Exploratory Data Analysis and Machine Learning Classification of Chronic Kidney Disease: Insights into Causes and Complications

Priyanka Gavade
Computer Science
KLE Technological University,
Belagavi campus
Belagavi, India
priyankagavade.mss@kletech.ac.in

Akash Halli
Computer Science
KLE Technological University,
Belagavi campus
Belagavi, India
02fe22bcs012@kletech.ac.in

Swapnil Laxman Shahapurkar
Computer Science
KLE Technological University,
Belagavi campus
Belagavi, India
02fe22bcs163@kletech.ac.in

Abhay Patil
Computer Science
KLE Technological University,
Belagavi campus
Belagavi, India
02fe22bcs002@kletech.ac.in

2025 10th International Conference on Signal Processing and Communication (ICSC) | 979-8-3315-1189-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICSC64553.2025.10967962

***Abstract*—As a component of the United Nations' 2030 Agenda, Sustainable Development Goal 3 (SDG 3) aims to improve access to healthcare and increase global health by tackling a variety of health issues, including both infectious and non-communicable diseases. Roughly 9.5% of people worldwide suffer from chronic kidney failure(CKD) which is a serious health disease, with prevalence rates differing across regions. CKD progressively reduces kidney efficiency in filtering blood, leading to the accumulation of toxins and fluids in the body. These complications can result in severe health outcomes, such as cardiovascular disease, elevated blood pressure, stroke, and reduced life expectancy. To deepen insights into CKD, we applied exploratory data analysis to explore correlations between key variables and class-specific outcomes. Using five machine learning models, Logistic Regression demonstrated superior performance, attaining 98% accuracy, 98% precision, and 97% recall. Logistic Regression outperformed Linear Regression, K-Nearest Neighbor, Naive Bayes, and Decision Trees, according to 10-fold cross-validation used to validate these models.**

***Index Terms*—Chronic kidney disease, Machine learning,Linear Regression, K Nearest Neighbor, Naïve Bayes, Decision Tree**

I. Introduction

Chronic Kidney Failure is a growing worldwide public health concern, affecting millions of individuals globally. CKD is diagnosed when there is evidence of kidney damage lasting three months or more, regardless of its cause, or when the glomerular filtration rate (GFR) is less than 60 mL/min/1.73 m². Diabetes and hypertension are the leading causes of chronic kidney disease, accounting for over two-thirds of reported cases. Other contributing factors include inherited genetic conditions, glomerular disorders, autoimmune diseases, infections, and abnormalities in the urinary tract. In the United States alone, nearly 37 million adults are living with CKD, with an alarming 90% unaware of their condition.

As people age, the prevalence of CKD rises, with almost half of those diagnosed being over the age of 70. CKD is a significant contributor to mortality, ranking as the ninth leading cause of death in the United States. Apart from diabetes and hypertension, other notable risk factors include heart disease, obesity, a history of acute renal injury, and genetic predispositions. Tobacco consumption has also been linked to an increased risk of CKD development, further underscoring the need to address modifiable risk factors.

Timely identification and effective management of CKD are critical for slowing the progression to End-Stage Renal Disease (ESRD) and reducing associated cardiovascular complications. Unfortunately, CKD often remains asymptomatic in its early stages, delaying diagnosis and treatment. Conventional diagnostic methods, such as laboratory tests and imaging, although effective, are often labor-intensive, time-consuming, and costly. These methods may also lack sensitivity for early detection, particularly in resource-constrained settings, leading to missed opportunities for timely intervention. Furthermore, reliance on human expertise for interpreting results introduces variability and potential biases in diagnosis and treatment planning.

Machine Learning (ML) techniques have emerged as a promising solution to these challenges, offering the ability to analyze complex datasets and uncover patterns that are not immediately apparent to human experts. ML models can integrate clinical, laboratory, and demographic data to make accurate

979-8-3315-1189-0/25/$31.00 ©2025 IEEE

Fig. 1. Proposed Methodology CKD classification

predictions, supporting healthcare providers in devising data-driven strategies for CKD diagnosis and management. Despite these advances, several issues remain in the current work. For example, selecting relevant features, managing imbalanced datasets, and ensuring model interpretability are ongoing challenges. Addressing these issues is crucial to developing robust and reliable systems that can be adopted in real-world clinical settings.

CKD, a debilitating condition that poses significant challenges to healthcare systems, is projected to become the fifth leading cause of death by 2040. This emphasizes the urgent need to adopt innovative techniques, such as machine learning, to enhance early detection and management of CKD. This study aims to bridge the gap by leveraging ML techniques and feature analysis to support clinical decision-making and improve patient outcomes.

The paper is organized as follows: Section 2 provides a brief review of the literature survey on recent works. Section 3 explains about data collection and preprocessing. Section 4 deep dives into EDA. Section 5 explains about different ML models implemented during our research. Finally, the paper concludes in Section 6, summarizing the findings and future directions.

II. Related Work

A comprehensive review of the relevant literature is required to understand the work and the associated challenges in the current advancements in the classification of Chronic Kidney Disease (CKD). This section highlights the methods, models, and techniques previously applied to CKD prediction and diagnosis, which forms the foundation for the proposed work. A detailed comparison table summarizing key contributions is provided at the end of this section.

R. H. Khan et al. [4] propose a hybrid ML model to predict CKD patients with 99% accuracy. The model combines Naïve Bayes, Random Forest, and Decision Tree using a stacking classifier. This hybrid approach reduces overfitting and improves accuracy by leveraging the strengths of individual algorithms. The UCI CKF dataset was utilized, demonstrating the effectiveness of ensemble techniques for robust predictions.

Bilal Khan et al. [5] evaluate seven machine learning models on the UCI repository dataset for CKD classification. Among the tested models, the CHIRP model achieved a remarkable accuracy of 99.75%, outperforming SVM, Logistic Regression, and Naïve Bayes. This study underscores the potential of CHIRP in reducing diagnostic errors and facilitating early detection of CKD.

