

DATA MINING ANALYSIS OF CHRONIC KIDNEY DISEASE (CKD) LEVEL

By:

MUHAMMAD HAFIZAM AFIQ BIN MOHD HARIZI

(Matrix no. 144209)

Supervisor:

Associate Professor Dr. Loh Wei Ping

July 2022

This dissertation is submitted to

Universiti Sains Malaysia

As partial fulfilment of the requirement to graduate with honors degree in

BACHELOR OF ENGINEERING (MECHANICAL ENGINEERING)




School of Mechanical Engineering

Engineering Campus

Universiti Sains Malaysia

DECLARATION


This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed.......... (Muhammad Hafizam Afiq Bin Mohd Harizi)

Date
..... 5/8/2022 (_____)

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by giving explicit references. Bibliography/references are appended.

Signed.......... (Muhammad Hafizam Afiq Bin Mohd Harizi)

Date
..... 5/8/2022 (_____)

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for interlibrary loan, and for the title and summary to be made available outside organizations.

Signed.......... (Muhammad Hafizam Afiq Bin Mohd Harizi)

Date
..... 5/8/2022 (_____)

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my supervisor, Associate Professor Dr. Loh Wei Ping who guided me throughout the Final Year Project. She provided me with invaluable advice and assisted me during my difficult time. Her motivation and inspiration have contributed tremendously to the successful completion of this thesis. I am deeply indebted to her for her fruitful ideas towards the completion of my project efficiently. I am very proud to say that I had an opportunity to work with an exceptional lecturer like her.

Besides, I am extremely grateful to my parents for their love, prayers, care, and sacrifice in educating and preparing me to achieve my future goals. I am also thankful to my siblings for supporting me spiritually throughout my life.

Last but not least, I am very grateful to all who have given me their love and friendship in supporting me physically and mentally.

TABLE OF CONTENTS

DECLARATION	ii
ACKNOWLEDGEMENT	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	vi
LIST OF FIGURES	vii
LIST OF ABBREVIATIONS	viii
LIST OF APPENDICES	ix
ABSTRAK	x
ABSTRACT	xii
CHAPTER 1 INTRODUCTION	1
1.1 Overview	1
1.2 Study Background	1
1.3 Problem Statement	3
1.4 Objectives	3
1.5 Scope of the Project.....	3
1.6 Thesis Outline	4
CHAPTER 2 LITERATURE REVIEW	6
2.1 Overview	6
2.2 Search Method.....	6
2.3 The Method and Algorithms used in CKD Studies.....	7
2.4 Factors Attributing to CKD	13
2.5 Summary	22
CHAPTER 3 METHODOLOGY	23
3.1 Overview	23
3.2 Data Collection.....	25

3.3	Statistical Analysis	28
3.4	Data Pre-processing.....	28
3.5	Data Classification	31
3.6	Classification Model Verification	33
CHAPTER 4 RESULTS AND DISCUSSION.....		34
4.1	Overview	34
4.2	Statistical Analysis	34
4.3	Classification.....	42
4.3.1	Outliers' treatment, before and after handling missing values.....	42
4.3.2	Uncertain Class	48
4.4	Comparative Analysis	53
4.5	Summary	54
CHAPTER 5 CONCLUSION.....		56
5.1	Concluding Remark.....	56
5.2	Study Contribution	58
5.3	Future Recommendation	58
REFERENCES.....		60
APPENDICES		

LIST OF TABLES

	Page
Table 2.1	Method or algorithm used in existing CKD studies..... 7
Table 2.2	The interpretation of blood pressure measurements in adults 18 years of age and older [28]..... 14
Table 2.3	Interpretation of BUN levels [40] 17
Table 2.4	Normal wc range for the following category [49] 19
Table 2.5	The normal range of rc [50] 19
Table 3.1	The data attribute description..... 25
Table 3.2	Upper and Lower Bound values for determining the outliers..... 30
Table 4.1	Statistics of numerical scale data 34
Table 4.2	Classification accuracies for outliers' treatment, before and after handling the missing values 43
Table 4.3	Classification accuracies for the condition: (a) and (b) 50

LIST OF FIGURES

	Page
Figure 1.1	The level of albuminuria and the stages of GFR for CKD stages [3]2
Figure 2.1	Schematic Diagram of Article Search Strategy7
Figure 2.2	The number of papers using particular algorithms for CKD studies9
Figure 3.1	Flowchart of the methodology process24
Figure 3.2	Missing values replacement using WEKA software.....29
Figure 4.1	Histogram of data attributes38
Figure 4.2	Graph of classification accuracy vs classifier algorithm.....44
Figure 4.3	Tree diagram for three classification conditions: (a), (b), and (c)46
Figure 4.4	Bar chart indicating the number of instances for each labeled class of condition when the number of missing values and outliers (a) ≥ 8 and (b) ≥ 749
Figure 4.5	Tree diagram for conditions when the number of missing values and outliers (a) ≥ 8 and (b) ≥ 751
Figure 4.6	Classification accuracy for conditions when the number of missing values and outliers (a) ≥ 8 and (b) ≥ 752

LIST OF ABBREVIATIONS

CKD	Chronic kidney disease
WEKA	Waikato Environmental for Knowledge Analysis
NB	Naïve Bayes
SVM	Support Vector Machine
GFR	Glomerular filtration rate
ESKD	End-stage kidney disease
AKI	Acute kidney injury
SMO	Sequential Minimal Optimization
ANN	Artificial Neural Network
KNN	k-Nearest Neighbours
DT	Decision Tree
PNN	Probabilistic Neural Network
MLP	Multilayer Perception
RBF	Radial Basis Function
XGBoost	Extreme Gradient Boost
LR	Linear Regression
RF	Random Forest
NN	Neural Network
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DL	Deep Learning
GBoost	Gradient Boosting
BUMAKI	Baskent University Model for AKI
BUN	Blood urea nitrogen
CBC	Complete blood count

LIST OF APPENDICES

- APPENDIX A** Detail information about the dataset
- APPENDIX B** Data Pre-processing using Microsoft Excel
- APPENDIX C** WEKA interface
- APPENDIX D** WEKA classification results
- APPENDIX E** Gantt Chart

