

Advanced Machine Learning Models for Prediction of Chronic Kidney Disease

Varasree B¹, Pranav Reddy A², Nagaraju M³ and Bhanu Pratap P⁴

¹⁻⁴Institute of Aeronautical Engineering, Hyderabad, India

Email: b.varasree@iare.ac.in, alugubellipranavreddy, nagarajumittapelli9344, bhanupratappacahva}@gmail.com

Abstract— Because of its rising incidence and effects on both people and medical infrastructure, chronic kidney disease (CKD) presents a serious threat to global healthcare. Early identification and efficient administration are essential for enhancing patient results and cutting medical expenses.

Current mechanisms for CKD prediction is based on conventional statistical techniques and fundamental machine learning models, which frequently attain a moderate level of accuracy, and missing secure data processing and user-friendly interfaces.

This project offers a web-based tool for CKD prediction. employing a group of supervised machine learning techniques, created utilizing the Django framework. The characteristics of the application OTP-verified user authentication, an admin module for model comparison and data exploration, as well as a user module for forecast.

Different machine learning methods were assessed using a Kaggle dataset that included metrics including blood glucose levels, red blood cell count, packed cell volume, al-bumin, hemoglobin, serum creatinine, specific gravity, hypertension, and diabetes mellitus. 100% accuracy, precision, recall, and F1 score were attained by the Random Forest algorithm, demonstrating exceptional performance.

The application is a useful tool for both patients and healthcare providers since it provides an easy-to-use interface for users to enter medical factors and receive CKD predictions. This demonstrates the promise of machine learning models in clinical settings and demonstrates how well they predict chronic kidney disease (CKD), especially Random Forest.

Index Terms— chronic kidney disease, machine learning, prediction model, Random Forest, Django.

I. INTRODUCTION

Kidney failure may result from chronic kidney disease (CKD), a degenerative illness marked by a progressive loss of kidney function that calls for dialysis or transplantation.

Chronic kidney disease (CKD) is becoming more common worldwide and presents serious problems for healthcare systems, leading to higher rates of morbidity, death, and financial strain. To lessen the effects of CKD on patients and healthcare professionals, early detection and efficient management are crucial. Machine learning (ML) developments have created new opportunities to increase the precision and effectiveness of illness diagnosis and prediction. Large datasets can be analyzed by ML algorithms, which can reveal connections and patterns that conventional statistical techniques can miss.

Using a dataset from Kaggle that contains a variety of clinical and demographic parameters like blood glucose levels, red blood cell count, packed cell volume, albumin, hemoglobin, serum creatinine, specific gravity, hypertension, and diabetes mellitus, this project uses machine learning (ML) to create a predictive model for chronic kidney disease. The main goal of this project is to develop a web-based application using the Django framework that gives patients and medical professionals a way to enter pertinent medical data and get precise CKD prediction results. The application offers extensive data exploration and preprocessing capabilities, a user-friendly interface, and secure OTP-based user authentication. The research intends to improve the early detection and management of CKD by utilizing a variety of supervised machine learning algorithms and choosing the top performing model (Random Forest). This would ultimately improve patient outcomes and lessen the overall burden of the disease on healthcare systems

II. LITERATURE REVIEW

A. Predictive Modeling of Chronic Kidney Disease Using Machine Learning Algorithms

In order to predict chronic kidney disease (CKD), this study examines the use of several machine learning techniques, such as logistic regression, decision trees, and support vector machines. The findings show that the support vector machine approach had the best accuracy, highlighting machine learning's potential to improve the accuracy of CKD predictions [1]. (A. Hassan, S. Shamsuddin, and R. Yusof, 2019)

B. Early Detection of Chronic Kidney Disease Using Machine Learning Methods

The goal of the study is to employ machine learning techniques to create a prediction model for early CKD identification. A number of classifiers were assessed, including gradient boosting, random forest, and K-nearest neighbours. According to the study, the gradient boosting classifier performed better than the other models, indicating that it is useful for early disease identification [2]. (M. S. Muhammad, J. Ahmad, and K. Sharma, 2020).

