Arab American University

Faculty of Graduate Studies

Prediction of Chronic Kidney Disease Depending on Patient Treatment History Using Machine Learning Techniques

By

Mohammad Hasan Ibrahim Khateeb

Supervisor

Prof. Dr. Mohammed Awad

Co-Supervisor

Dr. Yousef Mimi

This thesis submitted in partial fulfillment of the requirements for the Master's degree in Health Informatics

April/2021

Thesis Approval

Prediction of Chronic Kidney Disease Depending on Patient Treatment History Using Machine Learning Techniques

By

Mohammad Hasan Ibrahim Khateeb

This thesis was defended successfully on 23/6/2021 and approved by:

Committee Members	Signature
1. Supervisor: Prof. Dr. Mohammed Awad	....
2. Co-Supervisor: Dr. Yousef Mimi	.....
3. Internal Examiner: Dr. Rami Hadrob	
4. External Examiner: Mohammed Aldasht	

Declaration

This thesis was submitted in partial fulfillment of the requirement for Master's degree in Health Informatics.

I declare that the content of this thesis (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Students Name: Mohammad Khateeb

Signature: 

Dedication

I feel very grateful for my parents, my family, my supervisors, and my friends who supported me strongly to achieve the desired results of this research, thus; I dedicate this work to them.

Acknowledgments

All praise and gratitude to Allah Almighty the creator, the creator of the intelligence. I feel very grateful for my parents, my wife Reem, my children (Reeman, Yaman, and Zeina) who supported me morally. Also, many thanks and gratitude to Prof. Dr. Mohammed Awad and Dr. Yousef Mimi, who guided me with his wise advice to complete this work to the fullest.

I would like to thank the IT department of the Ministry of health for facilitating gaining the dataset for Chronic Kidney Disease patients in Palestine.

Abstract

Chronic diseases are among the most consuming diseases of health care budgets and are characterized by treatment difficulty. Chronic kidney disease (CKD) is a global health problem with a high morbidity and mortality rate. Since there are no visible symptoms during the early stages of CKD, CKD is one of the diseases that have no efficient treatment till now, and the cost of treatment is high. Prevention and prediction are the two ways that can help in improving CKD patients' health and reducing the treatment costs. Early detection of CKD allows patients to receive timely treatment, which improves the progression of this disease treatment.

Machine Learning (ML) models can effectively help health professionals to achieve this goal due to their fast and accurate recognition performance of the CKD. In this study, we proposed a technique based on Machine Learning methodology for the diagnosis of CKD.

Different ML techniques were used to classify and predict CKD using Patients' treatment history data collected from the Palestinian Ministry of Health (PMOH) repository. First, a preprocessing step was used for cleaning, transformation, and feature selection for the collected dataset. In the next step, we applied Eight ML algorithms, namely: Decision Forest (DF), Decision Jungle (DJ), Support Vector Machine (SVM), Locally Deep Support Vector Machine (DSVM), Logistic Regression (LR), Boosted Decision Tree (BDT), Bayes Point Machine (BM), Neural Networks (NNs). These ML models' performances will be compared to decide the best classifier model in predicting CKD for the given dataset.

Therefore, the research method depends on a list of steps, starting by collecting the data, data preprocessing, determining the prediction and classification factors, training and testing data will be applied, and implementing different ML models using the training and testing dataset to classify and predict CKD.

The dataset worked entirely in the cloud on the Microsoft Azure platform. The results of the models used showed the superiority of the model Boosted Decision Tree. It outperformed other models with the following results: Accuracy= 0.945, Precision= 0.95, Recall= 0.95, F1Score= 0.95, and AUC= 0.972. The experimental results indicate that using the ML techniques to predict and classify the CKD supports the early diagnosis before the disease grows to advanced stages that are difficult to treat. Many patients can avoid kidney dialysis or search for any donor to perform a kidney transplant.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ALT	Alanine Aminotransferase
AST	Aspartate Aminotransferase
AUC	Area Under Curve
BDT	Boosted Decision Tree
BI	Business Intelligence
BPM	Bayes Point Machine
BUN	Blood Urea Nitrogen
CKD	Chronic Kidney Disease
DAGs	Decision-Directed Acyclic Graphs
DF	Decision Forest
DJ	Decision Jungle
eGFR	Estimated Glomerular Filtration Rate
EHR	electronic health records
EMR	Electronic Medical Records
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GP	general practitioner
HGB	Hemoglobin

LD-SVM	Locally Deep Support Vector Machine
LR	Logistic Regression
ML	Machine Learning
NNs	Neural Networks
PMOH	Palestinian Ministry of Health
PPCA	Probabilistic Principal Component Analysis
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UCI	University of California, Irvine

Chapter 1: Introduction

1.1. General Introduction

Chronic Kidney Disease (CKD) most significantly impacts the quality of life of patients and exponentially increases the risk of death. If the CKD damage is severe, the kidneys may stop working, and that leads to forcing the patient to have dialysis for the rest of his life or a kidney transplant to live [1]. Kidney Disease ranks 6th in the list of the fastest-growing diseases causing death, as the number of cases reached 850 millions cases globally, and 2.4 millions deaths per year [2]. 10% of the world's population have CKD, and the age group from 60 to 75 years, one in five men and one in four women suffer from CKD [3]. CKD is a high-cost pathology for generating a powerful economic impact on the finances of the System, causing a dramatic effect on the quality of life of the patient and her family.

