The Prediction of Chronic Kidney Disease with Machine Learning

Patrick Oluyori Akinwumi

20019561

MSc Applied Artificial Intelligence and Data Analytics

School of Management, Law and Social Science, University of Bradford

Supervisor: David Peck

Abstract

The aim of this paper is to aid in the detection of Chronic Kidney Disease (CKD) by using different machine learning techniques to identify CKD at an early stage. Chronic Kidney disease is illness that affects the kidney's ability to function normally. Effective prediction approaches should be investigated as the percentage of individuals afflicted by CKD continues to rise. Multiple machine learning classification methods were applied on a dataset of 400 patients retrieved from the University of California, Irvine (UCI) repository and 9 attributes relevant to chronic kidney disease diagnosis were selected in this research. The 9 relevant features were selected using Pearson correlation. Multilayer Perceptron, Random Forest, Logistics Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbour, and Nave Bayes were among the classification techniques utilised in this work. To conduct the tests, the mean of the respective features was used to replace all missing values in the dataset. Then, after tuning the parameters and doing multiple tests, the optimal parameters for the procedures were found. The best-obtained parameters and characteristics were used to create the final models of the suggested approaches. The comparison of accuracy among the applied machine learning techniques was evaluated using the AUC score and confusion matrix. This study reveals the impact of feature selection including hypertension, diabetes, and other features on the prediction of CKD. The study findings showed that all the models give excellent performance where Random Forest of 99% and AUC score 0.925 has the highest accuracy, indicating that the study's conclusion is quite promising. The study showed that machine learning algorithms can help with the prediction of chronic kidney disease and can aid in the development of a reliable computer-based diagnosis system that can help healthcare professionals treat patients more effectively. The model's findings indicate that a higher prediction of CKD at an early stage is possible. The code for this study is available in my public repository https://github.com/Oluyori/Chronic-Kidney-Disease-Prediction/blob/main/CKDPrediction%20(10).ipynb

Keywords: Chronic Kidney Disease, Machine Learning

Acknowledgement

I acknowledge everyone who has been part of my academic journey especially my parents, Mr. and Mrs. Akinwumi, my brother, Bamikole Akinwumi, and my uncle, Segun Alebiosu, for their encouragement and support.

Dedication

This work is dedicated to God, my family, and friends.

Contents

List of Tables

List of Figures

Glossary of Terms

Words	Meaning
GFR	Glomerular Filtration Rate
CKD	Chronic Kidney Disease
ESRD	End STAGE Renal Disease
AI	Artificial Intelligence
DM	Diabetes Meletus
ML	Machine Learning
KNN	K- Nearest Neighbour
GNB	Gaussian Naïve Bayes
LR	Logistics Regression
DT	Decision Tree
UCI	University of California, Irvine
MLP	Multilayer Perceptron
SVM	Support Vector Machine
RF	Random Forest
RF	Reinforcement Learning
AKI	Acute Kidney Injury
HTN	Hypertension
CVD	Cardiovascular Disease
KDOQI	Kidney Disease Outcomes Quality
Hypertension	An unknown elevation in blood pressure that raises the risk of cerebral, cardiac, and renal problems
Diabetes	a chronic condition characterised by high blood glucose levels and a disruption in fat and protein metabolism
Glomerulonephritis	Inflammation of the kidney's small filters (glomeruli)
Nephrons	The kidney's minute or microscopic structure and functioning unit
Entropy	The criterion for determining the degree of unpredictability
Algorithm	A collection of instructions for completing a task
EMR	Electronic Medical Records

ROC Curve	Receiver Operating Characteristics
AUC	Area Under Curve
Python	A popular open-source programming language in the data science field.
Precision	A metric for assessing a machine learning model's precision.
Accuracy	The measures used to assess a machine learning model's accuracy
F1-Score	A metrics for evaluating a machine learning model's performance.
GDPR	General Data Protection Regulation
Glomerulonephritis	A collection of conditions that produce glomerulus inflammation

Chapter 1

1.1 Introduction

The human kidney is an important organ that filters blood and keeps the body's fluid balance in check. Chronic Kidney Disease (CKD), on the other hand, affects its basic function. The term "chronic" refers to the long-term gradual deterioration of kidney cells. This is a severe form of renal failure in which the kidneys lose their ability to filter blood and the body becomes bloated with fluid (Kalantar-Zadeh and Fouque 2017). However, in many situations, CKD requires kidney transplantation or permanent dialysis. A strong family history of renal disease, diabetes, high blood pressure, and hypertension among others raises the risk of getting CKD (Chittora et al. 2021). Early identification of high-risk patients with reasonable accuracy is significantly important for controlling the disease and allocating scarce clinical resources and suitable care (Rady and Anwar 2019). Thus, this study aims to develop and compare predictive machine learning models of CKD prediction for early clinical intervention using relevant features from patients’ health data generated from UCI for classification as CKD or NotCKD. This chapter will introduce this topic by first discussing the study background, the research problems, the research aim, objectives and questions, the significance and finally the limitations.

1.2 Research Background

Kidney disease is a critical health problem that may progress through different stages from acute (sudden development) to chronic (long term development) (Kellum and Prowle 2018). CKD is a global public health concern , linked with high mortality and morbidity (Nwaneri and Ugo 2022), with increased prevalence in cardiovascular disease and high health care costs in the different age groups of the population (Rady and Anwar 2019). This can cause heterogeneous disorders that may result to end-stage renal disease (ESRD). On estimation, kidney diseases affect about 850 million people across the world, where the majority suffer from CKD (Kramer et al. 2018; Murton et al. 2021) , and a global prevalence of 8-16% as a major non-communicable disease (Jha et al. 2013). In the United States, for example, there is a rising CKD incidence, which affects around 10% of adults and has poor results and high treatment cost (Salomon et al. 2015). China is facing a significant financial burden as the CKD incidence rises. However, patients with ESRD and kidney failure survive by undergoing dialysis or kidney transplant which barely provide a high quality of life because of the high cost and low survival rate (Wang et al. 2016b).

Complications include kidney disease progression, damage to the nervous system, anaemia, acute kidney injury (AKI), bone disorders and cognitive disorders. Increased global incidence and prevalence of CKD, diabetes Meletus, ageing, hypertension and glomerulonephritis are the most common causes of CKD (Wang et al. 2016a), but other factors such as herbal and environmental toxins are the likely causes in other regions (Adeniran et al. 2018). Therefore, it is important to look at the features of the diseases that are most likely to cause CKD (Foëx and Sear 2004). High blood pressure, anaemia and other disease including coronary artery disease, and albumin in the urine, medication problems, salt and potassium deficit in the blood, and a family history are all sub-factors (Wang et al. 2018).

CKD symptoms at the early stages are minimal and not evident, and the disease may not be discovered until the kidneys have lost roughly 25% of their renal function (Chen et al. 2020). The glomerular filtration rate (GFR), which measures the kidney's function in terms of glomerular filtration, is a well-accepted CKD screening test. Stages I – V of CKD are determined by the estimated GFR (Raju et al.). When levels fall below 60 mL/min/1.73 m2 for 3 months or longer, CKD is diagnosed (Levin and Stevens 2014). GFR is used to differentiate between various degrees of renal function, from normal to kidney failure (Murton et al. 2021), and be calculated using mathematical calculations that consider serum creatinine, age, gender, body size, and ethnicity (Li et al. 2020a).

In this context, the rising prevalence of CKD indicates that the underpinning issues must be addressed in the diagnostic and detection processes to improve decision accuracy. Sizable studies reveal that kidney illness is extremely severe and can be deadly if not diagnosed early. Doctors can treat patients promptly by accessing predicting system capable of detecting patients likelihood of renal disease diagnosis in the future (Wang et al. 2018). Furthermore, the task relies heavily on the expertise and techniques of nephrologists, which is susceptible to inaccuracy (Kore and Yohannes 2018).

Where the poorest population are at the highest risk, detection and intervention can prevent CKD and where management strategies have been adopted, the occurrence of end-stage kidney disease will be reduced (Sabath 2015). As a result, early identification, and management of the condition are critical. The dynamic and hidden structure of CKD in the early stages, as well as patient heterogeneity make it important to accurately predict CKD progression.

CKD is often described by severity stages (Rady and Anwar 2019). Therefore, machine learning has a significant role to play in identifying hidden patterns from electronic health record (EHR) datasets (Chittora et al. 2021) to execute accurate treatment options. Machine learning approaches are widely applied to enhance the decision making process (Khamparia et al. 2020). The increasing use of ML in medicine promotes medical innovation, lowers medical expenses, and improves medical quality (Jayatilake and Ganegoda 2021). Cancer prognosis and prediction are examples of efforts in this field (Kourou et al. 2015), and predict a year post-discharge mortality in clinical patients with acute coronary syndrome (Sherazi et al. 2020), and cardiovascular disease prediction (Yahaya et al. 2020).

Machine learning methods for detecting CKD and Risk Prediction of Chronic Kidney Disease are examples of works in this area by applying several ML classification algorithms on the patient's medical data (Aljaaf et al.; Dulhare and Ayesha 2016; Alloghani et al. 2020) . Evaluation of clinical data is feasible using machine learning techniques to diagnose CKD early (Wang et al. 2018).

1.3 Research Problem

According to Murton et al. (2021), In 2016, CKD, a non-communicable disease, claimed the lives of 1.2 million people and caused 35 million disability-adjusted life-years (DALYs). Having a dramatic global prevalence increase since 1990 (Xie et al. 2018), shortage of trained nephrologists, and the variation in the approach of serum creatinine concentration estimation, early detection of CKD will be effective in terms of cost and improve quality of life (Jha et al. 2013). Therefore, failure to detect CKD early can result in cardiovascular disease, which exacerbates the condition (Couser et al. 2011).

On a global level, the CKD burden continually increases with more prevalence in low and middle-income countries (Wolfe et al. 1999). Over the last decade, CKD has been recognised as a main cause of death among men in hospitals in El Salvador, with Honduras and Nicaragua, being one of the most afflicted areas in the world with similar excesses reported in India and Sri Lanka (Ramirez-Rubio et al. 2013). About 36.8 million people in Nigeria have CKD and require either pricey dialysis or a kidney transplant. Many CKD patients have died as a result of the citizen's inability to bear the financial burden of treatment because of the economic crisis (Ladi-Akinyemi and Ajayi 2017).

According to Luyckx et al. (2020), various contributions have been undertaken to promote early therapy to avoid the disease from progressing to the chronic illness stage. Recent research suggested that early identification and intervention might help to avert some of the negative outcomes (Ghosh et al. 2020). For CKD identification, numerous Artificial Intelligence (AI) and Machine-Learning-Based diagnostic methodologies have been employed, including KNN, DT, RF, LR, SVM, and ANN (Maurya et al. 2019). As gap most studies use all the features in the online dataset collected from university of California Irvine (UCI). Therefore, this study will use the most important features of the same dataset using Pearson correlation feature selection to generate more accuracy and performance. As a result, the life-threatening condition will be identified early enough for therapeutic care attributable to the aims and objectives listed below.

1.4. Research Aim, Objectives, and Questions

1.4.1. Aim

This study aims to develop and compare machine learning predictive model for CKD prediction.

To achieve the research aim, the following objectives will be required:

1.4.2. Objectives

1. To review existing literature on chronic kidney disease.
2. Conduct a review of the literature on machine learning (ML) and CKD prediction models.
3. Gather relevant data about the issue and extract key features
4. To develop a model that represents the knowledge of the area.
5. To compare and select the best machine learning model based on the accuracy of performance in the classification

1.4.3 Research Questions

The following research questions will be answered by this study

1. What are the predicting features for CKD prediction?
2. Which machine learning model is better for CKD prediction?
3. Is ML adoption an influence in disease prediction?
4. Can Machine Learning improve healthcare?

1.5 Research Justification

The review of the literature revealed that there have been a lot of research efforts in machine learning for CKD detection. Therefore, in mitigating the costly risk of dialysis or kidney transplant, early detection using fewer features will minimise the cost and time of diagnosis. This study will provide numerous benefits to various groups, particularly the patients and physicians will benefit from the accurate prediction of CKD using the developed model to reduce the mortality rate.

In addition to furthering my education, the study will benefit society by improving clinical outcomes, contributing knowledge in CKD prediction using machine learning and serving as a baseline for future academics interested in doing similar research. The findings will have a huge impact on society and health ministries all around the world.

1.7 Structural Outline

This paper is organised into six chapters. The introduction, literature review, and methods sections are covered in Chapters 1, 2, and 3, while the findings, discussion, conclusion, and recommendations sections are contained in Chapters 4, 5, and 6. The reference section and appendices section follow chapter 6.

Chapter 1contains background information on the research topic as well as the study's rationale, which includes the research aim, research objectives, and research questions. The literature review covers the core topics in CKD, machine learning algorithms and related studies in chapter 2; The research approach is covered in chapter 3; The findings, i.e., the results of the various machine learning classifications, are presented in Chapter 4; Chapter 5 contains a discussion of the research findings, and Finally, Chapter 6 depicts the conclusion as well as recommendations of the study.

Chapter 2

Literature Review

2.1 Introduction

This part aids in the extraction and expansion of knowledge about the study issue including background information about the general and specialised knowledge connected to the study topic. Relevant theoretical frameworks are included in this chapter. The study will explore important ideas concerning chronic kidney disease identification processes and machine learning technologies.

2.2 Theoretical Framework and Key Concept

Predicting the future state of individual patients in terms of their vulnerability to life-threatening conditions is a critical public health issue (Waezizadeh et al. 2018). Chronic Kidney Disease (CKD) is a dangerous condition marked by the progressive loss of kidney function. However, different models have been used in a variety of fields to identify high-risk individuals with diseases in their early stages (Baumann and Sandmann 2016). Among the fundamental theories relevant for this study are deterministic models for hypertension impact on renal failure, administrative data and graph theory, Artificial Intelligence Framework stimulating clinical decision making (Markov decision process model) (Bennett and Hauser 2013; Tierney and Lanford 2016) and machine learning model. Integration of these theories is essential due to the degree of uncertainty and probability aiding the clinical decision making (Bennett and Hauser 2013).

2.2.1 Deterministic models for hypertension and stress impact on renal failure

One of the major issues in modelling clinical systems is making an accurate prediction in the system’s future. In achieving this goal, different theories have been used as an approach to tackle the uncertainty problem arising in complex system. Hypertension and stress are regarded to be the primary causes of disease, and they are used as the foundation for the models' parameters. According to Waezizadeh et al. (2018), hypertension and stress are perceived as the damaging causes of CKD through the removal of nephrons thereby reducing the GFR and generate an increased irregular protein excretion in urine (proteinuria), that result into kidney failure or loss. Thus, this mathematical model is designed to identify high risk of disease in an individual in a target population (Farkas 2001). Building accurate predictive model that use biological information is of great importance.

This model evaluates the significance of clinical data in CKD prediction modelling, particularly the CKD stability and uncertainty by applying the entropy index to check the predictions for diverse states of the disease in correspondence to the initial conditions. In clinical data studies, the deterministic model of hypertension reveals that the selection of effective factors will be efficient for predictions. Specifically, in each prediction designed for a factor, the most reasonable prediction can be determined by finding the one with the lowest entropy. This model can help to make prediction for who are examined for chronic stress and hypertension, by studying the GFR, predictions of the effect of parameters on CKD can be made as Hypertension and stress can caused a reduced number of healthy nephrons. The model is given as:

$$\frac{\mathrm{d}R}{\mathrm{d}t} = -\alpha(x)g(t)\left(1-\frac{R}{k}\right)R \qquad (1)$$

α(x) and g(t) are considered as increasing functions and their values are between 0 and 1, k is number of healthy nephrons.

2.2.2 Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach

According to Bennett and Hauser (2013), the rapid expansion of cost, complexity, information, and treatment options in the modern healthcare that do not reach the frontline affect the ability to select the optimal decisions overtime. Through artificial intelligence framework this challenge can be tackled by:

(1) Stimulating the environment to explore healthcare policy and

(2) Making an AI that can do medical tasks.

By merging a Markov decision process with a dynamic decision network, the framework learns from clinical data while capturing the synergetic interactions of many components in the healthcare system. The AI framework outperforms existing treatment-as-usual (TAU) and leads to a greater change in result. The future potential of machine learning integration for personalised medicine, an AI simulation framework can approximate optimum judgments even in complex and unpredictable environments (Bennett and Hauser 2013). As Markov model provides an efficient probabilistic inference, this framework will improve treatment decision through modelling rather than intuition (Schaefer et al. 2005). Using randomly selected 500 patients including hypertension, diabetes, chronic pain, and cardiovascular disease by handling various physician agents for chronic and acute mental disorder and response and cost per unit change, integration of machine learning and medicine will optimise treatment option and reduce the complexity and uncertainty (Bennett and Hauser 2013).

Markov Decision Processes (MDPs) are well-equipped to deal with sequential and unpredictable judgments. The rise in the incidence rate and treatment costs of chronic illnesses and cancer over the last 10 years has sparked interest in treatment optimization research. The capacity of MDP models to accommodate many outcomes makes them particularly appealing to healthcare researchers. MDPs may be used to look at a problem from the patient's perspective as well as the population's. The framework has been used in the treatment procedure for cancer prognosis in developing cheaper and more effective prognostic tools by combining classifiers and clinical real time data for data driven clinical adaptive decision support (Sun et al. 2007; Boulesteix et al. 2008).

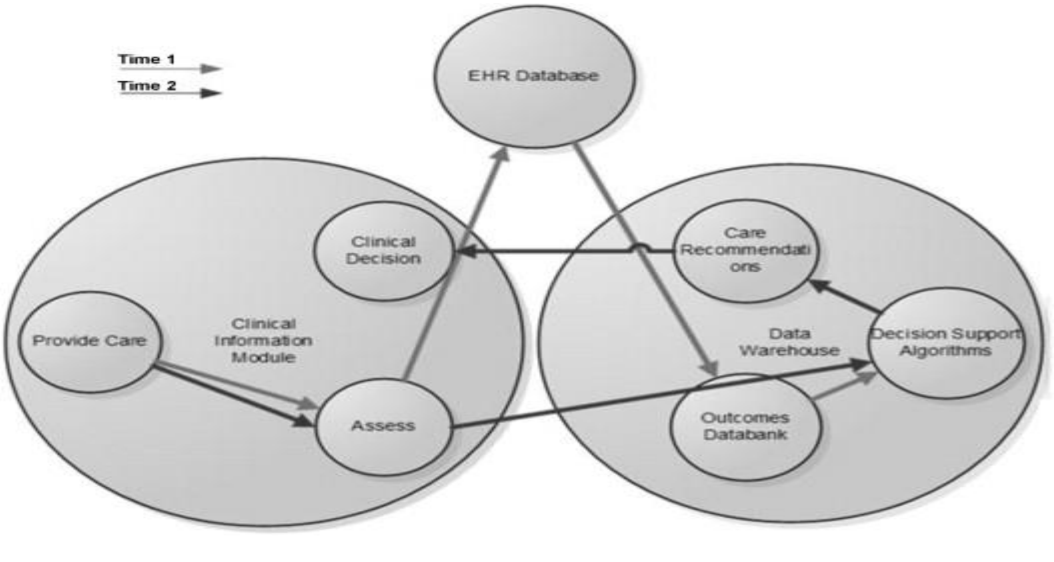


Figure 1 Clinical decision-making – flow diagram

A high-level illustration of information flow via an electronic health record (EHR) during the clinical procedure across time.

2.2.3 Administrative Data and Graph Theory

Chronic diseases have become a health and economic burden for countries globally (Yach et al. 2004). A major concern is the patient’s unawareness of disease progression and complications adding to the increased burden on limited healthcare resources (Tanno et al. 2016). The administrative data and graph theory emphasises how burden may be avoided in a large population by utilising routinely obtained administrative data and applying data mining and network analysis tools to extract potential insights regarding likely chronic illness existence and development. Particularly, data on hospital admissions and outpatient data that have been carefully gathered. The administrative data has the potential in understanding the comorbidities (Luijks et al. 2012), and disease prediction by identifying the rare patterns within the diabetes-related comorbidities (Piri et al. 2018). Khan et al. (2019), applied this theory on classifying diabetic patients by considering age, sex, and behavioural risk factors such as smoking while using binary tree classification, parameter estimation, binary logistic regression and achieved percentage that ranged between 82% and 87%.

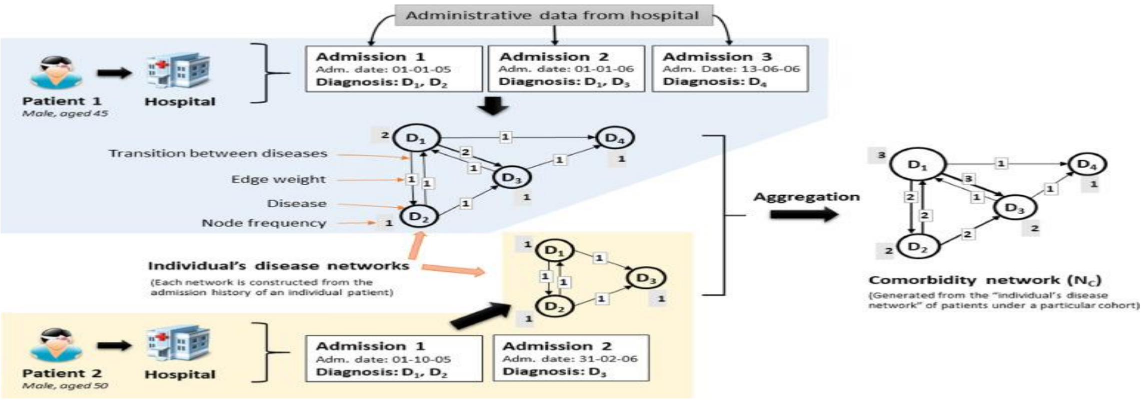


Figure 2 First component of the research framework - comorbidity network creation (Khan et al. 2019)

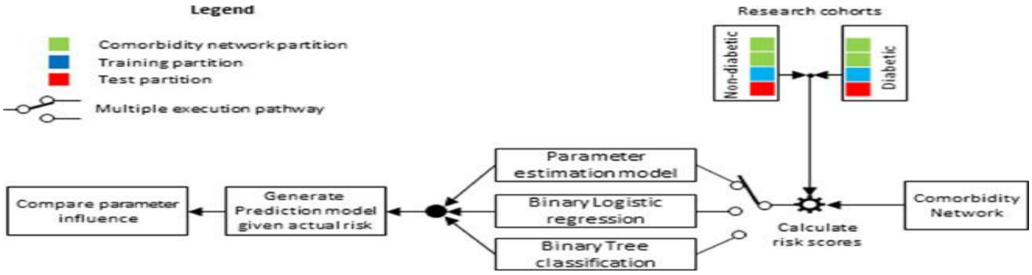

Legend
Comorbidity network partition
Training partition
Test partition
Multiple execution pathway
Research cohorts
Non-diabetic
Diabetic
Parameter estimation model
Binary Logistic regression
Binary Tree classification
Calculate risk scores
Comorbidity Network
Generate Prediction model given actual risk
Compare parameter influence

Figure 3 Prediction and performance analysis component of the framework.