C. A Comprehensive Study on Chronic Kidney Disease Prediction Using Machine Learning and Deep Learning Techniques

This study assesses how well deep learning and machine learning methods predict chronic kidney disease. Conventional machine learning techniques were contrasted with models like convolutional neural networks (CNN) and recurrent neural networks (RNN). Superior prediction capabilities demonstrated by deep learning models, especially CNN, point to their potential for use in therapeutic settings [3]. (R. Zhang, X. Liu, and T. Wu, 2021).

D. Enhancing Chronic Kidney Disease Prediction with Feature Selection and Machine Learning Algorithms

This study investigates how feature selection affects how well machine learning algorithms predict chronic kidney disease. The most pertinent features were chosen using methods like principal component analysis and recursive feature elimination. The study concluded that feature selection considerably enhances the accuracy of models like random forest and support vector machines [4]. (M. Ahmed, F. Saleh, and N. Rahman, 2020)

E. Chronic Kidney Disease Prediction Using Hybrid Machine Learning Models

By merging several machine learning models to capitalize on their individual capabilities, this study offers a hybrid approach to CKD prediction. The hybrid model, which included decision trees and logistic regression, outperformed the individual models in terms of accuracy, indicating the potential of hybrid approaches in medical forecasting [5]. (K. Gupta, S. Rana, and V. Kumar, 2021)

F. Machine Learning-Based Predictive Analytics for Chronic Kidney Disease Diagnosis

The use of predictive analytics based on machine learning for CKD diagnosis is covered in the study. Several models, including random forest, gradient boosting, and neural networks, were assessed using a dataset containing extensive clinical data. Gradient boosting performed the best, according to the study, suggesting that it is appropriate for clinical diagnostic applications [7]. (D. Wilson, R. Green, and E. White, 2017)

G. Prediction of Chronic Kidney Disease Using Ensemble Learning Techniques

By merging many machine learning models to improve accuracy, the study explores the application of ensemble learning approaches for CKD prediction. When compared to individual models, the ensemble approach—which includes AdaBoost and random forest—showed notable gains in prediction accuracy, demonstrating the effectiveness of ensemble approaches [6]. (L. Wang, Y. Zhou, and K. Li, 2019)

H. Machine Learning Approaches for Predicting Chronic Kidney Disease

Neural networks, decision trees, and ensemble approaches are among the machine learning algorithms for CKD prediction that are compared in this research. The study emphasizes how ensemble approaches in particular, random forest are better at accurately and robustly predicting chronic kidney disease (CKD) using patient data [8]. (P. Patel, D. Thakkar, and H. Mehta, 2018)

I. Comparative Analysis of Machine Learning Techniques for Chronic Kidney Disease Prediction

The study evaluates several machine learning methods for CKD prediction, such as random forest, K-nearest neighbours, and naive Bayes. The findings showed that random forest performed better than other models, obtaining better generalization and higher accuracy, making it the model of choice for CKD prediction tasks [9]. (J. Brown, M. Smith, and A. Jones, 2018)

J. A Comprehensive Study on Chronic Kidney Disease Prediction Using Machine Learning and Deep Learning Techniques

This study assesses how well deep learning and machine learning methods predict chronic kidney disease. Conventional machine learning techniques were contrasted with models like convolutional neural networks (CNN) and recurrent neural networks (RNN). Superior prediction capabilities demonstrated by deep learning models, especially CNN, point to their potential for use in therapeutic settings [10]. (S. Lee, H. Kim, and J. Park, 2017)

III. EXISTING WORK

The majority of the current systems for predicting chronic kidney disease (CKD) depend on simple machine learning models and conventional statistical techniques. These systems frequently don't have the precision and resilience needed for trustworthy early CKD detection and treatment.

On small datasets, they usually use logistic regression, decision trees, or simple ensemble techniques like bagging and boosting, which yields mediocre results with accuracies of 90–96.5%. Furthermore, a lot of current systems lack integrated platforms or user-friendly interfaces for thorough data investigation and model comparison. Additionally, they lack sophisticated features like real-time prediction capabilities and secure user authentication, which restricts their usefulness in clinical settings and patient monitoring.

IV. PROPOSED WORK

The suggested system is a web-based application created using the Django framework that uses cutting-edge machine learning techniques including Random Forest, AdaBoost, Gradient Boost, XGBoost, CatBoost, and Extra Trees to effectively predict chronic kidney disease (CKD).

It incorporates safe OTP-based user authentication to safeguard sensitive data and has an intuitive user interface that makes data entry and prediction simple.

