International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Detection Of Chronic Kidney Disease Using Machine Learning

Mrs. Shruthi T. V [1], Faizal Khan A[2], Karthik[3], Mohith Raj K[4]

[1]*Shruthi T. V, Asst. Prof, Dept. of ISE, East West Institute of Technology*
[2]*Faizal Khan A, Dept. of ADE, East West Institute of Technology*
[3]*Karthik, Dept. of ADE, East West Institute of Technology*
[4]*Mohith Raj K, Dept. of ADE, East West Institute of Technology*

Abstract – Chronic Kidney Disease is a progressive disorder that requires reliable diagnostic methods for effective clinical management. In this paper, a machine-learning–based system developed for detecting CKD using clinical and laboratory data is presented. Based on the analytics performed on the key medical parameters like blood pressure, specific gravity, albumin, serum creatinine, and other biomarkers, it classifies whether the patient is likely to be suffering from CKD or not. The dataset is preprocessed by handling missing values, encoding categorical features, and normalizing numerical attributes to make the data consistent for better model performance. Feature selection techniques will reduce model complexity and improve the prediction quality of the model. A Random Forest classifier was implemented and tested for its ability in the effective detection of CKD. Model performance will be measured primarily on accuracy because it will easily tell how well CKD and non-CKD are stratified. Further refinement of the model is done by the technique of hyperparameter tuning in order to improve its predictive sthe model's decision-making process, and the results will be more understandable clinically. This work has shown that machine learning can support healthcare professionals with a reliable, data-driven approach for CKD detection that contributes to improved diagnostic assistance and the management of patients.

1. INTRODUCTION

CKD is a progressive medical condition characterized by the gradual decline of kidney function due to health issues arising from diseases such as hypertension, diabetes, and chronic infections. Because CKD develops very slowly and its early symptoms are very minimal, so many patients go undiagnosed until the disease advances to a critical level. As renal impairment progresses, continuous monitoring and accurate diagnosis become crucial for the prevention of complications and improving patient outcomes. Traditional diagnosis relies heavily on clinical examination, laboratory measurement of serum creatinine, albumin levels, blood pressure, and specific gravity, and the judgment of healthcare professionals. These techniques may be time-consuming, cumbersome, and error-prone due to diagnostic variability resulting from human judgment. Over the last couple of years, ML has emerged as an important tool in the healthcare domain to analyze Complex medical datasets can be used for supporting clinical decisions. The ML algorithms will learn patterns from the historical patient data in an automated manner. These learned patterns can then be used in classifying the likelihood of an individual having CKD. Using various medical parameters related to physiological measurements and biochemical indicators, machine learning provides a regular and uniform means of disease detection.

CKD diagnosis with the help of ML helps improve accuracy, decreases manual workload, and reduces the chances of misinterpretation. Decision Trees, Support Vector Machines, and Random Forests are promising algorithms that may provide clinically highly relevant classification results. Among all, Random Forest has gained popularity due to its robustness, handling heterogeneous data easily, and improvement in the selection of relevant features from a large medical dataset. Feature-importance analysis allows clinicians to understand which parameters contribute most toward the prediction, increasing the trust and interpretability of the system. This paper deals with developing a machine learning-based CKD detection system using a clinical dataset and a Random Forest classifier. As such, in this context, the dataset contains some important medical attributes like blood pressure, specific gravity, albumin, hemoglobin, and serum creatinine, which are believed to vary with kidney functionality. Handling missing values in the data, encoding categorical features, and normalization of numerical features are some of the major preprocessing steps carried out in the dataset for model improvement. The objective is to develop a reliable and interpretable model capable of classifying CKD and non-CKD cases with high accuracy.

