**MACHINE LEARNING BASED CHRONIC KIDNEY DISEASE PREDICTION WITH SMART WEB APPLICATION**

*Report submitted to the SASTRA Deemed to be University*

*as the requirement for the course*

**CSE425 - MACHINE LEARNING ESSENTIALS PROJECT**

*Submitted by*

**SANJAI S**

**(126018042, B.TECH COMPUTER SCIENCE AND BUSINESS SYSTEMS)**

**November 2025**



**SCHOOL OF COMPUTING**

**THANJAVUR, TAMIL NADU, INDIA - 613 401**



**SCHOOL OF COMPUTING**

**THANJAVUR – 613 401**

# **Bonafide Certificate**

This is to certify that the report titled “**Machine Learning Based Chronic Kidney Disease Prediction with Smart Web Application**” submitted as a requirement for the course, **CSE425 : MACHINE LEARNING ESSENTIALS PROJECT** for B. Tech is a bonafide record of the work done by **Mr. Sanjai S (Reg. No. 126018042, B. Tech Computer Science and Business Systems),** during the academic year 2025-2026, in the School of Computing, under my supervision.



**Signature of Project Supervisor :**

**Name with Affiliation : Dr. Vidivelli S**, Asst. Professor-II, SoC

**Date :**   **07.11.2025**

Mini project *Viva voce* held on \_\_\_10.11.2025\_\_

**Examiner 1**  **Examiner 2**

**Acknowledgements**

We would like to thank our Honorable Chancellor **Prof. S. Sethuraman** for providing us with an opportunity and the necessary infrastructure for carrying out this project as a part of our curriculum.

We would like to thank our Honorable Vice-Chancellor **Dr. S.Vaidhyasubramaniam** and **Dr. S. Swaminathan**, Dean, Planning & Development, for the encouragement and strategic support at every step of our college life.

We extend our sincere thanks to **Dr. R. Chandramouli**, Registrar, SASTRA Deemed to be University for providing the opportunity to pursue this project.

We extend our heartfelt thanks to **Dr. V. S. Shankar Sriram**, Dean, School of Computing, **Dr. R. Muthaiah**, Associate Dean, Research, **Dr. K.Ramkumar**, Associate Dean, Academics, **Dr. D. Manivannan**, Associate Dean, Infrastructure, **Dr. R. Alageswaran**, Associate Dean, Students Welfare.

Our guide **Dr. Vidivelli S,** AP-ll School of Computing was the driving force behind this whole idea from the start. His deep insight in the field and invaluable suggestions helped us in making progress throughout our project work.

We also thank the project review panel members for their valuable comments and insights which made this project better. We would like to extend our gratitude to all the teaching and non-teaching faculties of the School of Computing who have either directly or indirectly helped us in the completion of the project.

We gratefully acknowledge all the contributions and encouragement from our family and friends resulting in the successful completion of this project. We thank you all for providing us an opportunity to showcase our skills through the project

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No.** |
| 1.1 | Data Preprocessing and Feature Selection |  |
| 1.2 | Input Webpage |  |
| 1.3 | Result Webpage |  |
| 3.1 | Preprocessing |  |
| 3.2 | Feature Selection |  |
| 3.3 | Classifier Selection |  |
| 3.4 | K-Fold Validation |  |
| 3.5 | Classification with Train Test Split |  |
| 4.1 | HeatMap Analysis |  |
| 4.2 | Accuracy of different classifiers for the split K=10 |  |
| 4.3 | Accuracy of different classifiers for the split K=20 |  |
| 4.4 | Accuracy of different classifiers for the split K=30 |  |
| 4.5 | AUC Scores of different classifiers for the split K=10 |  |
| 4.6 | AUC Scores of different classifiers for the split K=20 |  |
| 4.7 | AUC Scores of different classifiers for the split K=30 |  |
| 4.8 | Confusion matrix for the best two classifiers (RF,AdaB) performance in the Dataset\_1 |  |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Table Name** | **Page No.** |
| 1.1 | Selected features by different feature selection algorithm |  |
| 1.2 | Classifiers error rates |  |

**Abbreviations**

ML Machine Learning

DL Deep Learning

CKD Chronic Kidney Disease

UCI University of California

RFE Recursive Feature Elimination

SFS Sequential Forward Selection

KNN K-Nearest Neighbours

RF Random Forest

GB Gradient Boosting

XgB Extreme Gradient Boosting

AdaB Adaptive Boosting

LR Logistic Regression

DT Decision Tree

SVM Support Vector Machines

ANN Artificial Neural Network

RNN Recurrent Neural Network

CNN Convolutional Neural Network

LSTM Long Short-Term Memory

NLP Natural Language Processing

GFR Glomerular Filtration Rate

DTW Dynamic Time Warping

HMM Hidden Markov Model

**Abstract**

Chronic kidney disease (CKD) remains a critical global health concern, often leading to severe complications such as hypertension, anaemia, and renal failure if undiagnosed. Early detection and precise classification are essential for improving patient outcomes. This project focuses on implementing an existing machine learning-based CKD prediction model, which integrates advanced data preprocessing techniques, robust feature selection, and multiple classification algorithms to enhance diagnostic accuracy. The model employs standard preprocessing methods, including categorical-to-numeric transformations, missing value imputation, and data normalization to ensure consistency. Feature selection techniques such as Correlation Analysis, Chi-Square, and Recursive Feature Elimina on are applied to refine predictive attributes. The classification phase involves evaluating multiple machine learning models—including Random Forest, Gradient Boosting, AdaBoost, XGBoost, Support Vector Machine, Decision Tree and Logistic Regression—to determine the most effective approach for CKD prediction. As an extension of this work, we aim to adapt the model to a different dataset, tackling an unsolved problem in the medical domain. Additionally, a web-based application is developed to operationalize the model, enabling real-me disease prediction and enhancing accessibility for healthcare practitioners. By validating and extending the existing model, this research aspires to contribute to broader applications of machine learning in healthcare diagnostics.

**KEY WORDS**: Chronic kidney disease, Machine Learning, Prediction, Early Detection, Classification models, Web Application, Binary Classification, AI in Medicine, Predictive Modeling, Real Time Prediction.

**Table of Contents**

**Title Page No.**

Bonafide Certificate i

Acknowledgements ii

List of Figures iii

List of Tables iv

Abbreviations v

Abstract vi

1. Summary of the base paper 1

2. Merits and Demerits of the base paper 6

3. Source Code 9

4. Snapshots 18

5. Conclusion and Future Plans 23

6. References 24

7. Appendix -Base Paper 26

**CHAPTER-1**

**SUMMARY OF THE BASE PAPER**

**Title :** ML-CKDP: Machine Learning-Based Chronic Kidney Disease Prediction with Smart Web Application

**Journal Name :** Journal of Information Pathology

**Publisher :** Elsevier – ScienceDirect : 2024

**Year :** 2024

**Indexed in :** Scopus

Chronic Kidney Disease (CKD) is a silent and potentially life-threatening disease, frequently not detected until advanced stages. As a response to the increasing CKD burden, this project offers an end-to-end machine learning pipeline for early prediction of CKD along with a user-friendly real-time web application intended for clinical practice ease of use. The impetus for this work arose from the necessity of closing the gap between highly accurate prediction models and their usability in real-world settings.

We started by undertaking a wide survey of the literature that brought into focus significant drawbacks in current CKD prediction models—mostly missing strong feature selection, using limited datasets, bad management of missing values, and restricted usage of working tools in real-time settings. Taking a lesson from all these loopholes, we handpicked and preprocessed a 400-instance CKD dataset of the UCI repository by performing imputation of missing values, label encoding on categorical features, and min-max normalization. These features encompass critical medical indicators such as:

* **Numerical features**: Age, Blood Pressure (BP), Blood Glucose Random (BGR), Blood Urea (BU), Serum Creatinine (SC), Sodium (SOD), Potassium (POT), Hemoglobin (HEMO), Packed Cell Volume (PCV), and Red Blood Cell Count (RC).
* **Categorical features**: Red Blood Cells (RBC), Pus Cell (PC), Pus Cell Clumps (PCC), Bacteria (BA), Hypertension (HTN), Diabetes Mellitus (DM), Coronary Artery Disease (CAD), Appetite, Pedal Edema (PE), Anemia (ANE), etc.

