Chronic Kidney Disease Prediction Using Machine Learning Techniques

By: Dibaba Adeba Debal

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Masters of Science in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

January, 2021

Chronic Kidney Disease Prediction Using Machine Learning Techniques

By: Dibaba Adeba Debal

Advisor: Tilahun Melak (PhD.)

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing.

Presented in Partial Fulfillment of the Requirement for the Degree of Masters of Science in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

January, 2021

Declaration

I hereby declare that this MSc thesis is my original work and has not been presented for a degree in any other university, and all sources of material used for this thesis have been duly acknowledged.

Name: Dibaba Adeba

Signature: _____

This MSc thesis has been submitted for examination with my approval as a thesis advisor.

Name: Tilahun Melak (PhD.)

Signature: _____

Date of Submission: _____

Approval of Board of Examiners

We, the undersigned, members of the Board of Examiners of the final open defense by Dibaba Adeba have read and evaluated his thesis entitled "Chronic Kidney Disease Prediction Using Machine Learning Techniques" and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the Degree of Masters of Science in Computer Science and Engineering.

_____	_____	_____
Supervisor /Advisor	Signature	Date
_____	_____	_____
Chairperson	Signature	Date
_____	_____	_____
Internal Examiner	Signature	Date
_____	_____	_____
External Examiner	Signature	Date

ACKNOWLEDGMENTS

First and foremost I would like to thank God for all his blessing from the beginning to the end of this research. I am extremely grateful to my advisor Dr. Tilahun Melak for his friendly approach, invaluable advice, critical comments, and commitment to guide, and continuous support that greatly improved this thesis work.

I would also like to acknowledge the members of the Smart Software SIG of the CSE department. I am grateful to the SIG supervisor Dr. Mesifn Abebe for his continuous commitment to guide and support this research from its inception to completion. I would like to thank all my classmates who are working on their master's thesis for their cherished time spent together, ideas, and support until the end of my thesis.

I would also like to thank the Staff of the renal department for helping, allowing, and providing me valuable information to collect the data. I would also like to thank Mr. Negawo Cheka for his cooperation and willingness to help me through data collection. My appreciation also goes out to my family and friends for their encouragement and support all through my study. I would also like to express my gratitude and appreciation to Duresa Tamirat (PhD. Candidate at ASTU) and Tefera Regasa (PhD. Candidate at ASTU) for their support and continuous encouragement throughout my study.

Last but not least, my thanks goes to my lovely brother Dejene Adeba, whom without this would have not been possible and his wife Birke Terefa, for their every support and advice in my life.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
BMI	Body Mass Index
CAD	Coronary Artery Disease
CFS	Correlation based Feature Selection
CKD	Chronic Kidney Disease
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration
COVID_19	Corona Virus
CPU	Central Preprocessing Unit
CRF	Chronic Renal Failure
CSP	Common Spatial Pattern
CSV	Comma Separated Value
CV	Cross Validation
CVD	Cardiovascular Disease
DM	Diabetic Mellitus
DT	Decision Tree
ESRD	End Stage Renal Disease
FCBFS	Fast Correlation-Based Feature Selection
FN	False Negative

FP	False Positive
FS	Feature Selection
GFR	Glomerular Filtration Rate
HTML	Hypertext Markup Language
HTN	Hypertension
IG	Information Gain
KBS	Knowledge Based System
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LG	Logistic Regression
MBF	Markov Blanket Filter
MDRD	Modification Of Diet In Renal Disease
ML	Machine Learning
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
NB	Naïve Bayes
NCD	Non-Communicable Diseases
NKF-KDOQI	National Kidney Foundation's Kidney Disease Quality Outcome Initiative
OVR	One Versus Rest
PNN	Probabilistic Neural Network

RBC	Red Blood Cell
RBF	Radial Basis Function
RF	Random Forest
RFECV	Recursive Feature Elimination with Cross Validation
RRT	Renal Replacement Therapy
SBE	Sequential Backward Elimination
SFS	Sequential Forward Selection
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UCI	University of California, Irvine
UFS	Univariate Feature Selection
WBC	White Blood Cell
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

ABSTRACT

Chronic kidney disease is a major challenge for health care systems all over the world consuming a high percentage of health care budgets, mainly affects low-income countries. Early prediction of the stages and prevention of this disease based on severity level is one of the most important problems of health sectors especially in developing countries like Ethiopia. Machine learning plays a key role in analyzing huge medical data and solve complex problem for early prediction of diseases.

This study aimed to develop a chronic kidney disease prediction using machine learning techniques to predict the severity of the disease as notckd, mild, moderate, severe, or ESRD and also presence or absence of the disease as ckd or notckd. The data for this study purpose was collected from St. Paulo's Hospital. This research study have been conducted an experimental approach in order to determine the best performing model. This study used python programming language for the implementation purpose. The study employed three models such as Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT) and two feature selection methods such as analysis of variance (ANOVA) and recursive feature elimination using cross validation (RFECV). First, the models built on the whole dataset for both binary class and five class, and then feature selection methods applied to both datasets. Evaluation of the models was done using 10-fold cross-validation and classification performance was used in order to compare the models. Binary models and multiclass models were developed based on the two datasets. The results of this study show that Random forest based on recursive feature elimination with cross validation record better performance compared to support vector machine and decision tree models based on accuracy and F1-score. It record accuracy of 99.8 for binary class and 79.0% for multiclass and F1-score of 99.8% for binary class and F1-score of 77.9 % for multiclass. Finally, severity prediction model is recommended for the experts to provide appropriate prevention, treatment and diet recommendation based on the disease severity.

Keywords: Chronic Kidney Disease, Machine Learning, Random Forest, Feature Selection.

CHAPTER ONE

INTRODUCTION

1.1 Background

Non-Communicable Diseases (NCDs), chronic, especially chronic kidney diseases (CKD), cardiovascular, hypertension, and diabetes mellitus have now taken a place of communicable diseases that become a serious public health issue and economic cost issue worldwide, consuming a high percentage of health care budgets. CKD is the non-communicable disease (NCDs) that has significantly contributed to increased morbidity, mortality, and an admission rate of patients throughout the world. Chronic kidney disease is a condition in which the kidney structure is unusual or function reduces for 3 months and more with a reduced glomerular filtration rate. Nowadays, (CKD) is becoming a challenging public health problem with increasing prevalence worldwide, highly burdening low-income countries, where detection, prevention and treatment rates remain low[1]. According to the global burden disease project, CKD has now become a quickly expanding and killing disease all over the world. A report from 1990 to 2013 indicated the globally yearly life loss of CKD increased by 90% and it is known as the 13th leading death cause in the world [2]. Currently, 850 million people throughout the world are likely to have kidney diseases from different factors and the report world kidney day of 2019 indicated that the disease become a serious disease from which at least 2.4 million dies per year and now it is the 6th fastest-growing cause of death worldwide.[3].

These days, chronic kidney disease is becoming a serious public health problem and increasing than previously thought in Ethiopia affecting hundreds of thousands of people irrespective of age, sex. Peoples are suffering from this disease since there is lack of safe water, appropriate diet, and physical activities. Besides, most community living in rural area have no knowledge about the CKD due to lack of appropriate facilities such as hospitals, and enough medical doctors are not available. According to the WHO data published in 2017, the number of deaths in Ethiopia because of kidney disease reached 4,875 or 0.77% of total deaths that ranks Ethiopia 138 in the world and the age-adjusted death rate is 8.46 per 100,000 of the population and the death rate increased to 12.70 per 100,000 that ranks Ethiopia 109 in 2018 [4]. This World Health Organization (WHO) statistics report indicates

that kidney disease is increasing fast in Ethiopia from time to time. CKD is becoming an important contributor to morbidity and mortality caused by other diseases and risk factors including hypertension, cardiovascular disease, diabetes, obesity, as well as communicable diseases such as malaria, tuberculosis, and hepatitis. Furthermore, socio-demographic conditions such as age, sex, educational status, marital status, level of income contributes to the progress of the disease to end stage. Additionally, individual behavioral characteristics like a smoking habit, use of alcohol, use of traditional medicine, use of anti-pain medicine, and use of other addictive drugs have a great effect on the progress of CKD. The severity of the CKD can be identified by Glomerular Filtration Rate (GFR), which used to know how healthy the kidneys perform blood filtration, removing excess wastes and fluids by calculating it from serum creatinine values to age, sex, and ethnicity. Five stages of CKD were identified by national kidney foundation based on the abnormal kidney function and reduced Glomerular Filtration Rate (GFR), which measures a level of kidney function. There are different prevalent formulas such as MDRD-Equation, CKD-EPI formula, and Mayoquadratic formula used for estimating GFR.

The mildest stage (stage 1 and stage 2) is known with only a few symptoms and stage 5 on the other hand considered as end-stage or kidney failure and can potentially harm if it does not treat well. The Renal Replacement Therapy (RRT) cost needed for total kidney failure weighs much heavy on the health care budgets and it is even very expensive and even the treatment is not available in most developing countries like Ethiopia. In developing countries, including Ethiopia there is no awareness of early detection or identification of kidney diseases. Due to lack of facilities and unaffordable costs of treatment in Ethiopia, we always hear that peoples are asking for help through public media to get treatment of kidney renal replacement abroad and we hear death due to chronic kidney complications. Moreover, the management of kidney failure and its complications is very difficult in developing countries due to several reasons such as the shortage of facilities and physicians, and the high costs to get the treatment [5][6]. Therefore, early identification of CKD is very essential to minimize the economic burden of the population and also to bring quality service on treatment to the population[7]. If detected early, CKD could be treated and the best recommendation will be given to save the patient's life and make the patient stay alive more than expected time, so that it is possible to reduce complications of the health problem as well as the medical process, healthcare costs and time it takes to treat the patients.

Machine Learning is the AI subfield, which is currently becomes an interesting research area in health sectors for disease prediction and treatment. Machine learning has the most significant role in disease prediction based on patient history data with a high-performance rate and is used to help decision-makers to gather and understand the information easily. Machine learning can be supervised, unsupervised, or Semi-unsupervised. This study focuses on chronic kidney disease prediction using machine learning models based on the dataset collected from St. Paulo's Hospital in Ethiopia with five classes: notckd, mild, moderate, severe, and ESRD and binary classes: ckd and notckd by applying machine-learning models. Most previously conducted researches focused on two classes, which make treatment recommendations difficult because the type of treatment to be given is based on the severity of CKD. Predictive analysis with the help of efficient numerous machine learning algorithms helps to predict the disease more accurately and correctly help to treat patients within a few seconds and reduce waiting time [8]. Based on basic features such as age, gender, blood pressure, serum creatinine, diabetes, RBC count, WBC count, chloride, sodium, potassium, hemoglobin, hypertension, mean cell volume, platelet count, heart disease, anemia this research aims to classify the patient as notckd, mild, and moderate, sever or ESRD for five classes and ckd or notckd for binary class. Based on the result of the prediction necessary treatment and prevention method will recommend for the patient by the expert. Thus, different classifier algorithms are compared as per their prediction results and the best model is recommended based on the evaluation result to be used for the severity prediction of chronic kidney disease.

1.2 Motivation of the Study

The report from different health organizations and different statistical analysis indicates that the number of people suffers from CKD is increasing from time to time worldwide. Especially in developing countries like Ethiopia, chronic kidney disease is becoming a serious disease. Many researchers have conducted different research on how to manage the progress of chronic kidney disease locally and globally. Most globally conducted research on chronic kidney disease prediction focuses on the identification of the presence or absence of the disease but not on the identification of the severity of the disease. In Ethiopia, different statistical analysis research was conducted on the prevalence and the quality of life of the patient with chronic kidney disease. Furthermore, in Ethiopia, the treatment of kidney disease is given in only specified hospitals to which thousands of people run to get the

treatment and there is no enough nephrologists and computer-aided diagnosis to make the fast diagnosis. With this concern, nowadays, computer technology and machine learning techniques are highly in use to develop software to assist doctors in deciding the status of chronic kidney disease in the preliminary stage [9]. One paper was conducted by Siraj Mohammed and Tibebe Beshah [10] in Ethiopia on the treatment and diagnoses of chronic kidney disease using KBS and the machine-learning technique. However, the focus of the paper was only on the KBS. Therefore, we were inspired to work on the CKD prediction using machine learning techniques because machine learning can analyze medical data and predict the severity of the disease correctly by simply inserting common clinical predictors. Using machine-learning technique assists and enables the experts to make fast decisions and take necessary action to treat patients and give an appropriate recommendation based on the severity of the disease. Preparing the new dataset from locally collected patient history data, the study will contribute to the building of a better CKD prediction to save the time and cost of treatment with the help of machine learning predictive models.

1.3 Statement of the Problem

Throughout the world, regardless of geographic location, population size, or level of social and economic development, non-communicable diseases are highly responsible for the death and disability of millions in all countries[11]. For many years, non-communicable diseases were not been considered as priority problems in many low-level income countries including Ethiopia[12]. Chronic Kidney disease is a non-communicable disease that nowadays killing millions all over the world. Currently, from the global burden disease project, CKD is quickly expanding and killing disease throughout the world and consuming a huge proportion of health care finances. The prevalence of chronic kidney disease was raised rapidly since 1990 globally[13]. Chronic kidney disease can progress to kidney failure if it does not detect and treated early which consumes a huge budget and even difficult for developing countries like Ethiopia to get treatment locally.

In developing countries like Ethiopia, the patients get treatment after a serious case due to lack of medical resources, specialist doctors, and low income for treatment [14]. Treatment option of chronic kidney disease is not readily affordable in most sub-sub-Saharan Africa including Ethiopia. For many years, the treatment and diagnoses of chronic kidney disease in Ethiopia have not been studied and there is no national strategy for the prevention and

treatment of patients with chronic kidney disease [10]. Furthermore, many people do not know about the disease and only known that they had this kidney disease when their kidneys become chronic and affected their lives. They think that all symptoms that they have and the pain that they feel were just normal fever and can be fine when they take some drugs without being diagnosed. That is bad thinking and several factors encourage this bad attitude such as the traveling distance to the clinic or hospital which takes a lot of time and waiting time especially in the governmental hospital to get treatment. Thus the early detection enables the expert to diagnose the patients fast to reduce waiting time and can slow the progress and even stop the CKD and helpful in the management of other risk factors.

This day's modern healthcare system is faced with the problem of acquiring, analyzing, and applying a large amount of knowledge necessary to solve complex clinical issues because the clinical data is very complex. A major challenge facing healthcare institutions (hospitals, medical centers, etc.) is the provision of quality services at affordable costs because the quality service implies the timely diagnosis of patients and the timely management of treatments that are efficient and effective. Using the machine learning technique will solve the challenge of the health care system on disease prediction by training the machine with appropriate data. Thus, the proposed model could help physicians; health professionals determine the severity of chronic kidney disease. In Ethiopia, Siraj Mohammed and Tibebe Beshah[10] conducted their research on CKD prediction using KBS for the first three stages using small size data. Therefore, the contribution of this paper is to prepare a chronic kidney disease dataset with more instances that contain relevant features to predict chronic kidney disease with the severity level using a machine learning classifier algorithm and feature selection methods.

1.4 Research Questions

This study addresses the problem by answering the following research question:

RQ1: What are the relevant features to consider to design effective chronic kidney disease datasets in applying machine learning models?

RQ2: Which machine learning model and feature selection method is best appropriate for chronic kidney disease prediction?

RQ3: How to measure the performance of chronic kidney disease prediction models?

1.5 Objective

1.5.1 General Objective

The general objective of this study is to develop a Chronic Kidney Disease prediction using machine-learning techniques, which can help the practitioners to diagnose the patients fast.

1.5.2 Specific Objective

- To review related works on chronic kidney disease and its prediction using machine learning techniques.
- To collect chronic kidney disease patient history data and prepare the dataset.
- To identify the best feature selection method for chronic kidney disease prediction.
- To evaluate the performance of the chronic kidney disease prediction models and recommend the best performing model in detecting chronic kidney disease based on the evaluation result.
- To implement a chronic kidney disease prediction prototype.

1.6 Scope and Limitation of the Study

1.6.1 Scope

There are several approaches to develop the disease prediction system. This study focuses on applying the machine-learning classification algorithms to predict the risk level of disease as notckd, mild, moderate, severe, and ESRD that enables the experts to give appropriate treatment recommendations based on a dataset developed from locally collected patient history data from St. Paul's Hospital. The binary class dataset is also used to identify the presence or absence of the disease. In this study, the prediction of chronic kidney disease performed using 18 features on original dataset and supervised machine learning classification algorithm with feature selection.

1.6.2 Limitation

The research comes across different limitations on a different phase of the research processes. The most difficult thing in Ethiopia especially in the medical area, it is very difficult to find the organized data in electronic format and it is time-consuming to convert the data from manually recorded to electronic data. In Ethiopia, it is difficult even to collect

data because different hospitals are not interested to give the data for privacy purposes. Data were only collected from one hospital. However, building the model with a dataset collected from various medical centers and external validation could have better performance and more reliable. Unfortunately, COVID-19 came to our country while collecting the data. Thus, we could not collect data as much as we were intend to collect for our dataset preparation. We used the first stage and second stage of the CKD as mild stage due to data insufficiency of the first stage. Additionally, there is no standard rule to select relevant features to predict chronic kidney disease. Besides, there are no conducted studies in Ethiopia with the same dataset for comparison of the relevant features and performance of models for the prediction of chronic kidney disease and even globally, the researcher found only one study on stages prediction.

1.7 Application of Results

The main beneficiaries of the study are the health sectors. Mainly the practitioners or experts benefit from this prediction system as the system assists the experts on behave of predicting the severity of the disease depending on the given input data of patient's. Additionally, the patients also benefited from the system, since the predictive model reduces the number of predictive features to be used in CKD prediction. Consequently, the reduced predictive features minimize the cost for treatment and waiting time of the patients in the hospital.

1.8 Organization of the Thesis

This thesis work is organized into seven chapters.

Chapter 1 Introduction: describes the introduction and covers the motivation, the statement of problems, the objective of the study, the scope, limitation, and application of the study, and the organization of the study.

Chapter 2 Literature Review: states the literature pertinent to this study, reviews background information on several key concepts about chronic kidney disease, and algorithms used in chronic kidney disease prediction: feature selection and machine learning classifier, and finally, the related works are discussed.

Chapter 3 Methodology: the third chapter states, research methodology used in this study, methods used to build the dataset, methods used to develop chronic kidney disease prediction

system which are data preprocessing, feature selection methods, and machine learning classifier algorithm's, and finally, it states the evaluation methods.

Chapter 4 Proposed Solution: discusses the proposed solution for chronic kidney disease prediction, which is the proposed data preprocessing, feature selection, machine learning classifiers, and evaluation methods.

Chapter 5 Implementation: presents the implementation and experimentation of the proposed solution. Including working environment, dataset description, also the implementation of preprocessing, feature selection implementation, model implementations, and evaluation models.

Chapter 6 Results and Discussion: discusses the result of the dataset class distribution, feature selection, binary class, and five-class models and discusses the major results obtained by comparing the models based on their performance on both binary class and five-class dataset without and with feature selection.

Chapter 7 Conclusion: describes conclusion and forwards recommendation and feature works for further investigation of this research study.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORKS

In this chapter, the relevant literature related to the main objective of the paper is reviewed to clearly state and furtherly explore the research problem by providing a technical review on chronic kidney disease and existing method to detect chronic kidney disease. Started by stating the background chronic kidney disease, chronic kidney stages, and overview of machine learning, machine learning and health center, classification algorithm, and previously worked studies to solve detection problem chronic kidney disease.

2.1 Chronic Kidney Disease and Its Risk Factors

The kidneys are two bean-shaped organs, which are found to the right and left side inside the human being. The kidney is a very important and essential organ of the human being, which is used, for cleaning the wastes from the human body. Since it is a very essential part of the human being, the damage of the kidneys causes different risks from low risk to death. Kidney disease means when the kidneys are unable to filter blood the way the normal kidneys should. Kidney disease can affect the body's ability to clean the blood, filter extra water out of the blood, and help control the blood pressure. It can also harm red blood cell production and vitamin D metabolism important for bone health. There are five types of kidney disease. From these types, Chronic Kidney Disease and Acute Kidney Injury Disease are the most common kidney diseases. This study aims to develop a chronic kidney disease (CKD) prediction using machine learning.

Chronic Kidney Disease is a progressive kidney disease that kills millions of people silently all over the world regardless the age and sex. CKD disease is a condition when the kidney damaged and unable to filter wastes as much as the normal kidneys because wastes are built up in the kidneys for several time[15]. This can also cause other complications such as cardiovascular disease (CVD), anemia, and bone disease. It is a decreased function of the kidney over several years; it often goes undetected and undiagnosed until the disease is a well-advanced stage. From different report estimations, it indicated that the CKD prevalence is around 8 – 16 % globally, which shows it is becoming one of the serious public health problems. Chronic kidney disease (CKD) is a worldwide common health issue, with a high lifetime risk of greater than 50 percent, higher than that for invasive cancer, diabetes, and

coronary heart diseases [16]. Around 8 and 10 percent of the adult population all over the world have some complications of kidney damage, and every year millions die prematurely of complications related to chronic kidney disease (CKD) [17]. Among the people of developing countries, chronic kidney disease is the main causes of death, because, in developing countries, there are no well-established prevention methods of chronic kidney disease, the people have no awareness of the issue and a shortage of specialists and health facilities are there [10]. In most developed countries, CKD is mostly related to old age, diabetes, hypertension, obesity, cardiovascular disease, and diabetic glomerulosclerosis, and hypertensive nephroscerosis, whereas, in developing countries, the common causes of CKD are glomerular and tubulointerstitial diseases which result from infections and exposure to drugs and toxins. Moreover, in most developing countries low-level quality of life like lack of clean water, lack of appropriate diet is the main cause of CKD.

Our daily diet, what to eat, and drink can damage our kidneys. Therefore, since getting a balanced diet is difficult in developing countries like Ethiopia people are simply caught by kidney disease and other risk factors of this disease. Chronic kidney disease progress to kidney failure silently if do not detected early and accurately by practitioners. At the early stage of CKD, the may not has any symptoms. The only way to detect and identify the presence or absence of the CKD disease urine analysis and blood test needed to know the kidney function and kidney damage. The treatment option is not affordable on time and people have less knowledge about CKD in developing countries. CKD will grow to end-stage kidney failure slowly that is a very serious problem for which artificial filtering (dialysis) or a kidney transplant is needed [18]. Chronic kidney disease has different symptoms after sometimes though it has no identifiable symptoms at the early stage (stage 1 and stage 2). Based on these symptoms and additional urine analysis and blood taste the stages are identified to give treatment for the patients. Most people have no symptoms until the CKD gets advanced and has no awareness about the symptoms of whether the symptoms are related to chronic kidney disease or another disease. Most people even think the symptom is of the disease like flu, common cold, and other infectious diseases. Chronic kidney disease is known for its common symptoms such as sleep troubles, vomiting, loss of appetite, tiredness and weakness, nausea, changes in how much urinate, pain in the chest builds from fluid around the lining of the center, swelling of feet, and ankles and uncontrolled hypertension, etc.

The condition like high blood pressure and irritation of the kidney's filtering units, interstitial nephritis, inflammation within the encompassing structures and kidney tubules, urinary organ stones, and a few cancers, polycystic kidney disease damage the kidneys. Risk factors for chronic renal failure can be categorized into four types[19]; the first category includes older age, gender, family history of CKD, reduction in kidney mass, low birth weight, racial or ethnic minority status, low income, and education, the second category includes diabetes, high blood pressure, autoimmune diseases, hereditary diseases, etc. The third factors category includes a high level of proteinuria, poor glycemic control in diabetes, smoking, etc. and the fourth categories are anemia, CAD, and late referral to a nephrologist. Different researchers have researched on identifying common risk factors of chronic kidney disease locally and globally and they have put their general conclusion as follows.

Roshni PR and Mahitha Mathew [19] conducted their research on the common risk factor in chronic kidney disease. They have stated that among the risk factor, aging was a significant predictor for renal failure in both males and females, but more prominent in males and stated that hypertension and diabetes mellitus, main causes among diseases for causing renal failure. Jiayu Duan, Chongjian Wang, et al.[20] conducted their researchers to measure the prevalence and risk factors for chronic kidney disease (CKD) and diabetic kidney disease (DKD) in a Chinese rural population. According to their finding, they come up with age, gender, education, personal income, alcohol consumption, overweight, obesity, diabetes, hypertension, and dyslipidemia as prevalent risk factors.

Kabaye Kumela Goro, Amare DesalegnWolide, et al. [21] conducted their study on Patient Awareness, Prevalence, and Risk Factors of Chronic Kidney Disease among Diabetes Mellitus and Hypertensive Patients at Jimma University Medical Center. The finding of the study showed that the prevalence of CKD is associated with hypertension, diabetes mellitus, blood pressure, and high fasting blood sugar, long duration of hypertension, and age. Meaza Zeleke [22] conducted her study on the Magnitude of Chronic Renal Failure and Its associated factors among patients at St. Paulo’s Hospital. She conducted her study on 320 individuals. The study finding showed that age, sex, marital status average monthly income, smoking status, alcohol drinking status, family history of CRF, BMI, HTN, DM, cardiac problem, kidney infection history, and knowledge on kidney function and disease are the risk factors for chronic renal failure. The study illustrated that hypertension is a risk factor that is highly related to chronic renal failure.