Most people suffering from CKD have various complications, which affect many body functions. Among these complications is Gout, impaired kidney function accumulates uric acid in the body and may be reversible. The increased uric acid content in the body leads to impaired kidney function, and the two diseases are related [4]. Anemia is a condition in which the body lacks enough healthy red blood cells to carry adequate oxygen to your body's tissues [4]. Metabolic acidosis, the accumulation of metabolic acidosis in the body caused by kidney disease [4]. Bone disease and high phosphorus (hyperphosphatemia), the presence of phosphorous in the blood at a high level leads to the pull of calcium from the bones [4]. Heart disease has a common cause of death for

those who dialysis. High potassium (hyperkalemia) causes excess fluid to accumulate inside the body, leading to heart and lung problems [4].

The level of suffering from CKD is determined based on one of five stages, starting from the first stage in which the kidneys work almost typically in filtering the blood from waste products to the fifth stage, where the kidneys work harder to get rid of them, or they stop working completely [5]. Doctors describe these five stages by measuring the quality of the kidneys' filtering of wastes in the blood through the Estimated Glomerular Filtration Rate (eGFR) [5].

It has approved these five stages based on the following Table *1.1* below:

Table 1.1. Stages of CKD of all Types [5]

Stage	eGFR
Stage 1	90 or Greater
Stage 2	Between 60 and 89
Stage 3	Between 30 and 59
Stage 4	Between 15 and 29
Stage 5	Less than 15

Stage 3 is the boundary between the risk and delaying the transition to advanced stages, so stage 3 is considered the moderate stage. At this stage, the patient does not need to see a CKD specialist. However, it is sufficient for him to do a periodic test for blood and urine checks at a general practitioner (GP) to control the current situation and limit the transition stage 4. The patient may need other treatments to reduce other diseases, such as cardiovascular disease [6]. However, if the results of blood and urine tests

become a predictor of the transition to stage 4, it becomes necessary to see a doctor specializing in CKD [6].

Besides the health burden, CKD poses other financial, environmental, legal, and ethical burdens. Financially, in 2017 the total of what was spent on treating chronic kidney patients was over 84 billion dollars, over 22 thousand dollars per patient, while what was spent on late-stage patients was 36 billion dollars, equivalent to 80 thousand dollars per person (excluding medication costs) [7]. This cost is equivalent to 7% of the total health care costs paid [7]. As for the sustainability and preservation of the environment, dialysis needs to become more sustainable. 3.4 million people around the world dialysis use annually 2.5 million cubic meters of plastic, 204 billion liters of fresh water, and nearly 5 billion kilowatts of energy [8]. So far, more sustainable healthcare planning meetings are being organized and arranged [8].

The moral and legal problem is complex in managing the chronic kidney disease crisis because of several parties' presence and many cases that overlap. Among these issues, who is entitled to kidney transplantation with the increase in the number of people in need for kidney transplantation, the small number of donors, and the high cost of the kidney transplantation process [9]. Moreover, the cost of post-op immunosuppressive drugs, which not all patients can afford [9].

Besides these problems, the organ trade seen is spreading in many countries such as Pakistan, India, the Philippines, Egypt, South America, and Eastern Europe [9]. There are ethical, legal, and religious problems by the donor, as the donor may be minor and cannot decide in isolation from the family's influence [9]. In some countries, organ donation on death row is a moral issue because this is considered repugnant and morally

repulsive. Besides that, there is an ethical and religious problem in taking organs from the dead, or whoever has a brain dead while his heart is still beating [9].

In kidney living donation, physical harm is inevitable. There violates this ethical principle, as living kidney donation is associated with a mortality rate of 0.03% in donors and some surgical complications. However, it provides a significant benefit to the recipient [9]. By focusing on two aspects, the first: prevention and the second: prediction, the problem can be limited by reducing budget waste and improving the health status of chronic kidney patients [10].

To reduce the high mortality of CKD, research should be directed to diagnosis CKD in the initial stages of the disease, analyzing its risk group, with the help of laboratory tests, seeking those patients do not reach the final stages such as dialysis, transplantation, or death.

In the past two decades, the growth of and the increase in their power, with the availability of large data sets, and the development and improvement of algorithms rapidly led to the development and improvement of performing Machine Learning (ML) [10]. The basis for developing ML in the health sector is the availability of clinical data through Electronic Medical Records (EMR), and it is difficult for specialists and others to analyze this big data directly. The problem is not only a large amount of data or human errors but also the ability to elicit information and knowledge deeply. This can only be achieved through Artificial Intelligence (AI) and ML, as unparalleled features in humans characterize it of their mental abilities [10]. It bases using ML in CKD on four aspects: Alerting systems, Diagnostic assistance, Guiding treatment, and Evaluating prognosis [10].

