Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH
AND DEVELOPMENT

| RESEARCH ARTICLE

"Comparing the Effectiveness of Machine Learning Algorithms in Early Chronic Kidney Disease Detection"

Shuvo Dutta[1] ✉ Rajesh Sikder[2], Md Rasibul Islam[3], Abdullah Al Mukaddim[4], Mohammad Abir Hider[5] and Md Nasiruddin[6]

[1]Master of Arts in Physics, Western Michigan University, USA
[2]PhD student in Information Technology, University of the Cumberlands, KY, USA
[36]Department of Management Science and Quantitative Methods, Gannon University, USA
[45]Master of Science in Business Analytics, Grand Canyon University

Corresponding Author: Shuvo Dutta, **E-mail**: Shuvo.dutta@wmich.edu

| ABSTRACT

CKD is a gradual disease that affects millions of people throughout the United States and results in high morbidity and mortality rates. Chronic Kidney Disease is an ailment that culminates in a gradual loss of kidney function over t,ime. Early detection is essential since timely interventions may prevent the progression of CKD, improve outcomes and survival for patients with CKD, and reduce healthcare costs. In the recent decade, machine learning models have emerged as a game-changing tool in medical diagnostics, leveraging big data and complex algorithms to find patterns almost invisible to clinicians and physicians. This study deployed and evaluated various machine learning approaches for the early detection of CKD, focusing on their comparative performance, strengths, and weaknesses. Machine learning transforms medical diagnosis by leveraging big data and sophisticated algorithms to find patterns that might otherwise elude healthcare professionals. The dataset used for this research will be the CKD dataset, which was contributed to by the Cleveland Clinic in 2021. The dataset can be accessed publicly through the University of California, Irvine's UCI Machine Learning Repository. In this project, the analyst compared and contrasted the performance of Logistic Regression, Decision Trees, and Random Forests. Experimentation results demonstrated that logistic regression had the best performance, yielding a perfect F1 score and accuracy, closely followed by random forest. This result showed that the Logistic model ideally classified all the instances in the test set. Consolidating machine learning algorithms into the early detection of Chronic Kidney Disease (CKD) holds substantial promise for transforming clinical practice. Healthcare professionals can enhance diagnostic accuracy and facilitate timely interventions by leveraging proposed algorithms such as logistic regression.

| KEYWORDS

Chronic Kidney Disease; Early CKD Detection; Medical Diagnosis; Machine Learning algorithms; Logistic Regression; Random Forest; Decision Tree

| ARTICLE INFORMATION

ACCEPTED: 30 September 2024 **PUBLISHED:** 07 October 2024 **DOI:** 10.32996/jcsts.2024.6.4.11

1. Introduction

CKD (Chronic Kidney Disease) is a major worldwide health problem, and if not adequately treated, it may lead to kidney failure, cardiovascular diseases, and other lethal complications. CKD is a progressive condition that affects millions across the USA, causing significant morbidity and mortality. Early detection of CKD will be helpful for its early intervention, which can improve the patient's quality of life, further slowing down the disease process and lowering healthcare costs (Debal & Sitote, 2022). Machine learning techniques in the spectrum of medical diagnosis have brought a new revolutionary method of detecting and predicting diseases.

This invention opens up new vistas in identifying CKD at an early stage. This research project presents several comparative machine-learning techniques for the early detection of CKD and their efficacies, accuracy, strengths, and weaknesses.

1.1 Background and Motivation

Chittora et al. (2021) indicate that Chronic Kidney Disease is an ailment that culminates in a gradual loss of kidney function over time. The kidneys, instrumental organs for filtering excess fluids and waste from the blood, can gradually lose functionality because of various factors, such as hypertension, diabetes, and genetic predisposition. Early detection is essential since timely interventions may prevent the progression of CKD, improve outcomes and survival for patients with CKD, and reduce healthcare costs. The WHO estimates millions of people around the world are affected by CKD. However, it is asymptomatic in its early stage and thus mainly progresses undiagnosed (Debal & Sitote, 2022).

Dritsas & Trigka(2022), in the recent decade, machine learning models have emerged as a game-changing tool in medical diagnostics, leveraging big data and complex algorithms to find patterns almost invisible to clinicians and physicians. Machine learning methods applied to healthcare can increase precision and efficiency in disease detection, including CKD. These techniques use only blood tests, case histories, and demographic information to identify the at-risk individual with an accuracy far superior to traditional methods. Therefore, machine learning integrated with CKD early detection is perceived as one of the most promising avenues for improvement in patient care.

1.2 Problem Statement

Although machine learning algorithms inspire a brighter future, early CKD detection remains challenging. CKD is a multifactorial disease with complex pathophysiology, as detecting an early symptom is hard to determine. Most of the time, the early stages of CKD are asymptomatic; thus, the patients may not realize that they have it until significant kidney damage has taken place (Ebenezer, 2022). Besides, most conventional diagnostic tests are predisposed to variability in clinical measurement and late recognition of the decline in kidney function. Thus, not all traditional diagnostic tools can lead to timely detection.

Islam et al. (2020) posit that machine learning models offer an array of solutions to the healthcare domain. However, they also come with their challenges. These encompass the quantity and quality of available data, the choice of relevant features, and the interpretability of sophisticated models. Besides, different machine learning methods give variable levels of performance, computational cost, and complexity. Furthermore, an algorithm may perform excellently on specific datasets but fall behind on others. Therefore, a comparative study of different machine-learning methods is highly warranted to select the best techniques for early-stage CKD detection.

2. Literature

Jena et al. (2021) contend that CKD is a worldwide health problem-it affects millions of individuals throughout the world as well as in the US, literally bringing a high burden on healthcare systems. It is then imperative to detect CKD early because timely interventions prevent the progression of the disease and diminish the complications, such as the risk of kidney failure. In turn, the difficulty with early detection of CKD lies in the subtleness and asymptomatic features inherent in the disease process. Recently, ML has emerged as a promising tool to enhance diagnostic precision by deciphering complex patient data patterns. This literature review discusses current methods of diagnostics for CKD, the role of machine learning in medical diagnostics, and research on machine learning applications for detecting CKD, ending with an analysis of the gaps in the existing body of knowledge.

2.1 Current Approaches to Diagnosis of CKD

CKD is commonly diagnosed using several traditional methods, most of which are based explicitly on damaged or deteriorated kidney function. The most widely used markers include serum creatinine levels, GFR, urinalysis, and imaging biopsy, all of which carry critical significance in estimating the kidneys' capability of filtering blood.

Serum Creatinine Testing. The creatinine level in the blood is one of the most critical markers indicating kidney function. Creatinine is a waste product produced through muscle metabolism and is usually filtered out by the kidneys. When kidney function starts to deteriorate, serum creatinine levels increase. Although this is a standard test, its sensitivity has substantial limitations (Ghosh, 2024). Serum creatinine depends on factors other than kidney functions, such as muscular mass, age, sex, and diet. This component makes the test unreliable because, in the initial stages of CKD, the changes in creatinine are not significant enough to show changes in the kidney.

Glomerular Filtration Rate. GFR refers to an estimate of the blood volume that the kidneys normally can filter during one minute. It is estimated from serum creatinine, age, sex, and body size. GFR is a reliable measurement compared to creatinine for deciphering kidney function; it also has limitations in the early stages of CKD. GFR can be variable, and minimal losses in kidney function may not be clinically noticeable until gross damage is done (Prasad et al., 2024). The formulas for estimating GFR may also be incorrect in populations such as older adults or those with unusual body compositions.