2.1.1 Worldwide Prevalence of Chronic Kidney Disease

Kidney disease is a worldwide public health challenge from which over 750 million people suffer worldwide [23]. The burden of kidney disease differs substantially throughout the world. Chronic kidney disease (CKD) is a worldwide serious health problem that needs a large economic cost to health systems[24].

Luxia Zhang, et al. [25] researchers have done their research on the chronic kidney disease prevalence in China: a cross-sectional survey. The study was conducted on 47204 patients. The researchers stated that the number of patients with chronic kidney disease in China is estimated to be about 119.5 million (112.9–125.0 million). The researchers have also concluded that 119.5 million adults aged 18 years or older have chronic kidney disease as defined in this study, and only 12.5% of them are aware of the condition.

Dr. Marni Armstrong, et al. [26] have conducted their research on Prevalence and Quality of Care in Chronic Kidney Disease. The key findings of this study stated that Albertans identified with Stage 1-4 CKD people affected by CKD close to 191,000 based on laboratory measurement. Nevertheless, geographical difference exists across the province. Additionally, the researchers come up with the finding that shows the prevalence of Stage 3 and 4 CKD has increased by 7.1% over two years.

2.1.2 Prevalence of Chronic Kidney Disease in Ethiopia

In 2014 Temesgen F, Mehidi K, and Tilahun Y conducted their study on the Prevalence of Chronic Kidney Disease and associated risk factors among diabetic patients in southern Ethiopia. The researchers have conducted their study on 214 participants. The result of the study indicated that 18.2 % and 23.8 % of the study participants were found to have CKD, according to the MDRD and Cockroft-Gault equations, respectively [27].

Mekiya Hussein, Geremew Muleta, Dinberu Seyoum, Demeke Kifle, Dechasa Bedada conducted their study on Survival Analysis of Patients with End-Stage Renal Disease the Case of Adama Hospital, Ethiopia. The study result indicates that among the total of 500 ESRD patients, 72.40% of them had been alive until the end of the study period, while 27.60% have died. Out of those male individuals in this study 169 (33.80%), 32 (18.93%) died, and 106 (32.02%) died individuals reported from 331 (66.20%) female individuals [28].

2.1.3 Chronic Kidney Stages

Chronic kidney disease usually starts slowly and silently and progresses to kidney failure over several years. Chronic kidney disease has five stages identified by the GFR rate, which enable the experts to assess how much-advanced kidney disease is. The stages of chronic kidney disease are identified by the level of kidney function measured by a test called the glomerular filtration rate (GFR). In each stage, the function of the kidney reduces from stage to stage and the kidneys do not work as well as the stage before [29].

The figure below shows, the conceptual model of CKD that was developed by NKF-KDOQI in 2002, revised, and adopted by the international consensus in 2005.

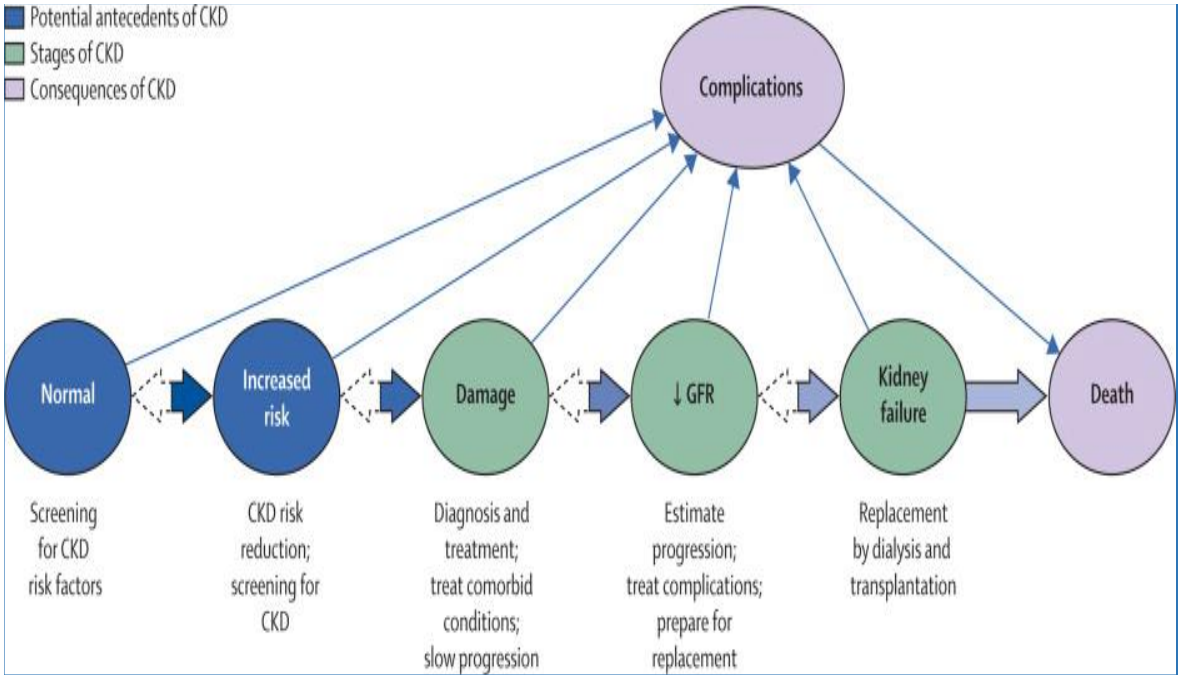

Figure 2. 1 Chronic Kidney Disease Complications Model [29]

There are different treatments such as reducing the progression of the disease, preventing the progression, treating the complications of decreased glomerular filtration rate, control of CKD risk factors, and treating cardiovascular disease For each stage of CKD [30].

Table 2. 1 CKD Stages according To National Kidney Foundation Based on GFR measurement value.

Stage	GFR (min/1.73m^2)	Description	Treatment
Stage 1	90+	Normal Kidney function (Mild stage)	Observation, control of blood pressure.
Stage 2	89-60	Digestible loss of kidney function (Mild stage)	Observation, control of blood pressure, and risk factors.
Stage 3a	59 – 45	Digestible to a slight loss of kidney function (Moderate)	Observation, control of blood pressure, and risk factors.
Stage 3b	44 -30	Slight to a severe loss of kidney function(Moderate)	Observation, control of blood pressure, and risk factors.
Stage 4	29-15	Severe loss of kidney function	Planning for dialysis
Stage 5	-15	Kidney failure or dialysis	Planning for transplant

Early detection of chronic kidney disease is used to identifying in which stage the kidney disease is and to take action. The treatment for chronic kidney disease is given to the patient according to the stage of the kidney disease identified based on the blood test and urine analysis.

2.1.4 Treatment and Management of Chronic Kidney Disease (CKD)

Early kidney disease has no indicators or symptoms related to kidney function. Hence it difficult to know that something is wrong with the kidney. Many people lose their kidney function before the symptoms appear. For this reason, it is very important and helpful to treat early and take action to prevent CKD by early diagnosis rather than going to the hospital after an advanced stage of kidney disease. For the CKD to be determined and diagnosed early, a different test such as blood test, urine test**,** hemoglobin, potassium, and sodium should be made by experts.

Management of chronic kidney disease aims to keep kidney function and homeostasis by treating any underlying condition, reduce the progression, and reduce the risk of developing CVD to kidney failure for as long as possible [31]. Once detected, the patient with CKD must follow the instruction, advice, and recommendation from doctors including choices about what to eat, drink. It is possible to prevent chronic kidney disease from getting worse stage by detecting it at an early stage and give treatment. There are several treatments and technologies to detect it, reduce the risk of complications of slow the progression of the disease. Treatment for CKD aims to reduce the risk of kidney disease and keep it from leading to the end-stage. Treatment includes anemia treatment, phosphate balance, control of high blood pressure, control of diabetes, etc. Furthermore, if the kidney reaches the final stage or failed dialysis or transplant treatment given to the patient.

2.2 Machine Learning

The invention of the computer and internet technology has led to several breakthroughs in the fields of science, medicine, law, and business [32]. Currently, most organizations worldwide are using the computerized system, which enables them to save their time and resources and makes them profitable. Machine learning is the subfield of artificial intelligence, which is currently, become an interesting area to work with a huge amount of data in different organizations. It is the most important and interesting methods in today's research to solve complex problems with huge data. Machine learning is used for the detection of target patterns from the data. It is concerned with training the machine what to do and make it do automatically based on the experience without explicit programming. According to different scholars, machine-learning approaches are categorized into supervised, unsupervised, and semi-supervised and reinforcement learning approaches.

2.2.1 Supervised Learning

Supervised learning is the type of machine learning algorithm concerned with developing a machine-learning model based on a known label. This algorithm consists of a target or dependent feature, which is predicted from a given set of predictor or independent features. The data used in the supervised learning algorithm is labeled or categorized data. Then learned rule is used to categorize unseen data with unknown outputs. Classification and regression are the two common tasks of supervised machine learning. Some examples of

supervised learning are Naïve Bayes, Support Vector Machine, Decision Tree, Random forest, K- Nearest Neighbor, Logistic Regression, etc.

2.2.2 Unsupervised Learning

Unsupervised learning is the algorithm used to work with unlabeled data. This algorithm is a very powerful tool for analyzing data and for identifying patterns and trends without the unknown label the instances belong to. It is the learning algorithm applicable to clustering a given dataset into different groups. K-means is the most common examples of unsupervised learning.

2.2.3 Semi-supervised Learning

It is a combination of both supervised and unsupervised machine learning methods. In this type of algorithm, the classified and unclassified data are mixed. This algorithm aims to learn a model that predicts targets of future test data better than that from the model generated by using the data with a known target alone.

2.2.4 Reinforcement Learning

This algorithm aims to use the data gathered from the interaction with the environment and the algorithm trains itself continually by using trial and error methods and feedback methods. This learning algorithm learns from previous experiences and tries to take the best possible knowledge to make accurate decisions. Markov decision process is one example of reinforcement learning. Here learning data gives feedback so that the system adjusts to dynamic conditions to achieve an aimed objective. The system evaluates its performance based on the feedback responses and replies according to the feedback responses. Since this study aimed to use classification techniques, which fall under the supervised learning category, other machine learning techniques were considered beyond the scope of this thesis.

2.3 Machine Learning and Health Sector

The health sector is a big sector for one country to take care of the life of the population. Good health care is very important for developing countries like Ethiopia, where predominantly rural and subsistence agriculture populations with poor access to safe water, housing, sanitation, food, and health service are living. In Ethiopia, there is no customer satisfaction from the existing healthcare system. One important issue for improving

healthcare is to give more focus on the prevention of disease. Not every disease is preventable, but in many cases, early diagnosis can bring better health outcomes and lower costs for treatment.

One way of preventive healthcare issues is an early prediction of the disease. Through this early prediction, makes the diagnosis of disease simple and enables the practitioners to treat the patients early. After screening the disease, healthcare providers or experts could then use this predictor to identify high-risk individuals and recommend tests or other interventions for the disease[33]. In healthcare, machine-learning can be applied in different tasks like identifying diseases and diagnosis, drug discovery and manufacturing, medical imaging diagnosis, personalized medicine, outbreak prediction, etc. Hence, to solve the diagnosis and treatment problems the machine learning techniques play a crucial role in the health sectors to identify correctly the presence or absence of the disease based on the patient history data. In this study, because this study aims to work on labeled data to predict CKD, a supervised machine-learning algorithm is used.

2.3.1 Machine Learning for Disease Prediction

Machine learning plays a crucial role in disease prediction based on patient history data. The machine learning technique showed success in the healthcare system in the prediction and diagnosis of numerous critical diseases. Machine learning in the healthcare system enables the practitioners to process huge and complex medical datasets and then analyze them into clinical insights [8]. Hence, machine learning implemented in the healthcare system has a great impact on fast disease prediction to give treatment and increase customer satisfaction. The most common machine learning approach used in chronic kidney detection is a supervised machine-learning algorithm. A supervised machine-learning algorithm is an algorithm that can be applied to the already labeled data and it consists of classification and regression tasks. Classification is the technique mostly used in disease prediction and diagnosis by different researchers. Most of the previous research work reviewed for this research have used the classifiers such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), K-Nearest Neighbors (K-NN), Logistic Regression (LR), Artificial-Neural Network (ANN). Most of the research studies for chronic kidney disease prediction uses more than two classifier algorithms for computational, comparison reason, and used to suggest the algorithm that has high performance and accuracy for their proposed prediction model [34] [35], [36], [16], [37], [38], [39], [40], [41], [42].

Random Forest: Random forest is ensemble learning, which is the supervised machine-learning algorithm that consists of several collections of decision trees and is used for classification, regression. This model consists of several decision trees and outputs the class target that is the highest voting results of the target output by each tree [43]. Random Forest uses both bagging and random feature selection to build the tree and creates an uncorrelated forest of trees whose prediction by the group is more accurate than that of any individual tree. After it builds the forest, test instances are permeated down through each tree and trees make their respective prediction of class [44].

Support Vector Machine: This algorithm is the highest prominent and convenient supervised machine-learning algorithm that can be used for data classification, learning, and prediction. Support Vector Machine builds a set of hyperplanes to classify all input in high dimensional data. In SVM data, classified into two data points in which hyperplane lies between two classes. SVM creates a discrete hyperplane in the signifier space of the training data and compounds are classified based on the side of the hyperplane located [45]. Hyperplanes are decision boundaries that help separate the data points and support vectors are data points that are closer to the hyperplane and determine the position and orientation of the hyperplane. SVMs were mainly proposed to deal with binary classification, but nowadays many researchers have tried to apply it to multiclass classification because there are a huge amount of data to be classified into more than two classes in today's world. Support Vector Machine (SVM) solves the multiclass problems through the two most popular approaches one-versus-rest (one-vs-all) and one-vs-one. In this study, one-versus-rest is used.

Decision Tree (DT): It is one of the most popular supervised machine-learning algorithms that can be used for both classification and regression problems. Decision Tree solves the problem of machine learning by transforming the data into a tree representation by sorting them by feature values. Each node in a decision tree denotes features in an instance to be classified, and each leaf node represents a class label the instances belong to. This model uses a tree structure to split the dataset based on the condition as a predictive model that maps observations about an item to make a decision on the target value of instances [46].

2.3.2 Deep Learning for Disease Prediction

In the healthcare system, deep learning is used to extract the hidden opportunities and patterns in clinical data helps doctors for diagnoses and treat patients fast [47]. In deep learning algorithms a huge amount of data, including patient-related records data, medical reports of patients collected, and on that, data neural network techniques are applied to provide better outcomes. The deep learning algorithm is used to solve complex problems, which are not solved by machine learning algorithms. Deep learning uses a different neural network technique and provides better results. In healthcare systems, deep learning provides the analysis of any diseases accurately to doctors and enables doctors to check for the presence or absence of disease, treat particular patients and give better medical decisions based on identified diseases and severity of the disease. Deep learning techniques can help physicians analyze information and detect multiple conditions such as, Analyze blood samples, track glucose levels in diabetic patients using image analysis to detect tumors, detecting cancerous cells and diagnosing cancer, detecting osteoarthritis from an MRI scan before the damage has begun, etc. However, deep learning is out of the scope of this study.

2.4 Feature Selection Methods

Feature selection is an important aspect of machine learning to select relevant features to develop the model. The goal of applying feature selection is to avoid noisy data, inconsistent data, and reduce the dimension of the dataset to get an efficient and accurate result. Besides complexity can be reduced by applying feature selection and extraction to the dataset.

Feature Extraction: is the method of extracting important or relevant features from the original features by applying a transformation that produces the projection of the original dataset to predict the outcome. Principal component analysis, linear discriminant analysis algorithm, and independent component analysis are popular feature extraction used in the clinical dataset.

Feature Selection: is a method of selecting the subset feature that contributes most to the prediction variable or output based on statistical analysis from the original dataset. Feature subset selection is a preprocessing technique that finds a minimum subset of features by removing as much of the irrelevant and redundant features from the original dataset as to enable adequate classification [48]. In the medical area, feature selection is used to select

and identify important and relevant features from patient history data that contribute most to the prediction of the disease in the patient. Feature selection applied in areas such as diabetes prediction heart disease prediction, hypertension prediction, and chronic kidney prediction of a medical dataset. The CFS and the chi-squared algorithms give a better performance for selecting the features on the medical dataset[49]. Feature selection approaches are classified into four categories, such as filter approach, wrapper approach, and embedded approach [50].

Filter Methods: It is the method of selecting and filtering features before implementing any machine-learning model and only after the best features are found, modeling algorithms can use them. Filter methods use statistical measures to score the dependencies between input features and target class so that most relevant features can be filtered. This method is efficient and fast to filter features and work with the dataset of a high number of features. The limitation of this method is that it does not identify the interaction and relationship between the dependent and independent features. The filter method can broadly be divided into Univariate Filter Methods and Multivariate filter methods categories. Univariate filter feature selection evaluates and usually rank a single feature based on its score, while multivariate filters evaluate an entire feature subset [51].

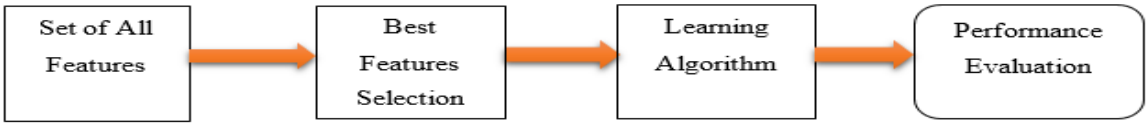

Figure 2.2 Filter Model

Wrapper Methods: The method considers the machine-lesarning model to randomly select a relevant feature and train the model with this feature subset to evaluate the accuracy of the model on the dataset. The wrapper method takes considers feature independencies, the interaction between each feature, and model selection [52].

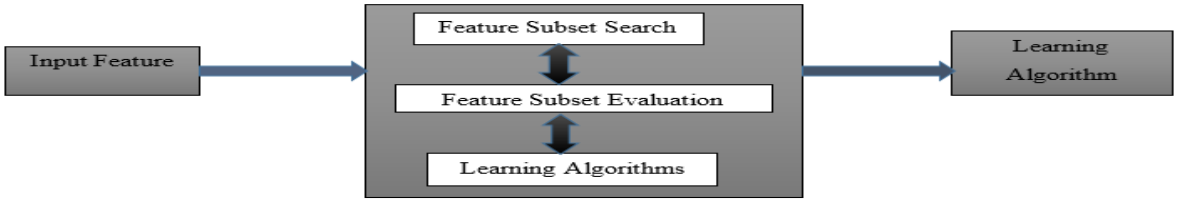

Figure 2.3 Wrapper Model [53]

Embedded Methods: This method is used to select the best features during the modeling algorithm execution by interacting with the learning models at a lower computational cost than the wrapper approach. This method considers feature dependencies and searches locally for features that allow better local discrimination and it uses the independent criteria to recognize the optimal feature subset for a known cardinality to select the final optimal subset among the optimal subsets [54].

Table 2. 2 Taxonomy of Feature Selection and Extraction Algorithms [55]

Model Search		Advantages	Disadvantages	Examples
Filter	Univariate	Fast, Scalable, Independent of the classifier	Ignores feature dependencies,	X2,Information gain, ANOVA
	Multivariate	Independent of the classifier, Better computational complexity	Slower, Less scalable, Ignores interaction with the classifier	(CFS), (MBF), (FCBF)
Wrapper	Deterministic	Simple, classifier dependent, Less computationally intensive	Risk of overfitting, Classifier dependent	(SFS), (SBE), Plus L Minus R, Beam search
	Randomized	Less prone to local optima, classifier dependent,	Computationally intensive, Classifier dependent selection, Higher risk of overfitting	Randomized hill-climbing, Genetic algorithms,
Embedded		Classifier dependent, Better computational complexity	Classifier dependent selection.	DT, Weighted NB, Feature importance

2.5 Related Works

This section discusses the review of necessarily related research papers to chronic kidney disease prediction to clearly understand general techniques, methods, and results of current studies. Many researchers have been researching the system that can predict the diseases and still the health sectors are left with huge researchable areas that need the intervention of machine learning suitable for the classification and prediction of disease such as machine learning, deep learning. Different related works have conducted locally and globally by different researchers on chronic kidney disease prediction using machine-learning techniques have reviewed.

A. Charleonnan *et al.*[38] presented the predictive models such as K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree (DT) classifiers to predict chronic kidney disease. Finally, they have concluded that SVM has the highest accuracy of the other classifiers.

A. Salekin and J. Stankovic [34] proposed three classifiers: K-NN, RF and ANN to predict CKD . The researchers have used a dataset with 400 patients form UCI with 24 attributes. They implemented wrapper feature selection to find the attributes that detect this disease with high accuracy. Applying the wrapper feature selection five attributes are selected and used in model construction. Finally they concluded that they can predict CKD with 98% F1 and 0.11 RMSE using RF.

F. Aqlan, R. Markle, and A. Shamsan[56] Proposed predictive models such as Decision Trees, Logistic Regression, Naive Bayes, and Artificial Neural Networks for CKD prediction purpose. The researchers conducted data preprocessing to increase the accuracy of the model and reduce overfitting. The researchers used a dataset with 400 patients form UCI with 24 attributes. Researchers have used and compared six predictive models to predict the presence of CKD. The researchers come up with Random Trees as the best method for the CKD prediction in their study.

S. Tekale *et al.*[39] worked on "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm" with a dataset consists of 400 instances and 14 features. They have used the Decision tree and support vector machine models to implement the prediction system. The researchers preprocessed the dataset and reduced the number of features from

25 to 14. Finally, they come up with SVM as the better model than DT with an accuracy of 96.75%.

The research conducted by S. Zeynu and S. Patil [40] proposed K-NN, J48, ANN, NB, and SVM classifiers to predict Chronic Kidney Disease. In the research Information gain attributes, an evaluator with ranker search engine and wrapper subset evaluator with the best first engine was used as the feature selection method. The result of their research showed ANN has highest accuracy by using Wrapper feature selection method with Best first search engine feature selection method. J48 and SVM classifiers have an accuracy of 98.75% and 98.25% with Info Gain Attribute Evaluator with a ranker search engine. Building the second method ensemble model by combing the five different classifiers based on a voting algorithm. They showed that the proposed ensemble model achieved 99% accuracy.

J. Xiao *et al.*[57] proposed the prediction of chronic kidney disease progression using logistic regression, Elastic Net, lasso regression, ridge regression, support vector machine, random forest, XGBoost, neural network and k-nearest neighbor and compared the models based on their performance. They have used 551 patients' history data with proteinuria with 18 features and classified the outcome as mild and moderate/severe. The researchers have concluded that Logistic regression ranked first, with AUC of 0.873, with a sensitivity and specifcity of 0.83 and 0.82, respectively.

S. Mohammed and T. Beshah [10] conducted their research to develop a self-learning knowledge-based system for diagnosis and treatment of the first three stages of chronic kidney using machine learning. The researchers have used a small number of data and they have developed the prototype, which enables the patient to query KBS to see the delivery of advice. They used decision tree modeling approach in order to model the rules. The researchers conclude that the overall performance of the prototype system registered a 91% accurate result. R. Kadam Vinay, K. L. S. Soujanya, and P. Singh [47]conducted their study on Disease Prediction by Using Deep Learning Based on Patient Treatment History. The study aimed to predict three human diseases such as Heart disease, Diabetes, Chronic Kidney Disease based on patient history data collected from a different source. 1471 no of records of the dataset used for the experimentation purpose. Two models were built and compared based on their accuracy level. They have compared the result of the ANN model with the traditional SVM method and the researchers concluded that ANN is more accurate than

SVM. They generalized that the ANN model gives 95%, 98%, 72% accuracy for predicting the heart, kidney, and diabetes diseases respectively.

A. Subas, E. Alickovic, and J. Kevric[58]conducted their study on the Diagnosis of Chronic Kidney Disease using Random Forest Algorithms. The researchers have used a dataset with 400 patients form UCI with 24 attributes. The researchers applied feature selection like CSP filter and LDA analysis and 10 features are selected out of 24 features.

A. B. K. A, K. Priyanka *et al.* [41] developed a chronic kidney disease prediction system based on the naive Bayes technique. They have tested using other algorithms such as KNN (K-Nearest Neighbor Algorithm), SVM (Support Vector Machines), Decision tree, and ANN (Artificial Neural Network) and they have got Naïve Bayes has higher accuracy results when compared to other algorithms. The Accuracy obtained is about 94.6%.

M. Almasoud and T. E. Ward [59] aimed in their work to test the ability of machine learning algorithms for the prediction of chronic kidney disease using the smallest subset of features. They used the feature selection technique such as Pearson correlation, ANOVA, and Cramer's V test to select predictive features. They have conducted their work using LR, SVM, RF, and GB machine learning algorithms. Finally, they concluded that Gradient Boosting has the highest accuracy with an F-measure of 99.1.

The research conducted by S. Y. Yashfi [42] proposed a system from which it will be possible to predict the risk of CKD using machine learning algorithms by analyzing the data of CKD patients. The researcher proposed random forest (RF) and artificial neural network. They have used a real-world dataset and conducted a feature selection method to extract relevant features. Finally, they have extracted 20 features out of 25 features and applied RF and ANN algorithms. They concluded that RF has the highest accuracy of 97.12% than ANN.

