A blue and white logo

Description automatically generated

UNIVERSITY OF HERTFORDSHIRE  
School of Physics, Engineering and Computer Science

**MSc Advanced Computer Science with Research**

**7COM1039-0509-2022**

**Advanced Computer Science Masters Project**

4th December 2023

**Early Detection of Chronic Kidney Disease Using Machine Learning Techniques**

Name: Anjal Sali

Student ID: 21020816

Supervisor: Thiago Matos Pinto

**DECLARATION STATEMENT**

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Advanced Computer Science Masters Project at the University of Hertfordshire (UH).

It is my work except where indicated in the report.

I did not use human participants in my MSc Project.

I hereby permit the report to be available on the university website, provided the source is acknowledged.

# Abstract

Chronic Kidney Disease (CKD) presents a significant global health challenge, prompting the need for innovative approaches to detect and intervene early. This study utilises a comprehensive dataset incorporating clinical and demographic variables to examine the effectiveness of traditional machine learning methods, including K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), alongside a sophisticated neural network model, in identifying early signs of CKD. Rigorous attention is given to handling and cleaning the data, involving methods such as exploratory data analysis (EDA) and visualisation.

Conducting a thorough comparative analysis of KNN, SVM, and a neural network, the study assesses their performance metrics, including accuracy, precision, recall, and F1-score. It delves into the interpretability and generalizability of these models concerning CKD detection, emphasising addressing ethical considerations related to sensitive health data throughout the research process.

The results shed light on the strengths and limitations of each algorithm, offering valuable insights into their practicality in real-world clinical scenarios. Moreover, the study contributes to the broader discourse on the convergence of machine learning and healthcare, underscoring the transformative potential of these techniques in redefining early disease detection strategies.

The findings from this research enhance our understanding of the role machine learning plays in proactive healthcare interventions, setting the stage for future developments in CKD detection methodologies.

Contents

[Abstract 3](#_Toc152505932)

[1. Introduction 6](#_Toc152505933)

[1.1 CKD Symptoms and Causes 6](#_Toc152505934)

[1.2 CKD Diagnosis and Challenges 6](#_Toc152505935)

[1.3 Role of Machine Learning 7](#_Toc152505936)

[1.4 Problem Statement 7](#_Toc152505937)

[1.5 Aim and Objectives 8](#_Toc152505938)

[1.6 Research Questions 8](#_Toc152505939)

[1.7 Hypothesis 8](#_Toc152505940)

[1.8 Tools and Techniques 9](#_Toc152505941)

[1.9 Report Summary 9](#_Toc152505942)

[2. Literature Review 9](#_Toc152505943)

[3. Methodology 10](#_Toc152505944)

[3.1 Data Collection and Cleaning 10](#_Toc152505945)

[Figure 1: Dataset Overview 10](#_Toc152507213)

[Figure 2: Columns before cleaning 10](#_Toc152507214)

[Figure 3: Columns after cleaning 10](#_Toc152507215)

# Introduction

Chronic Kidney Disease (CKD) is a significant global health challenge with profound medical, economic, and societal implications. This complex condition involves the progressive deterioration of renal function over time, leading to a range of debilitating health complications, including cardiovascular disease, anaemia, bone disorders, and even end-stage renal failure (Go *et al.*, 2004). With an estimated prevalence of around 10% worldwide and its gradual, often asymptomatic progression, CKD imposes an immense burden on healthcare systems and patients alike (Hill *et al.*, 2016).

Early detection of CKD is critical for two primary reasons: improving patient outcomes and alleviating the economic strain on healthcare resources. Detecting CKD in its early stages presents an opportunity to intervene with targeted treatments that can slow or even halt its progression, thereby reducing the risk of complications and improving patients' quality of life (Gansevoort et al., 2013). Additionally, early detection can lead to cost savings by reducing the need for costly interventions such as dialysis or kidney transplantation that become necessary in the advanced stages of the disease.

