**Pakistan Journal of Bioinformatics (PJB)**
Volume 01, Issue 02, Year 2026 — ISSN: 2222-7628

# Early Prediction of Chronic Kidney Disease Using Machine Learning Approaches

Muqaddas Shehzadi[1], Rubab Ahmad[2]
Department of Computer Science, University of Agriculture, Faisalabad, Pakistan
[1]edupak2004@gmail.com, [2]rubabahmad6156172@gmail.com

## Abstract

Chronic Kidney Disease (CKD), also known as chronic renal disease. It is defined as a problem in the structure and function of the kidney. It is a progressive disorder and a life-threatening condition that often remains undiagnosed in its stages due to the absence of clear symptoms. Early CKD prediction utilizing clinical information from a Kaggle dataset that is accessible to the public. The purpose of this study is to develop a machine learning-based framework for the early prediction of CKD. The following uses clinical data obtained from a publicly available Kaggle dataset. The methodology involves loading the dataset followed by data preprocessing, including handling missing values and data normalization. Techniques for feature selection are applied to the dataset to identify the most relevant clinical attributes, and principal component analysis is employed in the process of dimensionality reduction to enhance model efficiency and reduce redundancy to improve the process. The processed data are then used to train and evaluate machine learning classifiers. The proposed approach achieves an accuracy of 96 percent, hence proving that this approach gives effective performance in predicting CKD at an early stage. The results indicate that the integration of feature selection and PCA significantly improves model performance. The study indicates the capabilities of machine learning approaches as valid decision-support tools for early CKD prediction, which can assist healthcare professionals in improving patient outcomes are reducing disease progression.

**Keywords:** Chronic Kidney Disease; Machine Learning; Data Preprocessing; Principal Component Analysis; Early Prediction

# 1 Introduction

According to the World Health Organization (WHO), chronic kidney disease is defined as a progressive loss of kidney function lasting for at least three months, characterized either by reduced glomerular filtration (GFR ¡60 ml/min/1.73 m²) or by evidence of renal injury such as proteinuria or structural abnormalities. In 2023, a probable 788 million adults aged 20 years and older represented a marked increase from 378 million people with CKD in 1990. There were also an estimated 44.8 The global age-standardized prevalence rate is approximately 14.2 percent. CKD is a major risk factor for other serious conditions, with impaired kidney function contributing to probable 11.5 percent of cardiovascular disease deaths. Common causes of chronic kidney disease are diabetes, High Blood Pressure (Hypertension), Autoimmune disease, and Obesity. Chronic Kidney Disease (CKD) has serious consequences, primarily due to the kidneys' inability to filter waste, leading to fluid/toxin buildup and affecting nearly every body system, most notably causing high blood pressure, severe cardiovascular diseases. Because chronic kidney disease is in silent stages, early detection is essential for interventions that slow the disease's progression, prevent serious complications like heart disease and kidney failure (dialysis/transplant), lower medical costs, and enable patients to make lifestyle changes that greatly improve long-term health, quality of life, and save lives.

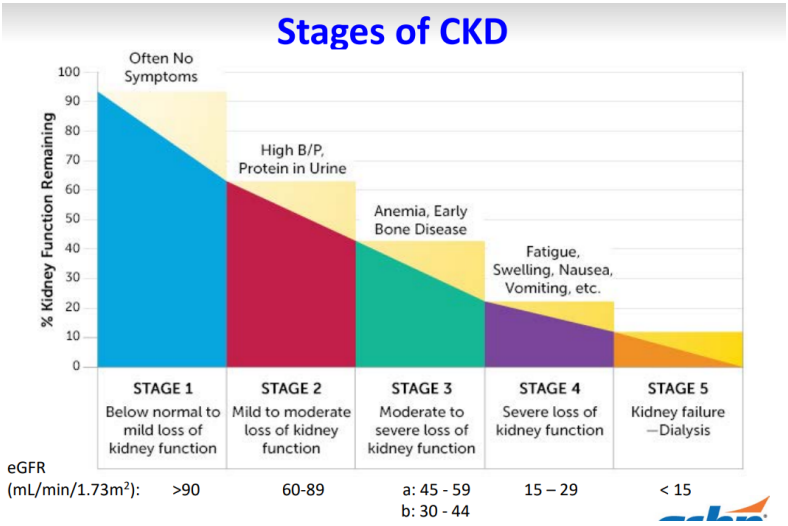


Figure 1: CKD progression

By examining large amounts of patient data (genetics, imaging, EHRs, symptoms) to identify subtle patterns, machine learning (ML) plays a crucial role in disease prediction. This allows for early, accurate diagnosis, individualized treatment, and proactive care for conditions like heart disease, cancer, diabetes, and infectious diseases, ultimately improving patient outcomes.The need for datasets in machine learning (ML) is fundamental because algorithms learn relationships and make predictions from data, rather than being explicitly programmed for a specific task. Datasets are the raw material that gives machine learning its power to learn, adapt, and

make informed, data-driven decisions. Machine models are trained on large datasets containing patient demographics and lab results (such as age, blood pressure, serum creatinine, albumin levels, and presence of diabetes or hypertension) to identify patterns and predict a patient's risk of developing CKD or its progression. (Google, 2025)[7] Google, Google Search, Google, 2025. [Online]. Available: https://www.google.com