E. A. Rady and A. S. Anwar [60] have used Probabilistic Neural Networks (PNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Radial Basis Function (RBF) algorithms to predict kidney disease stages. The researchers conducted their research on a small size dataset and few numbers of features. The result of this paper shows that the highest accuracy recorded with Probabilistic Neural Network.

Table 2. 3 Some Summary of Related Works

No.	Author	Title	Methodology and Accuracy	Draw back
1	A. Salekin and J.Stankovic (2016) [34]	Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes	K-NN, RF, and NN. Detection F1-score of RF 99.8	Small dataset size was used; Severity level prediction was not included.
2	S. Tekale *et.al.* [39] (2018)	Prediction of Chronic Kidney Disease Using Machine Learning Algorithm	DT and SVM with an accuracy of 91.75 and 96.75 respectively.	Small dataset size, Severity level prediction was not included.
3	A. B. K. A, K. Priyanka *et.al.* [41] (2019)	Chronic Kidney Disease Prediction Based On Naive Bayes Technique	NB, KNN, SVM, DT, and ANN. NB accuracy is 94.6%.	Small size dataset. No severity prediction.
4	S. Y. Yashfi [42] (2020)	Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms	RF and ANN with an accuracy of 97.12% and 94.5%.	Small size dataset, they recommend feature selections, stages prediction.
		Multiclass		
5	E. A. Rady and A. S. Anwar [60]	Prediction of kidney disease stages using data mining algorithms	PNN, MLP, SVM, RBF accuracy of 96.7, 60.7, 87, and 51.5 respectively.	Small data size. They used the algorithms, which is not appropriate for small dataset size.

The above reviews indicates that several studies have been conducted on chronic kidney disease prediction using machine-learning techniques. However, the previously conducted researches were not concerned with chronic kidney disease severity prediction and also the dataset used for previously conducted researches is small size, noisy and not recently collected data. Besides, chronic kidney disease prediction using machine learning were not conducted in Ethiopia. Therefore, in this study the machine-learning model that can predict the severity level of chronic kidney disease was built using dataset with large instances than real-world dataset.

CHAPTER THREE

METHODOLOGY

This chapter discusses the methodology and the process of developing chronic kidney disease prediction using machine-learning techniques. Nowadays, the application of machine learning is quickly increasing in various sectors. Health sectors are among those sectors, which are using machine-learning technology extensively to satisfy the customer's need. Because the health sectors have a huge amount of data, which is complex for the human for analysis, but machine learning makes it easy to make a pattern and analysis of the data. Different methods, procedures, and techniques have to be followed and used in the research build machine-learning models. The justification of the methods and procedures have been made in different phases. First, the data collection method and techniques are explained. The second phase explains the preprocessing method, which is a very important step in machine learning model building and feature selection methods, which is also a crucial part of building a machine-learning model. The third phase explains and justifies the supervised machine learning algorithms used in developing a chronic kidney disease prediction. Finally, the performance evaluation method used in this research study is explained.

3.1 Data Source

The data to conduct this research was collected from the patient history data from St. Paulo's Hospital. St. Paulo's Hospital is the second-largest public hospital found in Addis Ababa, the capital of Ethiopia. Additionally, by assessing different information about the hospitals St. Paul's Hospital is selected because there could a high number of patients with chronic diseases and dialysis treatment is given at the hospital. The hospital launched the kidney transplant center in 2015. It opens up for the service and has a good opportunity for researchers and policymakers for program monitoring and evaluation.

3.2 Data Collection

The data for this study was secondary and collected from the patient history data without direct communication with the patient. The data collected and used in this study, included patients' records of chronic kidney disease, who were admitted to the renal ward that attends their treatment or died during a period 2018 to 2019 an and some of them were obtained from

the same patient history data at different times. To collect data different features to predict the presence of chronic kidney disease are identified by asking the domain experts, assessing different local and global references on chronic kidney disease, and considering online available features of the chronic kidney disease dataset. The data were collected by reviewing the patient history of the renal cases by three nurses. Three of them work in the renal ward of St. Paul's Hospital and training was given for three of them. The collected data is depending on 18 features and 1 class, from which 6 is nominal, and 11 is numeric. Because the stages did not identified on some patient history data the stage was calculated using the CKD-EPI equation. The CKD-EPI equation is given in Appendix A.1.

3.3 Chronic Kidney Disease Prediction Modeling

3.3.1 Preprocessing

Preprocessing is the main part of building a prediction model in our study. Pre-processing is a means of converting the noisy and huge data into relevant and clean data. As the collected data contains missing values and categorical data, it has to handle missing values and categorical data to prepare the dataset in a suitable format for the models to use it. Preprocessing is a very essential part of developing a machine-learning model since the inconsistent data affect the accuracy of the model. After the collection and preparation of the dataset, performing preprocessing which is cleaning the data before they are fed into the classifiers has a vital role in accurately and correctly predicting the disease. Preprocessing is an important part to complete the prediction model. Following dataset preparation, pre-processing steps are conducted to feed the clean dataset to the machine-learning model:

Cleaning Noisy Data: Cleaning outlier, noisy data is an important part of preprocessing to build a prediction model with the best result. Outliers are the values in the dataset, which lies away from the range of the rest of the values. Therefore handling outliers is very important in building a better predictive model. In clinical data, outliers may arise from the natural variance of data. Researchers mostly use boxplots as an easy way to detect outliers. Boxplot uses median and interquartile range IQR to identify the outliers, where the median is the middle number of an ordered set of values and the interquartile range is the variance between the first and third quartiles[61]. To be more precise, outliers are the data points that fall above $Q3+1.5(\text{IQR})$ and below $Q1-1.5\ (\text{IQR})$, where Q1 is the first quartile, Q3 is the third quartile,

and IQR=Q3−Q1[61]. The outliers in this study are handled by replacing the value using the mean value.

Handling Missing Values: Some results had missing values due to some mistakes while collecting data and some unfilled values of patient's history data. Patient data often have missing diagnostic test results that would help to predict the likelihood of diagnoses or predict treatment effectiveness[62]. The missing values have an impact on the accuracy of the prediction model. There are several ways of handling missing values namely dropping missing values, filling missing values. Sometimes missing values are ignored if they are least in number i.e. if missing data under 10% for an individual case, but it is not considered healthy for the model because the missing value can be an important feature contributing to the model development. Sometimes the missing values can be also replaced by zero, which will bring no change to the model. To handle these missing values the mean, an average of the observed features used in this study, because the missing features are numeric and mean imputation is better for numerical missing values.

Handling Categorical Data: In this step, real data is transformed into the required format. The collected data consists of real data, decimal values, and nominal data. Therefore, in this step nominal data converted into numerical data of the form 0 and 1. For instance, 'Gender' has the nominal value that can be labeled as 0-for female and 1-for male. After preprocessing the data then the resultant CSV file comprises all the integer and float values for different CKD related features.

Feature Scaling: It is always important to scale numerical descriptive features before fitting any models, as scaling is mandatory for some important classes of models such as nearest neighbors, SVMs, and deep learning[63]. There are different techniques of scaling data, but in this study Standard Scaling that scales the descriptive feature to 0 mean and 1 standard deviation was used. Standard scaling is done as follows:

$$Z = \frac{X - \mu}{\sigma} \qquad \text{Equation 3. 1}$$

Where:

Z = Standard Scale

X = Score or value

μ = Mean and

$\boldsymbol{\sigma}$ = Standard Deviation

3.3.1.1 Feature Selection Methods

To generate the best results from the data, it is crucial to first identify the most relevant predictive features[64]. Feature selection is the process of selecting the most effective features among original dataset features and using the selected features as input data for the models. Feature selection is an important preprocessing method to solve the dimensionality problem. The main aim of the feature selection technique is to select the subset of features that are relevant and independent of each other for training the model [48].To develop a chronic kidney disease predictive model feature selection plays a crucial role in selecting the best and relevant features from original features and reduce the dimension of the dataset. This reduces the dimensionality and complexity of the data and makes the models operate faster, more effectively, and accurately. On the other hand, a feature selection algorithm is used to select relevant features after the construction of the dataset.

Feature selection can be either filter, wrappers, or embedded which are used for feature selection in different clinical datasets including for chronic kidney disease dataset. Filter and wrapper methods were used in this study. A filter method is a feature selection technique that is independent of a classification algorithm and uses general characteristics of data for evaluating and selecting relevant features. The filter method works independently of the learning algorithm to remove irrelevant features and analyses the properties of a dataset to chooses the relevant features[65]. The filter method is popular; it is the most widely used approach due to its less complex nature. The wrapper method selects relevant feature selection while using the classification algorithm. The wrapper method is the best feature selection in terms of accuracy than filter feature selection. However, it requires high processing time. Univariate feature selection method used from filter methods because it is fast, efficient, scalable, and recursive feature elimination with cross-validation (RFECV) is used from wrapper feature selection.

Univariate Feature Selection (UFS): This method is the most popular, simplistic, and fastest feature selection method used in medical data analysis. The univariate method is the feature selection technique that considers each feature separately to determine the strength of the relationship of the feature with the dependent variable. It is Fast, Scalable, Independent

of the classifier. Different options are there for univariate algorithms such as Pearson correlation, information gain, chi-square, ANOVA (Analysis of Variance). In this study, feature selection was done using the ANOVA method. It also discussed in next chapter section 4.3.3.1.

The Formula for ANOVA is:

$$F = \frac{MST}{MSE} \qquad \text{Equation 3. 2}$$

Where:

F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

Recursive Feature Elimination with Cross-Validation (RFECV): An optimization algorithm to develop a trained machine-learning model with relevant and selected features by repeatedly eliminating irrelevant features. It repetitively creates the model, keeps aside the best or worst performing feature at each iteration, and builds the next model with the left feature until the features are completed to select best feature selection[66]. It eliminates the redundant and weak feature whose deletion least affects the training error and keeps the independent and strong feature to improve the generalization performance of the model[67]. This method uses the iterative procedure for feature ranking and to find out the features that have been evaluated as most important. Because this technique work interacting with a machine learning model it first builds the model on the entire set of features and ranked the feature according to its importance. The detail procedure of RFECV discussed in chapter for section4.3.3.2.

Generally, the feature selection application is said to be successful if the dimension of the data is reduced and the accuracy of a model improves or remains the same as the model on original data.

3.3.2 Machine Learning Models

This study aims to develop automatic chronic kidney disease prediction using machine-learning classification algorithms based on already categorized or labeled data. Three

supervised machine-learning classification algorithms were used in this study for the comparison of the predictive models based on their performance. The algorithms have selected based on their popularity in chronic kidney disease prediction and their performance of classification on previously done research works[36], [37], [41], [43], [45], [56], [68], [47] [69], [40], [40], [35]. In this study, three machine-learning classification algorithms such as Random Forest, Support Vector Machine and Decision Tree were selected to develop the predictive model.

3.4 Evaluation

3.4.1 Prediction Model Evaluation

Performance evaluation is the critical step of developing an accurate machine-learning model. The prediction model must be evaluated by the model evaluation techniques to ensure that the model fits the dataset and work well on new unseen input data. The aim of the model performance evaluation is to estimate the generalization accuracy of a model on unseen/out-of-sample data. The performance evaluation method is divided into two categories; holdout and Cross-validation. Holdout evaluation aims to test a model on different data that it was trained on which provides an unbiased estimate of learning performance. In holdout, mothed data is randomly divided into a training set, validation set, and test set.

Hold-out is simple, flexible, and has high speed than cross-validation. However, this method has the problem of high variability since differences in the training and test dataset can cause differences in the performance. In the train-test-split method, the training might contain some repeated records and the model might miss out on parts of the data. Besides, there is lesser data available to train and test the model in the train test split method.

Cross-Validation is one of the performance evaluation methods for evaluating and comparing models by dividing data into partitions: one used to train a model and the rest used to test or validate the model. To avoid overfitting, the cross-validation technique is the most used evaluation technique. Therefore, we used the most popular K-fold cross-validation evaluation method to make sure that the models got the pattern for the training dataset for this study. Cross-validation is a technique that involves dividing the original dataset into a training set, used to train the model, and the test set used to evaluate the model. K-fold cross-validation is the basic form of the cross-validation technique. In k-fold

validation, the original dataset was partitioned into k equal size subsamples called folds. Therefore, we used the most popular 10-fold cross-validation technique in our study because k=10 result in a model with low bias and moderate variance.

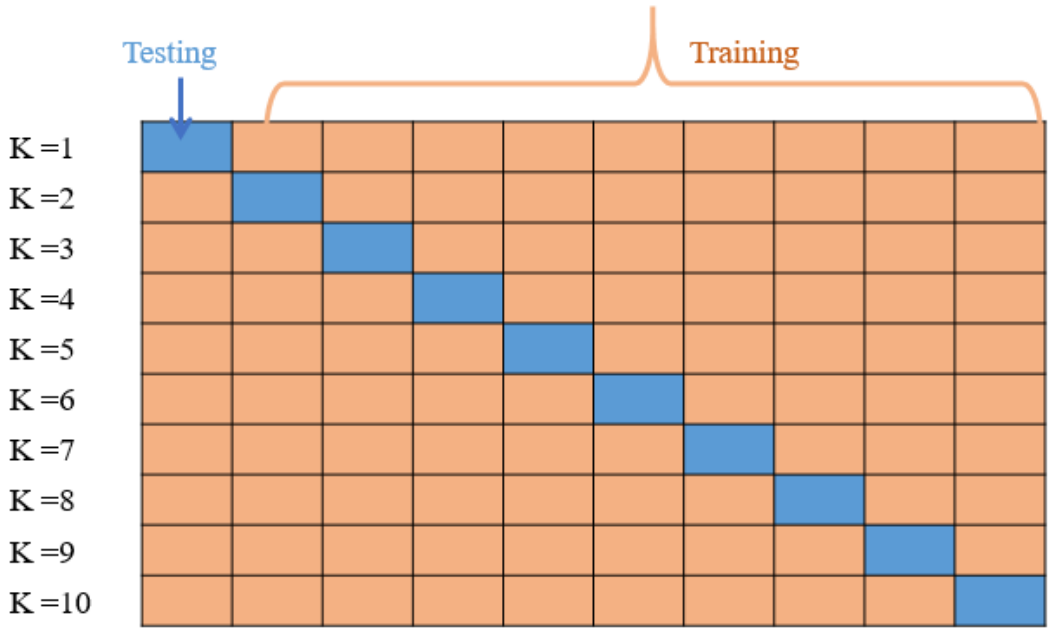

Figure 3. 1 Algorithm for the 10-Fold Cross-Validation[70]

3.4.2 Prediction Model Performance Evaluation Metrics

Prediction model evaluation metrics are required to measure model performance because classifier algorithms do not give the same result. Different performance evaluation metrics such as precision, recall, confusion matrix, and f1-score, sensitivity, specificity are used in this study with cross-validation.

The followings are the fundamental terms in performance measure:

- True positive (TP): are the condition when both actual value and predicted value are true.
- True negative (TN): are the condition when both the actual value of the data point and the predicted are False.
- False positive (FP): These are the cases when the actual value of the data point was False and the predicted is true.
- False negative (FN): are the cases when the actual value of the data point was true and the predicted is False.

3.4.2.1 Accuracy

Accuracy implies the ability of the classification algorithm to predict the classes of the dataset correctly. It is the measure of how close or near the predicted value is to the actual or theoretical value [71]. Generally, accuracy is the measure of the ratio of correct predictions over the total number of instances. Accuracy calculated using the equation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{Equation 3. 3}$$

3.4.2.2 Precision

Precision measure the true values correctly predicted from the total predicted values in the actual class. Precision quantifies the ability of the classifiers to not label a **negative** example as **positive.** The equation to calculate precision is:

$$Precision = \frac{TP}{TP + FP} \quad \text{Equation 3. 4}$$

In this study, the macro average is used for multiclass classification because it gives equal weight for each class. The macro average precision is calculated by:

$$Precision\_macro = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TPi}{TPi + FPi} \quad \text{Equation 3. 5}$$

3.4.2.3 Recall

Recall measure the rate of positive values that are correctly classified. Recall answers what proportion of **actual Positives** is correctly classified.

Recall calculated using the equation:

$$Recall = \frac{TP}{TP + FN} \quad \text{Equation 3. 6}$$

Since the macro average is used in order to compute the recall value of the models, macro average recall is calculated by:

$$Recall\_macro = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TPi}{TPi + FNi}$$ Equation 3. 7

Where: C = class

3.4.2.4 F-measure

F-measure is also called F1-score is the harmonic mean between recall and precision. The score is calculated by the equation:

$$F1_{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$ Equation 3. 8

The macro average of F1_score is calculated as:

$$F1_{score\_macro} = 2 * \frac{\text{Precision\_macro} * \text{Recall\_macro}}{\text{Precision\_macro} + \text{Recall\_macro}}$$ Equation 3. 9

3.4.2.5 Sensitivity

Sensitivity is also called True Positive Rate. Sensitivity is the mean proportion of actual true positives that are correctly identified[72].

$$Sensitivity = \frac{TP}{TP + FN}$$ Equation 3. 10

3.4.2.6 Specificity

Specificity is also called True Negative Rate. It is used to measure the fraction of negative values that are correctly classified.

$$Specificity = \frac{TN}{TN + FP}$$ Equation 3. 11

3.4.2.7 Confusion Matrix

The confusion matrix is the technique used to describe the performance of the classification model. The confusion matrix contains information about the actual value and predicted value used to evaluate the performance of the model based on the data in the matrix. The confusion matrix represents the number of correct and incorrect predictions by the model compared

with the actual classifications in the test data[73]. This study uses the confusion matrix with five class model (notckd, mild, moderate, severe, ESRD). The following tables show the binary class and five-class confusion matrix.

Table 3. 1 Confusion Matrix with Binary-Classes

		Predicted Values	
		Ckd	Notckd
Actual value	Ckd	True ckd	False notckd
	Notckd	False ckd	True notckd

Table 3. 2 Confusion Matrix with Five-Classes

		Predicted values				
		Notckd	Mild	Moderate	Severe	ESRD
Actual value	Notckd	True notckd	False mild	False moderate	False severe	False ESRD
	Mild	False notckd	True mild	False moderate	False severe	False ESRD
	Moderate	False notckd	False mild	True moderate	False severe	False ESRD
	Severe	False notckd	False mild	False moderate	True severe	False ESRD
	ESRD	False notckd	False mild	False moderate	False severe	True ESRD

CHAPTER FOUR

PROPOSED SOLUTION FOR CHRONIC KIDNEY DISEASE PREDICTION

This chapter discusses the proposed solution for Chronic Kidney Disease prediction using machine-learning techniques appropriate to solve the identified problem by preparing a chronic kidney disease dataset with its severity for multiclass for further understanding and usage of the model. Then the machine-learning model is deployed using a flask server. The prototype was developed, to help the practitioners or professionals to diagnosis their patients fast and help the patients by giving appropriate advice and treatment based on the status of the patients.

The data for this study was collected from St. Paul's Hospital, and then we have prepared the dataset for prediction based on the methodology discussed in chapter three. This chapter will discuss the overall proposed chronic kidney disease prediction using machine-learning techniques.

4.1 Proposed Chronic Kidney Disease Prediction Architecture

Following the identification and validation of the relevant features for the prediction of chronic kidney disease and the collection of patient history data, the predictive model was built using the machine learning algorithms, namely; Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT). The proposed multiclass architecture is used to predict chronic kidney disease as notckd, mild, moderate, severe, ESRD, and the binary for the prediction of ckd and notckd. The proposed method takes the dataset as input and then the dataset is preprocessed to make the proposed method complete. Dataset preprocessing process includes data cleaning, handling missing values, handling categorical data. After preprocessing the dataset feature selection such as univariate feature selection (ANOVA), recursive feature elimination with cross-validation (RFECV) to select relevant features for the prediction of chronic kidney disease are applied. Before and after the application of the feature selection methods, the model developed using RF, SVM, and DT machine learning algorithms. The proposed model was evaluated using K-fold cross-validation to avoid overfitting because it is appropriate for a small dataset. Other evaluation metrics such as

precision, recall, f-measure, sensitivity, and specificity are used with K-fold CV. The evaluation result is important to select the best model to predict chronic kidney disease. So that, the best model was selected based on the above-mentioned metrics level for the prediction of the disease. Finally, the prototype that can take new input data to predict chronic kidney disease is developed using the best-selected predictive model and flask server for deployment.

For model development, the most popular and most widely used programming language, which is open source, and free data science libraries are used. Additionally Flask, which is an open-source python package distribution used for web development and Anaconda navigator, consists of different IDEs such as Jupyter, Spyder, PyCharm IDE, and Notebook. The notebook is used for building our model and testing purpose and PyCharm IDE is the editor used to work with the web application part.

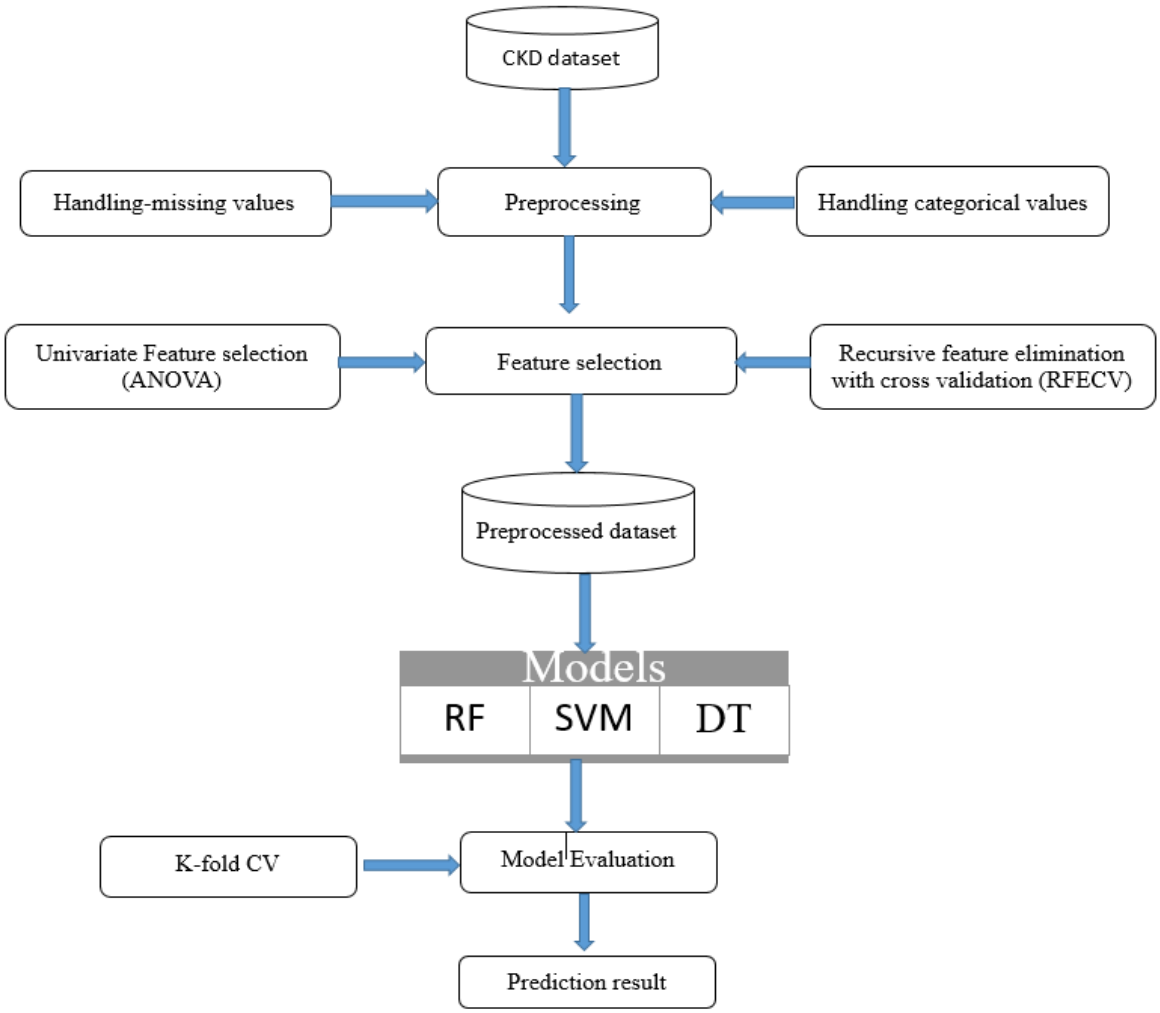

Figure 4. 1 Chronic Kidney Disease Prediction Architecture

4.2 Dataset Preparation and Features Description

This study aims to develop a chronic kidney disease prediction using machine-learning techniques. Hence, it needs to build a new chronic kidney disease dataset to build the model. This new dataset is needed because currently there is no published or annotated dataset for chronic kidney disease prediction in Ethiopia from a local hospital. The process of building the dataset for chronic kidney disease consists of the main steps of gathering data from patient history and prepare the dataset in a suitable way for model building. Preparing the data in a way that is appropriate for the specified machine learning tools, techniques, and tasks is one of the essential machine learning research part. Hence, after the data was collected, the researcher prepared the dataset in a way that the dataset is appropriate to the requirements of the selected machine learning tasks and the specific machine-learning tools.

