Title of the Thesis

Computer-aided Chronic Kidney Disease Detection: A Comparative Study of Machine Learning Algorithms

Course: **Thesis – Fall 2024**

Course Code**: CIS499**

By

Ismet Zahan Sithi

ID: 211-16-559

Department of CIS

Daffodil International University

Under the Guidance of

Mr. Israfil

Lecturer

Department of CIS

Daffodil International University

Date of Submission: 12-01-2025

APPROVAL

This thesis titled **"Computer-aided Chronic Kidney Disease Detection: A Comparative Study of Machine Learning Algorithms"**, Submitted by Ismet Zahan Sithi, ID No:211-16-559 to the Department of Computing and Information Systems, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computing & Information Systems and approved as to its style and contents. The presentation has been held on 12-01-2025.

BOARD OF EXAMINERS

12.01.25

Md Sarwar Hossain Mollah — **Chairman**
Associate Professor and Head
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

12.01.25

Md. Nasimul Kader — **Internal Examiner**
Assistant Professor
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

12.01.25

Md. Mehedi Hassan — **Internal Examiner**
Lecturer (Senior Scale)
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

Shahin 12.01.2025

Dr. Muhammad Shahin Uddin — **External Examiner**
Professor
Department of ICT
Mawlana Bhashani Science and Technology University

DECLARATION

I hereby declare that; this thesis has been done by me under supervision of **Mr. Israfil, Lecturer**, department of Computing and Information System (CIS) of Daffodil International University. I am also declaring that this thesis or any part of there has never been submitted anywhere else for the award of any educational degree like, B.Sc., M.Sc., Diploma or other qualifications.

Supervised By

12.01.25

Mr. Israfil
Lecturer
Department of CIS
Daffodil International University

Submitted By

Sithi

Name: Ismet Zahan Sithi
ID: 211-16-559
Department of CIS
Daffodil International University

ACKNOWLEGDEMENTS

First and foremost, I am deeply grateful to the Almighty Creator for granting me the strength and ability to reach this stage in my academic journey. His eternal blessings have been a source of inspiration and guidance throughout this process.

I would like to express my sincere gratitude to my esteemed supervisor, Mr. Israfil, Lecturer, Department of Computing and Information System, Daffodil International University. His unwavering support, invaluable advice, and continuous encouragement were instrumental in the successful completion of this thesis. Without his patience, guidance, and careful review, this work would not have been possible.

I would also like to extend my heartfelt thanks to all the faculty members of the Department of Computing and Information System for their valuable suggestions and encouragement throughout the course of my research. Special mention goes to Prof. Mr. Md. Sarwar Hossain Mollah, Mr. Md. Nasimul Kader and Mr. Md. Mehedi Hassan, whose support and inspiration greatly contributed to my academic progress.

Finally, I would like to express my deepest gratitude to my beloved family—my parents, sister, and brother—whose moral support and encouragement have been a constant source of strength throughout this journey.

ABSTRACT

Chronic Kidney Disease (CKD) continues to pose a significant healthcare challenge, especially in rural areas of developing countries like Bangladesh, where access to affordable and effective diagnostic services is extremely limited. Early detection of CKD is crucial for slowing disease progression and improving patient outcomes. However, the diagnostic methods currently available are often expensive and technologically advanced, making them inaccessible to rural populations. To overcome these challenges, this thesis explores the application of machine learning (ML) models to enhance CKD diagnosis in a cost-effective manner, specifically targeting rural communities with limited healthcare resources. The primary objective of this study is to leverage machine learning to improve the accuracy of CKD diagnosis in settings with small and imbalanced datasets. Many existing ML models are trained on datasets that are predominantly composed of healthy individuals, resulting in high accuracy but poor sensitivity, leading to missed diagnoses of CKD patients. To address this, we implemented data balancing techniques and fine-tuned hyperparameters to enhance the performance of the models for accurate CKD detection. After optimizing the Support Vector Classifier (SVC), we achieved an impressive 95% accuracy with an AUC of 0.9952. Logistic Regression also performed well, reaching 97% accuracy and an AUC of 0.9986. The Random Forest classifier outperformed all other models, achieving perfect classification with 100% accuracy and an AUC of 1.0. These results suggest that optimized machine learning models hold great potential as a low-cost, accurate, and accessible strategy for early CKD detection, particularly in rural regions of Bangladesh where healthcare services are scarce. By implementing such models, it is possible to significantly improve patient outcomes while reducing the financial burden on healthcare systems.

Keywords: Kidney Disease Detection, CKD Detection, Machine learning, CKD UCI.

PREFACE

Chronic Kidney Disease (CKD) is a growing health concern, particularly in rural areas of developing countries like Bangladesh, where access to affordable diagnostic services is limited. Early detection of CKD is critical to slow its progression and improve patient outcomes. This thesis explores the use of machine learning (ML) models to enhance CKD diagnosis in a cost-effective manner, focusing on rural populations with limited healthcare resources.

The thesis is structured into eight chapters:

Chapter 1: Introduction

This chapter introduces the problem of CKD, the importance of early detection, and the challenges in diagnosing CKD in rural settings. It outlines the study's aim to use ML models to improve diagnosis accuracy.

Chapter 2: Related Works

A review of previous research on machine learning techniques applied to CKD detection, highlighting the models used and the challenges addressed, such as imbalanced datasets.

Chapter 3: Methodology

Describes the machine learning models and techniques employed, including data preprocessing, model selection, and hyperparameter tuning to optimize performance.

Chapter 4: Evaluation Metrics

This chapter discusses the evaluation metrics used to assess model performance, such as accuracy, precision, recall, F1-score, and AUC**.**

Chapter 5: Results

Presents the results of the experiments, including model performance and comparison of different machine learning algorithms**.**

Chapter 6: Comparative Analysis

Compares the results with other similar studies, highlighting strengths and weaknesses in different approaches to CKD detection.

Chapter 7: Conclusion

Summarizes the findings of the study, emphasizing the effectiveness of machine learning models, particularly Random Forest, in achieving high accuracy for CKD detection.

Chapter 8: Future Work

Outlines potential directions for future research, including testing on larger datasets, exploring more complex models, and real-world deployment in resource-limited environments.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1Background and Motivation

Chronic Kidney Disease (CKD) is a progressive condition characterized by the gradual loss of kidney function, which can lead to severe health complications, including cardiovascular diseases and end-stage renal failure. Early detection of CKD is crucial to improving patient outcomes, reducing healthcare costs, and preventing disease progression. Traditional diagnostic methods rely heavily on clinical expertise and laboratory tests, which are often time-consuming and subject to human error. With the advent of machine learning (ML) in healthcare, there is growing interest in leveraging data-driven approaches to enhance the accuracy and efficiency of CKD detection.

The increasing prevalence of CKD globally has intensified the need for effective diagnostic tools. According to recent studies, millions of individuals remain undiagnosed or are diagnosed at later stages, leading to suboptimal management and poorer health outcomes. This scenario underscores the importance of developing computer-aided diagnostic systems capable of identifying CKD at its earliest stages, when interventions can significantly improve the quality of life and reduce mortality rates.

1.2Machine Learning in Healthcare

Machine learning has revolutionized healthcare by enabling automated, scalable, and accurate disease diagnosis. By analyzing large and complex datasets, ML algorithms can identify subtle patterns and correlations that may be overlooked by traditional methods. This capability is particularly valuable in chronic disease detection, where early and accurate identification can significantly impact patient management and prognosis. In the context of CKD, machine learning offers the potential to develop robust, cost-effective, and accessible diagnostic tools that support clinical decision-making.

Furthermore, ML models can incorporate a wide range of clinical, biochemical, and demographic data, enabling comprehensive analysis and risk prediction. Techniques such as

supervised learning, unsupervised learning, and ensemble methods have been employed in healthcare to address various diagnostic challenges. The ability to continuously improve and adapt these models through training with new data makes them an indispensable asset in modern medicine.

1.3 Contribution

This study aims to contribute to the field of CKD detection by developing a robust and efficient framework for automated diagnosis. By conducting a comparative analysis of various machine learning algorithms, the study seeks to identify models that achieve high diagnostic accuracy while addressing critical challenges such as data imbalance and the risk of false negatives and false positives. The proposed framework leverages advanced techniques in feature selection, model optimization, and performance evaluation to ensure reliability and clinical applicability.

Key contributions include:

1. Development of a comprehensive CKD detection framework that integrates data preprocessing, feature engineering, and model training.
2. Evaluation of multiple machines learning algorithms, including decision trees, support vector machines, neural networks, and ensemble methods, to determine their effectiveness in CKD detection.
3. Implementation of techniques to address class imbalance, such as resampling methods and cost-sensitive learning, to improve model performance.
4. Comparative analysis to identify the most suitable algorithm for CKD diagnosis, considering metrics such as accuracy, sensitivity, specificity, and F1-score.
5. Thorough exploratory data analysis (EDA), analysis of variance (ANOVA) tests, and feature selection to ensure optimal data preparation and model inputs.
6. Optimization of model parameters using grid search with cross-validation to enhance overall performance and robustness.

1.4 Challenges in CKD Detection

The detection of CKD presents several challenges, including:

- **High False Negative and False Positive Rates:** Misclassification can lead to delayed treatment or unnecessary interventions, compromising patient safety and increasing healthcare costs. False negatives pose a significant risk as they delay critical interventions, while false positives can lead to patient anxiety and unwarranted medical procedures.
- **Imbalanced Data:** CKD datasets often exhibit a skewed distribution, with fewer instances of the disease compared to non-disease cases. This imbalance can bias model performance, as algorithms may favor the majority class, leading to poor detection rates for CKD cases.
- **Heterogeneous Data Sources:** CKD detection relies on diverse data types, including clinical, biochemical, and demographic information. Integrating and preprocessing these heterogeneous data sources require robust methodologies to ensure consistency and accuracy.
- **Generalizability:** Ensuring that ML models perform consistently across different populations and healthcare settings is critical. Variations in patient demographics, clinical practices, and data quality can impact model performance, necessitating rigorous validation and testing.

1.5 Objective

The primary objective of this research is to design and evaluate machine learning models for accurate and early detection of CKD. The study focuses on identifying algorithms that can:

1. Minimize false negatives, ensuring that CKD cases are not overlooked.
2. Reduce false positives, avoiding unnecessary diagnostic procedures.
3. Handle imbalanced datasets effectively to ensure robust performance across diverse patient populations.
4. Provide reliable results to aid clinicians in understanding and applying the model's predictions.

CHAPTER 2: RELATED WORKS

Chronic Kidney Disease (CKD) remains a pressing global health issue, with an estimated 850 million people affected worldwide (Francis et al., 2024). Despite its prevalence, CKD often goes undetected until it reaches advanced stages, resulting in severe complications such as end-stage renal disease (ESRD). Kovesdy et al. (2022) underscore that at least one-tenth of the global population is affected by CKD, with millions progressing to ESRD annually, posing a major burden on healthcare systems. The challenge of early detection is compounded by limitations in current diagnostic methods, which frequently produce high false-positive rates and lead to unnecessary treatments. A significant barrier to improving diagnostic accuracy using machine learning (ML) is the imbalance in CKD datasets, where diseased cases are often underrepresented, making it difficult for models to correctly identify the minority class (Yildirim et al., 2017). Furthermore, hyperparameter optimization plays a crucial role in enhancing model accuracy, addressing both the false positives and the model's ability to detect CKD reliably.