ABSTRAK

Penyakit Ginjal Kronik (CKD) adalah keadaan apabila fungsi ginjal gagal dan bertambah teruk dari semasa ke semasa. Peningkatan bilangan pesakit CKD dan kematian akibat tahap akhir CKD telah memberi suatu cabaran yang ketara secara global. Kajian-kajian lepas secara meluasnya focus pada progress klinikal dan faktor-faktor penyebab penyakit ini. Namun, analisis perlombongan data untuk menangani CKD adalah rendah. Tidak banyak kajian yang mempertimbangkan faktor-faktor penyebab untuk mendapatkan klasifikasi tepat tahap CKD. Matlamat utama kajian ini adalah untuk menemui demografi pesakit CKD, penunjuk klinikal, dan faktor risiko yang berkait dengan peringkat CKD serta untuk membangunkan model perlombongan data mengenai faktor risiko untuk peringkat perkembangan CKD. Kes kajian CKD terdiri daripada rekod data klinikal yang diekstrak daripada domain umum UCI Machine Learning Repository. Analisis perlombongan data melibatkan empat peringkat analisis: pra-pemrosesan data, klasifikasi data, klasifikasi atribut mengikut peringkat CKD, dan pengesahan model klasifikasi dengan menggunakan perisian Microsoft Excel dan Waikato Environmental for Knowledge Analysis (WEKA) versi 3.8.5. Klasifikasi data dilakukan pada mod pengesahan silang 10 ganda dengan algoritma-algoritma Naïve Bayes (NB), Mesin Vektor Sokongan (SVM) dan Pokok J48. Algoritma ZeroR ditetapkan sebagai rujukan garis dasar. Terdapat tiga peringkat analisis klasifikasi: sebelum dan selepas mengendalikan nilai yang hilang, sebelum dan selepas rawatan data terpencil, dan penambahan kelas tidak pasti. Hasil kajian menunjukkan tiada perbezaan bagi sebelum dan selepas mengendalikan nilai yang hilang dengan NB (96.0%) and SMO (98.5%) kecuali J48 yang menunjukkan sedikit peningkatan sebanyak 1% kepada (97.8%). Ketepatan klasifikasi bagi rawatan data terpencil adalah sebanyak 97.4%, 95.4%, and 97.1% bagi NB, SMO, and J48 masing-

masing. Dengan penambahan kelas tidak pasti, ketepatan terbaik yang diperoleh adalah 98.5% dengan algoritma SMO. Sebuah model ramalan klasifikasi yang menentukan ketepatan untuk tiga kelas pengelasan telah dibangun dengan menggunakan algoritma SMO.

ABSTRACT

Chronic Kidney Disease (CKD) is the state when the kidneys' functions fail and worsen over time. The rise in the number of patients with CKD and the fatal consequences of end-stage CKD poses a significant challenge globally. Studies have widely focused on the clinical progression and attributing factors of the disease. However, data mining analysis to address the CKD is low. Few studies have considered the attributing factors that accurately classify the CKD levels. The main goal of this study is to discover the CKD patient demographics, clinical indicators, and risk factors attributing to CKD stages as well as to develop a data mining model on the risk factors for CKD progression stages. The CKD case study consists of clinical data records extracted from the UCI Machine Learning Repository domain. Data mining analysis takes place in four stages: data pre-processing, data classification, classification of attributes by stages of CKD, and classification model verification using Microsoft Excel and Waikato Environmental for Knowledge Analysis (WEKA) version 3.8.5 software. Data classifications are performed at a 10-fold-cross-validation mode using Naïve Bayes (NB), Support Vector Machine (SVM), and J48 Trees. The ZeroR algorithm was set as the baseline. There are three levels of classification analyses: before and after handling the missing values, before and after the outliers' treatment, and adding uncertain classes. Findings show no difference in classification accuracies before and after treating the missing values using NB (96.0%) and SMO (98.5%) except for J48 indicating a slight improvement of 1% to (97.8%). The classification accuracies for outliers' treatment are 97.4%, 98.4%, and 97.1% for NB, SMO, and J48 respectively. Adding the uncertain class the best accuracy obtained was 98.5% using the SMO algorithm. A predictive classification model that determines the

accuracy for three classification classes was developed accordingly using the SMO algorithm.

CHAPTER 1

INTRODUCTION

1.1 Overview

This chapter presents an introduction to data mining analysis of chronic kidney diseases (CKD) research study. The chapter is structured into 6 sections to present the study background, problem statement, objectives, scope of the project, and the thesis outline. The purpose of the study is basically to discover the CKD patient demographics, clinical indicators, and risk factors attributing to CKD stages as well as to develop a data mining model on the risk factors for CKD progression stages.

1.2 Study Background

CKD is a global health problem that is steadily increasing. Two million people with kidney failure receive treatment in only five relatively wealthy countries, representing 12% of the world's population [1]. According to Rady and Anwar [1] around half of the world's population is treated in just 100 developing countries, as compared to 20% in developed countries. In developed countries, HIV and exposure to toxins or heavy metals play an additional role in CKD, whereas hypertension and diabetes are the most common causes [2].

There are five stages of CKD ranging from mild to severe levels. However, referring to Romagnani et al. [3], there are actually six stages of CKD level according to glomerular filtration rate (GFR) with the level of albuminuria and the risk of each level and stage. Figure 1.1 shows the levels of albuminuria and the stages of GFR.