Due to the difficulty of getting a prepared dataset in Ethiopia, data were collected from manually recorded patient history data. Because the data was collected in hardcopy form, it should be converted to softcopy to prepare a machine-readable dataset. As python, programming is used for the development of the model CSV (comma-separated value) file is needed for the code. For the data to be read and design the classification model data should be available in the appropriate format. The collected data contains socio-demographic (Age, Gender) features, electrolyte (sodium, potassium, and chloride) features, laboratory test and urine analysis (serum creatinine, hemoglobin, blood urine nitrogen, mean cell volume, platelet count, red blood cell count, white blood cell count, blood pressure) features, risk factors (hypertension, diabetic mellitus, anemia, heart disease,) features, and dependent class (ckd_status). The prepared dataset had missing values and categorical data that needs to be handled for building predictive models.

Table 4. 1 Dataset Features Description

Symbols	Features full name	Type	Class	Missing values in %
Age	Age	Numeric	Predictor	0
Gender	Gender	Nominal	Predictor	0
Bp	Blood pressure	Numeric	Predictor	0.058207
Sg	Specific gravity	Nominal	Predictor	0.058207
Chl	Chloride	Numeric	Predictor	0.116414
Sod	Sodium	Numeric	Predictor	0.232829
Pot	Potassium	Numeric	Predictor	0.116414
Bun	Blood Urea Nitrogen	Numeric	Predictor	0.232829
Scr	Serum Creatinine	Numeric	Predictor	0.058207
Hgb	Hemoglobin	Numeric	Predictor	0.232829
Rbcc	Red blood cell count	Numeric	Predictor	0.232829
Wbcc	White blood cell count	Numeric	Predictor	0.232829
Mcv	Mean cell volume	Numeric	Predictor	6.111758
Pltc	Platelet count	Numeric	Predictor	7.275902
Htn	Hypertension	Nominal	Predictor	0
Dm	Diabetes Mellitus	Nominal	Predictor	0
Ane	Anemia	Nominal	Predictor	0
Hd	Heart disease	Nominal	Predictor	0
ckd_status	ckd_status	Nominal	Target	0

4.3 Preprocessing

Dataset preprocessing task carried out to prepare proper dataset to be fed into machine learning models to have accurate and correct results. This preprocessing method is performed based on the nature of the dataset. Technique such as cleaning noisy data, filling missing values, data transformation are applied to the dataset in this research as shown in Figure below.

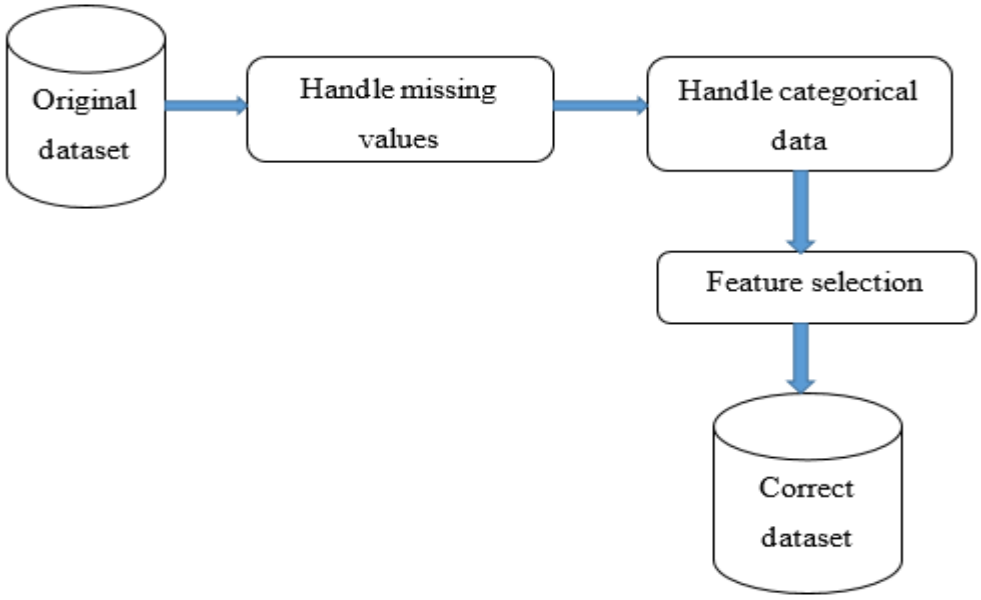

Figure 4.2 Chronic Kidney Disease Dataset Preprocessing Steps

4.3.1 Handling Missing Values

While collecting data some results have missing values due some error of encoding data. The dataset is created from the data gathered from patient history data, which contains some missing values and there were some missing values while encoding gathered data due to encoder mistake. The missing values in this study were handled using mean strategy. The detail of handling missing value was discussed in chapter three section 3.3.1.

4.3.2 Handling Categorical Data

The created dataset has some categorical values that need to be transformed to the form 1 and 0. Thus, this transformation task transforms all categorical values. Dummy variable encoding and replace method were used to handle categorical data in this study. It also discussed in chapter three section 3.3.1.

4.3.3 Feature Selection

The proposed feature selection method performs the selection of important features from the original dataset to improve the performance of predictive models. This action takes place after handling missing values and categorical data preprocessing step. Having a clean dataset the feature selection was applied to select significant features. The selected features used for training the model and predict the patient as notckd, mild, moderate, severe, and end-stage renal disease.

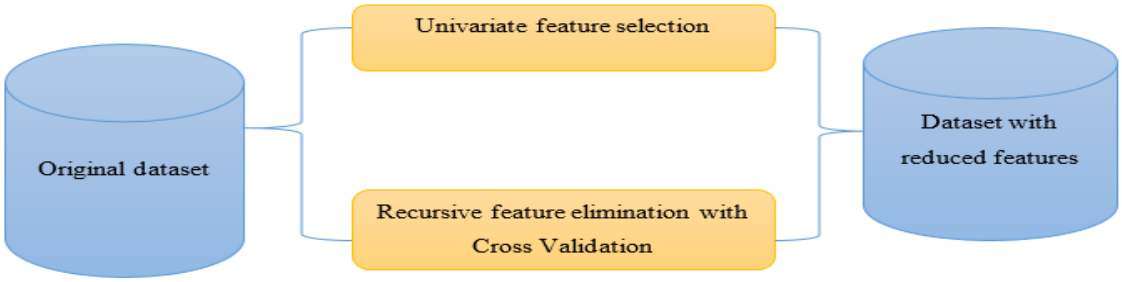

Figure 4. 3 Feature Selection Flow Diagram

4.3.3.1 Univariate Feature Selection (UFS)

In this research study, we proposed univariate feature selection, which is the approach used to examine each feature individually to determine the strength of the relationship of the feature with the target class using ANOVA which can calculate the F-measure to identify the strong feature. Based on this calculated F-measure value between each feature and the dependent variable the desired number of features with the best F-measure scores is selected. Since the selection of the feature is independent of machine learning algorithms in the univariate feature selection, we applied the machine learning classification algorithms after selecting the best features using ANOVA.

4.3.3.2 Recursive Feature Elimination with Cross-Validation

Recursive feature elimination is the feature selection method that works by searching for a subset of features by starting with all features in the training dataset and successfully eliminate irrelevant features until the relevant features remains. RFECV selects the best subset of features for the supplied estimator by removing irrelevant features using recursive feature elimination, then, select the best features based on the cross-validation score of the models. This approach is dependent on the machine-learning algorithm. Assume T is a

sequence number to store the feature ranking, at each iterative process of feature elimination, the Ti stores the top-ranked features on which the model refit and performance is measured[67].

Table 4.2 Steps of Recursive Feature Elimination with Cross-Validation[67]

Steps	Procedure
1	Train the ML model using all features with 10-fold cross-validation
2	Compute the model performance
3	Calculate the Feature importance or ranking
4	For each subset Ti, i=0,1,2,3,……., n do
5	Keep the Ti most important features
6	Train/Test model on Ti features
7	Recalculate model performance
8	Recalculate the importance of ranking each feature
9	End
10	Calculate the performance over Ti
11	Determine the optimal number of features
12	Use the model with the selected optimal features

Table 4.2 shows the procedure of the recursive feature elimination used to eliminate irrelevant features and select the strong and relevant features that finally fit with the machine-learning model. Because the technique is automatic feature selection and depends on the applied model, the number of selected features may differ from algorithm to algorithm.

4.4 Machine Learning Models Building

This study aims to develop a model that enables to predict whether the patient has chronic kidney disease or not and its risk level. This study builds a classification model mainly by using machine-learning models, RF, SVM, and DT. For creating a predictive model, a total instant of 1718 records was collected for training and testing using learning algorithms. In this study, the classification process outcome is predicted from a given input or futures relation to class.

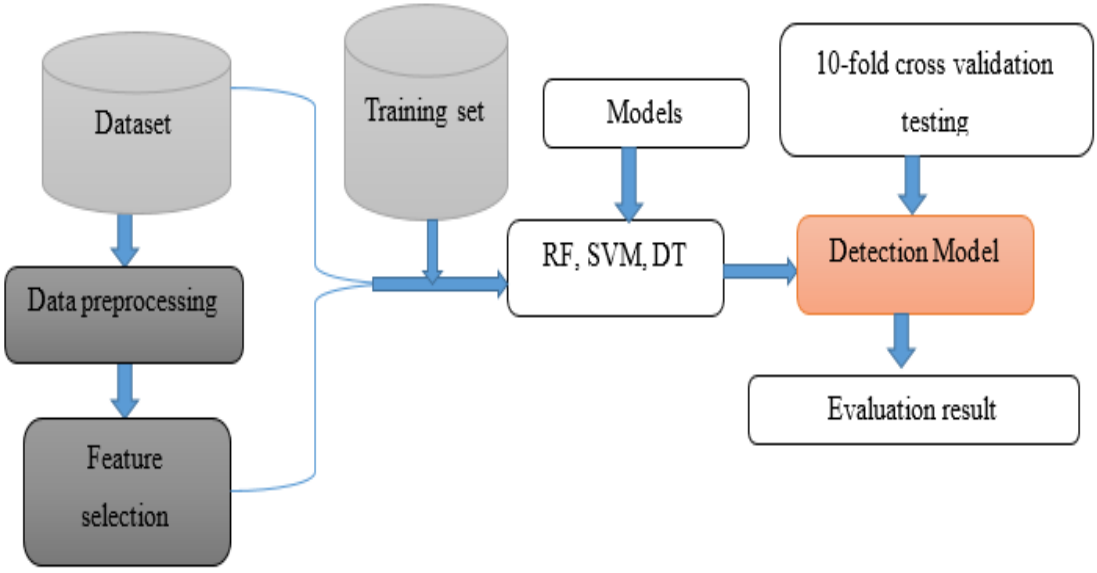

Figure 4. 4 Model Building Flow Diagram

Random Forest (RF) classifier that fits several decision tree classifiers on a subsample of the dataset was proposed in this study. It means RF consists of a large number of individual decision trees that operate as an ensemble by bagging and feature randomness. This study also proposed the One-vs-Rest (OVR) strategy of Support Vector Machine (SVM) machine learning classifier because a simple form of SVM algorithm is a binary classifier. To classify multiple groups of class, we applied an OVR method along with the SVM algorithm, which is used, for multiple class classifications. This method separates each class from the rest of the classes in the dataset. Besides, because LinearSVC is used in this study, one vs rest is an appropriate method with LinearSVC. The third proposed classifier, the Decision Tree (DT) classifier is the widely used machine-learning algorithm that makes decisions based on the conditions present in the data. A decision tree always considers the entire data as a root before starting partition. Then it starts iteratively splitting it up further on each of the

branches based on the specific condition and decides until it produces the outcome as a leaf. It starts partitioning from the tree root and split the data on the feature that has the largest information gain (IG).

4.5 Models Evaluation and Testing

Model evaluation and testing are critical in our study in order to test and estimate the performance of machine learning models trained on the chronic kidney disease dataset of both binary class and five class. Since more than two algorithms were used in this study model evaluation is very important to compare and select the best model for further prediction. In this study, the effectiveness of the predictive model is evaluated by using 10-fold cross-validation with different performance evaluation metrics appropriate for models; these metrics are precision, recall, F1_score, sensitivity, specificity, and confusion matrix are used as discussed in chapter three. The best model is with the best F1_score and accuracy is saved using pickle for deployment on the server after the evaluation.

4.6 Integrating Machine Learning Model with Server

This study aims to predict chronic kidney disease using three models and recommend the best model to be used for deployment on the server after it is evaluated with the best accuracy and f1-score. This is done by preparing the HTML form to insert the new input that is connected to the server and selected model that will allow the professionals to diagnose and give the best treatment and advice for the patient fast and correctly.

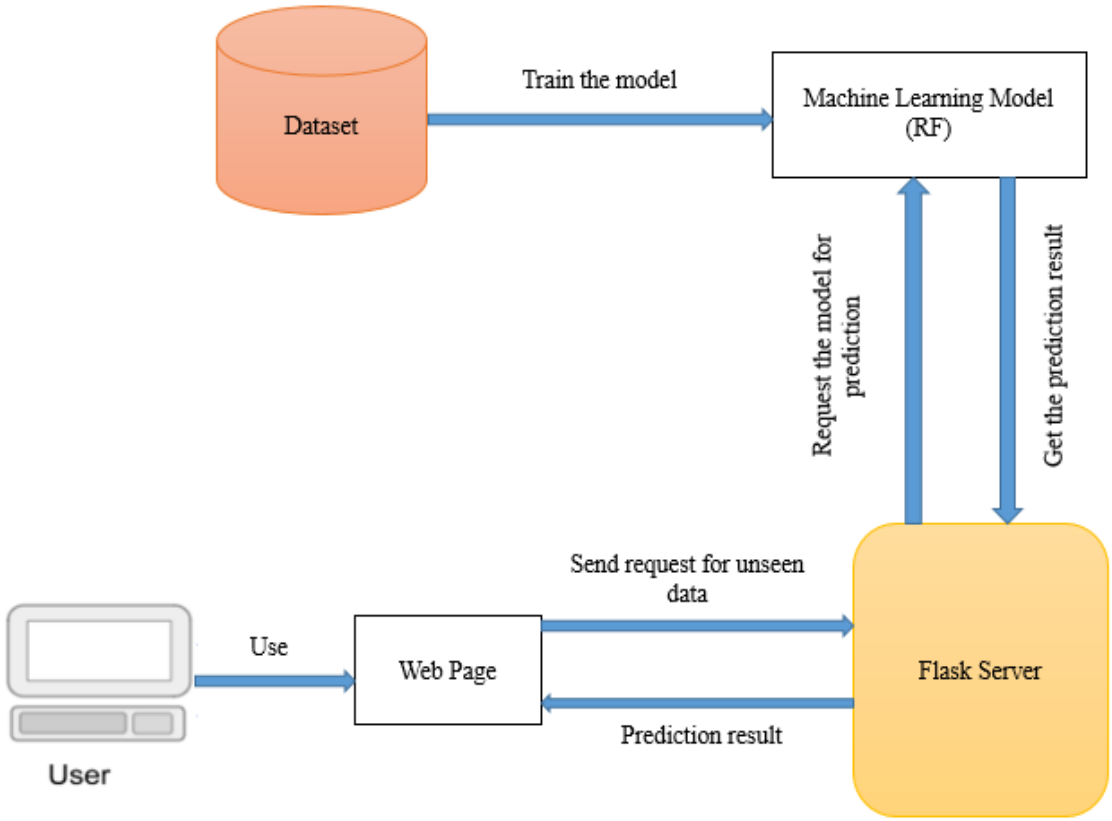

Figure 4. 5 Architecture of Model Integration with Server

CHAPTER 5

IMPLEMENTATION AND EXPERIMENTATION

The section describes the implementation of the proposed solution for chronic kidney disease prediction by using the methodology discussed in chapter three and implement the proposed solution presented in chapter four. In this study, different experiments were performed using two datasets. First, the experiment was done using a binary-class dataset, and the second performed using the five-class dataset. Finally, a prototype for chronic kidney disease prediction was developed by using the best model.

5.1 Implementation Environment

In this study, several tools and packages are used to implement the proposed solution. Python programming language has used this study for implementing and experimenting with each proposed solution from the data preprocessing to the final phase, which is the model-building phase and prototype development. The reason to choose python is that it is popular and it is the language of choice for developers, researchers, and data scientists that makes the machine learning models development simple and clear. Besides, Python is the general-purpose programming language that can be used for processing text, numbers, images, scientific data easily. It is also simple and easy to understand.

Table 5. 1 Description of the Tools and Python Packages Used During the Implementation

Tools and packages	Version	Description
Anaconda navigator	1.9.12	Help us to launch development applications and easily manage anaconda packages, environments, and channels without the need to use command-line commands.
Jupyter Notebook	6.0.1	It is a free web application that enables us to create and share documents. It is useful for data cleaning and transformation, modeling, data visualization, and machine learning.
Python	3.7.4	Python is a popular platform easy to learn, powerful programming language to develop a machine learning application.
Microsoft Excel	2016	For dataset preparation
Microsoft Word	2016	For documentation purpose
Scikit-learn	0.21.3	Is a python module used in machine learning for handling basic ML tasks clustering, regression, and classification.
Pandas	0.25.1	Pandas is an important package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data. It enables integrating and filtering of data, as well as loading it from other external sources like Excel and handling the dataset.
Numpy	1.16.5	Array processing for numbers, strings, and objects. This study uses it for handling converting the text to numeric data for features, training, and testing the model.
Matplotlib	3.1.1	Publication quality figures in python. This study uses it for data and results visualization.
Seaborrn	0.9.0	Statistical data visualization
Flask server	1.1.1	Microframework for making web services in python. This study uses it for implementing the prototype for selected models.

5.2 Deployment Environment

The tools and packages that are discussed above in section 5.1 have been deployed on a personal computer Processor Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz, 2601 Mhz, 2 Core(s), 4 Logical Processor(s), 8 Gigabyte of physical memory, 931.51 Gigabyte hard disk storage capacities and Microsoft Windows 10 Enterprise,64bit operating system.

5.3 Dataset Description

In order to build the dataset for this study, we collected the patient history data from manually recorded patient history. First, the researcher assessed different information about chronic kidney disease in Ethiopia and reviewed different material about the hospital, which works on chronic kidney disease treatment. Then St. Paulo's Hospital was selected to collect the data to build the dataset used for this research study. Then the patient history data from the year 2018 to 2019 was collected by considering the identified attributes for the dataset preparation purpose. Due to a lack of full information and time, not all-patient history data collected. This process results in a total number of 1718 patient data with ckd stages and notckd. The dataset had missing values. By applying the data preprocessing method to fill the missing values, the final five-class labeled dataset contains a total of 1718 of data which labeled with a class of notckd, mild, moderate, severe, and ESRD. The detailed method for building the dataset was discussed in chapter three and chapter four. In addition, the result labeling and size of each class in the dataset are presented in chapter six. Dataset labeling resulted in a distribution of five-classes and binary class dataset are also shown in chapter six.

The dataset contains the patient history data in Age, Gender, Bp, Sg, Chl, Sod, Pot, Bun, Scr, Hgb, Rbcc, Wbcc, Mcv, Pltc, Htn, Dm, Ane, Hd along with one of the five class. The dataset contains 19 columns with the last column as the target class is shown in Appendix A.

Table 5. 2 Dataset Features Description

Symbols	Features full name	Type	Class
Age	Age	Numeric	Predictor
Gender	Gender	Nominal	Predictor
Bp	Blood pressure	Numeric	Predictor
Sg	Specific gravity	Nominal	Predictor
Chl	Chloride	Numeric	Predictor
Sod	Sodium	Numeric	Predictor
Pot	Potassium	Numeric	Predictor
Bun	Blood Urea Nitrogen	Numeric	Predictor
Scr	Serum Creatinine	Numeric	Predictor
Hgb	Hemoglobin	Numeric	Predictor
Rbcc	Red blood cell count	Numeric	Predictor
Wbcc	White blood cell count	Numeric	Predictor
Mcv	Mean cell volume	Numeric	Predictor
Pltc	Platelet count	Numeric	Predictor
Htn	Hypertension	Nominal	Predictor
Dm	Diabetes Mellitus	Nominal	Predictor
Ane	Anemia	Nominal	Predictor
Hd	Heart disease	Nominal	Predictor
ckd_status	ckd_status	Nominal	Target

5.4 Preprocessing Implementation

Python programming language is used for this study to implement chronic kidney disease dataset preprocessing. The implementation of preprocessing activities includes handling missing values, handling categorical, and feature selection. All none numerical data should be converted to the numerical value. To fill the missing values mean the strategy is used and to transform the nominal and none numeric values the study used the Label Encoder and replace method. First, when the dataset is converted into a data frame using pandas, it replaces all blank values with NaN.

```
#Importing necessary library and unpreprocessed dataset
import pandas as pd
CKD_Dataset = pd.read_csv('New_CKD_Dataset.csv')
```

Figure 5.1 Python Code for Loading the Pandas Library and Dataset.

5.4.1 Implementation of Handling Missing Values in the Dataset

In order to fill the missing values in the dataset, we have used the fillna() method that replaces the missing values by calculating the mean value. The python SimpleImputer module is used to fill missing values using the mean strategy.

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(missing_values = np.nan, strategy = "mean")
imputer.fit(X)
X = imputer.transform(X)
```

Figure 5.2 Fill Missing Value

5.4.2 Implementation of Handling Categorical Data in the Dataset

To build a correct and appropriate dataset and to have an accurate model none numeric data should be transformed to numerical data. this data transformation method was implemented using python with the get_dummies module. The method accepts the feature value to be transformed and replace the categorical value with a numerical value.

```
X_encoded = pd.get_dummies(X, prefix_sep="_")
y_encoded = LabelEncoder().fit_transform(y)
X_scaled = StandardScaler().fit_transform(X_encoded)
```

Figure 5. 3 Handling Categorical Data

5.5 Implementation of Feature Selection

The feature selection method is used to select the relevant features from the original dataset in order to build the model. In this study, we used two feature selection methods to select relevant features to train the models. The study used sklearn module for RFECV and ANOVA First, the dataset must be preprocessed to apply the feature selection method to reduce the number of features.

5.5.1 Implementation of Univariate Feature Selection

The univariate feature selection method is applied in order to reduce the number of features in the dataset, is to inspect the correlation of our features with the output variable. using ANOVA via the f_classif() function:

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif
#K-Best Fit
X_new = SelectKBest(score_func=f_classif, k=15)
X_new = X_new.fit(X,y)
df_scores = pd.DataFrame({'features': X.columns, 'f_classifScore':X_new.scores_, 'pValue': X_new.pvalues_ })
df_scores
dfscores = pd.DataFrame(X_new.scores_)
dfcolumns = pd.DataFrame(X.columns)
#concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Features','Score']  #naming the dataframe columns
#Print the selected features
print(featureScores.nlargest(15,'Score'))
```

Figure 5. 4 ANOVA Feature Selection

5.5.2 Recursive Feature Elimination Implementation with Cross-Validation

```
#Applying Recursive Feature eleminattion
from sklearn.feature_selection import RFECV
RFmodel = RandomForestClassifier(n_estimators=100, random_state=0,max_features='auto',class_weight=None)
rfecv = RFECV(estimator=RFmodel,step=1,cv=10, scoring='accuracy',verbose=0)
rfecv = rfecv.fit(X,y)
print("Optimal number of features:%d" %rfecv.n_features_)
print("selected Features %s" %rfecv.support_)
print("Feature Ranking %s" %rfecv.ranking_)
```

Figure 5. 5 Recursive Feature Elimination

5.6 Machine Learning Models Implementation

To build machine-learning models using the proposed algorithm, we used a python sci-kit learning library package and followed the conventional method that is importing the important library package for modeling and metrics for model evaluation.

```
#Importing libraries and packages used for modeling and evaluating the model
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure
import seaborn as sns
import pickle
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif
from sklearn.feature_selection import RFECV
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import LabelEncoder,OneHotEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import LinearSVC,SVC
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.multiclass import OneVsRestClassifier
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score,cross_val_predict
from sklearn.metrics import classification_report,confusion_matrix,plot_confusion_matrix
from sklearn.pipeline import Pipeline
```

Figure 5. 6 Importing Necessary Packages for Modeling

After cleaning the dataset, the features split into independent and dependent features. Then the data without feature selection used as X training and the labels, which contain the class or labels dataset belong set as y. Then, the classifiers train by using the X and y of the entire dataset. To train using the *fit()* method with the appropriate set parameter for each classifier instantiate the class.

```
# Dataset loading and spliting it into independent and dependent
ckd_dataset = pd.read_csv("Preprocessed/Preprocessed_Final_Multiclass.csv")
#Independent and dependent variable X and y respectively
X = ckd_dataset.drop(['ckd_status'], 1) # attributes to determine dependent variable / Class
y = ckd_dataset['ckd_status']# dependent variable / Class
```

Figure 5. 7 Splitting Dataset into Dependent and Independent

RF model is implemented through RandomForestClassifier () method of the sklearn ensemble package is used to train the classifier. This classifier fits several decision tree classifiers as declare in n_estimators parameters.

```
#Instantiate Random Forest Classifier
RFmodel = RandomForestClassifier(n_estimators=200, random_state=10,max_features='auto',class_weight='balanced')
RFmodel.fit(X, y)
```

Figure 5. 8 Instantiate and Fit RF Model

Another proposed classifier in this study is Support Vector Machine. The study used the *LinearSVC()* classifier of the sklearn package to building the SVM model. This classifier is instantiating with a basic parameter like OVR to classify the multi-class dataset.

```
#Instantiate Support Vector Machine
SVM_model = LinearSVC(C=1,class_weight=None,multi_class ='ovr',random_state =0)
SVM_model.fit(X, y)
```

Figure 5. 9 Instantiate and Fit SVM Model

Another classifier used in the study is the Decision Tree classifier. To implement the DT classifier, the study again uses a sklearn *DecisionTreeClassifier*() function.

```
#Instantiate Decision tree Classifier
DTmodel = DecisionTreeClassifier (criterion='gini', max_depth=4,random_state =0)
DTmodel.fit(X_new, y)
```

Figure 5. 10 Instantiate and Fit DT Model

5.7 Model Testing and Evaluation

One of the important aspects of machine learning is evaluating the performance of learning algorithms. In this study, we used the most popular k-fold cross-validation of K = 10 that divided data equally to 10 folds, where 9-folds are used for training, and the remaining 1-fold is used for evaluation or testing. The evaluation technique implement using the sklearn cross_val_score() and cross_val_predict() methods using K value ten. These methods use the models and dataset to validate the learning skill of each model.

```
from sklearn.model_selection import cross_val_score
CVscore = cross_val_score(models,X,y,cv=10) # To train the models with 10-cv
CVscore #to get the accuracy of each fold
CVscore.mean() # to get the mean accuracy of all folds
y_pred = cross_val_predict(models,X,y,cv=10) # to get the prediction class by the models when testing 10-cv
```

Figure 5. 11 Applying 10-fold CV on Models

Additionally in this study, we used classification_report(),confusion_matrix() methods, which are used to evaluate the performance of the built models using predict the value of 10-fold CV for the entire dataset. These methods give a result of performance evaluation metrics such as precision, recall, and f1-score and other performance metrics like sensitivity and specificity value of the model.