Through ML, the aim is to find a valuable contribution so that a classification of the disease can carry out through the results of clinical laboratories, taking advantage of the great potential of ML, in the analysis and classification of data. In this thesis, the aim is to determine if the person is at risk of CKD using different ML techniques. We perform a comparison of classification algorithms to determine the more adequate classifier to improve the prediction accuracy in cases of CKD. During this thesis, all the experiments were performed on a local dataset from Palestine. The process starts with using the dataset as input and applies the preprocessing method, which is feature selection and data normalization. Then using the features to diagnose and classify CKD disease and to support the take of medical decisions.

1.2. Problem Statement

CKD is considered one of the critical diseases that accompany the patient for life, as it occurs through the loss of partial or complete kidney function, and dialysis or kidney transplantation is the only treatment that keeps the patient alive [11]. Early prediction using ML and discovering early treatment can reduce the risk of disease, or the patient may not reach the advanced stages of risk [11]. According to the Palestinian ministry of health , CKD is one disease that ranks high in the cause of death, and the death rate from this disease increases every year, according to the reports of the Palestinian Ministry of Health in 2017, 2018, and 2019 [12]. The causes of death because of chronic kidney disease was increasing, in 2017, it was the 12th cause of death by 2.2%, and in 2018 it rose to be the 9th cause of death by 2.9%, and in 2019 it reached the 8th cause of death by 3.2% [12]. Using ML techniques in healthcare has proven effective in

the accurate and economical diagnosis of diseases. Because of the presence of electronic health records in large quantities and their rapid development, the use of machine learning may be the most promising and ideal method for diagnosing CKD. This helps the specialist to make the appropriate decision at the right time.

Considering the above and that for each patient on dialysis there are people with some degree of probability of suffering from kidney disease. These facts generate concern in the national health system. Starting from the fact that if CKD is diagnosed in their early stages, they can receive low-cost treatments with a better quality of life for the patient and his family. For this, ML tools are necessary to help that by exploring the patterns on data can support the decision-making process in the diagnosis, quickly, with acceptable precision, and at low cost. Early diagnosis of CKD allowing the patient to receive treatment before its disease progresses to a dangerous stage.

1.3. Thesis Objectives

The research aims to design and implement different Machine Learning models (to design a clinical decision support system) based on a dataset collected from patients' treatment history stored in the Ministry of Health (MOH) repository to identify which one will have more accurate results in classifying and predicting of CKD. This allows predicting the possible diagnosis of CKD in its initial stages, helping to reduce the mortality rate and costs for the health system.

1.3.1. Primary Objectives

- To identify and analyze CKD features that could affect Palestinian patients.
- To study, analyze, and compare different ML models applied on a dataset collected from the MOH repository to identify the features to be used in constructing the ML model for the classification of CKD.
- Recommend which model produces the best accuracy to continue from this step toward a national tool (CKD clinical decision support system) for CKD risk factors classification and prediction.

1.3.2. Secondary Objectives

The secondary aim of the study is to improve the decision-making and to adopt CKD preventive factors.

1.4. Research Expected Outcomes

Using ML in the prediction and classification of CKD will produce a result as: will not have CKD or have CKD (0 or 1), which depends on the testing used evidence based on diagnostic accuracy or risk and connected it with how this would be expected to achieve better outcomes for patients through better early detection, better evaluation or better treatment of disease. on the other hand, the proposed model will be recommended to use as a decision support system for the CKD field's medical sector.

1.5. Research Significance

These predictions will be used for improving decision-making by healthcare providers for the patient's health. The main challenge is how to extract the features from these data because the number of features is high, and the data is vast, so some ML techniques can be used by applying a preprocessing of the collect data: data clustering, cleaning, filtering, and Normalization. Also, the expected result of this study is that if the CKD can be predicted so early treatment can be given to the patients, which can reduce the risk of life and save the life of patients and reduce the cost of treatment. The adoption of electronic health records (EHR) in MOH hospitals and facilities has created a wealth of data and information about patients, a goldmine for improving the service providing to patients. Selected ML methods used to predict and classify CKD using patient treatment history stored in the MOH repository.

1.6. Thesis Overview

Given the aim and objectives of the research, the material presented in this thesis falls into 4 chapters.

- **Chapter two** includes a literature review that studies previous related work and presents it.
- **Chapter three** where the research methodology is presented, including study population and sample size, data management and analysis framework (ML

Framework), study tools and data collection, confidentiality, and study limitations.

- **Chapter four** in which the results and discussion will be presented. This chapter will include descriptive and visualization for the CKD dataset. This chapter will also consider presenting and discussing the results of applying ML models on the selected dataset.

- **Chapter five** in which the conclusion, recommendations, future work, and strengths of the study, are presented

Chapter 2: Literature Review

2.1. Introduction

This chapter will review the literature relevant to CKD prediction, the wide application of ML in the medical field help to improve the quality of medical diagnosis and reduces its costs, the knowledge gap in the previous literature, and the conceptual framework of the study.

2.2. Chronic Kidney Disease

Chronic Kidney Disease (CKD) is the pathological clinical situation that can lead to end-stage renal disease. CKD is a condition that represents a high burden for the patient, the family, society, and the health system. There are several tests to detect CKD. One of them is the creatinine test [13].