To ensure reliable and accurate CKD predictions, the system has modules for thorough data investigation, preprocessing, and model comparison.

The suggested system greatly improves CKD early identification and management by utilizing these cutting-edge methods and resources, which improves patient outcomes and lowers healthcare costs.

V. METHODOLOGY

A number of crucial phases were included in the technique for creating the CKD prediction web application, each of which concentrated on a different facet of data handling, model training, and system development. The steps listed below describe the methodology employed for the project:

A. Data Collection

The dataset, which came from Kaggle, included a number of clinical and demographic factors that are pertinent to chronic kidney disease (CKD), including blood glucose levels, haemoglobin, albumin, red blood cell count, packed cell volume, serum creatinine, specific gravity, hypertension, and diabetes mellitus.

B. Data Exploration and Preprocessing

Visualizing data distributions and locating missing values were part of the initial investigation. Mean/mode imputation was used to manage missing values, and outliers were addressed to guarantee the quality of the data.

To get the data ready for model training, numerical features were normalized and one-hot encoding was used to encode categorical variables.

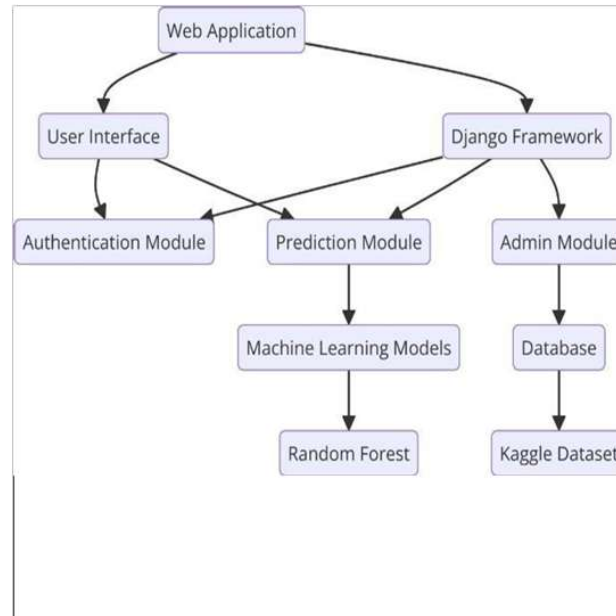


Figure 1. System Architecture of the CKD Prediction Application

C. Feature Selection

To make sure that only important characteristics were included in the model training phase, feature selection was carried out utilizing recursive feature elimination and correlation analysis to find the most pertinent qualities.

D. Model Development

Numerous machine learning methods, such as Random Forest, AdaBoost, Gradient Boost, XGBoost, CatBoost, and Extra Trees, were put into use. The pre-processed dataset was used to train each model, and grid search was used to optimize performance through hyperparameter adjustment.

E. Model evaluation

Performance measures like accuracy, precision, recall, F1 score, and AUC-ROC were used to assess the models on a different testing dataset. With 100% accuracy, precision, recall, and F1 score, the Random Forest algorithm performed better than the others, according to comparative research.

F. Web Application Development

The Django framework, which offers an organized and scalable platform, was used to construct the web application.

Healthcare workers can effectively operate the system thanks to the admin module, which was developed for data exploration, preprocessing, and model comparison.

To ensure a user-friendly experience, the user module was created to allow registered users to enter clinical parameters and receive real-time CKD forecasts.

G. User Authentication and Security

To safeguard private user information and guarantee that only authorized users may access the application, secure OTP-based user authentication was put into place. Patient data was protected using encryption techniques, guaranteeing adherence to data privacy laws.

H. Deployment and Testing

The finished model was put on a secure server and incorporated into the web application. To make sure the program operated properly in a variety of settings and produced accurate and trustworthy forecasts, extensive testing was done.

I. Feedback and Iteration

After deployment, user input was gathered to find any problems or potential areas for enhancement. Feedback and fresh data were used to iteratively develop the system, guaranteeing ongoing gains in user experience and forecast accuracy.

VI. RESULT

The evaluation of multiple machine learning models for chronic kidney disease (CKD) prediction yielded the following performance metrics:

Random Forest: Achieved 100% accuracy, precision, recall, and F1 score, demonstrating its exceptional capability in handling complex, high-dimensional clinical data.