Finally, a prototype for predicting chronic kidney disease is implemented using the selected best-performed model then deployed using flask web services for it to be able to accept a new input from the user and then return the prediction result.

CHAPTER SIX

RESULTS AND DISCUSSIONS

This chapter presents the result of the experiment of the proposed chronic kidney disease prediction using machine-learning techniques. In this section, the class distribution of the dataset and experiments to builds classification models for detection based on binary classes and five-class datasets are discussed. Then we finalize by discussing the result of each experiment in this study and illustrating the graphical user interface of the prototype.

6.1 Data Collection Result and Class Distribution Results

The total size of the data set after preprocessing is 1718. Out of the total dataset, 441 patients are end-stage renal disease stage or stage five, 399 patients are at a severe stage or stage four, 354 patients are at a moderate stage or stage three, 248 patients are at a mild stage or stage two, and 276 patients have no chronic kidney disease or normal. The class distribution in the five-class dataset is shown in the figure 6.1 below.

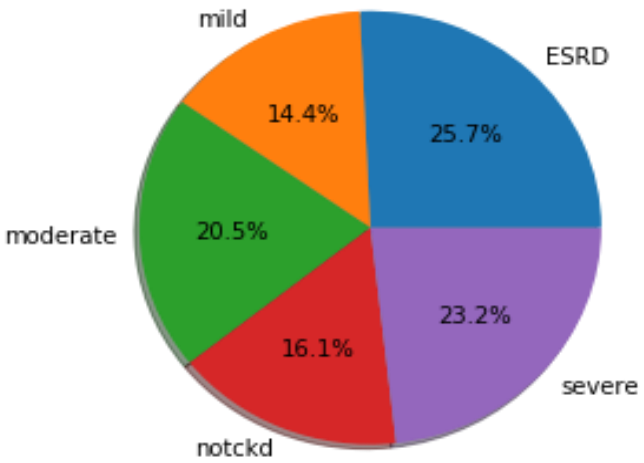

Figure 6. 1 Percentage of Class Distribution in Five-Class Dataset

Figure 6.1 indicates the percentage distribution of each class in the dataset that 16.1% is notckd, 14.4% is mild, 20.6% is moderate, 23.2% is severe and 25.7% is ESRD

The five-class dataset converted to two class datasets by considering the chronic kidney disease status like mild, moderate, severe, ESRD as ckd and notckd as it is for binary class dataset building purpose. All the classes except notckd class converted to ckd resulted in a

dataset with 1442 ckd and 276 notcckd classes. The result of the binary-class distribution of the dataset shown in figure 6.2 below.

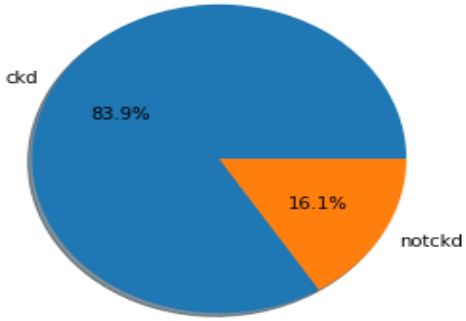

Figure 6. 2 Percentage of Class Distribution for Binary Class

Since the binary-class dataset is imbalanced, the data resampling technique was used to balance the label or class of the dataset. The oversampling data resampling technique was used to balance the value of the minority class with the value of the majority class. After the resampling method was used, the size of class labels for 'ckd_status' class attributes was balanced to 1444 ckd and 1444 notckd and total of 2888 as shown in figure 6.3

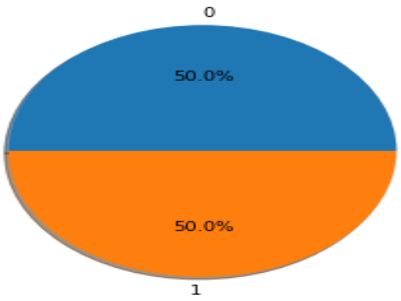

Figure 6. 3 Percentage of Class Distribution of Binary-Class Dataset after Balancing

Figure 6.3 represents that the class distribution after applying the resampling technique is balanced. The total size of a balanced dataset is 2884.

6.1.1 Feature Importance

The rank of the features based on their importance score with help of the Random Forest Classifier Algorithm is shown in Figure 6.4. The feature importance provides a score that indicates how useful each feature in the model. The more important feature in the model will have a higher score of importance. The rank of each feature is based on the comparison of other feature's importance scores in the dataset.

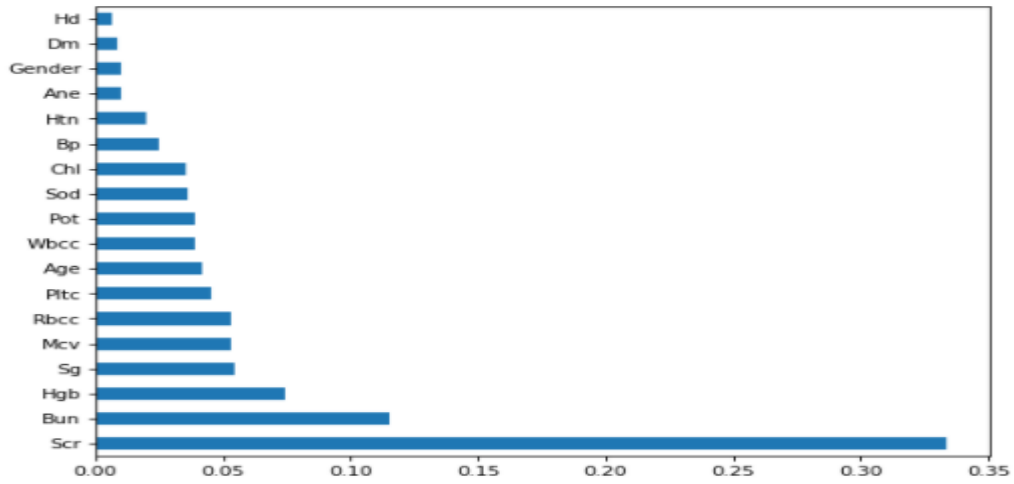

Figure 6. 4 Feature Importance Using Random Forest

6.2 Feature Selection Results

The features selection process using the two methods UFS and RFECV produced two different sets of features from the dataset, which are uses in the training of RF, SVM, and DT models. The selected features for both five-class and binary-class are different because recursive feature elimination with cross-validation can automatically eliminate less predictive features iteratively depending on the model. Table 6.1 below shows the number of selected features of binary-class and five-class that set the size of the dataset respectively.

Table 6. 1 Feature Selection Results

Feature selection method	Machine learning algorithm	Selected features (Binary)	Selected features (five-Class)
RFECV	RF	8	9
	SVM	9	10
	DT	16	7
UFS(ANOVA)	RF	14	14
	SVM	14	14
	DT	14	14

6.3 Model Building

In this study, three models namely; Random Forest, Support Vector Machine, and Decision Tree are proposed. First, the data set is split into independent and dependent. Independent features are the features whose values do not determine by others and the dependent feature is the target or the class whose value depends on the independent features. Then, the models were built on binary class and multiclass datasets without and with the implementation of feature selection methods namely; UFS (ANOVA) and RFECV.

6.4 Model Evaluation Results

The Experiment resulted in 18 models based on two feature selection methods and three classifiers for both binary and five-class classification. Testing and training these models is performed using 10-fold cross-validation with other performance evaluation metrics as discussed in chapter three and chapter four. 10-fold cross-validation method randomly split the dataset into ten equal size sets. Train the models using 10-1 folds and test using one remain fold. The process is iteratively for each fold. The obtained results are presented in binary and five-class classification models. In our thesis work, we divide the justification of results into two sets, first, we conduct the models on the original dataset for both binary class and multiclass, and second, we apply the two feature selection methods with models on both datasets.

6.4.1 Binary Classification Models Evaluation Results on Original Dataset

These classification models built using the two-class dataset that is converted from the five-class dataset that means the target classes are notckd and ckd. The models are trained and tested using the 10-fold CV and other performance evaluation metrics of CV. In our experiment, first, we train and test the model on the original dataset without applying feature selection methods. The result of each test accuracy score for RF, SVM, and DT models before feature selection and after feature selection method are presented in tables separately below as follows.

Table 6. 2 Models Accuracy of the Binary-Class Dataset

Models	**10-Folds accuracy (%) for RF, SVM, DT**										
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc (%)
RF	1.00	99.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	99.0	99.7
SVM	97.0	96.0	1.00	99.0	98.0	96.0	95.0	95.0	97.0	97.0	96.9
DT	99.0	98.0	1.00	99.0	99.0	1.00	98.0	98.0	99.0	98.0	98.5

Table 6.2 indicates the CV accuracy score for three classifiers with each fold of the binary class dataset before applying feature selection. The lowest average accuracy of 96.9% results was recorded on the SVM model, and the highest average accuracy of 99.7% resulted in the RF model. In addition to the accuracy of 10-fold cross-validation, this study uses other performance evaluation metrics such as precision, recall, f1_score, sensitivity, specificity, and confusion matrix as discussed in chapter three and chapter four. Table 6.3 below shows the result of each performance evaluation metrics for each model with the 10-fold CV. Normalized Confusion matrix for each model described in the following figures to show how correctly the models classify each class along with prediction error.

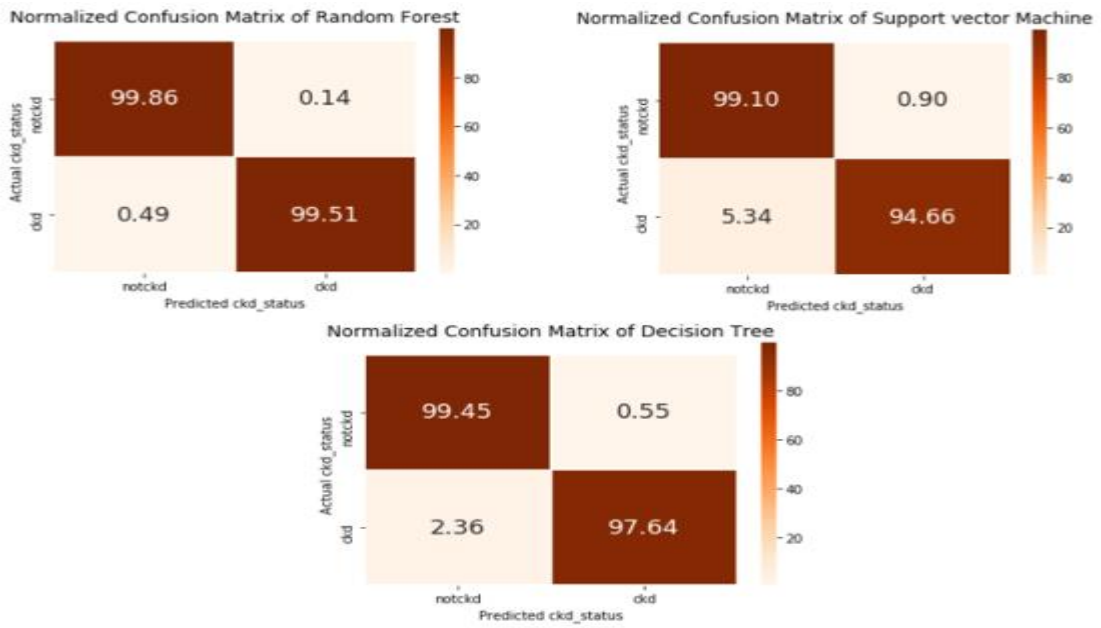

Figure 6. 5 Confusion Matrix of Binary RF, SVM, and DT before Feature Selection

Figure 6.5 represents the normalized binary confusion matrix for RF, SVM, and DT. RF classifies 99.86% of notckd and 99.51% of ckd correctly and 0.14% of notckd as ckd and 0.49% of ckd as notckd are misclassified. SVM classifies 99.10% of notckd and 94.66% of ckd correctly. SVM misclassify 0.9% of notckd as ckd and 5.34% of ckd as notckd. DT classify 99.45% of notckd, 97.64% of ckd correctly, and misclassifies 0.55% of notckd as ckd and 2.36% ckd as notckd. Finally, the result of binary models before applying the feature selection method demonstrates that the RF model shows better accuracy, f1_score, and sensitivity than the SVM and DT.

Table 6.3 Performance Result of Binary Classification Models

Models	Performance evaluation metrics				
	Precision	Recall	F1_score	Sensitivity	Specificity
RF	99.9	99.5	99.7	99.5	99.9
SVM	99.1	94.7	96.8	94.7	99.1
DT	99.4	97.6	98.5	97.6	99.4

Table 6.3 express the performance evaluation metrics results of the three classifiers. Based on the accuracy value because the dataset is balanced Random Forest has a higher score for each performance metric than the rest two models (SVM and DT).

6.4.2 Binary Classification Models Evaluation Results after Feature Selection

We have implemented the feature selection techniques to select the relevant and most predictive features after implementing the model on original dataset. The importance of using feature selection techniques is to reduce overfitting, reduce processing time, and improve the accuracy of the model. However, it does not mean that the implementation of feature selection always improves the accuracy of the model since the selected features set determine the accuracy of the model. We train and test the three RF, SVM, and DT models based on the selected features using the two UFS and RFECV. The result of each model on the best-selected features is evaluated using a 10-fold cross-validation method to identify better feature selection method with which model. The result of each test accuracy score for

RF, SVM, and DT models based on selected features and normalized confusion matrix is discussed as follows.

Table 6. 4 Binary Class Accuracy of RF Model with Each Feature Selection

FS methods	**10-Folds accuracy (%) for RF**										
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc. (%)
RFECV	1.00	99.3	1.00	1.00	99.7	1.00	1.00	99.7	1.00	99.3	99.8
UFS	1.00	99.3	1.00	1.00	99.7	1.00	1.00	99.7	99.7	98.6	99.7

Table 6.4 shows the CV accuracy result of each fold for Random Forest Classifier with feature selection techniques. As shown in table 6.1 in section 6.2 8 features and 14 features are selected using RFECV and UFS respectively. The result shows the higher mean CV accuracy of 99.8% recorded in RF with RFECV. The performance of the model was improved after implementing feature selection when compared to the performance of the model on the original dataset based on RF with RFECV.

Figure 6.6 below represent the visualization of the number of selected feature along with the cross-validation score using RFECV and RF.

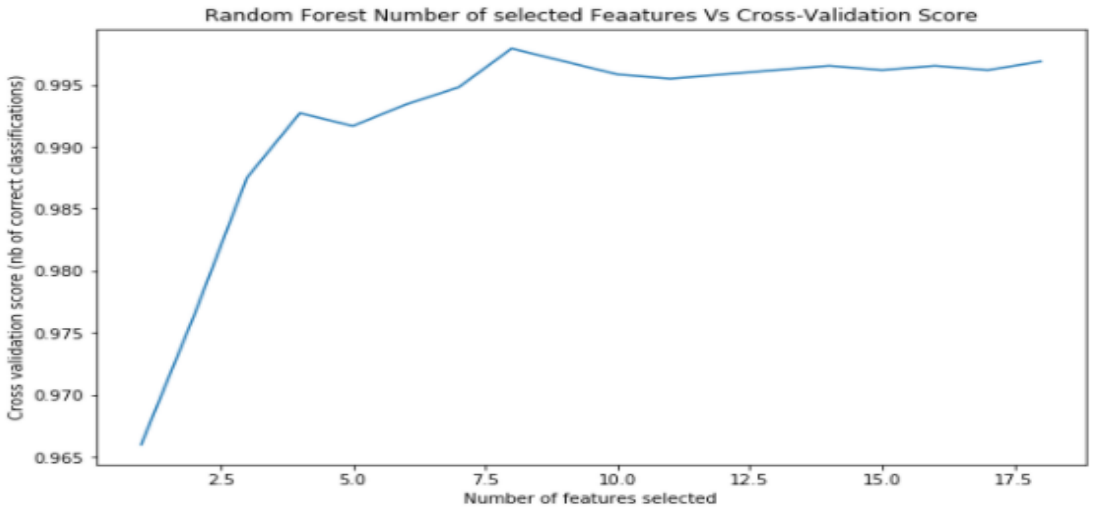

Figure 6. 6 RFECV with RF Number of Selected Features Vs Cross-Validation Score on the Binary Dataset

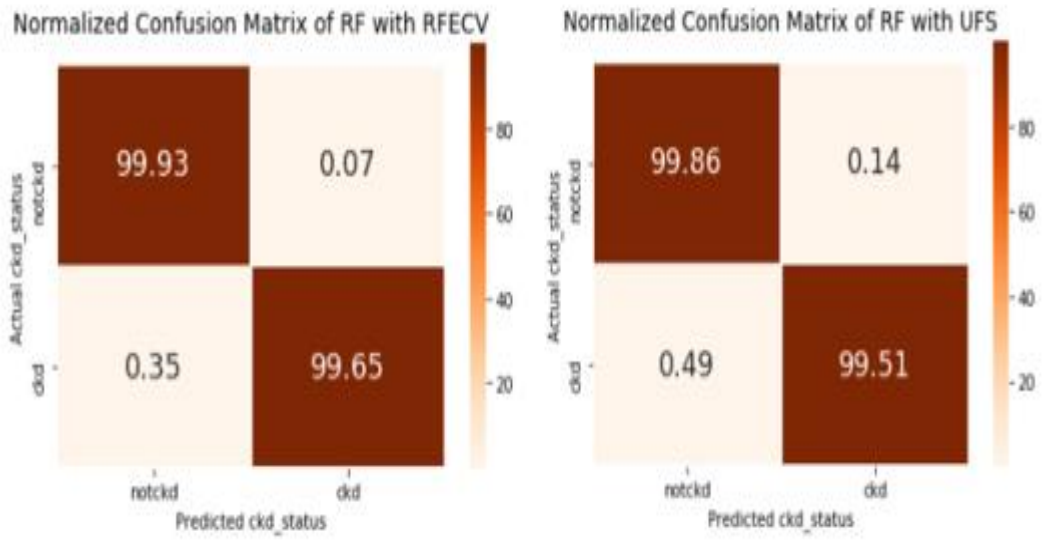

Figure 6. 7 Normalized Confusion Matrix of RF with RFECV and UFS

Figure 6.7 illustrates the normalized confusion matrix of the Random Forest model with the two feature selection methods namely RFECV and UFS respectively. RF with RFECV classify 99.65% of ckd and 99.93% of notckd correctly and misclassify 0.35% of ckd as notckd and 0.07 of notckd as ckd. RF with UFS classify 99.51% of ckd and 99.86% of notckd correctly and misclassify 0.49 % of ckd as notckd an 0.14% of notckd as ckd.

Table 6. 5 Binary Class Accuracy of SVM Model with Each Feature Selection

FS methods	**10-Folds accuracy (%) for SVM**										
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc. (%)
RFECV	95.50	93.77	99.31	98.27	97.22	94.10	93.40	94.44	95.14	93.40	95.5
UFS	97.23	95.16	99.65	98.62	97.57	95.83	94.10	95.83	96.88	96.53	96.7

Table 6.5 shows the CV accuracy result of each fold and mean accuracy for Support Vector Machine with feature selection techniques. As shown in table 6.1 9 and 14 features selected using RFECV and UFS respectively. The result shows a higher mean CV accuracy of 96.7% recorded in SVM with RFECV. Figure 6.8 below represent the visualization of the number of selected feature along with the cross-validation score using RFECV and SVM.

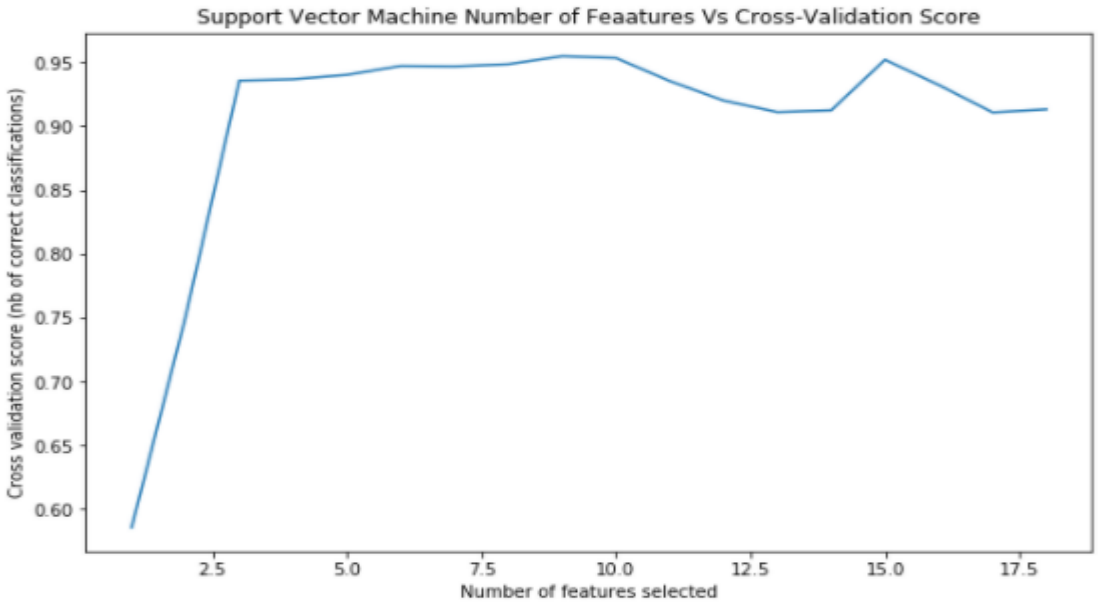

Figure 6. 8 RFECV with SVM Number of Selected Features Vs Cross-Validation Score on the Binary Dataset

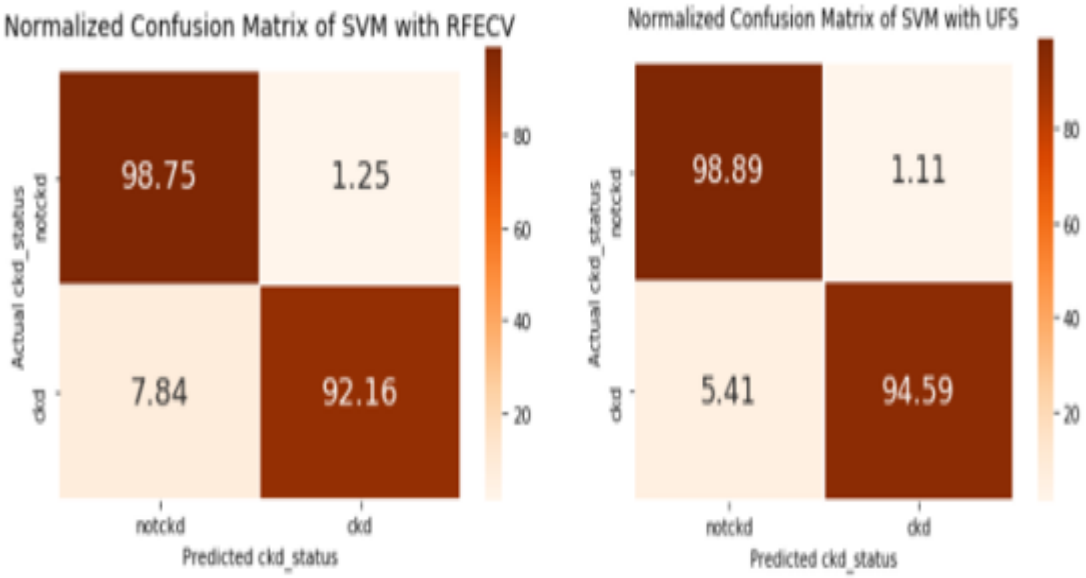

Figure 6. 9 Normalized Confusion Matrix of SVM with RFECV and UFS

The normalized confusion matrix of Support Vector Machine model with the two feature selection methods is shown on Figure 6.9. SVM with RFECV classify 92.16% of ckd and 98.75% of notckd correctly and misclassify 7.84% of ckd as notckd and 1.25% of notckd as ckd. SVM with UFS classify 94.59% of ckd and 98.89% of notckd correctly and misclassify 1.11% of notckd as ckd and 5.41%% of ckd as notckd.