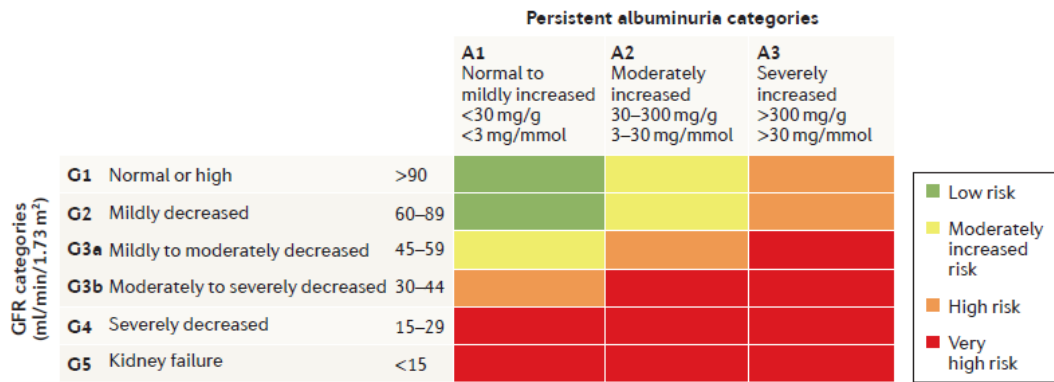


Figure 1.1 The level of albuminuria and the stages of GFR for CKD stages [3]

The functional status of the kidneys can be observed in Figure 1.1 categorized from low risk to very high risk.

Individuals with CKDs are often devalued and ignored as there are no symptoms at the beginning, leading to the lack of symptoms. The CKD stage is irreversible. As a result of CKD progression towards more severe stages, kidney failure develops as a pathological condition called end-stage kidney disease (ESKD) [4]. If the kidney function is impaired, patients would soon develop uremia or even die from the consequences until they seek for kidney transplant or dialysis.

Many related studies considered different contributing factors to CKD. As in Karakaya et al. [5], the factors that contribute to acute kidney injury (AKI) were studied in patients with serious burns. While Jayasekara et al. [6] found that heat stress and dehydration are risk factors in this form of CKD. An effective and timely diagnosis of CKD patients allows for more personalized care and better treatment planning [7].

In addition, a large database of chronic disease patients can be used to predict kidney failure without the need for a prior diagnosis of CKD [4]. Although many problems associated with CKD have been solved by previous studies using experimental or analytical methods. In terms of data mining analysis, there are few studies that diagnose CKD and predict the stage of CKD. Thus, this study aims to

discover the CKD patient demographics, clinical indicators, and risk factors attributing to CKD stages as well as to develop a data mining model on the risk factors for CKD progression stages

1.3 Problem Statement

CKD is the state when the kidneys' functions worsen over time. Although CKD has been medically solved and divided into five stages, there has not been an accurate method to determine the risk factors for CKD progression stages. Few studies have considered data mining analysis to diagnose CKD among the public and to predict the stage of CKD. Therefore, it is essential to know the risk factors for CKD progression stages utilizing a data mining approach. No study has developed a predictive model that determines the accuracy for three classification classes.

1.4 Objectives

The aims of the study are to

- i. discover the CKD patient demographics, clinical indicators, and risk factors.
- ii. determine the attributes that classify the stages of CKD.
- iii. develop a data mining model on the risk factors for CKD progression stages.

1.5 Scope of the Project

This study considers the CKD analysis involving a dataset from CKD and non-CKD patients. The attributes considered are age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cells, pus cells clumps, bacteria, blood glucose

random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia, and class. The attributes were divided into two data scales: 11 numerical and 14 nominals. Data is analyzed using a data mining approach. in five stages of analysis: data collection, data pre-processing, data classification, classification of attributes by stages of CKD, and model classification verification. Classification algorithms used include Naïve Bayes (NB), Sequential Minimal Optimization (SMO), and J48 methods aided by Waikato Environmental for Knowledge Analysis (WEKA) version 3.8.5 tool. Data is trained at 10-fold cross-validation for the classification analyses.

1.6 Thesis Outline

This thesis is structured into five chapters. Chapter 1 presents an overview of the CKD, study background, problem statement, objectives, and scope of the project. Chapter 2 discusses the state-of-the-art review emphasizing the methods used by the previous studies related to CKD problems and the factors attributed to CKD. The approaches used in the past, the advantages and disadvantages of the methods, limitations, and the issues or challenges also will discuss in chapter 2. Chapter 3 details the methodology of the CKD analysis. The section includes the chapter overview, data collection, data pre-processing, data classification, and classification model verification. Chapter 4 presents the three levels of data classification results which are the comparison of before and after handling missing values, treatment of the outliers, and after the uncertainty imputation. The section includes the chapter overview, statistical analysis, classification, and comparative analysis. Chapter 5 conclude the

overall outcome of this study. The section includes concluding remarks, study contributions, and future recommendations.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter examines the state-of-the-art review emphasizing the methods used by the previous studies related to CKD problems and the factors attributed to CKD. The strategy of the journal search filtration is detailed. The approaches used in the past, the advantages and disadvantages of the methods, limitations, and the issues or challenges are presented.

2.2 Search Method

A broad search was performed in the ScienceDirect search engine with the keyword "Data Mining on Chronic Kidney Disease" resulting in 6103 articles that can be identified from the year 1998 to 2022 (Figure 2.1). All papers published in English were included. The papers were refined to six recent years; 2017 to 2022 thus returning 2691 articles. Considering only the research articles category was included in this research while other categories such as review articles, encyclopedia, book chapter, and book review were excluded only 1313 articles were left. These articles were further screened by titles and abstracts. Only 23 total papers were identified as relevant for reviews after the screening process. Figure 2.1 shows the schematic diagram of the entire search strategy.