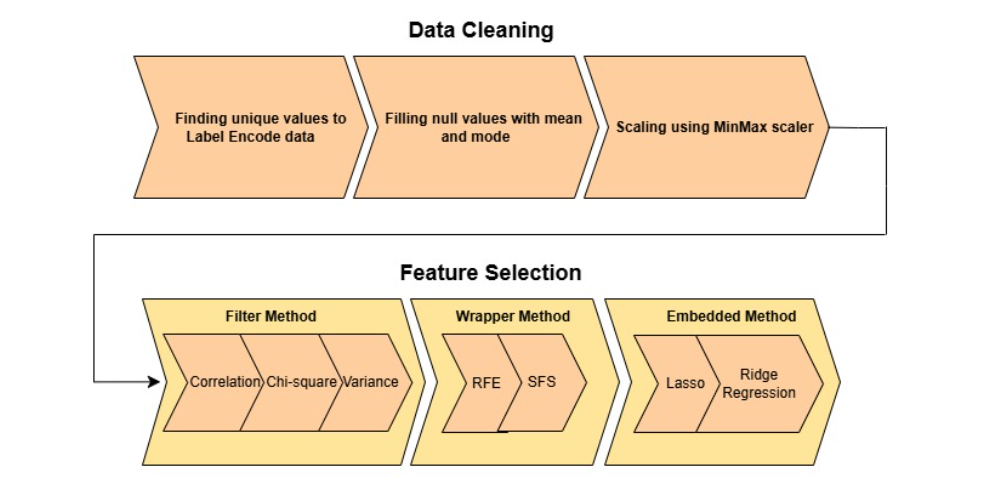


Fig 1.1 Data Preprocessing and Feature Selection

FEATURE SELECTION::

To ensure optimal feature selection, we implemented **seven techniques**, including:

* Correlation Analysis
* Chi-Square
* Variance Threshold
* Recursive Feature Elimination (RFE)
* Sequential Forward Selection (SFS)
* Lasso Regression
* Ridge Regression

These methods were applied to form seven uniquely refined datasets, each offering a different perspective on which attributes contributed most to CKD classification.

|  |  |
| --- | --- |
| SELECTION TECHNIQUES | SELECTED FEATURES |
| CORRELATION | AGE, BP, SG, AL, SU, BGR, BU, SC, SOD, POT, HEMO, WC, RBC, PC, PCC, BA, HTN, DM, CAD, APPET, PE, ANE |
| CHI-SQUARE | SG, AL, SU, BGR, BU, SC, HEMO, PCV, RC, PC, PCC, BA, HTN, DM, CAD, APPET, PE, ANE |
| VARIANCE THRESHOLD | RBC, PC, HTN, DM, APPET, PE, ANE |
| RECURSIVE FEATURE ELIMINATION | SG, AL, BGR, BU, SC, SOD, HEMO, PCV, RC, HTN, DM |
| SEQUENTIAL FORWARD SELECTION | BP, SG, AL, BGR, BU, SC, HEMO, HTN, DM, APPET, ANE |
| LASSO REGRESSION | HTN, DM |
| RIDGE REGRESSION | BP, AL, BGR, SC, POT, WC, PC, HTN, DM, APPET, PE |

Table 1.1 Selected features by different feature selection algorithm

CLASSIFIERS:

We then trained and evaluated **seven classifiers**:

* Random Forest (RF)
* AdaBoost (AdaB)
* Gradient Boosting (GB)
* XGBoost (XgB)
* Support Vector Machine (SVM)
* Decision Tree (DT)
* Naive Bayes (NB)

In terms of classification, **nine machine learning models** were evaluated using multiple performance metrics including **accuracy, AUC, confusion matrix, and error rates**. According to the error rate table, **Random Forest (RF)** and **AdaBoost (AdaB)** consistently achieved **very low error rates across all datasets**, often as low as **0.01**, highlighting their robustness and reliability in CKD prediction. Models like **Naive Bayes (NB)** and **K-Nearest Neighbors (KNN)** exhibited **comparatively higher error rates**, up to **0.375**, indicating their relatively lower suitability for medical diagnostics in this context refer Table 1.2

WEB INTERFACE:

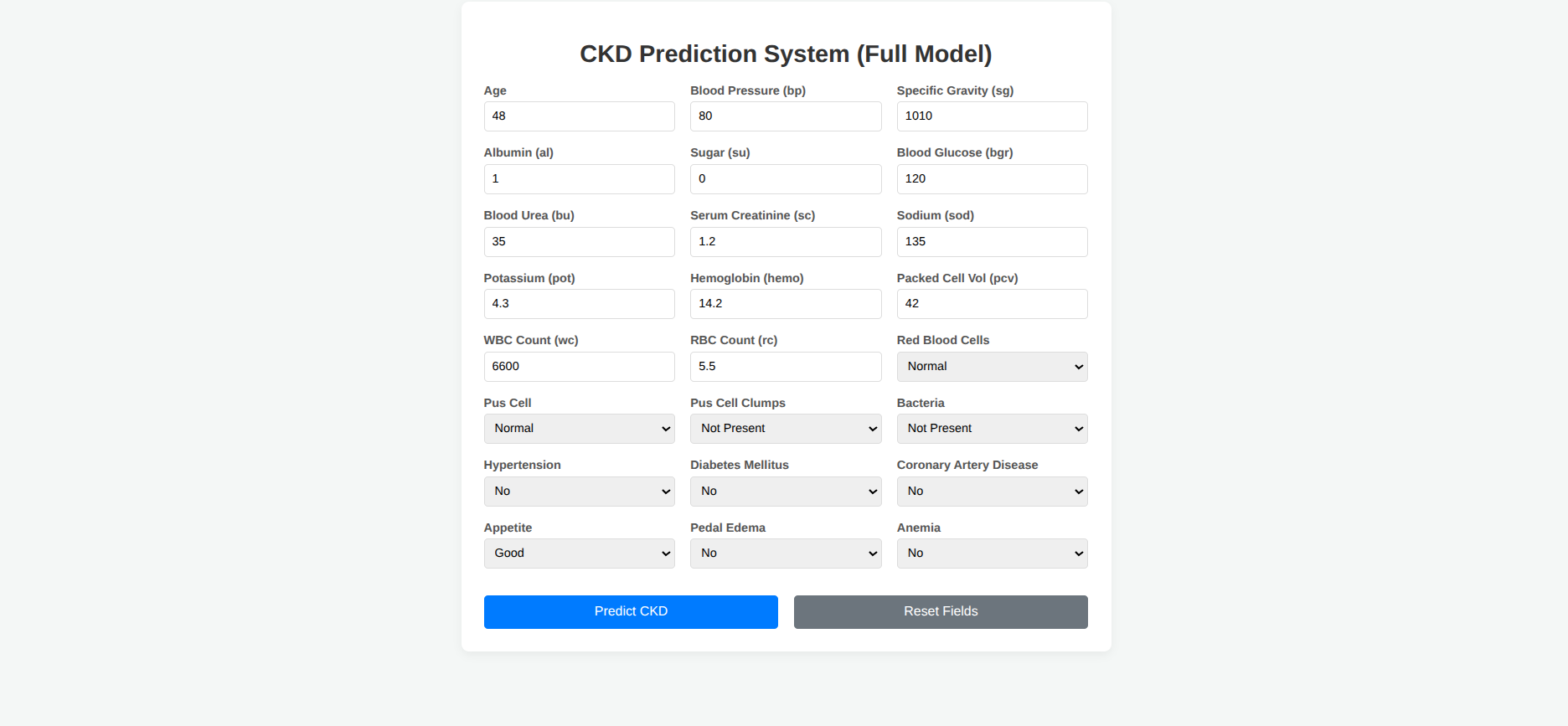


Fig 1.2 Input Webpage

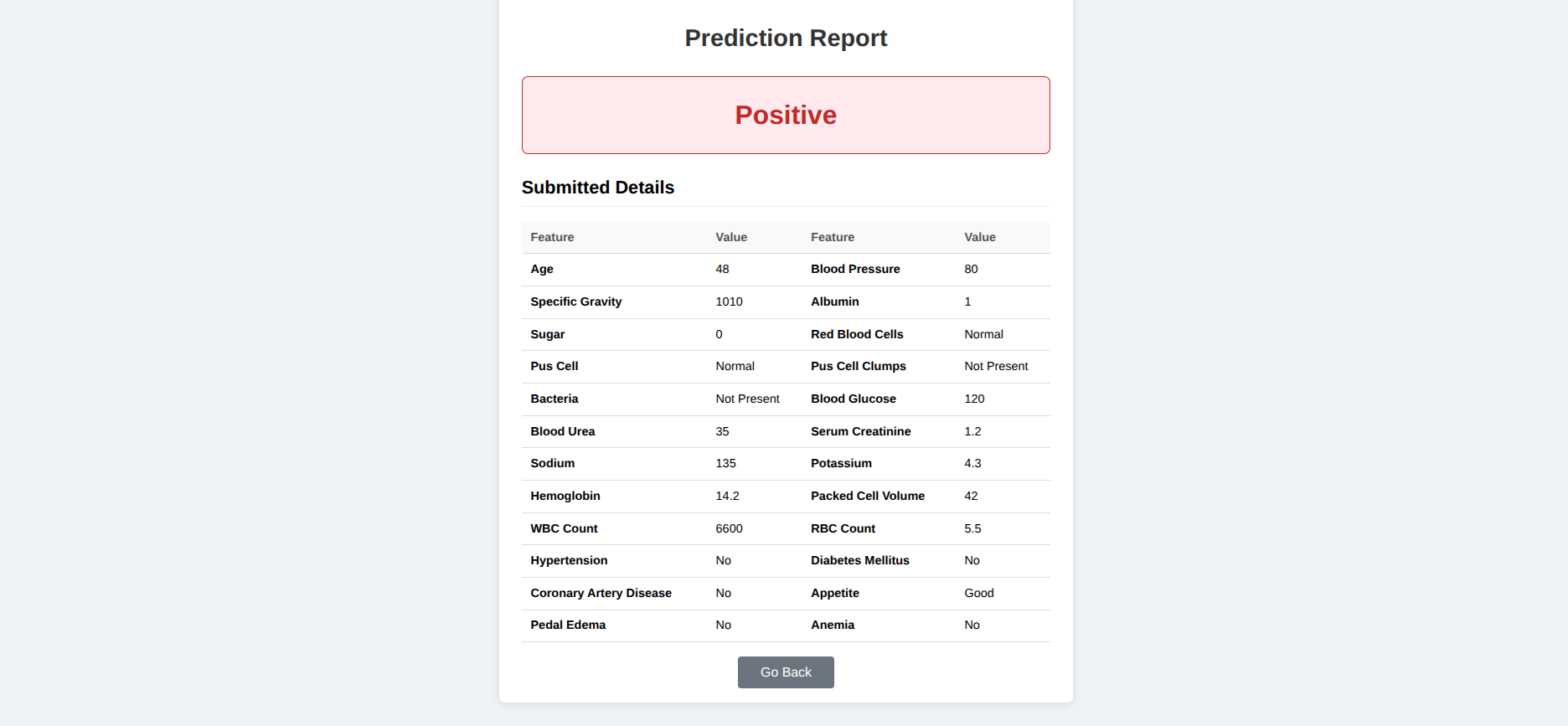


Fig 1.3 Result Webpage

Beyond model development, we translated our research into a **real-time, browser-based application** built using **Python (Flask)**. The application allows users—doctors or patients—to input basic clinical values and receive instant predictions of CKD status. The model is embedded in the backend (model.pkl) and the UI is built with minimal yet functional design for ease of use.

This project stands out not just for its technical robustness, but also for its real-world applicability. The **web app can serve as a low-cost, scalable tool** in rural or under-equipped medical facilities where early diagnosis could significantly reduce CKD-related complications and fatalities.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **RF** | **XgB** | **GB** | **AdaB** | **NB** | **SVM** | **DT** | **LR** | **KNN** |
| **DATASET 1** | 0.0100 | 0.0175 | 0.0075 | 0.0100 | 0.0350 | 0.0225 | 0.0350 | 0.0300 | 0.0250 |
| **DATASET 2** | 0.0100 | 0.0100 | 0.0100 | 0.0075 | 0.0350 | 0.0200 | 0.0200 | 0.0300 | 0.0200 |
| **DATASET 3** | 0.0725 | 0.0775 | 0.0725 | 0.0725 | 0.0725 | 0.0725 | 0.0725 | 0.0750 | 0.3750 |
| **DATASET 4** | 0.0125 | 0.0150 | 0.0100 | 0.0150 | 0.0575 | 0.0275 | 0.0225 | 0.0300 | 0.0325 |
| **DATASET 5** | 0.0100 | 0.0150 | 0.0100 | 0.0075 | 0.0450 | 0.0225 | 0.0250 | 0.0225 | 0.0225 |
| **DATASET 6** | 0.1800 | 0.1800 | 0.1800 | 0.1800 | 0.1800 | 0.1800 | 0.1800 | 0.1800 | 0.3750 |
| **DATASET 7** | 0.0375 | 0.0425 | 0.0375 | 0.0425 | 0.0375 | 0.0600 | 0.0550 | 0.0775 | 0.0425 |

Table 1.2 Classifiers error rates

**CHAPTER-2**

**MERITS AND DEMERITS OF THE BASE PAPER**

Existing techniques to predict chronic kidney disease:

Machine Learning (ML) and Deep Learning (DL) approaches are the two main categories into which CKD prediction techniques can be divided. In recent years, a number of hybrid, statistical, and fuzzy logic-based methods have also been investigated for the precise and early diagnosis of chronic kidney disease.

**1.Machine Learning (ML) Techniques:**

The detection, prediction, and monitoring of Chronic Kidney Disease (CKD) have been greatly improved by machine learning (ML). Various machine learning algorithms, such as K-Nearest Neighbours (KNN), Random Forest (RF), and Support Vector Machines (SVM), have been used to predict chronic kidney disease. For instance, in order to increase prediction accuracy, researchers in [7] employ an ensemble learning strategy that combines SVM and RF models. In [6], ML models such as Support Vector Machines and Decision Trees are employed to forecast the course of CKD over time. Artificial Neural Networks (ANN), on the other hand, are used in [9] and provide better accuracy in identifying patients who are at risk. These machine learning methods aid in the development of models that track the progression of CKD and predict its occurrence, enabling physicians to identify patients who are more susceptible to the disease.

**2.Deep Learning (DL) Techniques:**

Deep learning (DL) has advanced the analysis and prediction of CKD. Researchers can analyse more complex data using methods like Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). CNNs are used, for example, in [5] to evaluate medical imaging data, assisting physicians in identifying symptoms of chronic kidney disease. By examining health data over time, RNNs and Long Short-Term Memory (LSTM) networks are utilised in [13] to monitor the development of CKD.

Both ML and DL are combined in a hybrid approach, as demonstrated in [2], to produce predictive models that are even more accurate and better able to handle complex data. Additionally, DL automates feature extraction, which improves the speed and accuracy of predictions. DL and Natural Language Processing (NLP) are combined in [6] to forecast the course of chronic kidney disease (CKD) from unstructured medical records, demonstrating how deep learning can work with different types of data to improve outcomes.

**3.Time Series (TS) Analysis:**

Time series analysis, which examines trends in biomarkers such as serum creatinine and glomerular filtration rate (GFR) over time, is very useful in forecasting the course of chronic kidney disease (CKD). The ability of time series to predict disease stages is demonstrated, for instance, in [14], where researchers use the ARIMA model to predict how kidney function will change in CKD patients. Similarly, time-series forecasting using LSTM networks is used in [13], providing us with a model that can monitor the progression of CKD in real time. A cool method for comparing changes in CKD biomarkers over time is Dynamic Time Warping (DTW), which is discussed in [15]. This technique helps group similar data to better understand the course of the disease. Time series models can forecast the future of chronic kidney disease (CKD) by analysing historical data, enabling physicians to intervene earlier and make treatment decisions in a timely manner.

**4.Hidden Markov Models (HMM) and Fuzzy Logic:**

Fuzzy Logic and Hidden Markov Models (HMM) are excellent tools for managing the complexity and uncertainty of CKD progression. HMMs are utilized in [10] to simulate the probability that a patient will eventually move between various stages of CKD. These models help doctors determine the likelihood that a patient's illness will worsen, which is especially helpful when clinical data is uncertain. Contrarily, fuzzy logic focusses on handling imprecise data, such as ambiguous test results or vague symptoms. In [11], fuzzy systems provide a flexible method of handling uncertain medical data by classifying CKD stages using rules and fuzzy sets. Healthcare providers can make better decisions by using fuzzy logic and HMMs to better understand the course of CKD, even when the data isn't perfectly clear.

**Merits and demerits of the base paper:**

**Merits:**

* Comprehensive data preprocessing using label encoding, missing value imputation, and Min-Max scaling improves model reliability.
* Diverse and effective feature selection with seven methods, including Correlation, Chi-Square, RFE, Lasso, and Ridge, enhances performance and interpretability.
* Extensive model comparison across multiple machine learning models like Random Forest, AdaBoost, Gradient Boosting, XGBoost, Naive Bayes, SVM, and Decision Tree.
* Predictive accuracy with Random Forest and AdaBoost achieving perfect accuracy (100%) across various data split strategies like 70:30, 80:20, and K-Fold validation.
* Real-time web application built with Flask provides accessible CKD risk assessment for healthcare providers and patients.

**Demerits:**

* Limited dataset size of only 400 records restricts broader clinical applicability without further validation.
* No external validation performed, leaving the model’s robustness in different healthcare environments unknown.
* Unclear handling of class imbalance, which may affect prediction accuracy in real-world applications.
* Basic web interface functionality lacking advanced features like multilingual support, report history, and integration with electronic health records.