Machine learning has shown great promise in improving CKD detection, but it is critical to address issues such as data imbalance and hyperparameter optimization to achieve better diagnostic performance. Several studies have contributed significantly to this field by exploring various algorithms, feature selection techniques, and hybrid models. Qin et al. (2019) utilized the UCI CKD dataset and applied KNN imputation for handling missing values, achieving a diagnostic accuracy of 99.75%. This demonstrated the feasibility of using ML methods for CKD classification and set a benchmark for future research. On a similar note, Ilyas et al. (2022) explored the use of Random Forest and J48 algorithms for classifying CKD stages based on Glomerular Filtration Rate (GFR). While the achieved accuracy of 85.5% was lower, their research emphasized the potential of ML in CKD classification, offering valuable insights into the limitations and opportunities for model improvement. Fattah et al. (2022) further demonstrated that hybrid machine learning models could significantly improve early CKD detection, offering more reliable predictions by combining multiple techniques.

Arif et al. (2023) advanced the field with a novel approach that integrated iterative imputation, sequential data scaling, and the Boruta algorithm for feature selection, resulting in an impressive

accuracy of 100% on the UCI CKD dataset. This study highlighted the importance of preprocessing and effective feature selection strategies in machine learning. Similarly, Khalid et al. (2022) employed an ensemble logistic regression model that combined Naive Bayes-Gaussian, Gradient Boosting, and Random Forest, achieving perfect accuracy. Their work underscored the power of ensemble models in improving prediction accuracy for CKD detection.

The use of boosting algorithms for CKD detection was also explored by Ganie et al. (2023), who compared the performance of five different boosting algorithms: XGBoost, CatBoost, LightGBM, AdaBoost, and Gradient Boost. Their results revealed that AdaBoost performed the best, with an average accuracy of 98.47%. This highlighted the importance of selecting the right model for the task at hand. Building on this, Saif et al. (2024) developed a deep ensemble model that combined Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, achieving an accuracy of 0.993. Their work showcased the potential of deep learning approaches for improving CKD detection, offering a promising direction for future research.

The role of feature selection in enhancing model performance was emphasized by Dasgupta et al. (2023), who achieved an accuracy of 99.62% using Gradient Boosting along with effective feature selection. Yedilkhan Amirgaliyev et al. (2018) and Md. Mustafizur Rahman et al. (2024) also highlighted the significance of selecting the right features to improve the accuracy of CKD detection models. Elias et al. (2022) utilized feature ranking techniques to identify the most significant predictors for CKD detection, further reinforcing the importance of feature selection. Additionally, Al-Momani et al. (2022) demonstrated that various machine learning classifiers, when properly optimized, could achieve high accuracy rates for CKD prediction.

Chittora (2021) and Ahmed et al. (2018) emphasized the importance of predictive analytics and feature engineering, achieving remarkable performance metrics such as an AUC of 0.995. Their work further reinforced the need for careful parameter selection and preprocessing steps to ensure optimal model performance. Swain (2022) and Muhammad Shoaib et al. (2023) also underscored the importance of hyperparameter optimization, demonstrating that tuning these parameters is crucial for maximizing model accuracy. Tahsin et al. (2019) explored the use of ECG signals for CKD detection, achieving a validation accuracy of 97.6%, which opened up new possibilities for incorporating diverse data sources into CKD prediction models.

Azian et al. (2020) confirmed that combining feature selection methods with machine learning classifiers, such as Random Forest, resulted in an accuracy of 98.825%. Similarly, Razib et al.

(2023) demonstrated that hybrid models, which combine multiple machine learning techniques, can outperform traditional methods, achieving an impressive accuracy of 99%. These findings further support the integration of hybrid models in CKD detection, offering a pathway to more reliable and accurate diagnostic systems.

In summary, the reviewed studies collectively show that machine learning has the potential to significantly improve CKD detection. By addressing data imbalance, employing robust feature selection techniques, and optimizing hyperparameters, ML models can achieve higher accuracy rates and contribute to the early diagnosis of CKD. This research aims to build upon these advancements to develop more effective diagnostic systems, ultimately improving patient outcomes and reducing the burden of CKD on healthcare systems worldwide.

CHAPTER 3: METHODOLOGY

3.1 Data Collection

This study uses a dataset from the UCI Machine Learning Repository, which includes 25 features in total. These features are divided into 13 categorical, 11 numerical, and 1 Boolean attribute. The classification variable separates the data into two groups: individuals diagnosed with chronic kidney disease (CKD) and those without the condition, labeled as NotCKD.

Table 1 Summary of the dataset

Class	**Value**
CKD	250
NotCKD	150

The dataset consists of 400 samples, with 250 instances representing CKD patients and 150 representing individuals without CKD. This indicates an imbalance between the two classes, where the CKD cases are more prevalent than the NotCKD cases, highlighting a challenge often encountered in datasets used for classification tasks.

Table 2 presents a detailed description of each variable used in this study. Interpretation of how machine learning models perform in diagnosing CKD relies on this breakdown. By reading through this whole section, readers can appreciate the importance of each feature in predicting CKD and, more importantly, a understanding of how these variables interact with each other in the dataset.

Table 2 Description of Dataset features

Feature	**Description**
age	Patients’
bp	Blood pressure.
sg	Specific gravity of urine
al	Albumin levels in urine

su	Sugar levels in urine.
rbc	red blood cells in urine
pc	Pus cells in urine
pcc	pus cell clumps
ba	Bacteria in urine
bgr	Blood glucose level
bu	Blood urea
sc	Serum creatinine
sod	Serum sodium levels
pot	Serum potassium levels
hemo	Hemoglobin levels
pcv	Packed cell volume
wc	White blood cell count
rc	Red cell count
htn	Hypertension status
dm	Diabetes status
cad	Coronary artery disease status
appet	Appetite status
pe	Edema status
ane	Anemia status
classification	Indicates the patient has CKD or NotCKD

3.2 Preprocessing

Preprocessing plays a pivotal role in machine learning, as it directly impacts the performance and predictive accuracy of models. This study, focused on predicting chronic kidney disease (CKD), involves various data preparation stages to optimize model learning. Organizing the data correctly ensures that the model can efficiently identify patterns and make more reliable predictions.

The first step in the process is encoding the categorical variables. Since most machine learning algorithms require numerical inputs, it is essential to convert categorical data into a usable form. The CKD dataset consists of 25 features, 13 of which are categorical. These categorical features are transformed into numeric binary values using label encoding, making them compatible with machine learning models and setting the stage for more accurate analyses.

Once encoding is complete, the dataset is examined for missing values. As illustrated in **Figure 1**, the distribution of missing data across each column is visualized, showing which features have missing values and their proportion. Given the presence of missing values in several columns, mean imputation is applied, as there are no extreme outliers in the data. This ensures that the dataset remains consistent and statistically sound, preventing the model from being influenced by missing or incomplete data.

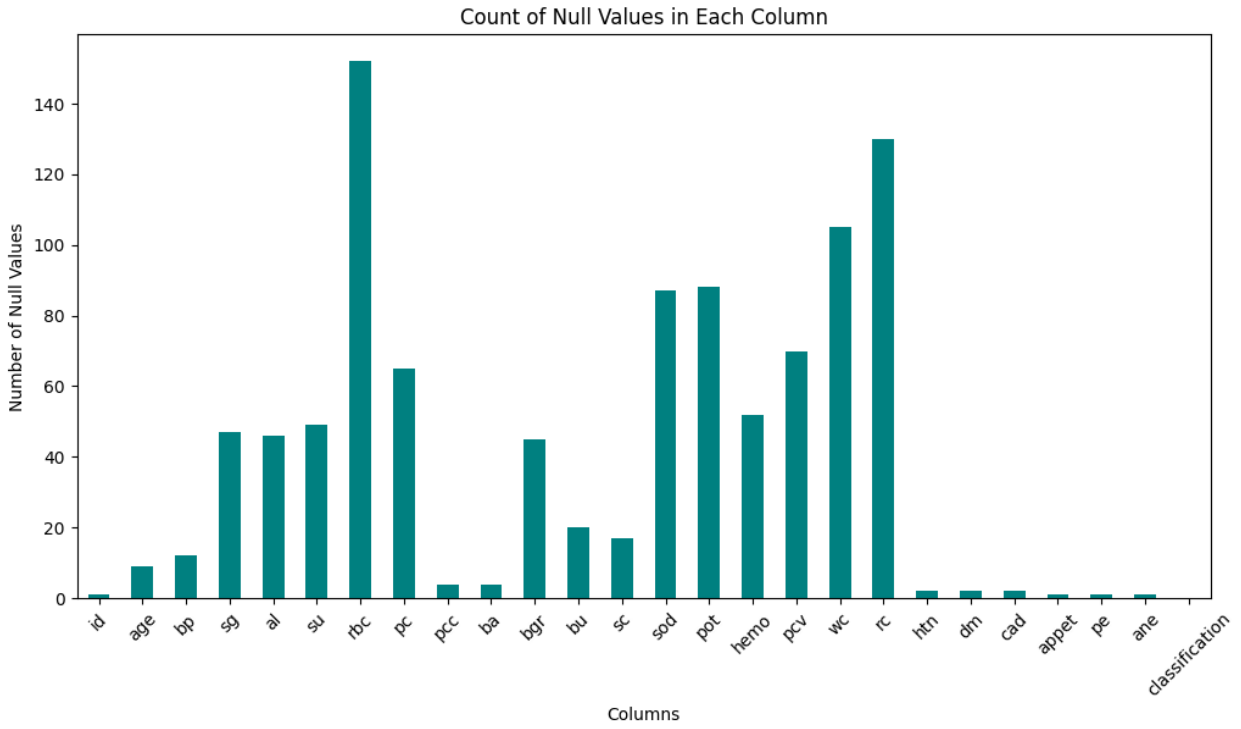

Figure 1 shows the missing data distribution across the column

After handling missing values, outlier detection is carried out using the Interquartile Range (IQR) method. This technique involves calculating the IQR between the first and third quartiles and setting a threshold of 1.5 times the IQR to identify potential outliers. As shown in **Figure 2**, the box plot of the dataset visually represents the distribution of values across each variable. No significant outliers are detected, indicating that the data points fall within acceptable ranges, maintaining the consistency of the dataset.

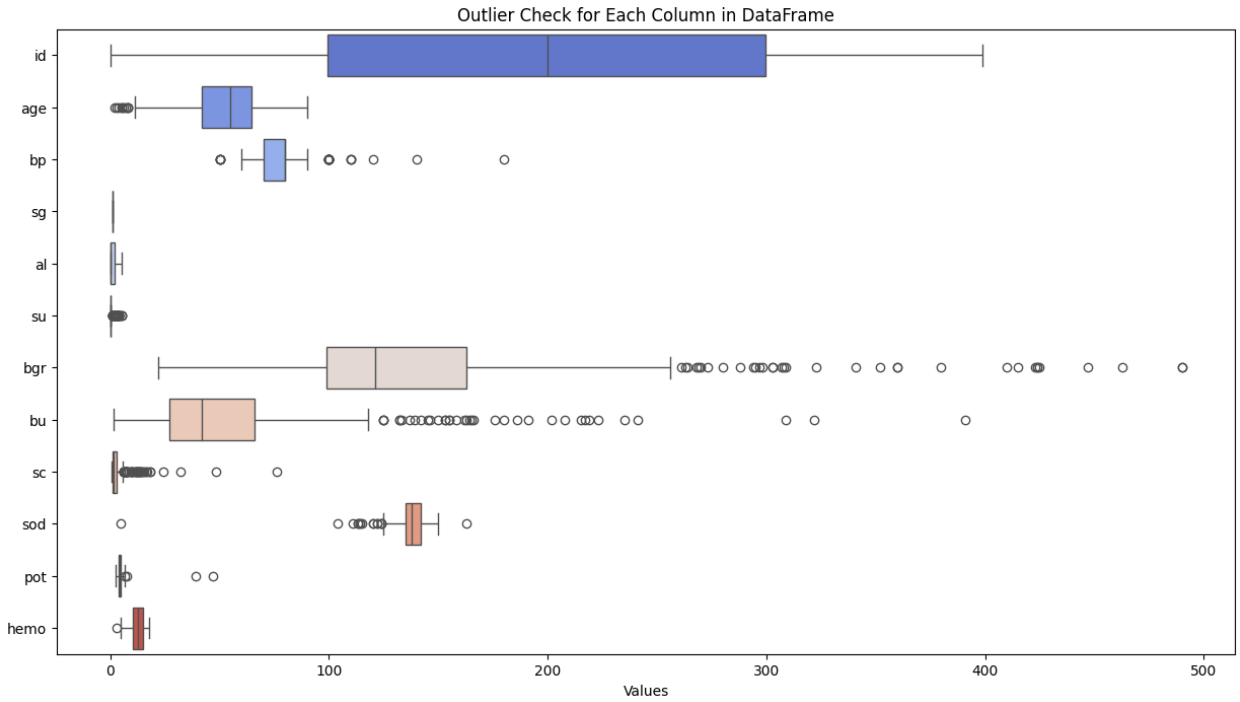

Figure 2 illustrates the box plot of the dataset

Next, standardization is applied to the data. This step is crucial to ensure that each feature contributes equally to the model's performance, particularly when the features have different scales or magnitudes. Standardization involves transforming the dataset so that each variable has a mean of zero and a standard deviation of one, achieved through Z-score normalization. This process reduces the impact of varying scales and enhances model accuracy and consistency.

