International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04779
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Chronic Kidney Disease Prediction Using Machine Learning

CH.VASUNDHARA, DALAI MURALI

Assistant Professor, MCA Final Semester, Master of Computer Applications, Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India

Abstract
This project is designed to predict Chronic Kidney Disease(CKD) using machine learning and provide an easy- to-use web interface for healthcare professionals. It begins by loading a medical dataset and cleaning it by handling missing values and converting incorrect data formats. Next, it performs basic data analysis through visualizations like class distribution, feature correlation, and hemoglobin levels to understand patterns in the data.The project uses a machine learning pipeline where numeric and categorical data are preprocessed separately—missing values are filled, and categorical data is encoded properly. A Random Forest model is then trained to classify whether a patient has CKD or not, using features like age, blood pressure, blood test results, and more. The model's accuracy and performance are evaluated using standard metrics like accuracy score, classification report, and a confusion matrix.After training, the model is saved and integrated into a user-friendly Gradio interface. This allows doctors or users to enter patient details using sliders and dropdown menus and instantly get a prediction about the presence of CKD along with the confidence level. This tool helps in early detection and supports doctors in making faster, data-driven decisions.

Index Terms: Chronic Kidney Disease (CKD) Machine Learning,Random Forest Classifier,Preprocessing,Medical Diagnosis, Feature Engineering,Biomedical Data ,Classification Algorithm, Python,Data Visualization,CKD Detection,Model Evaluation, Early Disease Detection

1. INTRODUCTION

Chronic Kidney Disease (CKD) is a serious and progressive condition that affects the kidneys' ability to filter waste and fluids from the blood [1]. If not detected early, it can lead to kidney failure, requiring dialysis or transplantation [6]. In many cases, CKD remains unnoticed during its initial stages due to the absence of obvious symptoms, which makes early diagnosis crucial for timely treatment and improved patient outcomes [3]. With the rise of healthcare data and recent advances in machine learning, there is now significant potential to create predictive models that support early disease detection [2]. This project focuses on building an intelligent system that predicts the likelihood of CKD using various clinical and laboratory parameters [5]. A Random Forest classifier is employed in the machine learning pipeline, trained on real patient data, to achieve accurate risk identification [14]. To ensure accessibility for non-technical users such as doctors and medical staff, the system is integrated with a Gradio-based web interface [15]. This interface enables users to input patient data via sliders and dropdown menus and receive immediate predictions along with confidence scores [20]. The overall goal is to provide healthcare professionals with a dependable and user-friendly tool that aids in faster and more accurate diagnosis of CKD [7].

1.1 Existing System

In the current healthcare system, the diagnosis of Chronic Kidney Disease (CKD) is largely dependent on manual analysis of various clinical and laboratory reports by healthcare professionals [3]. Doctors typically examine patient history and assess multiple test results, such as blood pressure, serum creatinine, blood urea, hemoglobin levels, and urinalysis, to determine the presence or progression of kidney disease [6]. This process is time-consuming and prone to human error, especially when dealing with large volumes of patient data [4]. Additionally, the existing system often lacks automation and early prediction capabilities [1]. Most cases of CKD are diagnosed only after the disease has progressed to a moderate or advanced stage, reducing the chances of successful treatment or management [5]. There is also no standardized or intelligent decision-support tool to assist clinicians in quickly identifying high-risk patients based on patterns in their clinical data [7]. Moreover, the absence of data-driven methods and predictive analytics in routine check-ups limits the ability of healthcare providers to take preventive measures [10]. In rural or resource-limited settings, where access to specialized nephrologists may be minimal, early detection becomes even more challenging [19].

1.1.1 Challenges:

- **Missing and Incomplete Data**

Many entries in the dataset contained missing values or placeholder symbols (like '?'), which required careful handling through imputation or removal to avoid bias or data loss.

- **Mixed Data Types**

Several features were stored as strings (e.g., 'pcv', 'wc', 'rc') even though they represent numeric values. These had to be correctly

converted to ensure accurate model training.

- **Imbalanced Dataset**

The dataset often has more CKD cases than non-CKD (or vice versa), which can lead to biased predictions. Special care (like stratified splitting) was needed to maintain class balance during training.

1.2 Proposed system:

The proposed system is an intelligent, machine learning–based solution designed to predict chronic kidney disease (CKD) at an early stage using clinical and laboratory test data [1]. It addresses the limitations of traditional manual diagnosis methods by automating the analysis process and providing quick, accurate predictions through a user-friendly interface [5]. This system uses a Random Forest Classifier integrated into a machine learning pipeline that performs automated data preprocessing, feature handling, model training, and evaluation [14]. It supports both numeric and categorical data inputs, with missing values handled using median and mode imputation, and categorical features encoded using one-hot encoding [4]. To make the system accessible to medical professionals without programming knowledge, it includes a Gradio-based web interface, where users can input patient data using sliders and dropdown menus [15].

