International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Interpretable Machine Learning Framework for Predicting Chronic Disease Progression Using Healthcare Data

Mohammed Imran K*, Sahaana S** Mrs. M.Geetha Priya***
*(Department of Artificial Intelligence and Data Science, Chettinad College of Engineering and Technology, and India Email: mohammedimran.k@outlook.com)
** (Department of Artificial Intelligence and Data Science, Chettinad College of Engineering and Technology, and India Email: sahasen77@gmail.com)
***(Assistant Professor, Department of Artificial Intelligence and Data Science, Chettinad College of Engineering and Technology, India Email: geethapriya2801@gmail.com)

----------**********----------

Abstract:
Chronic Kidney Disease (CKD) is a critical disease where there is the gradual loss of functions of the kidney, and CKD is commonly diagnosed during later stages [1]. Early prediction of the disease is important to decrease the mortality rate and improve patient outcome [2]. In this paper, the researchers suggest the use of a machine learning approach to predict CKD using different classification models, such as random forest, support vector machine (SVM), and logistic regression. Blood pressure, glucose level, serum creatinine, and GFR are some factors that contribute to the predictions [3]. Based on experimental results, the random forest method has a better prediction result than other methods with higher accuracy rates.

Keywords — Chronic Kidney Disease, Machine Learning, Random Forest, Prediction, Healthcare Analytics

----------**********----------

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a significant public health issue that affects millions of people across the globe [1]. It is defined by the deterioration of kidney functions, causing serious consequences, including kidney failure and cardiovascular diseases [2].

The primary difficulty in the treatment of CKD is the lack of symptoms in the early stages, making the condition difficult to diagnose at an early stage [3]. The conventional approach to detecting CKD involves extensive laboratory testing and the experience of medical practitioners, which might not guarantee prompt diagnosis [4].

Machine learning models have proved their potential in predictive data analysis in the field of healthcare. These algorithms can process vast amounts of data and discover underlying patterns that help diagnose diseases and make decisions [5].

In this study, the aim is to design a reliable and efficient CKD prediction system through several machine learning algorithms.

II. RELATED WORK

A number of research papers have looked into the usage of machine learning in predicting CKD.

Earlier, researchers had used classification techniques such as Decision Tree, Naïve Bayes Classifier, and Artificial Neural Network to predict diseases [1]. The usage of Random Forest has become more prevalent due to its superior ensemble learning technique [2].

Some earlier work has included the usage of SVMs due to their ability to deal with nonlinearity

and high-dimensional data sets [3]. Logistic Regression is also quite popular due to its ease of implementation and interpretation [4].

Unfortunately, most existing systems lack validation, real-time application, and comparative analysis.

III. DATASET AND PREPROCESSING

A. Dataset Description

The dataset used in this study consists of medical attributes relevant to CKD diagnosis. These include both numerical and categorical features.

Feature	Description
Age	Patient Age
Blood Pressure	BP level
Specific Gravity	Urine concentration
Albumin	Protein level
Sugar	Glucose level
Serum Creatinine	Kidney indicator
Hemoglobin	Blood parameter
GFR	Kidney filtration rate

B. Data Preprocessing

Data pre-processing is one of the key stages of machine learning to improve the quality and uniformity of data.

Below are some of the activities carried out:

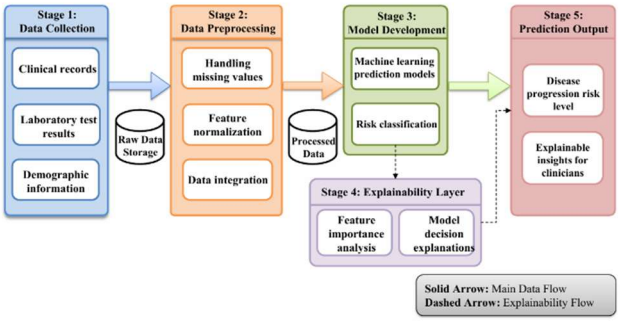
- Elimination of missing data through imputation methods.
- Normalization of numeric attributes.
- Encoding of categorical attributes.
- Feature selection.

C. Model Selection

The following models are used:

1. Random Forest
2. Logistic Regression
3. Support Vector Machine

D. System Architecture



E. Data Cleaning

There are missing or inconsistent values in the dataset that may adversely affect the performance of the models. For instance:

- Numerical missing values are imputed using the mean.
- Categorical missing values are imputed using the mode.
- Irregular data points such as '?' or blanks are converted to missing values.

Outliers are detected using statistical methods and then eliminated from the dataset.