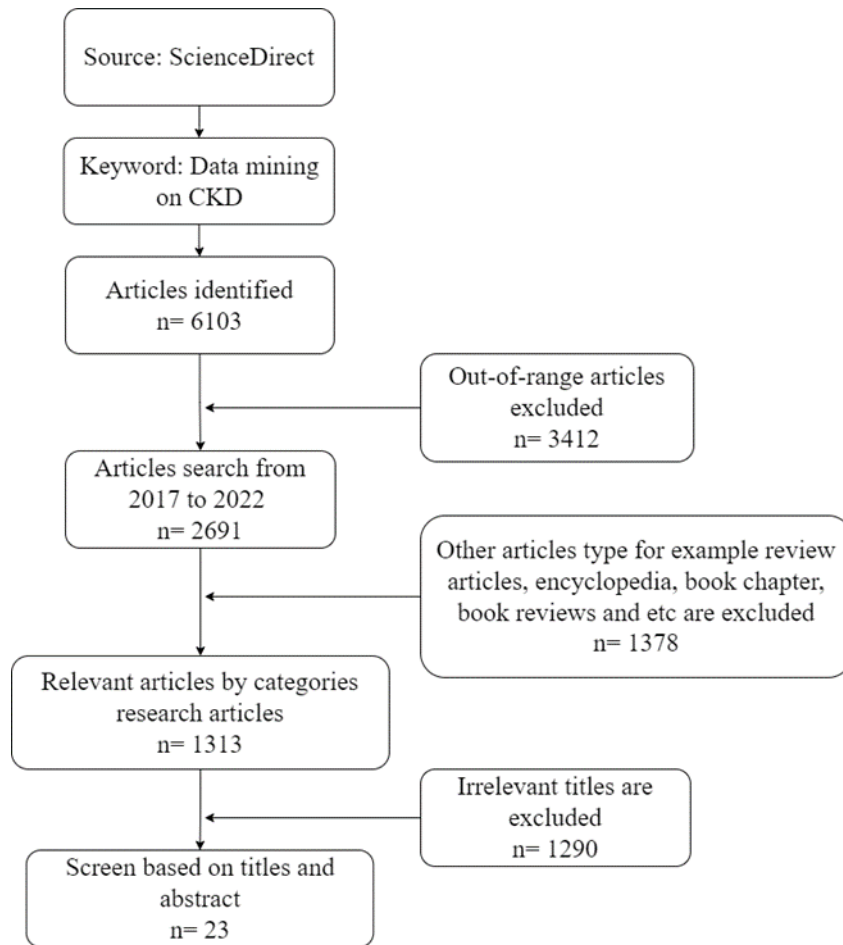


Figure 2.1 Schematic Diagram of Article Search Strategy

2.3 The Method and Algorithms used in CKD Studies

Out of the 23 articles, various algorithms were used for CKD analysis. The majority of studies applied more than one method or algorithm to achieve their study outcomes as shown in Table 2.1.

Table 2.1 Method or algorithm used in existing CKD studies

Method/algorithm	Number of papers	Paper
ANN	4	[8],[9],[10],[11]
NB	4	[8],[12],[13],[14]
KNN	4	[8],[12],[10],[14]

SVM	7	[8],[12],[1],[4],[9],[10],[15]
J48(DT)	1	[8]
PNN	1	[1]
MLP	1	[1]
RBF	1	[1]
XGBoost	1	[4]
LR	3	[4],[7],[16]
DT	3	[4],[5],[14]
RF	3	[4],[7],[10]
NN	2	[7],[16]
CNN	3	[17],[11],[15]
Improved Alex Net	1	[17]
Experimental	9	[6],[18],[19],[5],[20],[21],[22],[23],[24]
DNN	2	[25],[10]
DL	1	[11]
GBoost	1	[15]

Table 2.1 is summarized in Figure 2.2.

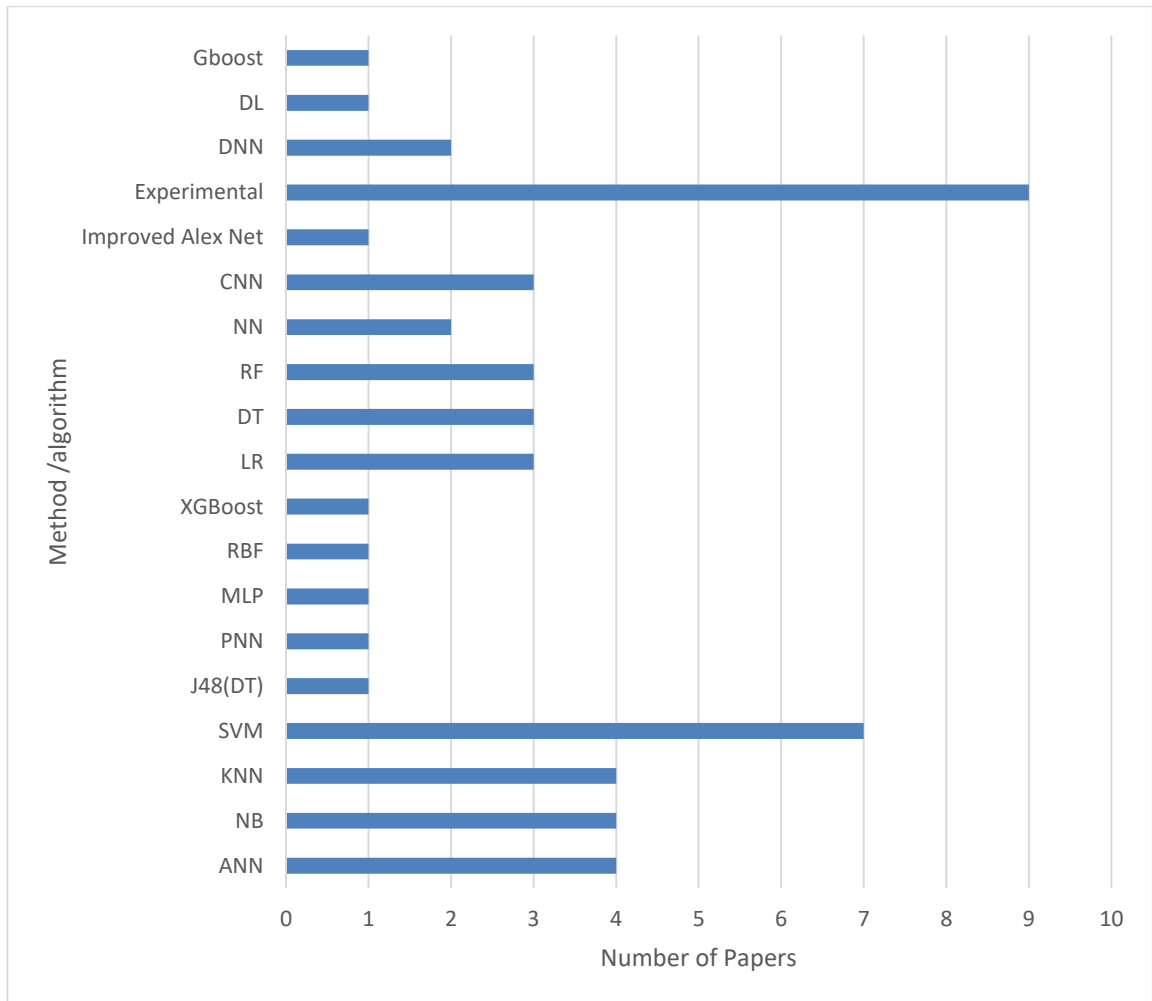


Figure 2.2 The number of papers using particular algorithms for CKD studies

Most articles used the experimental method which contributes to 9 out of the 23 papers. Out of the nine papers which applied experimental methods, there was one paper that combined experimental and analytical using a classification algorithm.