Upon submission, the system instantly predicts whether the patient has CKD and displays the confidence level of the prediction [20].

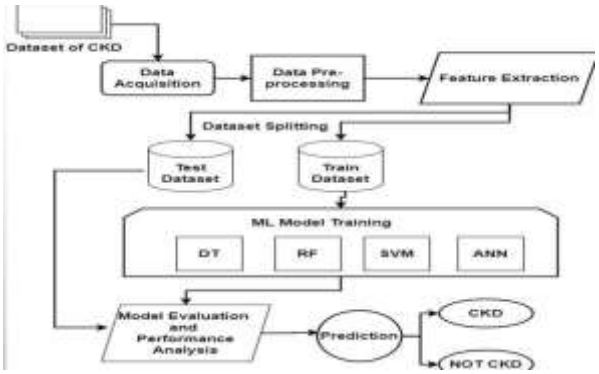

Fig: 1 Proposed Diagram

1.2.1 **Advantages**:

- **Early Detection of CKD**

The system enables timely identification of CKD, which helps in starting early treatment and improving patient outcomes.

- **Automation Reduces Manual Work**

Automates the prediction process, reducing the dependency on manual diagnosis and minimizing human error.

- **High Accuracy with Random Forest**

Utilizes a Random Forest model known for high performance and reliability, ensuring better classification results.

- **Handles Missing Data Efficiently**

Built-in preprocessing handles missing or incomplete medical records using imputation techniques, making the system robust.

2.1 **Architecture:**

The architecture of the CKD prediction system is designed as an end-to-end machine learning pipeline integrated with a user-friendly web interface [15]. It starts with a data layer where the system loads the CKD dataset (*kidney_disease.csv*) containing clinical and laboratory features such as age, blood pressure, hemoglobin, blood urea, and others [5]. The preprocessing stage handles missing values by replacing placeholders and applies appropriate type conversions, especially for columns incorrectly stored as text [4]. Non-informative columns like patient IDs are dropped to avoid noise and enhance model focus on relevant features [14]. Next, the cleaned dataset is passed to a processing pipeline built using Scikit-learn's Pipeline and ColumnTransformer [1]. Numerical features are processed using median imputation, while categorical features are handled using mode imputation followed by one-hot encoding, ensuring that the data is properly formatted for machine learning [10]. The processed data is then used to train a Random Forest Classifier, which is known for its high accuracy and robustness on structured healthcare data [2]. The model is evaluated using accuracy score, classification report, and a confusion matrix to measure its performance in predicting CKD [6].

Fig:2 Architecture

2.2 Algorithm:

The primary algorithm used in this project is the Random Forest Classifier, a powerful ensemble method widely adopted for classification tasks [14]. Random Forest builds multiple decision trees during training and outputs the class that is the mode of the individual trees' predictions, an approach that reduces overfitting, boosts accuracy, and accommodates both numerical and categorical features effectively [5]. For Chronic Kidney Disease (CKD) prediction, the algorithm evaluates patient data—such as age, blood pressure, blood glucose, hemoglobin levels, and other clinical indicators—to gauge the likelihood of CKD presence [1]. It is well-suited to medical datasets because it can tolerate missing values and remains robust against noisy data [2]. The model is trained on a cleaned, preprocessed dataset where numerical and categorical variables are handled separately to maximize learning quality [10]. Its performance is then assessed with metrics like accuracy, a detailed classification report, and a confusion matrix to verify reliability [11]. Finally, the trained model is deployed through a Gradio-based web interface, enabling healthcare professionals to obtain rapid, confidence-scored CKD predictions from user-entered data [15].

2.3 Techniques:

In this project, several essential data preprocessing techniques are applied to prepare the dataset for effective machine learning [4]. One of the key steps is imputation, which involves filling in missing values in the dataset [5]. For numerical features, missing values are typically replaced using statistical methods such as the mean or median, ensuring that the data remains consistent and usable [10]. For categorical features, the most frequent (mode) value is often used to fill in the missing entries, preserving the distribution of the original data [14]. Another important preprocessing step is encoding categorical variables, since machine learning models require numerical inputs [2]. This is achieved using techniques like Label Encoding—where each category is assigned a unique integer—or One-Hot Encoding, which creates binary columns for each category, allowing the model to understand categorical information without assuming any ordinal relationship [11]. These preprocessing steps are crucial for building a robust and accurate predictive model, especially in healthcare datasets that often include missing or categorical data [6].