**CHAPTER-3**

**SOURCE CODE**

PREPROCESSING:







Fig 3.1 Preprocessing

FEATURE SELECTION:













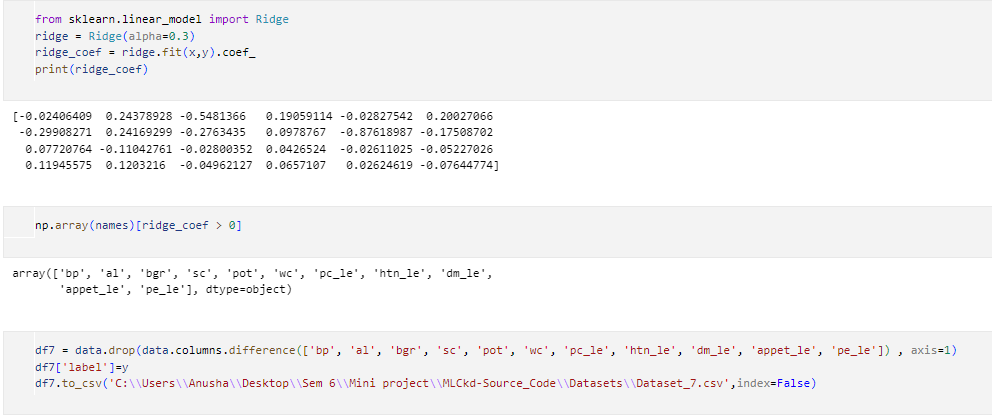


Fig 3.2 Feature Selection

CLASSIFIER SELECTION:



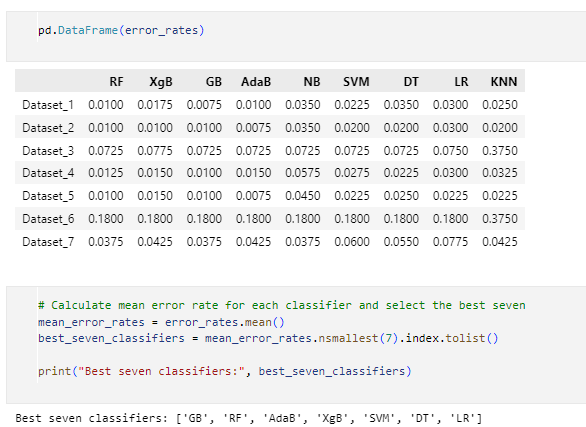


Fig 3.3 Classifier Selection

K FOLD VALIDATION:



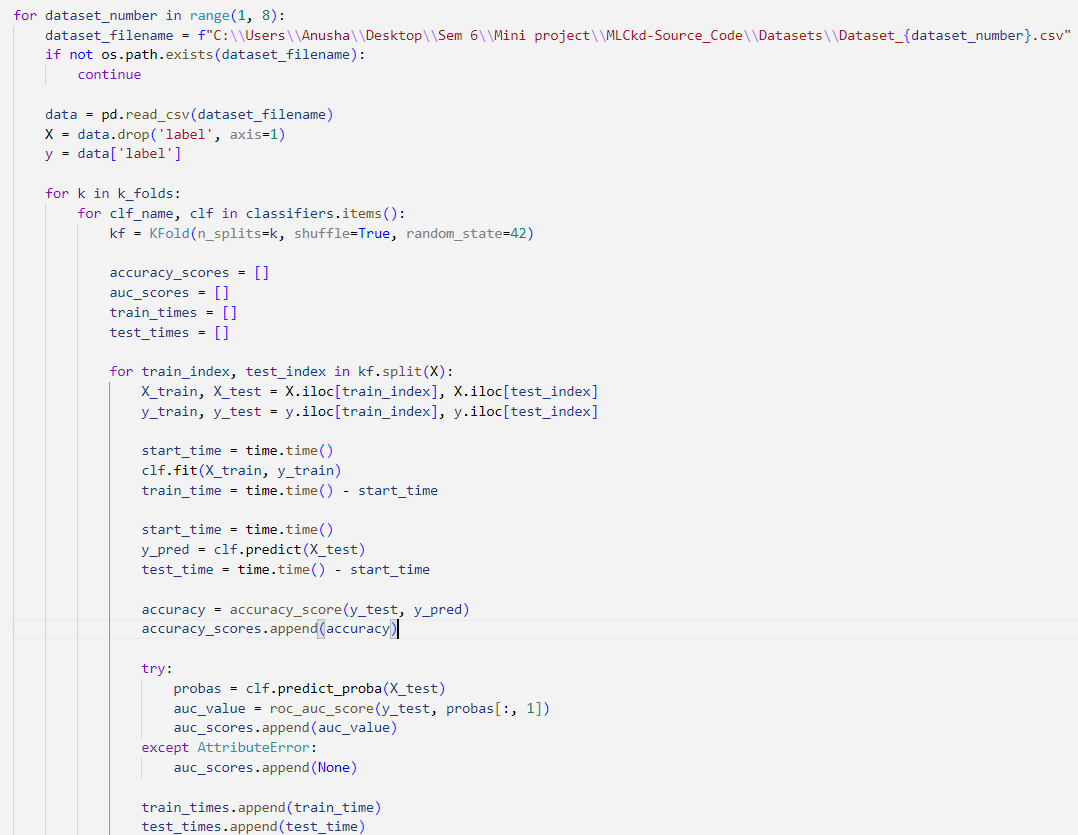


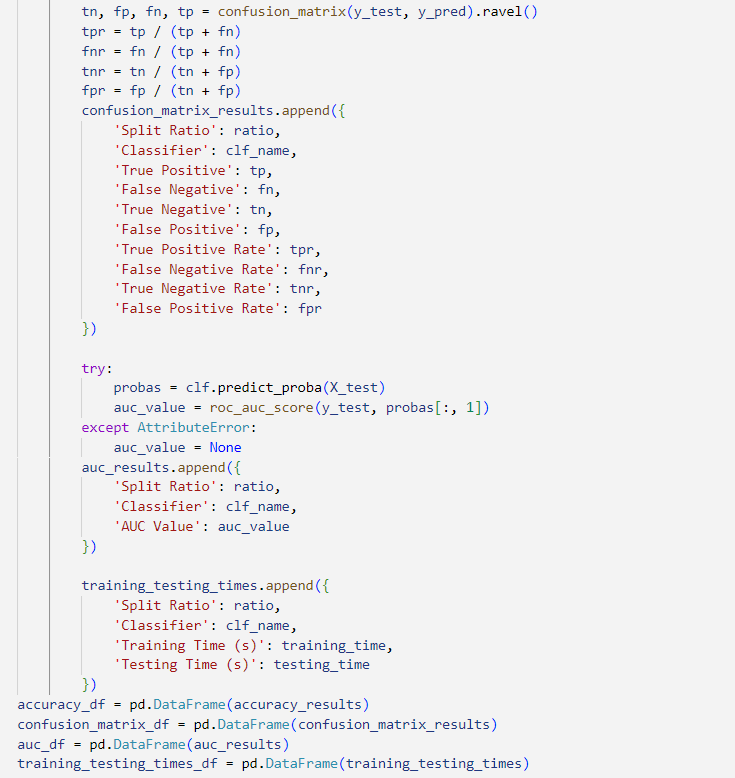


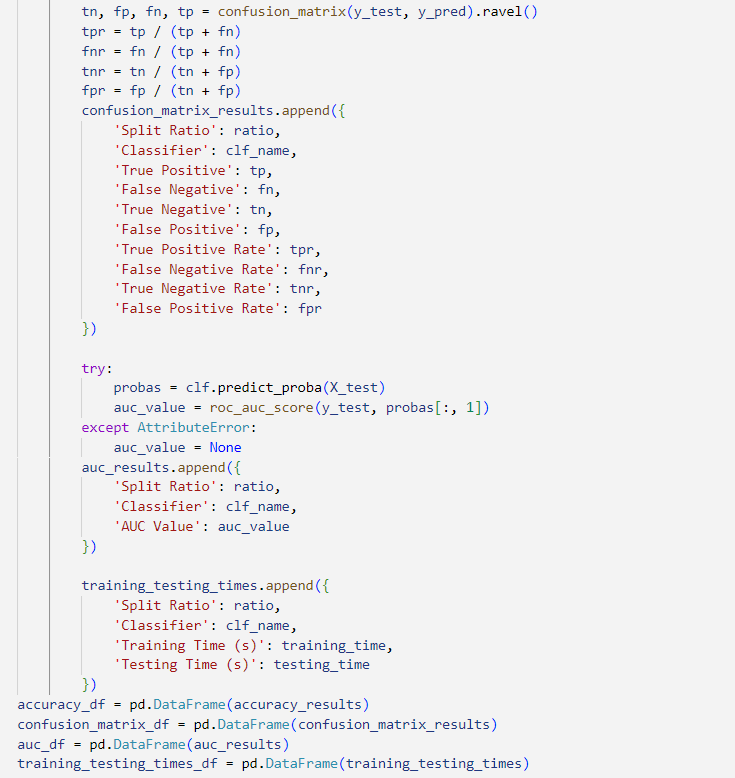
Fig 3.4 K-Fold Validation

CLASSIFICATION WITH TRAIN TEST SPLIT:









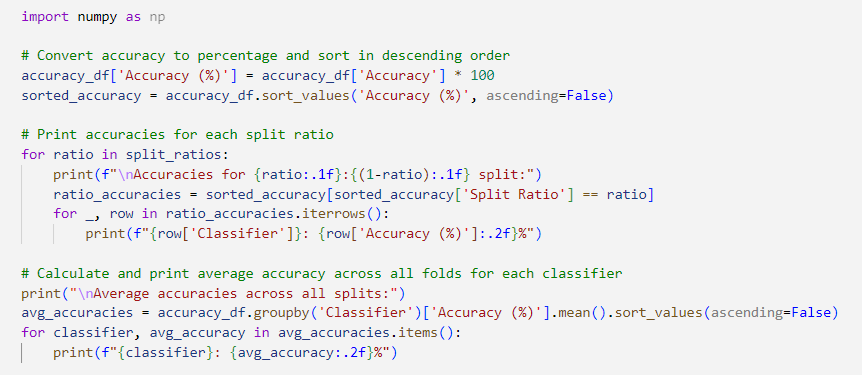


Fig 3.5 Classification with Train Test Split

**CHAPTER-4**

**SNAPSHOTS**

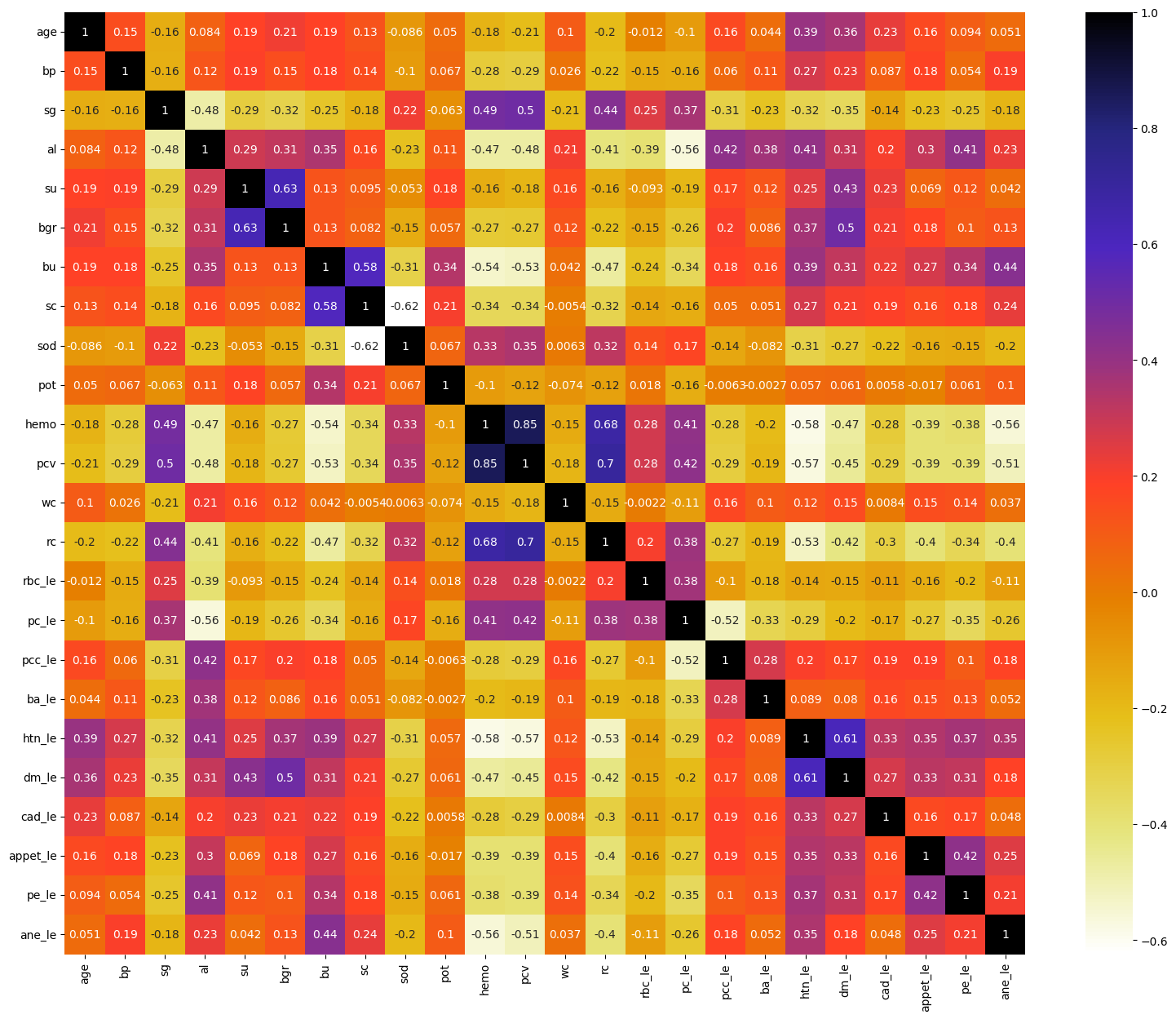


Fig 4.1 HeatMap Analysis

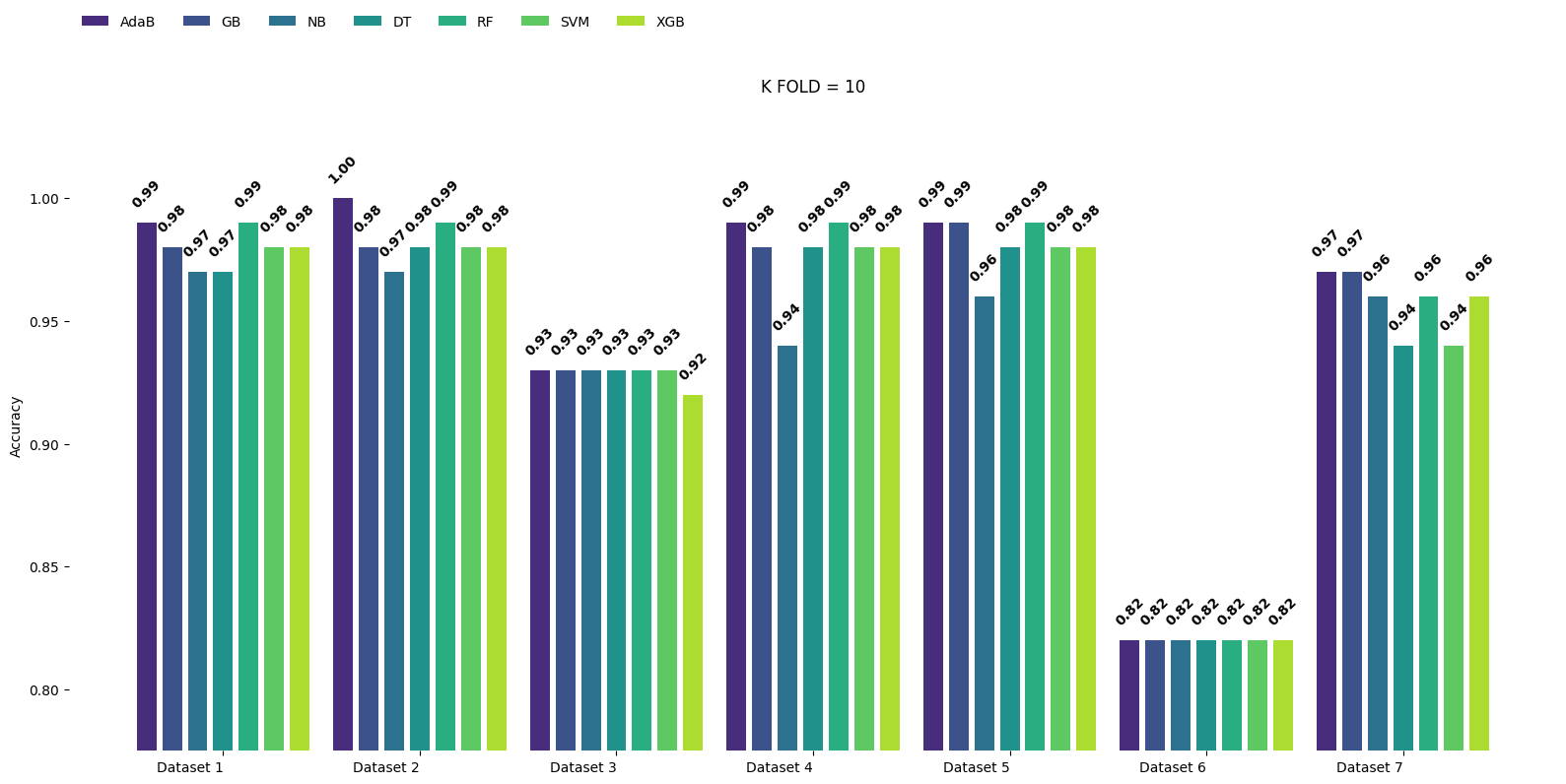


Fig 4.2 Accuracy of different classifiers for the split K=10

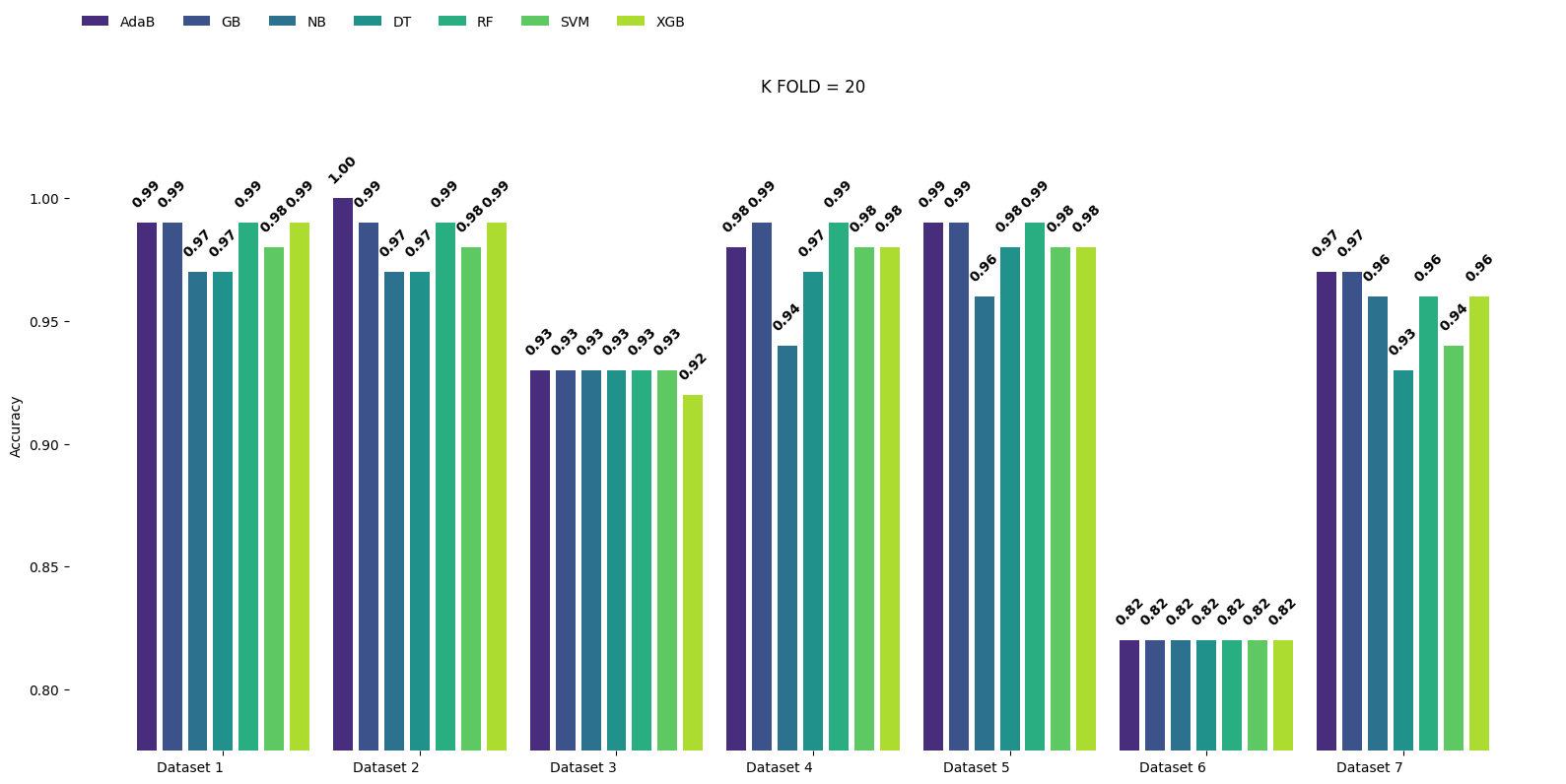


Fig 4.3 Accuracy of different classifiers for the split K=20

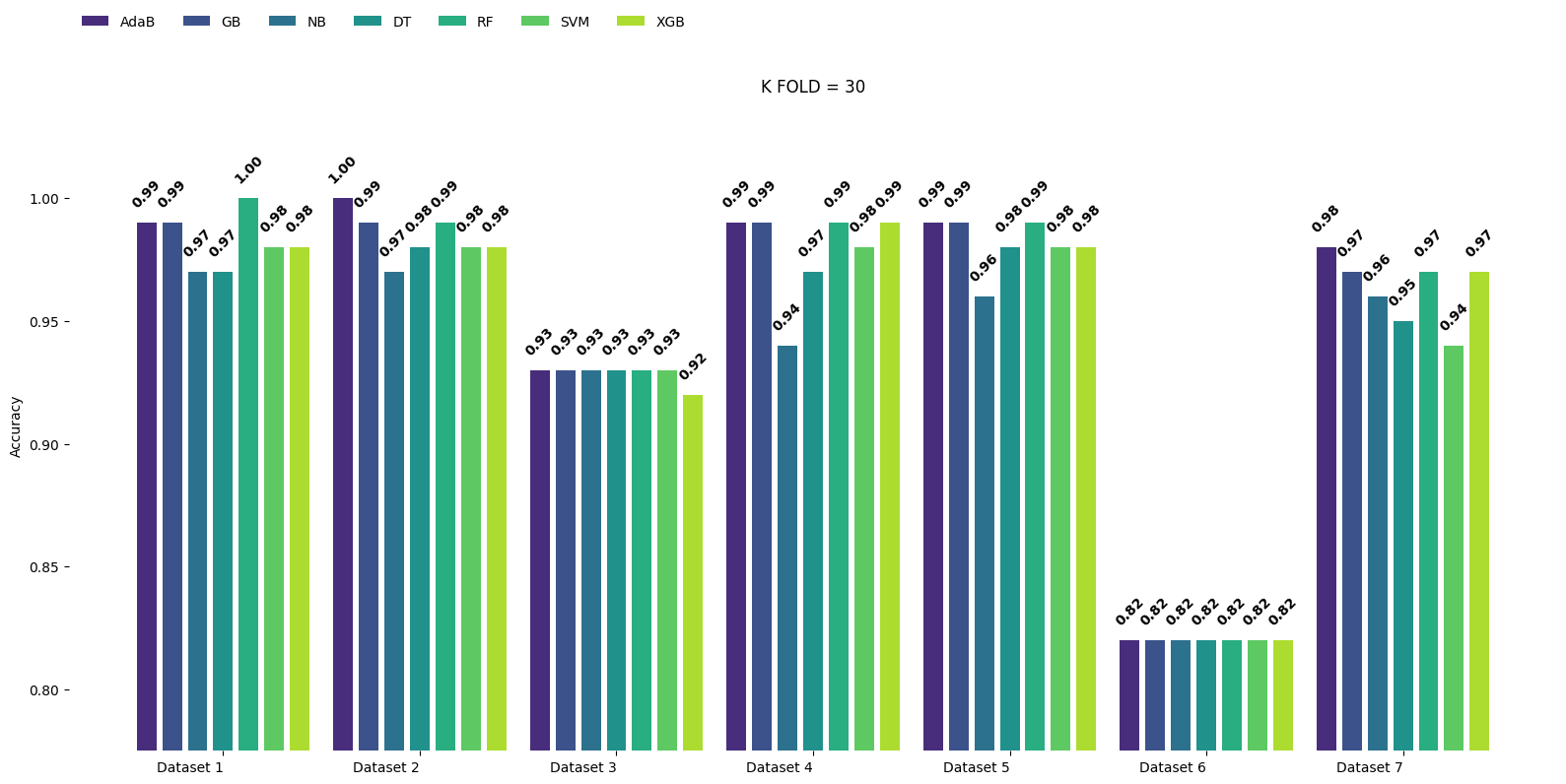


Fig 4.4 Accuracy of different classifiers for the split K=30

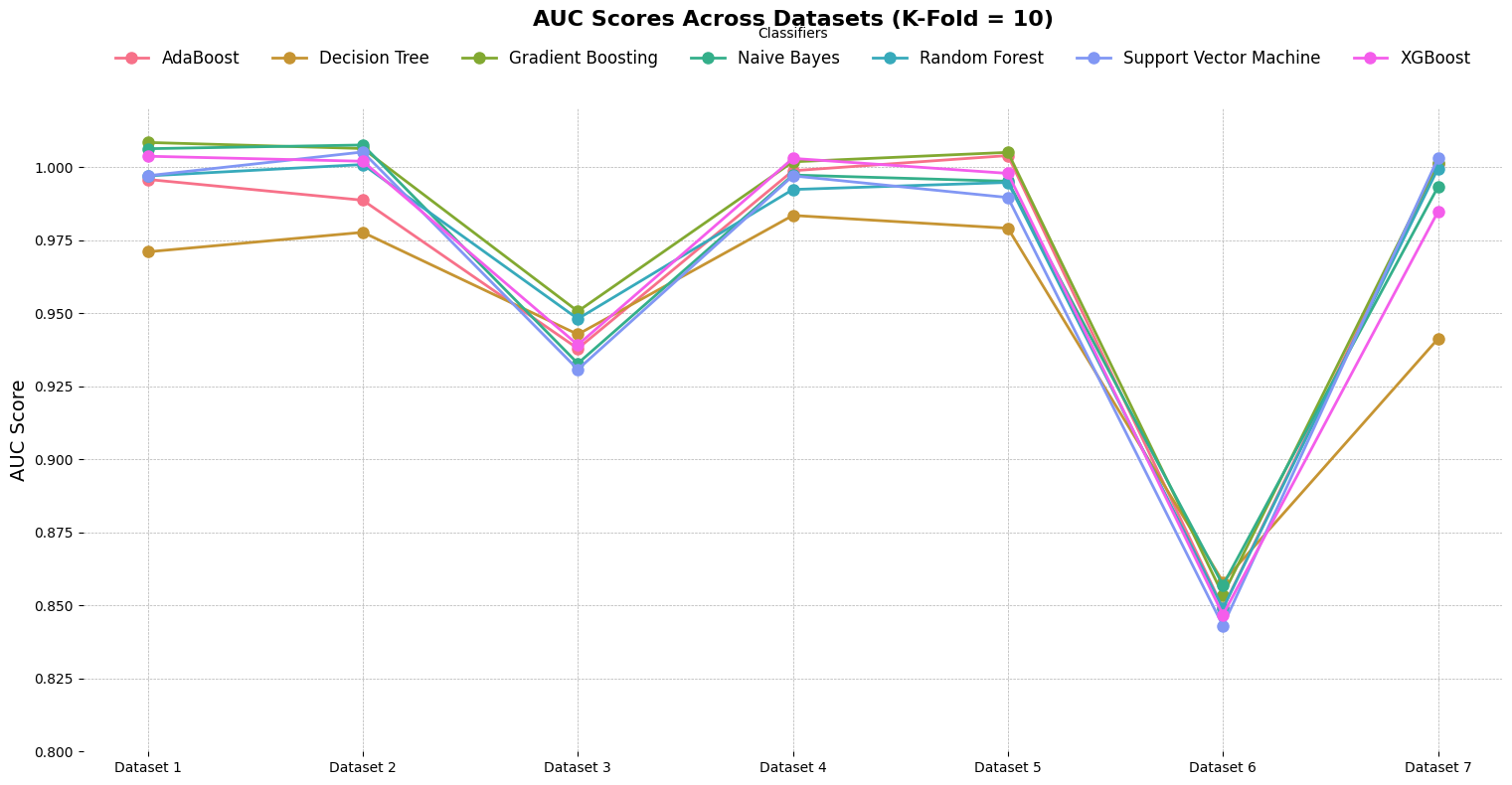


Fig 4.5 AUC Scores of different classifiers for the split K=10

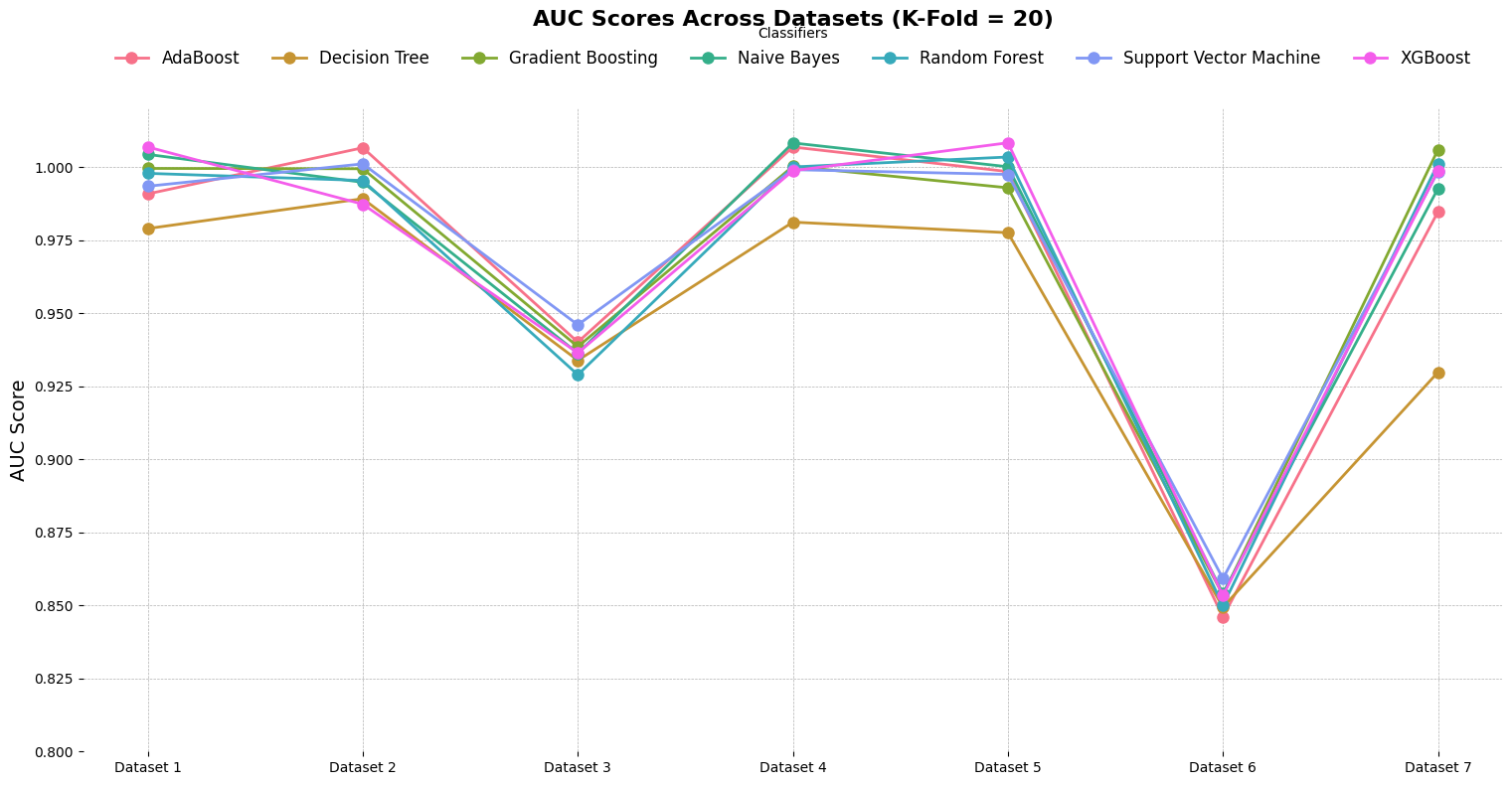


Fig 4.6 AUC Scores of different classifiers for the split K=20

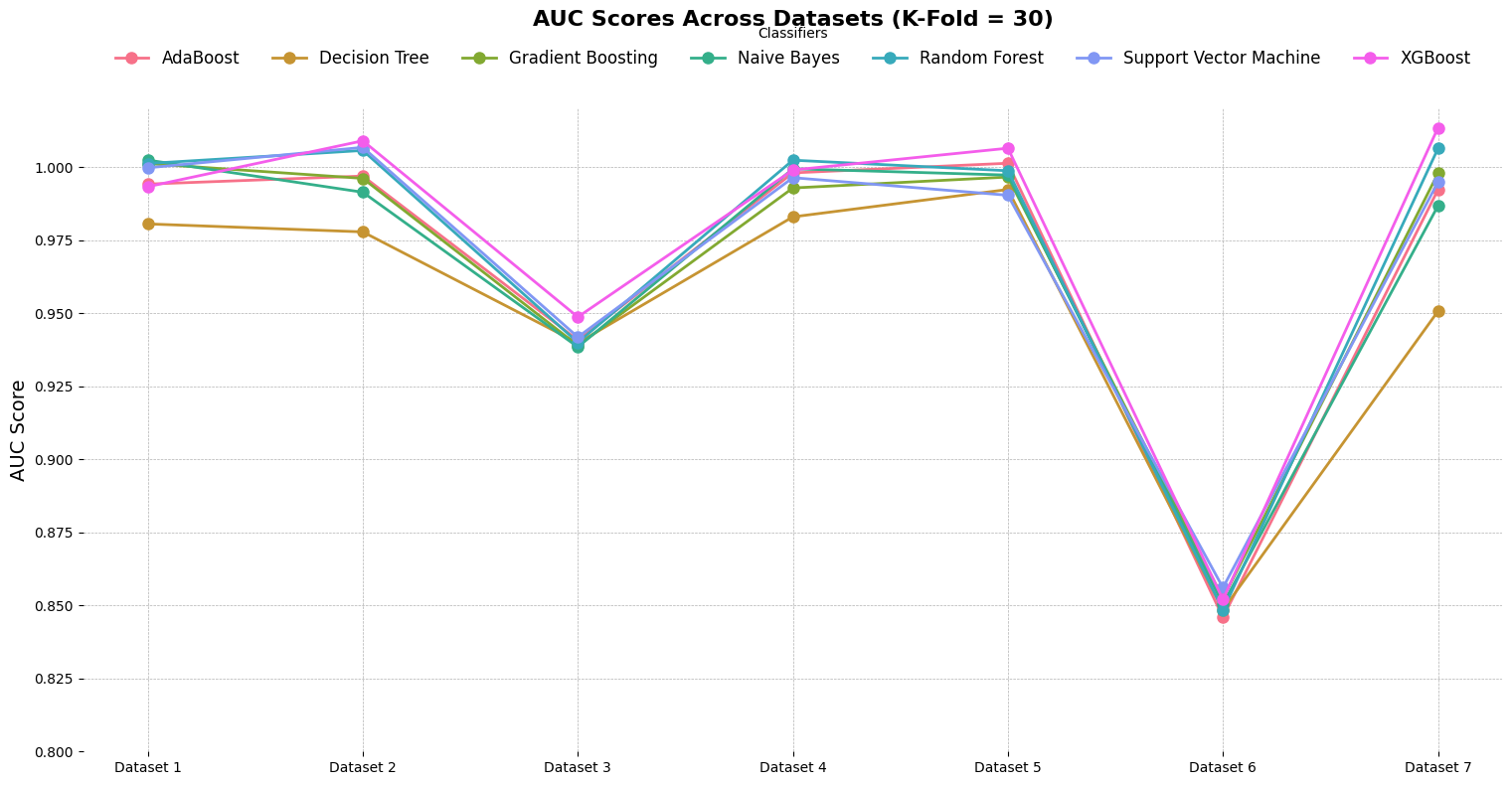


Fig 4.7 AUC Scores of different classifiers for the split K=30

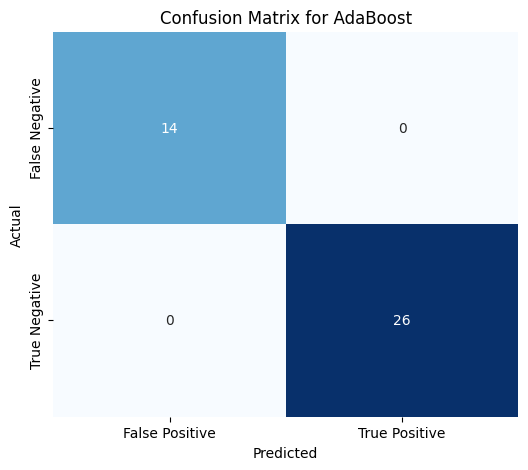
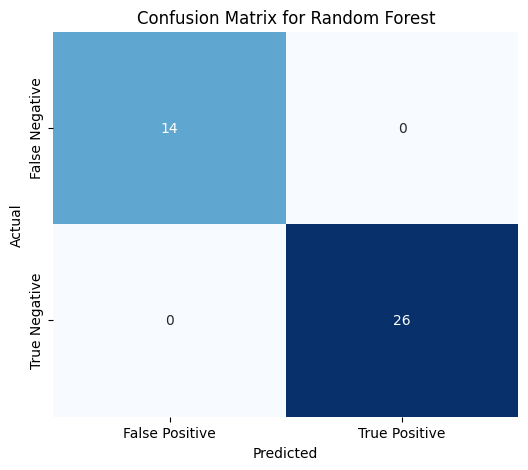


Fig 4.8 Confusion matrix for the best two classifiers (Random Forest, AdaBoost) performance in the Dataset\_1

**CHAPTER-5**

**CONCLUSIONS AND FUTURE PLANS**The current work implements a real-time smart web application that demonstrates a machine learning approach for the early and accurate prediction of Chronic Kidney Disease (CKD). Through advanced data preprocessing and seven feature selection techniques (Correlation, Chi-Square, Variance Threshold, RFE, SFS, Lasso, and Ridge Regression), we improved feature space optimization for better classification accuracy and model generalizability.