# 2 Methodology

The Chronic Kidney Disease (CKD) dataset, which is openly accessible on Kaggle, was utilized in this investigation. Shahzadi (n.d.) Each of the 400 patient records in the dataset has laboratory features. These characteristics include hemoglobin levels, blood pressure, blood sugar, demographic data, and other medical measurements. A binary classification that indicates whether or not CKD is present is the target variable. Due to its thorough coverage of significant clinical indicators, this dataset has been extensively utilized in earlier studies, making it appropriate for machine learning applications.

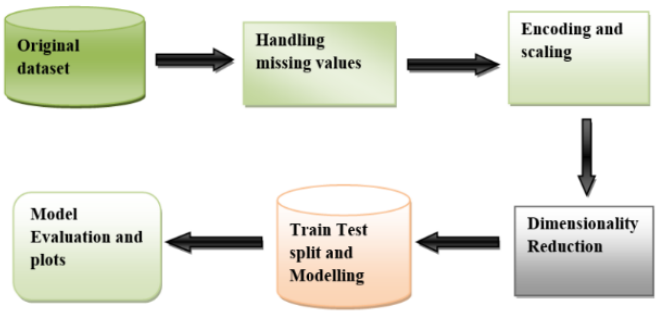


Figure 2: Flowchart of methodology

- **Original dataset:** The University of California, Irvine (UCI) Repository was used to gather CKD data. The dataset can be collected from Kaggle over 3 years ago, given 24 health-related characteristics taken over 3-year periods from 400 patients, predicting the outcome (if a patient has chronic renal disease) of the remaining 242 missing variables in their records using the data of 158 patients with complete records.The data set has 25 features that may predict a patient with chronic kidney disease.
- **Handling Missing Values:** The dataset contains 242 missing values in the whole record. This process in Preprocessing is important to check whether there is a missing value present in the dataset. The dataset can be cleaned by removing missing values using the statistical technique of SimpleImputer, which helps set up a strategy for imputation. So, use the strategy of "most frequent," which replaces the nan values with the mode of the

column. By this, we can confirm there are no null values in the dataset.(SMM Elkholy, n.d.)

- **Encoding and scaling:** Because no machine learning algorithm can accept numerical values as input, the categorical values must be encoded into numerical values. Categories like ”no” and ”yes” are represented by binary digits ”0” and ”1.”

- **Dimensionality Reduction:** After removing the missing values, the dimensionality reduction process should be performed. This reduces the irrelevant features(variables) of data so the information can be easily retrieved and more retained, given better data visualization, and also increases computational analysis. Using PCA (principal component analysis, it uses continuous numerical data, gene expression data, and visualizes in 2D/3D.

- **Label Balance Check:** Performing label imbalance, a binary classification task. Two kinds of labels are used here: 1) checked and 2) not checked. This step will give greater accuracy. It ensures that your model can generalize and doesn't just memorize the training data (overfitting values.

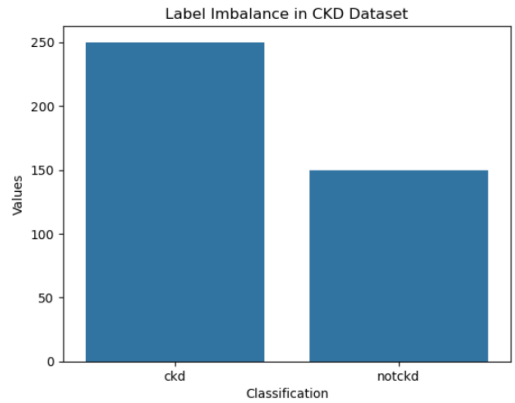


Figure 3: Label Imbalance

- **Correlation Analysis :** To examine the type of correlation between the independent variables, this section begins by creating a pairplot, which displays the relationship between pairs of attributes. Showing multiple graphs as shown in Figure 3, the straight lines in the graphs show categorical columns, while the scatters show numerical values

- **Distribution of data:** The data should be normally distributed; in this case, build a distribution plot of all the numerical columns.