Among the burdens of CKD is the high death rate, as the number of CKD cases worldwide in 2017 was 697.5 millions, an increase of 29.3% over 1990 [14]. The incidence of dialysis and kidney transplantation at all ages increased by 43.1% and 34.4%, respectively [14].

Also, from 1990 to 2017, the death rate of CKD increased globally, as the cause of death due to CKD increased from the seventeenth cause of death to the twelfth. In 2017, the number of deaths because of kidney disease was about 1.2 millions [14].

In Palestine, the death rate due to CKD increased from 2017 to 2019, from 2.2% to 3.2%, rising from the twelfth to the eighth place in the list of the most common causes of death as shown in Figure 2.1 below [12].

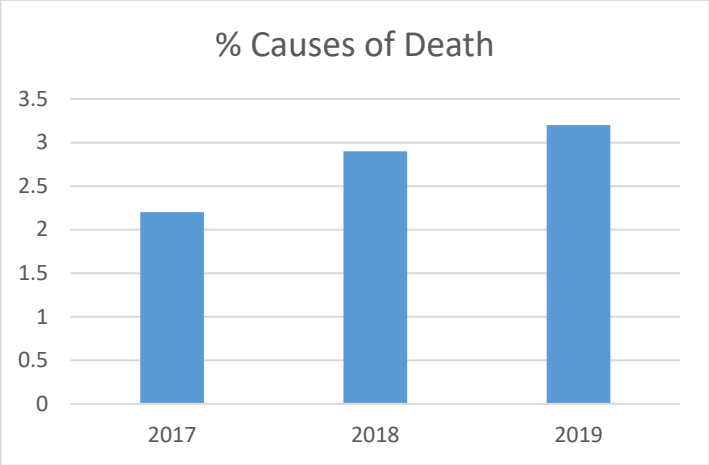


Figure 2.1 Death Rate Due to CKD in Palestine [12]

2.3. Machine Learning

Machine Learning is part of Artificial Intelligence that depends on exploring the patterns in the data, with the goal of prediction and classification of this data. ML models depend on algorithmic methods. The wide application of ML in the last 30 years in the medical field help to improve the quality of medical diagnosis and reducing its costs [15]. Thus, nowadays, technology affects all our life fields, it is clear that ML techniques and AI are one of those basics to obtain benefits in the daily problems of humanity.

This thesis aims to use various ML algorithms by using the patients' treatment history stored in the EHR to identify which one will have more accurate results. Moreover, the researcher will find out the correlations between various features. The study's objective

is to improve decision-making by finding the correct diagnosis and adopting CKD preventive factors.

2.4. Related Works

In this chapter, the previous studies related to CKD prediction reviewed, the ML methods used, and which method provide the best result, dataset sources, number of features, and features extraction and selection. In previous years, until now, many researchers have used various ML algorithms to predict CKD. Chittora et al.,2021 have developed a model for predicting CKD using ML. they have built the model by using all features and by selected features and tested the performance according to that [16]. The researchers depended on five steps for this research, dataset preprocessing, feature selection, classifier application, Synthetic Minority Oversampling Technique (SMOTE), and analyzing the classifier's performance.

The researchers used a dataset from the University of California (UCI) website. This dataset contains 400 clinical records and 24 features with one target class [16]. What distinguishes this research is the use of methods mentioned above to make feature selection: Filter Method by using "Correlation-based Feature Selection" method (CFS), Wrapper Method, and Embedded Methods by using "least absolute shrinkage and selection operator" method (LASSO) [16]. This research depended on seven classification techniques, Artificial Neural Network (ANN), Decision Tree (C5.0), Logistic Regression, Chi-square automatic interaction detection (CHAID), Linear Support Vector Machine (LSVM), K-Nearest Neighbors (KNN), and Random Tree

[16]. The researchers have used various tools. Weka tool used for CFS. R studio for LASSO. IBM SPSS Modeler used for performing classifiers. They built a deep neural network in IBM SPSS Modeler [16]. The researchers have observed that LSVM reached the highest accuracy of 98.86% in SMOTE with full features [16].

Almasoud et al., 2019 have developed detection of Chronic Kidney Disease using Machine Learning Algorithms with the least number of predictors [17]. The researchers used a dataset from the UCI website. This dataset contains 400 clinical records and 24 features with one target class. The researchers have used statistical tests to reduce features such as the Cramer's V test, the ANOVA test, and the Pearson's correlation, which led them to choose only three features to build prediction models: 1) hemoglobin, it was the highest contribution in detecting CKD 2) specific gravity and 3) albumin [17]. The study used python programming language using Jupyter Notebook and libraries from Sciket-learn [17]. The researchers have developed their prediction models by depending on 4 ML algorithms: Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Gradient boosting (GB) [17]. The result of each model has been evaluated using 10-fold cross-validation, and the experiments got higher performance with the GB by F1-measure (99.1%), sensitivity (98.8%), and specificity (99.3%) [17].

Aljaaf et al., 2018 developed their study by using four ML models for early prediction of CKD, which were, support vector machine (SVM), classification and regression tree (CART), logistic regression (LR), and multilayer perception neural network (MLP) [11]. The researchers used a dataset that was downloaded from the UCI website; and used different statistical tests to measure the correlation between the features like Chi-Square, Pearson's correlation for the association, and ANOVA for analysis of variance.