## 1.1 CKD Symptoms and Causes

Chronic Kidney Disease involves the progressive loss of kidney function over time, leading to health complications such as cardiovascular disease, anaemia, bone disorders, and end-stage renal failure (Hill et al., 2016). While CKD often remains asymptomatic in its early stages, symptoms may eventually arise, including fatigue, fluid retention, hypertension, and changes in urinary patterns. The insidious nature of CKD's progression often means that the disease remains asymptomatic until advanced stages, making early detection and intervention a critical imperative (Levey *et al.*, 2003). CKD can result from various factors, including diabetes, hypertension, glomerulonephritis, and polycystic kidney disease (Hill et al., 2016). Lifestyle factors such as smoking, obesity, and inadequate hydration also contribute to the risk of developing CKD. These causes highlight the multifaceted nature of the disease and the need for comprehensive strategies for its prevention and early detection.

## 1.2 CKD Diagnosis and Challenges

The diagnosis of Chronic Kidney Disease (CKD) presents a considerable challenge due to its subtle progression and diverse underlying causes. While CKD is a complex condition with various clinical manifestations, its early stages are often asymptomatic, making timely detection difficult. Furthermore, CKD shares symptoms with other diseases, necessitating precise diagnostic techniques to differentiate it from other potential causes. The early stages of CKD are characterised by a lack of specific symptoms, leading to its "silent" nature. As the disease progresses, symptoms such as fatigue, fluid retention, changes in urinary habits, and hypertension may become apparent. However, these symptoms are non-specific and can overlap with other health issues, hindering accurate diagnosis solely based on clinical presentation. Laboratory tests play a crucial role in CKD diagnosis. Serum creatinine levels and estimated glomerular filtration rate (eGFR) are commonly used biomarkers to assess kidney function (Levey *et al.*, 2003). Imaging techniques like ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI) provide insights into kidney structure and potential abnormalities. These methods are instrumental in identifying structural anomalies such as kidney stones or cysts. However, they might not capture functional changes occurring in the early stages of CKD (Schwartz *et al.*, 2009). One of the primary challenges in CKD diagnosis is its asymptomatic nature during the initial stages, resulting in delayed detection. Additionally, co-morbidities and the overlap of symptoms with other medical conditions can lead to misdiagnosis or delayed intervention.

## 1.3 Role of Machine Learning

Machine learning's potential in CKD detection and prediction lies in its ability to process diverse and multidimensional patient data, encompassing clinical records, laboratory results, imaging studies, and genetic information. Through the development of predictive models, machine learning algorithms such as Support Vector Machines (SVM), Random Forests, and Deep Neural Networks (DNN) can assist healthcare professionals in making informed decisions by identifying subtle correlations that might otherwise be overlooked (Rajkomar et al., 2019).

Machine learning algorithms can be trained to recognise intricate patterns indicative of early-stage CKD that elude human observation. These algorithms learn from historical data, improving their accuracy continuously as they encounter new cases. Such predictive capabilities promise to enable earlier intervention, thereby reducing the risk of disease progression and its associated complications. The multifaceted nature of CKD, influenced by genetic, environmental, and lifestyle factors, aligns well with machine learning's capacity to handle complex datasets and identify intricate relationships. Moreover, machine learning techniques can facilitate personalised medicine by tailoring CKD management strategies to individual patient profiles (Esteva et al., 2017).