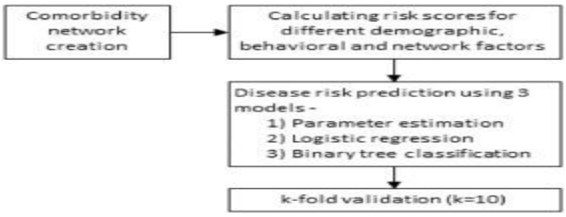

Comorbidity network creation
Calculating risk scores for different demographic, behavioral and network factors
Disease risk prediction using 3 models -
1) Parameter estimation
2) Logistic regression
3) Binary tree classification
k-fold validation (k=10)

Figure 4 Research framework for disease prediction.

2.2.4 Machine learning Framework

According to Moradi et al. (2015); Ramalingam et al. (2018); Antony et al. (2021) and Amirgaliyev et al. (2018), machine learning models have been used in disease prediction including CKD, Alzheimer and heart disease. Different supervised machine learning techniques such as support vector machine, Naïve Bayes, k nearest neighbour, random forest, and decision tree have been applied to clinical dataset for the disease prediction due to their high accuracy and simplicity. In the application of random forest algorithms, there is high hope of predicting cardiovascular diseases where the model achieved between 91.6% to 97% of accuracy at every level of feature selection. In the application of decision tree, the model performed poorly in some cases which could be due to overfitting (Ramalingam et al. 2018). For effective model performance, basic steps are followed in the algorithm development; data collection, data pre-processing by handling missing value, train/test data, model building, hyperparameter tuning and evaluation (Shah et al. 2020). These classifiers will be further detailed in the methodology chapter of this study.

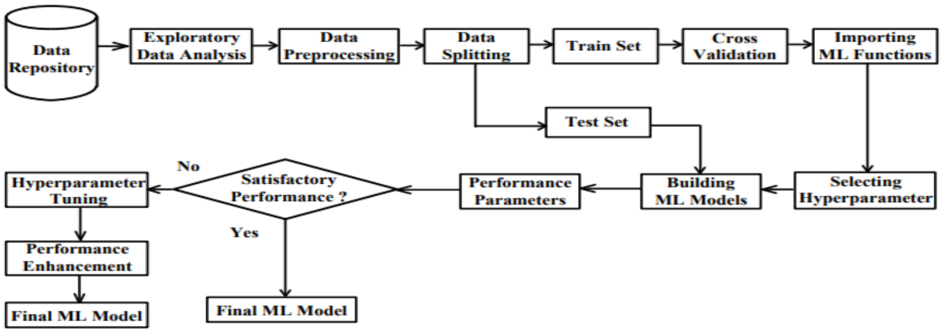


Figure 5 Sequential approach for Disease Prediction (Nishat et al. 2021)

2.3 Chronic Kidney Disease

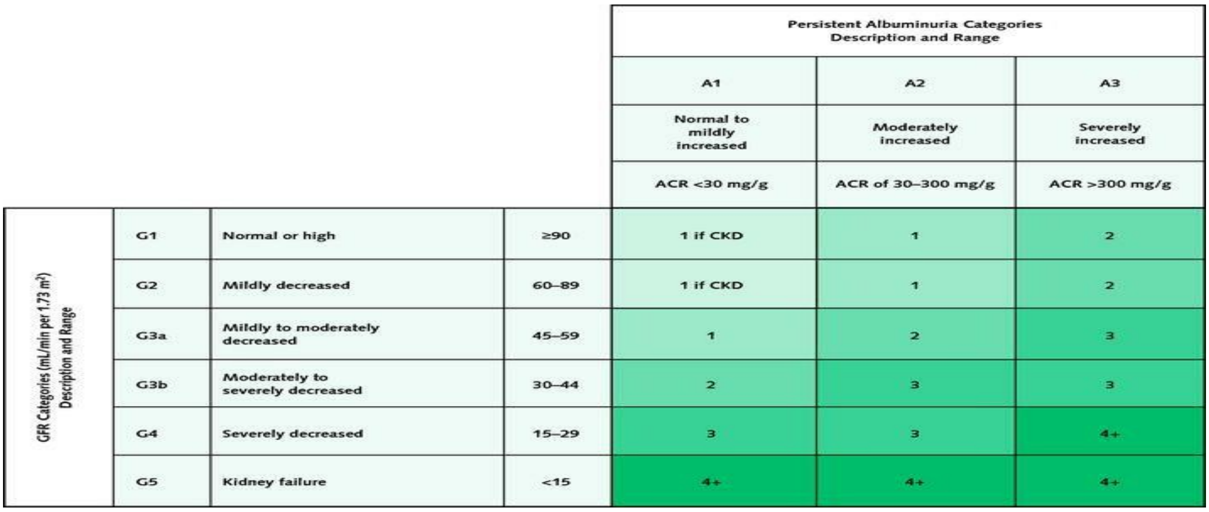
The chronic kidney disease (CKD) is described as a persistent impairment in renal function that results in the loss of functioning nephrons (Romagnani et al. 2017). The severity can be influenced by several factors, including underlying renal disorders, blood pressure, hypertension, proteinuria, and age. Early detection of CKD necessitates a great deal of attention from doctors, especially when it comes to picking the optimal time to start medical treatment and preventing disease progression to End Stage Renal Disease (ESRD), due to hypertension, proteinuria, and hyperphosphatemia effects (Rady and Anwar 2019).

2.3.1 Stages of CKD

The stages of chronic kidney disease (CKD) are primarily determined by the GFR (Waezizadeh et al. 2018), which may be calculated by the estimated Glomerular Filtration Rate (eGFR). There are 5 stages, with normal kidney function in Stage 1 and minimally impaired renal function in Stage 2 (Rady and Anwar 2019). Kidney disease is staged according to the KDOQI (Kidney Disease Outcomes Quality Initiative) as shown in table 1.

Table 1 Stages of CKD

Stage	GFR	Description	Treatment stage
1	90+	Normal kidney function, but anomalies in the urine, structural abnormalities, or a hereditary characteristic suggest renal illness.	Blood pressure monitoring and management.
2	60–89	Renal function is mildly impaired, and other results (as in stage 1) allude to kidney illness.	Blood pressure monitoring, regulation, and risk factors.
3A	45–59	Kidney function has deteriorated to a moderate degree	Blood pressure monitoring, regulation, and risk factors.
3B	30–44		
4	15–29	Kidney function is severely impaired.	preparing for end-stage renal failure.
5	<15 or on dialysis	Kidney failure that is very severe, or end stage kidney failure (sometimes call established renal failure)	Treatment options are available.



				Persistent Albuminuria Categories Description and Range		
				A1	A2	A3
				Normal to mildly increased	Moderately increased	Severely increased
				ACR <30 mg/g	ACR of 30–300 mg/g	ACR >300 mg/g
GFR Categories (mL/min per 1.73 m²) Description and Range	G1	Normal or high	≥90	1 if CKD	1	2
	G2	Mildly decreased	60–89	1 if CKD	1	2
	G3a	Mildly to moderately decreased	45–59	1	2	3
	G3b	Moderately to severely decreased	30–44	2	3	3
	G4	Severely decreased	15–29	3	3	4+
	G5	Kidney failure	<15	4+	4+	4+

Figure 6 CKD Persistent Albuminuria Categories

As defined by Levey et al. (2003), a person is considered chronic upon the collection of two samples at least 90 days apart. Creatinine measurement, sex, race, and age all influence the eGFR (Rady and Anwar 2019).

In many situations, the development and complications of CKD are associated with a substantial mortality risk and show no symptoms at first but usually detected during the examination of concomitant conditions (Romagnani et al. 2017). A rapid progression of CKD can lead to renal failure in a matter of months (Levey and Coresh 2012). The CKD symptoms vary depending on the stage. The treatment options include slowing or stopping the course of the illness, and controlling CKD risk factors

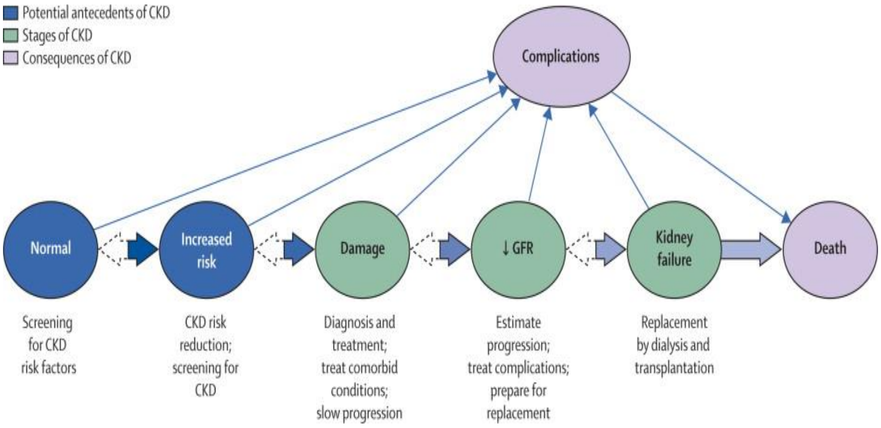

Potential antecedents of CKD
Stages of CKD
Consequences of CKD
Complications
Normal
Increased risk
Damage
↓ GFR
Kidney failure
Death
Screening for CKD risk factors
CKD risk reduction; screening for CKD
Diagnosis and treatment; treat comorbid conditions; slow progression
Estimate progression; treat complications; prepare for replacement
Replacement by dialysis and transplantation

Figure 7 Conceptual model for chronic kidney disease

2.3.2 Risk factors of CKD

All phases of CKD are most frequent in adults over the age of 65, while younger people have a larger risk of developing ESRD (≤65 years of age) with more prevalence in female than male, men are more to developing ESRD (Hill et al. 2016). Diabetes mellitus and hypertension are the most frequent underlying illnesses linked to CKD, especially in high- and middle-income countries. The incidence of CKD in people with diabetes is thought to be 30–40%. As described by Stanifer et al. (2017), infectious diseases, glomerulonephritis, and improper medication use such as traditional remedies with potential nephrotoxins are all connected to CKD. The number of people with CKD is expected to rise, and the diabetes and obesity epidemic may eventually overtake smoking as the major cause of CKD due to socioeconomic trends and ageing. Acute kidney injury (AKI), a highly prevalent sudden deteriorating state of functional kidney due to accumulation of metabolic waste and toxin causing loss of nephron number is a risk factor of CKD (Barton et al. 2015). As systemic hypertension is communicated to intraglomerular capillary pressure, a patient's high blood pressure and stress can damage the filters and induce harm to the glomeruli, leading to glomerulosclerosis and kidney failure (Kazancioğlu 2013). In the United States, half of those with ESRD have diabetes. Diabetic individuals are at risk for CKD due to nephropathy and hyperfiltration damage (Lea and Nicholas 2002). In a patient with an obstructed and infected urinary tract, too many sediments, such as salts (calcium), with improper solubility in the kidney can develop kidney stones, which can be addressed with prompt surgical intervention in a patient with deteriorating renal functions (Han et al. 2015).

2.3.3 Diagnosing and CKD Management

CKD is an asymptomatic disease with no symptoms at the early stage of poor renal function. As a result, detecting early clinical concern with the kidney is challenging. Many people's renal function deteriorates before symptoms develop. Therefore, it is critical and beneficial to treat early and take measures to avoid CKD. A variety of tests, including as blood tests, urine tests, haemoglobin, potassium, and sodium, should be performed by professionals to detect and diagnose CKD early. CKD management attempts to maintain kidney function and homeostasis for as long as feasible by addressing any underlying conditions, slowing the progression, and lowering the risk of developing cardiovascular disease and renal failure(Stevens et al. 2013).

There are a variety of therapies and technologies available to identify it, lower the risk of consequences, and limit the disease's course. Treatment for CKD aims to reduce the risk of kidney disease and keep it from advancing to the end stage. There are therapies for anaemia, phosphate balance, high blood pressure control, diabetes control, and more.

2.4 Machine Learning (ML)

Machine learning (ML) is an artificial intelligence subfield that learns from data to develop inference concepts (Mahesh 2020). ML uses estimated dependencies to anticipate future system outputs by estimating unknown system dependencies from a given dataset (Aslam et al. 2021). ML has also shown to be a fascinating area of biomedical science with several applications by exploring an n-dimensional space for a collection of clinical samples using various algorithms (Niknejad and Petrovic 2013). The widespread use of ML in the medical area aids in the promotion of medical innovation, the reduction of medical expenditures, and the improvement of medical quality. However, further research into using machine learning to solve clinical problems in nephrology is needed. Understanding the aim and technique of ML application, as well as the present state of its use in nephrology, is necessary for properly addressing and overcoming these issues. Machine learning has previously been used to identify human health status, assess disease-related characteristics, and diagnose a variety of illnesses including cancer, heart disease, diabetes and retinopathy, acute renal damage, and other disorders.

In clinical innovation, machine learning has been utilised in a lot of clinical prediction model research (Motarwar et al. 2020) by capturing vast volumes of information on individual patients that are difficult for human to comprehend and transforming the healthcare industry (Chen 2020). Machine learning allows healthcare providers to advance toward individualised care, often known as precision medicine, by automatically finding patterns and reasoning about data and improve patient outcomes through predictive analytics, resulting in more accurate diagnosis and treatment as well as improved clinician insights for tailored and cohort therapies (Chowriappa et al. 2014). According to diverse modelling demands, ML technology may be categorised into supervised learning, unsupervised learning, semi supervised learning, and reinforcement learning.

2.4.1 Supervised Machine Learning

Supervised machine algorithm consists of the target/outcome (dependent) variable to be predictor from the collection of predictors (independent variables). In supervised learning, a labelled set of training data is used to estimate or map the input data to the expected result (Hana 2021). The model is trained until it reaches the appropriate degree of accuracy on the training data. Regression, Decision Tree, Random Forest, KNN, Logistic Regression, and other examples of supervised learning (Nasteski 2017).

2.4.2 Unsupervised learning

In unsupervised learning approach the data is not given a labelled example and does not consider the output throughout the learning process (Lee et al. 2022). As a result, finding patterns or discovering groupings of input data is up to the learning model. This approach may be regarded of as a classification challenge in supervised learning. A learning process that categorises data into a collection of finite classes is referred to as classification. It is commonly used for segmenting customers into distinct groups for activities, as well as other fields like as probability density estimation, discovering associations among variables, and dimensionality reduction (Chen et al. 2020). Unsupervised Learning Examples: K-means algorithm, apriori algorithm. Regression and clustering are two more popular ML problems. A learning function transforms the data into a real-value variable in regression situations. Clustering is a frequent unsupervised activity in which the goal is to discover categories or clusters of data objects to explain them.

2.4.3 Semi Supervised Learning

Semi-supervised learning, which combines supervised and unsupervised learning, is another form of ML approach that has been frequently used. It creates an accurate learning model by combining labelled and unlabelled data. When there are more unlabelled datasets than labelled datasets, this method of learning is typically utilised (Nasteski 2017).

2.4.4 Reinforcement Learning (RL)

Here, the algorithm develops a policy on how to act by observing the environment, and the environment provides feedback and rewards to the learning algorithm, which define the agent's purpose in a particular situation (Sutton and Barto 1999). This system learns from its previous experiences and attempts to acquire the most relevant information to make appropriate decisions. Robot vision and movement, automatic chess player, and autonomous car driving are just a few of the decision-making problems that RL is employed to solve. Reinforcement learning is exemplified via the Markov Decision Process.

Many of the previous studies examined for this study used Decision Trees (DT), Random Forests (RF), and Artificial-Neural Networks (ANN), and most used various classification algorithms to develop CKD prediction models (Amirgaliyev et al. 2018; Khan et al. 2019; Ghosh et al. 2020; Almustafa 2021; Nishat et al. 2021) .

Table 2 The advantages and disadvantages of machine learning classification methods

classifiers	Benefits	Shortcomings
Naïve Bayes Classifier algorithm	Simple, quick, and easy. Not Sensitive to non-essential characteristics. Work with both continuous and discrete data. In practise, less training data is needed for binary and multi class classification	Takes every feature as independent
KNN	Simple to comprehend and adopt. Training time ranges from none to very little. Simple to use with multiple data sets. high degree of predictability. Perform admirably in reality.	Stage of testing that is computationally intensive. It is possible to have a skewed class distribution. When dealing with high-dimensional data, accuracy might suffer. A value for parameter k must be defined.

Table 3 The advantages and disadvantages of machine learning classification methods

Decision Tree	Simple to comprehend Rules are simple to create.	Overfitting might be a problem. Working with non-numerical data is difficult.
	There are just a few hyperparameters that can be tweaked.	In compared to other methods, it has a low prediction accuracy.
	The representation of complex decision tree models can greatly simplify them.	Calculation might be difficult when there are several class labels.
Random Forest	There is no such thing as an overfitting issue.	Complexity necessitates many computational resources.
	It may be used for feature engineering, which is the process of determining the most significant characteristics from many options.	It takes a long time.
	On a large dataset, it performs well.	Needs to decide on the variety of trees.
	Extremely adaptable with a great degree of precision	
	There is no need to prepare the input data.	
Neural Network	For situations involving regression and classification	Difficult on variable influence due to black box in neural network
	Modeling nonlinear data with a high number of inputs is useful for everything from speech recognition to cancer prediction.	Costly and intensive computational training with CPU
	High performance when classification challenge is broken down to layers of smaller pieces	Training data is extremely important for neural networks.
	Predictions work well with any additional layers once they have been trained.	
Support Vector Machine	Algorithms that are quick	Does not work well with massive amounts of data.
	In high-dimensional space, it is effective. It is also very precise.	It is not easy to programme.
	Kernels provide power and versatility.	Poor performance with noisy data as the target classes overlap.
	When there is a definite dividing margin, it works well.	
	There are several applications.	

The study aims to construct a predictive model using many supervised learning techniques, including Decision tree, Random Forest, Support vector Machine, K Nearest Neighbour, Naive Bayes, Logistics regression and Artificial Neural Network to predict chronic kidney disease.

2.5 Machine Learning in Healthcare

In the contemporary world, data across industries is growing exponentially. As the volume of data increases, new novel ways of extracting meaningful insights from the data are emerging. ML is playing a significant role in many fields like finance, medical science and in security (Shailaja et al. 2018). The incapability of human to find hidden patterns in large data is a challenge particularly in healthcare sector. ML as a method of analytics automate model building by using history to predict future events to predict diseases. In recent times, large amount of data has become available in healthcare including the electronic medical record (EMR) that contains both structured and unstructured form. Most of the medical data are unstructured in the forms of notes, discharge summary and images and since medicine is a story, new machine learning approaches must focus on organising and finding relationships between large volumes of unstructured data. The ability to collect and analyse this type of data on a wide scale will be highly useful when using machine learning technology in the healthcare field. When used effectively, machine learning may assist physicians in making near-perfect diagnoses, selecting the appropriate prescriptions for their patients, identifying individuals at high risk for poor pharmaceutical results, and improving patients' overall health while lowering costs (Devi and Balasubramanian 2022). Machine learning has been shown to assist diagnose patients and identify people who are more prone to repeated diseases. Furthermore, about 90% of emergency department visits may be avoided. Machine learning is used to better diagnose and send patients to appropriate therapy while also lowering expenses by avoiding the usage of expensive, time-consuming emergency rooms.

2.6 Related Work

The adoption of machine learning to solve health issue is increasingly gaining relevant attention and sizable research. This attention has been driven by the application of various machine learning models for disease prediction. With notable contributions, classification of the disease had been done on the CKD data by many researchers to provide a better diagnostic benefit. Using data mining classification on CKD dataset, Alasker et al. (2017), implemented ML classifiers including Naïve Bayes (NB), Neural Network(NNs), Decision Trees(DT), K Nearest Neighbour (KNN) and one rule classifier, with Naïve Bayes giving a better prediction accuracy support of 99.36% and 0.977 sensitivity than other classifiers.

Similarly, to find hidden patterns in our dataset, Gharibdousti et al. (2017) used varying features of the CKD data to understand the significance of features on classification results. As the number of features is reduced, a decrease in accuracy and a corresponding decrease in Area under curve (AUC) across the DT, NN, LR, SVM and NB ML techniques applied is observed upon the validation.

By the work of Polat et al. (2017), the detection of GFR was improved to detect CKD patients. The author used support vector machine by feature selection including wrapper and feature approach to minimise the dimension of data sets to 13 features. An accuracy of 98.5% of SVM was achieved compared to other selected techniques.

