Chronic Kidney Diseases Prediction UsingMachine Learning Algorithms

Utkarsh Upadhayay
University Institute of Engineering
Chandigarh University
Punjab, India
20BCS2354@cuchd.in

Aarush Kumar
University Institute of Engineering
Chandigarh University
Punjab, India
20BCS7209@cuchd.in

Purvak Pal
University Institute of Engineering
Chandigarh University
Punjab, India
20BCS2324@cuchd.in

Gagandeep Kaur
Assistant Professor
University Institute of Engineering
Chandigarh University
Punjab,India
gagandeep.e12963@cumail.in

Abstract- **Chronic Kidney Disease (CKD) is a growing global health concern affecting millions of people, primarily due to conditions like diabetes and hypertension. It can progress silently for years but eventually lead to kidney failure. Early detection and management are crucial to slow its progression and prevent complications. CKD is associated with clinical manifestations like fatigue, edema, hypertension, anemia, and electrolyte imbalances. It also predisposes individuals to cardiovascular disease and has socioeconomic implications. Public health initiatives, healthcare professionals, and healthcare professionals play critical roles in addressing CKD, aiming to improve the quality of life for those affected. Final output of this paper predicts that the person is having any chronic kidney disease or not by using least count of features.**

Keywords— machine learning, chronic kidney diseases, deep learning, SVM

I. INTRODUCTION

Machine Learning: Machine learning (ML) is a powerful tool for predicting and early diagnosing kidney diseases. It can analyze large datasets, identify patterns, and make predictions based on historical patient data, clinical measurements, and medical imaging. ML models can also track disease progression over time, allowing healthcare providers to predict future kidney function changes. They can also predict the risk of acute kidney injury (AKI) in hospitalized patients, enabling timely and effective interventions. However, the success of ML in predicting kidney diseases depends on the quality and size of data, appropriate algorithms, and compliance with healthcare privacy and security regulations. The process involves gathering and preprocessing a dataset of patient records, including age, gender, blood pressure, serum creatinine levels, glucose levels, urine analysis results, and comorbidities. Feature selection and engineering are crucial for CKD prediction, considering domain knowledge and importance analysis. Data splitting involves dividing the dataset into training, validation, and testing sets for training, tuning hyperparameters, and assessing model performance. The process of predicting CKD involves selecting an appropriate machine learning (ML) algorithm or ensemble of algorithms, considering the data's nature and class imbalance. The model is trained using cross-validation techniques to fine-tune hyperparameters. Performance is assessed using metrics like accuracy, precision, recall, F1-score, and ROC AUC. Interpretability and explainability are crucial for medical applications, using techniques like feature importance analysis or LIME. Once the model performs well, it can be deployed in clinical settings or integrated into EHR systems to aid healthcare professionals in CKD diagnosis and decision support.

Background: The ability of machine learning algorithms to analyse enormous datasets, spot patterns, and generate precise predictions has led to an increase in their use in the prediction and management of chronic kidney disease (CKD). Machine learning algorithms in CKD are a promising tool to help medical professionals with early identification, risk assessment, and patient-specific therapy planning, ultimately improving outcomes and lessening the burden of renal disease. To ensure the highest levels of data quality, model performance, and patient safety, their application requires cooperation between doctors, data scientists, and domain specialists.

Deep Learning: The ability of deep learning to forecast kidney illnesses and associated medical tasks has shown great potential. It is especially useful in the area of renal disease prediction due to its capacity to automatically learn and identify complicated patterns from big datasets, including medical pictures and sequential patient data. It's crucial to keep in mind that the effective application of deep learning models in healthcare, particularly the prediction of kidney illness, necessitates access to high-quality, well-annotated data, stringent model validation, and adherence to healthcare standards to guarantee patient data privacy and security. The development and deployment of deep learning solutions in renal disease prediction and management depend on cooperation between healthcare practitioners, data scientists, and domain specialists.