2.4 Tools:

In this project, several tools and technologies are used to build an end-to-end CKD prediction system [1]. The core development is done using Python, a versatile programming language widely adopted in data science and machine learning [5]. For data manipulation and analysis, libraries like Pandas and NumPy are used to handle structured data and perform numerical operations efficiently [14]. Matplotlib and Seaborn are employed for data visualization, enabling graphical analysis of feature distributions and correlations [4]. The machine learning model is built using scikit-learn (sklearn), which provides tools for preprocessing, model training, evaluation, and saving the trained model [10]. Specifically, the Random Forest Classifier from sklearn is used for prediction due to its robustness and accuracy in structured healthcare data [2]. To make the model accessible through an interactive user interface, the project uses Gradio, a Python-based library that enables the easy creation of web apps for machine learning models [15]. The trained model is integrated with Gradio to allow healthcare professionals to input patient data and receive CKD predictions instantly [20]. Together, these tools create a seamless pipeline from data preprocessing to user-friendly deployment [16].

2.5 Methods:

The project follows a structured approach involving multiple methods to ensure accurate CKD prediction [4]. Initially, data cleaning methods are used to handle missing values and correct inconsistent data types [5]. For missing data, imputation methods such as mean, median, or mode are applied depending on whether the feature is numerical or categorical [11]. Feature encoding methods, including label encoding and one-hot encoding, are used to convert categorical variables into a machine-readable numerical format [2]. The dataset is then split into training and testing sets using the train-test split method to evaluate model generalization [14]. A Random Forest Classifier is trained using supervised learning, where the algorithm learns patterns from labeled training data [1]. To evaluate the model's performance, methods like accuracy score calculation, classification report generation, and confusion matrix plotting are used [6]. After validating the model, model serialization methods (e.g., using joblib) are employed to save the trained model [15]. Finally, Gradio interface methods are used to create a web- based user interface, allowing real-time predictions by inputting patient data [16]. These combined methods ensure that the system is accurate, reliable, and easy to use for early CKD

detection [7].

III. METHODOLOGY

3.1 Input:

In this project, input information is provided through a Gradio web interface where users enter patient medical details required for CKD prediction [16]. Inputs include features like age, blood pressure, blood sugar, hemoglobin levels, and categorical values such as red blood cell type or pus cell presence [5]. These are entered using sliders for numerical values and dropdown menus for categorical ones, ensuring ease of use and consistency [15]. Once all fields are filled, clicking the "Predict CKD" button gives an instant prediction along with the model's confidence score, making the tool user-friendly and practical for healthcare professionals [20].

```
# ------------------------------------------------------------
# 4. Exploratory Data Analysis  – three quick plots
numeric_cols = ['age','bp','sg','al','su','bgr','bu','sc','sod','pot','hemo','pcv','wc','rc'

# 4.1 Class distribution
plt.figure(figsize=(6,4))
sns.countplot(x='classification', data=df, palette='Set2')
plt.title('CKD vs. Non-CKD Cases')
plt.show()
```

Fig 1: data analysis

```
def predict_ckd(*args):
    cols = list(input_components.keys())
    input_df = pd.DataFrame([args], columns=cols)
    pred = model.predict(input_df)[0]
    prob = model.predict_proba(input_df)[0][pred]
    label = 'CKD Detected' if pred == 1 else 'No CKD'
    return { 'Prediction': label, 'Probability': f"{prob:.2%}" }
```

Fig 2: prediction of C

3.2 Method of Process:

The method of process in this CKD prediction project follows a step-by-step machine learning pipeline to ensure accurate and reliable results [4]. It begins with data collection and cleaning, where missing values are filled and incorrect formats are corrected [5]. Next, data preprocessing is performed, including imputation and encoding of categorical features [11]. The dataset is then split into training and testing sets to evaluate the model's generalization capability [14]. A Random Forest Classifier is trained using the processed training data [1]. After training, the model is evaluated using accuracy, classification report, and confusion matrix to measure its performance [6]. Once validated, the model is saved using serialization techniques and deployed using a Gradio web interface [15]. This allows users to input patient data and receive real-time CKD predictions [16].