Lakshmi et al. (2014) applied ML tools on kidney dialysis survivability by using artificial neural networks, Decision Tree and Logistic Regression with ANN having the best accuracy of prediction of 93.521%.

Shrivas et al. (2018) compared the accuracy, sensitivity, and specificity of multilayer preceptor network and radial-basis function network, using five different feature selection techniques: chi-square, information gain, symmetrical uncertainty, gain ratio, and relief-f on the two classification algorithms. Multilayer preceptor network has 97 percent of accuracy and radial-basis function networks, with an accuracy of 98.5 percent.

Tazin et al. (2013) applied NB, KNN, SVM and DT on CKD dataset from UCI repository using ranking algorithm for the selection of 10 important features. Root absolute error and a ROC curve was used to calculate the algorithms accuracy where DT gave the highest accuracy of 99.75% followed by SVM of 97.75% using WEKA as data mining tool. The author suggested further application of ANN and fuzzy logic to analyse their accuracies.

Automation of CKD diagnosis was proposed by Akben (2018) by applying k-means for the data pre-processing and utilizing KNN, NB and SVM as classifiers. The highest accuracy of 97.8% of model performance was obtained, while a tremendous increase in accuracy was made using the same features from age group 35 and above, the study reveals features combination has a direct effect on accuracy ranging from 83.75% to 97.8%.

Ani et al. (2016) adopted different machine learning techniques while using KNN model to develop a decision support system because of its high percentage of 90% performance accuracy through random subspace. This will help clinicians for accurate and better diagnosis. The author suggested the future adoption of other classifiers like kernel and ANN in a parallel computing environment.

Using the UCI repository dataset, the application of clinical data and the validity of three multivariate models in CKD patients' clinical risk assessment were given by (Chen et al. 2016). K-Nearest Neighbour, Support Vector Machine, and Soft Independent Modeling of Class Analogy were the three multivariate models employed in this study. With an accuracy of over 93 percent, the patients were classified as having CKD or not having CKD by the authors. This study also found that SVM handled interference in the combined dataset better than the other two models and had the best CKD predictive accuracy, with 99 percent.

Vijayarani et al. (2015) examined the performance of Support Vector Machine (SVM) and Artificial Neural Network (ANN) in terms of accuracy and execution time (ANN). Based on the results of the experiment, it can be concluded that ANN 87.70% with the highest classification accuracy outperforms SVM 76.32% with minimum execution time.

A prediction algorithm for heart disease and renal failure has been developed by Chaudhary and Garg (2014). The A-priori and k-mean algorithms are used in this system. For the prediction, 42 attributes were considered. For data analysis, they employed ROC and calibration charts. The result of the two data mining techniques shows K-means clustering make system more accurate and can aid the classifier's performance in disease diagnosis when compared to the weighed association with A-priori algorithm. However, the integration of A-priori and K-means algorithm will help doctors in heart and kidney disease diagnosis.

Almansour et al. (2019) used different ML classifying algorithms to dataset of 400 patients using best obtained Artificial Neural Networks (ANN) and Support Vector Machine (SVM). The experiment was performed using WEKA. The results demonstrate that the study's conclusion is impressive, with ANN (accuracy: 99.75%) outperforming SVM (accuracy: 97.75 percent) while using 12 features.

Chatterjee et al. (2017) proposed the Cuckoo search (CS) trained Neural Networks (NNs) or NN-CS model to combat the financial concerns and future implications caused by CKD in emerging and undeveloped nations. This NN-CS model solves the challenge of training NNs using local based learning methods. When compared to well-known classifiers like the Multilayer Perceptron Feedforward Network (MLP-FFN), the NN-CS-based model is more effective in detecting CKD than any other model currently available.

Nusinovici et al. (2020) compared the logistics regression performance with other machine learning techniques using four features including, hypertension, diabetes Meletus, cardiovascular disease and the class. With logistics regression having a high percentage of accuracy using a 10033 sample size. The authors claimed where the number of incidences is low, ML superiority may not be assured compared to conventional regression.

Mohammed Siyad et al. (2016) used fusion and genetic algorithm feature selection to lower computational cost and achieved 100%, 99% and 98.25% of random forest, naïve bayes and logistics regression accuracy respectively. The authors utilised 6 selected features including haemoglobin, packed cell volume, specific gravity, hypertension, serum creatinine, diabetes Meletus from the 400 UCI CKD dataset.

Other researchers in the literature have utilised similar methodologies to identify diverse applications other than CKD-related work using the NL classifiers. Some of these applications are health related. Sharmila et al. (2018) used KNN proposing a hybrid ML method to detect epileptic seizure. For forecasting a chaotic crude oil time series, other non-health-related applications employed a feature selection technique (Karasu et al. 2020), for wind speed predictions, a neural network hybrid model was deployed (Altan et al. 2021). A lot of ML were applied to predict solar radiation(Hacioğlu 2017).

Table 4 Taxonomy of some related works

Author	Title/idea	Number of features	Methodology and Accuracy	Limitation/drawback
Almansour et al. (2019)	Utilising ML for early CKD diagnosis	24 features, UCI data, 400 patients	ANN 99.75% SVM 97.75%	Further research on deep learning
Yashfi et al. (2020)	Risk of CKD prediction using machine learning	455 patients UCI data + real time dataset	Used python , 10-fold cross validation RF 97.12%, ANN 95.5% Ch-square filter	Data collection, missing value. Inability to use sufficient real-life data
Salekin and Stankovic (2016)	CKD prediction with relevant feature selection	400 UCI data 24 features 14 features	KNN, RF, ANN, Wrapper approach Embedded approach, Cost analysis	Data collection
Rady and Anwar (2019)	Predicting kidney disease stages using data mining	361 patients	MLP 2second processing time. Probabilistic Neural Network- 96.7% best model with 12 seconds' computational time SVM, Radial Basic function	Sample size
Polat et al. (2017)	CKD diagnosis using SVM	400 UCI dataset	Wrapper, greedy step wise and filter approach, SVM 98.5%	Sample size

Despite contributions to CKD prediction using machine learning and feature selection from the previous work, much research do not apply the identified theories including the deterministic, and administrative and graph theory, and only few work in the literature has considered hypertension and diabetes as predictive features of CKD and only few have evaluated the models' accuracy and give the true positive or true negative value. These gaps are the study's key contribution, which we hope will assist society and healthcare by recognising some CKD patients at an early stage by utilising the Pearson correlation described in the paper's feature selection section. Some of these frameworks have been applied on diabetes classification using patients' behavioural characteristics. This study will integrate the frameworks on CKD using observed laboratory clinical data and seven different machine learning techniques to assess their performance and propose the optimal CKD prediction model based on the accuracy and AUC score.

Chapter 3
Research Methodology

3.0 Introduction

The philosophical underpinnings of the research are described in this part, as well as the research strategy used. The rationale for using the positivism paradigm as the research philosophy is explained. The explanation for why the study issue and certain research procedures were chosen as suggested by (Bell and Waters 2018) is presented. In addition, this chapter included thorough details on the data utilised and the data source. The justification for using secondary data from an open-source repository is presented. The techniques of analysis, as well as how and where the data was acquired, are all discussed. The tools that would be used to analyse the data, as advised by Denscombe, were also stated (Denscombe 2012). In Chapter one, our study topic was stated as well as our study aims and objectives.

3.1 Research Onion

The research onion, as proposed by Saunders et al. (2009) and represented in figure 9 below, was employed in this study. The issue underlying the data collecting technique and it is discussed here how it relates to the rest of the study onion's layers, such as research philosophy and knowledge generation process, research strategies, and methodological choice.

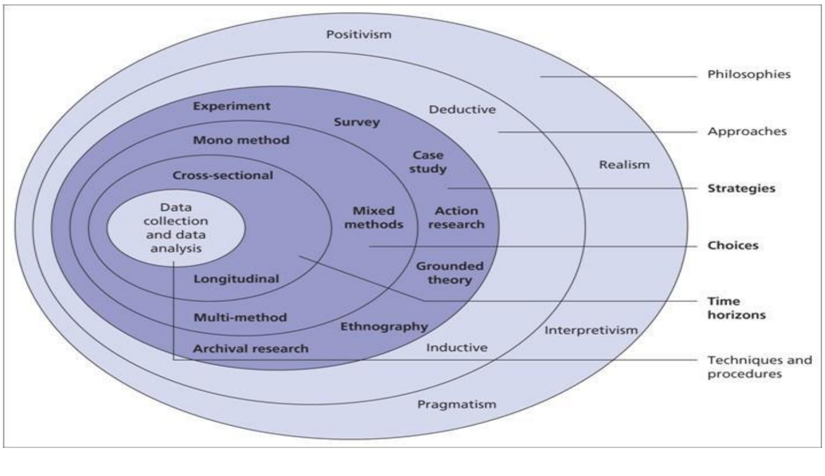


Figure 8 The research Onion (Saunders et al. 2009)

3.2 Research Philosophy

A research philosophy is a set of beliefs about how data about a phenomenon should be collected, analysed, and applied. The word epistemology (what is known to be true) rather than doxology (what is thought to be true) refers to the many study approaches. The process of changing what is believed into what is known is what science is all about: doxa to episteme. In the Western heritage of science, two fundamental research ideologies have been identified: positivist (also known as scientific) and interpretivist (also known as antipositivist) (Walsham 1995). According to Saunders et al. (2009), selecting the most relevant research underlying philosophy leads to selecting the best methodologies to use in this study. Depending on the ontology, epistemology, or axiology used, each research philosophy is unique.

Ontology is concerned with assumptions about the nature of reality (Saunders et al. 2009), while epistemology is concerned with knowing assumptions including what constitutes acceptable, valid, and true knowledge, as well as how information is delivered (Burrell and Morgan 2017). Axiology is concerned with ethics and value roles. While it may be unavoidable for a researcher to include his or her beliefs into the study process, it is critical that researchers reflect on them while doing their research (Heron 1996). As a result of the above, this study's ontology is Nave realism, which claims that reality can be comprehended using acceptable means (Moon and Blackman 2014).

For this research, positivist philosophy provides the best intellectual grounding. The research will be carried out utilising quantitative research methodologies because of the nature of data collected and a deductive methodology. The study will be based on applicable theory, which will drive the creation of hypotheses that are most appropriate for the research issue. The position will be approved or refused based on the results of the data analysis (Leavy 2017).

3.3 About the Data source

The data for the study was taken from the University of California, Irvine's Machine Learning repository (UCI). It contains a collection of datasets, domain theories, and data generators that the machine learning community regularly uses throughout the world to experimentally analyse various methods. It was founded by David Aha in the late 1980s and is a widely used database among academics and key source of machine learning datasets.

3.3.1 Source of Research Data Collection

This study's data comes from a machine learning data repository, making it a secondary data source based on previous research by (Bellomo et al. 2012; Aljaaf et al. 2018; Wang et al. 2018). The attributes and further information of the data will be discussed further down.

3.3.2 Dataset Attributes

The data for this study comprises 400 columns (400 instances) and 25 rows (24+ target variable "class") of unique variables, it may be considered a small dataset with 11 numerical and 14 nominal measurements from blood and urine test. The dataset is divided into two classes: CKD and NotCKD, which correspond to individuals with CKD and those who do not have CKD.

Table 5 Feature Description of Dataset

Attribute	Acronym	Used Value		Type
1	Age	age	years	numerical
2	Blood Pressure	bp	mm/Hg	numerical
3	Specific Gravity	sg	1.005, 1.010, 1.015, 1.020	nominal
4	Albumin	al	0, 1, 2, 3, 4, 5	nominal
5	Sugar	su	0, 1, 2, 3, 4, 5	nominal
6	Red Blood Cells	rbc	normal, abnormal	numerical
7	Pus Cell	pc	normal, abnormal	nominal
8	Pus Cell clumps	pcc	present, not present	nominal
9	Bacteria	ba	present, not present	nominal
10	Blood Glucose Random	bgr	mgs/dl	numerical
11	Blood Urea	bu	mgs/dl	numerical
12	Serum Creatinine	sc	mgs/dl	numerical
13	Sodium	sod	mEq/L	nominal
14	Potassium	pot	mEq/L	numerical
15	Hemoglobin	hemo	gms	nominal
16	Packed Cell Volume	PCV	mEq/L	numerical
17	White Blood Cell Count	wc	cells/cumm	nominal
18	Red Blood Cell Count	rc	millions/cmm	numerical
19	Hypertension	htn	yes, no	nominal
20	Diabetes Mellitus	dm	yes, no	nominal
21	Coronary Artery Disease	cad	yes, no	nominal
22	Appetite	appet	good, poor	nominal
23	Pedal Edema	pe	yes, no	nominal
24	Anemia	ane	yes, no	nominal
25	Class	class	ckd, notckd	nominal

3.4.3 Data Processing

Machine learning project relies heavily on data that can be noisy and erroneous and of poor quality, resulting in a poor mining result for the model. Data pre-processing might result in an unexpected increase in model accuracy. Therefore, undertaking data cleaning is important in managing missing values and converting nominal values to numbers and removing outliers in the CKD dataset. To clean the dataset, the following data architecture was followed.

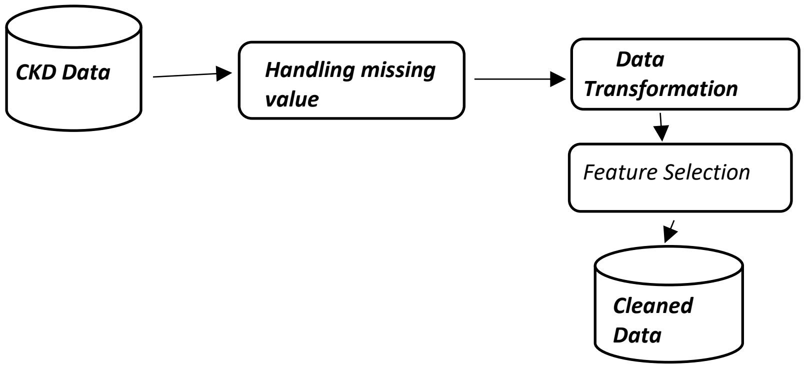


Figure 9 CKD data processing architecture

3.4.4 Data Cleaning

The data cleaning procedure ensures that the data in the actual world is accurate and consistent. Machine learning techniques necessitate a logical data structure, with each data sample following the same pattern. Algorithms are unable to handle even slight irregularities in data unlike the human brain.

3.4.5 Handling missing value

Some of the findings include missing values because of data encoding errors. The dataset was constructed using data acquired from patient lab test. The mean technique was used to deal with missing variables in this study.

According to Schmidt et al. (2015), medical data is frequently affected by missing and/or corrupted data. It happens when a variable in an experiment or test has no saved data value. The lack of data is a typical event that can have a big impact on the conclusions formed from it. Many missing values are common in medical data, making analysis challenging for researchers to develop a model using it. For a variety of reasons, medical data has missing values; for example, it is often difficult to acquire comprehensive data owing to personal privacy concerns (Huang and Cheng 2020). Algorithms have a difficulty capturing patterns in medical datasets because of these drawbacks. Missing data must be dealt with carefully since any conclusion based on a dataset with non-random missing values may be skewed (Antony et al. 2021). This study used the Imputation approach by replacing the missing data with the mean, which entails substituting approximation values for missing data.

3.4.6 Handling Categorical Data

Some categorical values in the produced dataset need to be changed to the form 1 and 0. As a result, all categorical values are transformed by this transformation task. In this investigation, categorical data was handled using dummy variable encoding and the replace approach. Data standardisation can increase model's accuracy. As a result, all nominal data were converted into numerical data as shown in table 6

Table 6 Feature Normalisation

Attributes	Category	Transformation
htn	yes	1
	no	0
dm	yes	1
	no	0
Anemia	Yes	1
	no	0
Artery disease	yes	1
	no	0
CKD	yes	1
	no	0

3.4.7 Feature Selection

Data or dimensionality reduction is applied to reduce the input variables to ML model by identifying the most important features in the data (Antony et al. 2021). When creating a predictive model, feature selection is the process of minimising the number of input variables (Huang and Cheng 2020). To improve the accuracy of our model, the research found related characteristics from a collection of data and remove less significant information that do not contribute much to our goal variable. Filter technique, Wrapper method, and Embedded approach are some of the many types of feature selection methods used by researchers.

According to Jain and Singh (2020), the most difficult aspect of the feature reduction technique is identifying the appropriate subset of characteristics to produce the best classification result. This study will use the relevant concept discussed in chapter two to hypothesise the importance of the selected features.

Correlation-based technique is the primary metric for dependency measurements. The Pearson's Correlation approach is used to determine the relationship between independent and dependent variables. The correlation approach is used in this study to choose features on threshold 0.5 where 1 is positive correlation and -1 is negative correlation for the binary classification. The Correlation approach examines all qualities associated to the target variable using Pearson's correlation approach, which ranks the attributes from high to low, reducing processing time and data dimension (Bahadir 2016).

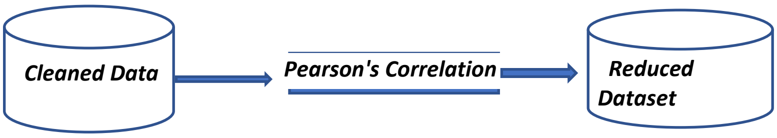


Figure 10 feature selection with highly correlated features

3.4.8 Scaling the Features

Feature scaling is an approach for normalising or standardising the range of independent variables characteristics, or for putting features on equal standing. Some machine learning methods, such as SVM, KNN, and NN, are affected by feature scaling (Pinto et al. 2020). Then, before employing in any model construction, we must scale features. The dataset characteristics of a health-care system may vary in scale, and unit, and then many machine learning algorithms are based on the measurement magnitude rather than the unit. The min-max method keeps features in a defined range of [0, 1].

3.5 Machine Learning Methods

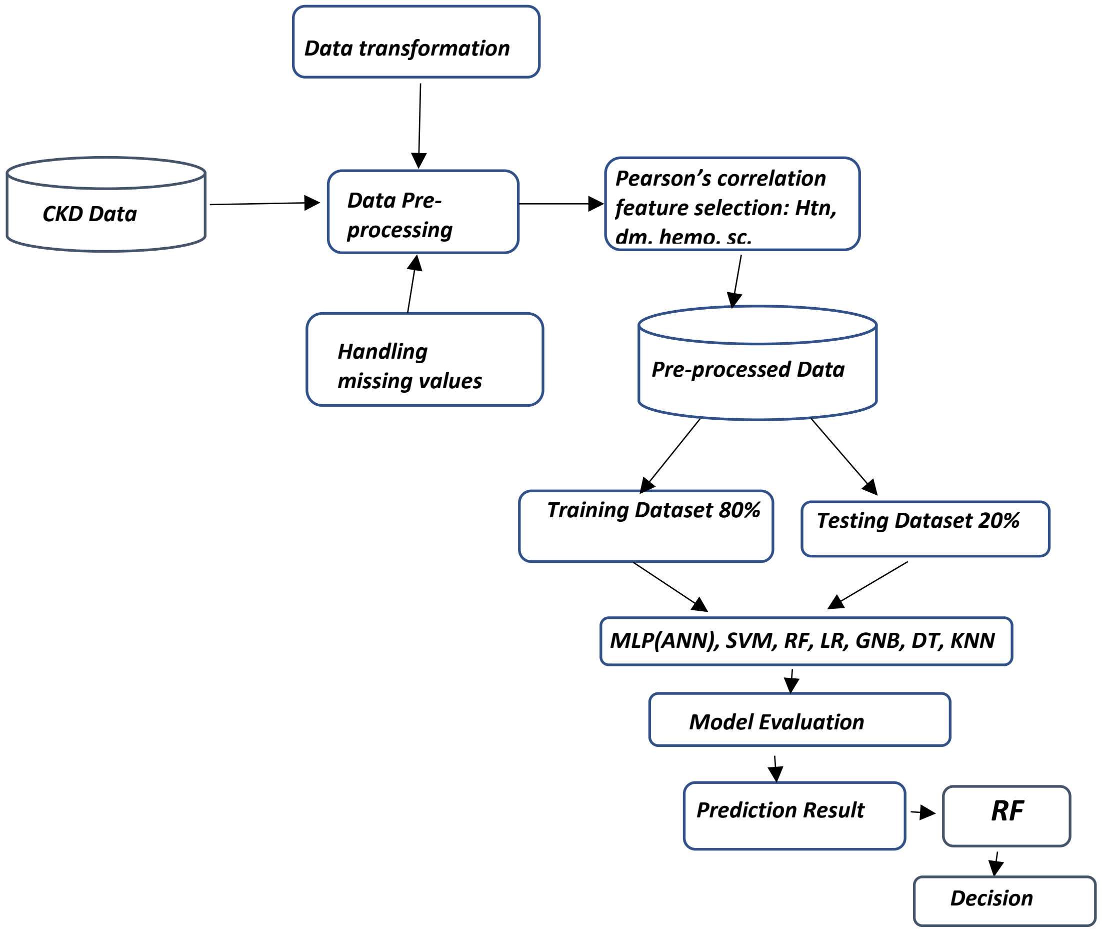
The goal of this study is to develop and compare machine predictive models for CKD prediction using CKD data that can predict if a person has chronic kidney disease. This objective was achieved by comparing the performance of 7 machine Learning classifiers including RF, KNN, ANN, Naïve Bayes, LG, and SVM. The algorithms were employed to compare the performance of the prediction models because of their usefulness in the prediction of chronic renal disease and their classification performance in prior research studies.

