CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING MODELS

M. Siri Chandana – 2203A51492 | K. Pavani – 2203A51482 | B. Sreeja – 2203A51467 | P. Trisha – 2203A51383 | M. Sravani – 2203A51019

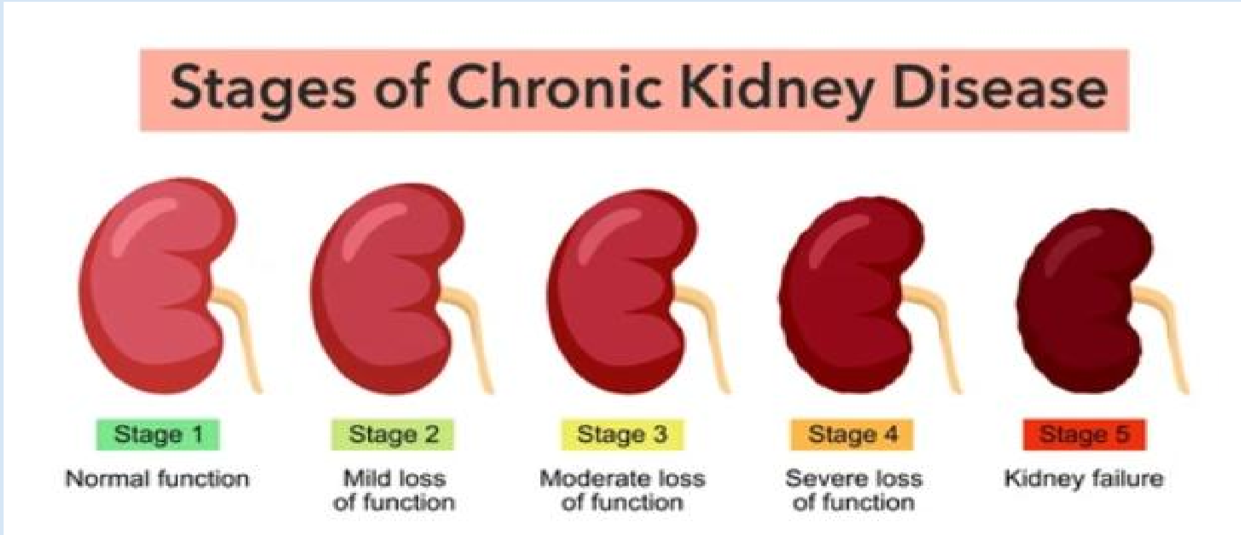
SR UNIVERSITY

ABSTRACT

Chronic Kidney Disease (CKD) is a progressive disorder characterized by gradual loss of kidney function. This project presents a Machine Learning-based system for early CKD prediction using clinical and demographic data. The study compares multiple models and identifies the optimal one based on performance metrics. The proposed Random Forest model achieved the highest accuracy (97%) with robust predictive capability. This system aims to support clinical decision-making, improve patient outcomes, and reduce healthcare costs.

INTRODUCTION

Chronic Kidney Disease (CKD) affects nearly 10% of the global population. Major risk factors include diabetes, hypertension, and obesity. Late detection leads to irreversible damage, costly treatments, and increased mortality rates. Traditional diagnostic methods rely on laboratory tests that are time-consuming. Machine Learning (ML) enables data-driven analysis of multiple parameters for early CKD detection. This study leverages algorithms like Decision Tree, Random Forest, and XGBoost to develop predictive models.



OBJECTIVES

- Develop an ML-based system for early CKD prediction.
- Analyze patient attributes influencing kidney health.
- Compare multiple models (Decision Tree, Random Forest, KNN, AdaBoost, XGBoost).
- Evaluate models using metrics like Accuracy, Precision, Recall, and F1-Score.
- Recommend the best model for healthcare implementation.

MATERIALS & METHODS

Dataset: The UCI CKD dataset with 400 patient records and 26 attributes (age, blood pressure, serum creatinine, hemoglobin, etc.).