Urine Tests. In urinalysis, abnormalities may be noted, including albuminuria, an essential marker of renal damage; however, the sensitivity and specificity of most tests to detect unsubtle changes in renal function until later in the course of the disease are variable. Moreover, confounding conditions will affect the presence of urinary tract infections (Prasad et al., 2024).

Imaging and Biopsy. This approach is often complemented by imaging studies, such as ultrasound or CT scans, which help assess kidney structure and rule out obstruction (Sherbiny, 2023). Kidney biopsy is considered the gold standard for diagnosing certain types of kidney disease; however, this is an invasive procedure with risks and is not indicated in all patients. Relying on these traditional methods outlines the need for better diagnostic tools that allow for more reliable and timely detection of CKD.

Although these traditional methods for diagnosing CKD have proved useful, several limitations exist with these conventional diagnostic techniques, especially at earlier stages of the disease. The insensitivity in determining slight changes to kidney function results in most cases of CKD remaining undiagnosed until the disease has progressed deeper into its stages, by which time treatment usually becomes rapidly more complex and less effective (Yashfi et al., 2020). Also, most traditional diagnostic methods must be repeated over time to confirm CKD, which may delay early interventions. These limitations underscore the need for more advanced diagnostic tools, specifically those that can detect subtle alterations in kidney function earlier in the disease process.

2.2 Machine Learning in Medical Diagnostics

Chittora et al. (2021) articulates that Machine learning (ML), a branch of artificial intelligence, has become increasingly pivotal in medical diagnostics. Vital strong points of ML algorithms involve analyses of big data, finding patterns therein, and making predictions using complex relationships in the data. This is all very relevant to health care since medical data is often high dimensional and multifactorial. These capabilities allow ML to deliver much more targeted diagnostic tools by analyzing a wide array of data, including clinical biochemical and genetic information, to identify patterns corresponding to specific diseases. Also, the integration of machine learning into medical diagnosis is set to revolutionize traditional health practices by offering more accurate, efficient, and expansible solutions for disease diagnosis and management.

Debal & Sitote (2022) contends that Machine learning has, in recent years, been related to various fields in the study of medical diagnostics, ranging from tumor detection to the identification of cardiovascular diseases. One of the most common examples is image analysis: The use of ML algorithms in analyzing medical images such as X-rays, CT scans, and MRI studies to depict tumors, fractures, and other abnormalities. Deep learning, a part of machine learning using neural networks, has offered unparalleled accuracy in interpreting medical images and giving diagnostic support. Conversely, ML algorithms have equally been developed to predict disease outcomes based on patient data to personalize treatment options. These models have advanced to predict outcomes for such patients with diagnoses related to heart diseases, diabetes, and cancer levels to help clinicians make more informed decisions about the treatment of such patients.

As per Dritsas & Trigka (2022), the application of machine learning algorithms in kidney disease diagnostics is imperative and strategic. Machine learning algorithms are appropriate for the treatment of CKD since this disease represents a complicated and multifactorial process, where many risk factors influence the disease, such as age, diabetes, hypertension, and genetic predispositions. By parsing large datasets on patient records, laboratory test results, and clinical histories, ML models' curate early warning signs for CKD can be identified, which would otherwise not be easily realized using traditional diagnostic means. Machine learning models are capable of aggregating data from various sources, including wearable devices or home-based monitoring systems, to provide timely updates on the status of a person's kidney function and detect variation over time.

2.3 Previous Studies on Machine Learning for CKD detection

There are a myriad of studies exploring the application of machine learning methods in the early detection of CKD. Many ML algorithms that have been employed in these studies include Decision Trees, SVM, Random Forests, Linear Regression, k-NN, and Neural Networks that classify patients according to their risk for developing CKD.

One of the noteworthy empirical studies was by Ebenezer (2022), where they deployed several machine learning techniques, such as Random Forests and Support Vector Machines, for CKD prediction using 400 patient datasets in the USA. This study reported that the Random Forest algorithm yielded a very good performance with an accuracy rate of 97% in detecting CKD. The authors remarked that Random Forests could handle imbalanced datasets, one common problem in medical data, and also provided feature importance rankings useful to clinicians aiming to understand the critical risk factors behind CKD.

Another study by Islam et al. (2020) focused on the possibility of a Neural Network model in predicting CKD in the USA from laboratory test results and demographic data of patients. The findings revealed that the neural network identified CKD with high accuracy, recording an AUC of 0.96. The authors have determined that deep learning models like Neural Networks can represent complex nonlinear relationships between variables. As such, they are apt for disease diagnosis at an early stage.

In a more recent study comparing the performances of several machine learning models, namely Decision Trees, k-NN, and SVM, by Ghosh (2024), they reported that for early detection of CKD, the highest performances in terms of overall accuracy belong to SVM among the different models tested, its precision, and recall. However, computationally expensive SVM-based models often only perform optimally after extensive parameter tuning. Feature selection was also important since this model was influenced more by values from features like blood pressure and serum creatinine.

2.4 Gap Analysis

Even though future research in the application of machine learning algorithms is promising, a myriad of issues calls for attention in the current research landscape on detecting CKD. Identifying these gaps will guide future studies and ensure the effective integration of machine learning tools into clinical practice.

Limited Generalizability. The generalization of most machine learning models in most current research is a significant concern across diverse populations. Various diverse studies are carried out on specific demographics or clinical settings that might not depict exactly or even closely the broader population in which one aims to generalize M. E. (Sherbiny, 2023). This aspect not only makes the study questionable regarding generalizing into very 'real' situations, where a patient or the characteristics of a particular disease can be highly variable, but it also pertains to the reproducibility of results.

Data Quality and Accessibility. The performance of all machine learning algorithms is highly dependent on the data quality and completeness used to train it. In most cases, the dataset may need to be completed, biased, or standardized. Besides, access to high-quality health data may be a challenge in its own right, especially in resource-limited settings (Sherbiny, 2023). Research should focus on constructing standardized datasets for a wide population base and a broad clinical spectrum that will allow or enable the excellent development of robust machine-learning models with widespread applicability.

Interpretability of Machine Learning Models. The development of machine learning into clinical practice is partly hampered by the fact that most intense learning algorithms are essentially "black boxes." Clinicians cannot inform themselves about how these algorithms come to their predictions (Sherbiny, 2023). That is why clinicians are suspicious and do not believe in models that are not transparent. For this reason, future studies must develop interpretable machine-learning techniques that allow clinicians to understand the underlying rationale for predictions, thus informing their decisions.

Integration into Clinical Workflow. Yashfi (2020) posits that applying any machine learning model to clinical practice needs to be embedded into healthcare workflows. Such integration involves technical challenges and cultural and organizational barriers to the health systems. Future studies need to consider how machine learning can be routinely conducted in clinical practice, which involves preparing health providers to integrate tools and developing algorithms to meet clinical practice guidelines.