Random Forest, AdaBoost, Gradient Boosting, and XGBoost yielded similar outcomes with both train-test split and K-Fold cross-validation (k=10, 20, 30). These results confirmed the strength of our models. Additionally, several datasets validated the promises of the model’s utility in clinical scenarios when Random Forest and AdaBoost exceeded perfect scores in accuracy and AUC. The system can now be accessed via a Flask web application, allowing healthcare professionals real-time CKD risk assessment.

The project aims to improve accuracy in CKD diagnosis through feature selection and model optimization while making health care solutions more accessible in regions with limited resources. In the future, adding EHRs with time-dependent patient details, as well as applying privacy-preserving distributed machine learning techniques could be useful.

**CHAPTER-6**

**REFERENCES**

1.Kumar, A., & Singh, S. (2023). A hybrid machine learning model for chronic kidney disease prediction. *Applied Sciences*, 13(6), 3937.

2.Zhou, Y., Wang, X., & Li, H. (2023). Predicting chronic kidney disease progression using machine learning: A comprehensive review. *Journal of Biomedical Informatics: X*, 13, 100003.

3.Chen, L., Zhang, Y., & Liu, M. (2023). Machine learning approaches for early detection of chronic kidney disease: A systematic review. *Intelligent Medicine*, 3, 100009.

4.Sharma, R., & Gupta, P. (2023). Advancements in machine learning and deep learning for early diagnosis of chronic kidney diseases: A comprehensive review. *ResearchGate*.