After that testing, they selected seven features out of 24 [11]. They examined the performance of these ML models. The results showed the MLP model had the best performance by highest AUC and sensitivity [11].

Snegha et al., 2020 have used data mining to predict chronic kidney disease; they downloaded a dataset from the Kaggle website, which contains 400 records and 24 features [18]. The researchers have used two data mining algorithms, Random Forest Algorithm (RF) and Back Propagation neural network, and the result they have was the Back Propagation neural network is more efficient than RF with an accuracy of 98.40% [18].

Pandit et al., 2020 try to reduce the death rate by CKD by using ML and several classifiers and clustering algorithms; they have used a dataset from the UCI repository, which has 400 records, 24 attributes [19]. The ML algorithms used in this paper were Random Forest, Decision Tree, K-Nearest Neighbors, and Naive Bayes, and the accuracy of these algorithms was 100%, 96.66%, 75%, and 91.66% consecutively [19].

Alloghani et al., 2020 focus on the prediction of CKD for high-risk cardiovascular diseases [20]; the study used data from the Ambulatory Electronic Medical Record (EMR) from Tawam Hospital outpatient clinics in Al Ain in UAE, and the number of records after the revision was 491, and it contained patients with CKD with late-stage, and their number was 56. At the same time, 435 were labeled as early-stage CKD cases. The study relied on 12 machine learning methods: KNN, DT, Radial Basis Function Support Vector Machine (RBF SVM), Stochastic Gradient Descent Neural Network (SGDNN), Ridge Logistic Regression, Random Forest, Polynomial SVM, Naive Bayes, Logistic Neural Network (LNN), Lasso Logistic Regression, CN2 rule induction, and

Boosted DT. The evaluation results showed that (Ridge and Lasso) logistic regression, SGDNN, LNN, and (Radial Basis Function and Polynomial) SVM had very high accuracy and efficiency. Moreover, the Polynomial SVM algorithm was the most efficient and accurate [20].

Chittora et al., 2020 study aim to predict CKD based on three features: Aluminum Level, Sugar Level, and Percentage of Red blood cells selected from the dataset downloaded from the UCI website and contained 24 attributes and one target variable, and 400 records [21]. In this research, five ML methods have been applied: Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), Logistic Regression (LR), and Random Tree (RT) and for each classifier, the results have been depended based on–(i) without preprocessing (ii) SMOTE with re-sampling and (iii) Class balancer. Furthermore, the results showed that the RF classifier in SMOTE with re-sampling was the highest accuracy with 98.93%.

Swathi et al., 2015 have got 690 cases and 49 attributes (Gender, Age, bath, blood group, body temp, breakfast, coffee, dinner items, diseases, drinking, fast foods, food habits, gender, height, job position, kidney stone, fruits intake, leafy-veg, lunch items, meals, milk, no drink, non-veg, coffee intake, milk, tea, smoke, place, fruits preferred, leaf, pregnant, relation members, salt consumption, scatter plot, sleeping, smoke, soft drinks, surgeries, sweat, tea, type of soil, type water, tablets used, water consumption, weight, and yoga) from dataset have collected from various hospitals from Visakhapatnam district in India. They have processed this dataset through the data mining tools such as Weka and Orange. Also, the researchers have used various ML algorithms such as K star, Alternating Decision Tree (ADTree), Naïve Bayes, J48, and

Random forest (RF). The researchers observed the K star and RF algorithms have the highest accuracy with ROC =1 [22].

Basar et al., 2018 have concentrated on this study to reduce the number of features used so that CKD can be diagnosed accurately and rapidly. Therefore, they tried 7 ML models (Naïve Bayes, Hoeffding Tree, Random Tree, REP Tree, Random Subspaces, Adaboost, and IBK), and the experiment repeated after reducing the number of features in each time according to the following sequence (24,12,6,3), The ML models performances have evaluated with five different performance metrics such as accuracy, kappa, mean absolute error (MAE), root mean square error (RMSE), and F measures. In the end, best accuracy and kappa values were calculated with Random Tree and IBk classifiers as 100% by using the six best features [23].

Pasadana et al., 2019 employed various decision trees algorithms such as Random Tree (RT), Decision Stump (DS), Hoeffding Tree, J48, CTC, J48graft, Random forest (RF), LMT, NB Tree, REP Tree, and Simple Cart to predict the CKD before the late stage [24]. These algorithms provide various experimental outcomes based on Accuracy, Mean Absolute Error, Precision, Recall, Kappa Statistics, and Runtime. The dataset has been collected from the UCI website. The selected algorithms have been evaluated and compared the performance of each algorithm by using WEKA. According to the analysis, RF was a higher performance than other algorithms, and its achieved accuracy is 100%.

Tekale et al., 2015, have studied various ML algorithms. They have used 14 specific attributes related to CKD patients collected from the UCI repository, and they have studied predicted accuracy for various ML algorithms like DT and SVM [25]. the

researchers have observed that the DT algorithms give an accuracy of 91.75% and SVM gives the accuracy of 96.75%.