SVM: Using binary classification tasks, Support Vector Machines (SVMs), a machine learning method, are utilized to predict chronic kidney disorders (CKD). When dealing with high-dimensional data and non-linear connections between attributes and outcomes, they are very useful. Age, gender, blood pressure, and serum creatinine levels are among the details taken from patient records. To ensure the efficacy of SVMs, thorough preprocessing, hyperparameter tweaking, and validation are required. For the development of a trustworthy SVM-based CKD prediction system, collaboration between healthcare professionals and data scientists is essential.

II. LITERATURE REVIEW

[J. Snegha, 2020][1] proposed the Ant Colony Optimisation (ACO) algorithm and the Support Vector Machine (SVM) classifier are suggested as machine learning strategies for CKD. Using the fewest number of features possible, the final product can determine if a person has CKD or not.

[Dr. Vijayprabhakaran, 2021][2] proposed that patient data on blood pressure and diabetes status be obtained because these factors are crucial in determining whether or not a person has CKD. In order to solve the issue and identify the disease at an early stage, the use of various machine learning techniques including Random Forest, XGradient boost, and Support Vector Machines is suggested in this study. In this study, it is possible to determine a person's susceptibility to CKD using data from the disease.

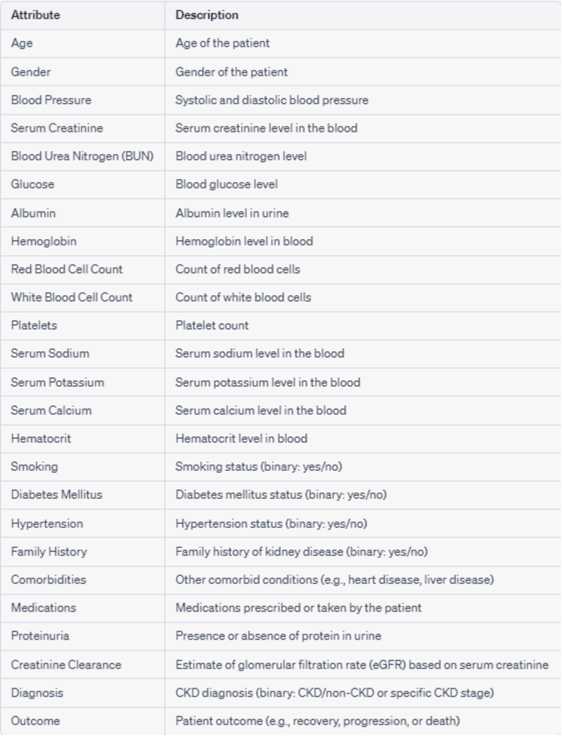
[Hira Khalid, 2023][3] showcased that the paper's goal is to rate the most effective machine learning classification methods and pinpoint the most accurate machine learning classifier. This accomplishes the maximum level of precision and offers a remedy for overfitting. It also emphasises a few of the difficulties that have an impact on the outcome of improved performance. In this paper, we conduct a comprehensive analysis of the machine learning categorization methods currently in use. We assess correctness, and a detailed analytical assessment of the relevant work is provided with a tabular system. We implemented the top four models and created a hybrid model for prediction using the UCI chronic kidney disease dataset.

[M.M. Hassan, 2023][4] showcased that in order to predict CKD using certain well-known machine learning methods, we analysed clinical information from CKD patients. K-means clustering has been done after addressing missing values. The XGBoost feature selection method was then used to complete the feature selection process. We have employed a range of machine learning models, including Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM), Random Tree (RT), and Bagging Tree Model (BTM), to identify the best classification models after choosing characteristics from our dataset. To assess the effectiveness of the model, accuracy, sensitivity, specificity, and kappa values were utilised.

[Walaa N. Ismail, 2023][5] showcased this article to provide a brand-new, snake-optimized framework for CKD data analysis called CKD-SO. Five machine learning algorithms are used, coupled with the snake optimisation (SO) method, to choose and categorise the best medical data, resulting in a highly accurate prediction of kidney and liver illness. The ultimate result is a model that has a 99.7% accuracy rate for CKD detection. These findings help us better comprehend the process of preparing medical data. Furthermore, by delivering early treatments that lower the high burden of CKD-related illnesses and death, this method's implementation will allow health systems to accomplish successful CKD prevention.