El Sherbiny et al. [6] developed a model to evaluate the performance of nine machine learning algorithms on a dataset containing 400 records (250 CKD and 150 non-CKD patients) from India. The ADA algorithm achieved the highest accuracy of 99.17%, with superior runtime, F1-score, and precision. This study highlights the importance of algorithm selection for precise CKD prediction.

Anurag et al. [7] utilized lab, clinical, and demographic data to investigate CKD prediction. The Random Forest (RF) model demonstrated the highest AUC-ROC of 0.89, benefiting from dimensionality reduction and preprocessing techniques. This work emphasizes RF's superiority over SVM, KNN, and Logistic Regression in early CKD detection.

Debal et al. [8] used data from St. Paul's Hospital in Ethiopia to predict CKD stages. By employing feature selection techniques like RFECV, the Random Forest model achieved 79% accuracy for multi-class classification and 99.8% for binary classification. This study emphasizes the necessity of customized models for resource-limited settings and recommends exploring deep learning for improved accuracy.

Charleonnan et al. [9] proposed a CKD prediction model using SVM, Decision Tree, KNN, and Logistic Regression. SVM achieved the best accuracy of 98.3%, enhanced by computationally efficient feature selection and data transformation. The study demonstrates SVM's effectiveness in diagnosing CKD and supporting early intervention efforts.

C. P. Kashyap et al. [10] conducted a study using machine learning classifiers, including SVM, Random Forest, Decision Tree, and KNN, on the UCI CKD dataset. Random Forest outperformed other classifiers, achieving an accuracy of 99.33%, followed by SVM at 98.67%. This work illustrates the potential of machine learning for accurate CKD diagnosis.

A. Tope-Oke et al. [11] focused on the role of toxic metals in urine as predictors for CKD. Using KNN, the study achieved

TABLE I
A COMPREHENSIVE LITERATURE SURVEY OF 2024 PUBLICATIONS ON CHRONIC KIDNEY DISEASE PREDICTION

Author	Dataset Description	Accuracy	Model Used	Limitations
Nikhil Verma et al. [1]	UCI CKD dataset with clinical and laboratory data (400 samples, 25 features)	97%	Random Forest, Decision Tree, KNN	Limited focus on scalability and interpretability of models in real-world healthcare applications.
Arpanpreet Kaur et al. [2]	CKD dataset with demographic and clinical data (400 samples, 25 features)	93%	Random Forest, Decision Tree	Random Forest shows high accuracy but lacks scalability and generalization to diverse populations.
Tanjim Mahmud et al. [3]	CKD dataset from UCI (400 samples, 24 features)	97%	Gradient Boosting, KNN, ANN	CNN performance was suboptimal; better suited for image data rather than clinical features.

an accuracy of 98.67%, outperforming Decision Tree and Random Forest. This research highlights the importance of environmental factors in enhancing CKD prediction systems.

L. Anifah et al. [12] classified CKD severity (Grades 1–5) using statistical analysis and Euclidean distance. With a dataset of 199 records, the method achieved a maximum accuracy of 76.88%, emphasizing the significance of laboratory parameters in CKD severity classification.

Verma et al. [1] investigated the application of machine learning models such as Random Forest and Decision Tree for CKD prediction. Using publicly available datasets, the Random Forest model achieved the highest accuracy of 97%, emphasizing the importance of early detection to prevent severe outcomes like dialysis or kidney transplantation.

Mahmud et al. [2] compared Decision Tree and Random Forest models for CKD diagnosis. Random Forest achieved a superior accuracy of 93% compared to 79% for Decision Tree. The findings suggest that advanced machine learning techniques can significantly enhance early diagnosis and management.

Kaur et al. [3] explored Gradient Boosting, KNN, and deep learning models such as ANN and CNN for CKD detection. With a dataset of 400 records, Gradient Boosting achieved the highest accuracy of 97%, showcasing its capability in identifying critical predictors for CKD.

III. DATA COLLECTION AND PREPROCESSING

The dataset on chronic renal disease utilized in this study was obtained from UC Irvine. A range of medical measurements taken from patients with and without chronic kidney disease are present in the data. The outcomes of laboratory tests and other clinical factors are also included in these measurements. The collection contains pertinent information about blood pressure, specific gravity, albumin, blood glucose levels, serum creatinine, age, and other important markers.

In our Exploratory Data Analysis phase, the data cleaning stage involves several crucial steps to make sure the datasets quality and reliability for further analysis and modeling. We began by examining the skewness of the features using histograms and statistical analysis. This step was essential to figure out the distribution of the data and identify any deviations from normality. Understanding skewness helped us recognize the need for potential transformations and guided our approach in the stage of handling missing values and outliers. We then

Symbol	Full Feature Name	Type	Class	Missing Values (%)
id	Id	Numeric	Predictor	0.00
age	Age	Numeric	Predictor	2.25
bp	Blood Pressure	Numeric	Predictor	3.00
sg	Specific Gravity	Nominal	Predictor	11.75
al	Albumin	Nominal	Predictor	11.50
su	Sugar	Nominal	Predictor	12.25
rbc	Red Blood Cells	Nominal	Predictor	38.00
pc	Pus Cell	Nominal	Predictor	16.25
pcc	Pus Cell Clumps	Nominal	Predictor	1.00
ba	Bacteria	Nominal	Predictor	1.00
bgr	Blood Glucose Random	Numeric	Predictor	11.00
bu	Blood Urea	Numeric	Predictor	4.75
sc	Serum Creatine	Numeric	Predictor	4.25
sod	Sodium	Numeric	Predictor	21.75
pot	Potassium	Numeric	Predictor	22.00
hemo	Hemoglobin	Numeric	Predictor	13.00
pcv	Packed Cell Volume	Numeric	Predictor	17.50
wc	White Blood Cell Count	Numeric	Predictor	26.25
rc	Red Blood Cell Count	Numeric	Predictor	32.50
htn	Hypertension	Nominal	Predictor	0.50
dm	Diabetes Mellitus	Nominal	Predictor	0.50
cad	Coronary Artery Disease	Nominal	Predictor	0.50
appet	Appetite	Nominal	Predictor	0.25
pe	Pedal Edema	Nominal	Predictor	0.25
ane	Anemia	Nominal	Predictor	0.25
classification	Class	Nominal	Target	0.00

Fig. 2. Dataset description.

checked for null values across the dataset. Given the skewness observed, we opted for the median to impute missing values in numerical columns.The median works well for a measure of central tendency as it is not affected by extreme values, hence good when there are skewed distributions. For categorical columns, we used the mode while imputing, thus it helped in retaining the most frequent category in each column. This approach ensured that the imputed values were representative of the existing data.