3. Methodology

3.1 Dataset Description

This research project uses the CKD dataset prepared in 2021 by Cleveland Clinic. The dataset is publicly available at the University of California, Irvine (UCI) machine Learning repository: This contains medical test results and records of 400 patients; 250 correspond to patients with CKD, while 150 correspond to no-CKD patients (Pro-AI-Rokibul, 2024). Thus, the dataset was imbalanced. There were 19 independent variables: 11 nominal and nine numeric, and a class variable, ckD or notckD. The attributes, along with their description, are presented in Table 1:

Feature Selection

No.	Attribute	Category	Description	Scale
F1	Age	Numerical	Age of the patient	Age in years
F2	su	Nominal	Blood Sugar	0,1,2,3,4,5
F3	bp	Numerical	Blood pressure	Mm/Hg
F4	rbc	Nominal	Red blood cells	Normal/abnormal
F5	ba	Nominal	Bacteria	Present, not present
F6	bgr	Numerical	Blood glucose random	Mgs/dl
F7	pcc	Nominal	Pus cell clumps	Present/not present
F8	pc	Nominal	Pus cell	Normal/abnormal

F9	Sc	Numerical	Serum Creatinine	Msgs/dl
F10	bu	Numerical	Blood urea	Msgs/dl
F11	hemo	Numerical	Hemoglobin	gms
F12	pot	Numerical	Potassium	mEq/L
F13	appet	Nominal	Appetite	Good,bad
F14	dm	Nominal	Diabetes mellitus	Yes, No
F15	ht	Nominal	Hypertension	Yes, no
F16	wbc	Numerical	White blood cells	Cells/cmm
F17	ane	Nominal	Anemia	Yes, no
F18	cad	Nominal	Coronary artery disease	Yes, no
F19	class	Nominal	Class	Ckd, notckd
F10	pvc	Numeric	Packed cell volume	-

By referring to the features in Tabe 1, *Specific gravity* refers to the ratio between the densities of urine and water. It enables the measurement of the concentration of particles in the urine. It can indicate one's hydration level or detect kidney malfunction. Albumin *i*s a protein in your blood. When your kidneys are damaged, they let albumin leak into your urine. High levels of albumin in the urine may indicate you have CKD. *Blood urea* gives particular views about the kidneys because a blood urea nitrogen test determines the quantity of urea nitrogen in the blood. High values of blood urea mean impaired kidney functions. The sugar measurement is carried out in the blood using a random blood glucose test. A reading of above 200 mg/dL indicates diabetes. *Serum creatinine*, which denotes a waste product resulting from the metabolic process of muscles, is also measured. High creatinine levels in the blood or urine also hint at poor kidney function while filtering out waste.

Sodium is an electrolyte and acts in significant ways in the functioning or working of muscles and nerves. The blood test related to sodium measures its level or concentration in the blood. It can be abnormally high when it leads to kidney problems, dehydration, and a few other conditions. *Potassium* is another kind of essential electrolyte; abnormal levels may mean that there is faultiness in one's health. *White blood cells* are known to be part of the immune system; they protect the body from infections. The average WBC count ranges typically between 4,000 to 11,000 cells per microliter of blood. In connection, high levels of WBCs are one of the common signs of CKD progression. *Red Blood Cells* deliver oxygen to the tissues and have an average count range of 4.7 to 6.1 million cells per microliter in men and 4.2 to 5.4 million cells per microliter in women. The most common complication with CKD is *anemia* or low counts of RBCs.

3.2 Pre-Processing

Meanwhile, the data had to be preprocessed to make it suitable for machine learning. Therefore, all the nominal or categorical data were codified. Specifically, the attributes whose scales are 'normal' and 'abnormal' were transformed to 1 and 0, respectively. The 'present' and 'not present' scales of the attributes have been converted to 1 and 0, respectively. Further, the 'yes' and 'no' scales were coded 1 and 0, respectively. And finally, the attribute with 'good' and 'poor' scales was transformed to 1 and 0, respectively (Pro-AI-Rokibul, 2024).

To deal with the missing values in the dataset before building ML models, imputation was applied to estimate and replace the missing values in a dataset. Since the number of missing values in our dataset was insignificant, the mean imputation technique was used to handle the missing values. In this case, the missing value treatment is decided by the mean imputation technique, which calculates the average of the available values w.r.t every variable, and the unavailable values are filled with the estimated mean value accordingly (Pro-AI-Rokibul, 2024). Meanwhile, leaving the 'age' and binary attributes aside, all the remaining attributes were scaled between 0 and 1 using the Min-Max Scaling technique.

3.3 Machine Learning Techniques Evaluated

The algorithms evaluated in detecting CKD are briefly described in this section. This process was done by deploying various classifiers and gauging how well they make CKD predictions. Classification algorithms such as Logistic Regression, Decision Trees, and Random Forests will be evaluated. The concept of the study was to investigate the efficiency of ML in predicting chronic renal disease. Linear regression Decision Trees & Random Forests, three ML strategies, were utilized in this research. The best overall performance among the three algorithms for each categorization was chosen as the best machine learning model.

1) Logistic Regression

Prasad (2024) contends that Logistic Regression is a type of supervised machine learning that is mainly adopted for binary classification-only problems; that is, predicting the probability of an event happening or not happening from one or more independent variables. Logit Regression: Despite the name, logistic regression models the probability of the dependent variable based on the independent variables using the logistic function, whose outputs are between 0 and 1. Since these fall between 0 and 1, it is well-suited for situations such as disease diagnosis based on patient tests, etc. The algorithm follows the principle of maximum likelihood estimation to come up with parameters of the logistic function, hence enabling the classification by applying a threshold to the predicted probabilities for classifying observations into one of two categories.

2) Decision Trees

Sherbiny (2023) articulates that a decision tree is a flexible, interpretable supervised learning model used for classification and regression tasks in machine learning. This model works by recursively partitioning the dataset into subsets based on the value of input features, creating a tree-like model of decisions. Each internal node of the tree represents some sort of feature test, with the branches corresponding to the outcomes of the test and leaf nodes representing the final decisions or predictions. This is why decision trees can visually and explicitly represent decision processes in a straightforward way that makes them understandable and interpretable (Sherbiny, 2023). First, decision trees are beneficial since they can operate with numerical and categorical variables, and their processing is required at a minimum. Besides, they can provide transparency regarding the importance of features in seminal prediction. However, decision trees can overfit, especially when dealing with complex datasets, which is why various techniques take place, such as pruning and ensemble methods like random forests, to improve their performance.

3) Random Forest

Yashfi et al. (2020) contend that the Random Forest algorithm is considered a powerhouse of Ensemble Learning. Random Forest works effectively for both classification and regression instances in Machine Learning. The training phase works by constructing many decision trees; every split is done for a random subset of the training data, and every split selects the features randomly. The randomness thereby helps in producing a diverse bunch of trees that are less correlated with each other. This capability may improve the robustness and accuracy of the overall model. When making predictions, the random forest sums up the outputs of all constituent trees by taking a majority vote in the case of classification tasks- or an average prediction in the case of regression tasks. This aspect boosts predictive performance and reduces overfitting variance, making random forests one of the most popular and successful algorithms in health, finance, and other areas. Easy to use, with the capability of handling big datasets with high dimensionality, it makes random forests more popular among data scientists and machine learning practitioners.

4. Evaluation Metrics

In this study, precision, recall, F1 score, averages, and a correlation matrix were employed to correctly approximate the performance of models in classification problems. The confusion matrix summarises prediction results by class concerning the number of correct and incorrect actual predictions. Primarily, a confusion matrix is the most basic summary of results with these notations: **TP, TN, FP, FN**.