The aforementioned paper combined an experimental study with the Decision Tree (DT) algorithm to develop a model of acute AKI risk in severe burn patients [5]. Their study involved patients' datasets from Konya Burn Centers and hospitals from Baskent University Ankara during the period January 2000 to March 2020. The limitations in Karakaya et al. [5] were that patients with high voltage burns and previously diagnosed kidney disease were not included. Sepsis, oliguria, and hypertension were the factors that were excluded including the myoglobinuria

parameter. Other weaknesses stated by Karakaya et al. [5] are that the myoglobinuria values could not be reached in some of their patients due to the myoglobinuria parameter limitation. Nevertheless, their study found that total burn surface area was the most important factor in estimating the occurrence of AKI. According to Karakaya et al. [5], the Baskent University Model for AKI (BUMAKI) standards found no correlation between inhalation injury and AKI. Furthermore, the interaction between AKI and inhalation injury in severe burn patients is not entirely clear and is one of the issues pointed out by Karakaya et al. [5]. Additionally, if myoglobinuria is included among the parameters of the study, the treatment algorithm may be improved and the determining power may increase.

Linear Regression (LR) and Random Forest (RF) algorithms were used by Yang et al. [4] and Ventrella et al. [7]. The study from Yang et al. [4] was mainly focused on predicting renal failure risk in patients suffering from high-incidence chronic diseases such as hypertension and diabetes and standardizing the management of these patients. There were a few limitations in the study which are a non-CKD patient's kidney failure is not emphasized when assessing and preventing renal failure, patients which do not have biochemical tests are omitted and the range of the patient's age is between 30 to 85 years were excluded. While the study from Ventrella et al. [7] was to identify when a CKD patient should undergo dialysis, allowing for individualized care and treatment planning. The method used in [7] includes LR, DT, RF, and Neural Networks (NN). The use of NN is known for its superior performance on a variety of different problems when sufficient training data are available. The limitations in [7] were ignored the stages cases G1, G2, G3a, and partially G3b and public datasets are hard to find, as the analysis of the CKD evolution needs to be more detailed. Issues highlighted by Ventrella et al.

[7] are patients at high risk can be more effectively tended to during the next clinical encounter by scheduling it within a shorter or longer period.

RF also was used in another study by Pradeepa and Jeyakumar [10] along with Deep Neural Network (DNN), Artificial Neural Network (ANN), SVM, and k-Nearest Neighbours (KNN). This paper proposes a self-tuning spectral clustering scheme to reduce the problem of large dimensions and redundancy in data. The advantages found in this study were machine learning performance can be enhanced by eliminating the values that share the same characteristics through clustering and it is possible to predict CKD using DNN and SVMs with clustering techniques. As a result of clustering with machine learning techniques, this research provides the best results in terms of accuracy, specificity, sensitivity, precision, F_score, and recall.

In Jayasekara et al. [6], the researchers examine the impact of heat stress and dehydration on CKD among Sri Lankan community members. There were a few limitations in which the study had excluded participants under 18 years old and pregnant women. Also, self-report information and only one urine sample were used without the serum measurements.

Based on Figure 2.2, the highest algorithm method used is the Support Vector Machine (SVM). There were 7 articles reportedly using SVM. As in Rady and Anwar [1], the researchers used SVM as one of their algorithm methods in order to help physicians accurately identify the severity stages of diseases. Probabilistic Neural Network (PNN), Multilayer Perception (MLP), and Radial Basis Function (RBF) are among other algorithms used by Rady and Anwar [1]. Nevertheless, these algorithms were among the least methods used (Figure 2.2). According to Rady and Anwar [1], the PNN technique can easily be adapted to classify patients' severity stages of CKD.

Although PNN is the most effective algorithm, it needs a longer time (12 seconds) to complete the analysis as compared to the MLP which required only 3 seconds. Physicians can use the PNN algorithm to eliminate diagnostic and treatment errors since the authors have [1] recommended this algorithm.

SVM was also reported in another study by Pinto et al. [8] besides the ANN, NB, KNN, and J48 (DT) to predict the early stage of CKD using clinical data. According to Pinto et al. [8], J48 (DT) is the most effective algorithm compared to other algorithms. However, the weakness of their study was having an imbalanced dataset. Such a phenomenon potentially leads to an inaccurate assessment of the model's performance. The authors suggested using more algorithms for better results and increasing the number of instances in the dataset so that it is more substantial and balanced. Besides, it is also possible to use other metrics to explore more evaluation parameters, thus achieving more solid decisions [8].

There are eight least used methods reported, which include Gradient Boosting (GBoost), Deep Learning (DL), Improved Alex Net, Extreme Gradient Boost (XGBoost), RBF, MLP, PNN, and J48(DT) (Figure 2.2). RBF, MLP, PNN, and J48(DT) were used in [8]. DL is one of the algorithms in order to focus on AKI, CKD, End-Stage Renal Disease (ESRD), dialysis, nephropathology, and kidney transplantation studied by Yao et al. [11]. Besides the authors also used ANN and Convolutional Neural Network (CNN) along with DL. There were issues identified in their study regarding medical responsibility, patient privacy, data quality and quantity, and Artificial Intelligence (AI) aspects like clinician connections, safety info, and physician-patient connections. It is noteworthy that ethical regulations and regulations regarding AI should be established especially in healthcare.