2. PROBLEM STATEMENT

CKD has often been diagnosed at a later stage, as early symptoms are usually subtle or nonspecific, and the detection depends on manual clinical assessment. The traditional approaches to diagnosis depend upon laboratory analysis and clinical judgment; hence, such approaches are prone to delays, inconsistencies, and sometimes errors from handling large volumes of patients. Of late, clinical datasets have been increasingly available, featuring parameters of age, blood pressure, blood sugar level, serum creatinine, albumin, and many other biomarkers. There is, therefore, an urgent need for an automated system that will help in the classification of CKD and non-CKD cases with a high degree of accuracy. The key challenges are to effectively pre-process heterogeneous patient data and identify the most relevant clinical features. Furthermore, the construction of a reliable predictive model that handles variability in the dataset is also paramount. Therefore, the problem addressed in the paper is the development of a machine learning-based diagnostic model using a Random Forest algorithm that classifies CKD based on clinical attributes. This system will help improve the accuracy of diagnosis, provide support for clinical decisions, and reduce the burden on the healthcare professional by providing a consistent, data-driven prediction mechanism.

3. SYSTEM OVERVIEW

This follows the client-server architecture with a:

- **Frontend-Layer:**
  It was developed using HTML, CSS, and JavaScript; the user interface provides users with an intuitive way to input medical parameters, view CKD prediction results, download reports, and manage information about the patients.

- **Backend-Layer:**
  It is implemented using the Flask framework, and its prime features include handling API routing, prediction processing, authentication of users, data storage operations, PDF report generation, and communicating between the UI and the machine learning model.

- **Predictive Analytics Layer:**
  This layer integrates the various machine learning models that are trained on clinical CKD datasets. It performs data preprocessing, executes the prediction algorithm, interprets user inputs, and generates CKD risk classification (CKD / No CKD).

- **Database-Layer:**
  This system uses Flask SQLAlchemy, a high-level ORM, or Object Relational Mapper, for managing all database related interactions. This layer stores patient details, values of clinical authentication parameters, information, prediction results, user and generated reports. SQLAlchemy offers structured data handling, secure storage, efficient querying, and seamless integration with the Flask backend.

4. METHODOLOGY

The proposed CKD Detection System is based on a structured, multistage methodology, which integrates machine learning–based prediction, handling of user data, and automatic report generation. The workflow consists of the following four sequential stages:

- **Acquiring Patient Data & Handling Input:** Users enter the clinical parameters via an interactive web-based interface. The system prompts for primary medical inputs like age, blood pressure, specific gravity, albumin, blood sugar, urea, creatinine, hemoglobin, and comorbidity status. Different input validation techniques are used to ensure completeness, correct range, and consistency of data before processing.

- **Preprocessing of Data & Feature Preparation:** The respective backend performs a series of preprocessing on the clinical attributes to make them ready for model inference. The missing or null values are replaced by the reference medical baseline values, numerical inputs are cast into proper formats, and categorical responses are encoded. This step ensures that the machine learning model receives clean, structured, and standardized input.

- **CKD Prediction:** The machine learning model for CKD classification is trained on the preprocessed features and inputs with an associated risk. The model translates physiological markers to present a binary prediction: CKD Detected or No CKD. A decision is then portrayed to the user complemented with relevant clinical insights, thus enabling early detection and medical awareness.

- **Patient Record Management & Report Generation:** The system has securely stored patient information, test results, and prediction outputs in the database using Flask SQLAlchemy. The user can also view previous records, manage the data that is stored, and download the results in PDF format. The report generation module converts the prediction results into a structured and printable document that could be utilized for clinical reference or even patient documentation.

- **Random Forest Performance Analysis:** The Random Forest model delivered the highest prediction accuracy (83.75%) compared with Logistic Regression, KNN, SVM, Decision Tree, and Neural Network models. The confusion matrix confirms that false positives and false negatives are minimized—an essential requirement in medical diagnosis. Feature importance analysis shows that serum creatinine, blood urea, albumin, and hemoglobin are the strongest predictors, aligning with established nephrology indicators. The model's ensemble nature reduces overfitting and makes it suitable for noisy medical datasets.