Charleonnan et al., 2016, have presented the predictive models by using ML models including KNN, SVM, LR, and DT classifiers to predict CKD, and they have collected the data from the UCI website for features selection, they used the best first method [26]. The experimental results after using Matlab and Weka tools, the researchers have seen that the SVM classifier gives the highest accuracy, and SVM has the highest sensitivity after testing. Therefore, they have concluded that the SVM model is appropriated for predicting CKD.

Gunarathne et al., 2017 try to predict the CKD based on 14 features selected from the dataset downloaded from the UCI website and contain 24 attributes, one target variable, and 400 records [27]. In this research, four ML methods have been applied: Multiclass Decision Forests, Multiclass Decision Jungle, Multiclass Logistic Regression, Multiclass Neural Network. The results obtained using Microsoft Azure Machine Learning Studio show that Multiclass Decision Forest performs as the best one with an accuracy of 99.1%.

Sharma et al., 2016 have studied 12 ML models by using a dataset downloaded from the UCI repository [28]. Moreover, they reached that the DT algorithm gives the highest performance than other algorithms in terms of three out of four performance metrics with the accuracy of 98.60%, precision of 1, sensitivity of 0.9720, and specificity of 1 with accuracy around 99% after reducing the features to 14.

Anantha et al., 2016 study aim to predict the CKD for diabetic based on 600 clinical records have collected from a leading Chennai-based diabetes research center. In this

research, two ML methods have been applied: Decision Tree (DT) and Naïve Bayes (NB), and the results showed that the DT classifier was the highest accuracy with 91% [29].

Qin et al., 2019 paper aim to predict CKD based on 400 clinical records have collected from the UCI repository [30]. In this research, six ML methods have been applied: logistic Regression (LR), random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN), naïve Bayes classifier, and feed-forward neural network. Besides, for filling the missing data in the dataset, the researchers have used KNN imputation. The results showed that the RF classifier was the highest accuracy with 99.75%, and the researchers have proposed an integrated model that combines LR and RF by using perception, and after ten times of simulation, that achieve an accuracy of 99.83%.

Charleonnan et al., 2016 tried to predict CKD based on 400 clinical records have collected from the UCI repository [31]. In this research, four ML methods have been applied: K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), decision tree (DT). They used Matlab and Weka tools to develop the predictive models. At the end, SVM classifier got the highest performance with accuracy= 98.3%, Sensitivity= 0.99, and Specificity= 0.98.

Al Imran et al., 2019 tried to predict the CKD based on three ML methods that have been applied, namely logistic Regression (LR), feed-forward neural networks (Feed-Forward ANN), and wide & deep learning [32]. They have evaluated the performance for LR by calculated f1-score, precision, recall, and AUC score, and additional they used loss score for the Feed-Forward ANN, and wide & deep learning models. After developing the models, they found the feed-forward ANN as the best performing model

with f1-score= 0.99, precision=0.97, recall= 0.99, and AUC score=0.99. LR was the lowest result of all.

Sossi et al., 2019 have tried to analyze the CKD dataset (downloaded from the UCI repository) using two types of data analytics using IBM SPSS: statistical and predictive analytics to reach a 100% accurate model based on ML algorithm [33]. The researchers tried to predict the CKD based on 10 ML methods that have been applied, namely XGBoost linear, XGBoost tree, LSVM, C&R tree, CHAID, Quest, C5, Random trees, Tree-AS, and Discriminant, and XGBoost linear reached the highest performance with accuracy 100%.

The objective of Sinha et al., 2015 study is to introduce a new decision support system to predict chronic kidney disease [34]. This work aims to compare the performance of the Support vector machine (SVM) and K-Nearest Neighbor (KNN) classifier based on its accuracy, precision, and execution time for CKD prediction. From the experimental results, it is observed that the KNN classifier's performance is better than SVM. The datasets are extracted from the UCI Machine learning repository. The result shows the KNN classifier performed better than the SVM.

2.5. The Knowledge Gap in the Previous Literature

The weaknesses are the majority of those studies used the same data source is almost one, and it is from the UCI website (which is a global warehouse of all types of data). It is not updated data; for example, many datasets are collected before ten years. The CKD

dataset in the UCI warehouse contains 400 records only, and this is a small dataset to get results closer to accuracy and can be generalized.

In this thesis, we will use an updated national data set from MOH. This dataset was collected from Palestinian cases.

In this thesis, we will use a set of ML models, using Microsoft Azure Machine Learning, to predict CKD, 8 models of ML will be used, namely: Decision Forest (DF), Decision Jungle (DJ), Support Vector Machine (SVM), Locally Deep Support Vector Machine (LDSVM), Logistic Regression (LR), Boosted Decision Tree (BDT), Bayes Point Machine (BM), Neural Networks (NNs), and a dataset will be fetched from the Palestinian MOH repository.

We will identify and analyze the features of CKD that may affect Palestinian patients, by implementing the steps of pre-processing the obtained data set, to improve its quality and usefulness for generating models, and studying, analyzing, and comparing different ML models that will apply to a dataset collected from the MOF repository, to determine the features that will be used in building the ML model for classification of CKD. Finally, we will recommend the model that produces the best performance from this step toward a national tool (CKD clinical decision support systems) for the classification and prediction of CKD risk factors.