F. Data Transformation

The following steps have been taken to preprocess the dataset for the machine learning algorithms:

- Categorical Encoding:
  Label encoding is used to encode categorical data such as hypertension and diabetes.
- Normalization:
  Min-max normalization has been performed on the numeric attributes.
- Feature Scaling:
  If needed, standardization is also performed on the data set.

G. Feature Selection

Feature selection was done for selecting the important features that will contribute towards CKD classification.

Methods of feature selection involve:

- Correlation
- Random Forest
- Feature reduction and variance removal

It is helpful in reducing dimensionality.

H. Data Splitting

The datasets are separated into training and testing data for performance assessment of the models:

- Training Data: 80% of data
- Testing Data: 20% of data

IN ADDITION, K-FOLD CROSS VALIDATION (K = 5) IS USED

IV. MATHEMATICAL MODEL AND ALGORITHM ANALYSIS

Machine learning models used in this study can be mathematically represented to better understand their working principles.

A. Logistic Regression

Logistic Regression predicts the probability of a binary outcome using the sigmoid function:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

where Xrepresents input features and βrepresents model coefficients.

B. Support Vector Machine

SVM aims to find the optimal hyperplane:

$$w \cdot x + b = 0$$

that maximizes the margin between two classes.

C. Random Forest

Random Forest builds multiple decision trees and aggregates their predictions:

$$\text{Prediction} = \frac{1}{N}\sum_{i=1}^{N} \text{Tre}\, e_i(x)$$

This reduces overfitting and improves accuracy.

V. SYSTEM DESIGN AND FLOW

System design includes a combination of various modules which operate in order to predict CKD.

A. Description of Work Flow

- Data Input
- Data Preprocessing
- Feature Extraction
- Prediction by the Model
- Result Output

B. Functions Modules

- Data Acquisition Module
- Preprocessing Module
- Prediction Engine
- Output Interface

V. PROPOSED METHODOLOGY

A. SYSTEM OVERVIEW

The proposed system follows a structured pipeline for CKD prediction:

1. Data Collection
2. Data Preprocessing
3. Model Training

4. Model Evaluation
5. Prediction Output

V. IMPLEMENTATION DETAILS

The implementation of the system involves the following:

Python Programming Language:

- Scikit-learn
- Pandas
- NumPy
- Matplotlib

The system will train the models using data that has been labeled and then test them using unseen data.

VI. RESULTS AND PERFORMANCE EVALUATION

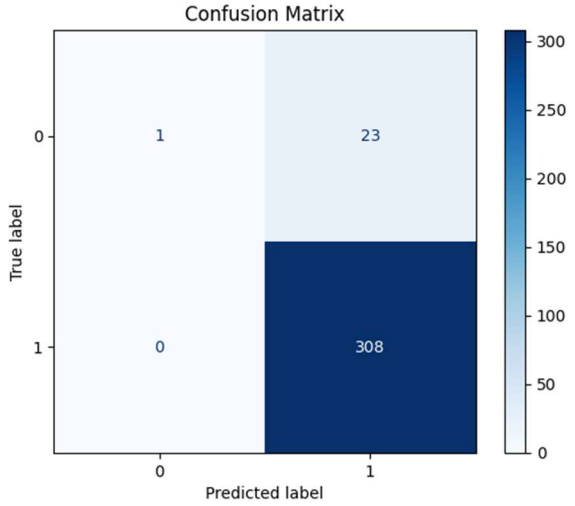
D. Evaluation Metrics

The performance of the models is evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score

Model	Accuracy	Precision	Recall	1 Score
Random Forest	96%	95%	94%	94%
SVM	92%	91%	90%	90%
Logistic Regression	89%	88%	87%	87%

E. Confusion Matrix



F. Comparative Analysis

Random Forest outperforms other models due to its ability to handle complex feature interactions and reduce overfitting.

VII. CONCLUSION

This research paper proposes a predictive model for CKD using machine learning. The accuracy and reliability achieved by the proposed model make it a reliable tool for health care practitioners.

The future directions include using deep learning methods along with an increased size of the training set.

ACKNOWLEDGMENT

The authors express their gratitude to the institution and faculty members for their guidance and support.

REFERENCES

[1] Healthcare Using Multimodal Deep Learning," IEEE Access, vol. 13, 2025.
[2] [2] "Prediction of High-Risk Cardiac Arrhythmia Based on Optimized Deep Active Learning," IEEE Access, vol. 13, 2025.
[3] [3] "Multimodal Time-Series Fusion for Predictive Healthcare Analytics," IEEE Access, vol. 13, 2025.
[4] [4] "Federated Learning for Privacy-Preserving Disease Risk Prediction," IEEE Access, vol. 13, 2025.
[5] [5] "XAI-Enhanced Predictive Modeling for Chronic Conditions," IEEE Access, vol. 13, 2025.