Data Preprocessing: Missing values were imputed using KNN imputation, and categorical features were label encoded. Data was normalized for consistency. Outliers were detected using boxplots.

Algorithms Used: Decision Tree, Random Forest, KNN, AdaBoost, and XGBoost.

Mathematical Model:

Random Forest prediction is defined as:

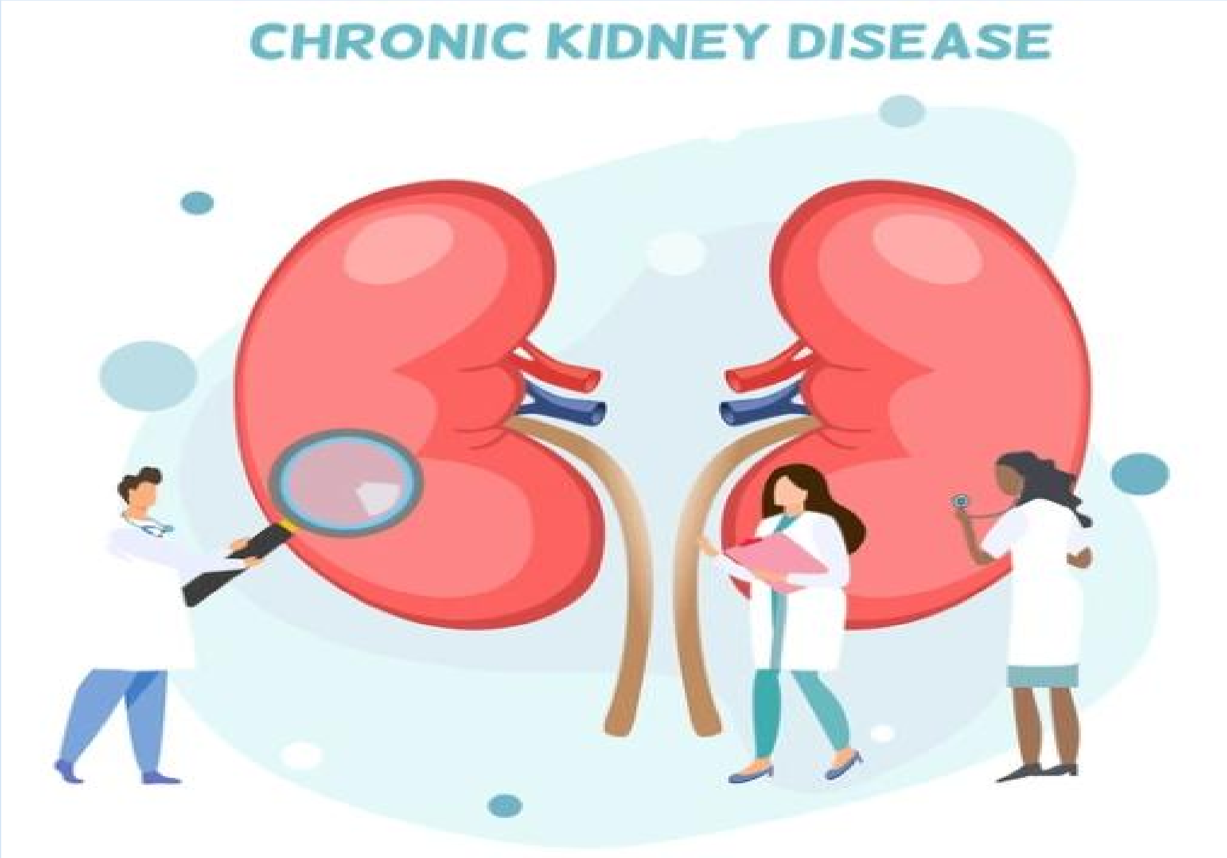
$f(x) = (1/N)\ \Sigma\ [T_i(x)]$ where T_i represents each decision tree.