2.6. Conceptual Framework

The study will consider that Chronic Kidney Disease as a dependent variable that is directly affected by risk factors as independent variables, as shown in Figure 2.2 below.

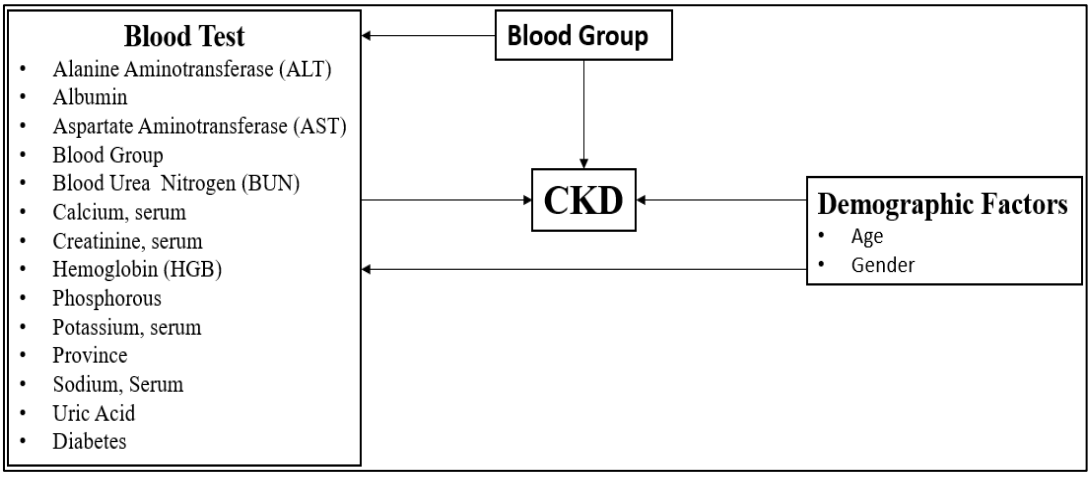


Figure 2.2 Research Conceptual Framework

The core risk factors included in the study are:

- CKD non-modifiable risk factors:
    - Demographic factors (such as gender and age).
    - Blood Group.
- CKD modifiable risk factors:
    - Blood Test:
        - Alanine Aminotransferase (ALT)
        - Albumin
        - Aspartate Aminotransferase (AST)
        - Blood Group
        - Blood Urea Nitrogen (BUN)

- Calcium, serum
- Creatinine, serum
- Hemoglobin (HGB)
- Phosphorous
- Potassium, serum
- Province
- Sodium, Serum
- Uric Acid
- Diabetes

2.7. Summary

The literature relevant to CKD prediction, the wide application of ML in the medical field helps to improve the quality of medical diagnosis and reduces its costs, the knowledge gap in the previous literature, and the conceptual framework of the study was presented in this chapter.

Chapter 3: Research Methodology

3.1. Introduction

In this chapter, the researcher will present study settings, study population and sample size, study tools, data collection methods, data management and analysis framework, ethical considerations, and limitation of the study. Furthermore, the researcher will explain how the four ML models developed on five feature selection methods.

Machine Learning can be divided into Supervised Learning, Unsupervised Learning, and Reinforcement Learning [35]. Supervised Learning is the most common form of ML used in the health care research field [15]. In general, supervised learning contains input data and output data. The most applied ML models in supervised learning type are Decision Forest (DF), Decision Jungle (DJ), Support Vector Machine (SVM), Locally Deep Support Vector Machine (LDSVM), Logistic Regression (LR), Boosted Decision Tree (BDT), Bayes Point Machine (BM), Neural Networks (NNs), which used as Classifier Sets.

In the execution of the ML to explore the pattern on the dataset, the whole process is divided into several phases; starting with focuses on understanding the objectives, then study and understanding of dataset, which begins with the dataset collection. Dataset analysis and feature selection will be the next step to build the final dataset from the initial dataset. Modeling techniques that related to the problem were selected and applied and their parameters are optimized. Finally, the evaluation or obtaining the results. In this thesis, different ML models have been applied.

3.2. Study Settings

The study was conducted by analyzing a dataset collected from the Palestinian Ministry of Health for Chronic Kidney Disease patients. The dataset is divided into two parts: records of patients with CKD, while the second section was for records of patients without this disease. Moreover, the data were chosen randomly after dividing the data source into three regions, North, Central, and South.

3.3. Study Population and Sample Size

The population from which the sample was taken can be considered all the medical records entered in the PMOH repository, as for the sample, it is a random sample containing 732 patient records, divided into two parts, the first part, files of patients without CKD, and there are 333 records. The second part is 398 patient records, and these are considered to patients have CKD.

3.4. Study Tools and Data Collection

The datasets were obtained from the PMOH, and before preprocessing the data, there were 732 records for patients with CKD and patients without that disease, and aged no less than 14 years, and 17 attributes included the class, which contain 0 for no CKD and 1 for CKD, and the other attributes are blood and urine tests, ages, gender, and address.