Table 6. 6 Binary Class Accuracy of DT Model with Each Feature Selection

FS methods	**10-Folds accuracy (%) for DT**										
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc. (%)
RFECV	98.96	98.27	99.65	99.31	98.61	1.00	97.22	97.92	97.57	98.26	98.6
UFS	98.96	98.27	99.65	99.65	98.61	1.00	97.57	97.92	97.57	96.88	98.5

Table 6.6 shows the CV accuracy result of each fold and mean accuracy for the Decision Tree with feature selection techniques. As shown in table 6.1, 16, and 14 features are selected using RFECV and UFS respectively. The result shows the higher mean CV accuracy of 98.6% recorded in DT with RFECV. Figure 6.10 below represents the number of selected features along with the cross-validation score using RFECV and DT.

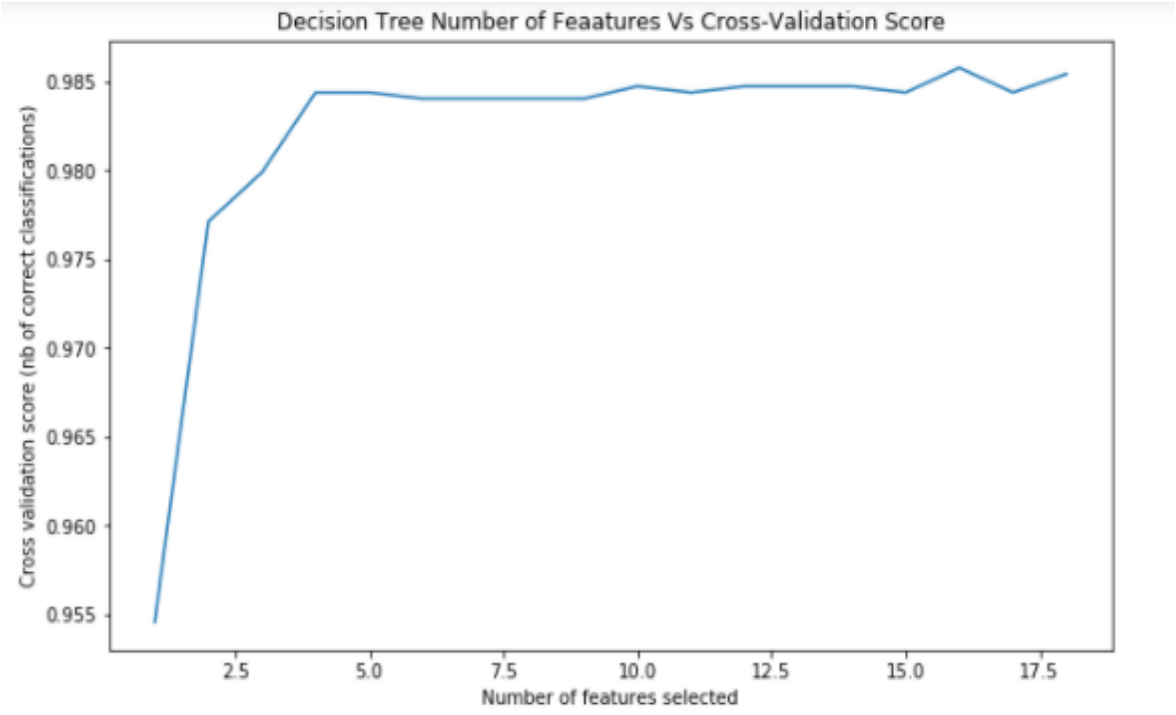

Figure 6. 10 RFECV with DT Number of Selected Features Vs Cross-Validation Score on the Binary Dataset

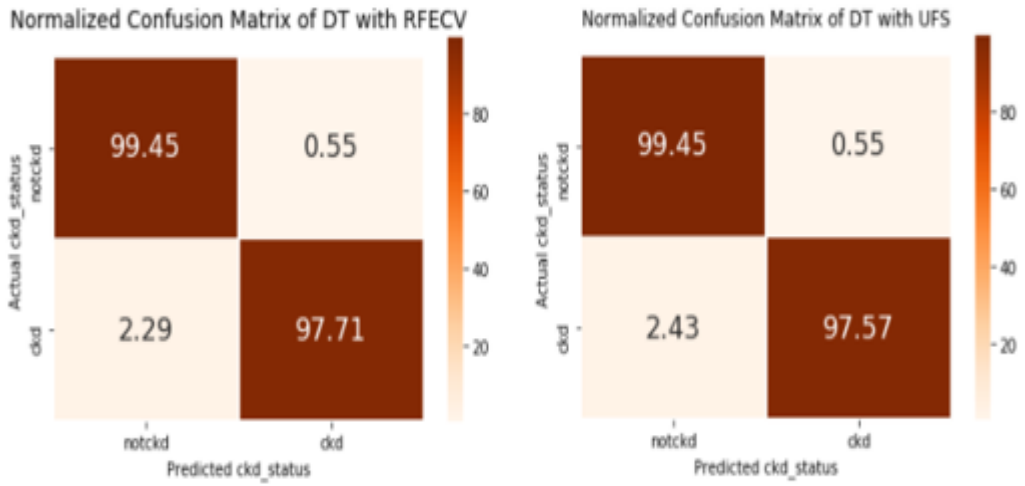

Figure 6. 11 Normalized Confusion Matrix of DT with RFECV and UFS

Figure 6.11 present the normalized confusion matrix of the Decision Tree model with the two feature selection methods. DT with RFECV classify 97.71% of ckd and 99.45% of notckd correctly and misclassify 2.29% of ckd as notckd and 0.55% of notckd as ckd. DT with UFS classify 97.57% of ckd and 99.45% of notckd correctly and misclassify 0.55% of notckd as ckd and 2.43% of ckd as notckd.

Additionally, we evaluated the performance of the model using other performance evaluation metrics discussed in chapter three and chapter four.

Table 6. 7 Performance Result of Binary Models with Feature Selection Methods

Feature selection method and models		**Performance evaluation metrics**				
		Precision	Recall	F1_score	Sensitivity	Specificity
RFECV	RF	99.9	99.7	**99.8**	**99.7**	99.9
	SVM	98.7	92.2	**95.3**	**92.2**	98.8
	DT	99.4	97.7	98.6	97.7	99.4
UFS	RF	99.9	99.5	99.7	99.5	99.9
	SVM	98.8	94.6	**96.6**	**94.6**	98.9
	DT	99.4	97.6	98.5	97.6	99.4

In Table 6.7 the result of the classification model performance with each feature selection method using performance evaluation metrics is shown. Based on F1_score and sensitivity value RF with RFECV score the highest result of 99.8% and 99.7 than RF with UFS, SVM with RFECV, SVM with UFS, DT with RFECV, and DT with UFS. Besides the accuracy 99.8 of the RF with RFECV is the highest than the rest models. Thus, based on the accuracy, sensitivity and F1_score value RF with RFECV perform better than the others do. Accuracy and F1_score are selected to compare the models because both are good performance evaluation metrics in balanced class distribution.

Table 6. 8 Summary of Binary Class Classification Models Results

Models	No of features	Acc.	Prec	Rec	F1	Sen	Spec
RF	18	99.7	99.9	99.5	99.7	99.5	99.9
RF with RFECV	8	99.8	99.9	99.7	**99.8**	**99.7**	99.9
RF with UFS	14	99.7	99.9	99.5	99.7	99.5	99.9
SVM	18	96.9	99.1	94.7	96.8	94.7	99.1
SVM with RFECV	9	95.5	98.7	92.2	**95.3**	**92.2**	98.8
SVM with UFS	14	96.7	99.4	97.7	98.6	97.7	99.4
DT	18	98.5	99.4	97.6	98.5	97.6	99.4
DT with RFECV	16	98.6	99.4	97.7	98.6	97.7	99.4
DT with UFS	14	98.5	99.4	97.6	98.5	97.6	99.4

Table 6.8 shows the summary of classification performance results of binary class models before and after feature selection from the result of each models stated in section 6.4.1 and 6.4.2 separately. Finally, the result of binary models after applying the feature selection method demonstrates that the RF model based on the RFECV method shows better accuracy, f1_score, and sensitivity than the SVM and DT and also than the performance models on the original dataset.

6.4.3 Multiclass Classification Models Evaluation Results

The multiclass models built using the prepared five-class dataset. The models are trained and tested using the 10-fold CV and evaluated with other performance evaluation metrics. The result of each trained model testing accuracy score presented for RF, SVM, and DT respectively, without and with feature selection method. In this section, first, we train and test the model with all features without applying feature selection methods and then we apply the feature selection methods.

Table 6. 9 Models Accuracy Scores of the Multiclass Dataset

Models	**10-Folds accuracy (%) for RF, SVM, DT**										
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc. (%)
RF	86.6	75.0	65.1	79.6	80.8	83.1	81.9	74.4	77.8	78.9	78.3
SVM	66.9	63.9	48.8	63.9	62.8	64.5	62.8	65.1	69.0	61.9	63.0
DT	78.5	70.9	63.9	81.4	83.7	80.8	86.0	75.6	78.4	76.0	77.5

From the 10-fold CV accuracy score given in Table 6.9 the highest accuracy, 78.3% scored in the RF model. The normalized confusion matrix of each model before the implementation of feature selection methods is given in the figure below.

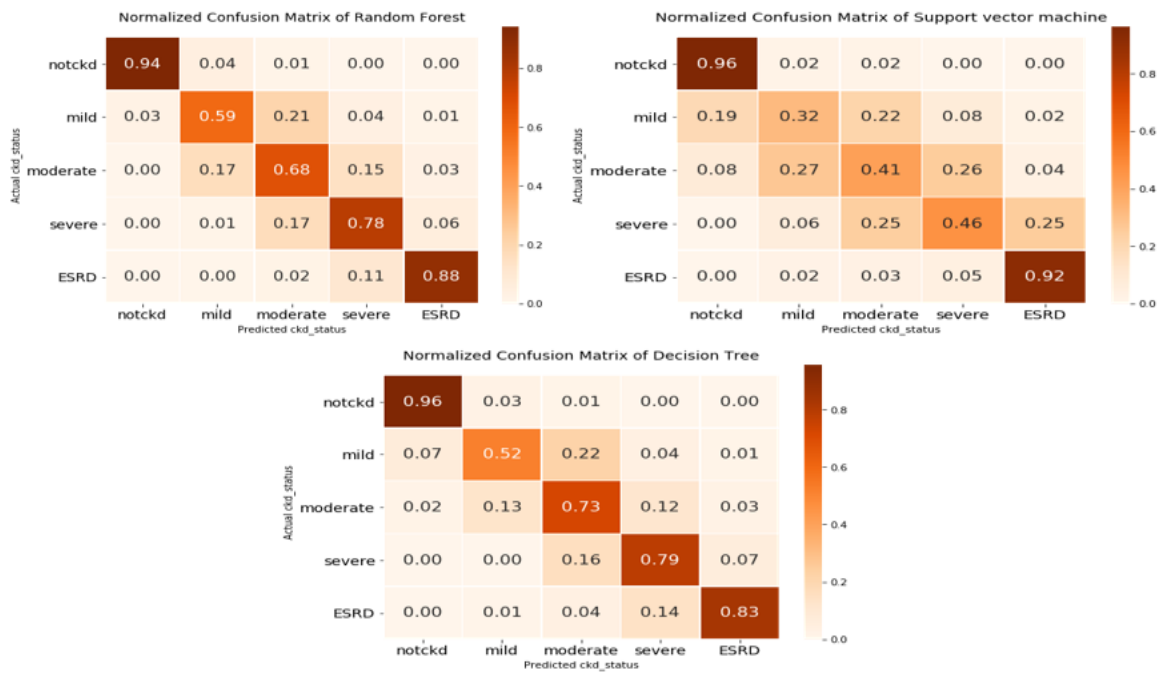

Figure 6. 12 Confusion Matrix of Multiclass RF, SVM, and DT on the Original Dataset

Figure 6.12 shows the normalized confusion matrix of three models before applying the feature selection method. RF classifies 94% of notckd, 59% of mild, 68% of moderate, 78% of severe and 88% of ESRD correctly. RF misclassify 4%, 1%, of notckd as mild and moderate, 3%, 21%, 4%, 1% of mild as notckd, moderate, severe and ESRD respectively, 17%, 15%, 3% of moderate as mild, severe and ESRD, 1%, 17%, 6% of severe as mild, moderate and ESRD respectively and 2% and 11% of ESRD as moderate and severe respectively.

SVM classifies 96% of notckd, 32% of mild, 41% of moderate, 46% of severe, and 92% of ESRD correctly. SVM incorrectly classify 2%, 2% of notckd as mild, moderate, 19%, 22%, 8%, 2% of mild as notckd, moderate, severe, and ESRD respectively. It also incorrectly classify 8%, 27%, 26%, and 4% of moderate as notckd, mild, severe, and ESRD, 6%, 25%, and 25% of severe as mild, moderate, and ESRD respectively, and 2%, 3% and 5% of ESRD as mild, moderate and severe respectively.

DT classifies 96% of notckd, 52% of mild, 73% of moderate, 79% of severe, and 83% of ESRD correctly. DT misclassify 3%, 1%of notckd as mild, moderate respectively, 7%, 22%, 4%, 1% of mild as notckd, moderate, severe and ESRD respectively, 2%, 13%, 12% and 3% of moderate as notckd, mild, severe and ESRD, 16%, 7% of severe as moderate and ESRD respectively and 1%, 4%, 14% of ESRD as mild, moderate and severe respectively.

In addition to 10-fold CV, mean accuracy, evaluation metrics such as precision, recall, f1-score, sensitivity, and specificity are used in order to evaluate the performance of the models. Table 6.10 shows the results of the model's performance evaluation metrics for each five-class classification model before applying feature selection.

Table 6. 10 Classification Performance Result of Five-Class Models

Models	Performance evaluation metrics				
	Precision	Recall	F1_score	Sensitivity	Specificity
RF	79.5	77.4	77.3	77.4	94.5
SVM	60.9	661.5	59.9	61.6	90.6
DT	79.6	76.4	76.2	76.5	94.3

Table 6.10 designates the classification report of each model and the RF model has the highest F1-score than the rest model. Thus Random Forest model is the best in our experiment before applying feature selection methods.

6.4.4 Multiclass Classification Results of Models with Feature Selection

After implementing the models on the original dataset then we apply the feature selection methods on the dataset and two feature sets are created. The result of each test accuracy score for RF, SVM, and DT models based on selected features is presented in tables separately below respectively.

Table 6. 11 Cross-Validation Result for Random Forest with Feature Selection

FS	**10-Folds accuracy (%) for RF**										
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc. (%)
RFECV	86.0	74.4	68.0	79.7	80.8	82.6	83.1	76.2	80.7	78.9	79.0
UFS	86.0	73.3	66.9	81.9	76.7	83.1	81.4	72.1	77.8	77.2	77.6

Table 6.11 represents the CV accuracy result of each fold and mean accuracy for Random Forest Classifier with feature selection techniques. As indicated in table 6.1 of section 6.2 9 features and 14 features are selected using RFECV and UFS respectively. The result shows the higher mean CV accuracy recorded in RF with RFECV 79.0%. Figure 6.13 and Figure 6.14 below represents a visualization of the number of selected features along with the cross-validation score using RFECV and RF and the normalized confusion matrix of RF with RFECV and UFS.

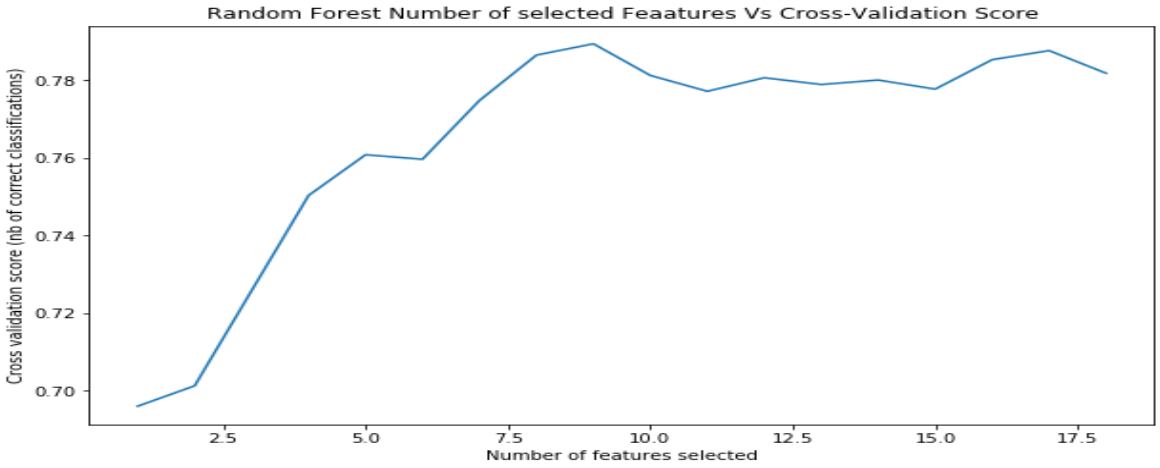

Figure 6. 13 RFECV with RF Number of Selected Features Vs Cross-Validation Score on the Multiclass Dataset

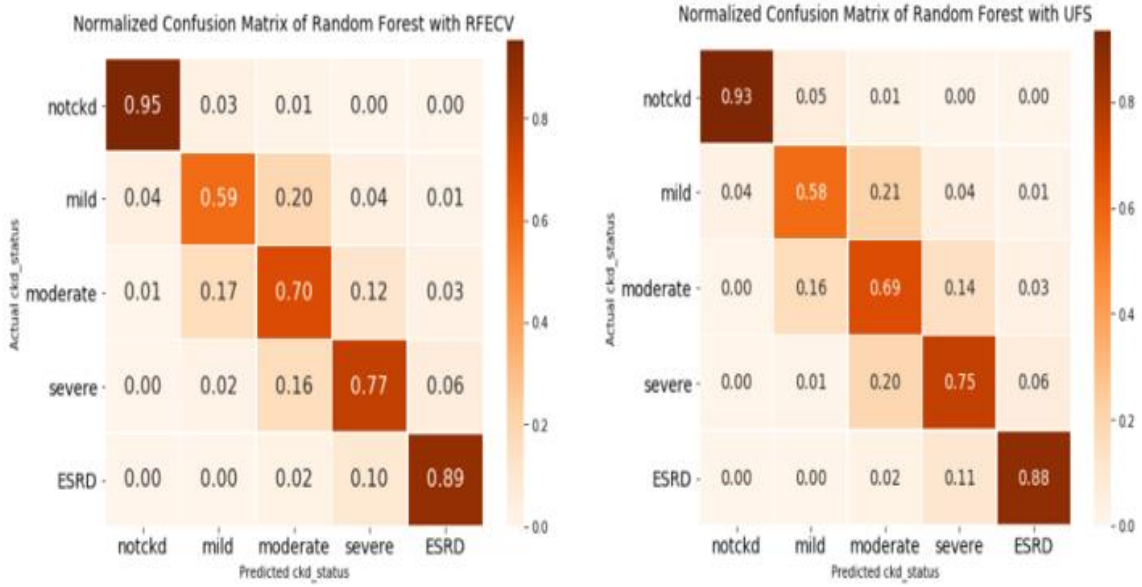

Figure 6. 14 Normalized Confusion Matrix of RF with RFECV and UFS Respectively on Multiclass Dataset.

Figure 6.14 shows the normalized confusion matrix of the Random Forest Classifier with RFECV and UFS. It represents the percentage of the correctly classified and incorrectly classified classes. RF with RFECV classifies 95% of notckd, 59% of mild, 70% of moderate, 77% of severe, and 89% of ESRD correctly.

But it misclassifies 3% and 1%of notckd as mild and moderate, 4%, 20%, 4%, 1%, of mild as notckd, moderate, severe and ESRD respectively, 1%, 17%, 12%, 3% of moderate as notckd, mild, severe and ESRD , 2%, 16%, 6% of severe as mild, moderate and ESRD respectively and 2%, 10% of ESRD as moderate and severe respectively.

RF with UFS classifies 93% of notckd, 58% of mild, 69% of moderate, 75% of severe and 88% of ESRD correctly. But it misclassifies 5% , 1% of notckd as mild, moderate, 4%, 21%, 4%, 1% of mild as notckd, moderate, severe and ESRD respectively, 16%, 14%, 3% of moderate as mild, severe and ESRD, 1%, 20%, 6% of severe as mild, moderate and ESRD respectively and 2%, 11% of ESRD as moderate and severe respectively.

Table 6. 12 Cross-Validation Result for Support Vector Machine with Feature Selection

FS	**10-Folds accuracy (%) for SVM**										
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc. (%)
RFECV	69.8	59.9	56.9	64.5	57.6	70.4	66.9	58.1	64.9	52.6	62.2
UFS	62.2	60.5	50.6	61.6	58.7	65.7	68.0	63.9	66.1	57.9	61.5

Table 6.12 shows the CV accuracy result of each fold and the mean accuracy of SVM with RFECV and UFS. As shown in table 6.3 in section 6.2 10 features and 14 features are selected using RFECV and UFS respectively. The result shows the higher mean CV accuracy recorded in SVM with RFECV 62.2. The figure 6.15 and figure 6.16 below represent the number of selected features along with the cross-validation score using RFECV and SVM and normalized confusion matrix of SVM with RFECV and UFS respectively.

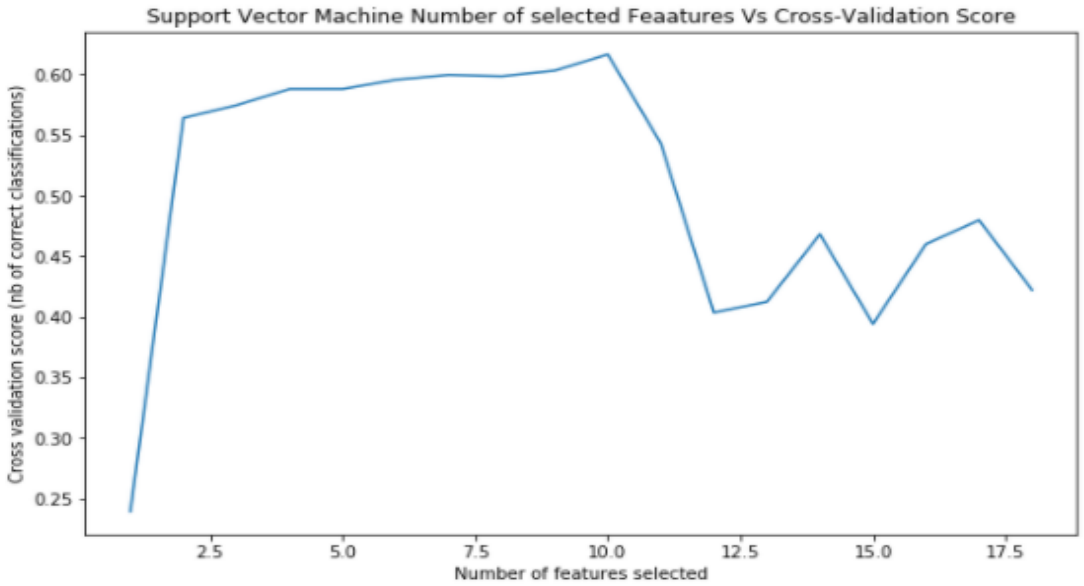

Figure 6. 15 RFECV with SVM Number of Selected Features Vs Cross-Validation Score on the Five-Class Dataset

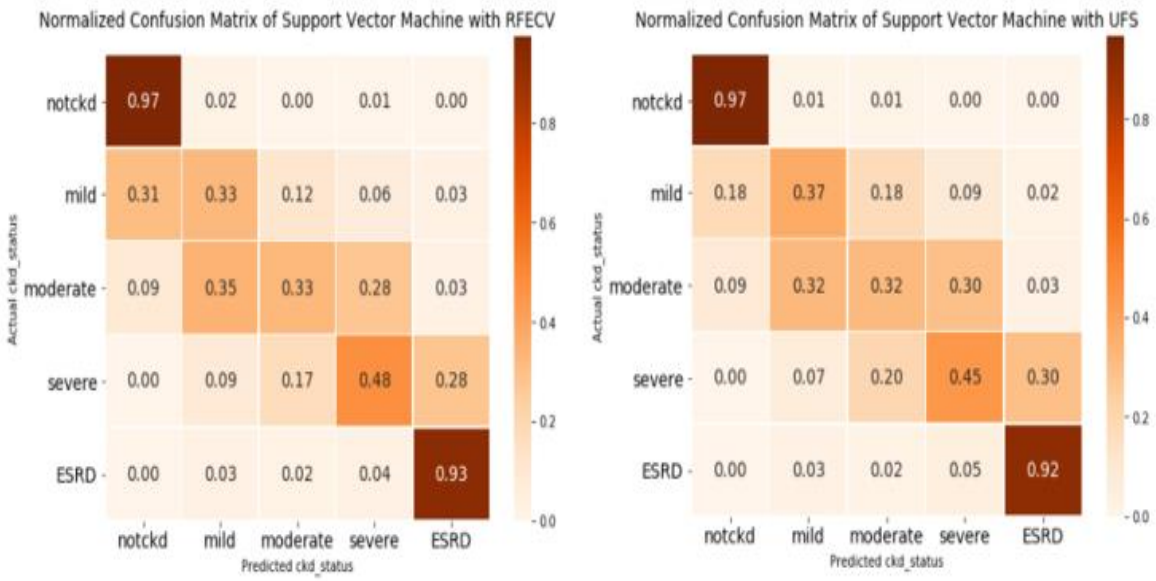

Figure 6. 16 Normalized Confusion Matrix of SVM with RFECV and UFS Respectively on the Five-Class Dataset.