Accuracy Formula:

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

DISCUSSION

This study demonstrates that ML can significantly improve CKD prediction accuracy compared to traditional methods. By analyzing multidimensional data, the model identifies subtle correlations between clinical parameters. The integration of Explainable AI (XAI) techniques such as SHAP and LIME enhances model interpretability for clinicians. Challenges include limited dataset diversity and real-time deployment constraints.

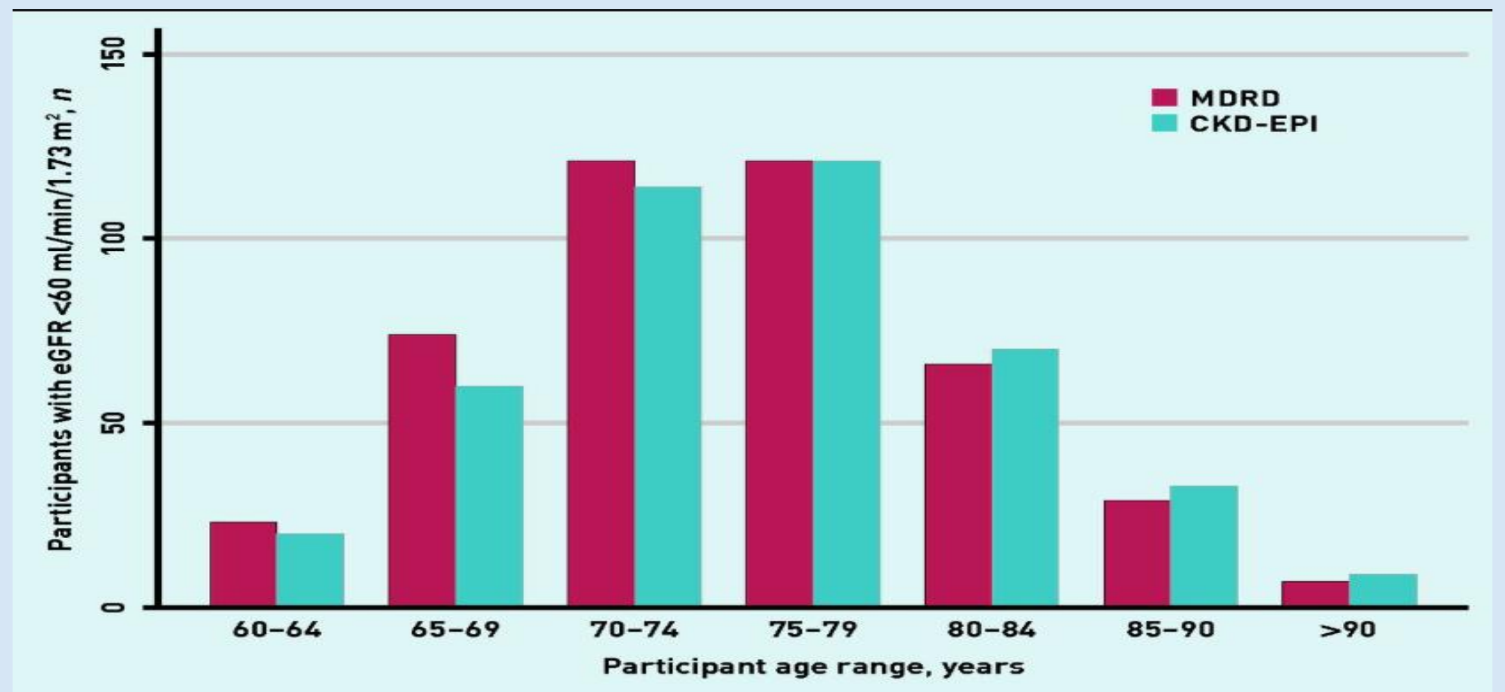


RESULTS

The Random Forest model achieved 97% accuracy, outperforming other algorithms. Precision and Recall were 96% and 95% respectively. The F1-score indicated excellent balance between sensitivity and specificity. Feature importance analysis revealed that serum creatinine, hemoglobin, and blood pressure were the most influential predictors.