## 1.4 Problem Statement

Chronic Kidney Disease (CKD) is a critical global health issue characterised by its often silent and asymptomatic progression, leading to delayed diagnosis and compromised patient outcomes (Levey et al., 2003a). The multifaceted nature of CKD, influenced by diverse factors, necessitates innovative diagnostic tools to capture its incipient stages effectively (Hill et al., 2016). Traditional diagnostic methods, although valuable, exhibit limitations in detecting subtle changes in kidney function and identifying early indicators of CKD (Levey et al., 2003a). To address this challenge, there is an imperative to harness the potential of machine learning techniques. As a powerful subset of artificial intelligence, machine learning has demonstrated remarkable capabilities in analysing complex datasets to identify intricate patterns and relationships. Leveraging machine learning's proficiency in data-driven analysis, this research project aims to explore and develop advanced predictive models that can revolutionise CKD detection. Integrating machine learning algorithms with comprehensive patient data, including clinical records and laboratory results, aims to create accurate and sensitive diagnostic tools. These tools would aid in the early identification of CKD and contribute to individualised patient care by tailoring interventions based on risk profiles (Esteva et al., 2017). However, achieving these advancements requires overcoming data heterogeneity, model interpretability, and ethical considerations. The successful implementation of machine learning in CKD diagnosis promises to enhance healthcare practices, reduce the burden on healthcare systems, and ultimately improve the quality of life for individuals at risk of or affected by CKD.

## 1.5 Aim and Objectives

This research project aims to implement and compare effective machine learning-based techniques for the early detection and prediction of Chronic Kidney Disease (CKD), aiming to improve patient outcomes and reduce the burden on healthcare systems.

Some objectives of the research are.

* Conduct an extensive review of existing literature on CKD, its progression, diagnostic challenges, and the role of machine learning in healthcare.
* Gather a comprehensive dataset encompassing clinical records, laboratory results, and patient history for CKD patients and controls.
* Identify relevant features and attributes that contribute to the early detection of CKD.
* Develop predictive machine learning models, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Deep Neural Networks (DNN), to detect CKD.
* Employ appropriate evaluation metrics to assess the models' accuracy, sensitivity, specificity, and generalisation capabilities.
* Address ethical considerations regarding patient privacy, data security, and informed consent under healthcare regulations.

## 1.6 Research Questions

RQ1- Can machine learning techniques accurately predict the onset of Chronic Kidney Disease (CKD) based on clinical attributes, laboratory test results, and patient demographics?

RQ2- How does the accuracy of different machine learning algorithms compare when applied to the same CKD dataset?

RQ3- How does the accuracy of machine learning-based CKD detection models compare to traditional diagnostic methods, such as clinical assessments and laboratory tests?

## 1.7 Hypothesis

**H0 (Negative Hypothesis):** Machine learning techniques cannot accurately predict Chronic Kidney Disease (CKD) based on clinical attributes, laboratory test results, and patient demographics.

**H1 (Positive Hypothesis):** Machine learning techniques can accurately predict Chronic Kidney Disease (CKD) based on clinical attributes, laboratory test results, and patient demographics.

## 1.8 Tools and Techniques

The project harnessed various vital tools and techniques crucial for effectively implementing machine learning to detect Chronic Kidney Disease (CKD) early. Data preprocessing and manipulation were executed using Python, leveraging Jupyter and Google Colab as interactive development environments. Libraries such as Pandas and NumPy were instrumental in data handling and manipulation, while Scikit-learn provided functionalities for data preprocessing, feature selection, and model evaluation. Visualisations were crafted using Matplotlib and Seaborn, enhancing data exploration and model insights. Feature engineering was streamlined using FeatureTools for automated feature generation, and dimensionality reduction was achieved through Principal Component Analysis (PCA).

## 1.9 Report Summary

# 2. Literature Review

Chronic kidney disease (CKD) is characterised by gradually losing kidney function over time. If not detected and treated early, CKD can lead to kidney failure, requiring dialysis or transplantation. Therefore, early detection of CKD is critical to prevent disease progression and improve patient outcomes. Recent research has explored applying machine learning techniques to predict CKD based on patient data to assist with early diagnosis.

The conventional diagnostic arsenal comprises blood pressure measurement, urinalysis, blood tests for Serum Creatinine and Blood Urea Nitrogen, Glomerular Filtration Rate (GFR) calculation, medical imaging techniques (ultrasound, CT scans, and MRI), kidney biopsy, medical history assessment, physical examination, and proteinuria testing.