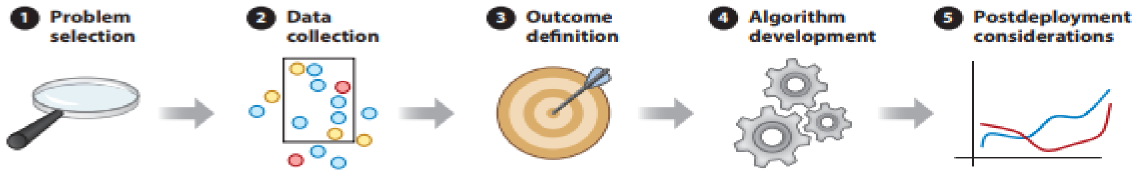
Step by step method for clinical machine learning

1. Identifying a clinical problem and obtaining the relevant to justify the solution.
2. Process the data
3. Data splitting into training and testing

4. Algorithm selection
5. Optimise the hyper-parameter
6. Handle overfitting
7. Model evaluation

Figure 11 Proposed Method





3.5.1 Identifying a clinical problem and obtaining the relevant data to justify the solution

This is one of critical problems to handle in machine learning project. Finding a viable research question typically necessitates a decent level of domain expertise in healthcare, and running a viable machine learning project on clinical data sets necessitates a significant degree of technical expertise and skill set in computers. These two skills sets in healthcare and computers is important for project execution (Domingos 2012).

Choosing the right tools in answering the research questions is important particularly if the data set are labelled and with a grand truth, then, supervised learning will be suitable(Tarca et al. 2007)

3.5.2 Check and arrange data properly

Data availability and understanding of the dataset are important in a machine learning project. According to Chicco (2017), the question of data availability, understanding and suitability is important. The data was cleaned to make it readable for the machine (Domingos 2012).

3.5.3 Splitting the data into independent subsets

Running a machine learning project requires the traditional method of splitting the dataset set into 4:1. Splitting the data into train and test was efficiently followed by cross-validation will (Refaeilzadeh et al. 2009) as well as optimised the hyper-parameter. Following the training of the model, the model is tested with test data and evaluated through the outcomes Powell et al. (2020) used the term "lockbox approach" to describe this method to check the success and failure in ML.

3.5.4 Choosing the right Algorithm

Choosing the right algorithm depend on different factors including cost evaluation, problem presentation and performance optimisation. According to Domingos (2012), supervised machine learning performs well on labelled data, and since the dataset is labelled, this study will use supervised machine learning. However, where choosing an ML algorithm is undecided, the simplest is preferred and complex algorithm be chosen if the data give reasonable justification for such selection (Hand 2006).

3.5.5 Optimising hyper-parameter:

Hyper-parameters are higher-level features in machine learning that can have a significant impact on the algorithm's statistical model's complexity, learning speed, and result application. The hyper-parameters must be defined before the training process begins.

3.5.6 Minimising overfitting

When a machine learning system learns correctly from the training dataset but performs badly on the validating and testing datasets, this is known as overfitting. Cross-validation and regularisation are two new potent techniques for dealing with the problem of overfitting. Cross-validation is a method that allows the trained model to learn from each data fold rather than overfitting to a given training dataset. This is accomplished by accumulating penalization values that rise in proportion to the weight of the learnt parameter (Neumaier 1998; Cogswell et al. 2015).

3.6 Model Evaluation

In the machine learning process, this is a critical step to evaluate the model and ensure it fits the dataset and works well with the completely new dataset. The goal of model performance evaluation is to determine a model's generalisation accuracy on unknown/out-of-sample data. There are two types of performance evaluation methods: holdout and cross-validation. The goal of holdout assessment is to put a model to the test on different data than it was trained on to provide an impartial estimate of learning performance. In holdout, mothed data is separated into three sets: training, validation, and test.

Hold-out is easier to use, more versatile, and faster than cross-validation. However, because variations in the training and test datasets might cause performance variances, this technique has a high level of unpredictability.

Cross-Validation is a performance evaluation approach that divides data into divisions, one of which is used to train a model and the remainder of which is used to test or validate the model. The cross-validation approach is the most often used assessment technique to minimise overfitting. To ensure that the models obtained the pattern for the training dataset for this study, we employed the most common K-fold cross validation assessment method.

3.6.1 Evaluation Metrics for Model Performance

Classifier methods do not provide the same results, prediction model assessment criteria are necessary to assess model performance. The test data set is used to assess the machine learning model's performance using statistical scores. Various assessment measures were utilised to check the performance of the classifiers in this study. Due to the dataset's two classifications, it's a 2*2 matrix. cross-validation was utilised to evaluate performance indicators such as accuracy, recall, confusion matrix, and f1-score, sensitivity, and specificity. Accuracy, precision, recall, F1 score, and AUC are all common machine learning assessment measures.

The following are some of the most keywords in performance measurement:

TP: True Positive indicates that the output is positive, and the projected result is appropriately identified.

True Negative (TN) denotes that the output is negative, ensuring that the predicted result is accurately categorised.

FP stands for False Positive, which denotes that the output is positive, but the predicted result is wrongly classified.

FN: False Negative denotes that the output is negative, causing the projected result to be classified wrongly.

3.6.2 Accuracy

The ability of the classifier to successfully forecast the dataset's classes is referred to as accuracy. It is a metric for determining how close the predicted value is to the real or theoretical value (Acharya 2017). The following equation was used to calculate accuracy:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (2)$$

accuracy: 0 to 1 (worst to best value)

3.6.3 Precision

Precision is used to assess the real values that were correctly predicted from the total predicted values in the actual class. Precision is a metric that assesses a classifier's ability to avoid mislabelling a negative example as positive(Chittora et al. 2021). Precision value ranges from 0 to 1. To calculate precision, the following formula can be used:

$$Precision = \frac{TP}{TP + FP + \mathrm{TN} + \mathrm{FN}} \qquad (3)$$

3.6.4 Recall

The recall is the proportion of positive instances that are actually positive. To put it another way, how many patients with a specific result was the model able to identify? With a value ranging from 0 to 1, True Positive Rate or sensitivity are other terms for recall.

$$Recall = \frac{TP}{TP + FN} \qquad (4)$$

3.6.5 F Measure

F Score is another name for it. The F-measure is used to determine the accuracy of a test. It is determined by combining the accuracy and recall together.

F Measure = 2∗ TP/2∗TP+FP+FN = 2 ∗ precision ∗ Recall/ precision + Recall (5)

3.6.6 Sensitivity

True Positive Rate is another name for sensitivity. Sensitivity refers to the average proportion of true positives that are properly classified.

$$Sensitivity = \quad TP \quad \underline{\qquad\qquad} \qquad (6)$$

$TP + FN$

3.6.7 Specificity

True Negative Rate is another name for specificity. It is used to measure the number of negative values that are correctly identified.

$$Specificity = \frac{TN}{TN + FP} \qquad (7)$$

3.6.8 AUC

The AUC stands for the area under the curve. The ROC Curve, also known as the Receiver Operating Characteristic Curve, is a graph that depicts a classification model's performance across all categorization levels. The True Positive Rate on the y-axis and the False Positive on the x-axis.

AUC is a composite measure of performance that considers all potential categorization thresholds. The AUC is a scale that spans from 0 to 1. The closest AUC value to 1 implies that the chosen model is better at differentiating between correct and incorrect classifications in the dataset. The ROC curve and AUC is depicted in Figure 14 and Figure 15 respectively.

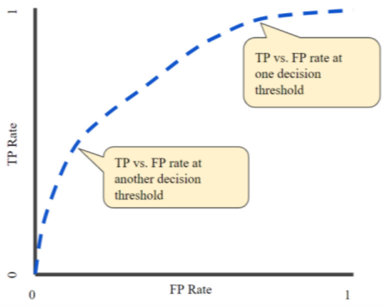


Figure 12 ROC Curve

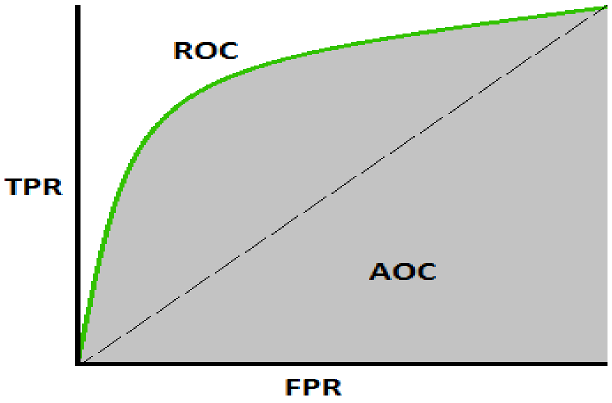


Figure 13 AUC

Key: TPR: True Positive Rate, FPR= False Positive Rate

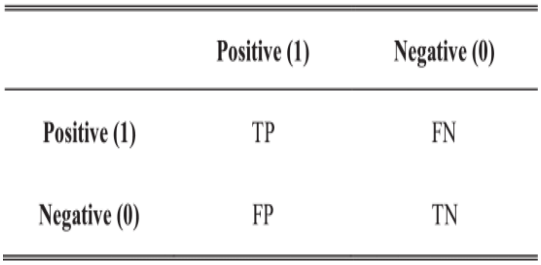
The AUC evaluation is beneficial for two reasons: it is scale invariant, which means

- it gauges how well predictions are scored rather than their absolute values.
- It is classification threshold invariant, which means it assesses the quality of the model's predictions independent of the class threshold used.

3.6.9 Confusion Matrix

A confusion matrix is a tool for describing the classification model's performance. The confusion matrix comprises data on the actual and predicted values that are used to evaluate the model's performance based on the data in the matrix. The confusion matrix represents the number of accurate and wrong predictions generated by the model in relation to the actual classifications in the test data (Sinha and Sinha). The confusion matrix shows two types of accurate classifier predictions and two types of wrong classifier predictions. Figure 16 illustrates the confusion matrix.

Figure 14 Confusion Matrix



	Positive (1)	Negative (0)
Positive (1)	TP	FN
Negative (0)	FP	TN

This work uses confusion matrix with two classifications (CKD, NotCKD) as shown in table 7

Table 7 Confusion Matrix

		Predicted Values	
		CKD	NotCKD
Actual Value	CKD	True CKD	False NotCKD
	NotCKD	False CKD	True NotCKD

3.7 The Machine Learning Algorithms

Machine learning is a branch of Artificial Intelligence that explores various algorithms and learns from them. The experience is obtained through the training of a set of data known as training data. Following the training, the system makes predictions without being programmed (Nishat et al. 2021). Seven machine learning approaches are used in this study, and the outcomes are compared thoroughly using various criteria. Just a few examples are K-nearest Neighbours, Support Vector Machine, Logistics Regression, Random Forests, Artificial Neural Networks, and the Naive Bayes Model.

3.7.1 Random Forest

Random Forest is a learning technique that creates numerous decision trees during training and provides output classes for each tree (Jiang et al. 2020). It may be used for regression as well as classification. RF is an ensemble classification approach that works by merging and voting across the outputs of many decision trees. This approach has the benefit of handling over-fitting and data set of high dimensionality (Subasi et al. 2017).

Each decision tree utilised in this study yields a unique prediction depending on the predictors and training dataset, both of which were picked at random. The findings of this model including its evaluation metrics such as f score recall, and AUC will be shown in the result chapter.

After training, predictions of unknown inputs may be expressed as:

$$f' = {}^{1}/_{B} \sum_{b=1}^{B} f_b(x') \qquad (8)$$

Where, B= optimal number of trees

3.7.2 Naïve Bayes

Nave Bayes is a supervised technique that enforces feature independence when categorising data (Jackins et al. 2021). This approach is particularly useful for datasets with many input characteristics. It considers all accessible features, including some that have minor influence on the final forecast. The Nave Bayes algorithm's probabilistic model is stated as the following equation, where A and B are two independent occurrences.

$$P(A \,/\, B) = \frac{P(B|A) \times P(A)}{P(B)} \qquad (9)$$

Where P is the posterior probability

This model is a Bayes' Theorem-based probabilistic classifier that is the fastest in training when compared to the others (Khomtchouk 2020). This is due to the model's lack of costly iterative parameter estimation, resulting in increased efficiency and practical usage for classification in high-dimensional datasets.

3.7.3 Support Vector Machine (SVM)

SVM is a robust method based on the statistical learning framework that can solve both regression and classification issues (Wang and Chen 2020). SVM can classify both linear and non-linear datasets using the kernel technique. A (n-1) hyper plane separates the datasets, with each data point being treated as an n-dimensional vector. Hyperplanes are decision boundaries that assist separate the data points (Vijayarani and Dhayanand 2015).

The following terms can be used to define a support vector classifier:

$$f(x) = \beta o + \sum_{i \epsilon S} a_i k(x_1, xi) \qquad (10)$$

Where, $\beta 0$ = bias, S= set of observations, α = model parameters that must be learned.

The Support Vector Machine (SVM) uses the one-versus-all and one-vs-one techniques to handle multiclass problems. This method is the most well-known and practical supervised machine-learning algorithm for data classification, learning, and prediction by regressing outliers (Ma et al. 2020). SVM classify all input in high-dimensional data and creates a collection of hyperplanes.

3.7.4 Decision Tree

Another supervised learning approach is Decision Tree which aims to train a model in classifying a target variable by learning simple chained decision rules from previous input features (Ghiasi et al. 2020). According to Nishat et al. (2021), the variables are separated recursively based on a set of impurity criteria until they reach a stopping point. The decision tree model looks like an upside-down tree, with the initial decision rule at the top and subsequent decision rules strewn around like branches throughout the tree.

Gini impurity is chosen as the utilised model among numerous impurities measuring techniques.

$$G(t) = 1 - \sum_{i=1}^{c} p_i^2 \qquad (11)$$

Where, G(t) = Gini impurity at node t

Pi = proportion of observation at class c of node t

3.7.5 Artificial Neural Network

Artificial neural network is a part of supervised machine learning model that attempts to replicate the way neuronal connections in the human brain operate (Chittora et al. 2021). The algorithm is made up of linked layers of neurons called nodes. The ANN is an ensemble approach, like the random forest, that is meant to learn complicated relationships with features in the dataset, in this case, codon frequencies. The ANN system is also capable of handling large datasets with high complexity, as well as capturing complicated relationships and forms across input layers. The multilayer perceptron, an analogous to a neural network blends several perceptions on a single layer is used in this work. A multilayer perceptron (MLP) is a feed-forward artificial neural network composed of several perceptron layers (Theerthagiri 2021). It has nodes designated input node, hidden layer, and output node from at least three layers. This network employs a nonlinear activation function to translate weighted inputs to the outputs of each neuron. The activation functions in this work were sigmoid functions.

The findings of the artificial neural networks model's performance will be shown in the research work's results section.

3.7.6 K-Nearest Neighbour (KNN)

One of the most basic and widely used supervised machine learning techniques is K-Nearest Neighbours (Li et al. 2020b). It does not actually train any dataset; instead, it predicts which classes an observation will fall into based on the fraction of k-nearest neighbours it has. The measure of distance is used to determine similarity. K-NN is a non-parametric machine learning classification technique. It works by classifying related items in an n-dimensional feature space and grouping them together. Distance is a metric to determine similarity, KNN is one of the simplest machine ML algorithms (Jiang et al. 2020). There are many other distance measures, such as Euclidean distance (d),

$$\mathrm{d}_{cuclidean} = \sqrt{\sum_{i=1}^{n}}\ (x_i^2 - y_i^2) \qquad (12)$$

The performance this model will be demonstrated in the research work's results section.

3.7.7 Logistic Regression

Logistic regression is a statistical classifier that assesses the probability of an event inside a given class. Even though the term "regression" appears in its name, logistic regression is a widely used binary classifier (Nusinovici et al. 2020). A decision boundary is a threshold that is set to anticipate which class a data belongs to and the logistic function, which is a sigmoid function, is used to compute the classification probability (Nishat et al. 2021). The mathematical model of the algorithm can be presented below:

$$P_i = \frac{1}{1+\mathrm{e}^{-\sum_{j=0}^{M} \beta_j x_i j}} \qquad (13)$$

Where,

i= 1…N (number of observations)

j= 1…M (number of individual variables)

pi = Probability of a ‘1’ at observation i

βj= Regression Coefficient

xij= The jth variable at observation i

3.7.8 Implementation Environment

Several tools and packages were employed in this study to accomplish the proposed approach as shown in table 8. This study has used the Python programming language to implement the recommended solution from data preparation through the last phase, which is model building. The rationale for using Python is because of its simplified approach in statistical analysis widely adopted by data scientists, researchers, and developers making the building of machine learning models simple and straightforward.

Table 8 Tool and Python Package Descriptions

Tools and packages	Description
Anaconda Navigator	help to launch application in an easily managed environment
Naas.ai	an open source for machine learning project
Jupyter Notebook	free application that allows to create document. Useful for data cleaning, modelling, visualisation, and machine learning
Google colab	an open source for machine learning project
Python	programming language for machine learning project

Microsoft Excel	Data preparation
Microsoft word	Documentation purpose
Scikit-learn	For handling classification and regression
Pandas	For loading and filtering data
Numpy	Array processing for number. Used in transformation of text values
Matplotlib	Quality figure in python for visualisation
Seaborn 0.9.0 Statistical data visualization	statistical data visualisation

3.8 Ethical Considerations

This research observed a full compliance to the rules and procedures from Health Insurance Portability and Accountability act (HIPAA) privacy rule as well as GDPR to obtain data by adhering to privacy, confidentiality and ensuring that the data is used for academic research purposes. The use of personal data must follow data protection principles by lawfully, fairly, and transparently using the obtained data. The dataset for this study was ethically taken from the UCI repository with necessary ethical approval. These machine learning algorithms were implemented in accordance with the instructions provided in the different documentations. Finally, the different machine learning python programmes that were utilised in this study can be seen on my GitHub repository at https://github.com/Oluyori/Chronic-Kidney-Disease-Prediction/blob/main/CKDPrediction%20(10).ipynb. The study's shortcomings are discussed in section 6.1, which follows the conclusions.

3. 9 Chapter 3 Summary

To summarise, this chapter looked at the research methods used in this study. We chose positivism as the philosophical framework most suited for our investigation, based on Saunders et al. (2009)'s research onion technique. The data source, as well as the nature and content of the information, were all detailed in length; steps on features selection and relevant machine learning tasks were given. The numerous machine learning algorithms were also explored, as well as the step-by-step strategies of machine learning in disease prediction. The performance measures for machine learning were also discussed. Finally, alongside the exact packages utilised in the study, information regarding the software packages, Python programming language, was presented. This brings us immediately to the work's following chapter, the results section.

Chapter 4

Results

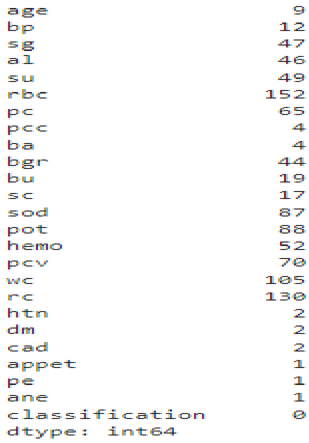
4.0 Introduction

The findings are presented in this section. It displays the performance metrics of the several machine learning algorithms employed in this study. Tables, charts, and figures are used to present the findings. Each machine learning model's performance metrics are shown in the table, while the confusion matrix is shown in the picture. All the models' accuracy is compared in chart.

4.1 Visualisation of missing value

Figure 15 shows the missing values in the CKD data set

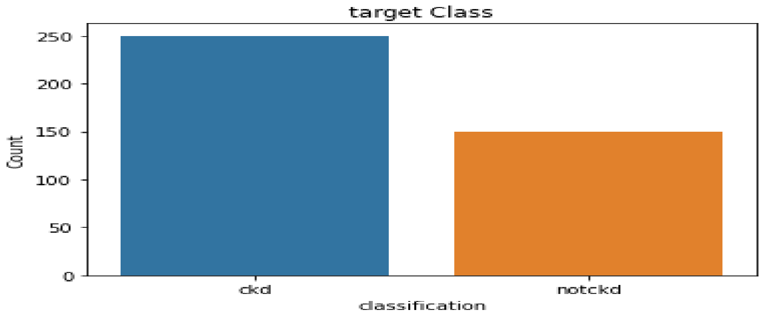
Figure 15 Missing Value



```
age                 9
bp                 12
sg                 47
al                 46
su                 49
rbc               152
pc                 65
pcc                 4
ba                  4
bgr                44
bu                 19
sc                 17
sod                87
pot                88
hemo               52
pcv                70
wc                105
rc                130
htn                 2
dm                  2
cad                 2
appet               1
pe                  1
ane                 1
classification      0
dtype: int64
```

After cleaning the data of 400 patients and 24 features, figure 16 shows the percentage of Class percentage

Figure 16 Class percentage



```
Percent of chronic kidney disease sample:  62.5 %
Percent of not a chronic kidney disease sample:  37.5 %
```

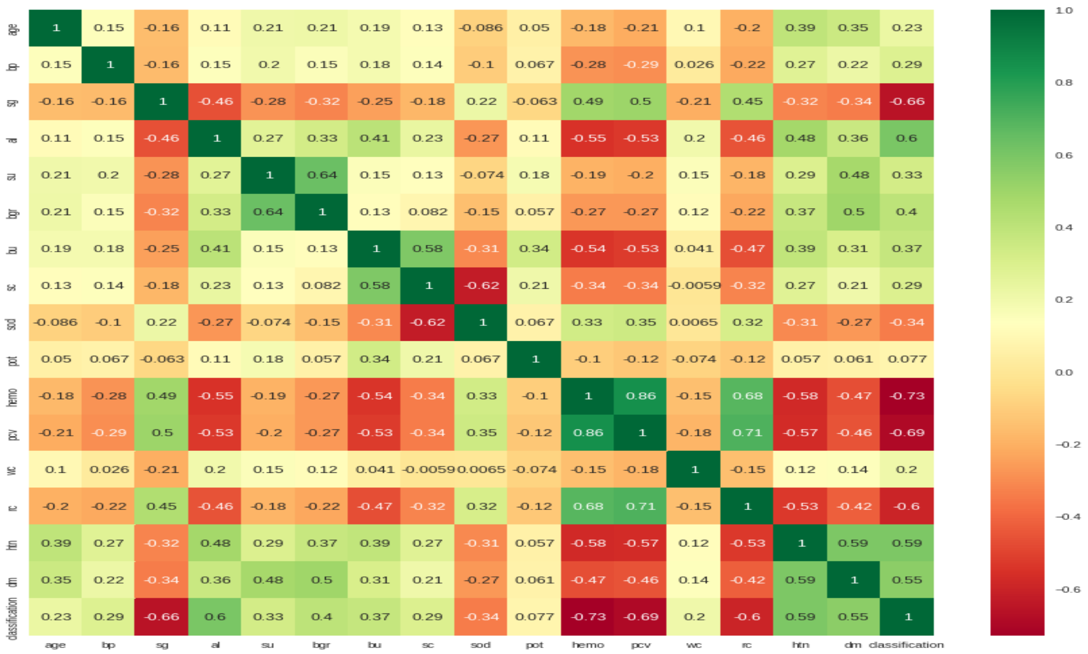
Table 9 Class percentage

Class	Number of patients	%
CKD	250	62.5
NotCKD	150	37.5

4.2 Feature selection after data transformation using Pearson

After transforming categorical values to nominal data, figure 17 depicts the correlation among features.