- **Encoding:** The final step of Preprocessing, since the dataset contains both numerical and categorical values, the algorithm requires numerical input, so numerical values should be converted into categorical values.

- **Find Correlation:** Finding correlation in the data is important because, if there is any correlation between independent variables, the values of the weight would be altered and give the wrong output. Plot the correlation between the variables using the Pearson Correlation method. Correlation between dependent and independent variables would not be a problem, but correlation between two independent variables would be a problem. (S Akter, n.d.) .

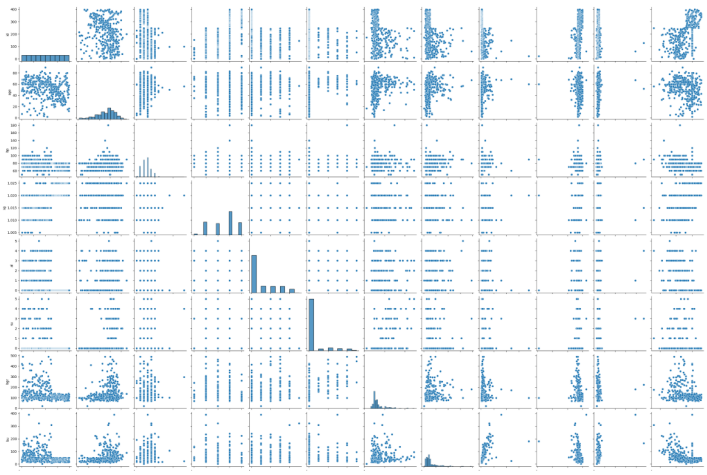

Figure 4: Correlation Plots

- **Scaling:** It is important in algorithms such as the support vector machine (SVM) and K-nearest neighbors, where the distance between the data points is necessary. Performed MinMax Scaler to ensure all the values in the columns lie in the range of -1 and 1.

- **Principal Component analysis (PCA):** The dimensionality reduction of independent variables should be performed in critical phases, especially while dealing with datasets of large columns, without losing the effect of the data, so the expected variance can be preserved.

- **Trainsplit:** The dataset is divided into training and testing sets to evaluate the performance of the model. The training set is used for learning the model parameters, while the testing set is used to assess the generalization capability of the trained model.

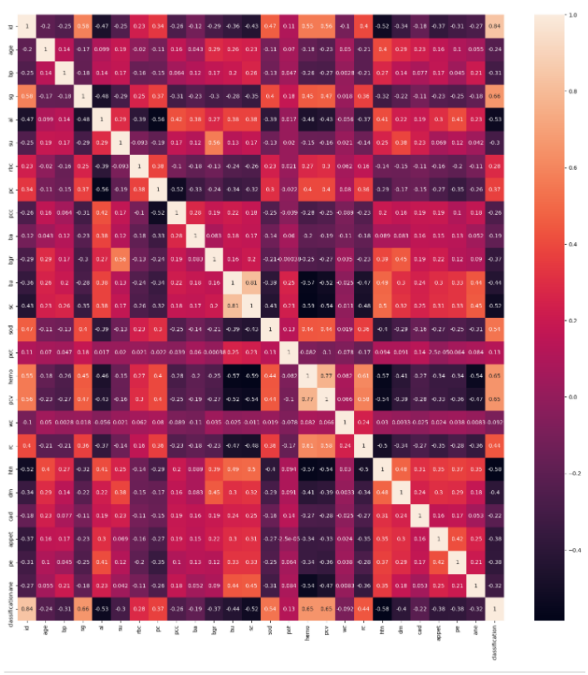

Figure 5: Correlation..

- **Model Building:** Creating a computational model of a system and a mathematical representation (model) of a dataset, which is a Keras model where layers are stacked over each other.

Model: "sequential_3"

| Layer (type) | Output Shape | Param # |
|---|---|---|
| dense_7 (Dense) | (None, 15) | 240 |
| dropout_4 (Dropout) | (None, 15) | 0 |
| dense_8 (Dense) | (None, 15) | 240 |
| dropout_5 (Dropout) | (None, 15) | 0 |
| dense_9 (Dense) | (None, 1) | 16 |

Total params: 496 (1.94 KB)
Trainable params: 496 (1.94 KB)
Non-trainable params: 0 (0.00 B)

Figure 6: Model summary

- **Model Evaluation:** Further evaluate the model by checking its accuracy of the model. That is calculated to measure the percentage of accurately categorized cases out of all the predictions. It provides an overall evaluation of the model's effectiveness and indicates how efficiently the model distinguishes between classes. (F Ma, n.d.)