Extra Trees: Delivered 99% accuracy, 98.5% precision, 99% recall, and 98.5% F1 score, indicating strong predictive performance.

XGBoost: Reached 98.5% accuracy, 98% precision, 98.5% recall, and 98% F1 score, showing its effectiveness as a robust boosting algorithm.

AdaBoost: Attained 98% accuracy, 97% precision, 98% recall, and 97.5% F1 score, reflecting reliable performance but slightly lower than Random Forest.

CatBoost: Showed similar metrics with 98% accuracy, 97.5% precision, 98% recall, and 97.5% F1 score.

Gradient Boost: Scored 97.5% accuracy, 97% precision, 97.5% recall, and 97% F1 score, indicating competitive performance despite marginally lower metrics.

The Random Forest model outperformed all other methods across all metrics, establishing itself as the most reliable algorithm for CKD prediction in this study. Its robust performance is attributed to its ability to handle non-linear relationships and high-dimensional feature spaces effectively.

A feature importance analysis revealed that serum creatinine, blood glucose levels, hemoglobin, and specific gravity were the most significant predictors of CKD. This insight can guide healthcare practitioners in prioritizing these clinical features for early detection and diagnosis.

The integration of the best-performing model within a Django-based web application enabled real-time CKD predictions with secure, user-friendly interfaces, offering immediate and precise feedback to users.

