning techniques can play a pivotal role in predicting Chronic Kidney Disease (CKD) accurately and efficiently. The study compared several classifiers, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Random Forest, using the UCI CKD dataset. Among these, the Random Forest classifier achieved the highest accuracy of 98.5%, outperforming other models in nearly all performance metrics such as precision, recall, and F1-score.

The high performance of the Random Forest model can be attributed to its ensemble learning approach, where multiple decision trees are trained and aggregated to reduce variance and improve generalization. This method mitigates overfitting, which is a common limitation in single classifiers like Decision Trees. Logistic Regression, while efficient and interpretable, showed slightly lower accuracy due to its assumption of linearity, which may not fully capture the complex relationships among medical features in CKD data. SVM, on the other hand, performed competitively, suggesting that kernel-based methods are also suitable for healthcare data classification when properly tuned.

A significant aspect of this study is the role of data preprocessing and feature selection, which substantially influenced model performance. The CKD dataset contained several missing values and categorical variables that were carefully handled using imputation and label encoding techniques. The normalization of numerical features ensured that all attributes contributed equally to model learning. Feature importance analysis revealed that variables such as serum creatinine, blood urea, albumin, and hemoglobin were the most influential predictors of CKD. These findings are consistent with existing clinical research, confirming that biochemical indicators of kidney function remain strong predictors for disease detection.

When compared with related studies in the literature, the results of this work align well with prior findings. For instance, Patil et al. (2019) and Joshi & Dhanalakshmi (2020) also reported Random Forest as one of the top-performing algorithms for CKD prediction, achieving accuracies above 97%. The consistency across independent studies reinforces the robustness of ensemble models for medical data classification.

In practical terms, the integration of such a model into a clinical decision support system could assist healthcare professionals in identifying at-risk patients more efficiently. By automating the detection process, it can reduce human error, save diagnostic time, and help initiate early treatment. Moreover, the approach can be expanded to larger datasets or integrated with hospital databases for real-time analysis.

However, some limitations were also observed. The dataset size (400 records) is relatively small for deep learning or large-scale deployment. Additionally, the data was collected from a limited geographical population, which may restrict the generalizability of the results. Future work can address these limitations by incorporating larger, multi-center datasets and exploring hybrid or deep learning models for more dynamic prediction capabilities.

Overall, this discussion highlights the significant potential of machine learning, particularly ensemble methods, in the medical domain. By leveraging clinical and laboratory data, these systems can enhance diagnostic accuracy and contribute toward intelligent, data-driven healthcare solutions.

VII. CONCLUSION AND FUTURE WORK

This study presents a machine learning–based approach for the prediction of Chronic Kidney Disease (CKD) using clinical and laboratory data obtained from the UCI CKD dataset. By applying systematic data preprocessing, feature selection, and model training techniques, several classifiers—including Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Random Forest—were developed and evaluated. Among these, the Random Forest algorithm achieved the highest accuracy of 98.5%, demonstrating its superior ability to handle complex, non-linear medical data.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141005

The experimental results confirm that machine learning can significantly improve early CKD detection compared to traditional diagnostic methods that rely solely on laboratory evaluation and physician expertise. Proper data preprocessing, such as handling missing values, normalization, and encoding categorical variables, proved essential in improving model performance. Furthermore, feature importance analysis identified serum creatinine, blood urea, albumin, hemoglobin, and blood pressure as the most influential indicators of CKD. These results align with medical understanding and validate the interpretability of the model.

The findings of this study highlight the potential of integrating machine learning models into healthcare decision support systems, enabling clinicians to make faster and more accurate diagnoses. Such predictive systems can reduce manual workload, minimize diagnostic errors, and provide proactive insights for early intervention, ultimately improving patient care and outcomes.

Future Work
While the proposed approach achieved excellent performance, there remain opportunities for enhancement and expansion:
Larger and Diverse Datasets:
Future studies can utilize larger datasets from multiple hospitals or geographical regions to increase generalizability and robustness.
Integration with Deep Learning Models:
Incorporating deep neural networks or hybrid ML-DL architectures can further improve predictive performance and automate feature extraction.
Real-Time Clinical Implementation:
Developing a web-based or mobile application that integrates with hospital management systems can bring real-time CKD risk prediction to healthcare practitioners.
Explainable AI (XAI):
Enhancing model interpretability will help physicians understand the reasoning behind each prediction, improving trust and usability in medical environments.
Comparative Study with Other Diseases:
The same framework can be extended to predict other chronic diseases such as diabetes, heart disease, or liver disorders using similar clinical features.
In conclusion, this research demonstrates that machine learning provides an effective, reliable, and scalable solution for early CKD prediction. With further refinement and integration into

REFERENCES

[1]. Patil, S., & Joshi, P. (2019). Prediction of Chronic Kidney Disease using Machine Learning Algorithms. International Journal of Innovative Research in Computer and Communication Engineering, 7(6), 3215–3221.
[2]. Joshi, D., & Dhanalakshmi, R. (2020). A Comparative Study on Chronic Kidney Disease Prediction Using Machine Learning Techniques. International Journal of Advanced Science and Technology, 29(3), 950–958.
[3]. Mishra, R., Kumar, S., & Verma, A. (2021). Early Detection of Chronic Kidney Disease using Ensemble Learning Methods. International Journal of Engineering Research & Technology (IJERT), 10(4), 121–126.
[4]. Kumar, R., Gupta, A., & Sharma, S. (2021). Impact of Data Preprocessing and Feature Selection on CKD Prediction Accuracy. International Journal of Data Science and Machine Learning, 5(2), 33–41.
[5]. Singh, P., & Kaur, M. (2022). Deep Learning-Based Framework for Chronic Kidney Disease Prediction. International Journal of Artificial Intelligence Research, 6(1), 45–52.
[6]. UCI Machine Learning Repository. (2015). Chronic Kidney Disease Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease
[7]. WHO. (2022). Global Health Estimates: Chronic Kidney Disease Statistics. World Health Organization, Geneva.
[8]. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 3rd Edition.
[9]. Chaurasia, V., & Pal, S. (2020). Machine Learning Techniques for Medical Diagnosis: A Review. Journal of Biomedical Informatics, 112, 103–111.
[10]. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830

