IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, May 2025
Impact Factor: 7.67

A Personalized Healthcare Approach for Chronic Kidney Disease using Machine Learning

Mr. K. Karthick Babu[1] and R. Saranyasri[2]
Assistant Professor (M.Tech), Electronics and Communication Engineering[1]
B.E, Electronics and Communication Engineering[2]
Anjalai Ammal Mahalingam Engineering College, Kovilvenni, Thiruvarur, India

Abstract: *Chronic Kidney Disease (CKD) is a growing global health concern requiring early detection for effective treatment. This paper presents the design, development, and implementation of a machine learning-based system to predict CKD using clinical data. A microcontroller-based system is not used here; instead, Python and its libraries handle data preprocessing, missing value imputation, and feature analysis. Various classification algorithms were implemented and compared, with Random Forest achieving the highest accuracy. A heatmap and feature importance graph were generated to identify the most influential attributes for prediction. The user interface provides predictions and insights on patient data such as blood pressure, albumin, serum creatinine, and hemoglobin levels. The model can assist healthcare providers in making data-driven, timely decisions, potentially improving patient outcomes. Simulation and testing were carried out on real-time datasets and validated using performance metrics.*

Keywords: Kidney Disease, Machine Learning, Classification, Medical Diagnosis, Random Forest

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a critical global health concern affecting millions of individuals. It is characterized by the gradual loss of kidney function over months or years and is associated with high risks of cardiovascular complications and premature mortality. The kidneys are responsible for filtering waste and excess fluids from the blood, regulating blood pressure, and maintaining electrolyte balance. When kidney function deteriorates, it can lead to serious systemic complications, culminating in end-stage renal disease (ESRD), which requires dialysis or kidney transplantation. Traditional methods for diagnosing CKD rely on clinical tests, including blood pressure measurement, urine analysis, and glomerular filtration rate (GFR) estimation. However, due to the often assymptomatic nature of early-stage CKD, patients frequently remain undiagnosed until the condition has progressed significantly. In this context, the use of machine learning techniques has emerged as a powerful solution to support early detection, improve diagnostic accuracy, and assist clinicians in decision-making.

II. LITERATURE SURVEY

1. CLINICALLY APPLICABLE MACHINE LEARNING APPROACHES TO IDENTIFY ATTRIBUTES OF CKD FOR USE IN LOW-COST DIAGNOSTIC SCREENING, MD. RASSHED-AL-MAHFUZ, ABEDUL HAQUE, AKM AZAD, SALEM A. ALYAMI, JULIAN M. W. QUINN, AND MOHAMMAD ALI MONI, 2021

The results of this analysis demonstrated that the SHAP-identified important features were consistent with the current clinical thinking. It also found that an RF classifier method provides significantly high classification accuracy with the pathologically categorized attributes sets. The proposed RF classifier and reduced test attributes can therefore be potentially applied to reduce diagnosis costs and enable better management of early treatment plans.

2. ARTIFICIAL NEURAL NETWORKS(ANN) FOR CKD PREDICTION, SAHU AND JENA, 2020

This research applies machine learning techniques ANN to predict the CKD. The model achieved 96.3% accuracy and demonstrated robustness against noise and irrelevant features. It was trained using backpropagation and stochastic

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, May 2025
Impact Factor: 7.67

gradient descent on a dataset with 400 patient entries. Features like serum creatinine, albumin, and Hemoglobin were found to be strong predictors.

3. LOGISTIC REGRESSION FOR EARLY CKD DIAGNOSIS, SHARAMA ET AL, 2019
The model achieved an accuracy of 88% and performed well in identifying high-risk patients based on linear relationships among features. The researchers highlighted the model's transparency and ease of deployment in real-world clinical settings. It allowed doctors to interpret the impact of each feature on the outcome, such as how an increase in blood pressure or decrease in hemoglobin directly affects the likelihood of CKD.

4. SUPPORT VECTOR MACHINE(SVM) FOR CHRONIC KIDNEY PREDICTION, MADHUR AND MAHAJAN, 2018
This paper focused on a CKD dataset from the UCI Machine Learning Repository, which contains 400 patient records. After preprocessing and normalizing the features, SVM achieved an accuracy of 94.1%. The model performed well in handling non-linear relationships between features such as blood pressure, albumin levels, and hemoglobin. However, SVM was sensitive to parameter settings like C (regularization) and gamma values, requiring grid search optimization.