Figure 6.16 illustrates the normalized confusion matrix of the SVM model with feature selection methods. SVM with RFECV classifies 97% of notckd, 33% of mild, 33% of moderate, 48% of severe, and 93% of ESRD correctly. But it incorrectly classifies 2%, 1% of notckd as mild and moderate, 31%,12%, 6%, 3% of mild as notckd, moderate, severe and ESRD respectively, 9%, 35%, 28%, 3% of moderate as notckd, mild, severe and ESRD, 9%,

17%, 28% of severe as mild, moderate and ESRD respectively. It classifies 3%, 2% and 4% of ESRD as mild, moderate, and severe respectively.

SVM with UFS classifies 97% of notckd 37% of mild, 32% of moderate, 45% of severe, and 92% of ESRD correctly. However, it misclassifies 1% of notckd as mild and moderate, 18%, 18%, 9%, 2% of mild as notckd, moderate, severe, and ESRD respectively. It also misclassifies 9%, 32%, 30%, and 3% of moderate as notckd, mild, severe, and ESRD, 7%, 20%, 30% of severe as mild, moderate, and ESRD respectively, and 3%, 2%, 5% of ESRD as mild, moderate and severe respectively.

Table 6. 13 Cross-Validation Result for Decision Tree Classifier with Feature Selection

FS	10-Folds accuracy (%) for DT										
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc. (%)
RFECV	79.1	70.9	63.9	81.4	85.5	80.8	86.6	76.7	78.4	76.6	78.0
UFS	79.7	70.4	65.1	81.4	84.9	80.8	84.9	78.5	77.2	76.6	77.9

Table 6.13 shows the CV accuracy result of each fold and the mean accuracy of the Decision Tree model with feature selection techniques. As shown in table 6.2 7 features and 14 features are selected using RFECV and UFS respectively. The result shows that higher mean CV accuracy of 78.00% recorded in DT with RFECV. Figure 6.17, 6.18 below represents the visualization of the number of selected features along with the cross-validation score using RFECV and DT and normalized confusion matrix of DT with RFECV and UFS.

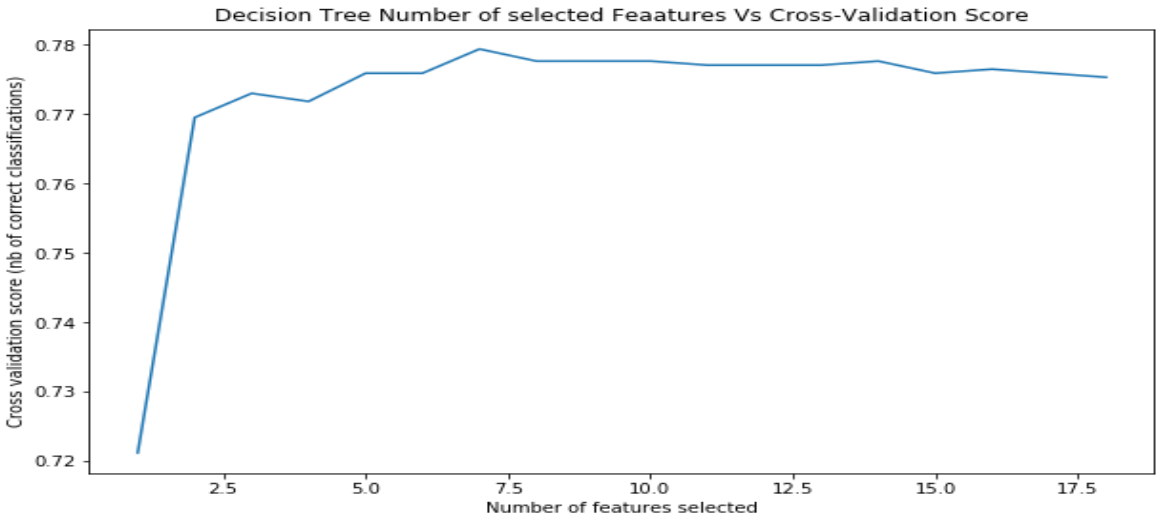

Figure 6. 17 RFECV with DT Number of Selected Features Vs Cross-Validation Score on the Five-Class Dataset

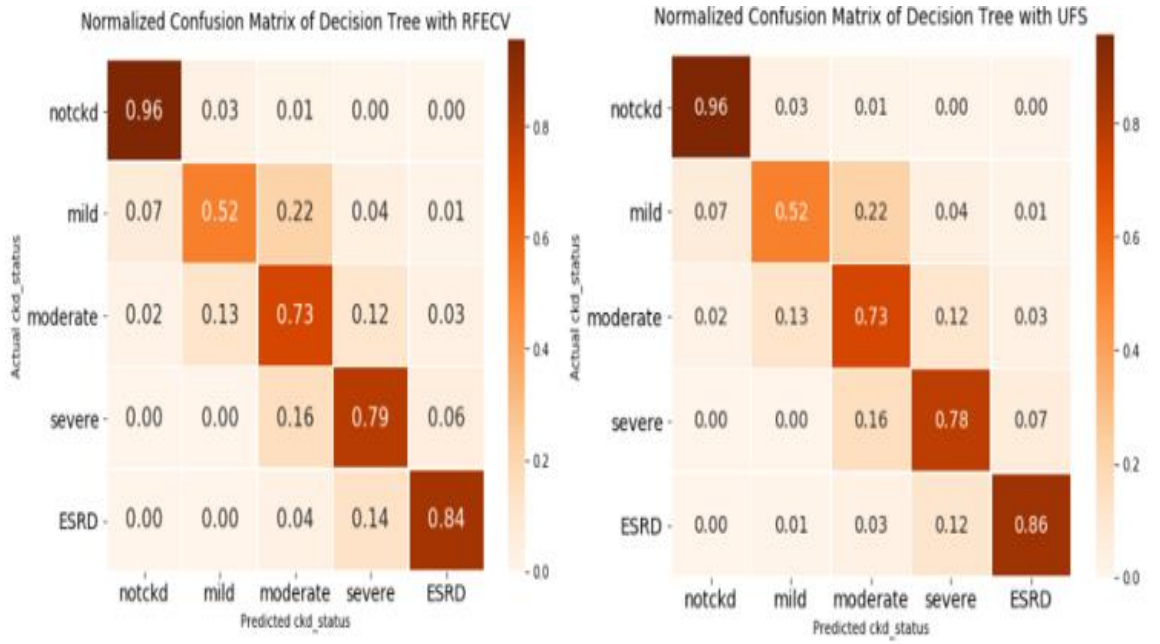

Figure 6. 18 Normalized Confusion Matrix of DT with RFECV and UFS Respectively on the Five-Class Dataset.

Figure 6.18 shows the normalized confusion matrix of the DT model with feature selection methods. DT with RFECV classifies 96% of notckd, 52% of mild, 73% of moderate, 79% of severe, and 84% of ESRD correctly. But it incorrectly classifies 3%, 1%, of notckd as mild an moderate, 7%, 22%, 4%, 1% of mild as notckd, moderate, severe and ESRD respectively, 2%, 13%, 12%, 3% of moderate as notckd, mild, severe and ESRD, 16%, 6% of severe as moderate and ESRD respectively and 4% and 14% of ESRD as moderate and

severe respectively. DT with UFS classifies 96% of notckd, 52% of mild, 73% of moderate, 78% of severe, and 86% of ESRD correctly. But it incorrectly classifies 3%, 1% of notckd as mild, moderate, 7%, 22%, 4%, 1% of mild as notckd, moderate, severe, and ESRD respectively, 2%, 13%, 12%, 3% of moderate as notckd, mild, severe and ESRD, 16%, 7% of severe as moderate and ESRD respectively and 1%, 3% and 12% of ESRD as mild, moderate and severe respectively. Besides, the accuracy value of CV in this study, performance evaluation metrics such as precision, recall, f1_score, sensitivity, and specificity are used to evaluate the model for the five-class dataset and given in table 6.14.

Table 6. 14 Performance Result of Models with Feature Selection Methods on Multiclass

Feature selection method with models		Performance evaluation metrics				
		Precision	Recall	F1_score	Sensitivity	Specificity
RFECV	RF	79.8	78.1	**77.9**	**78.1**	94.7
	SVM	58.7	60.7	56.5	60.8	90.5
	DT	80	76.9	76.6	76.9	94.4
UFS	RF	78.9	76.6	76.7	76.7	94.3
	SVM	59.2	60.5	**58.4**	**60.5**	**90.2**
	DT	79.8	76.7	76.4	76.7	94.4

Table 6.14 shows the result of each model based on the selected with feature selection techniques. Based on the F1-score and sensitivity, the RF model with RFECV obtains a higher score of 77.9% and 78.1% respectively than SVM and DT models. F1- score metrics selected to compare the models based on the features because it is more useful than accuracy when the dataset contains imbalanced class distribution.

Table 6. 15 Summary of Multi Class Classification Models Results

Models	No. of features	Acc.	Prec	Rec	F1	Sen	Spec
RF	18	78.3	79.5	77.4	77.3	77.4	94.5
RF with RFECV	9	79.0	79.8	78.1	**77.9**	**78.1**	94.7
RF with UFS	14	77.6	78.9	76.6	76.7	76.7	94.3
SVM	18	63.0	60.9	61.5	59.9	61.6	90.6
SVM with RFECV	10	62.2	58.7	60.7	56.5	60.8	90.5
SVM with UFS	14	61.5	59.2	60.5	**58.4**	**60.5**	**90.2**
DT	18	77.5	79.6	76.4	76.2	76.5	94.3
DT with RFECV	7	78.0	80	76.9	76.6	76.9	94.4
DT with UFS	14	77.9	79.8	76.7	76.4	76.7	94.4

Table 6.15 shows a summary of the performance of multiclass models before and after applying feature selection methods described in section 6.4.3 and 6.4.4 separately. It shows that RF with RFECV scored the highest accuracy of 79.0, f1_score of 77.9, and sensitivity of 78.1 of all models before and after employing feature selection methods.

6.5 Discussions

Now a day chronic kidney disease is become challenging throughout the world and it became a silent killer in Ethiopia[7]. Many people die or suffer the severe stage of this disease due to a lack of awareness about the disease and lack of early diagnosis. Thus, early prediction of chronic kidney disease is helpful to slow the progress of the disease. Machine learning plays a vital role in early disease identification and prediction. Machine learning supports medical experts as a decision support system, enabling them to diagnose the disease fast and accurately.

This study proposed chronic kidney disease prediction using machine learning techniques with a newly prepared dataset. The prepared datasets have 18 features with numerical and nominal values along with the class to which each instance belongs. The prepared dataset had missing values and these missing values were handled using the mean strategy. After preprocessing, we have prepared two datasets, a binary class, and a multiclass that has five-classes. The total dataset size of the five-class dataset is 1718. The binary class dataset is created by converting the five-class dataset as notckd and ckd. All classes except notckd converted to ckd in the binary class dataset. After making the binary class, it was an imbalanced dataset and we applied the data resampling technique to balance the dataset.

Three machine-learning models and two feature selection methods were used in this study. The model evaluation task is done using 10-fold cross-validation and other performance evaluation metrics such as precision, recall, F1_score, sensitivity, and specificity. The confusion matrix is used to show the correctly and incorrectly classified classes to evaluate the performance of the model. Initially, we applied the machine learning classifiers without feature selection on the original dataset for both binary class and five-class. The machine models used in this study are Random Forest, Support Vector Machine, and Decision tree. Then the feature selection techniques are applied along with the models in order to select predictive features for chronic kidney disease prediction. RFECV and UFS are applied to select the relevant features. UFS is the filtering method, it works independently of the machine learning model, and RFECV is the wrapper method that works depending on the machine-learning algorithm. Besides RFECV is an automatic feature selection method, can select the features without telling the number of features to select and the number of selected features is different from model to model. Using UFS 14 features for both binary class and five-class are selected and RF with RFECV select 8 for binary class and 9 for five class, SVM with RFECV select 9 for binary class and 10 for five-class and DT with RFECV select 16 and 7 for binary and five-class respectively. Here in this feature selection, there are features frequently selected in each feature sets, which indicates that they are the most predictive features and have strong predictive relation to the class. The most frequently selected features using both feature selection methods in all models are serum creatinine (Scr), blood urine Nitrogen (Bun), Hemoglobin (Hgb), and Specific Gravity (Sg). Additionally, the features such as Pltc, Rbcc, Wbcc, Mcv, Dm, and Htn are the most frequently selected next to the first-mentioned four features. Appendix B.1 and B.2 show the

selected features using UFS and RFECV for the Binary class and five class dataset respectively from the total 18 features.

Several related studies have done using machine learning and another algorithm in order to predict chronic kidney disease globally. Locally the researcher has found one paper conducted on chronic kidney disease prediction using KBS. For binary class classification, different studies done globally with different classifiers on the same data result in variable accuracy. However, one paper was found for multiclass classification to compare the result of the proposed models. However, the size of the dataset used and models are different from the dataset size and models of the proposed study.

Though the researcher couldn't find related papers conducted on chronic kidney disease prediction in Ethiopia, a number of papers have conducted in the medical area by Solomon[74], Bezahegn[75], Muluneh[76], and only one paper done by S. Mohammed and T. Beshah [10] on chronic kidney disease prediction using KBS. S. Mohammed and T. Beshah conducted on chronic kidney disease prediction using KBS and decision tree modeling approach for only the first three stages. They have used 20 cases for each stage and record 91% accuracy. Most papers conducted on chronic kidney disease focus on identifying only for the absence or presence of the disease, but not consider the prediction of the severity of the disease.

A. Salekin and J. Stankovic [34] proposed the method of detecting chronic kidney disease using K-NN, RF and NN, analysed the characteristics of 24 features, and sorted their predictability. The researchers have used the dataset with 400 instances, which contain 250 ckd and 150 notckd. The researcher evaluated the performance of the model using 10-fold cross-validation. The researchers used five features for final model construction. Compared to the proposed solution in this study, this paper used small size of dataset and only focus on predicting ckd and notckd. Even though dataset used in the previous work and dataset used in proposed study are not the same, the proposed model in this study has higher performance compared to the previous work.

In the study conducted by M. Almasoud and T. E. Ward [59] proposed Logistic regression, support vector machines, random forest, and gradient boosting algorithms and feature selection methods such as the ANOVA test, the Pearson's correlation, and the Cramer's V test to detect chronic kidney disease. A small dataset with small imbalance issue with 400

instances was used in the paper. Compared to the proposed model performance of binary class classification, the performance of the model in the paper is less than the proposed model.

The kidney disease stages prediction models developed by E. A. Rady and A. S. Anwar [60] using PNN, MLP, SVM and RBF. They evaluated and compared the performance of the models with accuracy. They reported accuracy accuracy result of 96.7% for PNN, 87% for RBF, 60.7% for SVM, 51.5% for MLP. They used real-world dataset consisted of 361 instances with 25 features including the target class. The total data used in the paper is the data of patient only with ckd and then they calculated the eGFR to identify stages of the disease. PNN with an accuracy of 96.7% and RBF with an accuracy of 87 % developed by E. A. Rady and A. S. Anwar [60] shows better performance than the model with best performance in this study which is RF based on RFECV with accuracy of 79% even though models in [60] and two models build in this study are different. However, the SVM model built in this study perform better than the SVM model built by E. A. Rady and A. S. Anwar [60]

The proposed model of this study was evaluated using 10-fold cross-validation and other performance evaluation metrics such as precision, recall, f1_scor, sensitivity, and specificity. In this study, different methodologies have conducted in order to do an experiment and analysis the result. Three machine-learning models, RF, SVM, and DT have implemented before applying feature selection, and RF records a high accuracy of 99.7% of binary class and 78.3% accuracy of five-class. SVM record accuracy of 96.7% for binary class and accuracy of 63% for five-class and DT record accuracy of 98.5% for binary class and accuracy of 77.5% for five-class. Then the two feature selection methods were applied to the two datasets. RF with RFECV record the highest accuracy for the two datasets 99.8% and 79% of the rest models with 8 features and 9 features for binary class and five-class respectively. The model with high accuracy and the least relevant predictive features support the expert to identify the disease fast and accurately and reduce the economic burden of patients. Thus, RF with RFECV is recommended in our study based on its accuracy, f1_score, and other performance evaluation for both binary class and five-class datasets. Table 6.16 shows the comparison of the proposed work and previously done works on chronic kidney disease prediction.

Table 6. 16 Comparison of Proposed Model and Previous Works

Reference	**Methodology**		**Target**	**Model with best Accuracy result (%)**
	Models	**Feature selection**	**Severity prediction**	
[34]	KNN, RF, ANN	Wrapper approach	No	RF with 99.8 F-measure
[59]	LG, SVM, RF, and GB	pearson correlation, ANOVA, Cramer's V test	No	Gradient Boosting (GB) with 99.1%
[41]	NB,KNN,SVM,DT,ANN	No feature selection	No	NB with 94.6%
[42]	RF, ANN	Chi- Square Test There	No	RF with 97.12%
[60]	PNN, MLP, SVM, RFB	No feature selection	Yes	PNN with 96.7
Our approach	**RF, SVM, DT**	**ANOVA , RFECV**	**Yes**	**RF+RFECV with 99.8% for binary class and 79% for multiclass.**

This study used original dataset with 18 features and one target class as clinical predictors. The final model recommended in this study is the model after feature selection method with 9 features with target class ckd_status as clinical predictors to predict the severity of the disease and developed a web app prototype. The disease severity can be quickly predicted to assist the physician in making decisions and provide patients with further appropriate examination and treatment. However, this study also has its own limitations.

6.5.1 User Interface

The user interface is the means of communication that enables the end-user (experts) to easily access the system with a prepared prototype using HTML. The server either initialized from the command prompt or using PyCharm IDE. After initializing the server it gives the HTTP address which contains the Html page from which the end-user insert the value of the feature and see the result as shown in fig 6.19.

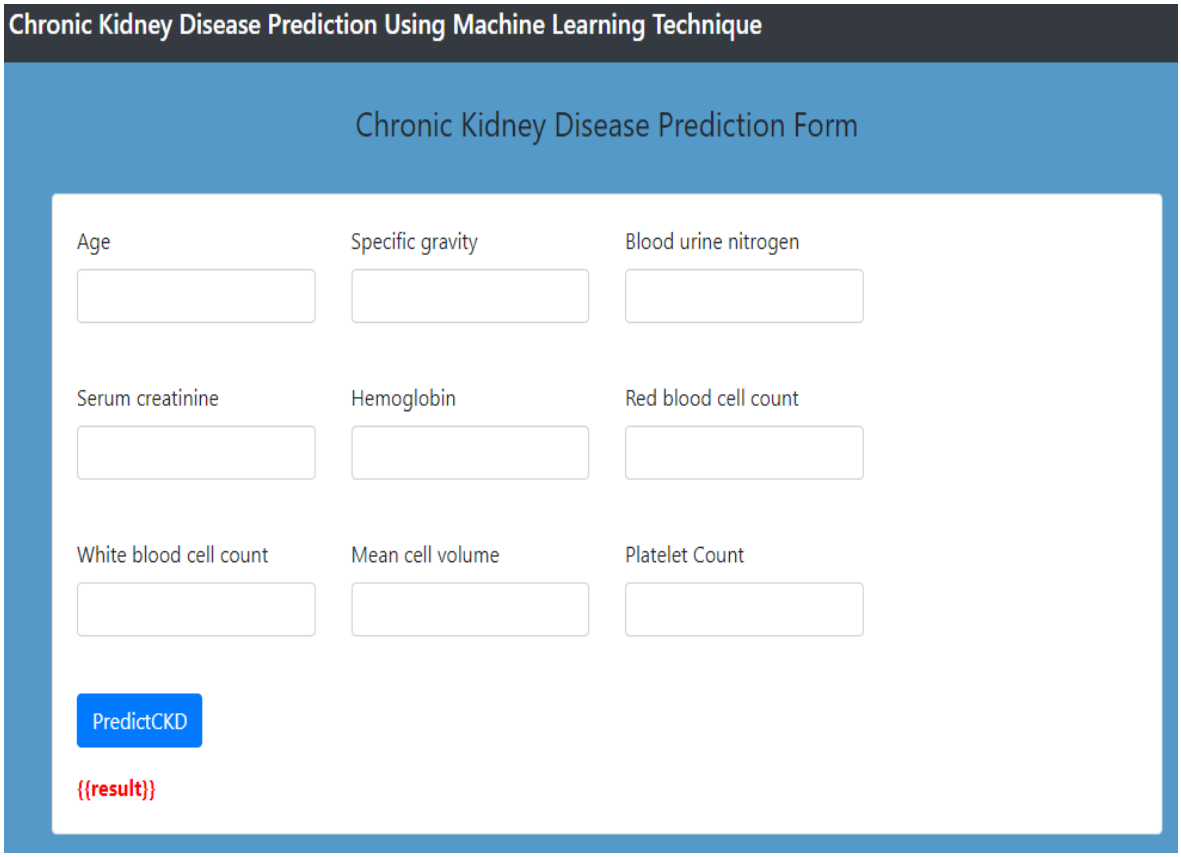

Figure 6. 19 GUI for Data Input

CHAPTER 7

CONCLUSION RECOMMENDATION AND FUTURE WORK

In this section, the proposed chronic kidney disease prediction system using machine-learning techniques concluded, recommendation, and future research directions have been described.

7.1 Conclusion

Early prediction is very crucial for both the experts and the patients to prevent and slow down the progress of chronic kidney disease to kidney failure. In this study, the chronic kidney disease prediction using the machine-learning technique is proposed with the deployment of the model in order to help the experts to diagnosis the disease fast. The proposed study also employs the feature selection method in order to select the most relevant and predictive features.

Preprocess action was taken in this study to make the data clean and appropriate for the machine learning models using machine learning and data science tool that is python. In this study, the missing values have handled using the mean replacement method and categorical values have been converted to binary except the target of the five-class dataset, which was assigned number using the replace method. The target of a five-class dataset, ckd_status replaced with 0, 2,3,4,5 for notckd, mild, moderate, severe and ESRD respectively. The binary class dataset was built converting the five class instances belong to four stages to ckd and notckd as it is. Therefore, there was a class imbalance while converting the five classes to binary classes. The resample technique was used to balance the dataset for binary class and the technique did not apply for the five-class dataset because the instant of class distribution is close to each other. The cleaned total dataset size for the five-class dataset is 1718 and the balanced and cleaned dataset for the binary class is 2884. The model is developed on the two datasets with and without applying feature selection methods. Three machine-learning models RF, SV, DT, and two feature selection methods RFECV and UFS were used to build the proposed model. The evaluation of the model was done using 10-fold cross-validation. First, the three machine learning algorithms applied to original datasets

with all 18 features. Applying the models on the original dataset, we have got the highest accuracy with RF scored accuracy of 99.7% for the binary class dataset and 78.3% for five-class dataset of the rest models. SVM and DT produced low performance compared to RF, but SVM scored the least accuracy than the rest models. The RF also produced the highest f1_score values than the rest models. Then the two feature selection methods with the models are applied on both datasets. RF with RFECV produced the highest accuracy of 99.8% and 79% for binary class and five-class datasets respectively and F1_score of 99.8% and 77.9% for binary class dataset and five-class dataset respectively. Finally, the model RF with RFECV has better performance than SVM and DT for both datasets and it was used to deploy on the server used for this study purpose, which is the flask server. Using five classes is very important to know the stages of the disease and suggest needed treatments for the patients in order to save their lives.

7.2 Recommendation

These days chronic kidney disease has become a challenge worldwide and many people die from this disease in Ethiopia even without knowing that they had chronic kidney disease before. Thus, the application of machine learning techniques in clinical data of chronic kidney disease is an important research area to improve the services provided as well as to come up with solutions to prevent and slow down the progress of this disease. In Ethiopia, there is no integration of health care system and technology, which leave a huge burden for the health experts and difficulties for the researchers to do their research, which needs more improvement in the health center for the experts to give appropriate service for the patients and for the researchers to do their work properly and efficiently.