One of the most critical aspects of our data cleaning stage was addressing outliers. Outliers can significantly skew the outcome of any analysis and lead to misleading conclusions. To visualize and identify outliers, we used box plots. Box plots are a compelling visual depiction of the data distribution that highlights the median, quartiles, and any outliers. The box plot's whiskers reach the least and greatest values, respectively, within 3/2 times the first and third quartile IQR. Data points that fall outside of this range are shown separately, as they are regarded as outliers. After identifying outliers, we employed the capping method to treat them. The capping method involves limiting extreme values to a certain percentile threshold, ensuring that the outliers do not distort the analysis while preserving the overall data distribution. Specifically, we capped the values at the 5th and 95th percentiles. This

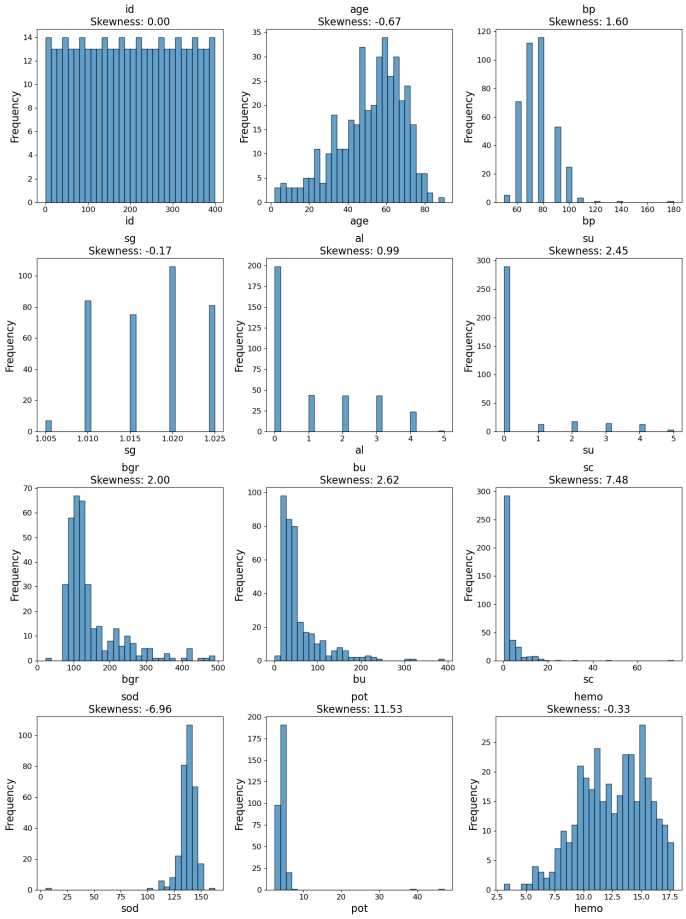

Fig. 3. Skewness of numerical Columns.

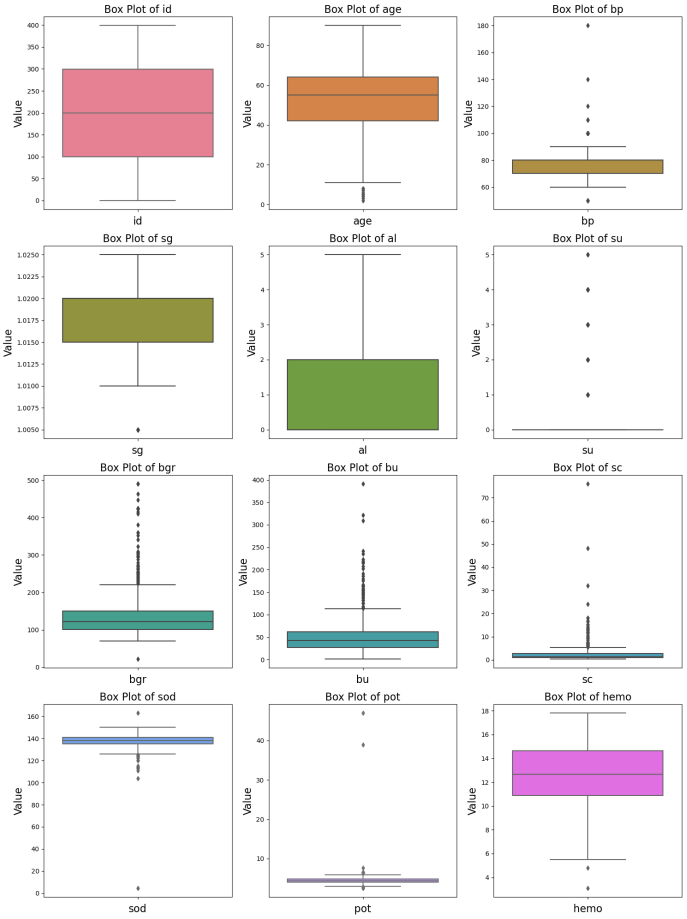

Fig. 4. Outliers detection.

approach retained most of the variability of the data and prevented extreme values from disproportionately influencing the results. Additionally, we eliminated the ID column, which contained unique values for each row. This column did not provide useful statistical information, as it merely served as an identifier. Similarly, we removed the sugar level(sg) column after the cleaning process revealed that all its entries had become zero. This rendered the column non-informative for any further analysis and justified its exclusion.