5. NAIVE BAYES CLASSIFIER FOR CHRONIC KIDNEY DISEASE DETECTION, PATEL AND PRAJAPATI, 2016
This study emphasized the model's speed and ease of implementation, making it suitable for rapid diagnostic systems. The model performed particularly well on categorical variables like hypertension, diabetes, and red blood cell count. However, its performance declined when features were correlated, as the assumption of independence was violated.

III. PROPOSED SYSTEM

- Data Collection: Utilizes open-source datasets (e.g., UCI CKD) with features like blood pressure, albumin, hemoglobin, and serum creatinine.
- Missing Value Handling: Imputes missing data using mean/mode substitution.
- Categorical Encoding: Converts categorical variables to numerical format.
- Normalization: Scales features to ensure uniformity.
- Feature Selection: Uses correlation heatmaps and tree-based feature importance to select relevant features.
- Model Training: Implements and compares classifiers like Random Forest, XGBoost, KNN, Decision Tree, and Logistic Regression.
- Optimization: Applies hyperparameter tuning and cross-validation for model enhancement.
- Model Evaluation: Assesses models using accuracy, precision, recall, F1-score, and confusion matrix.
- Best Performer: Random Forest achieves highest accuracy and generalization.
- User Interface: Provides a GUI for real-time CKD prediction using patient data input.

Demerits

- Data Dependency: Accuracy is limited by the quality and completeness of the input dataset.
- Imbalanced Data Risk: May underperform if the dataset has class imbalance (e.g., fewer CKD cases).
- Generalization Issues: May not perform well on unseen regional or demographic data without retraining.
- Limited Clinical Validation: Model predictions may not fully align with real-world clinical expertise.
- Black-Box Elements: Despite interpretability, ensemble models can still lack full transparency in decision-making.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, May 2025
Impact Factor: 7.67

Proposed work

This project proposes a work involves a structured, multi-phase approach to developing and deploying a CKD prediction model. Phase 1 begins with acquiring the CKD dataset from a trusted source like the UCI repository and analyzing its structure, including the number of instances, features, and missing values. In Phase 2, data preprocessing is performed through exploratory data analysis (EDA), missing value imputation, categorical encoding, and normalization of numerical data. Phase 3 focuses on feature engineering, using correlation analysis and feature importance metrics to select the most relevant attributes. Phase 4 involves training various classification algorithms—Random Forest, XGBoost, KNN, Logistic Regression, and Decision Tree—while optimizing hyperparameters via Grid or Random Search and validating models through K-fold cross-validation. In Phase 5, models are evaluated using accuracy, precision, recall, F1-score, and confusion matrix, and the best-performing model (anticipated to be Random Forest) is selected. Phase 6 includes developing a user-friendly interface using Tkinter or Streamlit for manual input and real-time prediction display. Phase 7 consists of testing the system with unseen data to validate its performance and usability. Finally, Phase 8 involves documenting all aspects of the system and preparing it for deployment in clinical or research settings.

IV. SYSTEM ARCHITECTURE

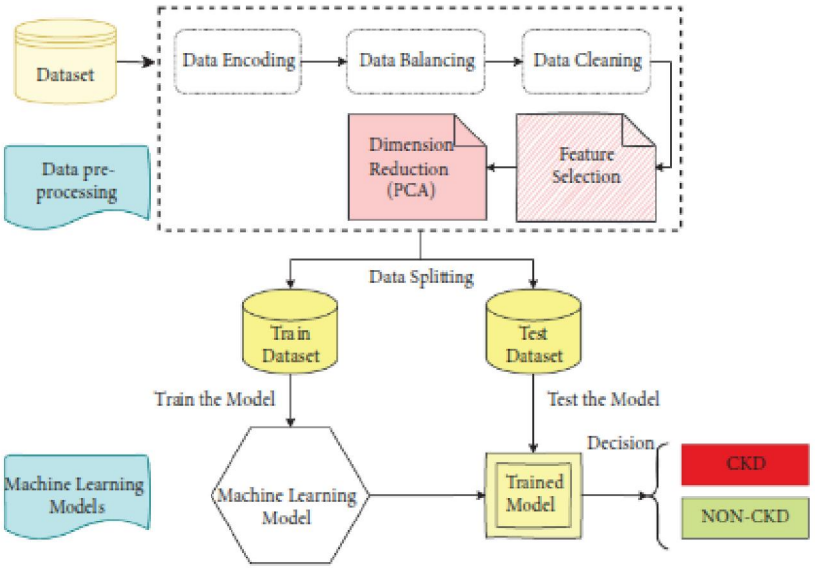

FIGURE 1: Proposed intelligent system for clinically early-stage chronic kidney disease diagnosis.