Figure 17 correlation among features

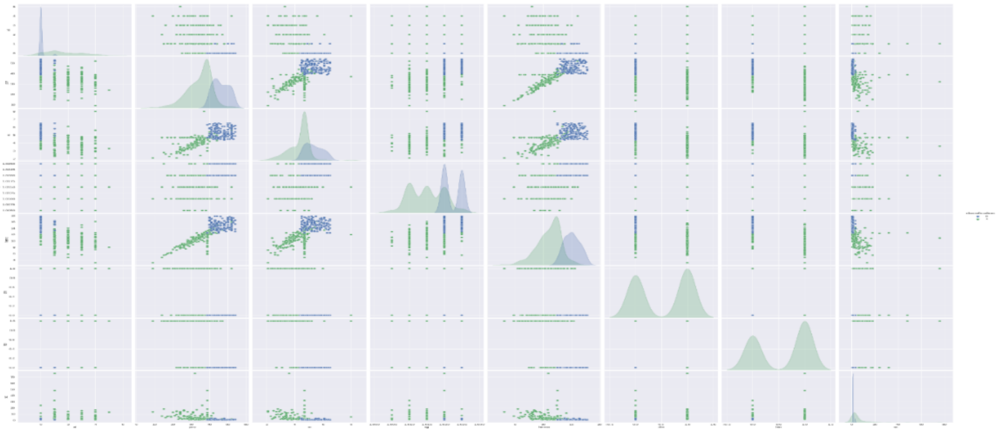


4.2.1 Outcome of features selection among the highly correlated features

Table 10 Correlation of features selected for the model building

Features	Acronym	Correlation
htn	Hypertension	0.59
dm	Diabetes meletus	0.55
al	Albumin	0.6
rc	Red blood cell count	-0.6
hemo	Hemoglobin	-0.73
pcv	Packed Cell Volume	-0.69
sg	Specific gravity	-0.66
su	Sugar	0.64
sc	Serum creatinine	0.58
class	Classification	1

Figure 18 Checking data distributions between some highly correlated Features



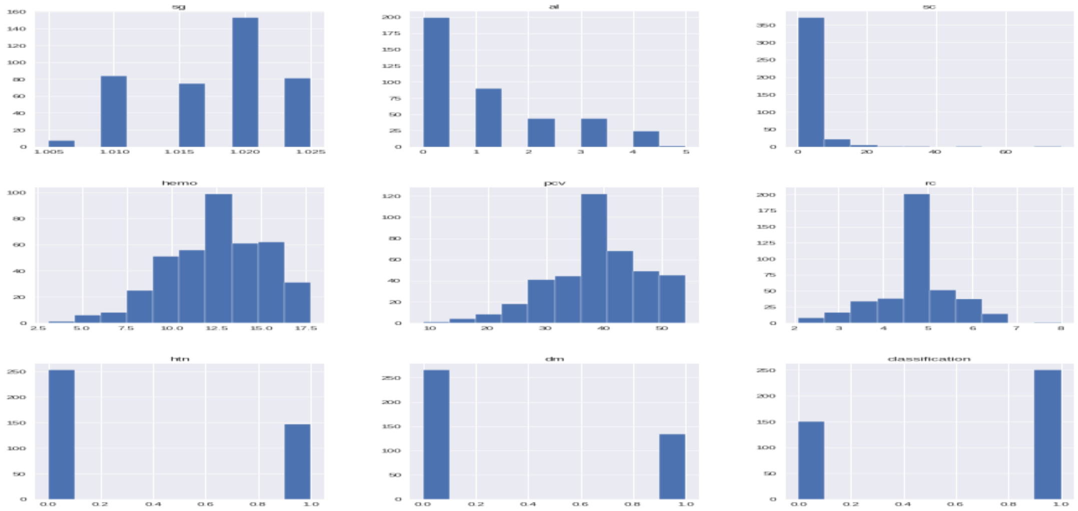


Table 11 Statistical Description of selected features

Features	count	mean	Std Deviation	min	max
sg	400	1.01771	0.005434	1.00500	1.025000
al	400	1.0173	1.272318	0.00000	5.000000
su	400	0.45013	1.02949	0.00000	5.000000
sc	400	3.07235	5.61749	0.40000	76.00000
hemo	400	12.5269	2.716171	3.100000	17.800000
pcv	400	38.86418	8.151199	9.000000	54.000000
rc	400	4.70173	0.840354	2.100000	8.000000
htn	400	0.3675	0.482728	0.000000	1.000000
dm	400	0.335	0.472582	0.000000	1.000000
class	400	0.625	0.484729	0.484729	1.000000

Table 12 Shape of the training and test data

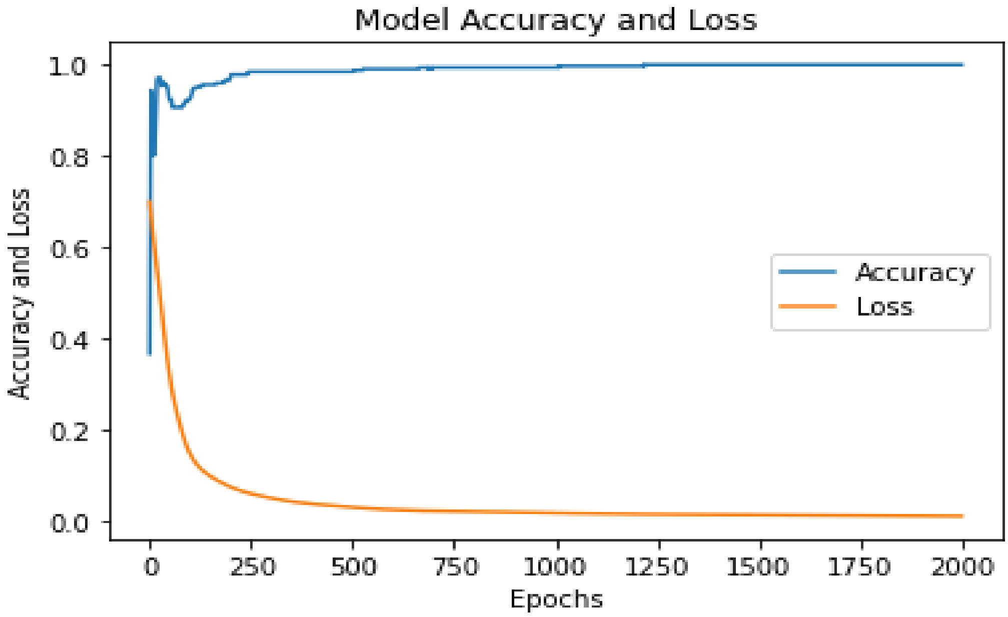
Training Data 80% with 9 features	Testing data 20% with 9 features
320	80

4.3 Model Results

4.3.1 MLP (Artificial Neural Network) Algorithm Performance and visualisation.

Figure 19 shows the visualisation of MLP when applying 2000 epoche, the loss and accuracy of the model.

Figure 19 MLP Accuracy and Loss Visualisation



The performance metrics of MLP model in classifying a patient as CKD or Notckd using clinical data is shown in table 13 below. The Notckd and Ckd are the class types indicated as 0,1 in the table, respectively.

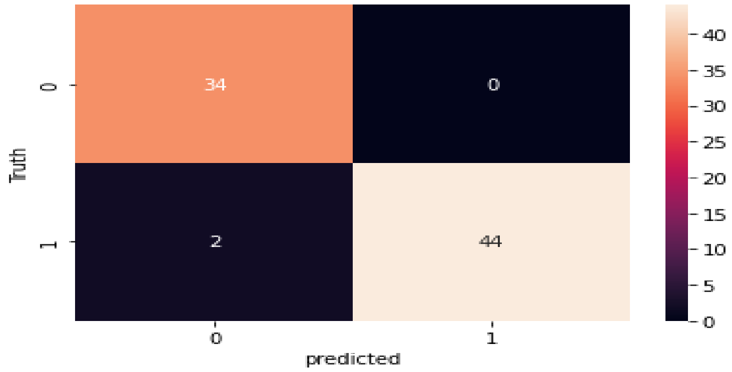
In the confusion matrix evaluation of the model, the classification accuracy for identifying the class of kidney disease is high at 0.98 (98%) accuracy. Precision, Recall, the F1-Score, and the macro average are among the additional assessment measures shown in table 13.

Table 13 Performance Metrics for MLP

			Evaluation Matrix					
Activation	solver	epoche	class		Precision	Recall	F1 score	Support
relu	adam	2000		0	0.94	1.00	0.97	34
				1	1.00	0.96	0.98	46
				Overall Accuracy	**0.98**			

Figure 20. below shows the heatmap representation of the confusion Matrix of the MLP for CKD classification

Figure 20 confusion Matrix of the MLP for CKD classification



True Neg: 34

False Pos: 0

False Neg: 2

True Pos: 44

Where,

TN: True Negative, FP: False Positive, FN: False Negative, TP: True Positive

4.3.2 K-Nearest Neighbours (KNN) Model Evaluation Metrics

The performance metrics of KNN model in classifying a patient as CKD or NotCKD using clinical data is shown in table 14 below. The NotCKD and CKD are the class types indicated as 0,1 in the table, respectively.

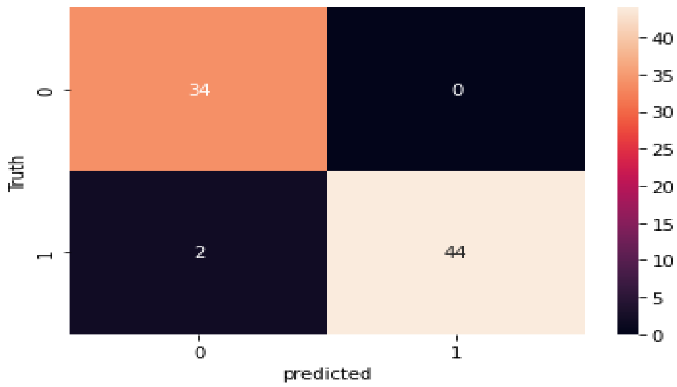
In the confusion matrix evaluation of the model, the classification accuracy for identifying the class of kidney disease is high at 0.98 (98%) accuracy. Precision, Recall, the F1-Score, and the macro average are among the additional assessment measures shown in the table.

Table 14 Performance Metrics for KNN

		Precision	Recall	F1 score	Support
class	0	0.94	1.00	0.97	34
	1	1.00	0.96	0.98	46
	Overall Accuracy	**0.98**			

Figure 21 below shows the heatmap representation of the confusion Matrix of the KNN for CKD classification

Figure 21 confusion Matrix of the KNN for CKD classification



True Neg: 34

False Pos: 0

False Neg: 2

True Pos: 44

4.3.3 Random Forest (RF) Algorithm Performance Metrics

The performance metrics of RF model in classifying a patient as CKD or Notckd using clinical data is shown in table 15 below. The NotCKD and CKD are the class types indicated as 0,1 in the table, respectively.

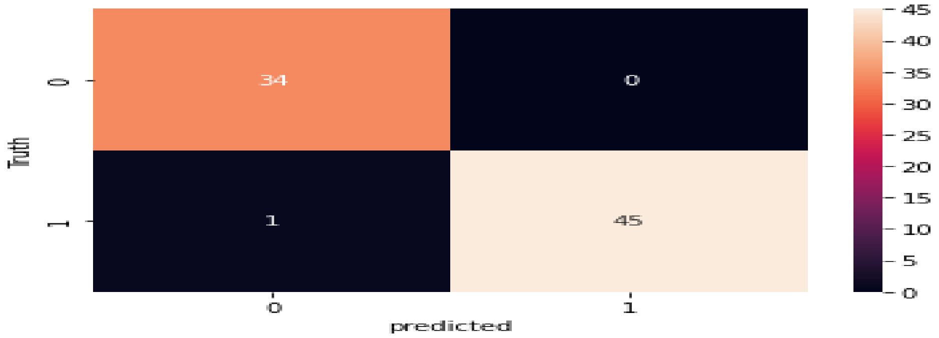
In the confusion matrix evaluation of the model, the classification accuracy for identifying the class of kidney disease is high at 0.99 (99%) accuracy. Precision, Recall, the F1-Score, and the macro average are among the additional assessment measures shown in table 15.

Table 15 Performance Metrics for Random Forest

		Precision	Recall	F1 score	Support
class	0	0.97	1.00	0.99	34
	1	1.00	0.98	0.99	46
	Overall Accuracy	**0.99**			

Figure 22 below shows the heatmap representation of the confusion Matrix of the RF for CKD classification

Figure 22 confusion Matrix of the RF for CKD classification



True Neg: 34

False Pos: 0

False Neg: 1

True Pos: 4

4.3.4 Support Vector Machine Performance Metrics

The performance metrics of SVM model in classifying a patient as CKD or Notckd using clinical data is shown in table 16 below. The NotCKD and CKD are the class types indicated as 0,1 in the table, respectively.

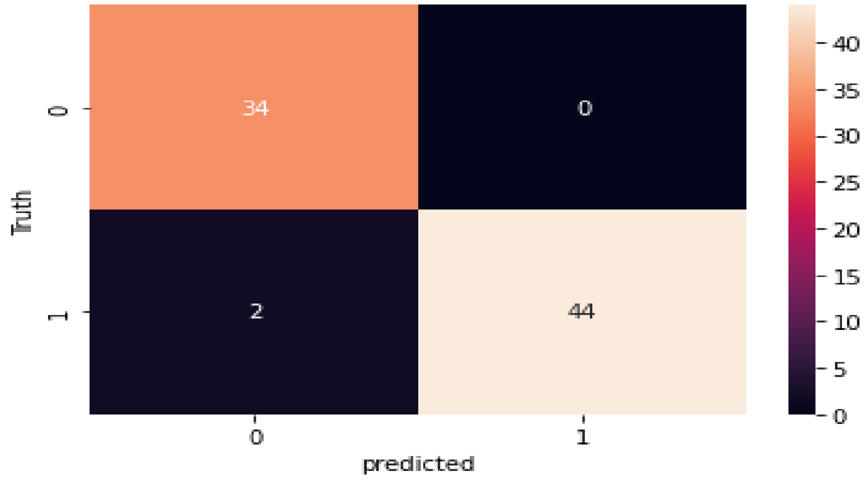
In the confusion matrix evaluation of the model, the classification accuracy for identifying the class of kidney disease is high at 0.98 (98%) accuracy. Precision, Recall, the F1-Score, and the macro average are among the additional assessment measures shown in table 16.

Table 16 SVM Evaluation Metrics

		Precision	Recall	F1 score	Support
class	0	0.94	1.00	0.97	34
	1	1.00	0.96	0.8	46
	Overall Accuracy	**0.98**			

Figure 23 below shows the heatmap representation of the confusion Matrix of the SVM for CKD classification

Figure 23 confusion Matrix of the SVM



True Neg: 34

False Pos: 0

False Neg: 2

True Pos: 44

4.3.5 Gaussian Naïve Bayes Algorithm Performance Metrics

The performance metrics of GNB model in classifying a patient as CKD or NotCKD using clinical data is shown in table 17 below. The NotCKD and CKD are the class types indicated as 0,1 in the table, respectively.

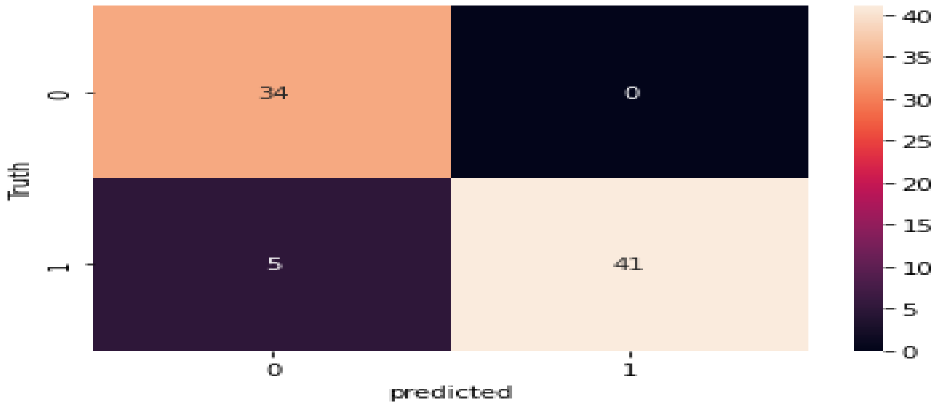
In the confusion matrix evaluation of the model, the classification accuracy for identifying the class of kidney disease is high at 0.95 (95%) accuracy. Precision, Recall, the F1-Score, and the macro average are among the additional assessment measures shown in table 17.

Table 17 GNB Performance Metrics

class		Precision	Recall	F1 score	Support
	0	0.87	1.00	0.93	34
	1	1.00	0.89	0.94	46
	Overall Accuracy	**0.95**			

Figure 24 below shows the heatmap representation of the confusion Matrix of the GNB for CKD classification

Figure 24 confusion Matrix of the GNB



True Neg: 34

False Pos: 0

False Neg: 5

True Pos: 41

4.3.6 Decision tress (DT) Algorithm Performance Metrics

The performance metrics of DT model in classifying a patient as CKD or NotCKD using clinical data is shown in table 18 below. The NotCKD and CKD are the class types indicated as 0,1 in the table, respectively.

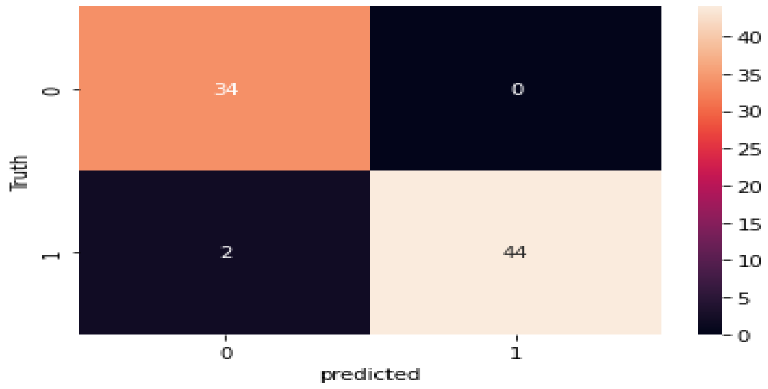
In the confusion matrix evaluation of the model, the classification accuracy for identifying the class of kidney disease is high at 0.98 (98%) accuracy. Precision, Recall, the F1-Score, and the macro average are among the additional assessment measures shown in table 18.

Table 18 DT Performance Metrics

class		Precision	Recall	F1 score	Support
	0	0.94	1.00	0.97	34
	1	1.00	0.96	0.98	46
	Overall Accuracy	**0.98**			

Figure 25 below shows the heatmap representation of the confusion Matrix of the DTC for CKD classification.

Figure 25 confusion Matrix of the DTC



True Neg: 34

False Pos: 0

False Neg: 2

True Pos: 44

4.3.7 Logistics Regression (LR) Algorithm Performance Metrics

The performance metrics of LR model in classifying a patient as CKD or NotCKD using clinical data is shown in table 19 below. The NotCKD and CKD are the class types indicated as 0,1 in the table, respectively.

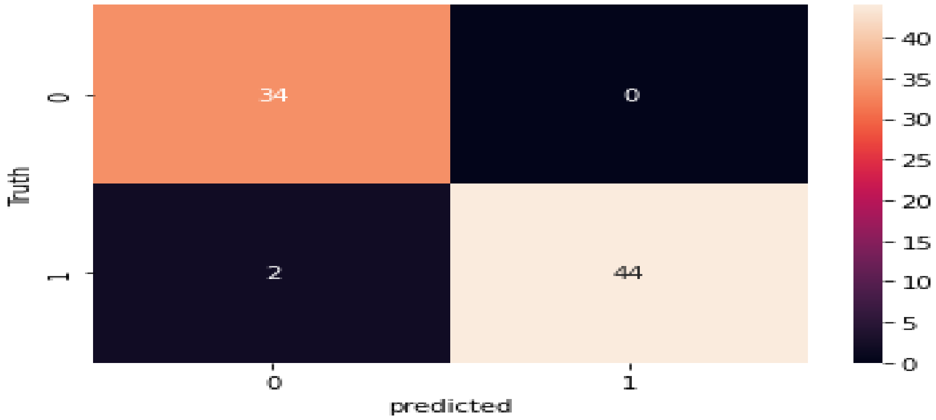
In the confusion matrix evaluation of the model, the classification accuracy for identifying the class of kidney disease is high at 0.98 (98%) accuracy. Precision, Recall, the F1-Score, and the macro average are among the additional assessment measures shown in table 19.