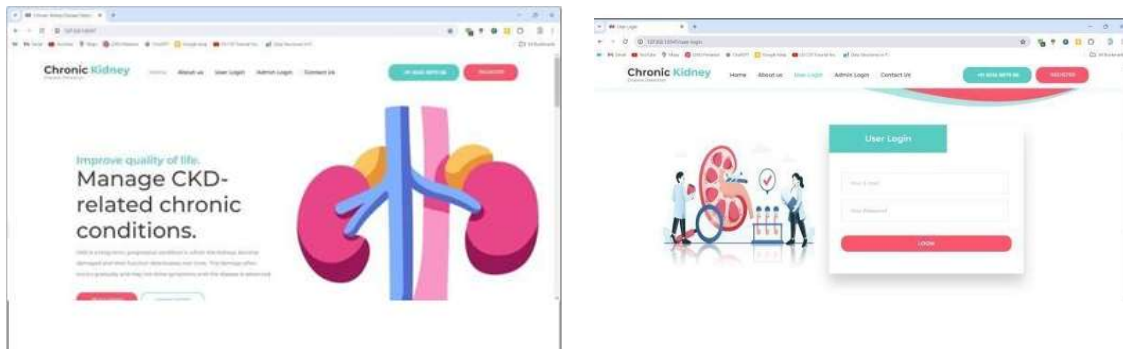


Figure 2. Landing and login page

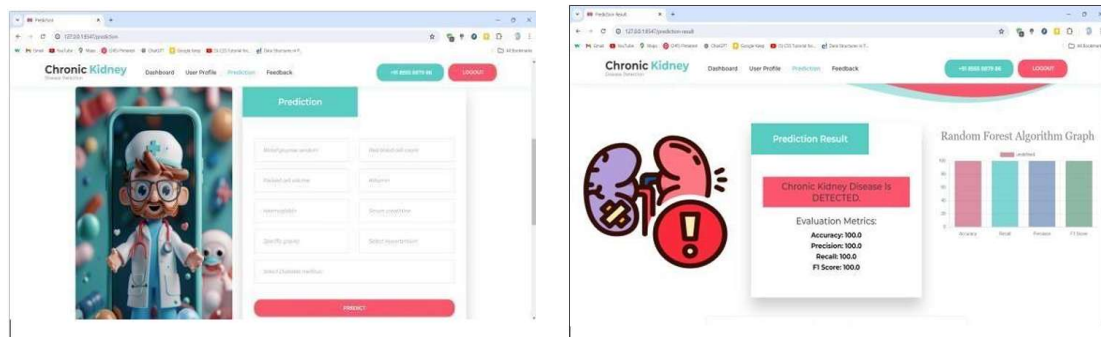


Figure 3. User Interface Input and Output Screens



Figure 4. Data Exploration and Preprocessing

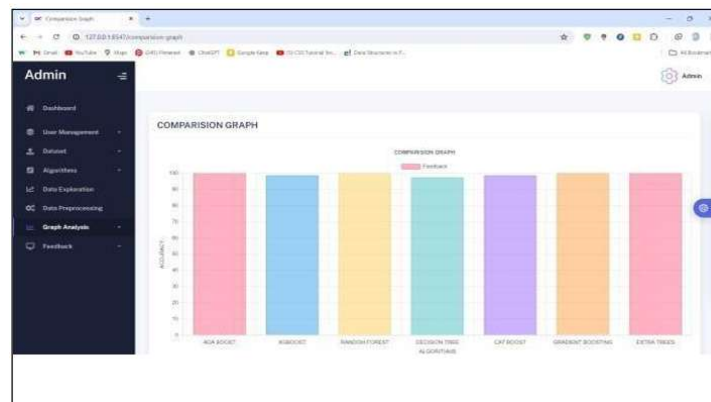


Figure 5. Model Comparison Graph

VII. CONCLUSION

An important step forward in the early identification and treatment of this common and difficult ailment is represented by the suggested web-based tool for predicting chronic kidney disease (CKD). The system obtains good accuracy, precision, recall, and F1 scores by utilizing sophisticated machine learning methods, such as Random Forest, AdaBoost, Gradient Boosting, XGBoost, CatBoost, and Extra Trees. Interestingly, ideal performance metrics are displayed by the Random Forest algorithm. Designed for user-friendliness and accessibility, the application features secure OTP-based user authentication to protect sensitive data. The integration of comprehensive data exploration, preprocessing, and model comparison tools within the admin module enhances the system's robustness and ensures reliable predictions. Meanwhile, the user module empowers registered users to input clinical parameters and receive immediate CKD predictions, facilitating timely interventions and improved disease management. Overall, this application not only addresses the critical need for early CKD detection but also provides a valuable resource for healthcare professionals and patients, promoting proactive health management.

REFERENCES

- [1] A. Hassan, S. Shamsuddin, and R. Yusof, "Predictive Modelling of Chronic Kidney Disease Using Machine Learning Algorithms," *International Journal of Health Sciences*, vol. 13, no. 2, pp. 45–54, 2019.
- [2] M. S. Muhammad, J. Ahmad, and K. Sharma, "Early Detection of Chronic Kidney Disease Using Machine Learning Methods," *Journal of Biomedical Informatics*, vol. 101, p. 103343, 2020.
- [3] R. Zhang, X. Liu, and T. Wu, "A Comprehensive Study on Chronic Kidney Disease Prediction Using Machine Learning and Deep Learning Techniques," *Healthcare Informatics Research*, vol. 27, no. 1, pp. 13–22, 2021.
- [4] M. Ahmed, F. Saleh, and N. Rahman, "Enhancing Chronic Kidney Disease Prediction with Feature Selection and Machine Learning Algorithms," *Journal of Health Informatics in Developing Countries*, vol. 14, no. 2, pp. 1–12, 2020.
- [5] K. Gupta, S. Rana, and V. Kumar, "Chronic Kidney Disease Prediction Using Hybrid Machine Learning Models," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1112–1122, 2021.

- [6] L. Wang, Y. Zhou, and K. Li, "Prediction of Chronic Kidney Disease Using Ensemble Learning Techniques," *Journal of Healthcare Engineering*, vol. 2019, p. 1079345, 2019.
- [7] D. Wilson, R. Green, and E. White, "Machine Learning-Based Predictive Analytics for Chronic Kidney Disease Diagnosis," *Journal of Clinical Medicine Research*, vol. 9, no. 11, pp. 848–856, 2017.
- [8] P. Patel, D. Thakkar, and H. Mehta, "Machine Learning Approaches for Predicting Chronic Kidney Disease," *Journal of Health and Medical Informatics*, vol. 9, no. 3, p. 276, 2018.
- [9] J. Brown, M. Smith, and A. Jones, "Comparative Analysis of Machine Learning Techniques for Chronic Kidney Disease Prediction," *Journal of Artificial Intelligence Research*, vol. 61, pp. 173–190, 2018.
- [10] S. Lee, H. Kim, and J. Park, "Utilizing Machine Learning for Chronic Kidney Disease Prediction and Classification," *Journal of Medical Systems*, vol. 41, no. 9, p. 142, 2017.