3.3 Output:

The output of this CKD prediction project is a clear and user-friendly result displayed through the Gradio interface [16]. Once the user inputs patient details and clicks the "Predict CKD" button, the model returns a prediction indicating whether the patient is likely to have CKD (Chronic Kidney Disease) or not [20]. Along with this prediction, the system also displays the confidence score (e.g., 92% confidence) to show how certain the model is about its result [15]. This output helps healthcare professionals make quick and data-driven decisions regarding early diagnosis and treatment [7].

.

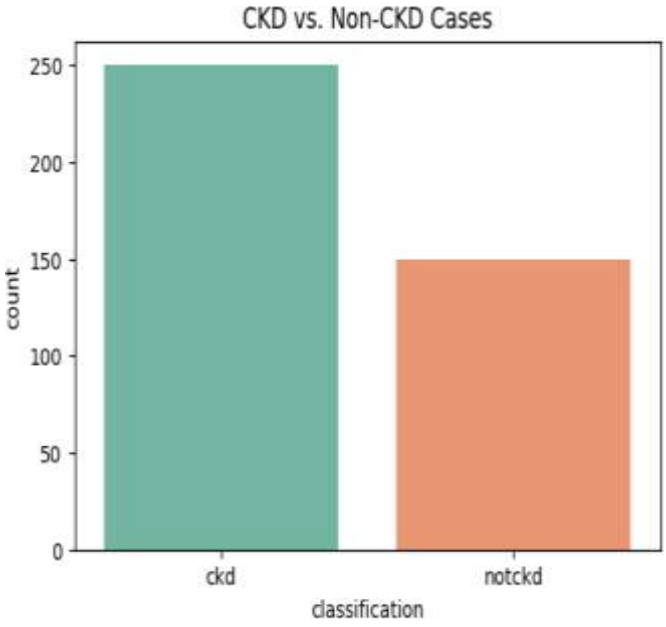

Fig:classification report

```
Accuracy: 1.0000

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        30
           1       1.00      1.00      1.00        50

    accuracy                           1.00        80
   macro avg       1.00      1.00      1.00        80
weighted avg       1.00      1.00      1.00        80
```

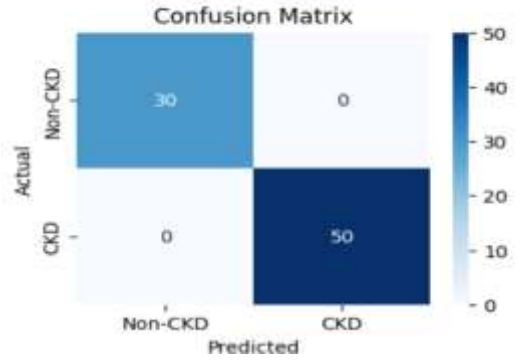

Fig:performance of the model

IV. RESULTS:

The result of the CKD prediction project is a successfully developed and deployed machine learning model that can accurately predict the presence of Chronic Kidney Disease based on patient medical data [1]. The model, built using a Random Forest Classifier, achieved high accuracy and reliable performance metrics such as precision, recall, and F1-score [14]. It was integrated into an interactive Gradio web interface, allowing users to input patient information easily and receive instant predictions with confidence scores [16]. Overall, the project demonstrates that machine learning can effectively support early detection of CKD, aiding healthcare professionals in making faster, data-driven clinical decisions [5].

V. DISCUSSION:

This project is designed to predict Chronic Kidney Disease (CKD) using machine learning and provide a simple web interface for doctors [1]. It starts by loading a medical dataset and cleaning it by fixing missing values and converting incorrect data formats [5]. Then, it uses charts and graphs to understand the data better, such as checking how features like hemoglobin or blood pressure relate to CKD [4]. A Random Forest model is used to learn from the data and predict whether a person has CKD or not [14]. The model is tested using accuracy and other evaluation methods to make sure it works well [6]. After training, the model is added to a Gradio web interface where users can enter patient details using sliders and dropdown menus [15]. The system gives an instant result with a confidence score [16]. The tools used in this project include Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, and Gradio [10]. Although some challenges like missing data and imbalanced classes were faced, the system helps in early detection of CKD and makes it easier for doctors to make fast and informed decisions [7].

VI. CONCLUSION

In this project, we successfully developed a machine learning-based system to predict Chronic KidneyDisease (CKD) using important medical features. By applying proper data preprocessing, visualization, and a Random Forest classification model, the system achieved accurate results. The integration of a Gradio web interface made the tool easy to use for doctors and healthcare workers, allowing them to enter patient data and receive instant predictions along with confidence levels. This system can support early detection of CKD, improve diagnosis speed, and assist in making better clinical decisions. Overall, this project shows how data science and machine learning can be effectively used in the medical field to save time and potentially save lives.