Figure 1: System Architecture

System Architecture Overview

1. Data Acquisition Layer

- Purpose: Collects patient health records for model training and prediction.
- Sources: UCI CKD dataset and real-time user input via GUI.
- Features: Includes 25+ clinical parameters (e.g., blood pressure, albumin, hemoglobin) with numerical and categorical types.

2. Data Preprocessing Layer

- Purpose: Cleans and transforms data for machine learning.
- Missing Value Handling: Uses mean, median, or mode imputation based on feature type.
- Encoding: Converts categorical data using Label Encoding or One-Hot Encoding.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, May 2025
Impact Factor: 7.67

- Normalization/Standardization: Applies Min Max Scaler or Standard Scaler for scaling.
- Outlier Detection: Optionally removes anomalous data points.

3. Feature Selection & Engineering Layer

- Purpose: Enhances model performance by selecting relevant features.
- Techniques: Uses correlation heatmaps, feature importance (Random Forest/XGBoost), and RFE.
- Output: Refined dataset with impactful features like albumin, serum creatinine, and hemoglobin.

4. Model Training Layer

- Purpose: Trains multiple ML models on preprocessed data.
- Models: Random Forest (primary), XGBoost, KNN, Logistic Regression, Decision Tree.
- Process: Data split (e.g., 80:20), hyperparameter tuning (GridSearchCV/RandomSearchCV), and K-fold validation.
- Output: Best model saved using joblib or pickle.

5. Prediction Layer

- Purpose: Predicts CKD status using trained model.
- Input: New clinical data via GUI or database.
- Process: Input preprocessing followed by model inference.
- Output: Binary prediction – CKD (Yes) or non-CKD (No).

6. Evaluation & Analytics Layer

- Purpose: Assesses model performance and offers explainability.
- Metrics: Accuracy, precision, recall, F1-score, and confusion matrix.
- Visualization: Uses Matplotlib/Seaborn; SHAP optionally for model interpretability.
- Insight: Supports validation before clinical deployment.

7. User Interface Layer

- Purpose: Provides an easy-to-use front-end for medical staff.
- Tools: Developed using Tkinter (desktop GUI) or Streamlit (web UI).
- Modular System Integration: Ensures seamless interaction between components in a sequential, dynamic processing flow.
- Real-Time Adaptation: Enables continuous learning model updates for accurate prediction and adaptive personalization.
- Outcome: Provides a scalable, intelligent framework to enhance e-learning effectiveness through behavioral analysis and tailored resource delivery.

V. ALGORITHMS

1. RANDOM FOREST

Random Forest is an ensemble learning algorithm that builds multiple decision trees using bootstrapped subsets of the data. Each tree is trained on a random subset of features, enhancing diversity and reducing overfitting. For classification, predictions are made by majority voting across all trees. It can compute feature importance by evaluating each feature's contribution to impurity reduction. The algorithm works well with both numerical and categorical data and handles large datasets efficiently. In the CKD prediction system, Random Forest classifies patients as CKD-positive or CKD-negative with high accuracy and robustness.

2. XGBOOSTING

XGBoost (Extreme Gradient Boosting) is a powerful and efficient implementation of gradient boosting machines, widely used for predictive modeling due to its high accuracy and scalability. It builds an ensemble of decision trees in a sequential manner, where each new tree attempts to correct the residual errors made by the previous trees using gradient descent optimization. The process begins with an initial prediction and iteratively improves the model by minimizing a

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, May 2025
Impact Factor: 7.67

loss function through gradient-based updates. XGBoost incorporates regularization techniques, including L1 (Lasso) and L2 (Ridge), directly into the objective function to prevent overfitting and enhance generalization. It also handles missing values inherently by learning the best direction for missing data during tree construction. Furthermore, XGBoost is highly optimized for speed, leveraging parallel processing, efficient memory usage, and sparsity-aware algorithms to accelerate training. Its combination of accuracy, interpretability, and computational efficiency makes it a top choice for structured data tasks in machine learning.