- **TP**-true positive means the observation is positive and correctly predicted as positive.
- **TN** stands for true negative, meaning a negative case is predicted correctly as negative.
- **FP** denotes a False Positive - Observation is falsely predicted as positive while it's negative.
- **FN** stands for False Negative: A truly positive observation inaccurately predicted to be negative.

These figures of a confusion matrix can ideally determine the model's accuracy.

4.1 Classification Accuracy

Classification accuracy was calculated by employing the expression displayed below (i):

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

4.2 Experimental Setup

This experiment setup involved importing the dataset into Python with the machine learning library sci-kit-learn. All experiments, including preprocessing, further data exploration, model training, and evaluation, were done using Python scripts. First was the opening and loading of this dataset for familiarization with the content structure and various attributes included in the data. The analyst obtained an overview of the number of samples, features, and a few characteristics of interest before more profound processing/model development. Understanding the raw dataset adds to the analyst's knowledge before commencing more advanced data exploration, analysis, and model training.

4.3 Cross Validation Techniques

In the machine learning performance estimation, we used k-fold cross-validation. This cross-validation divides the dataset into 'k' subsets or folds. In each run, one-fold serves as the validation set, and the remaining folders serve to train the model. The process is repeated 'k' times, with every fold once serving as a validation set. This technique will prevent overfitting and help estimate the model's performance more accurately by averaging the results across all folds.

5. Implementation

5.1 Data Pre-processing

Handling Missing Values

```
# Step 1: Handling missing values
# For numerical columns, we will use mean imputation
num_cols = ['age', 'bp', 'sg', 'al', 'su', 'bgr', 'bu', 'sc', 'sod', 'pot', 'hemo']
```

```
# For categorical columns, we will use the most frequent value for imputation
cat_cols = ['rbc', 'pc', 'pcc', 'ba', 'pcv', 'wc', 'rc', 'htn', 'dm', 'cad', 'appet', 'pe',
'ane', 'classification']
```

```
# Imputing missing numerical values with mean
num_imputer = SimpleImputer(strategy='mean')
df[num_cols] = num_imputer.fit_transform(df[num_cols])
```

```
# Imputing missing categorical values with the most frequent value
cat_imputer = SimpleImputer(strategy='most_frequent')
df[cat_cols] = cat_imputer.fit_transform(df[cat_cols])
```

Encoding categorical variables

```
# Label encoding for binary categorical features
label_enc_cols = ['rbc', 'pc', 'pcc', 'ba', 'htn', 'dm', 'cad', 'appet', 'pe', 'ane',
'classification']
```

```
# Apply LabelEncoder to each binary categorical column
le = LabelEncoder()
for col in label_enc_cols:
    df[col] = le.fit_transform(df[col])
```

Converting object types to numerical (for 'pcv', 'wc', 'rc')

```
# Convert 'pcv', 'wc', 'rc' to numeric (since they are stored as objects but represent
numbers)
df['pcv'] = pd.to_numeric(df['pcv'], errors='coerce')
df['wc'] = pd.to_numeric(df['wc'], errors='coerce')
df['rc'] = pd.to_numeric(df['rc'], errors='coerce')
```

```
# Impute any missing values in these columns after conversion
df[['pcv', 'wc', 'rc']] = num_imputer.fit_transform(df[['pcv', 'wc', 'rc']])
```

Step 1- Handling Missing Values Techniques: To manage the missing values, we imputed the missing values in our data. On numerical columns- 'age', 'bp', 'sg', 'al', 'us', 'bgr', 'bu', 'sc', 'sod', 'pot', 'hemo' missing value imputation was done with the mean. This technique replaces the missing values with the average of the column, keeping the data distribution intact as far as possible. For imputation of categorical columns, including 'rbc', 'pc', 'PCC', 'ba', 'pcv', 'wc', 'rc', 'htn', 'dm', 'cad', 'appet', 'pe', 'ane', 'classification', we used the most frequent value, i.e mode. This is a decent strategy for categorical variables because this way we aren't introducing new categories.

Step 2: Cleaning the Data—This involved identifying and correcting errors in the dataset. Outliers were detected using the Z-score method, which estimates erroneous data points. The outliers were either removed or transformed according to how their presence would affect the dataset.

Step 3: Feature Encoding - We then did label encoding on the binary categorical features: 'rbc', 'pc', 'PCC', 'ba', 'then', 'dm', 'cad', 'applet', 'pe', 'and', 'classification'. Label encoding assigns a unique integer for each category, appropriate for binary features or when there is a natural ordering. In this way, algorithms that require numerical encoding as input can be implemented.

Step 4: Normalization and Scaling—Most features, like serum creatinine levels and GFR itself, are usually on different scales. Standardization or Min-Max scaling can make all of them equally contribute to model predictions.

Step 5: Data Splitting: This was the final step, where the dataset is divided into three sub-datasets: 'training', 'validation', and 'test' datasets. A good rule of thumb was to split the data into 70% for training, 15% for validation, and 15% for testing to ensure it generalizes well.

5.2 Models Training

The first step involved importing the necessary libraries, including the kidney disease dataset, to conduct exploratory data analysis (EDA) to understand the features and target variable as showcased below:

```
import numpy as np
import seaborn as sns
import pandas as pd
from matplotlib import pyplot as plt

import warnings
# Ignore all warnings
warnings.filterwarnings('ignore')
```

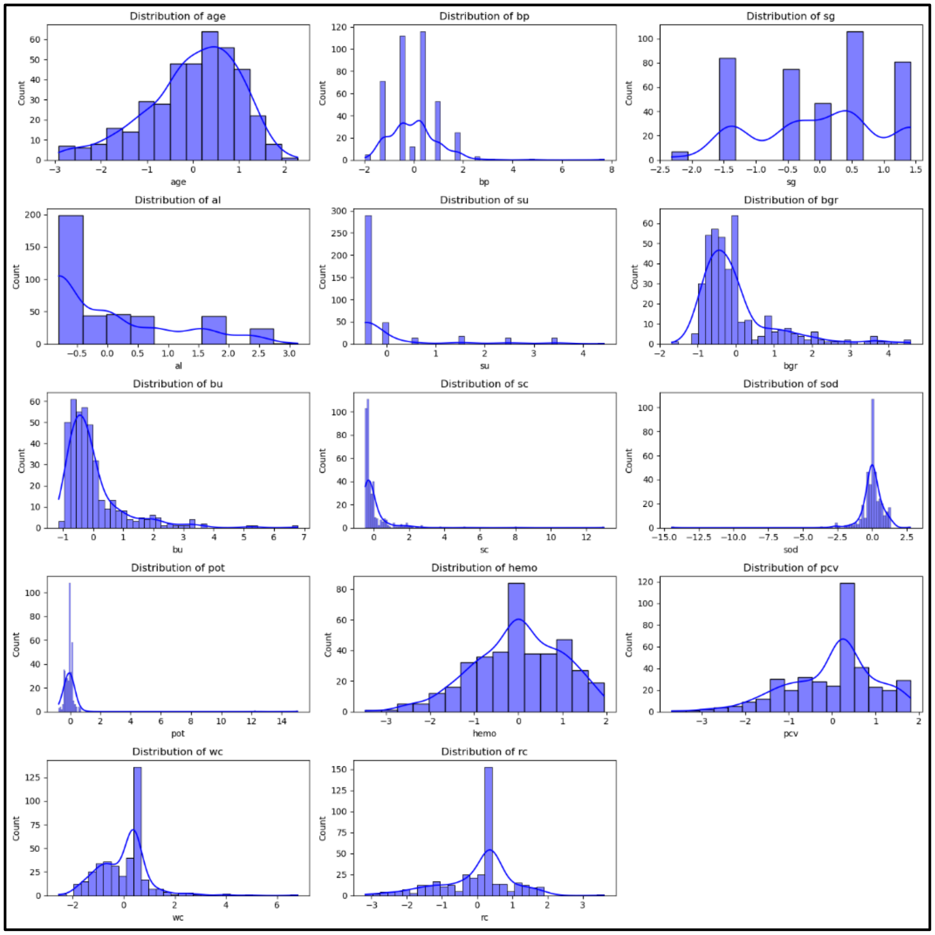
```
df = pd.read_csv("kidney_disease.csv")
```

```
df
```