The summary plot in Fig. 6, shows the impact of each feature on the model's predictions. Feature sugar level(sg) has low absolute SHAP values and minimal dispersion, indicating it has little influence on the model's output across different data instances. Therefore, sg can be dropped as it doesn't contribute significantly to prediction accuracy.

By following these steps, we guaranteed that our dataset was clean, free of inconsistencies, and ready for the subsequent stages of our EDA project. This thorough data cleaning process is essential for the accuracy and reliability of any analytical or machine learning models built on this data. Addressing skewness, handling missing values appropriately, and treating outliers effectively, shown in Fig. 4 are critical to maintaining the integrity of the dataset and deriving meaningful insights from the analysis. Through these efforts, we laid a solid foundation for exploring CKD and contributing valuable findings to the research community.

IV. EXPLORATORY DATA ANALYSIS (EDA)

In this part of the paper, we delve deep into the dataset to gain insights into the relationships between variables and identify patterns.

Our exploratory data analysis (EDA) reveals essential insights into the dataset and its features for chronic kidney disease (CKD) classification. Using ydata_profiling, we generate detailed reports that uncover data distributions, highlight missing values, and show feature correlations. For example, features such as serum creatinine (sc), albumin (al), and sugar (sg) demonstrate strong relationships with the target variable. Meanwhile, red blood cell count (rbc) shows substantial missing data, prompting us to handle these gaps with median imputation.

The correlation matrix and visualizations help us identify redundant and less relevant features. We exclude these during preprocessing to reduce noise and improve the focus on impactful predictors. By combining statistical analysis and domain knowledge, we select features that are both highly relevant and practical for training models. This streamlined

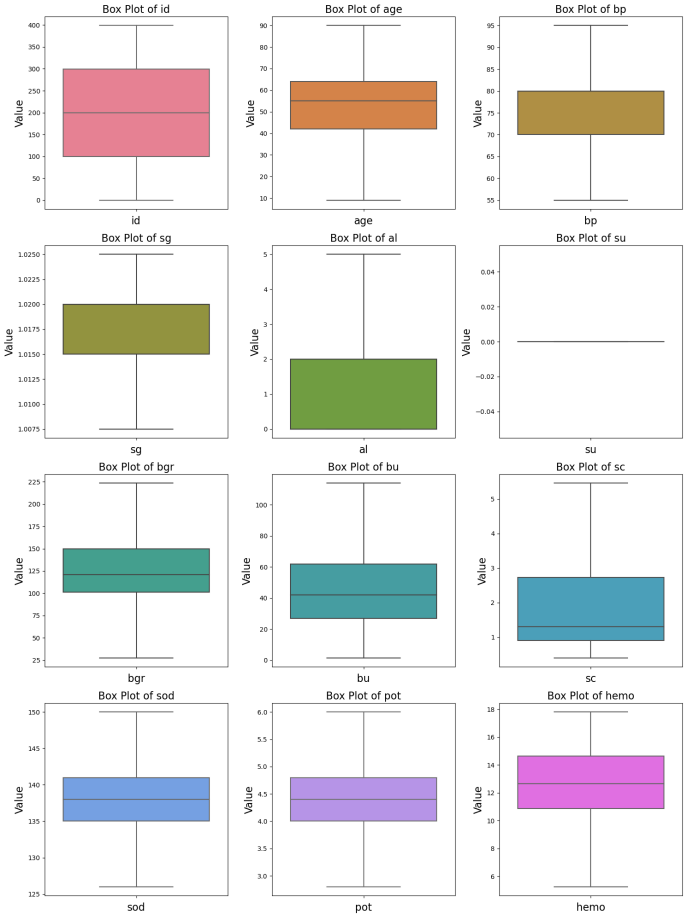

Fig. 5. Boxplot after handling outliers.

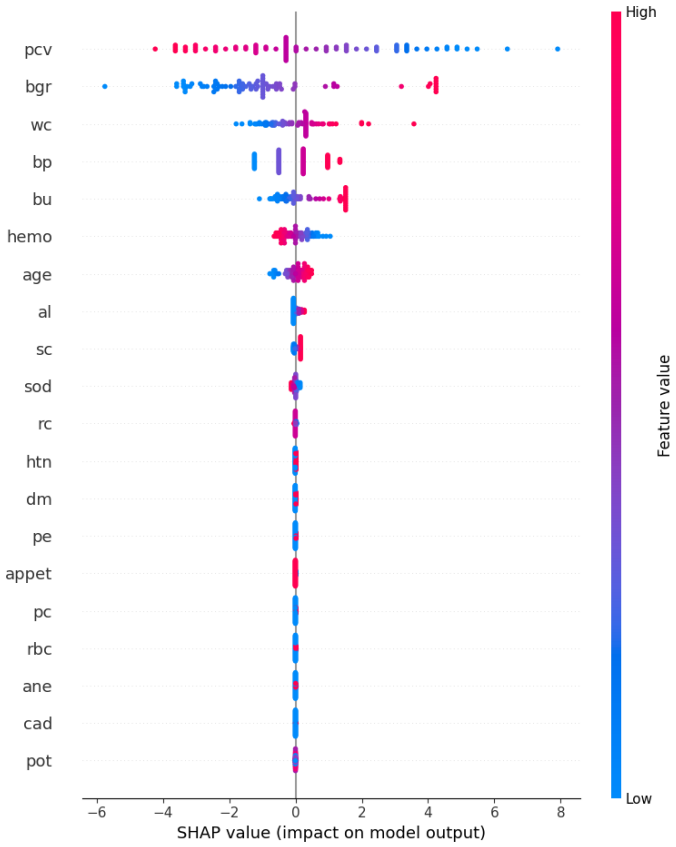

Fig. 6. SHAP Summary plot

approach ensures our models are efficient and capable of capturing the critical factors that influence CKD classification.

