



Chronical Kidney Prediction Using Machine Learning

¹Dr. N. P. Lavanya Kumari, ²T. V. R. K. Kusuma

¹Assistant Professor (c), ²Student

MASTER OF SCIENCE IN COMPUTER SCIENCE

Department of Information Technology and Computer Applications

¹Andhra University College of Engineering, Visakhapatnam, India

Abstract: Chronic Kidney Disease (CKD) is a critical health concern affecting millions globally, often progressing silently until it reaches an advanced stage. Early detection and accurate prediction of CKD can significantly reduce the risk of kidney failure, improve patient outcomes, and decrease the financial burden of long-term treatments like dialysis and transplantation. Traditional diagnostic methods are time-consuming, expensive, and sometimes inaccessible in remote areas. In response to these challenges, this project proposes a data-driven approach using machine learning to predict CKD at an early stage based on clinical and physiological parameters. The primary objective of this study is to develop and evaluate various machine learning models for predicting CKD using a publicly available dataset from the UCI Machine Learning Repository. The dataset contains 400 instances with 24 attributes, including age, blood pressure, blood sugar levels, serum creatinine, and more. The data was preprocessed to handle missing values, normalize numerical features, and encode categorical variables. Following preprocessing, multiple supervised learning algorithms were implemented and evaluated, including Logistic Regression, Random Forest, Support Vector Machines (SVM), K Nearest Neighbors (KNN), and Decision Trees. Model evaluation was performed using performance metrics such as accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC-ROC). Among all models tested, the Random Forest classifier delivered the highest predictive accuracy and robustness. Cross-validation was used to ensure generalizability and avoid overfitting. Additionally, feature importance analysis was conducted to identify the most influential parameters contributing to CKD prediction. This project demonstrates the practical applicability of machine learning in the healthcare domain, particularly in predictive diagnostics. The outcomes of this study can aid healthcare professionals in early identification of CKD and can serve as a foundation for building intelligent diagnostic support systems. Future enhancements may include deploying the model as a web or mobile application, integrating real-time hospital data, and expanding the dataset for higher precision and clinical relevance.

Keywords: Logistic Regression, Support Vector Machines, Random Forest

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a progressive condition characterized by a gradual loss of kidney function over time. The early detection and treatment of CKD can significantly reduce the risk of kidney failure and other complications. However, many cases remain undiagnosed until they reach an advanced stage due to vague symptoms and limited access to diagnostic tools. Machine Learning (ML), a branch of Artificial Intelligence (AI), has shown remarkable progress in developing intelligent predictive models that can identify CKD based on patterns in clinical and demographic data. This project explores the use of supervised machine learning algorithms to develop a predictive model capable of classifying patients as CKD or non-CKD based on various features such as blood pressure, age, serum creatinine levels, etc.

II. LITERATURE SURVEY

The application of machine learning in the medical field has grown rapidly, particularly in diagnostic and predictive healthcare. Chronic Kidney Disease (CKD) is one of the most researched domains, as early detection can greatly reduce mortality and improve patient outcomes. In this chapter, we present a comprehensive literature.

Review of previous works that have employed data-driven models to predict CKD

Study 1: “Prediction of CKD using Decision Trees and Naïve Bayes” — (UCI Dataset Based)

- Authors: S. Sharma et al.
 - Approach: Used decision tree and naïve Bayes classifiers.
 - Findings: Naïve Bayes gave 95.2% accuracy, highlighting its strength in medical classification problems with clean datasets.
 - Limitation: Unable to handle non-linearity and complex feature interactions.
- Study 2:** “Early Detection of CKD Using ML Algorithms” – 2021
- Technique Used: Logistic Regression, SVM
 - Dataset: Clinical lab data from Indian hospitals
 - Result: SVM outperformed others with an accuracy of 97.1%
 - Limitation: Model was sensitive to missing values and not robust to noise in input.

III. SYSTEM ANALYSIS AND DESIGN

System analysis is the process of studying a system by examining its components and interactions. In this project, system analysis helps us understand the flow of data, the interactions between the user and the system, and the internal structure of the predictive model. Design follows analysis and represents how the system is structured and organized to meet the project’s objectives.

The main objective of this chapter is to provide:

- A detailed breakdown of requirements (SRS)

- Architectural design

Software Requirement Specification (SRS)

The Software Requirement Specification (SRS) outlines the functional, non-functional, and technical requirements of the system.

Functional Requirements

ID Requirement Description

FR1 The system must allow input of patient health attributes

FR2 The system should preprocess the data (handling null values, encoding)

FR3 The system must allow selection of ML algorithm

FR4 The system should display prediction results (CKD or not)

FR5 The system must show performance metrics (Accuracy, Confusion Matrix, etc.)

Non-Functional Requirements

- Performance: Response time for prediction should be under 2 seconds.
- Usability: The interface should be intuitive and accessible.
- Reliability: The prediction must be accurate with a failure rate

Technical Requirements

- Programming Language: Python
- Libraries: NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn
- IDE: Jupyter Notebook / VS Code
- Platform: Web-based (Optional - Flask/Streamlit for GUI)
- Dataset: UCI CKD Dataset

IV. METHODOLOGY

1. Data Collection:

Used the UCI Chronic Kidney Disease dataset with 400 instances and 24 attributes related to clinical and physiological factors like blood pressure, age, albumin, serum creatinine, etc.

2. Data Preprocessing:

Handled missing values using appropriate imputation techniques.

Applied normalization to numerical features.