Table 3 shows data before and after being standardization

Before Standardization

Column	Max Value	Min Value
Age	90.0	2.0
bp	180.0	50.0
sg	1.025	1.005
al	5.0	0.0
su	5.0	0.0
rbc	2.0	0.0
pc	2.0	0.0
pcc	2.0	0.0
ba	2.0	0.0
bgr	490.0	22.0
bu	391.0	1.5
sc	76.0	0.4
sod	163.0	4.5
pot	47.0	2.5
hemo	17.8	3.1
pcv	44.0	0.0
wc	92.0	0.0
rc	46.0	0.0
htn	2.0	0.0

After Standardization

Column	Max Value	Min Value
age	2.27186668	-2.9187301
bp	7.69206700	-1.9665802
sg	1.41572746	-2.3137635
al	3.134467519	-0.8002895
su	4.425073711	-0.437796
rbc	1.12652052	-1.9284503
pc	1.7324816	-1.6397453
Pcc	5.2128603	-0.347524
ba	6.4390634	-0.250872
bgr	4.5784912	-1.687483
bu	6.810428	-1.136130
sc	12.99848	-0.476333
sod	2.770793	-14.47103
pot	15.04577	-0.755344
hemo	1.943974	-3.474833
pcv	1.345334	-2.838776
wc	0.980635	-2.29170
rc	1.076261	-2.57881
htn	3.27798	-0.7626

Table 3 shows data before and after being standardized and normalizing the maximum and minimum values

The dataset also suffers from class imbalance, with 250 instances of CKD and only 150 instances of non-CKD. Such imbalance could lead to decreased performance for the minority class, particularly in terms of precision, recall, and F1 score. To address this issue, Synthetic Minority

Over-sampling Technique (SMOTE) is applied to the training dataset, ensuring that both classes are balanced. **Figure 3** provides a visual representation of the class distribution before and after applying SMOTE, demonstrating the effect of this balancing technique. The test dataset is left unchanged to reflect real-world occurrences and ensure the validity of the evaluation metrics.

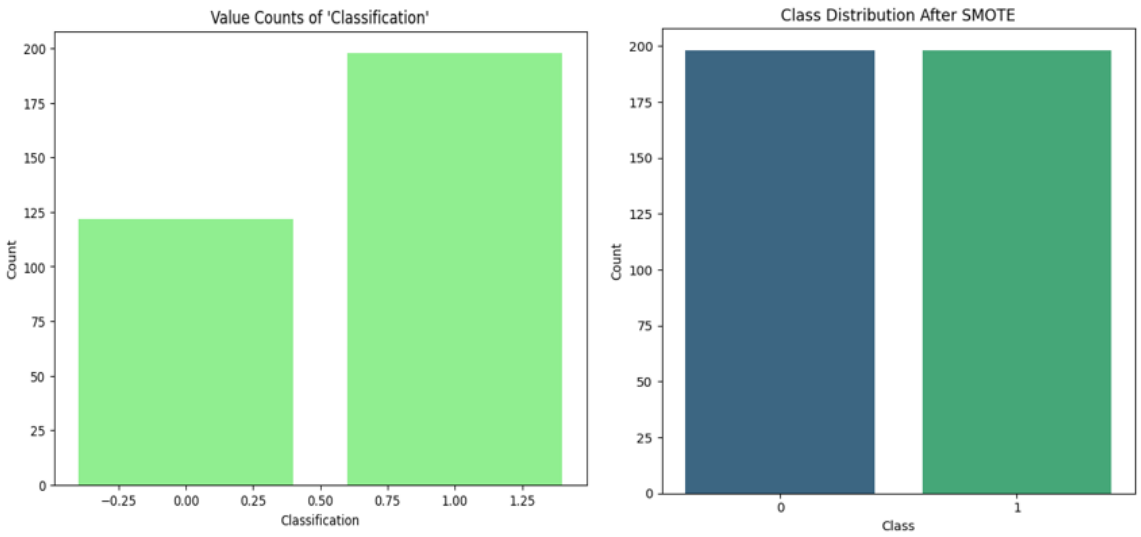

Figure 3 shows the class distribution before and after balancing using SMOTE

Feature selection is another key step in the preprocessing pipeline. An ANOVA test is performed to identify which features are significant for classifying CKD and non-CKD patients. The p-values for each feature tested in the ANOVA test are presented in **Table 4,** which highlights those features that are statistically relevant for prediction. Features with p-values greater than 0.05 are excluded from the dataset, such as "rc," "ba," and "pot," which had p-values of 0.138, 0.263, and 0.125, respectively.

Table 4 contains the ANOVA test below

Feature	P-Value	Important
hemo	1.134e-60	Important
sg	6.44e-60	Important
al	2.398e-40	Important

htn	7.181e-33	Important
bgr	6.437e-17	Important
dm	1.939e-14	Important
bu	2.7591e-14	Important
appet	2.6000e-14	Important
pe	6.04340e-13	Important
pcv	6.48600e-13	Important
sod	1.9530e-12	Important
su	1.7880e-11	Important
ane	1.25770e-09	Important
sc	2.0190e-09	Important
bp	3.1695e-09	Important
rbc	1.24960e-06	Important
age	5.29560e-06	Important
pcc	1.970e-03	Important
cad	2.08250e-03	Important
pc	2.225630-02	Important
wc	3.65024e-02	Important
pot	1.26510e-01	Not Important
rc	1.38030e-01	Not Important
ba	2.62654e-01	Not Important

Figure 4 shows the Pearson correlation heatmap, visualizing the relationship between each feature in the dataset. The heatmap uses color to indicate positive and negative correlations, with red representing high positive correlations and blue indicating high negative correlations. This analysis provides valuable insights into feature interdependencies and aids in feature reduction to improve the model's predictive capabilities.

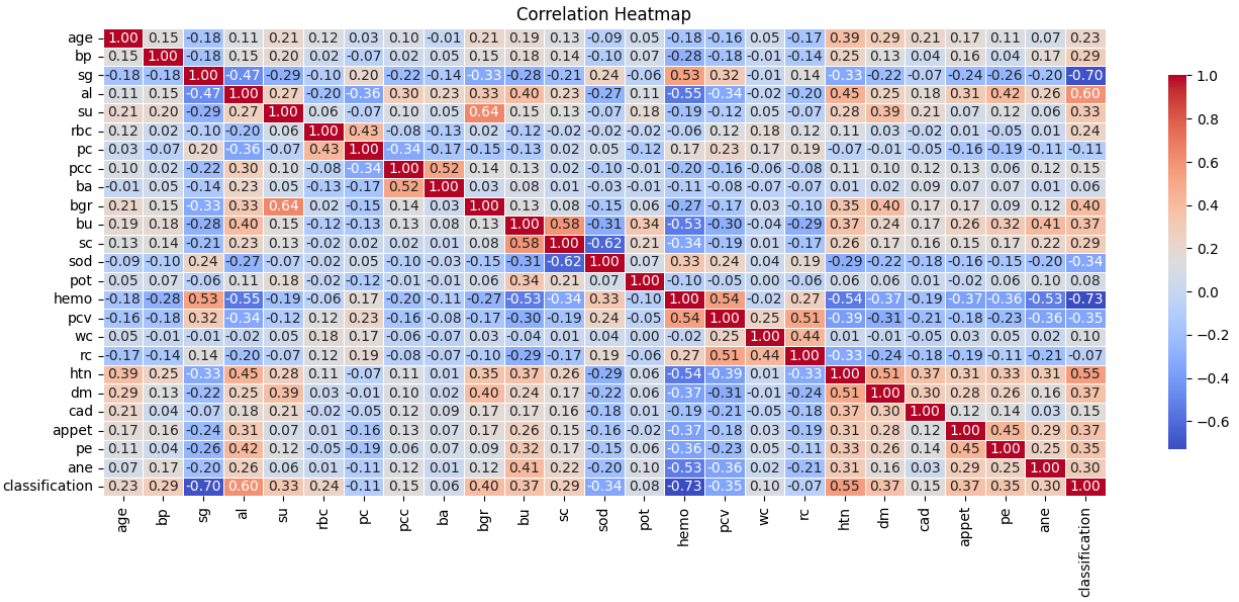

Figure 4 shows the correlation of the feature

Finally, the dataset is split into training and testing sets. The training set, consisting of 198 CKD instances and 122 non-CKD instances, is used to train the model. The testing set, with 52 CKD instances and 28 non-CKD instances, is reserved for evaluating model performance. This 80-20 split ensures that the model is trained on a substantial portion of the data while preserving a separate set for performance evaluation, reducing the risk of overfitting.

Table 5 displays the value counts of the 'classification' in both the training and testing datasets

Dataset	Classification (1/0)	Value Count
Training	1	197
Training	0	123
Testing	1	51
Testing	0	29

Also **Figure 5** is added to get a clear idea about overall work flow.

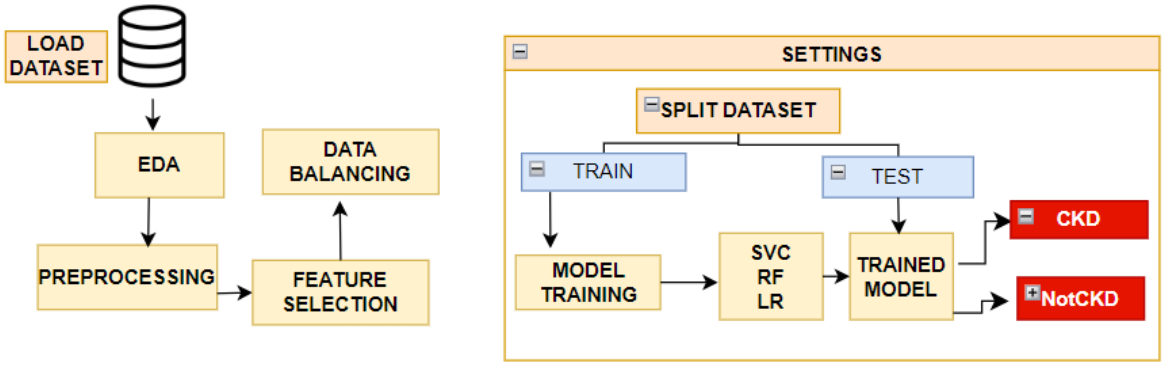

***Figure* 5** Workflow Diagram

3.3 Model Training and Selection

For the model training phase, three well-known machine learning models were employed: Random Forest (RF), Support Vector Classifier (SVC), and Logistic Regression (LR). The purpose was to train these models on the preprocessed data to enable accurate predictions of chronic kidney disease (CKD) and to fine-tune hyperparameters for optimal model performance. Random Forest operates by creating multiple decision trees, each built using random subsets of the dataset. These trees are constructed by considering all the features in the data, which allows the model to uncover both linear and nonlinear relationships within the dataset. The strength of Random Forest lies in its ability to aggregate the predictions from all the trees, which leads to a more robust and stable final prediction. This characteristic is particularly useful in handling complex clinical data, such as the various markers related to CKD. In the case of CKD detection, Random Forest uses features like blood pressure, serum creatinine, and blood urea to better understand the intricate relationships within the data. The model's feature importance score, shown in **Figure 6**, highlights the contribution of each feature to the final prediction, with higher scores indicating features that play a more significant role in detecting CKD.