The statistical analysis of blood pressure data shown in Fig. 7, revealed that individuals with chronic kidney disease (CKD) exhibit significantly higher blood pressure levels compared to those without the disease. This finding supports the notion that hypertension is a common CKD symptom and highlights the significance of blood pressure management in CKD control.

Building on the knowledge of high blood pressure (BP) in individuals with chronic kidney disease, we investigated the connection between blood pressure and blood glucose levels in Fig. 8. The investigation found a small positive correlation among the two variables, suggesting that blood glucose levels tend to rise as increase in blood pressure. This view points to a compounding risk factor, according to which higher blood pressure may be correlated to higher blood glucose levels in patients with chronic kidney disease.

Later, the higher risk of diabetes mellitus among CKD individuals with elevated blood glucose levels was analyzed, taking into account a confirmed connection between blood pressure and blood glucose levels. According to Fig. 9, a significant portion of patients with CKD who had blood glucose levels more than 150 mg/dL—roughly 20.6 percent

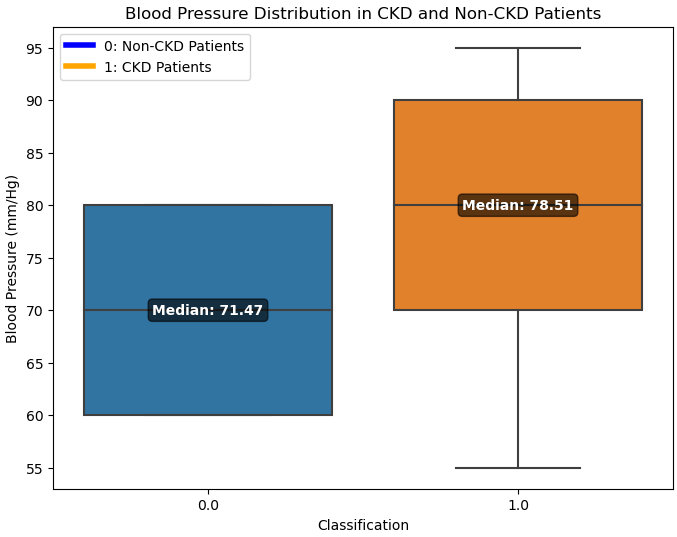

Fig. 7. Blood pressure distribution in ckd and non-ckd patients.

have diabetes mellitus. Fig. 9 highlights how diabetes and CKD are interconnected.

Next, we looked into the differences in serum creatinine levels between individuals with chronic kidney disease who were diabetic and those who weren't. The results shown in

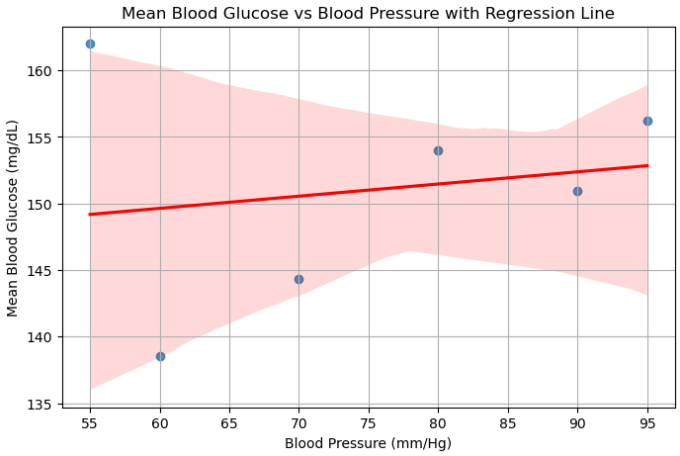

Fig. 8. Mean blood glucose vs blood pressure with regression line.

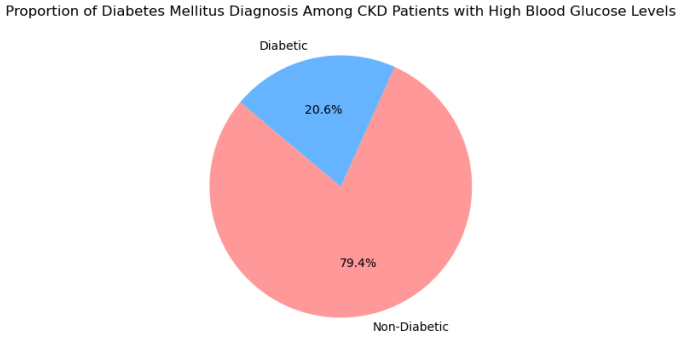

Fig. 9. Proportion of diabetes mellitus diagnosis among ckd patients with high blood glucose levels.

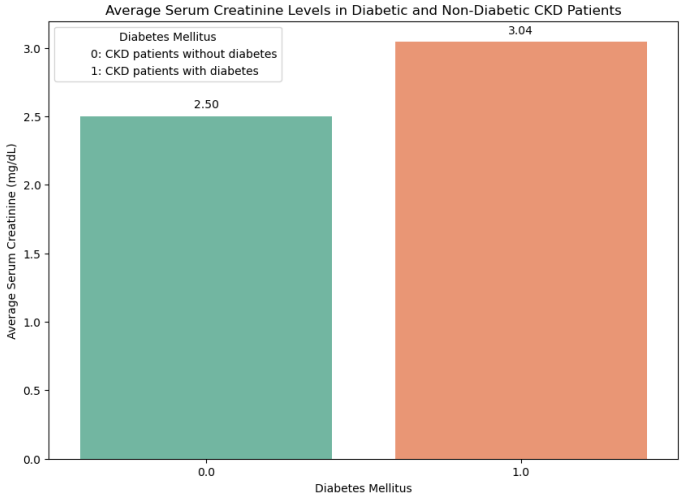

Fig. 10. Average serum creatinine levels in diabetic and non-diabetic ckd patients.