Regular blood pressure monitoring remains pivotal, considering the correlation between hypertension and CKD. Urinalysis and blood tests, focusing on serum creatinine and BUN levels, provide valuable insights into kidney function. GFR, calculated through various parameters, is a crucial indicator of renal health. Imaging modalities and kidney biopsies offer detailed structural information. Additionally, proteinuria testing and risk factor assessment, encompassing diabetes and cardiovascular diseases, contribute to early detection strategies.

Simple machine-learning approaches like logistic regression, SVM, decision trees, etc., have been tested for CKD prediction using the commonly studied UCI CKD dataset (Avci et al., 2018). The UCI repository dataset contains 400 patient records with 25 input attributes. D. Pavithra & R. Vanithamani (2021) compared the CKD classification performance of decision tree, KNN and SVM models. They found that SVM achieved the highest accuracy of 98.23% versus 94.69% for decision trees and 90.27% with KNN. Singh et al. (2022) also evaluated various traditional classifiers - SVM attained an accuracy of 93% in their experiments, decision tree had 96% accuracy, while KNN performed relatively poorly with 65% accuracy.

A 2022 study by Rajeshwari and Yogish developed models using Naive Bayes, Random Forest, Decision Tree, and Support Vector Machine algorithms to predict CKD using a dataset of 400 patients (Rajeshwari & Yogish, 2022). The Random Forest model provided the highest accuracy at 98.75%. Nikhila (2021) also found that ensemble methods like Random Forest and AdaBoost performed better for CKD prediction than other classifiers.

Machine learning models, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, Naive Bayes, and Artificial Neural Networks (ANN), have been extensively investigated (Singh et al., 2023; Polat et al., 2022). A pioneering study by Singh et al. introduced a deep neural network for early CKD detection, outperforming conventional classifiers with 100% accuracy (Singh et al., 2023). Critical features identified through Recursive Feature Elimination (RFE) included Haemoglobin, Specific Gravity, Serum Creatinine, Red Blood Cell Count, Albumin, Packed Cell Volume, and Hypertension (Singh et al., 2023).

The high accuracy demonstrates the potential of machine learning, and specifically Random Forest models, to accurately predict CKD using patient data. This could assist doctors in early detection of the disease to prevent progression to kidney failure. The authors note that further improvements can be made by incorporating more training data and additional algorithms.

The machine learning models have been only tested on the small UCI dataset of 400 patients. Evaluating model performance on more extensive, diverse CKD datasets is necessary. Moreover, newer deep learning architectures can be explored to enhance accuracy further. Identifying correlations between predictive input features and incorporating domain knowledge in model development also needs investigation.

# 

# 3. Methodology

## 3.1 Data Collection and Cleaning

The Chronic Kidney Disease (CKD) dataset is a multivariate dataset containing clinical and laboratory data related to the diagnosis of chronic kidney disease from the University of California Irvine (UCI) Repository. The data is collected from patients who have been diagnosed with CKD as well as those who do not have the disease. The dataset features 24 clinical attributes, Comprising 400 patient records crucial for chronic kidney disease prognosis. A distinctive class attribute designates the patient's likelihood of chronic renal failure, with two expected feature diagnostic values: "ckd" and "notckd." The dataset distribution includes 250 instances of the "ckd" class (62.5%) and 150 instances of the "notckd" class (37.5%).

**Dataset Details:**

**Source:** UCI Machine Learning Repository

**Dataset Name:** Chronic Kidney Disease Data Set

**Data Type:** Multivariate

**Number of Instances:** 400 instances

**Number of Attributes:** 24 attributes

**Attribute Types:** The attributes include numerical and categorical variables, representing a range of patient characteristics and laboratory measurements.

Notable attributes such as age, blood pressure, specific gravity, albumin, and an additional 21 relevant features contributed to the dataset's comprehensiveness. Several features in the dataset, including age, blood pressure, random blood glucose, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, and red blood cell count, exhibit continuous values. Additionally, to attributes such as specific gravity, albumin, sugar, red blood cells, pus cell clumps, bacteria, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal oedema, and anaemia, binary values are assigned. This diverse dataset captures a wide range of clinical indicators, enabling a comprehensive analysis of CKD-related factors in the research project.