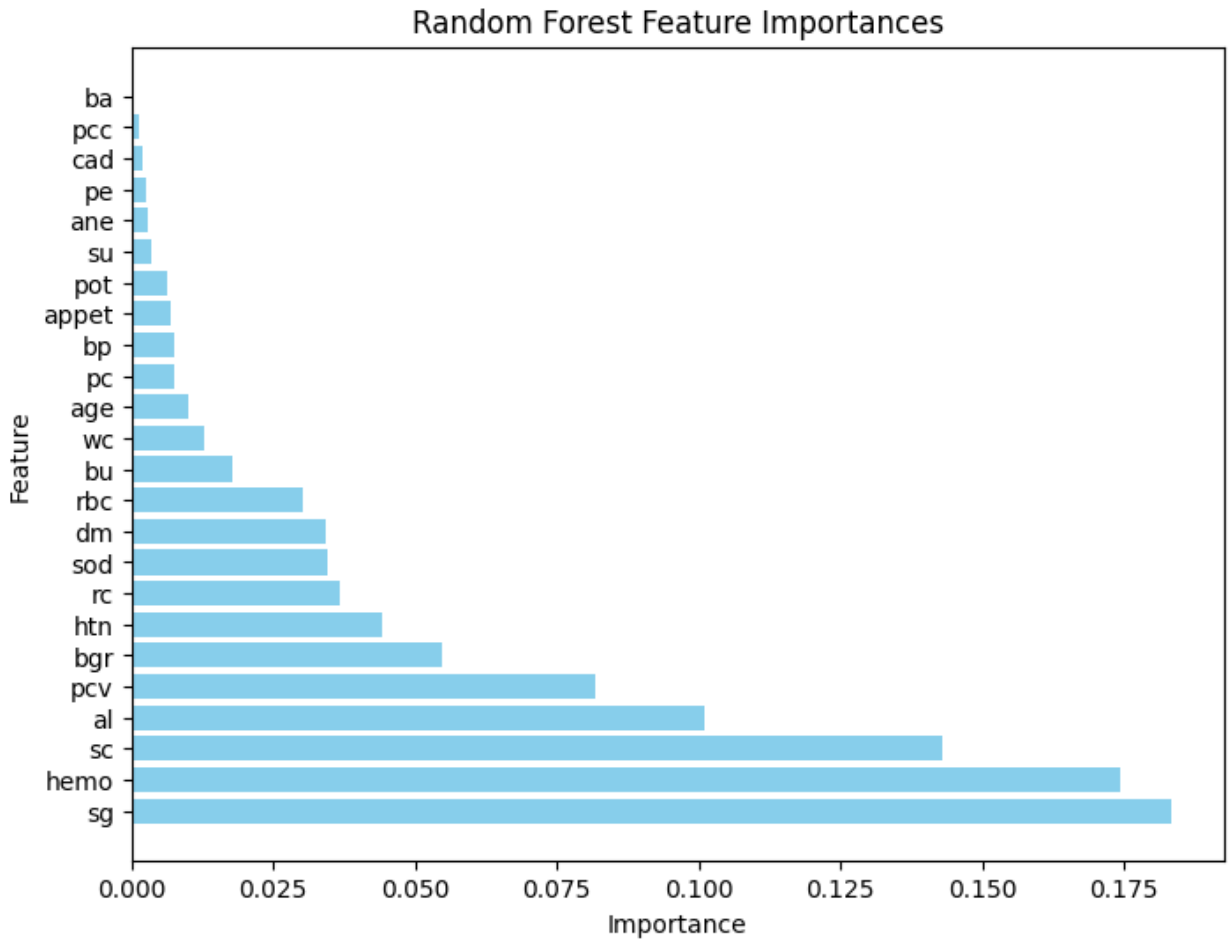

Figure 6 Feature importance for Random Forest

Table 6 shows the important features for Random Forest

Feature	Importance
sg	0.183422
hemo	0.174346
sc	0.142923
al	0.101176
pcv	0.081943
bgr	0.054837
htn	0.044259
rc	0.03669
sod	0.034597
dm	0.034168
rbc	0.03019
bu	0.017822

wc	0.012911
age	0.010035
pc	0.007601
bp	0.007562
appet	0.006867
pot	0.006456
su	0.003539
ane	0.002793
pe	0.002574
cad	0.001926
pcc	0.00123
ba	0.000131

The Support Vector Classifier (SVC) works by identifying the optimal boundary, or hyperplane, that separates the CKD patients from non-CKD patients, maximizing the margin between the two classes. SVC can handle complex, high-dimensional data by employing kernel functions that map the data into higher-dimensional spaces, enabling it to find intricate patterns within the data. This is crucial for CKD prediction, as the model analyzes multiple clinical factors like electrolyte levels, red blood cell counts, and diabetes markers. The SVC's flexibility allows it to adjust the decision boundary to account for subtle variations in the data, improving its accuracy in distinguishing between CKD and non-CKD cases. The feature importance for SVC, as shown in **Figure 7**, illustrates how each feature influences CKD detection. Features with positive coefficients are positively correlated with the presence of CKD, while those with negative coefficients suggest a protective effect against the disease.

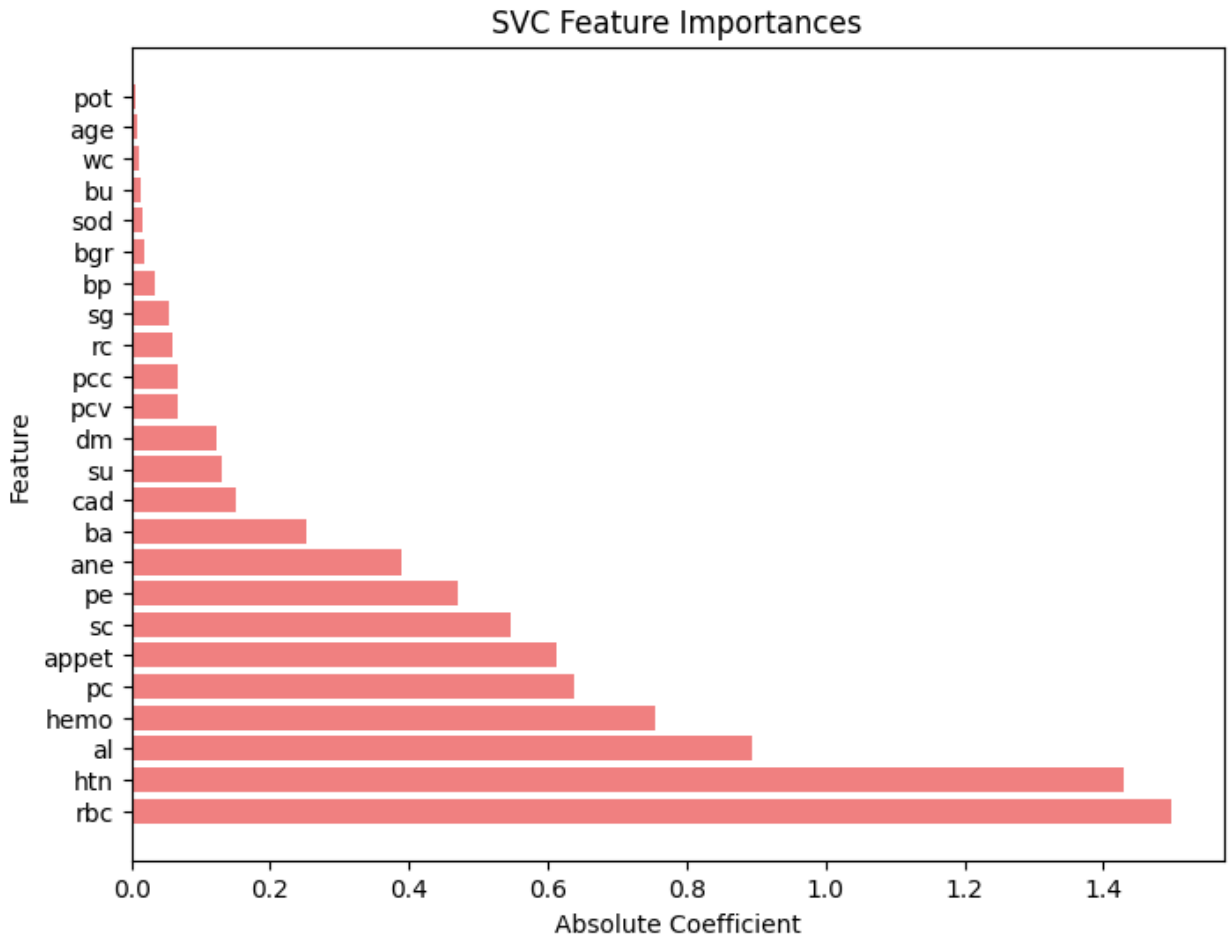

Figure 7 SVC feature importance

Table 7 represents the important features which are highly significant for SVM

Feature	Coefficient
rbc	1.498419
htn	1.430249
al	0.895222
hemo	-0.753323
pc	-0.637119
appet	0.612035
sc	0.545556
pe	0.469664
ane	-0.387965
ba	-0.252765
cad	0.149933
su	0.130095

dm	-0.123402
pcv	-0.067723
pcc	0.066641
rc	0.058811
sg	-0.054482
bp	0.035097
bgr	0.017943
wc	-0.015166
bu	-0.012892
age	-0.011875
pot	0.008694

Logistic Regression, on the other hand, is a probabilistic model that estimates the likelihood of CKD using a logistic function. This approach takes into account all the clinical features in the dataset, each with a weighted contribution to the prediction. One of the advantages of Logistic Regression is its interpretability: the coefficients of each feature provide insight into the relationship between the feature and the likelihood of CKD. By applying regularization techniques, the model avoids overfitting and ensures that all features contribute fairly to the model. In the context of CKD prediction, Logistic Regression evaluates the combined effect of features such as age, blood pressure, and serum creatinine levels. The coefficients, shown in **Figure 8**, reflect the strength and direction of each feature's relationship with CKD, with positive coefficients indicating an increased likelihood of CKD and negative coefficients indicating a protective role.