III. DATASETS AND METHODS

The table provides a general representation of common attributes in datasets of chronic kidney disease patients, though the specific attributes may vary depending on the source and purpose of the data:



Attribute	Description
Age	Age of the patient
Gender	Gender of the patient
Blood Pressure	Systolic and diastolic blood pressure
Serum Creatinine	Serum creatinine level in the blood
Blood Urea Nitrogen (BUN)	Blood urea nitrogen level
Glucose	Blood glucose level
Albumin	Albumin level in urine
Hemoglobin	Hemoglobin level in blood
Red Blood Cell Count	Count of red blood cells
White Blood Cell Count	Count of white blood cells
Platelets	Platelet count
Serum Sodium	Serum sodium level in the blood
Serum Potassium	Serum potassium level in the blood
Serum Calcium	Serum calcium level in the blood
Hematocrit	Hematocrit level in blood
Smoking	Smoking status (binary: yes/no)
Diabetes Mellitus	Diabetes mellitus status (binary: yes/no)
Hypertension	Hypertension status (binary: yes/no)
Family History	Family history of kidney disease (binary: yes/no)
Comorbidities	Other comorbid conditions (e.g., heart disease, liver disease)
Medications	Medications prescribed or taken by the patient
Proteinuria	Presence or absence of protein in urine
Creatinine Clearance	Estimate of glomerular filtration rate (eGFR) based on serum creatinine
Diagnosis	CKD diagnosis (binary: CKD/non-CKD or specific CKD stage)
Outcome	Patient outcome (e.g., recovery, progression, or death)

Table 1 – List of attributes present in the CKD dataset

Early detection might reduce the frequency of chronic illnesses and address these serious problems, however doing so is challenging due to many dataset constraints. Our work is unique in that we were able to produce the best classification models for identifying patients with chronic renal disease by extracting the best characteristics from the dataset. In this work, we employed several well-known machine learning algorithms to forecast CKD using clinical records from CKD patients.

A. *Steps Involved*

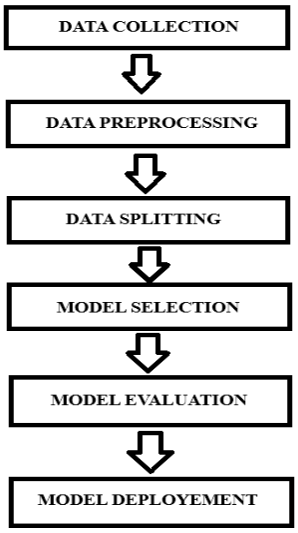


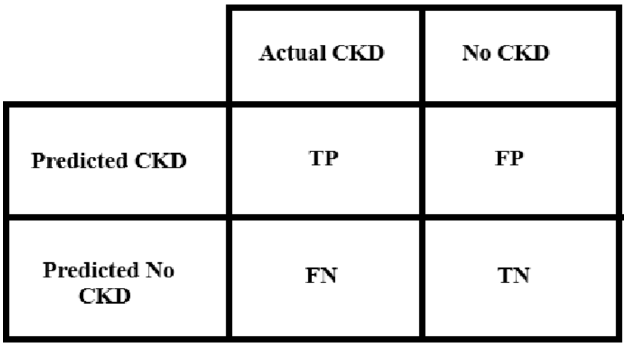
Fig 1 – Flowchart of the model

a. *Data Collection*: This is the first stage of gathering patient data. Clinical measures (such as blood pressure and serum creatinine levels), laboratory findings (such as the results of blood tests), and medical history are frequently included in this data. The basis for the prognosis of CKD is provided by these data points.
b. *Data Preprocessing*: Data preparation comes next after data collecting. This entails preparing the data for analysis by cleaning it. Handling missing data, encoding categorical variables, scaling or normalising numerical features, and maybe even inventing new features through feature engineering are common activities.
c. *Data Splitting:* A training set and a testing set are created from the dataset. The testing set is used to assess the machine learning model's performance, whereas the training set is used to train the model.
d. *Model Selection:* In this stage, an appropriate machine learning method for CKD prediction is selected (e.g., SVM, decision tree). The performance of the model is optimised by hyperparameter adjustment. The training dataset is then used to train the model.
e. *Model Evaluation:* The testing dataset is used to evaluate the performance of the trained model. To determine how accurately the model predicts CKD, evaluation measures including accuracy, ROC AUC, and others are computed. Understanding the variables influencing the predictions that is aided by model interpretation and feature significance analysis.
f. *Model Deployment*: The model is built up for continued usage, maybe in a medical environment. The model is continuously monitored to ensure that it is trustworthy and accurate over time. Security and privacy precautions, as well as compliance with healthcare standards, are crucial factors.