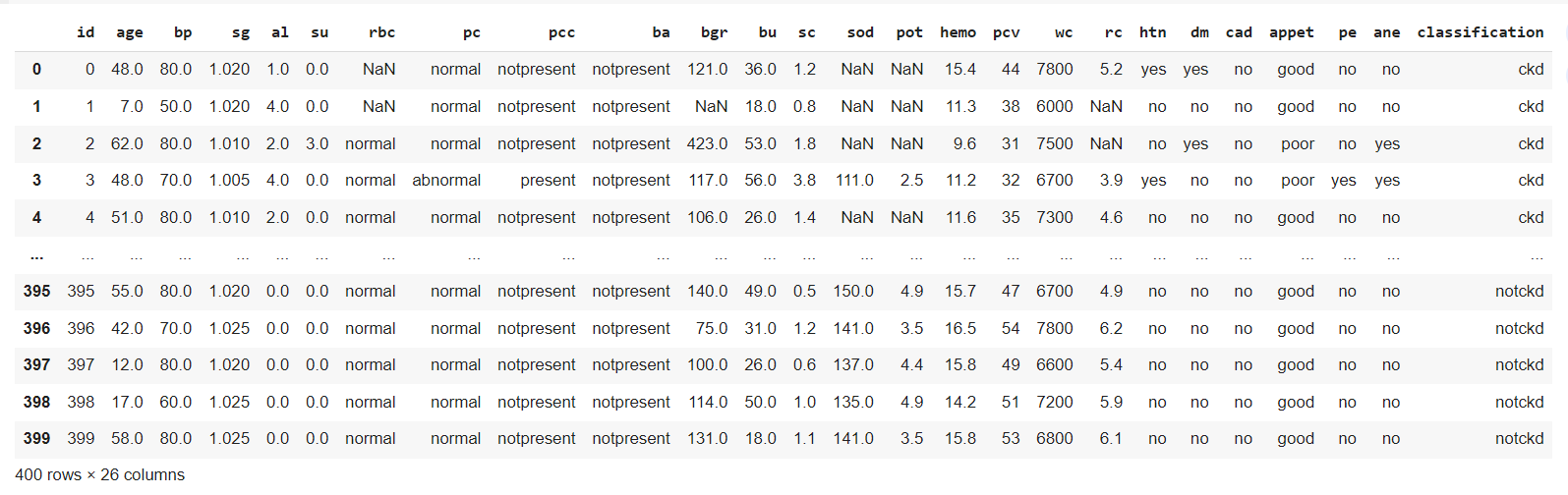


Figure 1: Dataset Overview

The initial stages of the project focused on data cleaning and processing to ensure the accuracy, consistency, and meaningfulness of the Chronic Kidney Disease (CKD) dataset. The "id" column, devoid of useful information, was dropped from the dataset. Additionally, the column names were renamed to more comprehensible and informative labels, enhancing the readability of the dataset. Categorical values in columns such as 'packedCellVolume', 'whiteBloodCellCount', and 'redBloodCellCount' were transformed into numerical representations, aligning with the numeric nature of the values.

|  |  |
| --- | --- |
| Figure 2: Dataset before cleaning | Figure 3: Dataset after cleaning |

## 3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical step in understanding the characteristics and patterns within the dataset before delving into predictive modelling. This section presents key observations and insights obtained through EDA:

The research examined the distribution of each feature to determine its data type (numerical or categorical) and identify any outliers or skewness. This analysis helped to understand the range of values for each feature and select appropriate preprocessing techniques.

Distribution plots were employed to gain a deeper understanding of the distribution of numerical values within the chronic kidney disease (CKD) dataset. Distribution plots provide a comprehensive visualisation of the frequency distribution of numerical attributes. By displaying a histogram along with kernel density estimation, it also offers insights into the data's shape, central tendency, and spread.