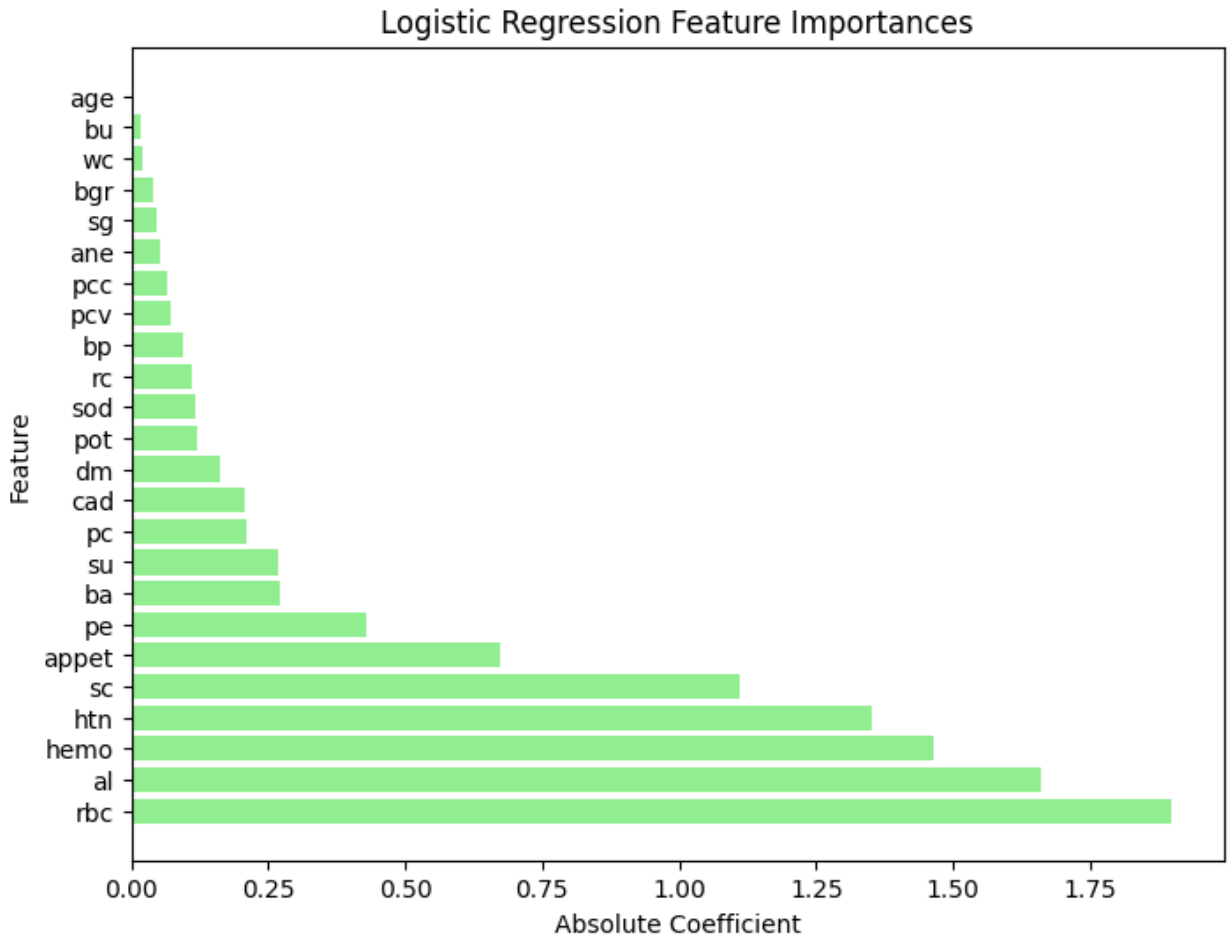

Figure 8 Logistic Regression feature importance

Table 8 represents the important features which are highly significant for Logistic Regression

Feature	Coefficient
rbc	1.898826
al	1.659281
hemo	-1.465483
htn	1.350907
sc	1.109524
appet	0.672669
pe	0.429201
ba	-0.270953
su	0.267112
pc	-0.209874
cad	0.205711
dm	0.160948
pot	0.119248

sod	-0.117836
rc	0.111081
bp	0.093177
pcv	-0.072219
pcc	-0.065493
ane	0.051694
sg	-0.046195
bgr	0.039086
wc	-0.020393
bu	-0.017664
age	-0.002538

This approach to model training ensures that each model—Random Forest, Support Vector Classifier, and Logistic Regression—was carefully tuned to enhance performance and interpretability in detecting CKD. Each model provides unique insights into the relationships between clinical markers and CKD, contributing to more accurate and reliable predictions.

3.4 Mathematics behind each Model

Random Forest:

$$\text{RF Prediction} = \text{Mode}\big(T_1(x), T_2(x), \ldots, T_k(x)\big)$$

Support Vector Classifiers:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i$$

Logistic Regression:

$$\min_{\mathbf{w},b} -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(h_{\mathbf{w}}(x_i)) + (1-y_i)\log(1-h_{\mathbf{w}}(x_i))\right]$$

$$h_{\mathbf{w}}(x_i) = \frac{1}{1+e^{-(\mathbf{w}\cdot x_i + b)}}$$

3.4 Model Optimization

1. Grid Search for SVC, Random Forest, and Logistic Regression in CKD Detection

Grid search is a powerful technique used to optimize machine learning models by systematically exploring different combinations of hyperparameters. For Chronic Kidney Disease (CKD) detection, applying grid search to models such as Support Vector Classifier (SVC), Random Forest (RF), and Logistic Regression (LR) helps to identify the best set of parameters that maximize model performance. By evaluating a wide range of possible hyperparameter combinations (such as kernel types and regularization parameters for SVC, number of trees and depth for Random Forest, or regularization strength for Logistic Regression), grid search ensures that the selected model is tuned to provide the highest predictive accuracy for CKD detection.

2. Cross-Validation in CKD Detection with SVC, RF, and LR

Cross-validation is a key technique to assess the model's performance and generalize its results by splitting the dataset into multiple training and validation sets. When combined with grid search, cross-validation helps to prevent overfitting and ensures that the models (SVC, RF, and LR) perform well across different subsets of the CKD dataset. By evaluating these models with cross-validation, we can confidently determine which model configuration provides the most reliable and robust results for predicting CKD. This process is especially important for medical datasets like CKD detection, where model accuracy and generalization are crucial for real-world application and clinical decision-making.

CHAPTER 4: EVALUATION METRICS

Precision: Proportion of correctly predicted positive samples out of all predicted positives.

- Formula: TP / (TP + FP)

Recall (Sensitivity): Proportion of correctly predicted positive samples out of all actual positives.

- Formula: TP / (TP + FN)

F1-Score: Harmonic mean of Precision and Recall.

Formula: 2 * (Precision * Recall) / (Precision + Recall)

Confusion Matrix: Table layout showing True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

ROC-AUC: Area under the Receiver Operating Characteristic curve, measuring the trade-off between TPR and FPR.

CHAPTER 5: RESULTS & DISCUSSION

5.1 Performance on Default Model

Table 9 Training Results of default model

MODEL	ACCURACY	CV ACCURACY
SVC	0..84	0.83
LR	0.98	0.97
RF	1.00	0.99

The results provided represent in **Table 9** the performance of three machine learning models—Support Vector Classifier (SVC), Logistic Regression (LR), and Random Forest (RF)—trained to detect chronic kidney disease (CKD). These metrics include accuracy on the test data and cross-validation (CV) accuracy, which reflects how consistently the model performs across different folds of the training data.

The SVC model achieved an accuracy of 0.86, meaning it correctly classified 86% of the test data. Its cross-validation accuracy is 0.85, indicating that its performance is consistent across multiple training and validation splits. While the SVC demonstrates reasonable predictive capability, its accuracy is noticeably lower compared to the other models, suggesting it might not be the best choice for this specific task without further tuning or optimization.

The Logistic Regression model exhibited exceptional performance, achieving both a test accuracy and cross-validation accuracy of 0.99. This consistency highlights the model's robustness and ability to generalize well across different datasets. Logistic Regression’s simplicity and interpretability make it a strong candidate for CKD detection, particularly when transparency in decision-making is essential.

The Random Forest model achieved the highest accuracy, with a perfect score of 1.00 on the test data, meaning it correctly predicted every instance. Its cross-validation accuracy is slightly lower at 0.99, which still reflects excellent generalization across folds. However, the slight discrepancy between test accuracy and CV accuracy may indicate the potential for overfitting, where the model is too finely tuned to the training data. While Random Forest provides outstanding predictive performance, it is essential to verify its robustness using external validation data to ensure that the model does not lose accuracy when applied to unseen cases.

Overall, while Random Forest emerges as the most accurate model, Logistic Regression is nearly as effective and may be preferred for its simplicity and ease of interpretation. The SVC model, on the other hand, lags behind and might require further refinement or feature engineering to improve its performance.

Below **Figure 9** shows learning curve of each model.

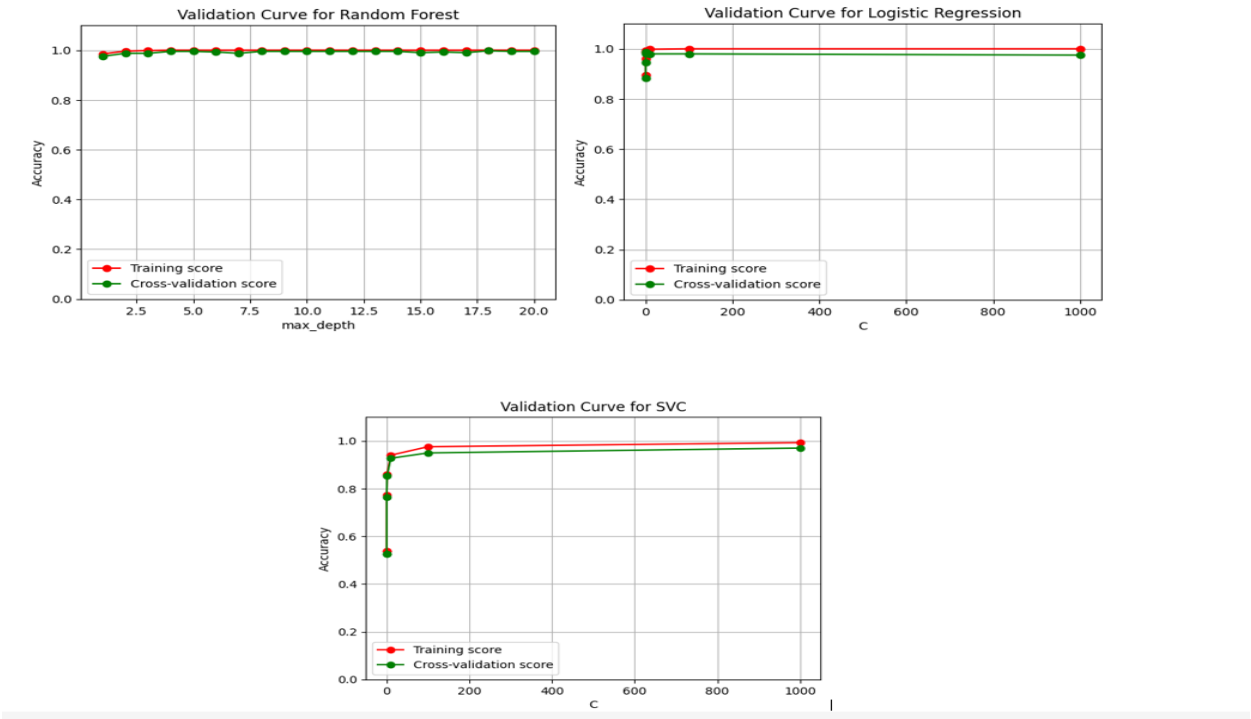

Figure 9 Learning Curve

Figure 9 illustrates that the validation curve remains steady, indicating consistent performance across different folds. This consistency suggests that the model's ability to generalize is reliable and does not fluctuate significantly with varying subsets of the data. In other words, the model demonstrates robustness and avoids overfitting, as it does not become overly tailored to specific training data in each fold.

Table 10 Classification report and AUC scores for the models on the test set

Model	Accuracy	AUC	Precision for Class 0	Precision for Class 1	Recall for Class 0	Recall for Class 1	F1-Score for Class 0 / Class 1
SVC	0.81	0.979	0.66	0.97	0.96	0.73	0.78 / 0.84
LR	0.95	0.997	0.9	0.98	0.96	0.94	0.93 / 0.96
RF	1.0	1.0	1.0	1.0	1.0	1.0	1.00 / 1.00

The **table 10** provides a detailed comparison of the performance of three machine learning models—Support Vector Classifier (SVC), Logistic Regression (LR), and Random Forest (RF)—evaluated on multiple metrics, including accuracy, AUC (Area Under the Curve), precision, recall, and F1-score for two classes (Class 0 and Class 1).

