CHRONIC KIDNEY DISEASE CLASSIFICATION

24EEE431 – AI and Edge Computing

MID-COURSE REPORT

Submitted by

PRIYEN S I (CB.EN.U4EEE22039)
AKSHIT GUNAWAT (CB.EN.U4EEE22059)
LINGESWAR S (CB.EN.U4EEE22160)
MITUN T S (CB.EN.U4EEE22165)

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING
AMRITA SCHOOL OF ENGINEERING,
AMRITA VISHWA VIDYAPEETHAM,
COIMBATORE - 641112

CONTENTS

ABSTRACT

Chronic Kidney Disease (CKD) is a global health challenge, often diagnosed late due to non-specific symptoms. This project leverages machine learning (ML) to enable early and accurate CKD classification using clinical and demographic data. Five ML models—Logistic Regression, SVM, Random Forest, Gradient Boosting, and Decision Tree—were evaluated on the UCI CKD dataset. After rigorous preprocessing and hyperparameter tuning, Random Forest emerged as the top performer with 96.2% accuracy and 97% specificity, minimizing false negatives. The study highlights the potential of integrating ML into healthcare systems for real-time screening, while acknowledging limitations such as dataset size and feature gaps. Future work includes expanding data collection and exploring advanced models like neural networks.

INTRODUCTION

Chronic Kidney Disease affects 10% of the global population, with delayed diagnosis leading to irreversible damage and high mortality. Early detection is critical but challenging due to non-specific symptoms like fatigue and swelling. This project addresses this gap by developing an ML-based classification system using 24 clinical features (e.g., age, blood pressure, serum creatinine). The primary objectives are to optimize diagnostic precision, reduce false negatives, and provide a tool for early intervention. The motivation stems from the need to curb healthcare costs, avoid unnecessary treatments from false positives, and improve patient outcomes through timely diagnosis.

PROBLEM STATEMENT

Problem Description

➢ The kidney is one of the most important body organs that filtrates all the wastes and
water from human body to make urine

➢ Chronic Kidney Disease (CKD), also commonly known as chronic renal disease or
chronic kidney failure is a life-threatening disease

➢ It leads to the continuous decrease of Glomerular Filtration Rate (GFR) for a period of
3 months or more and is a universal health problem

➢ CKD is caused by a variety of underlying factors, including diabetes, high blood pressure
and other diseases that damage the kidneys

➢ Early symptoms of CKD can be subtle and may include fatigue, swelling and decreased
urine output which is why it often goes undiagnosed until the later stages

➢ Early detection and treatment can help for slow the progression of the disease and
prevent complications

➢ Machine Learning (ML) techniques can be used to predict, diagnose and monitor
Chronic Kidney Disease (CKD)

Requirement Specification

➢ Use historical dataset (Dataset Attached) on chronic kidney disease from UC Irvine
Machine Learning Repository which consists of information such as age, blood pressure, Specific Gravity, Albumin, Sugar, Red blood cells and etc.

➢ Develop an appropriate ML model to classify whether chronic kidney disease is present or
not as per the given dataset.

METHODOLOGY

1. **Dataset**: The UCI CKD dataset (397 samples, 26 features) was used, containing demographic, clinical, and categorical variables.

2. **Preprocessing**:
    - Missing values: Median imputation for numerical features, mode imputation for categorical features.
    - Encoding: Categorical variables (e.g., hypertension, diabetes) mapped to binary values.
    - Scaling: StandardScaler applied to numerical features for SVM and Logistic Regression.
    - Class distribution: Adjusted to 62.4% CKD-positive and 37.5% CKD-negative after preprocessing.

3. **Model Selection**: Five models were tested, including Logistic Regression (baseline), SVM (RBF kernel), Random Forest, Gradient Boosting, and Decision Tree.

4. **Hyperparameter Tuning**: GridSearchCV optimized Random Forest (n_estimators=200, max_depth=20) and Decision Tree (max_depth=10).

5. **Validation**: An 80-20 stratified train-test split and 5-fold cross-validation ensured robust evaluation.

RESULT

Model	Accuracy	Precision	F1 - score
Logistic Regression	95.20%	95.34%	0.95
SVM	94.60%	94.78%	0.95
Random Forest	96.20%	96.30%	0.96
Gradient Boosting	95.50%	95.62%	0.95

CONCLUSION & FUTURE WORKS

The Random Forest model demonstrated superior performance, making it suitable for clinical deployment in electronic health records (EHR) for real-time CKD screening. Key limitations include a small dataset (180 samples in the PDF vs. 397 in code) and omitted genetic/lifestyle factors.

Future Directions:

1. Expand datasets with multi-center collaborations to improve generalizability.
2. Investigate neural networks for capturing complex patterns.
3. Address potential data leakage (e.g., removing id during training).
4. Include lifestyle and genetic features for holistic risk assessment.

This project underscores ML's transformative potential in healthcare, enabling proactive CKD management and reducing societal healthcare burden.

REFERENCES

- UCI Machine Learning Repository. (2023). Chronic Kidney Disease Dataset.
- Breiman, L. (2001). Random Forests. Machine Learning.
- Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. JMLR.