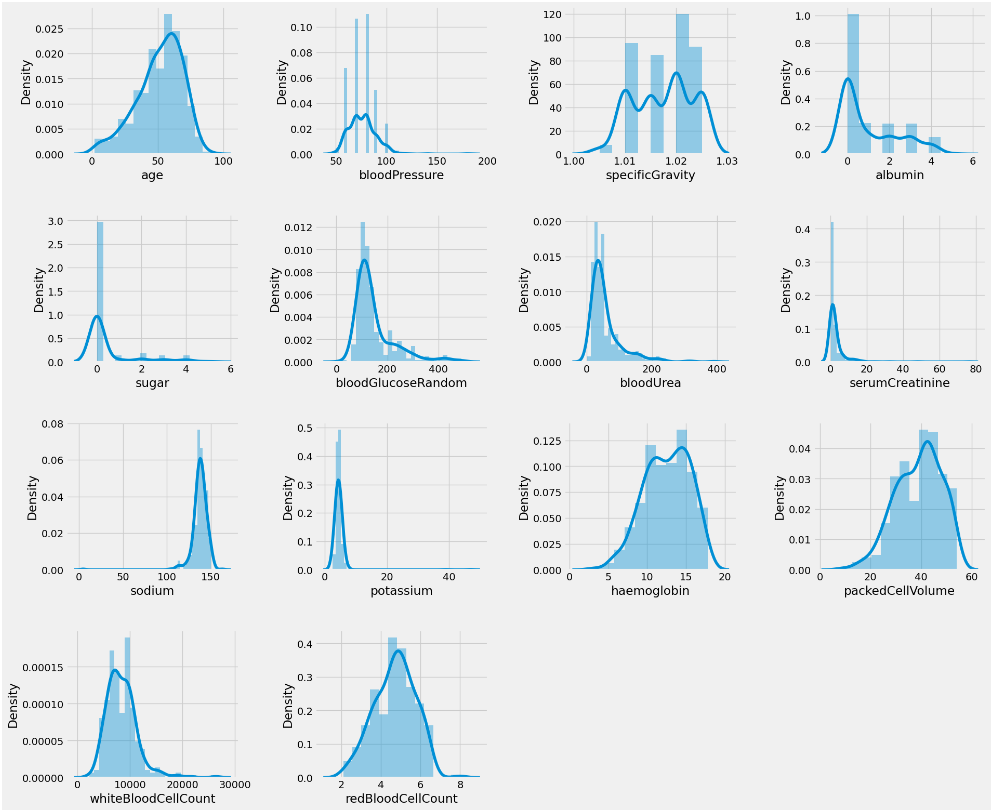


Figure 4: Feature distribution of numerical values

In the CKD dataset, counterplots were generated for categorical attributes such as 'diabetesMellitus', 'coronaryArteryDisease', 'class', and others. Each counterplot represents the frequency of occurrences for each unique category within the attribute.

A group of purple and orange bars

Description automatically generated

Figure 5: Categorical values distribution

By visualising the distribution of categorical values through counterplots, the project gained insights into the prevalence of specific categories and potential class imbalances. This type of visualisation aids in understanding the composition of the dataset and identifying any skewed distributions that could impact subsequent analysis or model performance.

A Heatmap displays a matrix of colour-coded values, where colours represent the degree of correlation between pairs of numerical attributes. Darker colours indicate higher positive or negative correlations, while lighter colours signify weaker or no correlations. By generating a heatmap, the project was able to quickly identify which attributes have strong correlations, potentially indicating redundant information or multicollinearity.

The Heatmap facilitated the detection of attribute pairs that might be highly correlated, which is crucial for data preprocessing and feature selection. Identifying such correlations aids in making informed decisions about which attributes to include in the modelling process and whether any feature engineering steps, such as dimensionality reduction techniques, are warranted.

A colorful squares with white text

Description automatically generated

Figure 6:The Heatmap of the dataset with correlation profile.