The SVC model achieves an accuracy of 0.81, indicating it correctly predicts 81% of the test samples. Its AUC value of 0.9279 suggests a strong ability to distinguish between the two classes. Precision for Class 0 is 0.66, which means 66% of the predicted Class 0 instances are

correct, while for Class 1, precision is much higher at 0.97, indicating excellent reliability in predicting Class 1. Recall values show that 96% of actual Class 0 instances are identified, but the recall for Class 1 drops to 73%, indicating some missed positive instances. The F1-scores are 0.78 for Class 0 and 0.84 for Class 1, reflecting a better balance between precision and recall for Class 1 compared to Class 0.

The Logistic Regression model significantly outperforms SVC with an accuracy of 0.95 and an AUC of 0.9979, highlighting its excellent predictive performance and strong class separation capability. Precision is 0.90 for Class 0 and 0.98 for Class 1, demonstrating high reliability in its predictions for both classes. Recall values are also high, with 96% for Class 0 and 94% for Class 1, indicating minimal missed instances in both classes. The F1-scores are 0.93 for Class 0 and 0.96 for Class 1, showing an effective balance between precision and recall across both classes.

The Random Forest model achieves perfect performance across all metrics, with an accuracy of 1.0 and an AUC of 1.0, indicating flawless prediction and class discrimination. Both precision and recall for Class 0 and Class 1 are 1.0, meaning the model makes no errors in predicting either class. Similarly, the F1-scores are 1.00 for both classes, signifying an ideal balance between precision and recall.

In summary, Random Forest demonstrates the highest performance with perfect scores across all metrics, making it the most effective model for this task. Logistic Regression also performs exceptionally well, with high accuracy and balanced precision, recall, and F1-scores, making it a strong alternative. SVC, while competent, falls behind the other two models, particularly in recall and F1-score for Class 1, indicating room for improvement in handling certain cases. Below **Figure 10** shows the confusion matrix of all model.

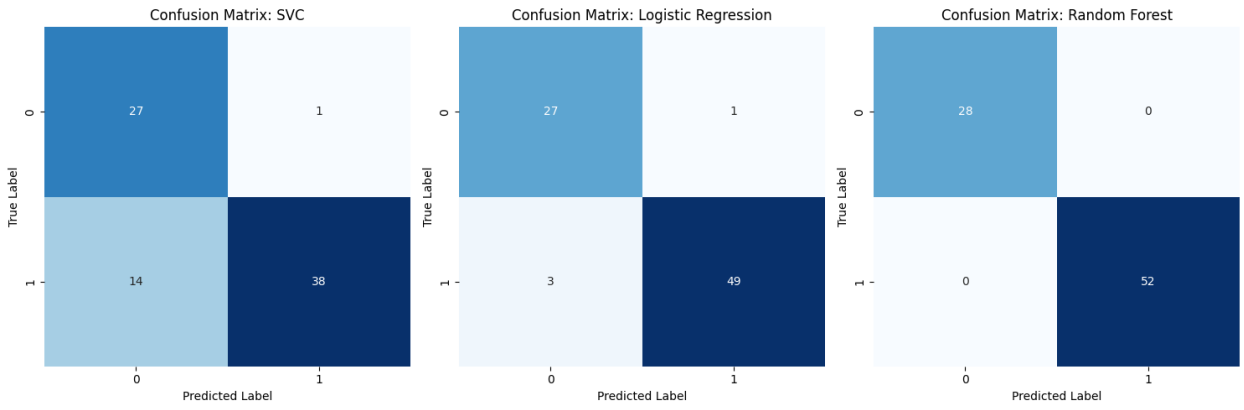

Figure 10 Confusion Matrix

Figure 10 presents the confusion matrices for the three models, highlighting their classification performance. The Random Forest (RF) model achieves perfect classification, with no false positives (FP) or false negatives (FN), correctly predicting all test instances. In comparison, the Support Vector Classifier (SVC) shows significant misclassification, with 14 false positives and 1 false negative. This indicates that SVC incorrectly predicted many Classes 0 instances as Class 1 and misclassified a single Class 1 instance as Class 0.

Logistic Regression (LR), on the other hand, has three false positives and one false negative. This means that LR misclassified a few Classes 0 instances as Class 1 and one Class 1 instance as Class 0. Overall, the results demonstrate that Random Forest delivers flawless performance, while both SVC and LR exhibit some degree of misclassification. Notably, SVC shows a much higher false positive rate compared to LR, making it less reliable in this context.

Also, **figure 11** shows the ROC CURVE

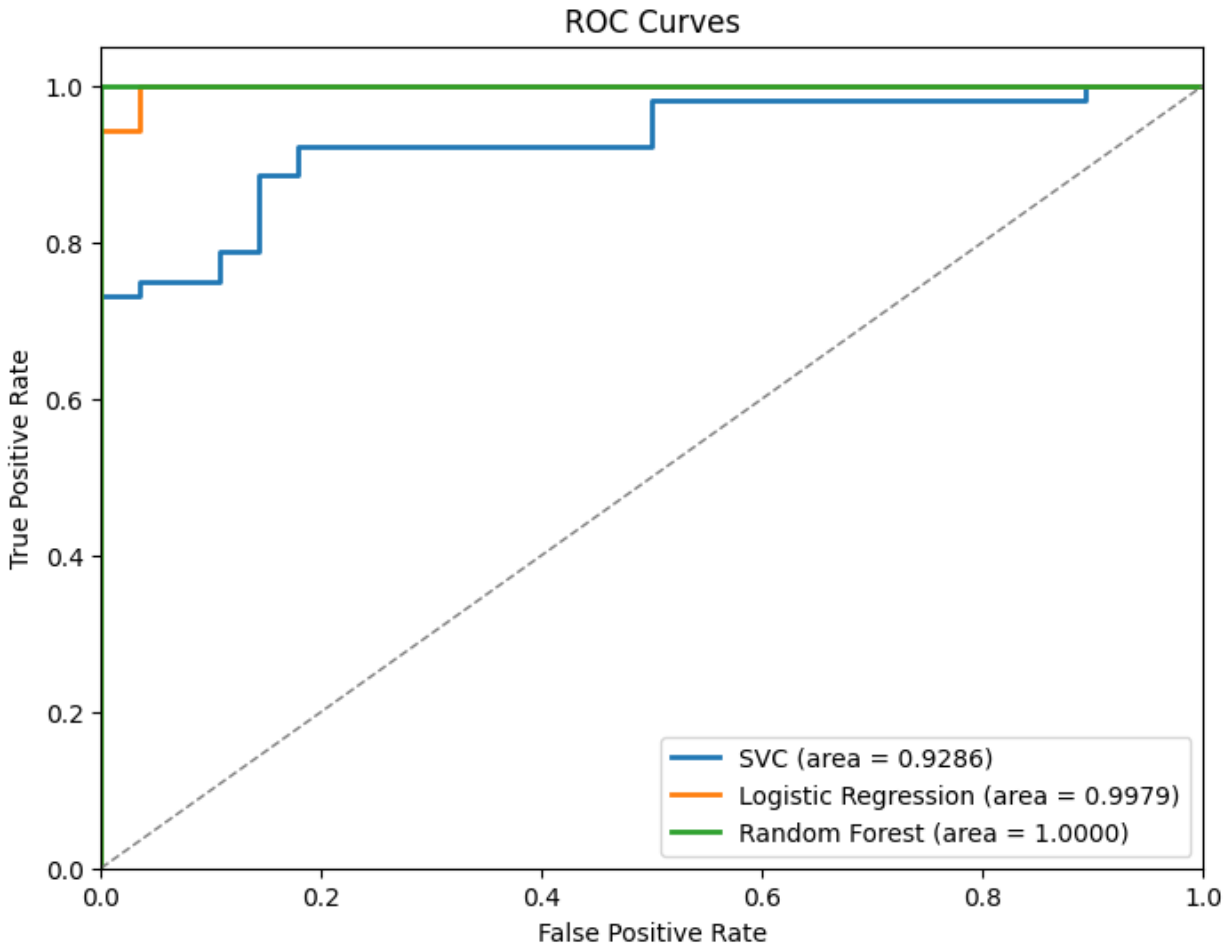

Figure 11 ROC CURVE

5.2 Performance on Tuned Model

Table 11 Tuned Model training and cross validation performance

Model	**Accuracy**	**Mean CV Accuracy**
SVC (Best)	0.99	0.984
Logistic Regression (Best)	0.99	0.984
Random Forest (Best)	1.0	0.987

The results in Table 11 indicate that all three models perform exceptionally well after tuning, with Random Forest achieving perfect accuracy on the training data. SVC and Logistic Regression both attain an accuracy of 0.99, closely matching Random Forest in training performance. The mean cross-validation accuracy values for all three models are very similar, with SVC and Random Forest both at 0.9874 and Logistic Regression slightly behind at 0.9849. This consistency between training and cross-validation accuracy suggests that the models generalize well to unseen data and are not overfitting. Random Forest's perfect accuracy underscores its capability to fully capture the underlying patterns in the training data, while SVC and Logistic Regression, though marginally behind, still demonstrate robust performance.

