Chronic Renal (Kidney) Disease Prediction Using Machine Learning

ASSISTANT PROFESSOR MRS.BHAGYASHRI WAKDE [1], DEVIKA M[2], KIVUDI CHIKKAPLA PRIYANKA[3], GANGOTRI SUBHASH BASAGOUDAR[4], SANDESH MALLAPPA KALAGI[5]
[1,2,3,4,5]Department of Computer Science and Engineering Rajiv Gandhi Institute of Technology Bangalore, India

Abstract- Chronic Kidney Disease (CKD) is a progressive medical condition that, if left undiagnosed or untreated, can lead to kidney failure and severe health complications. Early detection is crucial to managing the disease and improving patient outcomes. This project aims to develop an intelligent system that accurately predicts the likelihood of CKD based on various clinical parameters such as age, blood pressure, serum creatinine, and others. The system employs advanced data preprocessing techniques, exploratory data analysis, and feature selection methods including correlation heatmaps, LASSO regularization, and wrapper-based techniques to identify the most significant features.Multiple machine learning algorithms—including Random Forest, Support Vector Machine (SVM), Decision Tree, Logistic Regression, XGBoost, and deep learning hybrid models—are trained and evaluated using performance metrics like accuracy, precision, recall, F1-score, and AUC-ROC. The best-performing model is integrated into a user-friendly web application built using Flask, with a front end developed using HTML, CSS, and JavaScript. The application allows users to input medical parameters and instantly receives a CKD risk prediction. This system not only aids healthcare professionals in early diagnosis but also empowers users with a tool for proactive health monitoring. The solution is scalable, interpretable, and can be continuously updated with new data to improve prediction accuracy and reliability.

Index Terms- Chronic Kidney Disease (CKD),Machine Learning (ML),Healthcare AnalyticsData Mining, Early Detection Classification Algorithms,Clinical Decision Support, Predictive Modeling,webapp.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a major public health concern affecting millions of people worldwide. Early detection and timely intervention can significantly improve patient outcomes and reduce healthcare costs. This project leverages machine learning and deep learning techniques to develop an intelligent CKD prediction system. The system analyzes patient health metrics such as age, blood pressure, creatinine levels, and other physiological data to assess the risk of CKD. The project includes a comprehensive pipeline—from data collection and preprocessing to model development and deployment in a web-based application for real-time predictions.

Chronic Kidney Disease (CKD) is a progressive condition characterized by the gradual loss of kidney function over time. It has become a major global health concern, often leading to kidney failure, cardiovascular diseases, and increased mortality if not detected and treated early. According to medical studies, early prediction and timely intervention can significantly reduce the risk of complications and improve patient outcomes.

With the growing availability of healthcare data, machine learning techniques have emerged as powerful tools for medical diagnosis and prognosis. By analyzing patient records—including parameters such as blood pressure, blood sugar, serum creatinine, hemoglobin levels, and lifestyle factors—machine learning models can effectively predict the likelihood of CKD. These models can assist healthcare professionals in decision-making by providing accurate, data-driven insights for early detection.

This project aims to develop and evaluate machine learning models for predicting CKD based on clinical and demographic data. The study involves data preprocessing, feature selection, model training, and performance evaluation using various algorithms. The ultimate goal is to create a reliable and efficient predictive system that can support healthcare practitioners in identifying patients at risk of CKD at an early stage.

II. LITERATURE SURVEY

i. The study utilized the Chronic Kidney Disease (CKD) dataset from the UCI Machine Learning Repository, consisting of 400 patient records with 24 attributes collected from hospitals in 2015. Since the dataset contained missing values, preprocessing steps such as imputation, normalization, binary transformation, and standardization were applied to ensure data quality. To enhance prediction accuracy and minimize overfitting, three feature selection techniques were employed: Correlation-based Feature Selection (CFS), Forward Feature Selection, and LASSO regularization. Several classification algorithms were implemented, including Artificial Neural Network (ANN), C5.0 Decision Tree, Logistic Regression, CHAID, Linear SVM, K-Nearest Neighbors, and Random Tree. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic minority samples. Model performance was evaluated using metrics such as accuracy, precision, recall, F1-score, ROC-AUC, and GINI coefficient, while McNemar's Test was used to validate statistical significance between classifiers.

ii. The CKD dataset from Kaggle, containing 400 samples with 13 clinical features, was preprocessed and split into 340 training and 60 testing records. Models including SVM, ANN, Linear Regression, and KNN (k=7) were evaluated, with Gaussian SVM using 10-fold cross-validation achieving the highest accuracy. The ANN model with 15 hidden neurons also showed strong performance, effectively capturing non-linear patterns in the data.

iii. Chronic Kidney Disease (CKD) is a progressive condition where early detection is critical to prevent kidney failure. Traditional tests like GFR are costly, so machine learning (ML) provides a reliable alternative for early prediction using medical records. Studies show algorithms such as KNN (76.96% accuracy), Decision Tree (98.60%), and ANN (95.20%) are effective in CKD detection. Key attributes like age, blood pressure, and specific gravity play a vital role in prediction. Model performance is evaluated using metrics such as accuracy, precision, recall, F-measure, MAE, RMSE, and RAE to ensure robust and accurate results.