VII. FUTURE SCOPE:

This project can be further improved and expanded in several ways to enhance its clinical value and usability [7]. In the future, more advanced machine learning algorithms like XGBoost or deep learning models can be explored to improve prediction accuracy and performance [2]. Training the system on larger, more diverse datasets collected from multiple hospitals will make it more generalizable and reliable across different patient populations [5]. Integration with hospital management systems or electronic health records (EHR) would enable seamless, real-time use in clinical environments [10]. A mobile application version of the interface could increase accessibility, especially in rural or remote areas where desktop use may be limited [19]. Additionally, the model can be enhanced to predict not only the presence but also the stage of CKD, enabling better treatment planning and disease management [20]. Incorporating explainable AI (XAI) techniques will help clinicians understand the reasoning behind each prediction, thereby increasing transparency and trust in the system [6].

VIII. ACKNOWLEDGEMENT:

Chinthagingala Vasundhara working as an Assistant professor in master of computer application in sanketika vidya parishad engineering college, Visakhapatnam Andhra Pradesh. With 2 years of experience in computer science and engineering (CSE), accredited by NAAC. with her area of interest in java full stack. She is dedicated and emerging academician in the field of Computer Science, currently serving as a faculty member. With a strong foundation in technical concepts and a passion for teaching, She has begun his academic journey by actively mentoring students in practical and innovative projects. As a faculty, She has already made a significant impact by guiding student teams through their academic projects with clarity, enthusiasm, and technical proficiency. His commitment to student success and interest in research-driven teaching make him a promising contributor to academic excellence.

Dalai Murali is pursuing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine learning Dalai Murali has taken up his PG project on Chronic Kidney Disease Prediction Using Machine Learning and published the paper in connection to the project under the guidance of Ch.Vasundhara Assistant Professor SVPEC.

REFERENCES

[1] Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics https://ieeexplore.ieee.org/abstract/document/8477876

[2] A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease https://www.mdpi.com/2075-4418/12/1/116

[3] Early recognition and prevention of chronic kidney disease https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(09)62004-3/abstract

[4] Prediction of Chronic Kidney Disease - A Machine Learning Perspective http://ieeexplore.ieee.org/abstract/document/9333572

[5] Chronic kidney disease prediction based on machine learning algorithms https://www.sciencedirect.com/science/article/pii/S2153353923000032

[6] Machine learning algorithm for early detection of end-stage renal disease https://link.springer.com/article/10.1186/s12882-020-02093-0

[7] Enhancing the Early Detection of Chronic Kidney Disease: A Robust Machine Learning Model

https://www.mdpi.com/2504-2289/7/3/144

[8] Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study

https://www.sciencedirect.com/science/article/abs/pii/S0010482519301258

[9] Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data

https://www.nature.com/articles/s41591-018-0239-8

[10] Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis

https://www.mdpi.com/2075-4426/10/2/21

[11] Comparison and development of machine learning tools in the prediction of chronic kidney disease progression

https://link.springer.com/article/10.1186/s12967-019-1860-0

[12] Performance Analysis of Conventional Machine Learning Algorithms for Identification of Chronic Kidney Disease in Type 1 Diabetes Mellitus Patients

https://www.mdpi.com/2075-4418/11/12/2267

[13] Early detection of feline chronic kidney disease via 3-hydroxykynurenine and machine learning

https://www.nature.com/articles/s41598-025-90019-x

[14] An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy

https://ieeexplore.ieee.org/abstract/document/9040562

[15] Kidney Failure Detection and Predictive Analytics for ckd Using Machine Learning Procedures

https://link.springer.com/article/10.1007/s11831-022-09866-w

[16] A Machine Learning Model for Predicting of Chronic Kidney Disease Based Internet of Things and Cloud Computing in Smart Cities

https://link.springer.com/chapter/10.1007/978-3-030-01560-2_5

[17] Predicting Chronic Kidney Disease Using Hybrid Machine Learning Based on Apache Spark

https://onlinelibrary.wiley.com/doi/full/10.1155/2022/9898831

[18] Machine Learning Framework for Early Detection of Chronic Kidney Disease Stages Using Optimized Estimated Glomerular Filtration Rate

https://ieeexplore.ieee.org/abstract/document/10979939

[19] Machine learning-based warning model for chronic kidney disease in individuals over 40 years old in underprivileged areas, Shanxi Province

https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2022.930541/full

[20] Early Prediction of Chronic Kidney Disease Using Deep Belief Network

https://ieeexplore.ieee.org/abstract/document/9543546