Table 19 LR Performance Metrics

		Precision	Recall	F1 score	Support
class	0	0.94	1.00	0.97	34
	1	1.00	0.96	0.98	46
	Overall Accuracy	**0.98**			

Figure 26 below shows the heatmap representation of the confusion Matrix of the LR for CKD classification.

Figure 26 confusion Matrix of the LR



True Neg: 34

False Pos: 0

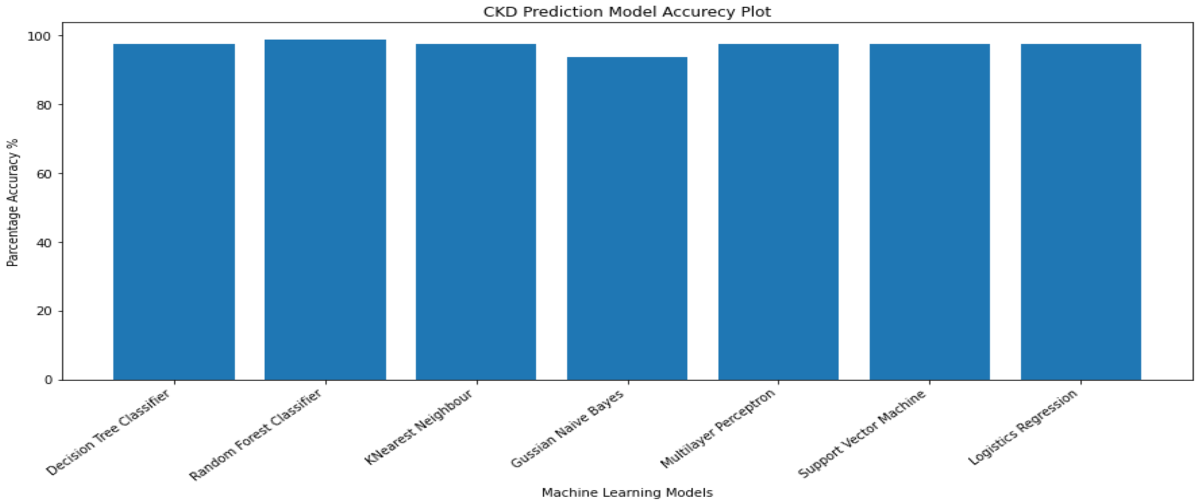
False Neg: 2

True Pos: 44

4.4 Performance Comparison of Machine Learning Classifiers

The graph below compares the performance characteristics of the seven machine learning models (accuracy).

Figure 27 A bar graph comparing the accuracy of machine learning models.



A bar graph comparing the accuracy of machine learning models on two classification tasks.

The accuracy of all the seven models in classifying the CKD is high in all the classifiers. Where RF has the highest percentage, followed by SVM, MLP, KNN, LR and DT; GNB has the least accuracy in the classification of CKD.

Table 20 Model comparison

Model	Accuracy	TN	FP	FN	TP	AUROC score
RF	99%	34	0	1	45	0.925
SVM	98%	34	0	2	44	0.940
MLP	98%	34	0	2	44	0.969
KNN	98%	34	0	2	44	0.580
LR	98%	34	0	2	44	0.947
DT	98%	34	0	2	44	0.969
GNB	95%	34	0	5	41	0.993

4.5 Model evaluation

As discussed in chapter 3, the performance of the models will be evaluated using the confusion matrix and the Area under Receiver operating characteristic curve AUC ROC curve

4.5.1 Receiver Operating Characteristics (ROC) Curve

The ROC curve of the seven machine learning classifiers is shown in figure 28 and 29 below after the models were evaluated on completely new and noisy data set to validate their performance.

Figure 28 ROC curve of RF, GNB, KNN, DT, LG, and MLP

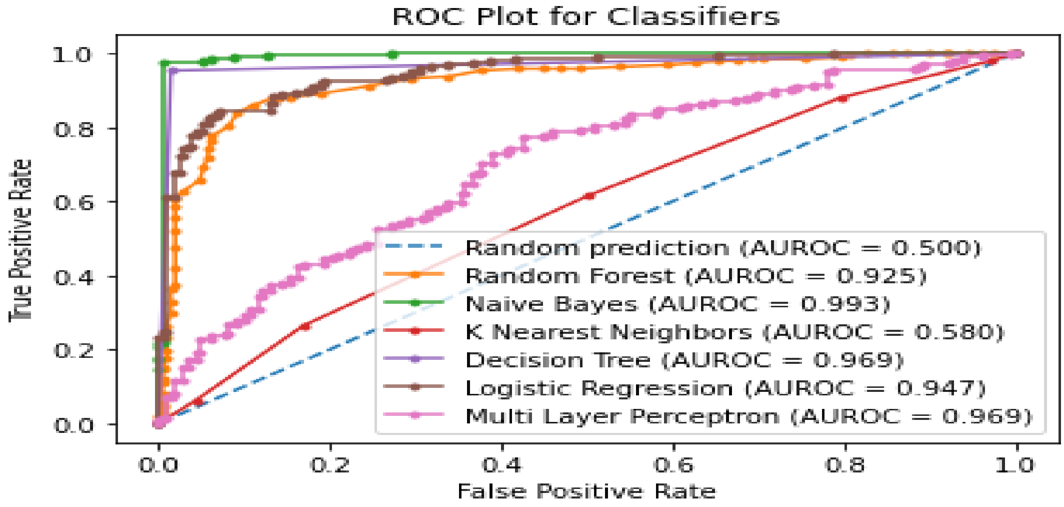
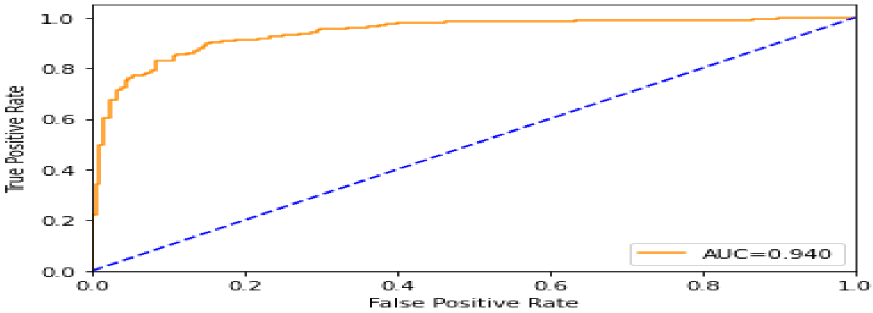


Figure 29 ROC curve of SVM



From the Area Under Curve (AUC) scores obtained from the ROC graphs in figure 4.9 and 4.10 above, the best performing model in classifying CKD is GNB with 0.993 AUC, followed by Multilayer Perceptron and Decision Tree with AUC of 0.969. K Nearest Neighbour has the least AUC of 0.580 but still above our threshold of 0.500.

Chapter 5

5.0 Discussion

In this chapter, we discuss the results of utilizing several machine algorithms and 10-fold cross validation to investigate the influence of selected features on prediction and evaluating the model performance.

The research aims to predict CKD using different machine learning techniques by comparing their performances. The findings of the study reported in Chapter 4 provide crucial and convincing evidence that supervised machine learning methods may be used to grasp medical concepts buried in clinical data. Biomedical scientists and physicians may use machine learning techniques and approaches to assist them grasp complicated diagnostic ideas, according to this study. Several supervised machine learning approaches were employed in this study to categorise and identify chronic kidney disease.

The overall research result is summarised by describing and addressing each of the research questions (RQ) posed at the start of the study, utilising explanations from the literature review and our research findings.

5.1 RQ1 What are the predicting features for CKD prediction?

The finding of the feature selection from the obtained CKD data indicates that Hypertension, diabetes Meletus, sugar, Albumin, Red blood cell count, Hemoglobin, Packed cell volume, specific gravity and serum creatinine are the most relevant predictors following the high correlation among the features and the target variable “classification”. According to Sivasankar et al. (2015), applying features reduction will eliminate irrelevant data from the original dataset. In this study, relevant features were selected from the already cleaned and transformed data. A threshold of 0.5 was adopted to select the attributes that are either positively or negatively correlated with the target variable by defining the absolute function because negative correlation is also very important, indicating that a feature's value increases while the value of another feature decreases which is important in a machine learning project. On the reviewed theories, systemic hypertension from patient's high blood pressure and stress can damage the filters and induce harm to the glomeruli, leading to glomerulosclerosis and kidney failure (Kazancioğlu 2013; Waezizadeh et al. 2018). According to Lea and Nicholas (2002), diabetic individuals are at risk for CKD due to nephropathy and hyperfiltration damage. Hypertension and stress are perceived as the damaging causes of CKD through the removal of nephrons thereby reducing the GFR and generate an increased irregular protein excretion in urine. These findings confirm these existing concepts. On the visualisation among the highly correlated features, hemo, pcv and rc are more evenly distributed. The feature reduction process is important to reduce the cost of running clinical test and time, and building cost effective model for predicting disease (Rajeswari et al. 2013).

5.2 RQ2 Which machine learning model is better for CKD prediction?

From the findings all the machine learning models give satisfactory results on accuracy ranging from 95% to 99% and have negligible difference between precision and recall values. In comparison among the models, accuracy, precision, recall and AUC score of KNN is low which indicates that the model gives false negative with low capacity to distinguish between positive and negative class. Random forest with the selected features achieved the highest percentage of 99% of accuracy and AUC score of 0.925. After data pre-processing, 250 of the 400 patients have CKD occupying 62.5% of the instances and 150 patients with 37.5% do not have CKD. According to Ramyachitra and Manikandan (2014) and Sun et al. (2009), data imbalance can affect classification algorithm. Model evaluation including F1 score, Precision, Recall and confusion matrix were used to measure for performance.

The activation function “relu” in MLP, epoch of 2000 and loss function of 0.0203 play a role in getting higher accuracy, evaluating how the model process the data with less computational time. Applying these parameters on MLP, we get 98% of accuracy, Precision and F-measure; 97% and of Recall and 0.969 AUC score which evaluates the high performance of MLP model.

The seven machine learning models are various classification methods that have been thoroughly detailed in this work's third chapter. Except for the GNB model and KNN, all the seven models performed well in terms of prediction, validation, and classification. GNB has the lowest accuracy of 95%, while KNN has the lowest AUC score of 0.580. These analyses have significant promise in demonstrating the CKD risk factors, in addition to the classification of CKD utilising data and machine learning. Where Random Forest has the highest accuracy of 99.0% of 34 and 46 support, the model absolutely predicts 34 True negative, 0 False Positive, 0 False Negative and 45 True Positive, the further evaluation of RF with 0.925 AUC score checks the model performance for CKD prediction.

From the result, Support Vector Machine, Multilayer perceptron, KNN, logistic regression and decision tree all have 98% of accuracy with True Negative:34, False Positive:0, False

Negative:2, True Positive:44 and ROC score of 0.940, 0.969, 0.580, 0.947 and 0.969 in that order. Gaussian Naïve Bayes with the least percentage of accuracy of 95% shows the highest ROC score of 0.993 with True Negative:34, False Positive:0, False Negative:5, True Positive:41. The finding indicates the efficiency of machine learning with potential benefits in disease prediction if adopted.

In the previous work by Salekin and Stankovic (2016), 99.8% of RF performance accuracy was recorded on the same dataset using 24 features, while this study give 99% of RF accuracy using 9 features. Chen et al. (2016) achieved 99% of SVM accuracy, through feature selection in this study, the 98% accuracy of SVM is evaluated using confusion matrix and AUC score. The confusion matrix justifies the accuracy through the accurate True Negative:28, False Positive:0, False Negative:1, True Positive:52, the AUC score of 0.940 further check the efficiency of the accuracy of RF. Therefore, some of the covered metrics may not be sufficient to evaluate model performance. According to Rácz et al. (2019), performance metrics are usually the most important consideration factors when choosing the best machine learning classification algorithm; however, only a few performance metrics were evaluated in this study which are confusion matrix, Area under ROC curve and the True Positive, True Negative, False Positive and False Negative; and they may not be enough to strongly suggest the model that will perform best in all cases. Choosing the optimal machine learning model does not only depend on the metrics identified in this study but also on other metrics such as the R squared score (R^2), that is not covered in this study.

The high accuracy obtained in this study ranging from 95% to 99% reflects the relevant features selected. The reduction in the features reduce the computational time particularly the time to build MLP of 13 milli second. Overall, the findings of this study show that all the machine learning classifiers give excellent performance accuracy and validation score. In similar research by Mohammed Siyad et al. (2016), feature selection was adopted including hypertension, hemoglobin and diabetes Meletus and achieved 100% accuracy in random forest classifier. Therefore, from this study's result, adopting of Random Forest with the highest accuracy of 99% and AUC of 0.925 on Medical Health Record data may yield valuable clinical insights, which can help physicians and patients understand the likely risk of CKD and take quick diagnosis upon early detection.

5.3 RQ3 Is ML adoption an influence in disease prediction?

The results confirm that availability of data and features selection will aid high model performance accuracy. This approach is important for clinical decision making. According to Luijks et al. (2012); Piri et al. (2018), the application of machine learning and data mining approach on routinely obtained data from hospital admissions will help in understanding the comorbidities thereby improving treatment option through the integration of the Markov Decision making processes (MDPs) (Bennett and Hauser 2013). Effective use of administrative data will help in disease risk prediction and reduce the pressure on limited healthcare resources.

The result of this study suggests that the application of machine learning on disease prediction will improve healthcare through high accuracy. However, where this study shows high model accuracy, for example, our MLP model gives 98% accuracy, previous work by Lakshmi et al. (2014) show 93.521% of ANN(MLP) accuracy. Tazin et al. (2016) achieved 99.75% model with decision tree where our study give 98%. This in turn reflects the impact of feature selection in machine learning project. Overall, this work gives high model accuracy ranging from 95% to 99% and ROC score from 0.580 to 0.993 which further indicates the performance of the model. Data availability, Machine learning, and MDPs will benefit clinicians, patients and healthcare in effective decision making.

5.4 RQ4 Can Machine Learning improve healthcare?

Machine learning developments are having an influence on healthcare, just as they are in other fields. Through the vast quantity of data that generated, ML is widely used in healthcare for diagnosis of diseases such as cancer and neuroimaging for brain disease. According to Devi and Balasubramanian (2022), data collection will be useful in healthcare for machine adoption and when used effectively, machine learning will assist physicians in making near-perfect diagnoses, selecting the appropriate prescriptions for their patients, identifying individuals at high risk for poor pharmaceutical results, and improving patients' overall health while lowering costs. Through the study’s finding on model accuracy, machine learning can assist diagnose patients and identify people who are more prone to repeated diseases. Furthermore, about 90% of emergency department visits may be avoided. Relying on human clinicians to estimate the probabilistic effects of multiple actions over time may affect the provision of genuine patient care, modelling can be used to tackle the complex medical treatment decisions than relying just on intuition (Schaefer et al. 2005).

Even though several machine learning algorithms have been created, machine learning is still more prominently featured in research than in implementation. For the greater benefit of society, more deployable kinds of machine learning must be developed.

To summarise, we have shown that machine learning can be used to classify patients into CKD or NotCKD for clinical intervention. The relevant features including hypertension, serum creatinine and diabetes are great predictors of CKD, and this was fully covered. The performance of the machine learning models employed for the classification tasks was also examined, with the top and worst performing approaches being identified.

Chapter 6
6.0 Conclusion

The goal of the research is to predict chronic kidney disease using key variables from the CKD dataset and compare the algorithms' performance. Random Forest, Logistics regression, Multilayer perceptron, Nave Bayes, K closest neighbour, decision tree, and support vector machine were among the seven classifiers used. Random Forest was shown to be the most accurate, with a 99 percent accuracy rate. In the study, the impact of features selection on disease prediction was demonstrated. Hypertension, diabetes, and albumin are risk factors of CKD. This was clearly justified using different machine learning techniques. The dominance on labelled data that give clearer criteria for model optimisation and the popularity of the applied supervised machine learning models enhance the comparison of the models' performance (Nasteski 2017). All the techniques gave excellent percentage and have been applied in multiple fields. RF which has the highest accuracy of 99% in this study has previously displayed a high performance in other related works including the 100% accuracy achieved in CKD prediction (Mohammed Siyad et al. 2016). Therefore, RF model will be proposed on selected features to minimise cost and time thereby improving clinical support. When feature selection is used, the result revealed a significant improvement in classifier performance.

Data pre-processing carried out to eliminate the noise in the dataset and the missing value, and 10-fold cross validation on the 80% trained data and 20% test proportion of the data contributed to the model accuracy. Accuracy percentage of 99; 98; and 95 are obtained in RF; SVM, MLP, KNN, LR, and DT; and GNB model in that order. Using performance metrics, all the implemented models achieved excellent AUC score greater than the threshold ranging from 0.580 to 0.993. Overall, Random Forest has the highest percentage of accuracy of 99%, followed by SVM, MLP, KNN, LR, and DT of 98%, and Naïve Bayes with 95% of accuracy. Evaluation on the model accuracy using confusion matrix and AUC score validate the model performance with each model performing beyond the 0.5 ROC threshold.

6.1 Practical Implication

This study has demonstrated the application of machine learning in disease prediction that will grant physician more efficiency and improve patients' quality of life. As hypertension and diabetes have significant impact of CKD, clinician can early diagnose hypertensive and diabetic patients with the likelihood of developing CKD that will save cost and time. However, physicians have an important role in affecting the quality of administrative data, which necessitates a shift in attitude towards chart documentation and coding. Ethical concerns on machine learning adoption will require cross-disciplinary collaboration among the stakeholders at every stage. GDPR guidelines and principle of fairness will regulate data security and privacy when it comes to the costs of incorrect classifications, such as false negative or false positive diagnoses (Char et al. 2018).

6.2 Theoretical Implication

The study has added knowledge in the prediction of chronic kidney disease with machine learning. This study indicates that the integration of administrative data and graph theory, deterministic theory, artificial intelligence, and machine learning models are suitable for CKD prediction and will be essential for clinical support and decision making. Through this study, it is confirmed that application of various machine learning techniques on collected clinical data is promising.

6.3 Recommendation

Despite the significant contribution made in this study, additional work, and exploration into the adoption of machine learning approaches for CKD prediction is needed. Institutions that handle these data, on the other hand, should make high-quality data publicly available. Data collection is important and should be unbiased. The availability of these data will aid in the development of more effective machine learning algorithms that will reveal hidden information in clinical data. Since CKD is non-communicable and asymptomatic, physicians can detect hypertensive and diabetic patients who are at risk of developing CKD before it becomes deadly using Random Forest.

6.4 Limitation and Future work

At various phases of the research development process, this study encounters several limitations. Access to relevant data was a major issue. The confidentiality and organization of primary data posed ethical concerns regarding patients’ privacy. In this regard, secondary data from UCI was alternatively considered for this study. Challenges associated with machine learning project is data availability. Finding the relevant data for this study was a challenge that modified the study from initially proposed topic. For example, prediction of acute kidney injury

(AKI) was initially proposed for this study. The unavailability of AKI data and research timeline inspired the consideration CKD prediction. This study applied supervised machine learning. However, progressive research on CKD prediction with deep learning and large dataset on the selected features will help to reduce CKD. With numerous research on CKD prediction, further work can focus on more adoption of these models in the society and the challenges associated with the technique's adoption.

7.0 References

Acharya, A. (2017) Comparative study of machine learning algorithms for heart disease prediction.

Adeniran, A., Stainbrook, S., Bostick, J. W. and Tyo, K. E. J. (2018) Detection of a peptide biomarker by engineered yeast receptors. *ACS synthetic biology* 7 (2), 696-705.

Akben, S. B. (2018) Early stage chronic kidney disease diagnosis by applying data mining methods to urinalysis, blood analysis and disease history. *IRBM* 39 (5), 353-358.

Alasker, H., Alharkan, S., Alharkan, W., Zaki, A. and Riza, L. S. (2017) Detection of kidney disease using various intelligent classifiers. *2017 3rd International Conference on Science in Information Technology (ICSITech).* 25-26 Oct. 2017.

Aljaaf, A. J., Al-Jumeily, D., Haglan, H. M., Alloghani, M., Baker, T., Hussain, A. J. and Mustafina, J. Early prediction of chronic kidney disease using machine learning supported by predictive analytics. 2018. IEEE.

Aljaaf, A. J., Al-Jumeily, D., Haglan, H. M., Alloghani, M., Baker, T., Hussain, A. J. and Mustafina, J. (2018) Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics. *2018 IEEE Congress on Evolutionary Computation (CEC).* 8-13 July 2018.

Alloghani, M., Al-Jumeily, D., Hussain, A., Liatsis, P. and Aljaaf, A. J. (2020) Performance-Based Prediction of Chronic Kidney Disease Using Machine Learning for High-Risk Cardiovascular Disease Patients. In Yang, X.-S. and He, X.-S. (editors) *Nature-Inspired Computation in Data Mining and Machine Learning.* Cham: Springer International Publishing. 187-206.

Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., Alrashed, S. and Olatunji, S. O. (2019) Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. *Computers in Biology and Medicine* 109, 101-111.

Almustafa, K. M. (2021) Prediction of chronic kidney disease using different classification algorithms. *Informatics in Medicine Unlocked* 24, 100631.

Altan, A., Karasu, S. and Zio, E. (2021) A new hybrid model for wind speed forecasting combining long short-term memory neural network, decomposition methods and grey wolf optimizer. *Applied Soft Computing* 100, 106996.

Amirgaliyev, Y., Shamiluulu, S. and Serek, A. (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).* 17-19 Oct. 2018.