Key Insights from Literature:
Most studies used the UCI CKD dataset or variants.
Decision Tree and ANN showed consistently high accuracy.Use of SMOTE and feature selection significantly improved model performance.The lack of user-facing applications in these systems highlighted the need for a deployable web-based model.

III. METHODOLOGY

A. Data Collection

The dataset was sourced from the UCI Machine Learning Repository and Kaggle, comprising 400 patient records with 24 clinical features, such as age, blood pressure, hemoglobin, serum creatinine, albumin, and specific gravity. Each record contains a target label indicating CKD or non-CKD.

B. Data Preprocessing

Data preprocessing ensured high quality and consistency: Handling Missing Values: Missing entries were imputed using mean/median values; rows with excessive nulls were removed. Normalization: Min-Max scaling and StandardScaler were applied to numerical features. Categorical Encoding: Variables like yes/no and normal/abnormal were encoded using label encoding and one-hot encoding. Outlier Detection: Outliers were identified using Z-score and IQR methods. Class Balancing: The SMOTE algorithm was optionally applied to address data imbalance.

C. Exploratory Data Analysis (EDA)

EDA was performed to analyze feature trends and correlations. Histograms, box plots, and pair plots were used to visualize distributions, while a Pearson correlation matrix highlighted multicollinearity and strong relationships with the target variable.

D. Feature Selection
To avoid overfitting and improve performance:Filter Methods: Correlation-based Feature Selection (CFS) ranked attributes by relevance.Wrapper Methods: Forward feature selection identified optimal subsets.Embedded Methods: LASSO regression selected important predictors during training.

E. Model Development
The following algorithms were implemented and trained: Random Forest, Support Vector Machine (SVM), Decision Tree, LogisticRegression, XGBoost, Deep Neural Network (DNN), Hybrid Models (Voting Classifier, Stacking, DNN + XGBoost), Hyperparameter tuning was performed using Grid Search and Random Search to optimize model performance.

F. Model Evaluation
Models were evaluated using the following metrics: Accuracy – overall correctness of classification, Precision & Recall – critical for minimizing false negatives in medical diagnosis, F1-Score – harmonic mean of precision and recall, AUC-ROC Curve – ability of the model to distinguish between CKD and non-CKD classes

G. Web Application Development
A web application was developed for real-time use: Backend: Flask framework handled model integration and prediction logic., Frontend: HTML, CSS, and JavaScript provided an interactive interface., Output: Predictions were displayed with probabilities, supporting interpretability.

H. Deployment and Testing
The system was tested on local environments with sample patient inputs. The backend design allows easy model updates without modifying the frontend. End-to-end testing validated prediction accuracy and system robustness.

IV. EXISTING SYSTEM

Several research studies have demonstrated the potential of machine learning techniques in detecting Chronic Kidney Disease (CKD) at early stages. The existing systems primarily focus on the following:

The prediction of Chronic Kidney Disease (CKD) commonly uses datasets from UCI or Kaggle, which contain around 400 patient records with features such as blood pressure, albumin, hemoglobin, and creatinine. Since these datasets often have missing values, preprocessing steps like normalization, binary encoding, and standardization are essential. Machine learning models such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), and Random Trees are frequently applied and evaluated using metrics like accuracy, precision, recall, F1-score, and AUC-ROC. To further enhance model performance, feature selection techniques including Correlation-Based Selection, LASSO, and Wrapper Methods are used, while data imbalance issues are addressed using Synthetic Minority Oversampling Technique (SMOTE). Despite these advancements, most existing systems remain limited to offline research settings, lack real-time web-based interaction, require manual updates, and often rely on single models instead of leveraging ensemble or hybrid approaches.

V. PROPOSED SYSTEM

To overcome the limitations in existing approaches, our proposed system introduces a comprehensive, intelligent, and user-interactive CKD prediction platform, which combines powerful machine learning techniques with a web-based deployment for accessibility and ease of use.