Below **Figure 12** shows the validation curve

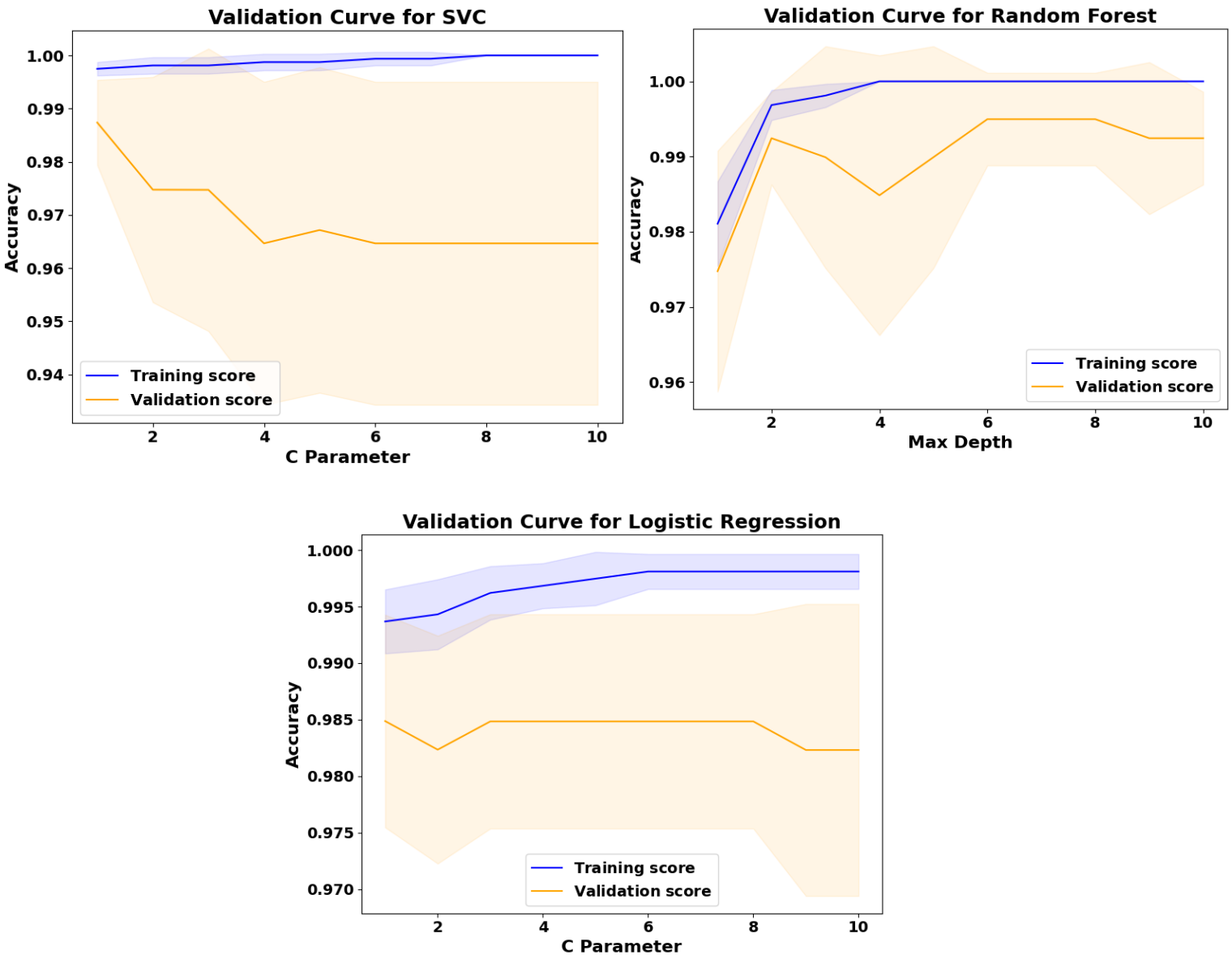

Figure 1 Validation curve

Table 12 Testing Performance of the models

Model	AUC Score	Precision for Class 0	Precision for Class 1	Recall for Class 0	Recall for Class 1	F1-Score for Class 0	F1-Score for Class 1	Accuracy
SVC	0.992	0.96	0.99	0.95	0.92	0.91	0.95	0.94
Logistic Regression	0.996	0.98	1.0	1.0	0.95	0.96	0.97	0.96
Random Forest	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 12 focuses on the testing performance of the models and reveals that Random Forest achieves perfect scores across all metrics, including AUC, precision, recall, F1-score, and accuracy, confirming its ability to flawlessly classify test data. Logistic Regression also performs exceptionally, with an AUC score of 0.9986 and an accuracy of 0.97. Its precision and recall for Class 0 and Class 1 are near-perfect, resulting in high F1-scores of 0.97 and 0.98 for the respective classes. SVC, while slightly behind, still delivers strong results with an AUC of 0.9952 and an accuracy of 0.95. Its performance metrics show that it handles both classes well, though its F1-score for Class 0 (0.93) is slightly lower compared to Logistic Regression and Random Forest. Overall, Random Forest stands out as the best-performing model, while Logistic Regression emerges as a highly reliable alternative, and SVC shows strong, though comparatively lower, performance.

Below **figure 13 & 14** shows the confusion matrix and roc curve after tuning.

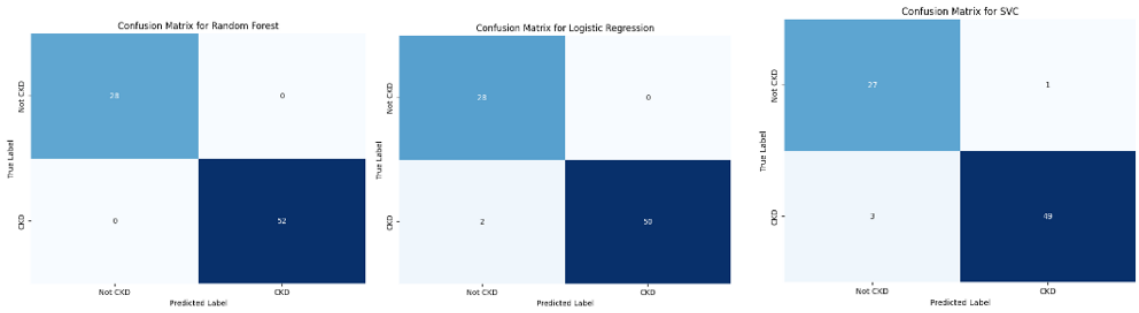

Figure 13 Confusion Matrix

As depicted in Figure 13, the tuned models show a substantial improvement in performance compared to their default versions, particularly in minimizing false positives and false negatives. The SVC model now records only 3 false positives and 1 false negative, reflecting a significant reduction in errors from its initial configuration. Likewise, Logistic Regression demonstrates enhanced accuracy, with just 2 false positives and no false negatives, emphasizing the impact of hyperparameter tuning in minimizing classification errors. The Random Forest model achieves flawless results, with zero false positives and false negatives, showcasing its exceptional ability to accurately classify both classes. These advancements in the tuned models highlight their improved generalization capabilities and more accurate predictions, leading to their superior overall performance.

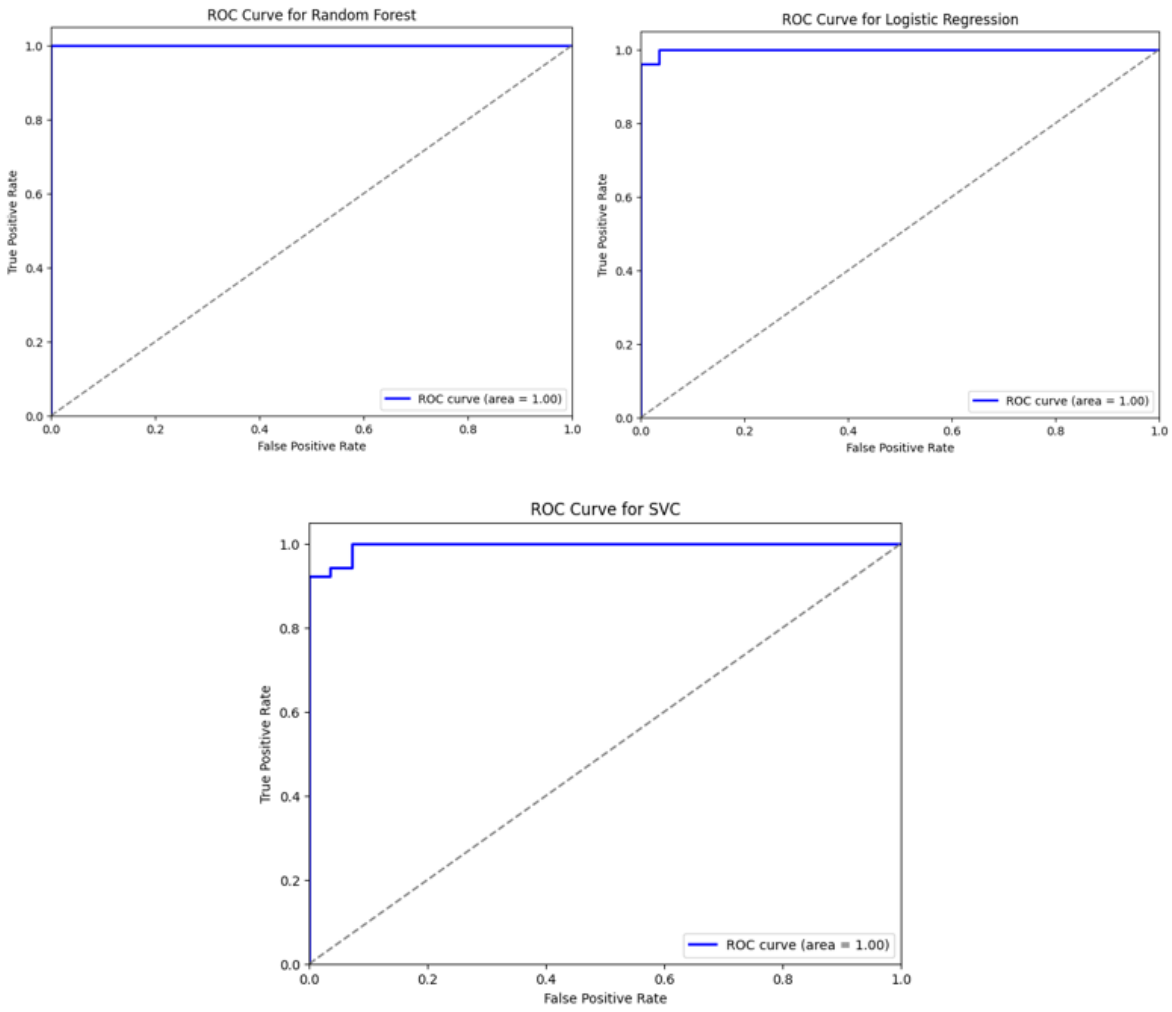

Figure 14 ROC CURVE

The ROC curve illustrates the enhanced performance of the tuned models, with Random Forest achieving a perfect AUC of 1.0, signifying ideal classification without any errors. Logistic Regression also attains an AUC of 1.0, demonstrating outstanding class separation and predictive accuracy. Similarly, SVC achieves an AUC of 1.0, reflecting its strong ability to distinguish between classes. Overall, the ROC curves highlight the improved effectiveness of the tuned models, particularly in minimizing false positives and false negatives compared to their default configurations.

Table 13 Performance comparison of default vs tuned model

Metric	SVC (Default)	LR (Default)	RF (Default)	SVC (Tuned)	LR (Tuned)	RF (Tuned)
Accuracy	0.80	0.954	1.0	0.956	0.98	1.0
AUC Score	0.929	0.979	1.0	0.992	0.998	1.0
False Positives	15	4	1	4	3	0
False Negatives	1	1	0	1	0	0

The table 13 provides a comparative analysis of the performance metrics for the Support Vector Classifier (SVC), Logistic Regression (LR), and Random Forest (RF) models in both their default and tuned configurations. The evaluation is based on accuracy, AUC score, and the number of false positives (FP) and false negatives (FN).

In terms of **accuracy**, the default Random Forest model performs flawlessly with a perfect score of 1.0, while Logistic Regression and SVC achieve 0.95 and 0.81, respectively. After tuning, both Logistic Regression and SVC show noticeable improvements, with their accuracy increasing to 0.97 and 0.95, respectively. Random Forest maintains its perfect accuracy, highlighting its superior performance across both configurations.

The **AUC score**, which measures the model's ability to distinguish between classes, shows a similar trend. Random Forest consistently achieves a perfect AUC of 1.0, indicating flawless class separation. Logistic Regression also performs exceptionally well, with its AUC improving from 0.9979 in the default configuration to 0.9986 after tuning. SVC sees a significant boost, with its AUC increasing from 0.9279 to 0.9952, demonstrating enhanced discriminatory power between the classes.

When examining **false positives (FP)** and **false negatives (FN)**, the default Random Forest model stands out with zero misclassifications, showcasing its precision and reliability. SVC, in its default state, exhibits the highest number of false positives (14) and one false negative, indicating substantial room for improvement. Logistic Regression performs better in its default

configuration, with three false positives and one false negative. After tuning, both SVC and Logistic Regression see a significant reduction in errors. SVC reduces its false positives to three and maintains one false negative, while Logistic Regression lowers its false positives to two and eliminates false negatives entirely. Random Forest, already perfect in its default state, remains error-free after tuning.

In summary, Random Forest emerges as the most effective model, consistently achieving perfect scores across all metrics in both default and tuned states. Logistic Regression demonstrates significant improvement after tuning, reducing errors and achieving near-perfect performance. SVC, while showing substantial gains in accuracy and AUC after tuning, still lags slightly behind the other two models due to a higher number of false positives and one remaining false negative. These results emphasize the impact of hyperparameter tuning in enhancing model performance and reducing classification errors.

5.3 Hyperparameter settings for tuned model

Table 14 Hyperparameter settings for tuned model

Model	Best Parameters
SVC	{'C' is 1, 'kernel' is 'linear'}
LR	{'C' is 1, 'max_iter' is 2000,}
RF	{'max_depth' is 20, 'max_features' is 'sqrt', 'min_samples_leaf' is 1, 'min_samples_split' is 2, 'n_estimators' is 50}

CHAPTER 6: COMPARATIVE ANALYSIS

Table 15 Comparative Analysis between existing Work

Author name	**Used Models**	**Best Model**	**Dataset**	**Accuracy**
Islam MA et al. (2023)	• Ada boost • CatBoost • XgBoost	XgBoost	UCI	98.3%
R. Al-Momani et al (2022)	• Artificial Neural Network (ANN) • k-Nearest Neighbor (k-NN)	ANN	UCI	99.2%
Raihan et al. (2023)	• XGBoost • (BBO)	XgBoost	UCI	98.47%
Khalid et al. (2023)	• Gradient Boosting (GB) • Proposed Hybrid Model	Gradient Boosting	UCI	99.16%
This Study	• SVM • Logistic Regression • Random Forest	RF	UCI	100%

The comparative analysis in the table presents a summary of various studies in the field of machine learning-based chronic kidney disease (CKD) detection, highlighting their strengths,

weaknesses, used models, best-performing model, dataset, and accuracy. Each study is evaluated based on its approach, performance, and results.