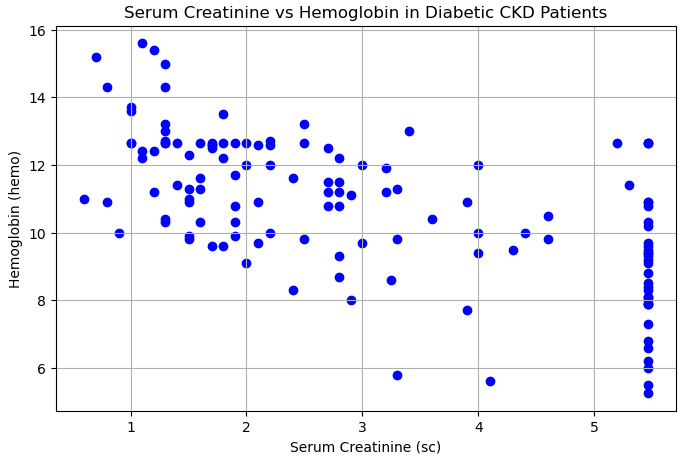

Fig. 11. Serum creatinine vs hemoglobin in diabetic ckd patients

Patients with diabetes who also have chronic kidney disease (CKD) are depicted in Fig. 10 have significantly higher average serum creatinine levels than their non-diabetic individuals. Fig. 10, implies that diabetes causes kidney disease, mostly due to increased levels of creatinine, an essential indicator of kidney function.

Later, we looked into the relationship between hemoglobin levels and serum creatinine in chronic kidney disease individuals with diabetes. The scatter plot shown in Fig. 11, revealed a poor correlation between the two variables, showing that hemoglobin levels tend to decrease as serum creatinine levels increase. This style highlights the link between kidney function and diabetes related anemia in people with CKD.

In Fig. 12, In anemic individuals with kidney disease, we looked at hemoglobin levels and their distribution. The analysis supported the hypothesis that anemia normally corresponds to lower hemoglobin in CKD by showing that a large percentage of anemic CKD patients exhibit low hemoglobin levels. The relationship between white blood cell (WBC) count and packed cell volume (PCV) in individuals with chronic renal disease anemia. The results showed a small positive connection, with the optimal data factor population density centered around a PCV of 25 and a WBC count of 10,000 cells/cumm. This implies that while a relationship exists, it could possibly be effected independently by other factors.

By concentrating on specific gravity levels, we looked at how they related to WBC counts in patients with disease. The violin plot shown in Fig. 14, indicated no significant variations in WBC counts across several specific gravity categories, suggesting a stable WBC distribution irrespective of specific gravity. This suggests that WBC counts in individuals with disease are not as greatly affected by specific gravity stages.

Lastly, we looked at the relation between specific gravity and serum creatinine levels and appetite variation in patients with chronic kidney disease in Fig. 15.

According to the findings of the research, patients who had low specific gravity and normal serum creatinine also reported having lower appetite. For those with elevated serum creatinine, the pattern was less noticeable, but in specific gravity, a trend toward less appetite maintained. This discovery demonstrates the connection between kidney function, specific

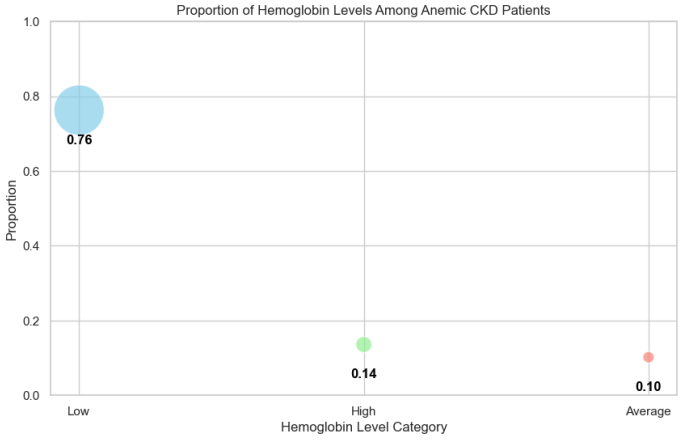

Fig. 12. Proportion of hemoglobin levels among anemic ckd patients.

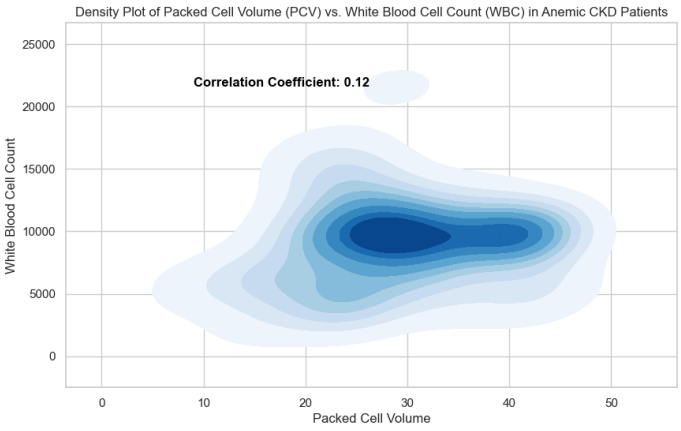

Fig. 13. Density plot of PCV vs WBC in anemic ckd patients.

gravity, and appetite in CKD patients and establishes the strong relationship.

Through EDA revealed that CKD individuals often have high blood pressure and high blood glucose levels, with a significant portion also being diagnosed with diabetes. Diabetic CKD patients have higher serum creatinine levels, which negatively correlate with haemoglobin levels. Anaemic CKD patients have lower haemoglobin levels and positively correlate with packed cell volume. The relationship between PCV and WBC count is weak, suggesting other factors may influence these variables independently. Specific gravity levels do not significantly impact WBC count distribution but are related with appetite variations.

V. Machine Learning Models

In this paper, Machine learning techniques such as Decsion Trees, Naive Bayes, K Nearest Neighbours and Logistic regression is used for classification.