VI. MODULES

This project utilized a comprehensive suite of Python libraries to build an effective and interpretable chronic kidney disease (CKD) prediction system. Pandas was employed for data manipulation and preprocessing, enabling efficient loading, cleaning, transformation, and feature extraction through its DataFrame structure. NumPy supported numerical operations and array transformations essential for statistical computations. Scikit-learn provided machine learning utilities including classification algorithms (e.g., Random Forest, KNN), preprocessing tools, evaluation metrics, and hyperparameter tuning methods. XGBoost was used for high-performance gradient boosting, offering superior predictive accuracy and integrated regularization. For data visualization, Matplotlib and Seaborn were utilized to produce informative plots such as histograms, heatmaps, and feature importance visualizations. Tkinter and Streamlit facilitated the development of graphical and web-based user interfaces, allowing real-time user interaction with the model. Joblib and Pickle were used for model serialization, enabling the storage and retrieval of trained models for deployment. SciPy contributed statistical testing and optimization functionalities. Additionally, SHAP (SHapley Additive exPlanations) was optionally integrated to enhance model interpretability by quantifying the contribution of individual features to specific predictions.

VII. CONCLUSION

Chronic Kidney Disease (CKD) is a growing global health issue that requires early diagnosis to prevent critical outcomes like kidney failure, dialysis, or transplantation. Traditional diagnostic methods often fall short due to increasing data complexity and volume. To address these limitations, this study presents a machine learning-based approach using Random Forest and XGBoost for effective CKD prediction.

The clinical dataset was carefully preprocessed by handling missing values, normalization, and feature selection, ensuring optimal model training. Performance evaluation using accuracy, precision, recall, and F1-score revealed Random Forest as the most effective classifier. Data visualization tools, such as heatmaps and feature importance plots, helped identify key predictors including serum creatinine, albumin, hemoglobin, and blood pressure.

A user-friendly interface was developed using Tkinter and Streamlit, allowing real-time predictions based on user-inputted clinical values. This system is scalable and practical for deployment in healthcare settings including hospitals and telemedicine platforms. Ultimately, the project highlights the potential of machine learning to enhance early detection, support clinical decision-making, and improve outcomes for CKD patients.

REFERENCES

[1]Levey, A. S., & Coresh, J. (2012). Chronic kidney disease. The Lancet, 379(9811), 165–180.
[2] Jha, V., et al. (2013). Chronic kidney disease: Global dimension and perspectives. The Lancet, 382(9888), 260–272.
[3] KDIGO. (2013). KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney International Supplements, 3(1), 1–150.
[4] Saran, R., et al. (2020). US Renal Data System 2019 Annual Data Report: Epidemiology of Kidney Disease in the United States. American Journal of Kidney Diseases, 75(1), A6–A7.
[5] Tomasev, N., et al. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. Nature, 572(7767), 116–119.
[6] Almansour, F. S. (2018). A machine learning model for predicting chronic kidney disease. International Journal of Advanced Computer Science and Applications, 9(6), 520–525.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, May 2025
Impact Factor: 7.67

[7] Kora, P., & Kalva, S. K. (2015). Chronic kidney disease prediction using machine learning techniques. International Journal of Engineering Research in Africa, 18, 185–195.
[8] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for chronic kidney disease using machine learning techniques. Procedia Computer Science, 167, 674–686.
[9] Tseng, C. C., et al. (2019). Predicting chronic kidney disease using a machine learning algorithm and population-based data. Medicine, 98(23), e16186.
[10] Huang, C., et al. (2020). A machine learning approach for predicting major chronic diseases using medical records. IEEE Access, 8, 94474–94483.
[11] Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. IJCSNS International Journal of Computer Science and Network Security, 8(8), 343–350.
[12] Sudharsan, B., et al. (2020). Machine learning algorithms for early detection of chronic kidney disease. Journal of Healthcare Engineering, 2020.
[13] Khan, Y., et al. (2019). Data mining in healthcare for prediction of chronic diseases: A review of literature. Journal of Computer Science, 15(4), 486–500.
[14] Santos, M. D., et al. (2019). Predicting CKD using data mining techniques. Procedia Computer Science, 164, 103–110.
[15] UCI Machine Learning Repository. (2020). Chronic Kidney Disease Dataset.
[16] Rakesh, S., & Kumari, S. (2017). Classification of chronic kidney disease using hybrid SVM and KNN classifier. Procedia Computer Science, 125, 390–396.
[17] Levey, A. S., et al. (2005). Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO). Kidney International, 67(6), 2089–2100.
[18] Goyal, M., et al. (2020). Early detection of chronic kidney disease using machine learning techniques. Journal of Applied Science and Computations, 7(4), 498–504.
[19] Behrens, T., et al. (2018). Artificial intelligence for early detection of chronic kidney disease. Kidney360, 1(6), 563–574.
[20] Dey, V., et al. (2021). A novel hybrid model for chronic kidney disease prediction using machine learning and bio-inspired optimization. Computers in Biology and Medicine, 132, 104308.