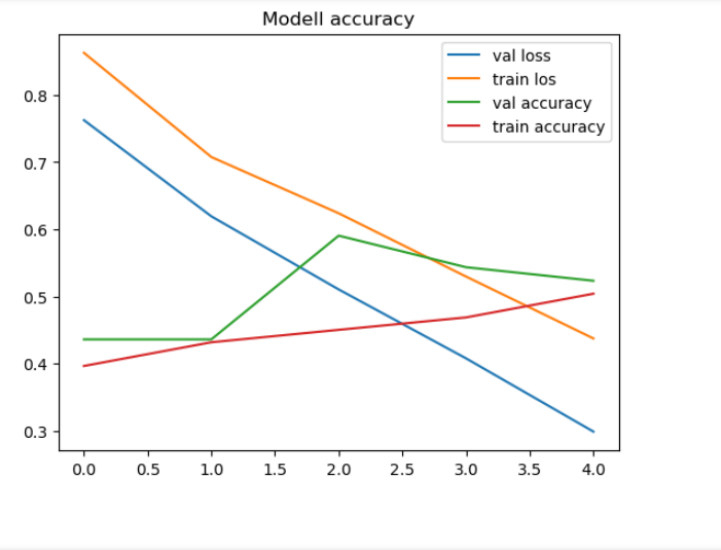


Figure 7: Model evaluation(Shehzadi,2025)

# 3 Results

The chronic kidney disease dataset comprised 400 samples with 24 features. Median imputation was used for numerical variables and mode imputation for categorical characteristics to deal with missing values. Categorical variables were encoded using label encoding to prepare the data for machine learning models.

# 4 Discussion

Principal component analysis was applied to decrease the dataset's complexity while maintaining maximum variance. A total of 10 principal components were selected, which together explained approximately 95 percent of the variance. The PCA-transformed features were used to train a random forest classifier. The model achieves an accuracy of 96 percent. The original features effectively summarized the dataset and captured the key patterns necessary for early CKD prediction. While earlier research has shown encouraging outcomes, the majority of these studies lacked a thorough comparative analysis and concentrated on a small number of machine learning models. Furthermore, problems like handling missing data and model generalizability continue to be difficult. Thus, by comparing several machine learning classifiers using standardized preprocessing and evaluation methods, this study seeks to create a reliable CKD prediction framework. (F. Ma, n.d)

# 5 Conclusion

The study shows that using frequently gathered data, machine learning models can effectively predict chronic kidney disease. To increase model accuracy, preprocessing strategies like handling missing values, encoding categorical features, scaling numerical features, and dimensionality reduction were essential. A more effective and comprehensible machine learning model was made possible by applying PCA to the CKD dataset, which effectively decreased the dimensionality while keeping the majority of the data. Dimensionality reduction can improve the CKD model's accuracy, according to the random forest classifier trained on PCA-transformed features.To further increase prediction accuracy, future research may concentrate on validating the model on huge datasets, investigating different dimensionality reduction techniques, and incorporating more clinical features.

# References


[1] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2019.

[2] D. Mansoor, "Chronic Kidney Disease Dataset," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/mansoordaku/ckdisease

[3] D. Dua and C. Graff, "UCI Machine Learning Repository: Chronic Kidney Disease Data Set," University of California, Irvine, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

[4] J. Smith *et al.*, "Prediction of chronic kidney disease using machine learning algorithms," *Journal of Medical Systems*, vol. 44, no. 6, p. 112, 2020, doi: 10.1007/s10916-020-01610-1.

[5] M. S. Arif, A. Mukheimer, and D. Asif, "Enhancing the early detection of chronic kidney disease: A robust machine learning model," *Big Data and Cognitive Computing*, vol. 7, no. 3, p. 144, 2023.

[6] M. Shehzadi, "CKD prediction using machine learning," Jupyter Notebook, 2025. [Online]. Available: https://www.kaggle.com/username/ckd-prediction

[7] Google, "Google Search," 2025. [Online]. Available: https://www.google.com

[8] N. A. Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Computers in Biology and Medicine*, vol. 109, pp. 101–111, 2019.

[9] S. Akter *et al.*, "Comprehensive performance assessment of deep learning models in early prediction and risk identification of chronic kidney disease," *IEEE Access*, vol. 9, pp. 165184–165206, 2021.

[10] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using a deep learning-based heterogeneous modified artificial neural network," *Future Generation Computer Systems*, vol. 111, pp. 17–26, 2020.

[11] S. M. M. Elkholy, A. Rezk, and A. A. E. F. Saleh, "Early prediction of chronic kidney disease using a deep belief network," *IEEE Access*, vol. 9, pp. 135542–135549, 2021.

[12] R. Sawhney, A. Malik, S. Sharma, and V. Narayan, "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease," *Decision Analytics Journal*, vol. 6, p. 100169, 2023.