GBoost algorithm was used together with SVM and CNN in Manonmani and Balakrishnan [15]. They also applied Teacher Learner Based Optimization (TLBO) and Improved Teacher Learner Based Optimization (ITLBO) features to be applied together with the algorithms in order to compare the results obtained between the algorithm with TLBO features and ITLBO features. An analysis of CKD data using these algorithms revealed a feature reduction of 36% as compared to the 25% achieved when the original TLBO algorithm was applied. Results presented by Manonmani and Balakrishnan [15] showed that the proposed feature selection method is useful for the early diagnosis of CKD, and it can be applied to other chronic illnesses as well. The study could be extended into the possibility of applying a filter approach for ranking the features within the medical dataset, and subsequently selecting the optimal features using the proposed ITLBO algorithm. In addition, comparative analysis of the superiority and performance of the proposed system can be done with semantic annotation based on an ontology and the proposed ITLBO algorithm.

2.4 Factors Attributing to CKD

There are various clinical and demographic factors considerably affecting the levels of CKD. Among the reported factors in existing studies include age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cells, pus cells clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia, and class [26].

Age: This attribute is measured in years under the numerical category. A person's age is a strong predictor of chronic kidney disease, and 11% of persons older

than 65 years without hypertension or diabetes have levels of creatinine that fall into stage 3 or worse [27]. Blood pressure (bp): It is defined as the measurement of the pressure within the major arterial system of the body, in millimeters of mercury, mm/Hg [28]. Conventionally, it is divided into systolic and diastolic measurements. The systolic blood pressure is the maximum recorded during a contraction of the ventricles, while the diastolic blood pressure is the pressure just before the next contraction. Nevertheless, there is no information as to whether the value of bp is for diastolic pressure or systolic pressure. Following Brzezinski [28], bp in this dataset is assumed to be diastolic pressure considering the range of this dataset which is 50 to 180 mm/Hg. According to Brzezinski [28], diastolic pressure below 85 mm/Hg indicates the person is normal, 85 to 89 mm/Hg, the person is high normal while between 90 to 104 mm/Hg, the person is under mild hypertension (Table 2.2). The person also can be in the moderate hypertension category if the reading of the diastolic pressure falls between 105 to 114 mm/Hg [28]. Severe hypertension measures the diastolic pressure above 115 mm/Hg [28]. From these diastolic ranges, it can be observed that high bp reading will lead to hypertension.

Table 2.2 The interpretation of blood pressure measurements in adults 18 years of age and older [28]

Diastolic Pressure (mm/Hg)	Category
<85	Normal
[85,89]	High normal
[90,104]	Mild hypertension
[105,114]	Moderate hypertension
>115	Severe hypertension

Specific gravity (sg): This attribute was recorded on a nominal scale. The specific gravity of urine is a measure of urine's density in relation to distilled water [29].

Additionally, this assessment measures the kidney's ability to excrete waste and balance its water content [30]. Generally, the specific gravity of urine is normally between 1.005 and 1.030 [30]. For example, 1.001 may be normal for someone who drinks a great deal of water. In some cases, above 1.030 may be normal for people who refrain from drinking fluids. Based on this study dataset, the highest value for specific gravity is 1.025 which shows the patient drinks a sufficient amount of water.

Albumin (al) and sugar (sg): These attributes were in the range of nominal scales {0, 1, 2, 3, 4, 5}. In blood plasma, the most predominant protein is albumin. According to Gurarie [31], al enables blood to circulate freely in arteries and veins and carries hormones, vitamins, and enzymes throughout the body. A normal level of albumin is between 3.4 and 5.4 grams per deciliter (g/dL) in blood supply [31]. Blood albumin levels that are too low may be an indicator of liver and kidney problems, and elevated levels, also called hyperalbuminemia, can be a sign of dehydration, diarrhea, or other conditions. Sugar (also called glucose) is another substance found in a person's urine. Normally, a small amount of sugar is found in a person's urine, but due to some health conditions, the level could rise above normal [32]. There are several causes of high glucose levels in urine, including medical conditions, genetic mutations, the effect of certain medications, and pregnancy. A high level of sugar in urine also may lead to CKD.

Red blood cells (rbc): It is also labeled on a nominal scale. Red blood, also known as erythrocytes, is a cellular component of blood, which carry oxygen from the lungs to tissues as well as give the blood its characteristic colour [33]. Red cells and hemoglobin are responsible for carrying oxygen throughout the body and for carrying carbon dioxide, waste products of metabolism, to the lungs, where it is expelled. It is

generally considered normal for adults to have 4.35 to 5.65 million red blood cells per microliter (mcL) of blood for men, while women are considered to have 3.92 to 5.13 million red blood cells per mcL [34]. Red blood cell counts in children vary with age and gender. A low reading of red blood cells can lead to anemia. The red blood cell level of an adult who has anemia is less than 100 grams per liter (g/L) while for a child is 75 g/L or less [35].

Pus cells (pc): This attribute is also a nominal scale. The presence of pus cells in urine also referred to as pyuria, may be a symptom of bacteria in the urine or can indicate an underlying urinary tract infection [36]. If the number of pus cells or HPF ≥ 4 in a centrifuged urine sample, pyuria is significant. **Pus cell clumps (pcc):** This attribute is the same as pc.

Bacteria (ba): In every environment, both inside and outside other organisms, bacteria exist in their millions, as for microscopic, single-celled organisms [37]. An infection of the kidneys is usually caused by bacteria, commonly the type E. coli, getting into the tube that carries urine out of the body (urethra) [38]. After entering a person's bladder, bacteria cause cystitis, which then spreads up into your kidneys.

Blood glucose random (bgr): It is measured on a numerical data scale. Blood glucose random is the glucose test to check glucose levels in a person's blood at any given point in time. A person who needs to perform a random glucose test does not need to fast before the test is run according to [39]. A person may have diabetes if their random blood glucose test results are 200 mg/dl or higher. However, most doctors will have to repeat the test a second time for a more accurate diagnosis. The doctor could order another test, such as a fasting glucose test, to confirm the diagnosis.