We carefully fine-tune the hyperparameters of our models to maximize their performance and generalization. For Logistic Regression, we adjust the regularization parameter (C) using a grid search with cross-validation to find the right balance

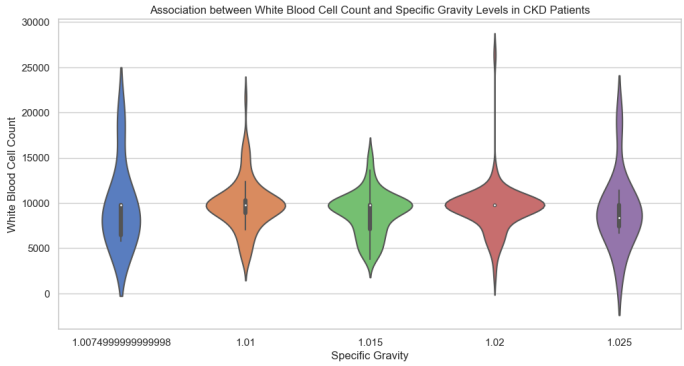

Fig. 14. Association between WBC and specific gravity levels in ckd patients.

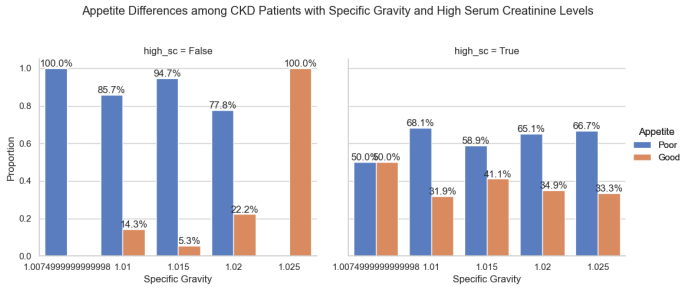

Fig. 15. Appetite difference among ckd patients with specific gravity and high serum creatinine levels.

between simplicity and overfitting. Similarly, for the Random Forest Classifier, we tune key hyperparameters like the number of trees (n_estimators), the maximum depth of each tree (max_depth), and the minimum number of samples needed to split a node (min_samples_split). Our tuning process starts with a randomized search to explore a wide range of parameter combinations. Once we identify the most promising ranges, we conduct a finer grid search to pinpoint the best values. Throughout this process, we evaluate the models using metrics like accuracy, precision, recall, and F1-score. By optimizing these parameters, we significantly improve the models' accuracy and ensure they perform reliably on new, unseen data.

Utilizing their non-parametric nature, decision trees split the dataset based on input feature values, creating a tree structure where branches represent decision rules, nodes represent features, and leaves represent outcomes. Mathematically, the decision tree algorithm recursively splits the training set into subsets based on feature X_i that maximizes the information gain or minimizes the Gini impurity. For instance, the information gain IG for a feature X_i is defined as:

$$IG(T, X_i = H(T) - \sum_{v \in values(x_i)} (\frac{|T_v|}{|T|}) H(T_v) \qquad (1)$$

where, H(T) is the entropy of the set T, and T_v is the subset for a particular value v of X_i.

Naive Bayes classifier, based on the Baye's theorem, assumes feature independence given, the class label. Despite its simplicity, it is effective for certain types of data and

calculates the probability of each class making predictions. Mathematically, the probability of a class C_k given a feature vector x is:

$$P(C_k|x) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(x)} \quad (2)$$

Where, $P(C_k)$is the prior probability of class C_k , and $P(x_i$ —$C_k)$ is the likelihood of feature x_i given class C_k.

K-Nearest Neighbours (KNN) is an instance-based learning algorithm that classifies data points based on their adjacency to K-Nearest Neighbours, with the sample being assigned to the most common class among its neighbours. The distance metric, often Euclidean distance, is used to find the nearest neighbours. For example, the euclidean distance between two points and is:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

The sample is then assigned to the most common class among its - nearest neighbours.

The logistic regression models a binary dependent variable by using the logistic function. It approximates the probability of an input point belonging to a class; the output of the logistic regression is a probability value mapped into discrete classes. The logistic function is defined by:

$$P(y = 1|x) = \frac{1}{1 + e^{-}(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)} \quad (4)$$

where β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the features.

In this study, feature selection was employed to identify and retain the most relevant features, thereby reducing the risk of overfitting, improving model accuracy, and decreasing training time. A heatmap (Fig. 15) was plotted to visualize the correlation among features, facilitating the identification and removal of redundant or irrelevant features. Specifically, the features **Sugar (su)** and **id** were dropped from the dataset based on the heatmap analysis because of their low relevance or redundancy in the context of the classification task. Feature engineering was utilized to transform and create new features, enhancing the model's ability to represent the underlying patterns in the data. Normalization and Standardization processes were applied to adjust the scales of the features, ensuring that they fall within a similar range and improving the performance of distance-based algorithms. The standard score formula used for standardization.

Later, categorical features were undergone through One-hot encoding, which generates binary columns for every category, was used to transform them into numerical representation so that these features could be efficiently used in machine learning models. Anemia (ane), Pedal Edema (pe), Coronary Artery Disease (cad), Hypertension (htn), Diabetes Mellitus (dm), an appetite (appet), a pus cell (pc), pus cell clumps (pcc), bacteria (ba), and red blood cells (rbc), and Class (class) were among the encoded features. This transformation

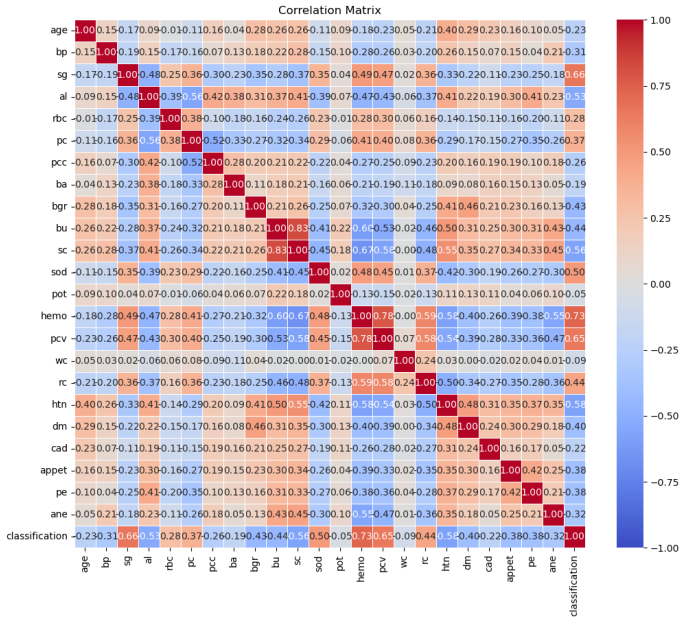

Fig. 16. Correlation analysis using heatmap

facilitated incorporating these characteristics into the study, which enhances the predictive capabilities of the model.