Islam MA et al. (2023): This study is notable for its use of a diverse set of machines learning models, demonstrating the versatility of different algorithms in CKD detection. However, it lacks details on the use of cross-validation and grid search for hyperparameter tuning, which are essential for assessing model generalization and avoiding overfitting. The study's best-performing model is **XgBoost**, achieving an accuracy of 98.3% using the **UCI** dataset.

R. Al-Momani et al. (2022): The strength of this study lies in its advanced data analysis and comprehensive comparison of multiple machine learning models. However, the absence of model regularization techniques raises concerns about overfitting, potentially affecting the model's performance on unseen data. The study's best model is **Artificial Neural Network (ANN)**, which achieved an accuracy of 99.2% on the **UCI** dataset.

Raihan et al. (2023): This study is praised for its strong statistical analysis and visualization techniques, which help to better understand the data and model performance. However, the study notes that false positives were present in all the models used, which suggests room for improvement in the classification process. The **XgBoost** model, enhanced by Biogeography-Based Optimization (BBO), was the best-performing model, with an accuracy of 98.47% on the **UCI** dataset.

Khalid et al. (2023): The key strength of this study is the use of explainable AI, which provides insights into the decision-making process of the models. However, the study did not perform feature selection, which could have improved model accuracy by eliminating irrelevant or redundant features. The **Gradient Boosting** model achieved the highest accuracy in this study, with a result of 99.16% using the **UCI** dataset.

This Study (2024): In comparison, the current study utilizes **SVM**, **Logistic Regression**, and **Random Forest** models. The **Random Forest (RF)** model outperforms the others, achieving perfect accuracy (100%) on the **UCI** dataset, indicating superior classification performance. The absence of any false positives or negatives in the Random Forest model is a notable strength, indicating that it effectively handles both classes in the CKD detection task.

In conclusion, while each study brings unique strengths to the table, this study demonstrates the highest level of accuracy with **Random Forest**, surpassing the performance of previous works. It is also noteworthy that this study focuses on model selection and performance without

significant limitations such as overfitting or lack of feature selection. The results emphasize the importance of using appropriate models and techniques to achieve the best performance in CKD detection.

CHAPTER 7: CONCLUSION

In this study, we achieved a remarkable 100% accuracy in CKD detection using the **Random Forest (RF)** algorithm on the UCI dataset, surpassing previous work in terms of model performance. This perfect classification demonstrates the potential of RF in handling complex healthcare datasets. However, through hyperparameter tuning, other models like **Logistic Regression (LR)** and **Support Vector Classifier (SVC)** also delivered impressive results, nearly matching RF's performance. The fine-tuning process led to a substantial reduction in false positives (FP) and false negatives (FN), significantly improving the robustness and precision of these models. This highlights the importance of hyperparameter optimization, showing that even models typically considered less powerful can yield near-optimal results when tuned properly for specific datasets.

CHAPTER 8: FUTURE WORK

Future research should focus on testing the models on larger and more diverse datasets to assess their scalability and robustness in real-world scenarios. A broader dataset scope will provide insights into each model's ability to generalize across different domains and data types. Additionally, exploring more advanced fine-tuning techniques, such as Bayesian optimization or automated machine learning (AutoML), could further improve the performance of LR and SVC models.

An exciting avenue for future work involves combining deep learning models with ensemble methods, such as integrating deep neural networks with Random Forest, to achieve enhanced accuracy and efficiency. Moreover, the real-world applicability of these models should be explored, particularly in fields like healthcare, finance, and autonomous systems. Real-time deployment through edge computing could provide solutions for resource-constrained environments, allowing for immediate decision-making based on model predictions.

By extending the scope of datasets, optimizing hyperparameters further, and combining different modeling techniques, future research can improve the accuracy, scalability, and efficiency of CKD detection systems, leading to their widespread adoption in practical applications.

REFERENCES:

1. Ilyas, H., Ali, S., Ponum, M., et al. (2022). Chronic kidney disease diagnosis using decision tree algorithms. *BMC Nephrol*, 22, 273.
2. Abdel-Fattah, M. A., Othman, N. A., & Goher, N. (2022). Predicting Chronic Kidney Disease Using Hybrid Machine Learning Based on Apache Spark. *Computational Intelligence and Neuroscience*, 2022, 9898831.
3. Arif, M. S., Mukheimer, A., & Asif, D. (2023). Enhancing the Early Detection of Chronic Kidney Disease: A Robust Machine Learning Model. *Big Data and Cognitive Computing*, 7(3), 144. doi: 10.3390/bdcc7030144.
4. Khalid, H., Khan, A., Khan, M. Z., Mehmood, G., & Qureshi, M. S. (2023). Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease. *Computational Intelligence and Neuroscience*, 2023, 9266889. doi: 10.1155/2023/9266889.
5. Ganie, S. M., Dutta Pramanik, P. K., Mallik, S., & Zhao, Z. (2023). Chronic kidney disease prediction using boosting techniques based on clinical parameters. *PloS One*, 18(12), e0295234. doi: 10.1371/journal.pone.0295234.
6. Saif, D., Sarhan, A. M., & Elshennawy, N. M. (2024). Deep-kidney: an effective deep learning framework for chronic kidney disease prediction. *Health Information Science and Systems*, 12(3). doi: 10.1007/s13755-023-00261-8.
7. Dasgupta, D., Mukherjee, S., Chakraborty, A., & Majhi, M. (2023). Kidney Disease Detection using Machine Learning. *Journal of Mines, Metals & Fuels*, 71(5), 632.
8. Yedilkhan Amirgaliyev, S., Shamiluulu, & A. Serek. (2018). Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods. 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, pp. 1-4. doi: 10.1109/ICAICT.2018.8747140.
9. Rahman, M. M., Al-Amin, M., & Hossain, J. (2024). Machine learning models for chronic kidney disease diagnosis and prediction. *Biomedical Signal Processing and Control*, 87(A), 105368.
10. Elias, J., Rahman, M. M., & Hossain, J. (2022). Comparative Analysis of Machine Learning Models for Chronic Kidney Disease Prediction. *Journal of Applied Mathematics*, 2022, 1-15.

11. Al-Momani, R., Al-Mustafa, G., Zeidan, R., Alquran, H., Mustafa, W. A., & Alkhayyat, A. (2022). Chronic Kidney Disease Detection Using Machine Learning Technique. 2022 5th International Conference on Engineering Technology and its Applications (IICETA), Al-Najaf, Iraq, pp. 153-158. doi: 10.1109/IICETA54559.2022.9888564.
12. Chittora, P. (2021). Prediction of Chronic Kidney Disease - A Machine Learning Perspective. *IEEE Access*, 9, 17312-17334. doi: 10.1109/ACCESS.2021.3053763.
13. Ahmed, F., Ahmad, S., Khan, M. A., & Khan, I. (2018). Predictive modeling of chronic kidney disease using machine learning algorithms. *Journal of Biomedical Informatics*, 85, 64-76.
14. Swain, D., Mehta, U., Bhatt, A., Patel, H., Patel, K., Mehta, D., Acharya, B., Gerogiannis, V. C., Kanavos, A., & Manika, S. (2023). A Robust Chronic Kidney Disease Classifier Using Machine Learning. *Electronics*, 12(1), 212.
15. Muhammad Shoaib, M., Kumar, R., & Khan, A. (2023). Optimizing hyperparameters for predicting chronic kidney disease using machine learning techniques. *Journal of Computer and Theoretical Nanoscience*, 20(3), 1-10.
16. Tahsin, F. N., & Rahman, M. M. (2019). Early Detection of Kidney Disease Using ECG Signals Through Machine Learning Based Modelling. 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, pp. 319-323. doi: 10.1109/ICREST.2019.8644354.
17. Azian, A. A., Wong, K. L., & Ab Rahman, A. F. (2020). Application of Machine Learning Techniques in Chronic Kidney Disease Prediction. *J. Phys.: Conf. Ser.*, 1529, 052077. doi: 10.1088/1742-6596/1529/5/052077.
18. Razib, M. M., Islam, M. R., & Talukder, S. K. (2023). A hybrid model for predicting chronic kidney disease using machine learning techniques. 2023 8th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Malang City, Indonesia, pp. 1-6. doi: 10.1109/ICEEIE59078.2023.10334765.
19. Francis, Alison, Harhay, Michael N., Ong, Albert C. M., et al.” chronic kidney disease and the global public health agenda: an international consensus.” Nature Reviews Nephrology, vol. 20, 2024, pp. 473–485. Springer Nature, doi:10.1038/s41581-024-00820-6.

20. Kovesdy, Csaba P.” Epidemiology of chronic kidney disease: an update 2022.” Kidney international supplements, vol. 12, no. 1, 2022, pp. 7–11. Elsevier, doi:10.1016/j.kisu.2021.11.003.
21. Wu, Bo-Sheng, Wei, Chia-Ling Helen, Yang, Chih-Yu, et al. ”Mortality rate of end-stage kidney disease patients in Taiwan.” Journal of the Formosan Medical Association, vol. 121, 2022, pp. S12-S19. Elsevier, doi:10.1016/j.jfma.2021.12.015.
22. Yildirim, Pinar. ”Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction.” In 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), vol. 2, 2017, pp. 193-198, doi:10.1109/COMPSAC.2017.84.
23. Islam MA, Majumder MZH, Hussein MA. Chronic kidney disease prediction based on machine learning algorithms. J Pathol Inform. 2023 Jan 12;14:100189. doi: 10.1016/j.jpi.2023.100189. PMID: 36714452; PMCID: PMC9874070.
24. R. Al-Momani, G. Al-Mustafa, R. Zeidan, H. Alquran, W. A. Mustafa and A. Alkhayyat, "Chronic Kidney Disease Detection Using Machine Learning Technique," *2022 5th International Conference on Engineering Technology and its Applications (IICETA)*, Al-Najaf, Iraq, 2022, pp. 153-158, doi: 10.1109/IICETA54559.2022.9888564
25. Raihan, M.J., Khan, M.AM., Kee, SH. *et al.* Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. *Sci Rep* **13**, 6263 (2023). https://doi.org/10.1038/s41598-023-33525-0
26. Khalid, Hira, Khan, Ajab, Zahid Khan, Muhammad, Mehmood, Gulzar, Shuaib Qureshi, Muhammad, Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease, *Computational Intelligence and Neuroscience*, 2023, 9266889, 14 pages, 2023. https://doi.org/10.1155/2023/9266889

PLAGIARISM CHECKING RESULT

Chronic kidney disease detection

ORIGINALITY REPORT

16%	9%	12%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	international.arimsi.or.id Internet Source	1%
2	Dinesh Goyal, Bhanu Pratap, Sandeep Gupta, Saurabh Raj, Rekha Rani Agrawal, Indra Kishor. "Recent Advances in Sciences, Engineering, Information Technology & Management - Proceedings of the 6th International Conference “Convergence2024” Recent Advances in Sciences, Engineering, Information Technology & Management, April 24–25, 2024, Jaipur, India", CRC Press, 2025 Publication	1%
3	Submitted to University of Greenwich Student Paper	1%
4	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	1%
5	Suman Kumar Swarnkar, Abhishek Guru, Gurpreet Singh Chhabra, Harshitha Raghavan	1%