Blood urea (bu): In general, this kind of test is called the blood urea nitrogen (BUN) test. An analysis of BUN can determine how well a person's kidneys are functioning [40]. This is done by determining BUN levels. Proteins break down in the liver into urea nitrogen. The kidneys are normally responsible for removing this waste from the body through urination. Having too much urea nitrogen in the blood can indicate kidney or liver disease, as BUN levels have a tendency to increase when the kidneys or liver are damaged [40]. Table 2.3 below shows the interpretation of normal BUN levels.

Table 2.3 Interpretation of BUN levels [40]

Category	BUN levels (mg/dL)
Adult men	[8,24]
Adult women	[6,21]
Children (1-17 years old)	[7,20]

Based on Table 2.3, these values differ greatly from those values recorded in the dataset ranging from 1.5 mg/dl to 391 mg/dl. Such reflects an obvious condition of CKD patients.

Serum creatinine (sc): This attribute is on a numerical data scale. The breakdown of muscle tissue generates creatinine as a waste product according to [41]. The body excretes urine when it is produced; it is filtered by the kidneys and excreted by the body. In order to identify the level of serum creatinine in the body, a creatinine test should be implemented. By doing a creatinine test, doctors can measure kidney function, which is also known as a serum creatinine test. The creatinine clearance rate is a measurement of how effectively your kidneys handle creatinine. It helps estimate how quickly blood is getting through your kidneys, which is known as GFR. High creatinine levels could

indicate that kidney function is poor. Creatinine levels typically fall within the following ranges 0.74 to 1.35 mg/dl for men and 0.59 to 1.04 mg/dl for women [42].

Sodium (sod): This attribute is measured in milliequivalents per liter (mEq/L). As well as assisting electrical signals between cells, sodium also plays a large role in controlling fluid balance according to [43]. A majority of foods contain sodium in the form of sodium chloride, which is found in table salt. Health care providers use a sodium blood test (also called a serum sodium test) to measure how much sodium is in someone's blood. Sodium levels in the blood normally range from 135 to 145 mEq/L [44]. According to Borrelli et al. [45], CKD patients develop hypertension due to sodium and fluid retention, which causes hypervolemia.

Potassium (pot): Potassium which is also measured in milliequivalents per liter (mEq/L) is on a numerical scale. Potassium is one of the most important nutrients that people get from consuming high in potassium food, such as broccoli, bananas, and oranges [46]. The kidneys flush any extra potassium from the body when a person urinates. Therefore, having excessive potassium in the blood is a sign that the kidneys malfunction. The potassium level of normal adults is between 3.5 and 5.5 millimoles per liter (mmol/L) [46]. Hemoglobin (hemo) is a protein that is carried in the red blood cells along with iron [47]. Oxygen is held in this iron, making hemoglobin an essential component of the body. A cell in a body may not receive enough oxygen when the blood does not contain enough hemoglobin. In adults, the average hemoglobin level for males is 13 g/dL or higher while 12 g/dL or higher for women. Packed cell volume (pcv) is a numerical data attribute. The packed cell volume is a measure of the number of red blood cells in whole blood [48]. A pcv test is commonly referred to as a hematocrit test if the automated machine is used in the determination, but a pcv test is sometimes

referred to as a pcv test if the manual method is used [48]. Human pcv is normal between 40% and 50% in males while it is between 36% and 46% in females [48].

White blood cell count (wc): This attribute is measured in cells/cumm. It refers to the number of white blood cells in a person's body [49]. There is a test to determine the value of white blood cells in a person's body called a leukocyte test. The patient is often given this test as part of a complete blood count (CBC), which screens for a variety of conditions [49]. Table 2.4 shows the normal range of wc. When wc measures beyond this range, infection, or underlying health condition triggers.

Table 2.4 Normal wc range for the following category [49]

Category	WC range per μL of blood
Adult men	[5000,10000]
Adult women	[4500,11000]
Children	[5000,10000]

Red blood cells count (rc): It is also known as the erythrocyte count and is used to determine the number of red blood cells a person's body possesses [50]. Red blood cells contain hemoglobin, which helps deliver oxygen to the body's tissues. The lower the number of red blood cells in a person's body, the lower will be the oxygen received by tissues. The normal range for rc differs by age and gender as shown in Table 2.5.

Table 2.5 The normal range of rc [50]

Category	Normal range (million cells/ μL)
Adult (female)	[4.2,5.4]
Adult (male)	[4.7,6.1]
Child, 1 – 18 years	[4.0,5.5]
Infant, 6 – 12 months	[3.5,5.2]
Infant, 2 – 6 months	[3.5,5.5]

Infant, 2 – 8 weeks	[4.0,6.0]
New-born	[4.8,7.1]

Hypertension (htn): This attribute is measured on a nominal data scale. It is a condition when excessive force is exerted against the arterial walls of the heart according to the article causing high blood pressure [51]. As blood pressure rises, a variety of problems may occur, such as the formation of small bulges, called aneurysms, in blood vessels. The common way of diagnosing high blood pressure is with an arm cuff that is wrapped around the upper arm using a sphygmomanometer. As the blood beats against the arteries, an inflatable cuff measures the pressure of the blood. Since high blood pressure has deleterious effects on kidney vessels, hypertension is one of the leading causes of CKD [52].

Diabetes mellitus (dm): This attribute is a nominal data scale. A category of metabolic illnesses known as diabetes is defined by hyperglycemia brought on by deficiencies in insulin secretion, action, or both [53]. As a result of chronic hyperglycemia due to diabetes, different organs become damaged, dysfunctional, or fail, including the eyes, kidneys, nerves, heart, and blood vessels. Referring to Taylor [54], types of diabetes vary depending on their causes. The first type of diabetes is known as insulin-dependent diabetes. It was also called juvenile-onset diabetes because it normally begins as a child. The second type of diabetes is known as non-insulin-dependent or adult-onset diabetes. But it's become more common in children and teens over the past 20 years, largely because more young people are overweight or obese. Statistic shows that type 2 diabetes is about 90% of people who suffer from diabetes [54]. The third type of diabetes is gestational diabetes. Generally, it occurs in pregnant women which normally causes some kind of insulin resistance.