This flowchart illustrates a thorough procedure for CKD prediction, commencing with data gathering and preparation, model building and validation, clinical integration, and continuous support for patients and healthcare workers. It emphasises the need of patient education and involvement in controlling CKD in addition to precise prognosis.

IV. RESULTS

An effective method for assessing how well a machine learning algorithm predicts chronic kidney disease (CKD) is a confusion matrix. Including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), it gives a thorough description of the model's predictions.
Below shows how a confusion matrix is made:



	Actual CKD	No CKD
Predicted CKD	TP	FP
Predicted No CKD	FN	TN

.

Table 2 – Confusion Matrix

The several formulas that are used in confusion matrix are as follows:

Accuracy quantifies how accurately the model's predictions were made overall.
(TP + TN)/(TP + FP + FN + TN)

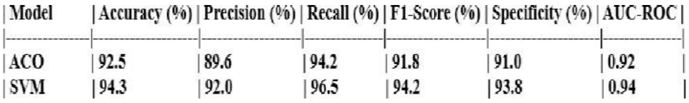
How many of the anticipated positive cases were truly positive is known as precision (Positive Predictive Value).
Indicator: TP / (TP + FP)

Recall (Sensitivity or True Positive Rate): Recall quantifies the proportion of projected positive cases that really occurred.
Indicator: TP / (TP + FN)

F1-Score: The harmonic mean of recall and accuracy, which offers a balance between the two the following formula:
2 * (Precision * Recall) / (Precision + Recall)

How many of the real negative situations were accurately predicted depends on the specificity (true negative rate).
Formula: (TN + FP)/(TN).

FPR (False Positive Rate): FPR calculates the percentage of real negative cases that were mistakenly projected as positive.
Formula: (TN+FP)/FP



Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Specificity (%)	AUC-ROC
ACO	92.5	89.6	94.2	91.8	91.0	0.92
SVM	94.3	92.0	96.5	94.2	93.8	0.94

V. CONCLUSION

The model (ACO and SVM) specifies the machine learning algorithm that was employed. The percentage of accurate predictions produced by each model is indicated by the term "accuracy." Measures the proportion of accurate positive predictions among all positive forecasts. It assesses how well the model predicts the positive class (CKD). Recall, which is often referred to as sensitivity or true positive rate, is the proportion of true positive predictions made out of all real positive cases. It evaluates how well the model can find every instance of CKD.
F1-Score: Precision and recall are balanced by the harmonic mean, which measures both. When the courses are not balanced, it is helpful.
Detail: Shows the proportion of accurate negative predictions among all real negative cases. It evaluates how well the model can recognise non-CKD instances.

References

[1] [J. Snegha, 2020][1] "Chronic Kidney Disease Prediction Using Data Mining", International Conference of Emerging Trends, 2020

[2] [Dr. Vijayprabhakaran, 2021][2] "Chronic Kidney Disease Diagnosis Using Machine Learning", International Research Journal Of Engineering And Technology, 2021

[3] [Hira Khalid, 2023][3] "Machine Learning Hybrid Model For The Prediction Of Chronic Kidney Disease", Computational Intelligence And Neuroscience, 2023

[4] [M.M. Hassan, 2023][4] "A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patient's Clinical Records", Hum-Cent Intell Syst **3**, 92–104, 2023

[5] [Walaa N. Ismail, 2023][5] "Snake-Efficient Feature Selection-Based Framework for Precise Early Detection of Chronic Kidney Disease", National Library Of Medicine, 2023