Ani, R., Sasi, G., Sankar, U. R. and Deepa, O. S. Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification. 2016. IEEE.

Antony, L., Azam, S., Ignatious, E., Quadir, R., Beeravolu, A. R., Jonkman, M. and Boer, F. D. (2021) A Comprehensive Unsupervised Framework for Chronic Kidney Disease Prediction. *IEEE Access* 9, 126481-126501.

Aslam, M. A., Xue, C., Liu, M., Wang, K. and Cui, D. (2021) Classification and Prediction of Gastric Cancer from Saliva Diagnosis using Artificial Neural Network. *Engineering letters* 29 (1).

Bahadir, E. (2016) Using Neural Network and Logistic Regression Analysis to Predict Prospective Mathematics Teachers' Academic Success upon Entering Graduate Education. *Educational Sciences: Theory and Practice* 16 (3), 943-964.

Barton, A. L., Mallard, A. S. and Parry, R. G. (2015) One year's observational study of acute kidney injury incidence in primary care; frequency of follow-up serum creatinine and mortality risk. *Nephron* 130 (3), 175-181.

Baumann, H. and Sandmann, W. (2016) Structured modeling and analysis of stochastic epidemics with immigration and demographic effects. *PloS one* 11 (3), e0152144.

Bell, J. and Waters, S. (2018) *Ebook: doing your research project: a guide for first-time researchers*. Mcgraw-hill education (uk).

Bellomo, R., Kellum, J. A. and Ronco, C. (2012) Acute kidney injury. *The Lancet* 380 (9843), 756-766.

Bennett, C. C. and Hauser, K. (2013) Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. *Artificial intelligence in medicine* 57 (1), 9-19.

Boulesteix, A.-L., Porzelius, C. and Daumer, M. (2008) Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics* 24 (15), 1698-1706.

Burrell, G. and Morgan, G. (2017) *Sociological paradigms and organisational analysis: Elements of the sociology of corporate life.* Routledge.

Char, D. S., Shah, N. H. and Magnus, D. (2018) Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *New England Journal of Medicine* 378 (11), 981-983.

Chatterjee, S., Banerjee, S., Basu, P., Debnath, M. and Sen, S. Cuckoo search coupled artificial neural network in detection of chronic kidney disease. 2017. IEEE.

Chaudhary, A. and Garg, P. (2014) Detecting and diagnosing a disease by patient monitoring system. *International Journal of Mechanical Engineering And Information Technology* 2 (06).

Chen, G., Ding, C., Li, Y., Hu, X., Li, X., Ren, L., Ding, X., Tian, P. and Xue, W. (2020) Prediction of chronic kidney disease using adaptive hybridized deep convolutional neural network on the internet of medical things platform. *IEEE Access* 8, 100497-100508.

Chen, L. (2020) Overview of clinical prediction models. *Annals of Translational Medicine* 8 (4).

Chen, Z., Zhang, X. and Zhang, Z. (2016) Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. *International urology and nephrology* 48 (12), 2069-2075.

Chicco, D. (2017) Ten quick tips for machine learning in computational biology. *BioData mining* 10 (1), 1-17.

Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., Jasiński, M., Ł, J., Gono, R., Jasińska, E. and Bolshev, V. (2021) Prediction of Chronic Kidney Disease - A Machine Learning Perspective. *IEEE Access* 9, 17312-17334.

Chowriappa, P., Dua, S. and Todorov, Y. (2014) Introduction to machine learning in healthcare informatics. *Machine Learning in Healthcare Informatics.* Springer. 1-23.

Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L. and Batra, D. (2015) Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*.

Couser, W. G., Remuzzi, G., Mendis, S. and Tonelli, M. (2011) The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. *Kidney international* 80 (12), 1258-1270.

Denscombe, M. (2012) *Research proposals: A practical guide: A practical guide.* McGraw-Hill Education (UK).

Devi, K. G. and Balasubramanian, K. (2022) *Machine Learning and Deep Learning Techniques for Medical Science.* CRC Press.

Domingos, P. (2012) A few useful things to know about machine learning. *Communications of the ACM* 55 (10), 78-87.

Dulhare, U. N. and Ayesha, M. (2016) Extraction of action rules for chronic kidney disease using Naïve bayes classifier. *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).* 15-17 Dec. 2016.

Farkas, M. (2001) *Dynamical models in biology.* Academic press.

Foëx, P. and Sear, J. W. (2004) Hypertension: pathophysiology and treatment. *Continuing Education in Anaesthesia Critical Care & Pain* 4 (3), 71-75.

Gharibdousti, M. S., Azimi, K., Hathikal, S. and Won, D. H. Prediction of chronic kidney disease using data mining techniques. 2017. Institute of Industrial and Systems Engineers (IISE).

Ghiasi, M. M., Zendehboudi, S. and Mohsenipour, A. A. (2020) Decision tree-based diagnosis of coronary artery disease: CART model. *Computer methods and programs in biomedicine* 192, 105400.

Ghosh, P., Shamrat, F. M. J. M., Shultana, S., Afrin, S., Anjum, A. A. and Khan, A. A. (2020) Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm. *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP).* 18-20 Nov. 2020.

Hacioğlu, R. (2017) Prediction of solar radiation based on machine learning methods. *The journal of cognitive systems* 2 (1), 16-20.

Han, H., Segal, A. M., Seifter, J. L. and Dwyer, J. T. (2015) Nutritional Management of Kidney Stones (Nephrolithiasis). *cnr* 4 (3), 137-152.

Hana, E. (2021) MACHINE LEARNING BASED CHRONIC KIDNEY DISEASE PREDICTION MODEL.

Hand, D. J. (2006) Classifier technology and the illusion of progress. *Statistical science* 21 (1), 1-14.

Heron, J. (1996) *Co-operative inquiry: Research into the human condition.* Sage.

Hill, N. R., Fatoba, S. T., Oke, J. L., Hirst, J. A., O'Callaghan, C. A., Lasserson, D. S. and Hobbs, F. D. R. (2016) Global prevalence of chronic kidney disease–a systematic review and meta-analysis. *PloS one* 11 (7), e0158765.

Huang, S.-F. and Cheng, C.-H. (2020) A Safe-region imputation method for handling medical data with missing values. *Symmetry* 12 (11), 1792.

Jackins, V., Vimal, S., Kaliappan, M. and Lee, M. Y. (2021) AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing* 77 (5), 5198-5219.

Jain, D. and Singh, V. (2020) A novel hybrid approach for chronic disease classification. *International Journal of Healthcare Information Systems and Informatics (IJHISI)* 15 (1), 1-19.

Jayatilake, S. M. D. A. C. and Ganegoda, G. U. (2021) Involvement of machine learning tools in healthcare decision making. *Journal of Healthcare Engineering* 2021.

Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., Saran, R., Wang, A. Y.-M. and Yang, C.-W. (2013) Chronic kidney disease: global dimension and perspectives. *The Lancet* 382 (9888), 260-272.

Jiang, N., Fu, F., Zuo, H., Zheng, X. and Zheng, Q. (2020) A Municipal PM2. 5 Forecasting Method Based on Random Forest and WRF Model. *Engineering Letters* 28 (2).

Kalantar-Zadeh, K. and Fouque, D. (2017) Nutritional management of chronic kidney disease. *New England Journal of Medicine* 377 (18), 1765-1776.

Karasu, S., Altan, A., Bekiros, S. and Ahmad, W. (2020) A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. *Energy* 212, 118750.

Kazancioğlu, R. (2013) Risk factors for chronic kidney disease: an update. *Kidney International Supplements* 3 (4), 368-371.

Kellum, J. A. and Prowle, J. R. (2018) Paradigms of acute kidney injury in the intensive care setting. *Nature Reviews Nephrology* 14 (4), 217-230.

Khamparia, A., Saini, G., Pandey, B., Tiwari, S., Gupta, D. and Khanna, A. (2020) KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network. *Multimedia Tools and Applications* 79 (47), 35425-35440.

Khan, A., Uddin, S. and Srinivasan, U. (2019) Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes. *Expert Systems with Applications* 136, 230-241.

Khomtchouk, B. B. (2020) Codon Usage Bias Levels Predict Taxonomic Identity and Genetic Composition. *bioRxiv*.

Kore, C. and Yohannes, H. M. (2018) Prevalence of chronic kidney disease and associated factors among patients with kidney problems public hospitals in Addis Ababa, Ethiopia. *J Kidney* 4 (01), 1-5.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. and Fotiadis, D. I. (2015) Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* 13, 8-17.

Kramer, A., Pippias, M., Noordzij, M., Stel, V. S., Afentakis, N., Ambühl, P. M., Andrusev, A. M., Fuster, E. A., Arribas Monzón, F. E. and Åsberg, A. (2018) The european renal association–european dialysis and transplant association (ERA-EDTA) registry annual report 2015: a summary. *Clinical kidney journal* 11 (1), 108-122.

Ladi-Akinyemi, T. W. and Ajayi, I. (2017) Risk factors for chronic kidney disease among patients at Olabisi Onabanjo University Teaching Hospital in Sagamu, Nigeria: a retrospective cohort study. *Malawi Medical Journal* 29 (2), 166-170.

Lakshmi, K. R., Nagesh, Y. and Krishna, M. V. (2014) Performance comparison of three data mining techniques for predicting kidney dialysis survivability. *International Journal of Advances in Engineering & Technology* 7 (1), 242.

Lea, J. P. and Nicholas, S. B. (2002) Diabetes mellitus and hypertension: key risk factors for kidney disease. *Journal of the National Medical Association* 94 (8 Suppl), 7S.

Leavy, P. (2017) *Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches.* Guilford Publications.

Lee, J., Warner, E., Shaikhouni, S., Bitzer, M., Kretzler, M., Gipson, D., Pennathur, S., Bellovich, K., Bhat, Z., Gadegbeku, C., Massengill, S., Perumal, K., Saha, J., Yang, Y., Luo, J., Zhang, X., Mariani, L., Hodgin, J. B., Rao, A. and the, C. P. S. (2022) Unsupervised machine learning for identifying important visual features through bag-of-words using histopathology data from chronic kidney disease. *Scientific Reports* 12 (1), 4832.

Levey, A. S. and Coresh, J. (2012) Chronic kidney disease. *The Lancet* 379 (9811), 165-180.

Levey, A. S., Coresh, J., Balk, E., Kausz, A. T., Levin, A., Steffes, M. W., Hogg, R. J., Perrone, R. D., Lau, J. and Eknoyan, G. (2003) National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Annals of internal medicine* 139 (2), 137-147.

Levin, A. and Stevens, P. E. (2014) Summary of KDIGO 2012 CKD Guideline: behind the scenes, need for guidance, and a framework for moving forward. *Kidney international* 85 (1), 49-61.

Li, Q., Fan, Q.-L., Han, Q.-X., Geng, W.-J., Zhao, H.-H., Ding, X.-N., Yan, J.-Y. and Zhu, H.-Y. (2020a) Machine learning in nephrology: scratching the surface. *Chinese Medical Journal* 133 (06), 687-698.

Li, W., Chen, Y. and Song, Y. (2020b) Boosted K-nearest neighbor classifiers based on fuzzy granules. *Knowledge-Based Systems* 195, 105606.

Luijks, H., Schermer, T., Bor, H., van Weel, C., Lagro-Janssen, T., Biermans, M. and de Grauw, W. (2012) Prevalence and incidence density rates of chronic comorbidity in type 2 diabetes patients: an exploratory cohort study. *BMC medicine* 10 (1), 1-10.

Luyckx, V. A., Cherney, D. Z. I. and Bello, A. K. (2020) Preventing CKD in Developed Countries. *Kidney International Reports* 5 (3), 263-277.

Ma, F., Sun, T., Liu, L. and Jing, H. (2020) Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. *Future Generation Computer Systems* 111, 17-26.

Mahesh, B. (2020) Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]* 9, 381-386.

Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R. and Dakshayani, R. (2019) Chronic Kidney Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning. *2019 International Conference on Nascent Technologies in Engineering (ICNTE).* 4-5 Jan. 2019.

Mohammed Siyad, B., Manoj, M., Mohammed Siyad, B. and Manoj, M. (2016) Fused features classification for the effective prediction of chronic kidney disease. *International Journal* 2, 44-48.

Moon, K. and Blackman, D. (2014) A guide to understanding social science research for natural scientists. *Conservation biology* 28 (5), 1167-1177.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H. and Tohka, J. (2015) Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage* 104, 398-412.

Motarwar, P., Duraphe, A., Suganya, G. and Premalatha, M. (2020) Cognitive Approach for Heart Disease Prediction using Machine Learning. *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE).* 24-25 Feb. 2020.

Murton, M., Goff-Leggett, D., Bobrowska, A., Garcia Sanchez, J. J., James, G., Wittbrodt, E., Nolan, S., Sörstadius, E., Pecoits-Filho, R. and Tuttle, K. (2021) Burden of Chronic Kidney Disease by KDIGO Categories of Glomerular Filtration Rate and Albuminuria: A Systematic Review. *Advances in Therapy* 38 (1), 180-200.

Nasteski, V. (2017) An overview of the supervised machine learning methods. *Horizons. b* 4, 51-62.

Neumaier, A. (1998) Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM review* 40 (3), 636-666.

Niknejad, A. and Petrovic, D. (2013) Introduction to computational intelligence techniques and areas of their applications in medicine. *Med Appl Artif Intell* 51, 2113-19.

Nishat, M. M., Dip, R. R., Faisal, F., Nasrullah, S. M., Ahsan, R., Shikder, M. F., Asif, M. A.-A.-R. and Hoque, M. A. (2021) A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology* 18 (e6).

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y. and Cheng, C.-Y. (2020) Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology* 122, 56-69.

Nwaneri, S. C. and Ugo, H. (2022) Development of a Graphical User Interface Software for The Prediction of Chronic Kidney Disease. *Nigerian Journal of Technology* 41 (1), 175-183.

Pinto, A., Ferreira, D., Neto, C., Abelha, A. and Machado, J. (2020) Data mining to predict early stage chronic kidney disease. *Procedia Computer Science* 177, 562-567.

Piri, S., Delen, D., Liu, T. and Paiva, W. (2018) Development of a new metric to identify rare patterns in association analysis: The case of analyzing diabetes complications. *Expert Systems with Applications* 94, 112-125.

Polat, H., Danaei Mehr, H. and Cetin, A. (2017) Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. *Journal of medical systems* 41 (4), 1-11.

Powell, M., Hosseini, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H. and Wyble, B. (2020) I tried a bunch of things: the dangers of unexpected overfitting in classification. *BioRxiv*, 078816.

Rácz, A., Bajusz, D. and Héberger, K. (2019) Multi-level comparison of machine learning classifiers and their performance metrics. *Molecules* 24 (15), 2811.

Rady, E.-H. A. and Anwar, A. S. (2019) Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked* 15, 100178.

Rajeswari, K., Vaithiyanathan, V. and Pede, S. V. (2013) Feature selection for classification in medical data mining. *International Journal of Emerging Trends and Technology in Computer Science (IJETTCS)* 2 (2), 492-7.

Raju, N. V. G., Lakshmi, K. P., Praharshitha, K. G. and Likhitha, C. Prediction of chronic kidney disease (CKD) using Data Science. 2019. IEEE.

Ramalingam, V. V., Dandapath, A. and Raja, M. K. (2018) Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology* 7 (2.8), 684-687.

Ramirez-Rubio, O., McClean, M. D., Amador, J. J. and Brooks, D. R. (2013) An epidemic of chronic kidney disease in Central America: an overview. *Postgraduate Medical Journal* 89 (1049), 123-125.

Ramyachitra, D. and Manikandan, P. (2014) Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)* 5 (4), 1-29.

Refaeilzadeh, P., Tang, L. and Liu, H. (2009) Cross-validation. *Encyclopedia of database systems* 5, 532-538.

Romagnani, P., Remuzzi, G., Glassock, R., Levin, A., Jager, K. J., Tonelli, M., Massy, Z., Wanner, C. and Anders, H.-J. (2017) Chronic kidney disease. *Nature Reviews Disease Primers* 3 (1), 17088.

Sabath, E. (2015) 18 - Arsenic, Kidney, and Urinary Bladder Disorders. In Flora, S. J. S. (editor) *Handbook of Arsenic Toxicology.* Oxford: Academic Press. 429-442.

Salekin, A. and Stankovic, J. (2016) Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. *2016 IEEE International Conference on Healthcare Informatics (ICHI).* 4-7 Oct. 2016.

Salomon, D. R., Langnas, A. N., Reed, A. I., Bloom, R. D., Magee, J. C., Gaston, R. S. and a, A. A. I. W. G. (2015) *AST/ASTS workshop on increasing organ donation in the United States: creating an "arc of change" from removing disincentives to testing incentives*. Wiley Online Library.

Saunders, M., Lewis, P. and Thornhill, A. (2009) *Research methods for business students.* Pearson education.

Schaefer, A. J., Bailey, M. D., Shechter, S. M. and Roberts, M. S. (2005) Modeling medical treatment using Markov decision processes. *Operations research and health care.* Springer. 593-612.

Schmidt, D., Niemann, M. and von Trzebiatowski, G. L. The Handling of Missing Values in Medical Domains with Respect to Pattern Mining Algorithms. 2015.

Shah, D., Patel, S. and Bharti, S. K. (2020) Heart disease prediction using machine learning techniques. *SN Computer Science* 1 (6), 1-6.

Shailaja, K., Seetharamulu, B. and Jabbar, M. A. (2018) Machine Learning in Healthcare: A Review. *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA).* 29-31 March 2018.

Sharmila, A., Madan, S. and Srivastava, K. (2018) Epilepsy detection using dwt based hurst exponent and SVM, K-NN classifiers. *Serbian Journal of Experimental and Clinical Research* 19 (4), 311-319.

Sherazi, S. W. A., Jeong, Y. J., Jae, M. H., Bae, J.-W. and Lee, J. Y. (2020) A machine learning–based 1-year mortality prediction model after hospital discharge for clinical patients with acute coronary syndrome. *Health informatics journal* 26 (2), 1289-1304.

Shrivas, A. K., Sahu, S. K. and Hota, H. S. Classification of chronic kidney disease with proposed union based feature selection technique. 2018.

Sinha, P. and Sinha, P. Performance evaluation of Classification Techniques on Prediction of Chronic Kidney Disease.

Sivasankar, S., Nair, S. and Judy, M. V. (2015) Feature reduction in clinical data classification using augmented genetic algorithm. *International Journal of Electrical and Computer Engineering* 5 (6).

Stanifer, J. W., Kilonzo, K., Wang, D., Su, G., Mao, W., Zhang, L., Zhang, L., Nayak-Rao, S. and Miranda, J. J. Traditional medicines and kidney disease in low-and middle-income countries: opportunities and challenges. 2017. Vol. 37. Elsevier.

Stevens, P. E., Levin, A. and Kidney Disease: Improving Global Outcomes Chronic Kidney Disease Guideline Development Work Group, M. (2013) Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. *Annals of internal medicine* 158 (11), 825-830.

Subasi, A., Alickovic, E. and Kevric, J. (2017) Diagnosis of Chronic Kidney Disease by Using Random Forest. *CMBEBIH 2017.* Singapore, 2017//. Springer Singapore.

Sun, Y., Goodison, S., Li, J., Liu, L. and Farmerie, W. (2007) Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* 23 (1), 30-37.

Sun, Y., Wong, A. K. C. and Kamel, M. S. (2009) Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence* 23 (04), 687-719.

Sutton, R. S. and Barto, A. G. (1999) Reinforcement learning. *Journal of Cognitive Neuroscience* 11 (1), 126-134.

Tanno, L. K., Calderon, M. and Demoly, P. (2016) Supporting the validation of the new allergic and hypersensitivity conditions section of the World Health Organization International Classification of Diseases-11. *Asia Pacific Allergy* 6 (3), 149-156.

Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R. and Drăghici, S. (2007) Machine learning and its applications to biology. *PLoS computational biology* 3 (6), e116.

Tazin, N., Sabab, S. A. and Chowdhury, M. T. Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique. 2016. IEEE.

Theerthagiri, P. (2021) Prognostic Analysis of Hyponatremia for Diseased Patients Using Multilayer Perceptron Classification Technique. *EAI Endorsed Transactions on Pervasive Health and Technology* 7 (26), e5.

Tierney, W. G. and Lanford, M. (2016) Conceptualizing Innovation in Higher Education. In Paulsen, M. B. (editor) *Higher Education: Handbook of Theory and Research.* Cham: Springer International Publishing. 1-40.

Vijayarani, S. and Dhayanand, S. (2015) Data mining classification algorithms for kidney disease prediction. *Int J Cybernetics Inform* 4 (4), 13-25.

Vijayarani, S., Dhayanand, S. and Phil, M. (2015) Kidney disease prediction using SVM and ANN algorithms. *International Journal of Computing and Business Research (IJCBR)* 6 (2), 1-12.

Waezizadeh, T., Mehrpooya, A., Rezaeizadeh, M. and Yarahmadian, S. (2018) Mathematical models for the effects of hypertension and stress on kidney and their uncertainty. *Mathematical Biosciences* 305, 77-95.

Walsham, G. (1995) The emergence of interpretivism in IS research. *Information systems research* 6 (4), 376-394.

Wang, H., Naghavi, M., Allen, C., Barber, R. M., Bhutta, Z. A., Carter, A., Casey, D. C., Charlson, F. J., Chen, A. Z. and Coates, M. M. (2016a) Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The lancet* 388 (10053), 1459-1544.

Wang, M. and Chen, H. (2020) Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis. *Applied Soft Computing* 88, 105946.

Wang, V., Vilme, H., Maciejewski, M. L. and Boulware, L. E. (2016b) The Economic Burden of Chronic Kidney Disease and End-Stage Renal Disease. *Seminars in Nephrology* 36 (4), 319-330.