Coronary artery disease (cad): It is on a nominal data scale. According to Higuera [55], cad occurs when there is impaired blood flow in the arteries that provide blood to the heart. This cad can lead to death based on statistics in the United States from [55] article. Uncontrolled cad also can lead to heart attacks. According to Cai et al. [56] the most common cause of death and morbidity among CKD patients is coronary artery disease.

Appetite (appet): According to the article by Fletcher [57], a person's appetite is their craving to eat food. Unlike hunger, which is the reaction of the body to a lack of food, cravings are not a reaction to hunger. The type and amount of food people want to eat might be affected by their appetite. Even if the body isn't showing signs of being hungry, a person can have an appetite. People's appetites can fluctuate as a result of many factors, causing them to eat less or more than their bodies need.

Pedal edema (pe): In pedal edema (pe), ankles, feet, and lower legs accumulate abnormally large amounts of fluid, and this causes swelling [58]. The cause of foot edema can be attributed to two factors which are venous edema and lymphatic edema. Capillary leakage causes fluid from the venous system to leak into the interstitial space resulting in venous edema. In lymphatic edema, the lymph drains from the legs, either due to lymph malfunctions or it is blocked. The condition of edema can also result from other conditions, like kidney disease according to Hoffman et al. [59]. It is a kidney disorder called nephrotic syndrome that causes edema in the legs as well as swelling throughout the body.

Anemia (ane): The condition of anemia refers to lower than normal red blood cell counts or hemoglobin levels. Anemia causes a variety of symptoms, including fatigue and shortness of breath, which are caused by the body's organs and tissues not

receiving enough oxygen according to Lights et al. [60]. The level of hemoglobin in the blood determines whether a person has anemia. A variety of causes and types of anemia exist, and it can range from mild and treatable to potentially serious health complications. Referring to Lam [61], red blood cell destruction decreased or impaired production, and blood loss are the common causes of anemia. However, dietary changes may be able to resolve the issue for some people with anemia.

2.5 Summary

This section summarizes the outcomes of this chapter. The strategy of the journal search filtration was detailed with the keyword "Data Mining on Chronic Kidney Disease" by utilizing the ScienceDirect webpage. In short, 23 out of 6103 research articles could be used in this study as references to prove the reliability of this study. Next, the approaches used in the past, the advantages and disadvantages of the methods, limitations, and the issues or challenges were discussed. From section 2.3, most articles used the experimental method which contributes to 9 out of the 23 papers. Besides, there were 7 articles reportedly using SVM indirectly being the highest algorithm used out of 23 papers. Furthermore, the factors attributing to CKD were successfully discussed in section 2.4. The normal range for the concerned attribute was detailed such as blood pressure, specific gravity, albumin, red blood cells, blood urea, potassium, packed cell volume, white blood cell count, and red blood cell count. This information was useful for the data pre-processing level.

CHAPTER 3

METHODOLOGY

3.1 Overview

This chapter details the methodology to mine the risk factors attributing to CKD progression stages. The approach begins with a related dataset collected from a publicly available domain. Data pre-processing is carried out to handle potential missing values, outliers, and uncertain data. Two software were used for data pre-processing efforts which include WEKA 3.8.5 version and Microsoft Excel. The pre-processed data will undergo data classification analysis whereby the dataset will be categorized into three predefined stages of CKD: CKD, non-CKD, and uncertain. Three algorithms: NB, SVM, and J48 to perform classification at a 10-fold cross-validation model. In this study, the classification accuracy threshold was set at 80%. In short, the model of risk factors for CKD progression stages will be developed if the accuracy of the classification is exceeding 80%. Figure 3.1 shows the flow chart of the stages involved.

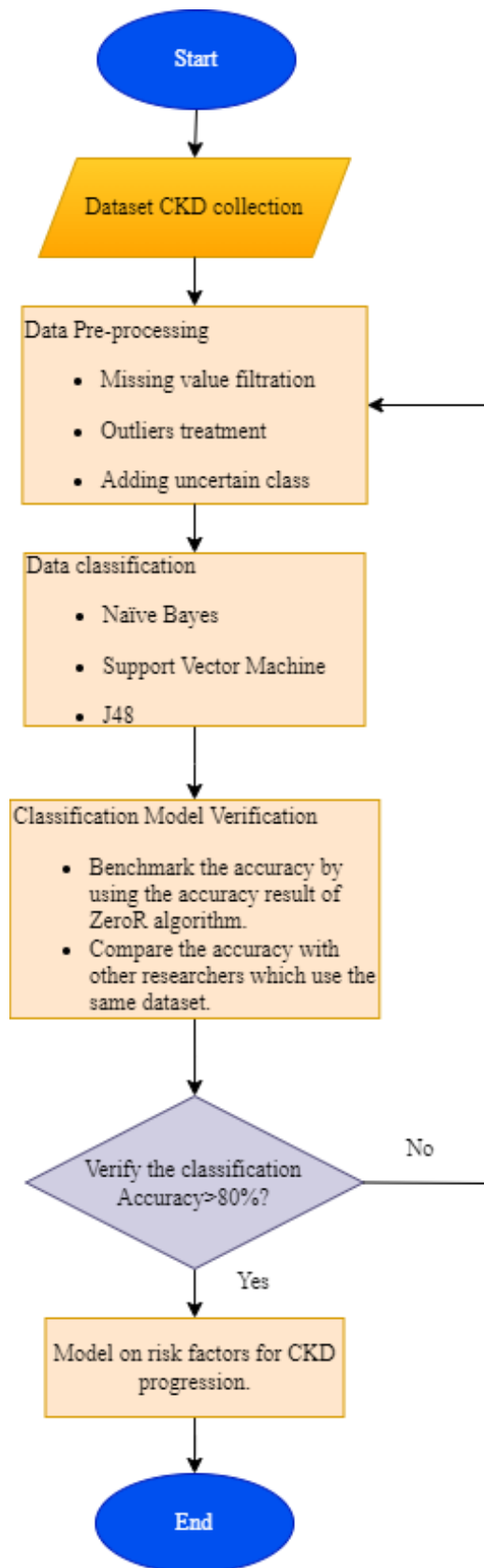


Figure 3.1 Flowchart of the methodology process