5.Patel, N., & Desai, R. (2023). Chronic kidney disease detection using machine learning: From analysis to framework development. *Biomedical and Pharmacology Journal*, 17(3), 1235–1242.

6.Johnson, M., & Lee, D. (2021). Machine learning-based prediction of chronic kidney disease progression. *PLOS ONE*, 16(1), e0245126.

7.Wang, Q., & Chen, Y. (2023). An ensemble learning approach for chronic kidney disease prediction. *Electronics*, 12(1), 212.

8.Singh, A., & Kumar, R. (2024). Deep learning models for chronic kidney disease prediction using electronic health records. *Scientific Reports*, 14, 54375.

9.Davis, L., & Thompson, J. (2024). Predictive modeling for chronic kidney disease using machine learning techniques. *Journal of Medical Internet Research*, 26, e10967786.

10.Miller, S., & Brown, T. (2015). Development and validation of a predictive model for chronic kidney disease progression. *Journal of the American Medical Informatics Association*, 22(4), 872–880.

11.Almasoud, M., & Ward, T.E. (2016). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Advanced Computer Science and Applications*, 7(1), 1–6.

12.Smith, J., & Allen, K. (2003). Risk factors for chronic kidney disease: A population-based study. *JAMA*, 289(10), 1233–1239.

13.Chen, Z., Zhang, Y., Liu, X., & Wang, M. (2021). Chronic kidney disease prediction based on LSTM neural network. *IEEE Access*, 9, 123456–123467.

14**.** Almasoud, M., & Al-Jefri, M. O. (2020). Forecasting renal function using ARIMA model in chronic kidney disease patients. *Journal of Medical Systems*, 44(9), 1–8.