The proposed system enhances CKD prediction by integrating advanced data processing, hybrid modeling, and real-time usability. It begins with detailed exploratory data analysis (EDA) using visualizations to extract insights, followed by feature selection through correlation heatmaps, LASSO, and wrapper methods. Multiple machine learning models, including Random Forest, SVM, XGBoost, Decision Tree, Logistic Regression, and Deep Neural Networks (DNN), are evaluated alongside hybrid approaches such as voting classifiers, stacking ensembles, and a DNN + XGBoost hybrid, with performance measured using accuracy, precision, recall, F1-score, and AUC-ROC. To make the system accessible, a Flask-based web application is developed with HTML, CSS, and JavaScript, allowing users to input medical parameters and receive real-time predictions with probability scores and interpretation. The software stack includes Anaconda, Jupyter Notebook, Python, Flask, and Visual Studio Code, ensuring robust development and deployment. Compared to existing systems, this approach provides real-time predictions through a user-friendly web interface, improved accuracy via hybrid models, scalability for future datasets, and a practical bridge between machine learning models and end-users such as healthcare professionals and patients.

CONCLUSION

This project successfully demonstrates the application of machine learning and deep learning techniques for early detection of Chronic Kidney Disease (CKD). By leveraging patient health data and implementing a comprehensive pipeline—ranging from data collection and preprocessing to model training and deployment—a robust and intelligent CKD prediction system was developed.

The system includes exploratory data analysis to uncover meaningful patterns, followed by effective feature selection techniques such as correlation analysis and LASSO regularization. Multiple models, including Random Forest, SVM, XGBoost, Decision Tree, Logistic Regression, and hybrid ensemble techniques, were evaluated to ensure optimal performance. Performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC were used to assess and compare models.

In addition to backend development, a user-friendly web application was built using Flask, HTML, CSS, and JavaScript. This enables end-users to easily input patient health data and receive real-time predictions regarding CKD risk. The final system provides not only accurate results but also an interactive and accessible platform for both healthcare professionals and individuals.

Overall, the project showcases how data science and AI can be effectively utilized in the healthcare domain to support early diagnosis, promote preventive care, and improve patient outcomes.

REFERENCES

[1] P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," in IEEE Access, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[2] R. Al-Momani, G. Al-Mustafa, R. Zeidan, H. Alquran, W. A. Mustafa and A. Alkhayyat, "Chronic Kidney Disease Detection Using Machine Learning Technique," 2022 5th International Conference on Engineering Technology and its Applications (IICETA), Al-Najaf, Iraq, 2022, pp. 153-158, doi: 10.1109/IICETA54559.2022.9888564.

[3] S. M. Anusha, V. Chaurasia, and A. Pal, “Chronic kidney disease prediction using machine learning techniques,” Journal of Big Data, vol. 9, no. 1, Art. no. 109, Nov. 2022.

[4] M. Shoaib Arif, A. Ur Rehman, and D. Asif, “Explainable machine learning model for chronic kidney disease prediction,” Algorithms (MDPI), vol. 17, no. 10, Art. no. 443, Oct. 2024.

[5] Md. Ehsanul Haque et al., “Improving chronic kidney disease detection efficiency: Fine tuned CatBoost and nature-inspired algorithms with explainable AI,” arXiv, Apr. 2025.

[6] Z. Dana, A. A. Naseer, B. Toro, and S. Swaminathan, “Integrated machine learning and survival analysis modeling for enhanced chronic kidney disease risk stratification,” arXiv, Nov. 2024.

[7] M. Zisser and D. Aran, “Transformer-based time-to-event prediction for chronic kidney disease deterioration,” arXiv, Jun. 2023.

[8] Kumar K., Pradeepa M., Mahdal M., Verma S., RajaRao M. V. L. N., and Ramesh J. V. N., “A deep learning approach for kidney disease recognition and prediction through image processing,” Applied Sciences, vol. 13, no. 6, Art. no. 3621, Jun. 2023.

[9] K. M. Almustafa, “Prediction of chronic kidney disease using different classification

algorithms," Information in Medicine Unlocked, vol. 24, Art. no. 100631, 2021.

[10] S. P. Singh and P. Yadav, "Machine learning hybrid model for the prediction of chronic kidney disease," Computational Intelligence and Neuroscience, vol. 2023, Art. no. 9266889, 2023.

[11] Rashid, "Artificial neural network and ML techniques for CKD diagnosis," Biomedical & Pharmacology Journal, 2022.

[12] Diagnosis of chronic kidney diseases using machine learning, CISCON 2018 Conference Proceeding, May 2024, pp. 49–61.