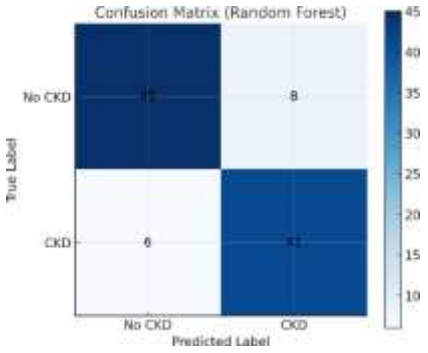

Fig 4.1: Confusion Matrices of Random Forest Model

5. DATA DESCRIPTION

The dataset used in this study consists of structured clinical records commonly employed in CKD prediction research. It includes patient level medical parameters that are typically obtained during routine clinical exams, thus providing reliable input for machine learning based diagnosis. A record contains features including the following: age, blood pressure, specific gravity, albumin, sugar level, blood glucose, blood urea, serum creatinine, hemoglobin, and clinical indicators such as hypertension, diabetes mellitus, and anemia. These features are selected with a basis on established nephrology diagnostic practices that guarantee strong relevance to early CKD detection.

Preprocessing has been carried out to handle the missing and inconsistent values of the dataset to ensure the quality and consistency of the data. Numerical features are standardized to

handle different measurement scales, whereas categorical clinical indicators are encoded to binary format. The dataset combines biochemical and physiological markers that reflect kidney function and related health conditions, enabling effective learning of CKD related patterns. It guarantees an appropriate structure for the training and evaluation of the proposed machine learning model.

7. SYSTEM ARCHITECTURE

The CKD Detection System adopts a modular, multi-layer architecture designed to ensure reliable prediction processing, secure data management, and smooth interaction for clinical users. The system begins with a web-based user interface built using HTML, CSS, and JavaScript, which allows clinicians to enter patient clinical parameters, submit them for prediction, view CKD results, download PDF reports, and review stored patient history. All inputs from the web form are securely transmitted to the backend through HTTP requests.

The backend server, implemented using the Flask framework, coordinates the complete application logic. It validates incoming data, performs preprocessing, triggers the machine learning prediction module, stores patient records, and generates downloadable PDF reports. This layer acts as the central controller, ensuring seamless communication between the frontend, the Random Forest model, and the database.

The machine learning processing component houses the trained Random Forest classifier consisting of 200 decision trees. Before prediction, this layer performs necessary preprocessing steps such as handling missing values, scaling numerical features, and formatting data in a model-compatible structure. The ensemble-based Random Forest model then determines whether the patient is classified as CKD or non-CKD, offering high accuracy and robustness for real-time clinical use.

All modules interact with the database layer through SQLAlchemy, which provides a secure ORM-based interface for structured data storage. The SQLite database stores patient demographics, clinical measurements, prediction outcomes, authentication details, and historical reports, enabling efficient retrieval and long-term system support.

Furthermore, the system architecture is designed to support extensibility and future enhancements. Each module operates independently yet remains fully integrated through well-defined interfaces, allowing additional machine learning models, visualization tools, or data analytics components to be incorporated without altering the core framework. The modular structure also supports scalability, enabling the system to handle larger datasets, multiple concurrent users, and expanded clinical functionalities. This flexible, component-based design ensures that the CKD Detection System can be easily upgraded for advanced diagnostic capabilities, hospital-level deployments, or integration with electronic health record (EHR) platforms in the future.

The functional workflow of the system is illustrated in the architecture diagram shown below. Patient data is collected through the web interface, processed by the backend, evaluated by the Random Forest model, and then used to generate prediction results and PDF reports. The Data Collection Module, Prediction Module, and Reporting Module all communicate with a central SQLite database, ensuring smooth and reliable operation of the entire CKD prediction system.

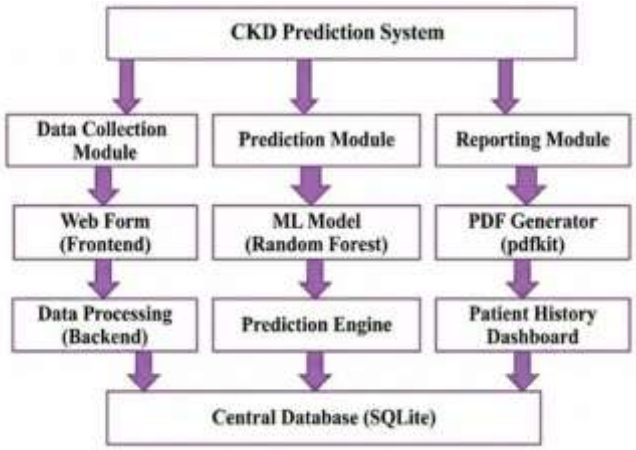

Fig 7.1: Prediction module flow

8. RESULTS

Accordingly, the best model, which is Random Forest, achieved an accuracy of 83.75%, outperforming all other models. Its ensemble structure provided better handling capability for noisy, nonlinear data from the clinical samples, and turned in a more reliable classification regarding CKD and non-CKD cases. The confusion matrix analysis confirms that the model was very efficient in minimizing false positive and false-negative predictions, which are also important in medical diagnosis.