The following code fragment was used in Exploratory Data Analysis, exploring the distribution of numerical features. This prepared a series of histograms for each numeric feature in the array 'num_cols'. It used matplotlib and seaborn libraries to create a 5x3 grid of subplots where each subplot would depict the distribution of one numerical feature. These histograms are created using the seaborn histplot function with the kernel density estimation enabled for a smoother distribution representation. The blue color was chosen for consistency purposes. Each subplot is titled with the name of the corresponding feature. Finally, adjust the layout for better visualization with plt.tight_layout() and display the whole figure using plt. show().

Expolatory Data Analysis (EDA)

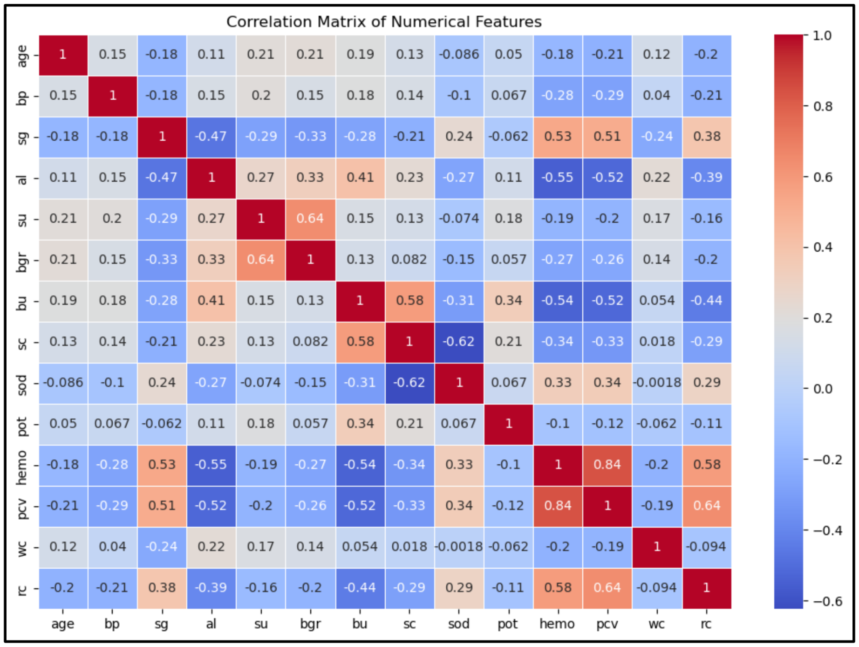
```
# Step 3: Distribution of Numerical Features
num_cols = ['age', 'bp', 'sg', 'al', 'su', 'bgr', 'bu', 'sc', 'sod', 'pot', 'hemo',
'pcv', 'wc', 'rc']

plt.figure(figsize=(15, 15))
for i, col in enumerate(num_cols, 1):
    plt.subplot(5, 3, i)
    sns.histplot(df[col], kde=True, color='blue')
    plt.title(f'Distribution of {col}')
plt.tight_layout()
plt.show()
```



The next code snippet focused on creating a correlation matrix visualization. The code started with setting up a figure in matplotlib, which was 12x8 inches in size. Then, the analyst computed and visualized the correlation matrix for numeric features in 'num_cols' by calling the corr() function on them. The computed correlation matrix was then visualized as a heatmap using Seaborn's heatmap function. This heatmap was configured to enable annotation, showing the correlation values in each cell, to use a color map 'coolwarm' to depict both positive and negative correlations respectively, and to set the line width between cells to 0.5. It gave the title "Correlation Matrix of Numerical Features" to the plot using a plot. title().This process was an important visualization in further Exploratory Data Analysis because it allowed the data analysts to identify and comprehend the relations between different numerical features in the dataset in a matter of seconds. These provide substantial information on feature selection, the assessment of multicollinearity, and overall data understanding for further modeling steps.

```
# Step 4: Correlation Matrix
plt.figure(figsize=(12, 8))
corr_matrix = df[num_cols].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix of Numerical Features')
plt.show()
```

Output:



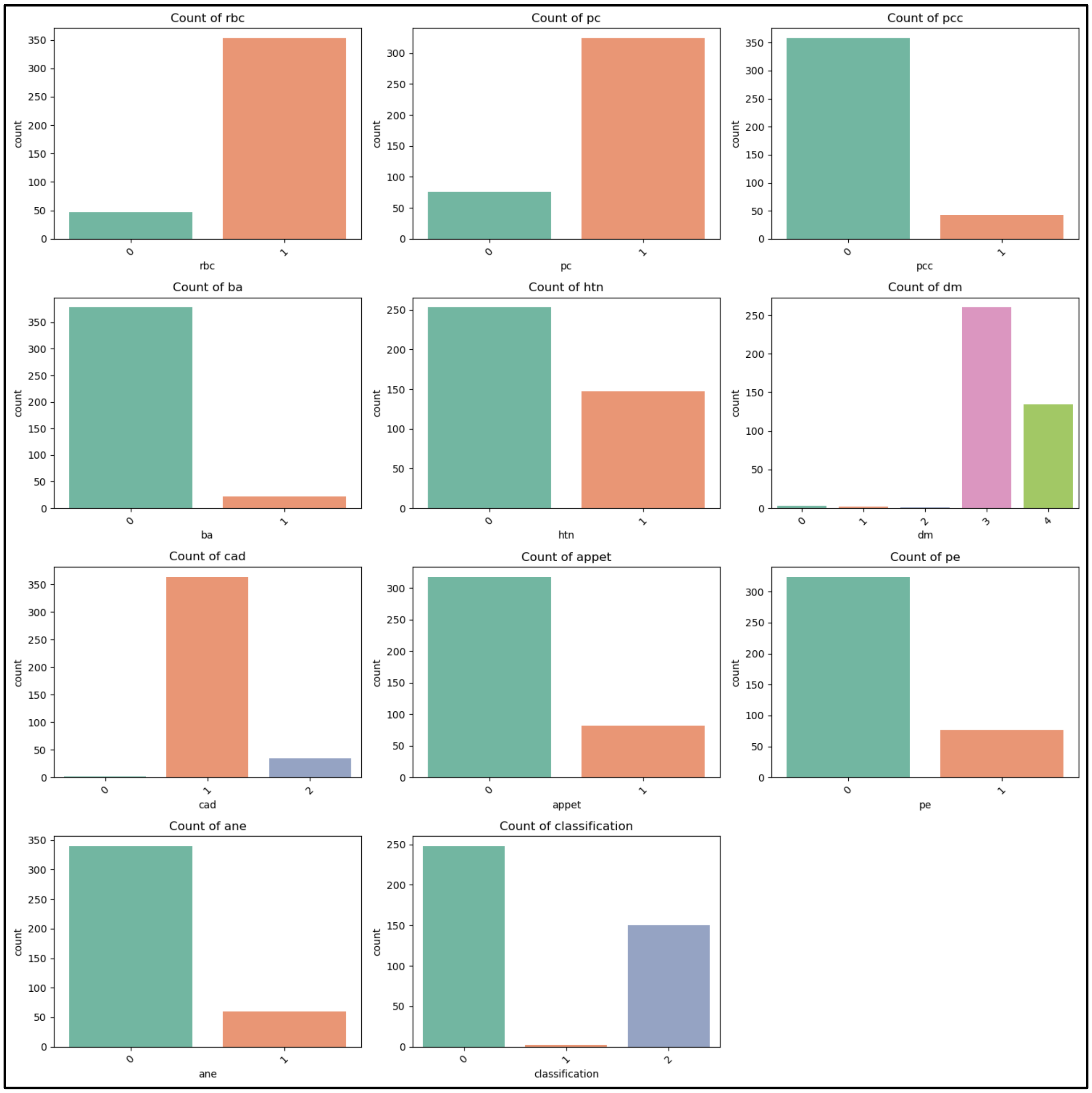
The strongest positive correlation of 0.84, was between 'hemo' and 'pcv', likely hemoglobin and packed cell volume respectively. This outcome was attributed to the fact that both were under the umbrella of red blood cell measures. On the other hand, 'su' ('sugar') and 'bgr' ('blood glucose') were highly correlated at 0.64, which was expected since both deal with blood sugar levels. By contrast, the 'bu' and 'sc' - probably blood urea and serum creatinine were middling positively correlated at 0.58, which makes sense because they were both measures of kidney function.