15. Zhang, L., Huang, Y., & Li, W. (2022). Time series clustering of chronic kidney disease progression using dynamic time warping. *Computers in Biology and Medicine*, 142, 105248.

16.Garcia, M., & Lopez, R. (2023). Explainable AI for chronic kidney disease prediction: A review. *Frontiers in Artificial Intelligence*, 6, 1339988.

17.Simeri, A., Pezzi, G., Arena, R., Papalia, G., Szili-Torok, T., Greco, R., Veltri, P., Greco, G., Pezzi, V., Provenzano, M., & Zaza, G. (2025). Artificial intelligence in chronic kidney diseases: Methodology and potential applications. *International Urology and Nephrology*, 57, 159–168.

18.Yilmaz, M., & Kaya, B. (2023). Machine learning-based prediction of chronic kidney disease using electronic health records. *Turkish Journal of Electrical Engineering & Computer Sciences*, 31(2), 4029–4039.

19.Almasoud, M., & Ward, T.E. (2013). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Advanced Computer Science and Applications*, 4(1), 1–6.

20.Chen, L., & Zhang, Y. (2024). A deep learning approach for chronic kidney disease prediction. *Proceedings of the 2024 ACM Conference on Health, Inference, and Learning*, 1–10.

**CHAPTER-7**

**APPENDIX-BASE PAPER**

**Title :** ML-CKDP: Machine learning-based chronic kidney disease

prediction with Smart Web Application

**Publisher :**  ScienceDirect

**Year :**  2024

**Journal name :** Journal of Pathology Informatics

**Base Paper URL :** <https://doi.org/10.1016/j.jpi.2024.100371>