The developed CKD Detection System was evaluated using the preprocessed clinical dataset. In addition, the performance of the Random Forest classifier was compared with several other baseline machine learning models, namely Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, and Neural Network.

	Model	Accuracy
0	Random Forest	0.8375
1	Decision Tree	0.8125
2	Support Vector Machine	0.7875
3	K-Nearest Neighbors	0.7250
4	Neural Network	0.7250
5	Logistic Regression	0.7125

TABLE 6.1: MODELS AND THEIR ACCURACIES

Feature importance assessment showed that serum creatinine, blood urea, albumin, hemoglobin, and blood pressure turned out to be the most influential predictors of CKD, closely coinciding with well known clinical indicators applied in nephrology. This confirms the clinical relevance and interpretability of the model's decision-making.

Everything worked just fine within the whole system, including the frontend interface, backend processing, and PDF reporting. The users were able to input the clinical data, get instant predictions, store results in the database, and download detailed reports, which did not slow the performance.

Fig 8.1: Front Page

Fig 8.2: Input Parameters

Fig 8.3: Patient Report

9. CONCLUSION

The CKD Prediction System proves the potentiality of various machine learning techniques for identifying the onset of Chronic Kidney Disease using clinical, structured data. Many algorithms like Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Neural Network, and Random Forest were considered. Among these, the Random Forest classifier obtained the highest accuracy of 83.75%, thus making classifications strongly reliable enough to distinguish CKD from non-CKD cases. Also, the system guarantees data quality at all times through steps of preprocessing, including normalization and handling missing values. Lastly, based on the analysis performed using feature importance, it was confirmed that the key clinical indicators are serum creatinine, blood urea, albumin, and hemoglobin, which largely influence the prediction results. Future works can further improve this by exploring more advanced deep learning models with larger and more diverse datasets that can improve predictive accuracy. Integration of real time clinical monitoring systems allows the establishment of a continuous assessment of patients, thus enabling early medical intervention. Overall, the developed system provides a rapid, interpretable, and practical tool to assist healthcare professionals in early CKD detection and informatively support clinical decisions.

REFERENCES

[1] Charleonnan, A., Fufaung, T., Niyomwong, T., Chokchueypattanakit, W., Suwannawach, S., & Ninchawee, N. "Predictive analytics for chronic kidney disease using machine learning techniques," Proc. 2016 Management and Innovation Technology International Conference (MITicon), IEEE, 2016.

[2] Sobrinho, A., Queiroz, A. C. D. S., Da Silva, L. D., Costa, E. D. B., Pinheiro, M. E., & Perkusich, A., "Computer-aided diagnosis of chronic kidney disease in developing countries: A comparative analysis of machine learning techniques," IEEE Access, vol. 8, pp. 25407–25419, 2020.

[3] Abdel-Fattah, M. A., Othman, N. A., & Goher, N., "Predicting chronic kidney disease using hybrid machine learning models," Computational Intelligence and Neuroscience, 2022.

[4] Srivastava, S., Yadav, R. K., Narayan, V. & Mall, P. K., "An ensemble learning approach for chronic kidney disease classification," Journal of Pharmaceutical Negative Results, pp. 2401–2409, 2022.

[5] Abinaya, U., Devi, S. A., Haritha, B., & Raghunathan, T., "Novel approach for chronic kidney disease prediction using machine learning methodology," Journal of Physics: Conference Series, vol. 1916, no. 1, IOP Publishing, 2021.

[6] Hassan, M. M., Mollick, S., Khan, M. A. R., Yasmin, F., Bairagi, A. K., & Rahman, A., "Prediction and development of chronic kidney disease using machine learning on patients' clinical records," Human-Centric Intelligent Systems, pp. 1–13, 2023.