Conversely, 'al' denoting albumin, had remarkable negative correlations with features such as 'sg' (-0.47), 'hemo' (-0.55), 'pcv' (-0.52). That outcome implied that the higher the albumin level, the lower these values are, which could point toward specific conditions of the kidneys. In contrast, 'Age' and 'bp' standing for blood pressure; both have relatively low correlations in most of the rest of the features, which may be independent factors. 'pot' denoting potassium, had a weak correlation with most of the features; the only possible exception is with 'sod', where it had a weak negative correlation of -0.62, which may have something to do with electrolyte balance.

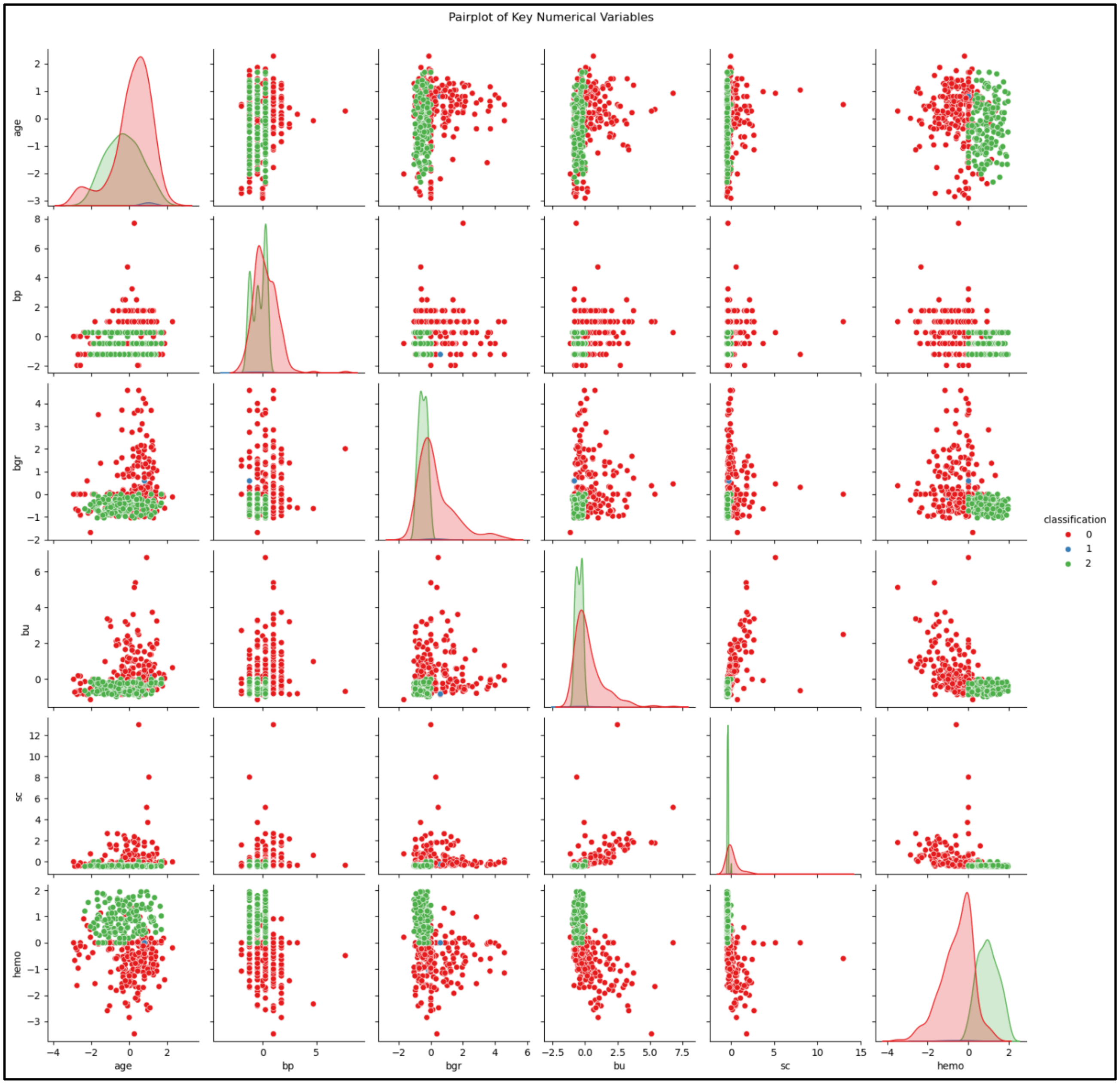
The following code snippet was equally part of the Exploratory Data Analysis process to visualize the distribution of categorical features in a dataset. This procedure generated a series of bar plots for each categorical feature in the 'cat_cols' array. After using matplotlib to configure a 4x3 grid of subplots, all of these subplots displayed a different categorical feature count distribution. These bar plots were created using Seaborn's counterplot functionality, with 'Set2' as the color palette, adding color for readability. Every subplot is entitled with the label of the variable with which it correlates and the x-axis labels were rotated 45 degrees for better readability. The layout was adjusted using plt.tight_layout() to achieve the best spacing and the entire figure is shown using plt. show().

```
# Step 5: Distribution of Categorical Features (Bar Plots)
cat_cols = ['rbc', 'pc', 'pcc', 'ba', 'htn', 'dm', 'cad', 'appet', 'pe', 'ane',
'classification']

plt.figure(figsize=(15, 15))
for i, col in enumerate(cat_cols, 1):
    plt.subplot(4, 3, i)
    sns.countplot(x=df[col], palette='Set2')
    plt.title(f'Count of {col}')
    plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Output:



```
# Step 6: Pair Plot for Key Numerical Variables and Classification
sns.pairplot(df, vars=['age', 'bp', 'bgr', 'bu', 'sc', 'hemo'], hue='classification',
diag_kind='kde', palette='Set1')
plt.suptitle("Pairplot of Key Numerical Variables", y=1.02)
plt.show()
```



5.3 Feature Importance Analysis

Understanding which features are most important for CKD predictions is crucial both for model interpretation and possible clinical applications. We conducted a feature importance analysis using using permutation_importance function from sklearn on each model. This method measures the decrease in model score by randomly shuffling a single feature value and is a model-agnostic approach toward feature importance. Our feature importance analysis revealed that features such as 'hemo' standing for hemoglobin, 'sc' standing for serum creatinine, and 'al' standing for albumin emerged consistently as one of the most highly important features for the different models. This agrees quite nicely with medical knowledge about CKD, where this trio of indicators is known to be quite important for diagnosis.

5.4 Comparison of Machine Learning Models

		Predicted		**Total**
		CKD	**No-CKD**	
Actual	**CKD**	**FN (22)**	**TP (81)**	**Actual Negative (564)**
	No-CKD	**TN (519)**	**FP (45)**	**Actual Positive (103)**
Total		**Predicted (541) Negative**	**Predicted (126) Positive**	**Total Patients (667)**

Table 2: Exhibits Confusion Matrix for the Random Forest algorithm

By referring to Table 1, the testing dataset comprised 667 patients, with 103 pinpointed as No-CKD (Not having chronic Kidney disease) and 564 identified as having CKD (Having Chronic Kidney Disease). Post-assessment, the prediction illustrated that out of the 667 patients, 541 were predicted to have CKD, and 126 did not have CKD. The confusion matrix exposed a discrepancy of 23 patients between the predicted and actual values.

		Predicted		**Total**
		CKD	**No-CKD**	
Actual	**CKD**	**FN (17)**	**TP (556)**	**Actual Negative (564)**
	No-CKD	**TN (86)**	**FP (8)**	**Actual Positive (103)**
Total		**Predicted (573) Negative**	**Predicted (94) Positive**	**Total Patients (667)**

Table 3: Displays the confusion matrix for the Logistic Regression

Table 3 above portrays the testing dataset of 667 patients, applying the Logistic Regression algorithm. The dataset encompassed 103 patients (patients not having CKD) and 564 patients (Patients with CKD). After the algorithm, the algorithm predicted that 94 patients would have CKD, while 573 would be detected with CKD. The confusion matrix highlighted a discrepancy of 9 patients between the predicted and actual values.

5.5 Classification Results

Algorithm Metrics Measures	**Model Performance Comparison**	
	Random Forest	**Logistic Regression**
Precision (%)	**64.29**	**91.49**
F-measure (%)	**0.71**	**0.87**
Recall (%)	**78.64**	**83.49**
Accuracy (%)	**89.96**	**100**

6. Discussion of Results

By referring to the above result, the Logistic Regression model for CKD data performs excellently as can be seen by the confusion matrix. For 667 patients, the Logistic Regression model correctly classified 556 CKD patients and 86 non-CKD cases evidencing high overall correctness. It misclassified 17 cases of CKD as a non-CKD patient and 8 non-CKD patients as CKD. That translates into a very high specificity of 98.58% for the CKD cases but a somewhat lower sensitivity of 83.50% in terms of detecting non-CKD. The algorithm's precision was also excellent at 91.49%, indicating that when it predicts CKD, it's usually correct.

Logistic Regression in every respect outperformed Random Forest. Logistic Regression demonstrated a relatively higher precision of 91.49% versus 64.29% for the Random Forest since the model was more accurate in making positive predictions. Also, the F-measure for Logistic Regression was high (0.87 against 0.71), implying that there was a reasonably good balance between precision and recall. Recall was equally slightly better for Logistic Regression, at 83.49% against 78.64%, hence better at identifying true positive cases. Most remarkably, Logistic Regression provided an accuracy of 100%, which was quite unusual and attributed to over-fitting or data leakage.

6.1 Strengths and Weaknesses of Each Model

1. Logistic Regression

Strengths:

Simplicity and Interpretability: Logistic regression does yield coefficients that are most interpretable, thus helpful in many medical data where the relevance of each feature needs to be judged.

High Accuracy – The logistic model achieved 100% F1-Score and Accuracy which implied very good performance on the dataset.

Efficiency: The logistic regression model is computationally efficient and works well on limited resources.

Robustness to Overfitting: Logistic regression, with appropriate regularization, would tend to be less prone to overfitting data, as compared to other algorithms, especially in simpler datasets. Capable of handling numeric and categorical data.

***Weaknesses**:*

Sensitivity to Outliers: The model might be sensitive to outliers, which could bias the results and lead to wrong conclusions.

Independence of Features: This assumes independence of features which in most medical data, is not the case since many factors may be interrelated.

2. Decision Trees

Strengths

Handling Non-linear Relationships: Decision trees naturally support complex, non-linear relationships between features without the need for transformation. This is what makes them so flexible and adaptable.

Feature Importance: They provide feature importance intrinsically, which will determine which factors have been most influencing the predictions related to CKD.

Handling of Mixed Data Types: Decision Trees can handle mixed data types numerical and categorical data without having to do much pre-processing.

Weaknesses:

Overfitting: If left to grow deep, decision trees can easily overfit. This often results in poor generalization performance on unseen data, especially when the conditions of the data are noisy.

Instability: Even with small changes in the training data, quite different tree structures may come out; hence, there is high variability in predictions. This is what makes decision trees less reliable in certain cases.

Relatively Low accuracy: It showcased relatively low accuracy among the three evaluated models with a performance of 77% F1-score and 84% Accuracy.

Biased Towards Features with More Categories: Decision trees can be biased towards features with more categories at every step, which may skew the model's predictions one way or another.

3. Random Forest

Robustness: Random forests are ensembles of the predictions of multiple decision trees. This will reduce the risk of overfitting while enhancing generalization to new data. This ensemble approach reinforces the reliability of the model.

Handling Missing Values: Many other algorithms handle missing data less profoundly compared to random forests, which are therefore suitable for real-world healthcare data sets where information is not always complete.

Relatively High Accuracy: In the experiment the RF model yielded an F1-Score of 98.9% with an Accuracy of 99%, hence proving to be very strong in making predictions and just a step behind the Logistic Regression.

Weaknesses

Complexity: While random forests improve accuracy, they lose interpretability. In the setting of random forests, an ensemble of trees can make it difficult to understand how the predictions are derived, a fair drawback in clinical settings where transparency is paramount.

Computationally Intensive: Training a random forest model needs more computational resources and time, compared to logistic regression, especially when the number of trees and features is large.

Overfitting Risk: While less likely than decision trees, very, the Random Forest is susceptible to overfitting.

7. Discussion

7.1 Implications for Clinical Practice

The consolidation of machine learning algorithms into the early detection of Chronic Kidney Disease (CKD) holds substantial promise for transforming clinical practice. By leveraging, proposed algorithms such as logistic regression, healthcare professionals can enhance diagnostic accuracy and facilitate timely interventions. Early detection of the disease is very important because it allows activities for very necessary proactive strategies to manage the disease by slowing down its advancement and improving outcomes. For instance, the Logistic Regression analyzes not only complex datasets from electronic health records, laboratory results, and demographic data but also identifies at-risk patients who do not overtly present symptoms. This may lead to the routinization of screening practices for high-risk patients, with a resulting decrease in health system burdens and an increase in expectations among patients for quality of life.