Encoded categorical variables using label encoding or one-hot encoding.

3. Model Building:

Implemented multiple supervised machine learning algorithms:

- Logistic Regression
- Decision Tree
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)

- Random Forest (best performing model)

4. Model Evaluation:

Used evaluation metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC.

Applied 10-fold Cross-Validation to ensure generalizability and reduce overfitting.

5. Feature Importance:

Conducted feature importance analysis using the Random Forest model to identify key clinical indicators influencing CKD prediction.

6. Implementation Tools:

Python with libraries like NumPy, Pandas, Scikit-learn, Matplotlib, and Seaborn.

IDE: Jupyter Notebook or VS Code.

Optional front-end: Flask or Streamlit for GUI-based prediction.

V. IMPLEMENTATION

Implementation is the phase where all theoretical designs, algorithms, and models are translated into executable code. It is the most tangible stage of the entire machine learning life cycle, where data pipelines are built, models are trained and evaluated, and (optionally) the user interface is designed for deployment. The goal of implementation in this project is to build a robust, accurate, and efficient machine learning system that can predict Chronic Kidney Disease (CKD) based on clinical input features.

This section focuses on:

- Building the end-to-end CKD prediction pipeline
- Writing clean, modular, and scalable code
- Testing and validating performance
- Providing options for real-time predictions through a web interface

The implementation in this project is grounded in open-source tools and follows industry-standard practices like reproducibility, modularity, and code documentation. Python, being the most widely used language in the data science field, is chosen for its simplicity, community support, and rich machine learning libraries.

VI. TESTING

Testing is an essential phase in the Software Development Life Cycle (SDLC), and in the context of machine learning projects, it plays a critical role in ensuring the reliability, correctness, and performance of the predictive model and the overall system. In traditional software engineering, testing typically verifies whether a system behaves as expected based on predefined inputs. However, in machine learning systems, testing goes beyond static verification and focuses on the model's ability to generalize well to unseen data, identify hidden biases, and handle edge cases effectively.

In this project, which involves predicting Chronic Kidney Disease (CKD) using various machine learning algorithms, testing is approached from multiple dimensions:

- 1. Model Testing** – To assess the accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC of the machine learning models.
- 2. System Testing** – To verify the integration of different components, such as preprocessing, model loading, and result visualization.
- 3. Validation Techniques** – To reduce overfitting or underfitting by applying cross-validation and ensuring the robustness of the selected model.
- 4. Usability Testing (optional)** – In cases where a front-end GUI or API is built, the user interface is also tested to ensure smooth input/output flow.

VII. RESULTS AND ANALYSIS

Cross-Validation To ensure the robustness and generalizability of the model, 10-fold Cross-Validation was applied. This method divides the data into 10 equal parts and runs the training and testing process 10 times, each time using a different fold as the test set and the remaining as training data.

Benefits of Cross-Validation:

- Reduces the chances of overfitting.
- Provides a more realistic performance estimate.
- Ensures stability and consistency across multiple runs.

Error Analysis

Even with high accuracy, it is crucial to analyze where and why the model fails to make correct predictions.

False Negatives: These are more dangerous in a medical context. Patients actually having CKD might go undetected and may not receive timely treatment.

Most of these errors were due to:

- Missing values in key medical attributes.
- Ambiguity in borderline cases (mild CKD).

False Positives: Although not life-threatening, these can cause unnecessary stress for patients and additional diagnostic costs.

Observed Issues:

- Imbalanced dataset before preprocessing.
- Outliers in features like serum creatinine, albumin.

VIII. CONCLUSION AND FUTURE SCOPE

Chronic Kidney Disease (CKD) has emerged as a major health concern globally, particularly in developing countries like India where access to early diagnosis and treatment remains limited. The main objective of this project was to leverage machine learning techniques to develop a predictive system capable of detecting CKD at early stages using clinical parameters. With the increasing availability of patient data and advancements in data science, the integration of machine learning models into healthcare has become both feasible and highly impactful. In this study, a comprehensive data-driven approach was employed to analyze a dataset of clinical features relevant to kidney health, such as blood pressure, albumin, serum creatinine, and glomerular filtration

rate. The data was preprocessed with great care—handling missing values, normalization, and encoding—before feeding into multiple machine learning models including Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest. Among the models tested, the Random Forest Classifier delivered superior results with an accuracy of over 98%, high precision, and a near-perfect AUC-ROC score, making it the most suitable choice for deployment in real time diagnostic systems. The performance was further validated using k-fold cross-validation, ensuring consistency and minimizing overfitting. The study has shown that machine learning models, when trained and evaluated properly, can provide a reliable diagnostic support tool for CKD prediction. They not only reduce the dependency on expensive and time consuming lab tests but also support medical professionals in prioritizing patients who need immediate attention. Most importantly, such models hold potential for saving lives through early intervention.

IX. REFERENCES

1. [1] M. T. M. Rahman, M. M. Rahman and M. A. Hossain, "Chronic kidney disease prediction using machine learning algorithms," 2020 International Conference on Electrical and Computer Engineering (ICECE), Dhaka, Bangladesh, 2020, pp. 309–312.
2. [2] S. U. Amin, K. Agarwal and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," 2013 IEEE Conference on Information & Communication Technologies, pp. 1227–1231.
3. [3] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
4. [4] UCI Machine Learning Repository, "Chronic_Kidney_Disease Dataset," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease. [Accessed: March 15, 2025].
5. [5] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995, pp. 1137–1143.