Based on the challenges and finding of this study the researcher recommend the tasks need to be improved for future works especially for machine learning and deep learning related medical research work. Finding organized data is difficult, the data was transformed from the paper-based file to electronic format, and it was a boring and time-consuming task. Thus, the health care organization should try to store patient history data in a computer system that enables the experts and concerned body to retrieve the data easily when it is needed. Though the dataset used for this study purpose is big in size when compared to real world chronic kidney disease dataset, it is better to implement the models using a dataset with more instances from various sources. Even if promising accuracy is obtained in this study, there

might be a probability to obtain more accurate and better performing results using other models for prediction techniques, especially for severity prediction. Thus, it is recommended that these models should be applied and proved to this data. Even if the researcher believes the most predictive attributes are included in this research work, there are several features that are important in the identification of chronic kidney disease but excluded in this study and recommended to be included for further work.

7.3 Future Works

This study used a supervised machine-learning algorithm, feature selection methods to select the best subset features to develop the models. It is better to see the difference in performance results using unsupervised or deep learning algorithms models. The proposed model supports the experts to give the fast decision, it is better to make it a mobile-based system that enables the experts to follow the status of the patients and help the patients to use the system to know their status. It is better to work with real-time dataset to achieve higher accuracy.

REFERENCES

[1] C. George, A. Mogueo, I. Okpechi, J. B. Echouffo-Tcheugui, and A. P. Kengne, "Chronic kidney disease in low-income to middle-income countries: The case f increased screening," *BMJ Glob. Heal.*, vol. 2, no. 2, pp. 1–10, 2017.

[2] J. Radhakrishnan and S. Mohan, "KI Reports and World Kidney Day," *Kidney Int. Reports*, vol. 2, no. 2, pp. 125–126, 2017.

[3] "Kidney Health for Everyone Everywhere." [Online]. Available: https://www.worldkidneyday.org/2019-campaign/2019-wkd-theme/. [Accessed: 07-Feb-2020].

[4] "ETHIOPIA: KIDNEY DISEASE." [Online]. Available: https://www.worldlifeexpectancy.com/ethiopia-kidney-disease. [Accessed: 07-Feb-2020].

[5] J. W. Stanifer *et al.*, "The epidemiology of chronic kidney disease in sub-Saharan Africa: A systematic review and meta-analysis," *Lancet Glob. Heal.*, vol. 2, no. 3, pp. e174–e181, 2014.

[6] S. Abd Elhafeez, D. Bolignano, G. D'Arrigo, E. Dounousi, G. Tripepi, and C. Zoccali, "Prevalence and burden of chronic kidney disease among the general population and high-risk groups in Africa: A systematic review," *BMJ Open*, vol. 8, no. 1, 2018.

[7] M. D. Molla *et al.*, "Assessment of serum electrolytes and kidney function test for screening of chronic kidney disease among Ethiopian Public Health Institute staff members, Addis Ababa, Ethiopia," *BMC Nephrol.*, vol. 21, no. 1, p. 494, 2020.

[8] A. Agrawal, H. Agrawal, S. Mittal, and M. Sharma, "Disease Prediction Using Machine Learning," *SSRN Electron. J.*, no. May, pp. 6937–6938, 2018.

[9] T. T. Mir Mohammad Jaber, Tahmid Imam Raad, "Comparison of Machine Learning Techniques to Predict Cardiovascular Disease," no. December, 2018.

[10] S. Mohammed and T. Beshah, "Amharic based knowledge-based system for diagnosis and treatment of chronic kidney disease using machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 11, pp. 252–260, 2018.

[11] "ETHIOPIA STEPS REPORT ON RISK FACTORS FOR NON-COMMUNICABLE DISEAES AND PREVALENCE OF SELECTED NCDs ETHIOPIA STEPS REPORT ON RISK FACTORS FOR CHRONIC NON- COMMUNICABLE DISEASES AND," no. December, 2016.

[12] E. Public and H. Problems, “Emerging Public Health Problems in Ethiopia : Chronic Non-Communicable Diseases.”

[13] K. T. Mills *et al.*, “A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010,” *Kidney Int.*, vol. 88, no. 5, pp. 950–957, 2015.

[14] P. Tuso, “SERVE Ethiopia,” *Perm. J.*, vol. 13, no. 3, pp. 61–64, 2009.

[15] M. American, “National Chronic Kidney Fact Sheet: general information and national estimates on chronic kidney disease in the United States,” *Centers Dis. Control Prev.*, no. Cvd, 2010.

[16] N. Radha and S. Ramya, “Performance Analysis of Machine Learning Algorithms for Predicting Chronic Kidney Disease,” no. 8, pp. 72–76, 2015.

[17] “Chronic Kidney Disease.” [Online]. Available: https://www.worldkidneyday.org/facts/chronic-kidney-disease/. [Accessed: 07-Feb-2020].

[18] V. Chaurasia, S. Pal, and B. B. Tiwari, “Chronic kidney disease: A predictive model using decision tree,” *Int. J. Eng. Res. Technol.*, vol. 11, no. 11, pp. 1781–1794, 2018.

[19] R. P.R. and M. M., “Risk factors associated with chronic kidney disease: An overview,” *Int. J. Pharm. Sci. Rev. Res.*, vol. 40, no. 2, pp. 255–257, 2016.

[20] J. Duan *et al.*, “Prevalence and risk factors of chronic kidney disease and diabetic kidney disease in Chinese rural residents: a cross-sectional survey,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, 2019.

[21] K. K. Goro *et al.*, “Patient Awareness , Prevalence , and Risk Factors of Chronic Kidney Disease among Diabetes Mellitus and Hypertensive Patients at Jimma University Medical Center , Ethiopia,” vol. 2019, 2019.

[22] M. Zeleke, “The Magnitude of Chronic Renal Failure and Its Associated factors among patients at St. Paulo’s Hospital, Addis Ababa, Ethiopia,” no. August, 2016.

[23] D. C. Crews, A. K. Bello, and G. Saadi, “Burden, Access, and Disparities in Kidney Disease,” *Can. J. Kidney Heal. Dis.*, vol. 6, pp. 126–133, 2019.

[24] T. Anothaisintawee, S. Rattanasiri, A. Ingsathit, J. Attia, and A. Thakkinstian, “Prevalence of chronic kidney disease: A systematic review and meta-analysis,” *Clin. Nephrol.*, vol. 71, no. 3, pp. 244–254, 2009.

[25] L. Zhang *et al.*, “Prevalence of chronic kidney disease in China: A cross-sectional survey,” *Lancet*, vol. 379, no. 9818, pp. 815–822, 2012.

[26] N. P. Marni Armstrong, Robert Weaver, “Prevalence and Quality of Care in Chronic Kidney Disease,” no. February, 2019.

[27] T. Fiseha, “Prevalence of Chronic Kidney Disease and Associated Risk Factors among Diabetic Patients in Southern Ethiopia,” *Am. J. Heal. Res.*, vol. 2, no. 4, p. 216, 2014.

[28] M. Hussein, “Survival Analysis of Patients with End Stage Renal Disease the Case of Adama Hospital, Ethiopia,” *Clin. Med. Res.*, vol. 6, no. 6, p. 201, 2017.

[29] N. K. Foundation, “The kidneys and kidney disease,” pp. 1–3, 2010.

[30] N. N. Shah, S. Nagpal, and P. Y. Zowj, “Chronic Kidney Disease Prediction.”

[31] B. Kefale, “Assessment of management and quality of life among patients with chronic kidney disease at Tikur Anbessa Specialized A Thesis Submitted to the Department of Pharmacology and Clinical Pharmacy , School of Pharmacy , College of Health Sciences , Addis Ababa,” no. January, 2018.

[32] F. Rabbi, “A review of the recent trends in the use of machine learning in business,” vol. 1, no. 1, pp. 1–6, 2019.

[33] A. J. Frandsen, “Machine Learning for Disease Prediction,” *Brigham Young Univ. Sch.*, p. Paper 5975, 2016.

[34] A. Salekin and J. Stankovic, “Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes,” *Proc. - 2016 IEEE Int. Conf. Healthc. Informatics, ICHI 2016*, pp. 262–270, 2016.

[35] M. Kumar and M. Kumar, “International Journal of Computer Science and Mobile Computing Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm,” *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, no. 2, pp. 24–33, 2016.

[36] S. Drall, G. S. Drall, and S. Singh, “Chronic Kidney Disease Prediction Using Machine Learning : A New Approach Bharat Bhushan Naib,” vol. 8, no. 278, pp. 278–287.

[37] F. Aqlan, R. Markle, and A. Shamsan, “Data mining for chronic kidney disease prediction,” *67th Annu. Conf. Expo Inst. Ind. Eng. 2017*, no. March, pp. 1789–1794, 2017.

[38] A. Charleonnan *et al.*, “Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques,” pp. 80–83, 2016.

[39] S. Tekale, P. Shingavi, S. Wandhekar, and A. Chatorikar, “Prediction of Chronic

Kidney Disease Using Machine Learning Algorithm," vol. 7, no. 10, pp. 92–96, 2018.

[40] S. Zeynu and S. Patil, "Journal of Data Mining in Genomics Prediction of Chronic Kidney Disease Using Data Mining Feature Selection and Ensemble Method," vol. 9, no. 1, pp. 1–9, 2018.

[41] A. B. K. A, K. Priyanka, R. B. T. M, and B. E. C. Science, "CHRONIC KIDNEY DISEASE PREDICTION BASED ON NAIVE BAYES TECHNIQUE," pp. 1653–1659, 2019.

[42] S. Y. Yashfi, "Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms," 2020.

[43] A. Subas, E. Alickovic, and J. Kevric, "Diagnosis of chronic kidney disease by using random forest," *IFMBE Proc.*, vol. 62, no. 1, pp. 589–594, 2017.

[44] S. Ramya and N. Radha, "Diagnosis of Chronic Kidney Disease Using," pp. 812–820, 2016.

[45] V. S and D. S, "Data Mining Classification Algorithms for Kidney Disease Prediction," *Int. J. Cybern. Informatics*, vol. 4, no. 4, pp. 13–25, 2015.

[46] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, 2017.

[47] R. Kadam Vinay, K. L. S. Soujanya, and P. Singh, "Disease prediction by using deep learning based on patient treatment history," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 745–754, 2019.

[48] T. Zar Phyu and N. N. Oo, "Performance comparison of feature selection methods," *MATEC Web Conf.*, vol. 42, pp. 2–5, 2016.

[49] S. Vanaja, "Analysis of Feature Selection Algorithms on Classification : A Survey," vol. 96, no. 17, pp. 28–35, 2014.

[50] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6371–6385, 2014.

[51] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.*, pp. 1200–1205, 2015.

[52] S. Beniwal and J. Arora, "Classification and feature selection techniques in data mining," no. August 2012, 2014.

[53] G. H. John, R. Kohavi, and K. Pfleger, “Irrelevant Features and the Subset Selection Problem,” *Mach. Learn. Proc. 1994*, pp. 121–129, 1994.

[54] V. Kumar, “Feature Selection: A literature Review,” *Smart Comput. Rev.*, vol. 4, no. 3, 2014.

[55] Y. Saeys, P. Larra, P. Systems, and B.- Ghent, “A review of feature selection techniques in bioinformatics,” no. November 2007, 2015.

[56] F. Aqlan, R. Markle, and A. Shamsan, “Data mining for chronic kidney disease prediction,” *67th Annu. Conf. Expo Inst. Ind. Eng. 2017*, no. May, pp. 1789–1794, 2017.

[57] J. Xiao *et al.*, “Comparison and development of machine learning tools in the prediction of chronic kidney disease progression,” *J. Transl. Med.*, vol. 17, no. 1, pp. 1–13, 2019.

[58] A. Subas, E. Alickovic, and J. Kevric, “Diagnosis of chronic kidney disease by using random forest,” *IFMBE Proc.*, vol. 62, no. 3, pp. 589–594, 2017.

[59] M. Almasoud and T. E. Ward, “Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors,” vol. 10, no. 8, pp. 89–96, 2019.

[60] E. A. Rady and A. S. Anwar, “Informatics in Medicine Unlocked Prediction of kidney disease stages using data mining algorithms,” *Informatics Med. Unlocked*, vol. 15, no. December 2018, p. 100178, 2019.

[61] A. Jasim and M. Kaky, “INTELLIGENT SYSTEMS APPROACH FOR CLASSIFICATION AND MANAGEMENT OF by,” no. July, 2017.

[62] M. Saar-tsechansky and F. Provost, “Handling Missing Values when Applying Classi … cation Models,” vol. 1, 2007.

[63] “Data Preparation for Statistical Modeling and Machine Learning.” [Online]. Available: https://www.featureranking.com/tutorials/machine-learning-tutorials/data-preparation-for-machine-learning/. [Accessed: 12-Oct-2020].

[64] Oliver Theobald, *Machine Learning For Absolute Beginners*. 2017.

[65] S. Koshy, “Feature Selection for Improving Multi-Label Classification using MEKA,” vol. 12, no. 24, pp. 14774–14782, 2017.

[66] analytics vidhya, “Introduction to Feature Selection methods with an example (or how to select the right variables?).” [Online]. Available: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-

methods-with-an-example-or-how-to-select-the-right-variables/. [Accessed: 24-Mar-2020].

[67] P. Misra and A. S. Yadav, “Improving the classification accuracy using recursive feature elimination with cross-validation,” *Int. J. Emerg. Technol.*, vol. 11, no. 3, pp. 659–665, 2020.

[68] S. Kapoor, R. Verma, and S. N. Panda, “Detecting kidney disease using Naïve bayes and decision tree in machine learning,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 1, pp. 498–501, 2019.

[69] M. S. D. Dr. S. Vijayarani1, “Liver Disease Prediction using SVM and Naïve Bayes Algorithms,” *Int. J. Sci. Eng. Technol. Res.*, vol. 4, no. 4, pp. 816–820, 2015.

[70] J. Dantas, “The importance of k-fold cross-validation for model prediction in machine learning.” [Online]. Available: https://towardsdatascience.com/the-importance-of-k-fold-cross-validation-for-model-prediction-in-machine-learning-4709d3fed2ef.

[71] A. Acharya, “Comparative Study of Machine Learning Algo- rithms for Heart Disease Prediction,” no. April, 2017.

[72] Y. Amirgaliyev, “Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods,” *2018 IEEE 12th Int. Conf. Appl. Inf. Commun. Technol.*, pp. 1–4, 2010.

[73] P. Sinha, “Performance evaluation of Classification Techniques on Prediction of Chronic Kidney Disease,” pp. 1–6.

[74] Solomon Gebremariam, “School of Graduate Studies School of Information Science a Self-Learning Knowledge Based System for School of Graduate Studies,” *Aau*, no. January, 2013.

[75] B. ZERIHUN, “DEVELOPING A PREDICTIVE MODEL FOR PRE- DIABETES SCREENING BY USING DATA MINING TECHNOLOGY A,” 2017.

[76] M. Endalew, “EXPLORING THE PREVALENCE OF DIARRHEAL DISEASE USING DATA MINING TECHNOLOGY (A CASE OF TIKUR ANBESSA HOSPITAL),” pp. 10–14, 2011.

APPENDICES

Appendix A: Dataset Features Description

- **Age** is the age of the patient in years. Age is one factor of chronic kidney disease. The prevalence and severity of CKD is higher among patients >60 years old.
- **Gender** is the gender of the patient (Male or Female).
- **Bp** is the blood pressure in mm/Hg. High blood pressure can damage blood vessels in the kidneys, reducing their ability to work properly.
- **Sg** is the specific gravity of the patient (nominal).
- **Chl** is chloride (mmol/L), which is the CKD predictive attribute.
- **Sod** is sodium (mmol/L), which is essential for the human body. A person with CKD cannot eliminate excess sodium and fluid from his or her body. High sodium increases blood pressure.
- **Pot** is potassium (mmol/L), which is important for the human body. A person with a damaged kidney cannot eliminate excess potassium and fluid from his or her body.
- **Bun** Blood Urine Nitrogen (mg/dL) is the CKD predictive attribute.
- **Scr** serum creatinine (mg/dL) is a waste product that comes from muscle activity. Damaged kidneys cannot remove creatinine from the blood as much as the proper kidney. Serum creatinine is the most predictive parameter for CKD detection.
- **Hgb** hemoglobin (gm/dl) is important CKD predictive parameter.
- **Rbcc** red blood cell count (m/cumm).
- **Wbcc** red blood cell count (k/cumm).
- **Htn** hypertension (yes or no) is the factor that increases the progress of CKD.
- **Dm** diabetic mellitus (yes or no) is a good predictive attribute of CKD.
- **Ane** anemia (yes or no) is the factor of CKD.
- **Mcv** Mean cell volume (fl).
- **Pltc** platelet count (k/cumm).
- **Hd** heart disease (yes or no) is the factor of CKD.
- **Ckd_status** the class assigned to the patient data history. The ckd_status for five class dataset are notckd, mild (for stage1 and stage2), moderate (for stage3), severe (for Stage4), ESRD (for stage5) and ckd_status for binary dataset are ckd and notckd.

Appendix A.1: CKD-EPI Creatinine Equation

eGFR =141xmin(SCr/κ,1)α x max(SCr /κ, 1)-1.209 x0.993Age x1.018 [if female] x1.159 [if Black].

Where:

eGFR (estimated glomerular filtration rate) = mL/min/1.73 m^2

S_{Cr} (standardized serum creatinine) = mg/dL

κ = 0.7 (females) or 0.9 (males)

α = -0.329 (females) or -0.411 (males)

min = indicates the minimum of S_{Cr}/κ or 1

max = indicates the maximum of S_{Cr}/κ or 1

Age = years

Appendix B: Selected features using UFS and RFECV

Table B. 1 Selected Features Using UFS for Binary Dataset and Five-Class Data Set

Selected features using UFS for binary dataset	**Selected features using UFS for five-class dataset**
Hgb, Scr ,Bun, Rbcc, Htn, Sg, Sod , Dm, Pltc, Ane, Bp, Pot, Hd, Mcv	Scr, Bun, Hgb, Rbcc, Ane, Sg, Htn, Sod, Pltc, Dm, Hd, Bp, Pot, Age,

Table B. 2 Selected Features Using RFECV for Binary Dataset and Five-Class Data Set with Respective Models

Models	**Selected features using RFECV for binary dataset**	**Selected feature using RFECV for five class dataset**
RF	'Bp', 'Sg', 'Bun', 'Scr', 'Hgb', 'Rbcc', 'Mcv', 'Htn'	'Age', 'Sg', 'Bun', 'Scr', 'Hgb', 'Rbcc', 'Wbcc', 'Mcv', 'Pltc'
SVM	'Gender', 'Pot', 'Scr', 'Hgb', 'Rbcc', 'Htn', 'Dm', 'Ane', 'Hd'	'Gender', 'Sg', 'Pot', 'Scr', 'Hgb', 'Rbcc', 'Htn', 'Dm', 'Ane', 'Hd'
DT	'Bp', 'Sg', 'Chl', 'Sod', 'Pot', 'Bun', 'Scr', 'Hgb', 'Rbcc', 'Wbcc', 'Mcv', 'Pltc', 'Htn', 'Dm', 'Ane', 'Hd'	'Age', 'Gender', 'Sg', 'Scr', 'Hgb', 'Wbcc', 'Htn'

Appendix C: Sample Code

Appendix C.1 Sample Code for Preprocessing

#Preprocessing

```
from sklearn.preprocessing import LabelEncoder,OneHotEncoder

#Handling Categorical data

X = pd.get_dummies(X, prefix_sep='_')

Y = LabelEncoder().fit_transform(Y)

#Filling missing values

from sklearn.impute import SimpleImputer

X = SimpleImputer(missing_values= np.NAN, strategy= 'mean', fill_value=None, verbose=0, copy=True)

X =X.fit(X)
```

Appendix C.2 Sample Source Code for Building and Evaluating Models

 # RF Model

```
from sklearn.ensemble import RandomForestClassifier

RFmodel = RandomForestClassifier(n_estimators=100,
random_state=0,max_features='auto',class_weight=None)

CVscoreRF = cross_val_score(RFmodel,X,y,cv=10)

print(CVscoreRF)

print("Average accuracy of Random Forest:",round(CVscoreRF.mean() * 100,2))

precision = cross_val_score(RFmodel,X,y,cv =10,scoring ='precision_macro')

print('Average Precision of Random Forest: %0.2f' %(precision.mean()))

recall = cross_val_score(RFmodel,X,y,cv =10,scoring ='recall_macro')

print('Average Recall of Random Forest: %0.2f' %(recall.mean()))

f_score = cross_val_score(RFmodel,X,y,cv =10,scoring ='f1_macro')

print('Average F_score of Random Forest: %0.2f' %(f_score.mean()))

y_predRF = cross_val_predict(RFmodel,X,y,cv=10)

confusion_matrixRF = pd.crosstab(y, y_predRF, rownames=['Actual ckd_status'],
colnames=['Predicted ckd_status'])

print(confusion_matrixRF)
```

#SVM Model

```
from sklearn.svm import LinearSVC

SVM_model = LinearSVC(C=1, class_weight=None,multi_class='ovr', random_state=0)

SVMaccuracy = cross_val_score(SVM_model, X, y, scoring='accuracy', cv = 10)
```

```
print(SVMaccuracy)

#get the mean of each fold

print("Accuracy of Suppoert Vector Machine Model with Cross Validation is:",round(SVMaccuracy.mean() * 100,2))

precision = cross_val_score(SVM_model,X,y,cv =10,scoring ='precision_macro')

print('Average Precision of Support Vector Machine: %0.2f' %(precision.mean()))

recall = cross_val_score(SVM_model,X,y,cv =10,scoring ='recall_macro')

print('Average Recall of Support vector Machine: %0.2f' %(recall.mean()))

f_score = cross_val_score(SVM_model,X,y,cv =10,scoring ='f1_macro')

print('Average F1_score of Support vector Machine: %0.2f' %(f_score.mean()))

y_predSVM = cross_val_predict(SVM_model,X,y,cv=10)

confusion_matrixSVM = pd.crosstab(y, y_predSVM, rownames=['Actual ckd_status'], colnames=['Predicted ckd_status'])

print(confusion_matrixSVM)
```

 # DT Model

```
DTmodel = DecisionTreeClassifier (criterion='gini', max_depth=4,random_state =0)

CVscoreDT = cross_val_score(DTmodel,X,y,cv=10)

print(CVscoreDT)

print("Average accuracy of Decision Tree:",round(CVscoreDT.mean() * 100,2))

precision = cross_val_score(DTmodel,X,y,cv =10,scoring ='precision_macro')

print('Average Precision of Decision Tree: %0.2f' %(precision.mean()))

recall = cross_val_score(DTmodel,X,y,cv =10,scoring ='recall_macro')
```

```
print('Average Recall of Decision Tree: %0.2f' %(recall.mean()))

f_score = cross_val_score(DTmodel,X,y,cv =10,scoring ='f1_macro')

print('Average F1_score of Decision Tree: %0.2f' %(f_score.mean()))

y_predDT = cross_val_predict(DTmodel,X,y,cv=10)

confusion_matrixDT = pd.crosstab(y, y_predDT, rownames=['Actual ckd_status'], colnames=['Predicted ckd_status'])

print(confusion_matrixDT)
```

Appendix C.2 Sample Code for Model Saving

```
# Saving model RF with RFECV

RFmodel = RandomForestClassifier(n_estimators=100,
random_state=0,max_features='auto',class_weight=None)

rfecv = RFECV(estimator=RFmodel,step=1,cv=10, scoring='accuracy',verbose=0)

rfecv = rfecv.fit(X,y)

print("Optimal number of features:%d" %rfecv.n_features_)

print("selected Features %s" %rfecv.support_)

print("Feature Ranking %s" %rfecv.ranking_)

X_rfecv = scaler.fit_transform(X_rfecv)

RFmodel.fit(X_rfecv,y)

pickle.dump(RFmodel, open('RF_RFECV_model.pkl', 'wb'))
```

Appendix C.3 Sample Code for Integrating ML Model with Flask Server

#chronic kidney disease prediction web app

```
import numpy as np
import pandas as pd
from flask import Flask, request, render_template
import pickle
app = Flask(__name__)
model = pickle.load(open('RF_RFECV_model.pkl', 'rb'))
@app.route('/')
def home():
    return render_template('index_multi.html')
@app.route('/predictckd', methods =['POST'])
def predictckd():
    input_features = [float(x) for x in request.form.values()]
    features_value = [np.array(input_features)]
    features_name = ['Age', 'Sg', 'Bun', 'Scr', 'Hgb', 'Rbcc', 'Wbcc', 'Mcv', 'Pltc']
    df = pd.DataFrame(features_value, columns=features_name)
    output = model.predict(df)
    if output == 0:
        res_val = "Patient has no Chronic kidney disease, the patient need to take care to stay
safe "
    elif output == 2:
        res_val = "Patient has mild stage (2) Chronic kidney disease, "
    elif output == 3:
        res_val = "Patient has moderate stage (3) Chronic kidney disease"
    elif output == 4:
        res_val = "Patient has severe stage (4) Chronic kidney disease"
    else:
        res_val = "Patient has end stage (5) Chronik kidney disease"
    return render_template('index_multi.html', prediction_value=format(res_val))
if __name__ == "__main__":
    app.run(debug=True)
```