Wang, Z., Chung, J. W., Jiang, X., Cui, Y., Wang, M. and Zheng, A. (2018) Machine learning-based prediction system for chronic kidney disease using associative classification technique. *International Journal of engineering & Technology* 7 (4.36), 1161-1167.

Wolfe, R. A., Ashby, V. B., Milford, E. L., Ojo, A. O., Ettenger, R. E., Agodoa, L. Y. C., Held, P. J. and Port, F. K. (1999) Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *New England journal of medicine* 341 (23), 1725-1730.

Xie, Y., Bowe, B., Mokdad, A. H., Xian, H., Yan, Y., Li, T., Maddukuri, G., Tsai, C.-Y., Floyd, T. and Al-Aly, Z. (2018) Analysis of the Global Burden of Disease study highlights the global, regional, and national trends of chronic kidney disease epidemiology from 1990 to 2016. *Kidney international* 94 (3), 567-581.

Yach, D., Hawkes, C., Gould, C. L. and Hofman, K. J. (2004) The global burden of chronic diseases: overcoming impediments to prevention and control. *Jama* 291 (21), 2616-2622.

Yahaya, L., Oye, N. D. and Garba, E. J. (2020) A comprehensive review on heart disease prediction using data mining and machine learning techniques. *American Journal of Artificial Intelligence* 4 (1), 20-29.

Yashfi, S. Y., Islam, M. A., Pritilata, Sakib, N., Islam, T., Shahbaaz, M. and Pantho, S. S. (2020) Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT).* 1-3 July 2020.

8.0 Appendix

Research Proposal

UB: 20019561

AI Transforming Healthcare: The integration of Artificial Intelligence and Neuroscience for early Acute Kidney Injury Detection (AKI) - an Empirical Study from Nigeria and the UK

Table of contents

Introduction

The human kidney is an essential organ responsible for filtering blood and maintaining a balance in the body fluid. However, this primary function is affected by Acute Kidney Injury (AKI), a clinical condition where patients experience a decline in renal function through a decrease in urine output or increase in serum creatinine (Kellum and Prowle, 2018; KDIGO, 2012). AKI is prominent in critically ill hospitalized patients with a notable effect on mortality and resource consumption. AKI is usually caused by inadequate flow of blood to the kidneys, which can be as result of ailments such as sepsis and complications of other health condition thereby aggravating the patient illnesses (Gong et al., 2021). The increase in the rate of AKI is characterized by early and long-term patient morbidity and development of chronic kidney disease (CKD) (Tao Li et al., 2013).

According to International Society of Nephrology, there are about 13.3million cases of AKI yearly and this burden is on the increase in emerging countries with an annual estimation of 11.3million (ISN, 2020). In about 1.7million annual death caused by AKI, about 1.4 million occur in low and developing countries. In the UK, there are about 100,000 annual deaths where research shows that 30% can be prevented with estimated costs ranging from 434 million to 620 million yearly (NHS, 2021). This life-threatening disease is often preventable if detected early. Although early detection of AKI can limit the threatening outcome to the patients' health, the prediction and diagnosis of the clinical condition remain difficult (Le et al., 2021). The early identification of high-risk patients for allocation of insufficient clinical resources and appropriate intervention is important and has generated sizable research to promote the prediction models for AKI risk stratification (Gong et al. 2021).

In tackling the difficulties associated with AKI detection and diagnosis, Artificial Intelligence techniques have been developed and used to predict AKI. The increasing database availability of electronic health record (EHR) and electronic medical record (EMR) system are attracting substantial research for the development of AI predictive model for AKI prevention and management (Gameiro et al. 2020). AI has the capacity to execute tasks and make decision like humans (Pan, 2016). Different AI techniques such as Machine learning and Deep learning have been applied to detect AKI due to their computational power to identify patterns and make predictions.

Various researchers have made contributions to AKI prediction for improved clinical outcomes. Koyner et al. (2018) worked on gradient boosting model that predicts AKI in the emergency department and Intensive care unit. The research demonstrated high accuracy in predicting the AKI severity. Also, Tran et al. (2019) developed a Machine learning technique using k-nearest neighbour (k-NN) to predict AKI in burn patients by checking the neutrophil gelatinase-associated lipocalin (NGAL), urine output, serum creatinine, and other laboratory measurements in the first 24 hours of admission, the method functioned well with more than 90% accuracy for detecting AKI showing that the effect of urine output and creatinine for predicting AKI can be supported by a ML algorithm using K-NN.

Statement of Research Problem

The cases of AKI linked with high mortality and morbidity is high in both the developed and the developing countries especially in Nigeria and the UK. Treatment of AKI can be expensive leading to high deterioration. Millions of people die yearly from preventable diseases if detected earlier (ISN, 2020. The effect of AKI can destroy primary functions and structures of the kidney. It is therefore identified that the high hospital mortality due to AKI will create a change in the understanding of this disorder. AKI causes abnormality in the immune, coagulation and central nervous system resulting to multisystem failure in the patient. By the aims and objectives mentioned below, the life-threatening disease will be detected earlier enough for clinical intervention. As a result of its economic importance, bringing awareness to the impact, detection and implications of AKI will be beneficial to the society.

The Aim and Objectives

The aim of the project is to develop a predictive model for Acute Kidney Injury (AKI) detection in the UK and Nigeria.

To accomplish the research aim, the following objectives will be needed:

1. To review existing literature on Acute Kidney Injury
2. To review the existing literature on the relationship between Artificial Intelligence (AI) and neuroscience for AKI detection.
3. Understand and extend the theoretical and practical concepts of AI techniques in Kidney diagnosis
4. Based on study gap, to carry out research using a secondary retrospective data to validate research studies and investigate factors aiding AKI disease and identify the potential AKI predictors.

5. To analyse and synthesize data using quantitative analytical research
6. To analyse and synthesize data using AI tools (Machine Learning and Artificial Neural Networks) to revolutionize predictions and improve clinical outcome.
7. Conduct a comparative study between AKI detection in Nigeria and the UK
8. To analyse, make recommendations and conclusion based on the results of research analysis on the AKI detection.

Justification of study

Literature study indicates the sizable work carried out on AKI detection. However, in addition to the pursuance of my education, the study will be beneficial to the society in improving clinical outcome and serve as a baseline information for future researchers willing to conduct relevant study. The analysis will be great significance for the UK and Nigeria ministry of Health in general.

This research will make the following additional contributions:

1. The study will contribute knowledge in AKI detection
2. The study will contribute knowledge in AI adoption for clinicians by identifying.
3. Identify more predicting features of AKI from a developing and developed context such as Nigeria and UK, respectively.
4. The study will give more insights for health patients and clinicians to understand AI techniques and benefits.

Research Focus

This research focuses on the AKI detection in the UK and Nigeria health sector.

Research Questions

The following research questions will be answered by this study

1. What are the major predictors of AKI?
2. What are the patients’ needs awareness level toward AKI?
3. How important is the influence of Artificial Intelligence (AI) towards healthcare?
4. How important is the influence of recommendation toward AI adoption in healthcare?
5. Does the number of AKI increase or decrease over a period in Nigeria and the UK?
6. Does the number of AKI depend on country?
7. In what country does AKI become more rampant?

Literature Review

The modernization, complexity in healthcare and increasing number of treatment options interfere with the decision-making ability of optimal treatment. This study aims to develop an AI enabled framework to address the challenges through the available clinical data and interconnection among different healthcare system. This framework will reason like a doctor in making decision making especially in AKI detection. Machine Learning as a algorithm capable of drawing inferences and identifying patterns will aid the framework. AKI is common and characterized with mortality. Emerging evidence indicates that the severe acute renal dysfunction has a relationship with survival in both inpatients and outpatients. The review of literature conducted to note key distinct unanswered question.

Following the conceptual framework of AKI there is a need for recommendation in the changes that appeared in the 3 stages of AKI which permits testing and development of research questions.

Does kidney damage without AKI lead to poor clinical outcomes?

Does AKI without proof of kidney damage lead to poor clinical outcomes?

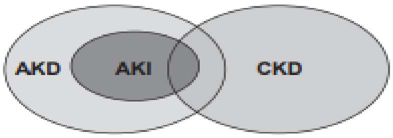
Does the degree of kidney damage predict poor outcome?

The definition of AKI is based on the reduction in kidney function that informs clinical attention. The conceptual framework shows a research recommendation for AKI staging.

According to the kidney disease: Improving Global Outcomes (KDIGO), the conditions affecting kidneys can be classified acute or chronic depending on the span. AKI is one of the acute kidney diseases (AKD) and can occur without any other kidney diseases as shown in fig 1 (Levey et al. 2011). AKI is a subset of CKD. There are 3 stages of AKI that are formed on Risk, Injury, Failure, Loss and End stage kidney disease (RIFLE) (Bilgili et al., 2014), and Acute Kidney Injury Network (AKIN) which are stage 1, 2, and 3 as shown in figure 2 (Fatoni and Kestriani, 2018). This disease is damaging, and potentially improvable through early detection (ISN, 2012). According to Wang et al. (2013), based on the serum creatinine (SCr) and urine output, AKI can be defined as:

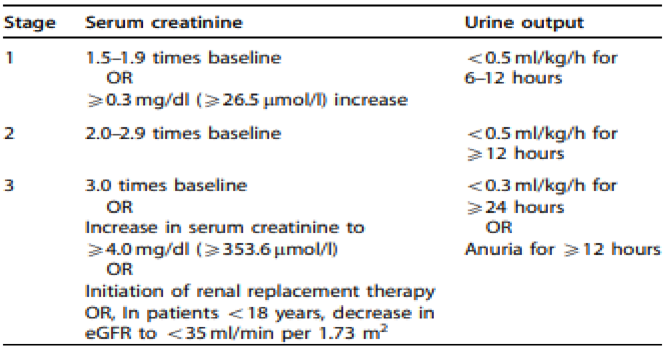
Increase in SCr by X0.3 mg/dl (X26.5 lmol/l) within 48 hours; or increase in SCr to X1.5 times baseline, which is known or presumed to have occurred within the prior 7 days; or Urine volume o0.5 ml/kg/h for 6 hours (KDIGO, 2012)

Figure 1. Overview of AKI, CKD, and AKD



Each stage with varying burden of illness and high cost of managing the disease is amendable to early detection and potential prevention. The current clinical practice has the capability to minimize the variations, give improved outcome and while reducing cost.

Figure 2. Staging of AKI



Stage	Serum creatinine	Urine output
1	1.5–1.9 times baseline OR ≥0.3 mg/dl (≥26.5 μmol/l) increase	<0.5 ml/kg/h for 6–12 hours
2	2.0–2.9 times baseline	<0.5 ml/kg/h for ≥12 hours
3	3.0 times baseline OR Increase in serum creatinine to ≥4.0 mg/dl (≥353.6 μmol/l) OR Initiation of renal replacement therapy OR, In patients <18 years, decrease in eGFR to <35 ml/min per 1.73 m^2	<0.3 ml/kg/h for ≥24 hours OR Anuria for ≥12 hours

Prior to the current clinical approach, various researchers have contributed to the improvement of AKI diagnosis.

AI as the machine intelligence like that of humans is transformative and set to revolutionize the healthcare industry including its subfield such as ML and Artificial neural networks. Where Machine Leaning as the statistical models has been used in medical field to recognize patterns, classify, and predict algorithms (Hamet and Tremblay, 2017). In the integration of AI and neuroscience, this study will be adopting the Neural Networks context. The Neural networks (NNs) is the kind of ML that is influenced by the human brain. It is made up of linked units of neurons that process information and react to external inputs. The NNs applies large network of data huge data in NN interrelatedly sending data to each other. The Artificial Neural

Networks (ANN) is however a computing system structurally developed to imitate biological neural networks (Xie et al., 2020). The high magnitude of performance and computational resources available for data calculation, processing and training make ML an ANN make them suitable for prediction.

Koyner et al. (2018) worked on gradient boosting model that predicts AKI in the emergency department and Intensive care unit. The research demonstrated high accuracy in predicting the AKI severity. The model involved data generated from patients with different demographics, laboratories and vital signs and provide great AUROC value that is above 0.900 suitable for renal replacement therapy within 72 hrs

Tran et al. (2019) developed a Machine learning tool using k-nearest neighbour (k-NN) to detect AKI in burn patients by calculating the neutrophil gelatinase-associated lipocalin (NGAL), urine output, serum creatinine, and N-terminal B-type natriuretic peptide (NT-proBNP) measured within the first 24 h of admission, the method worked greatly giving above 90% AKI prediction accuracy showing that the action of urine output and creatinine for predicting AKI can be supported by a Machine learning algorithm using K-NN (Tran et al., 2019). The limitation is on the $^{\text{retrospective}}$ feature of the analysis and the sample size.

Maurya et al. (2014) conducted a statistical analysis on the rate of kidney renal failure while taking into consideration age and sex as variables with 10 years secondary data collected from University of Maiduguri Teaching Hospital (UMTH) medical record (1998-2007). The research was conducted in Nigeria and was analyzed using t-test and chi square as the statistical tool. The finding shows a significant difference between the male and female sexes with high positive correlation between male and female.

Adedamola et al. (2019) studied the AKI among pediatric patients on admissions in a hospital situated in southwest Nigeria. The researchers conducted a prospective study of paediatric AKI in developing country and identified the importance of providing renal replacement therapy (RRP) for AKI low-resource settings while developing evidence based clinical approach to manage AKI among inpatients. This is potentially to reduce preventable deaths. The researchers found sepsis, malaria, and primary renal disorders as the most common causes of AKI with a limitation of detecting the statistically significant relationship of AKI with inpatient mortality due to the sample size and incomplete data (Adedamola et al., 2019).

Kate et al. (2016) conducted comparative research on prediction and detection of AKI using data of patients above 60 years and four different ML model including support vector machines, naïve Bayes, linear regression, and decisions tree. Demographics and laboratory measurements were considered in the models with the classifiers achieving AUCs that ranges from 0.621 to 0.664. linear regression has the highest detection performance and in prediction.

Despite the efforts, there is limited research on AKI detection using AI technique in Nigeria. In the UK, few research has been conducted on a comparative study of AKI detection between a developed and a developing country. This study plan to link the gap by conducting research for clinical decision support.

Research Methodology

This study will adopt quantitative research and a retrospective analysis of the clinical data of all adult AKI patients from age 18 to 90 from 2015 to 2020 in Nigeria and the UK would be conducted. Analysis would be done using python.

Machine Learning and Artificial Neural Network will be used for the prediction due to their high predictive performance and computational accuracy. The study utilizes a machine learning algorithm in conjunction with ANN methodologies to process data and make a prediction on the type of AKI predicting features. ANN improves in performance with more data set which makes it a suitable technique for the study.

Data Description

Secondary patients' data captured from UKRR which integrates detailed clinical data of UK patients will be used. Also, a retrospective data sourced from Nigeria hospitals will be used for the comparative study. The data will be de-identified and adhere to the EU General Data Protection Regulation (GDPR).

Following the literature review and deliberation with health professionals, the following potential predictors will be selected for analysis

Demographics: age, sex, ethnicity

Comorbidity: sepsis, diabetes, infection, hypertension, heart failure, chronic kidney disease.

Vital signs: heart rate, pulse rate, blood pressure, respiration rate, body temperature, body mass index (BMI)

Medications: vancomycin, mechanical ventilation, diuretics.

Laboratory measurements: which includes blood urea nitrogen, estimated glomerular filtration rate (eGFR), SCr, glucose, lactate.

Other features such as smoking and exercise, obesity, and special diet.

Data Analysis

The obtained clinical data will be analysed using ML and Artificial Neural Network enabled python to check the validity and effectiveness of the data for AKI prediction

Ethical considerations

This research will observe a full compliance to the rules and procedures from Health Insurance Portability and Accountability act (HIPAA) privacy rule as well as GDPR to obtain data by adhering to privacy, confidentiality and ensuring that the data is used for academic research purposes. GDPR (2018), the use of personal data must follow data protection principles by lawfully, fairly, and transparently using the obtained data.

Timeline

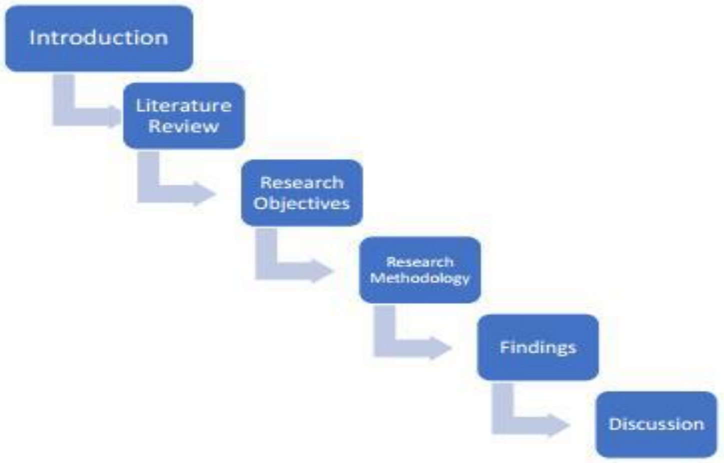
There is a projection of 9 months extensive study and analysis to complete this project.

Expected outcome

By the end of this project, the developed model will be able to predict likely AKI to improve clinical outcome. The study will be a base line for future researchers.

Overview of Remainder of Study

The research will contain five chapters, with proposal as the first chapter, literature review will be the second chapter. Literature review is segmented into important definitions, models and theoretical foundations. And the chapter will evaluate different models and theoretical frameworks relevant to AKI detection. Research methodology appears in the third chapter which will reflect how the study will be conducted in Nigeria and UK including sampling size, data collection and analysis, and justification. Ethical consideration and limitation will be included. The fourth and fifth chapters will be results of the study and discussion respectively. The results will be based on theories, hypothesis and study objective. The conclusion summarizes the key findings by responding to research questions and give recommendations for future execution.



Research Pitfalls

Poor electronic medical record (EMR) in developing country can affect secondary data collection. There is already a consultation with health professionals in Nigeria to get the required data for analysis.

References

Ademola, A.D., Asinobi, A.O., Ekpe-Adewuyi, E., Ayede, A.I., Ajayi, S.O., Raji, Y., Salako, M.J., Zappitelli, M. and Samuel, S.M. (2019) Acute kidney injury among paediatric emergency room admissions in a tertiary hospital in South West Nigeria: a cohort study. Clinical Kidney Journal, 12(4), 521-526.

Bilgili, B., Haliloglu, M. and Cinel, I. (2014) Sepsis and Acute Kidney Injury. 42(6), 294-301.

Fatoni, A.Z. and Kestriani, N.D. (2018) Acute Kidney Injury (AKI) pada Pasien Kritis. 36(2)

Gameiro, J., Branco, T. and Lopes, J.A. (2020) Artificial intelligence in acute kidney injury risk prediction. Journal of Clinical Medicine, 9(3), 678.

Gong, K., Lee, H.K, Yu, K., Xie, X. and Li, J (2021)A prediction and interpretation framework of acute kidney injury in critical care, Journal of Biomedical Informatics, 113, 103653

Hamet, P. and Tremblay, J. (2017) Artificial intelligence in medicine. Metabolism, 69, S36-S40

International Society of Nephrology (2012) Section 2: AKI definition, Kidney international Supplements, 2(1) 19-36.

Kate, R.J., Perez, R.M., Mazumdar, D., Pasupathy, K.S., and Nilakantan, V. (2016) Prediction and detection models for acute kidney injury in hospitalized older adults
BMC Medical Informatics and Decision Making, 16 (1), 39

Kellum, J.A., Prowle, J.R. (2018) Paradigms of acute kidney injury in the intensive care setting. Nature Review of Nephrology 14, 217–230.

Kidney Disease: Improving Global Outcomes (2012) Kidney international supplements. Official Journal of the International Society of Nephrology, 2 (1), 14-23.

Koyner, J.L., Carey, K.A., Edelson, D.P. and Churpek, M.M. (2018) The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model. Critical Care Medicine, 46, 1070–1077

Le, S., Allen, A., Calvert, J., Palevsky, P.M., Braden, G., Patel, S., Pellegrini, E., Green-Saxena, A., Hoffman, J. and Das, R. (2021) Convolutional neural network model for intensive care unit acute kidney injury Prediction. Kidney International Report, 6(5), 1289-1298

Levey, A.S., Jong, P.E., Coresh, J., Nahas, M.E., Astor, B.C, Matsushita, K., Gansevoort, R.T., Kasiske, B.L. and Eckardt, KU. (2011) The definition, classification, and prognosis of chronic kidney disease: a KDIGO Controversies Conference report. Kidney International, 80, 17–28.

Maurya, V.N., Singh, V.V. and Yusuf, M.U (2014) Statistical analysis on the rate of Kidney (renal) failure. America Journal of Applied Mathematics and Statistics, 2(6A), 6-12.

Pan, Y. (2016) Heading toward artificial intelligent 2.0. Engineering 2(4), 409-413

Tao Li, P.K., Burdmann, E.A. and Mehta, R.L. (2013) Acute kidney injury: global health alert. Kidney International, 83(8), 372-376.

Tran, N.K., Sen, S., Palmieri, T.L., Lima, K., Falwell, S., Wajda, J., Rashidi, H.H. (2019) Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: A proof of concept. Burns, 45, 1350–1358.

Wang, H.E., Jain, G., Glassock, R.J., Warnock, D.G (2013) Comparison of absolute serum creatinine changes versus Kidney Disease: Improving Global Outcomes consensus definitions for characterizing stages of acute kidney injury. Nephrology Dialysis Transplantation, 28(6), 1447-1454.

.Xie, G., Chen, T., Li, Y., Li, X. AND Liu, Z. (2020) Artificial intelligence in nephrology: How can artificial intelligence augment nephrologists' intelligence? Kidney Disease, 6, 1-6.