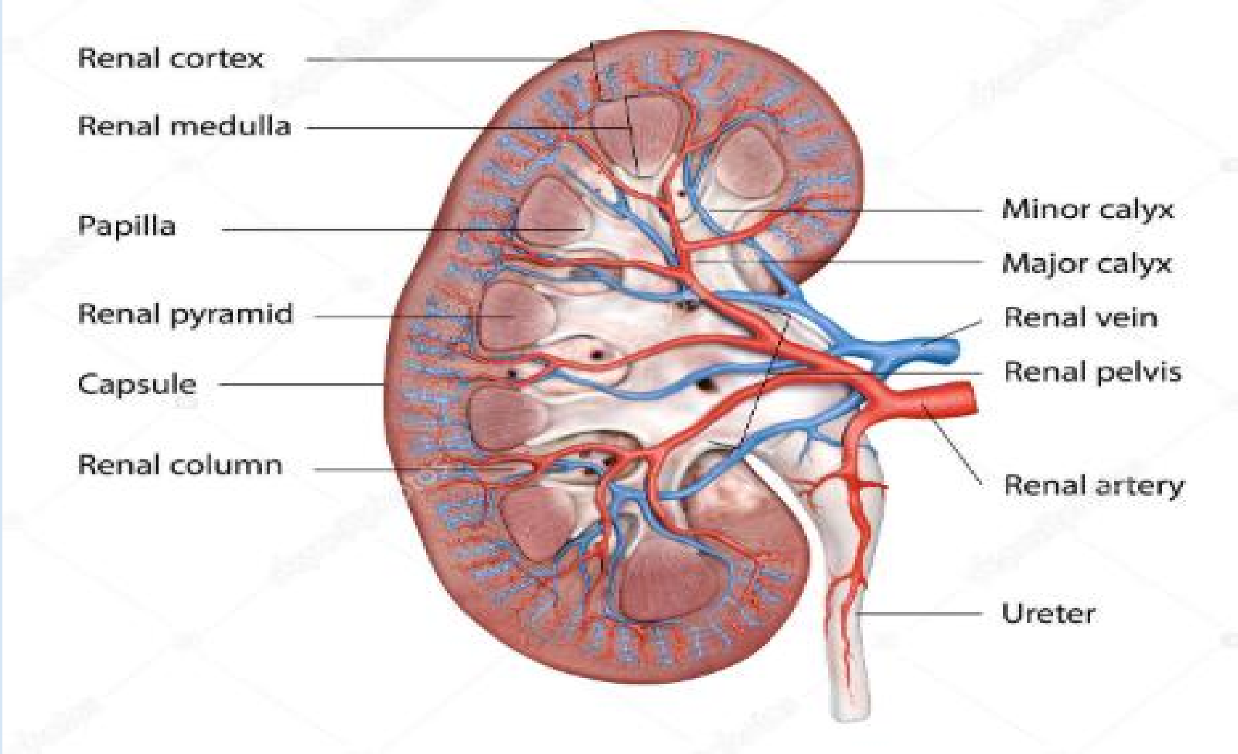
Performance Metrics Summary:

- Accuracy – 97%
- Precision – 96%
- Recall – 95%
- F1-Score – 95.5%



CONCLUSION & FUTURE WORK

The proposed Random Forest-based CKD prediction model provides a reliable, accurate, and interpretable framework for early detection. It can be integrated into clinical systems to assist doctors in making timely decisions. Future work involves integrating real-time wearable health data, expanding datasets across demographics, and enhancing model transparency using Explainable AI tools.



REFERENCES

[1] UCI CKD Dataset Repository.

[2] Sharma et al., 'Predicting CKD using Machine Learning,' IEEE Health Informatics, 2023.

[3] World Health Organization, 'Global CKD Report,' 2024.

[4] J. Brown et al., 'Explainable AI in Healthcare,' Springer, 2023.