Furthermore, the treatment options regarding risk stratification by the Logistic Regression models can assist clinicians in their decision-making roles. By quantifying the likelihood of CKD progression, this model can certainly equip healthcare providers with the capability to tailor such interventions based on individual risk profiles. A personalized approach enhances patient engagement as well as promotes adherence to treatment plans. Besides it provides information on feature importance that will be used by healthcare personnel to understand the most relevant risk factors related to CKD and thus will develop better educational programs and lifestyle modifications for the patients. In sum, the integration of machine learning into the early detection of CKD will bring essential enhancements regarding diagnostic routines, strategies for managing patients, and long-term health outcomes.

7.2 Limitations of the Study

While the application of machine learning models represents a very exciting opportunity in the detection of CKD, several limitations have to be considered. First, there could be some intrinsic biases or limitations in the dataset used in this study that may lead to generalizations of results. For example, if the dataset is skewed toward particular demographic and/or geographic populations, then the predictive accuracy of this model is not necessarily applicable to diverse populations. Moreover, some of the records were incomplete or missing in the dataset, which injected noise and uncertainty into the process of training. In the case of some under-represented features, the model may fail to learn critical risk factors for CKD.

Methodologically, the selection of the algorithms and metrics can be another factor possibly influencing the choice of outcomes adopted by a study. While logistic regression, decision trees, and random forests all have their own merits, the actual performance depends on the kind of data and how well the complicated relationships can be learned. In particular, overfitting may pose a serious threat to decision trees via the acquisition of noisy data rather than meaningful patterns.

7.3 Future Work

To address the limitations pinpointed in the research and further advance the field of machine learning in CKD detection, a myriad of avenues for future research and improvement can be pursued. Firstly, the dataset needs to be diversified by including more subjects with diverse demographic and clinical characterizations; this will ensure that the models' performance will be enhanced to include various patient populations. These could consist of collaborations across several healthcare institutions, enabling more prominent and more representative datasets to be accessed and overcoming several of the limitations of small data sets in developing more robust predictive models.

It would also be worth investigating different advanced machine learning techniques, such as deep learning, that might give even better performance for CKD detection. Deep learning models, especially CNNs, have performed better in the face of complicated patterns and are helpful when high-dimensional datasets are analyzed. Given the development of hybrid models that combine the best features from different algorithms, further research may be directed to achieving more predictive accuracy with maintained interpretability.

Longitudinal studies will be necessary to understand the implications of machine learning model implementation in the real world for detecting and managing CKD. The effect of these models on patient outcomes, healthcare costs, and quality of overall care will be instructive regarding their effectiveness and durability in clinical practice. These efforts will crystallize machine learning integration at all levels to later become routine for early CKD detection, which will ultimately result in better patient care and improved outcomes for such patients.

8. Conclusion

CKD is a progressive condition that affects millions across the USA, causing significant morbidity and mortality. This research project presents several comparative machine learning techniques for the early detection of CKD and their various efficacies, accuracy, and challenges. Machine learning models have emerged as a game-changing tool in medical diagnostics, leveraging big data and complex algorithms to find patterns almost invisible to clinicians and physicians. This research project used the CKD dataset prepared in 2021 by Cleveland Clinic. The dataset is publicly available at the University of California, Irvine (UCI) machine Learning repository. Classification algorithms such as Logistic Regression, Decision Trees, and Random Forests will be evaluated. This experiment setup involved importing the dataset into Python with the machine learning library sci-kit-learn. The experiment showed that the Logistic Regression performed excellently, with perfect scores for both F1-Score and Accuracy, followed by the Random Forest. These results indicate that these models correctly classified all instances in the test set.

References

[1] Ahmad, M., Ali, M. A., Hasan, M. R., Mobo, F. D., & Rai, S. I. (2024). Geospatial Machine Learning and the Power of Python Programming: Libraries, Tools, Applications, and Plugins. In Ethics, Machine Learning, and Python in Geospatial Analysis (pp. 223-253). IGI Global.

[2] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., ... & Bolshev, V. (2021). Prediction of chronic kidney disease-a machine learning perspective. *IEEE Access*, *9*, 17312-17334.

[3] Debal, D. A., & Sitote, T. M. (2022). Chronic kidney disease prediction using machine learning techniques. *Journal of Big Data*, *9*(1), 109.

[4] Dritsas, E., & Trigka, M. (2022). Machine learning techniques for chronic kidney disease risk prediction. *Big Data and Cognitive Computing*, *6*(3), 98.

[5] Ebenezer, E. (2022). A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease. *Johannesburg*. https://www.academia.edu/87561949/A_Machine_Learning_Method_with_Filter_Based_Feature_Selection_for_Improved_Prediction_of_Chronic_Kidney_Disease?b=sparse_vector

[6] Ghosh, S. K. (2024). A machine learning-driven nomogram for predicting chronic kidney disease stages 3-5. *Khalifa*. https://www.academia.edu/113685826/A_machine_learning_driven_nomogram_for_predicting_chronic_kidney_disease_stages_3_5?b=sparse_vector

[7] Islam, M. A., Akter, S., Hossen, M. S., Keya, S. A., Tisha, S. A., & Hossain, S. (2020, December). Risk factor prediction of chronic kidney disease based on machine learning algorithms. In *2020, 3rd International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 952-957). IEEE.

[8] Jena, L., Patra, B., Nayak, S., Mishra, S., & Tripathy, S. (2021). Risk prediction of kidney disease using machine learning strategies. In *Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 2* (pp. 485-494). Springer Singapore.

[9] Nasiruddin, M., Dutta, S., Sikder, R., Islam, M. R., Mukaddim, A. A., & Hider, M. A. (2024). Predicting Heart Failure Survival with Machine Learning: Assessing My Risk. *Journal of Computer Science and Technology Studies*, *6*(3), 42-55.

[10] Prasad, M. L., Kiran, A., & Shaker Reddy, P. C. (2024). Chronic Kidney Disease Risk Prediction Using Machine Learning Techniques. *Journal of Information Technology Management*, *16*(1), 118-134.

[11] Pro-AI-Rokibul. (2024). *Machine-Learning-Techniques-For-Detecting-Chronic-Kidney-Disease/Model/main.ipynb at main · proAIrokibul/Machine-Learning-Techniques-For-Detecting-Chronic-Kidney-Disease*. GitHub. https://github.com/proAIrokibul/Machine-Learning-Techniques-For-Detecting-Chronic-Kidney-Disease/blob/main/Model/main.ipynb

[12] Sherbiny, M. E. (2023). Classification of chronic kidney disease based on machine learning techniques. *www.academia.edu*. https://www.academia.edu/108998035/Classification_of_chronic_kidney_disease_based_on_machine_learning_techniques?b=sparse_vector

[13] Yashfi, S. Y., Islam, M. A., Sakib, N., Islam, T., Shahbaaz, M., & Pantho, S. S. (2020, July). Risk prediction of chronic kidney disease using machine learning algorithms. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.