Table.I lists the accuracy of various machine learning classifiers. **Logistic Regression:**

TABLE II
CLASSIFIER ACCURACY

Name of Classifier	Accuracy
Logistic Regression	98.8%
Decision Tree	97%
K-Nearest Neighbour	95%
Naive Bayes	87%

To ensure reliability, a 10-fold cross-validation was used, splitting the dataset into 10 subsets, with each subset serving as a test set once while the others formed the training set, thus reducing variability and risks of overfitting. Additionally, stratified sampling was used to divide the dataset into training (70%) and testing (30%) sets to preserve class distribution. Feature selection and encoding were used to eliminate redundant features and transform categorical variables into numerical representations, and normalization was used to ensure that the scales of features were consistent for models such as KNN. Logistic Regression achieved the highest cross-validation accuracy of 0.97, outperforming Decision Trees, Naive Bayes, and KNN, while Linear Regression performed suboptimally for binary classification.

VI. CONCLUSION

Exploratory Data Analysis (EDA) revealed that CKD individuals often have high blood pressure and high blood glucose levels, highlighting management challenges and a link to diabetes. Also, showed that elevated serum creatinine in

TABLE III
ACCURACY COMPARISON BEFORE AND AFTER K-FOLD CROSS-VALIDATION

Algorithm	Before K-Fold	After K-Fold
Logistic Regression	0.99	0.98
Decision Tree	0.97	0.97
Linear Regression	0.82	0.95
Naive Bayes	0.84	0.97
KNN	0.91	0.80

diabetic CKD patients indicates severe kidney damage, and the negative correlation between serum creatinine and hemoglobin suggests anemia with declining kidney function. Additionally, EDA found that anemia is associated with low hemoglobin levels and significant correlations with packed cell volume, affecting overall health. Five machine learning classifier were trained and the algorithm logistic regression was the best fit. A comprehension of the impact of diseases on the headway of chronic kidney disease will help in the development of comprehensive treatment methods. Quality of life assessments can intervene in a patient's social support networks, physical functioning, and mental health.

VII. FUTURE WORK

Future directions include the implementation of a web application that is integrated with all the predictive models developed in the study for the chronic kidney disease. The online application will include the input mechanisms for clinical and demographic data entered by healthcare practitioners and patients such that real-time machine learning algorithms shall predict the chances of CKD risk and progress. The platforms will be highly user-friendly to guide decision making in early diagnoses and management processes. This tool may be used to enhance CKD management through early intervention and increased access to healthcare.

REFERENCES

[1] N. Verma, T. Sharma, and B. Kaur, "Analysis for predicting chronic kidney disease with the application of machine learning models," in *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, 2024, pp. 1113–1119.

[2] T. Mahmud, M. F. Bin Abdul Aziz, B. Uddin, A. Majumder, T. Akter, N. Sharmen, M. S. Hossain, and K. Andersson, "Utilizing machine learning for early detection of chronic kidney disease," in *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, 2024, pp. 1–6.

[3] A. Kaur, K. S. Gill, S. Malhotra, and S. Devliyal, "Enhancing chronic kidney disease prediction through comparative analysis of decision tree and random forest algorithms," in *2024 3rd International Conference for Advancement in Technology (ICONAT)*, 2024, pp. 1–4.

[4] R. H. Khan, J. Miah, M. A. R. Rahat, A. H. Ahmed, M. A. Shahriyar, and E. R. Lipu, "A comparative analysis of machine learning approaches for chronic kidney disease detection," in *2023 8th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*. IEEE, 2023, pp. 1–6.

[5] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An empirical evaluation of machine learning techniques for chronic kidney disease prophecy," *IEEE Access*, vol. 8, pp. 55 012–55 022, 2020.

[6] M. M. El Sherbiny, E. Abdelhalim, H. E.-D. Mostafa, and M. M. El-Seddik, "Classification of chronic kidney disease based on machine learning techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, pp. e945–e955, 2023.

[7] Anurag, N. Vyas, V. Sharma, and D. Balla, "Chronic kidney disease prediction using robust approach in machine learning," in *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, 2023, pp. 1–5.

[8] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," *Journal of Big Data*, vol. 9, no. 1, p. 109, 2022.

[9] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *2016 Management and Innovation Technology International Conference (MITicon)*, 2016, pp. MIT–80–MIT–83.

[10] C. P. Kashyap, G. S. D. Reddy, and M. Balamurugan, "Prediction of chronic disease in kidneys using machine learning classifiers," in *2022 1st International Conference on Computational Science and Technology (ICCST)*. IEEE, 2022, pp. 562–567.

[11] A. Tope-Oke, B. Badeji-Ajisafe, A. Oguntimilehin, M. V. Inyang, O. Aweh, and O. Abiola, "K-nearest neighbour-based chronic kidney disease prediction system: A case of toxic metals in urine," in *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*. IEEE, 2024, pp. 1–6.

[12] L. Anifah *et al.*, "Chronic kidney disease severity identification using template matching feature selection statistics based," in *2022 International Conference on Electrical Engineering, Computer and Information Technology (ICEECIT)*. IEEE, 2022, pp. 20–24.