However, to select critical and important attributes as factors and symptoms for diagnosing CKD, an expert physician was consulted. After reviewing all attributes within the dataset, he sorted the attributes from most important to less important according to attributes associated with the eGFR (Creatinine, Gender, and age) covered on the dataset. Besides, using questionnaires for data collection was not possible and not suitable because tests of kidney functions are very accurate. No ordinary people can provide accurate information about their kidney function tests, wherefore the thesis was not depending on questionnaires, so the data taken from the source. The final selected attributes and their types and nominal values as shown in Table 3.1below.

Table 3.1. List of Selected Features

Name of Feature	**Type**	**Nominal Values**
Age	Scale	
Alanine Aminotransferase (ALT)	Scale	
Albumin	Scale	
Aspartate Aminotransferase (AST)	Scale	
Blood Group	Nominal	A+, A-, B+, B-, AB+, AB-, 0+, 0-
Blood Urea Nitrogen (BUN)	Scale	
Calcium, serum	Scale	
Creatinine, serum	Scale	
Gender	Nominal	Male, Female
Hemoglobin (HGB)	Scale	
Phosphorous	Scale	

Potassium, serum	Scale	
Province	Nominal	
Sodium, Serum	Scale	
Uric Acid	Scale	
Diabetes	Nominal	0, 1
CKD	Nominal	0, 1

Alanine Aminotransferase (ALT) is an enzyme found in the liver that helps convert proteins into the liver cells' energy. When the liver is damaged, ALT is released into the bloodstream, and levels increase [36]. Aspartate Aminotransferase (AST) is an enzyme that helps metabolize amino acids. Like ALT, AST is normally present in blood at low levels. An increase in AST levels may indicate liver damage, disease, or muscle damage [36].

3.5. Data Management and Analysis Framework: ML Framework

After extracting data from the dataset, the study framework to perform machine learning models to predict the probabilities of CKD occurrence was identified, and then compare these models' results according to Accuracy, Precision, Recall, F1 Score, and Area Under Curve (AUC). The proposed framework is shown in Figure 3.1 below.

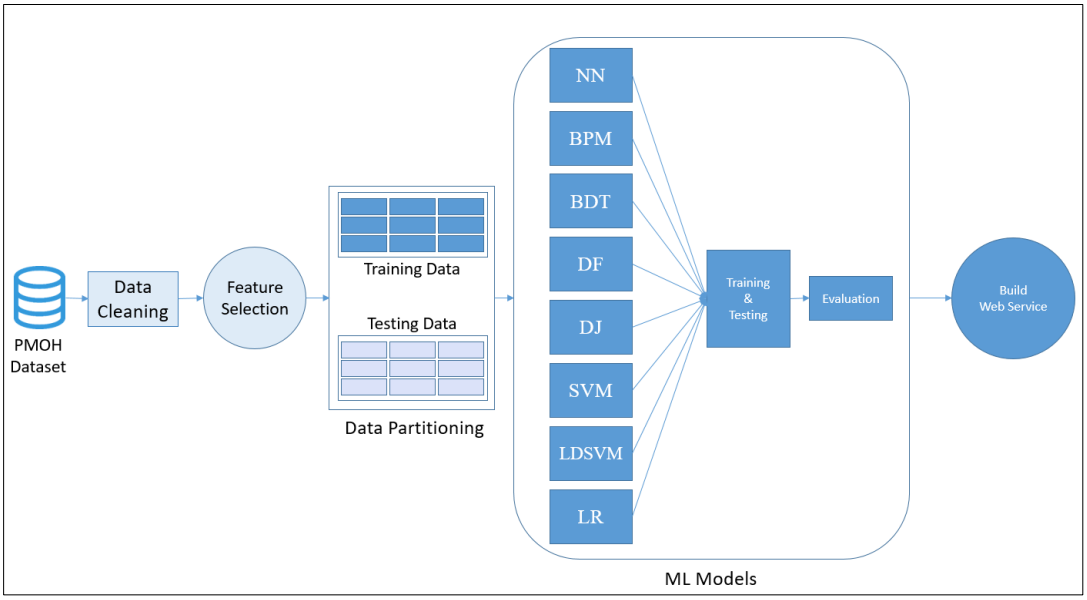


Figure 3.1. ML Proposed Framework

Within this framework, the dataset will be collected from the PMOH repository, as this data will be unprocessed and has a lot of duplicated records, deficient or entered as wrong data, so it will be cleaned. Then we will carry out selecting features, and high correlation features will be selected and exclude poorly correlation features. The data produced will then be divided into two parts, the first for training and the second for testing, and it will be delivered to each of the eight machine learning models until each model is trained and tested. After that, the results of each model will be evaluated. In the end, the best-evaluated model will be chosen, and the Web Service will be built from the chosen model to be used in the Clinical Decision Support System application.

The upcoming sections will show the details of each step in the ML framework:

3.5.1. Data Preprocessing

Data preprocessing in ML is a decisive step that raises data quality to elevate significant insights from the data. Data preprocessing in ML mentions the technique of preparing (cleaning, transforming, and organizing) the crude data to make it appropriate for building and training ML models. In simple words, data preprocessing in ML is a data mining technique that converts crude data into a readable and comprehensible format. In this thesis, Data preprocessing applied in three stages, which are:

- Data Cleaning
- Data Transformation
- Feature Selection

3.5.1.1. Data Cleaning

The goal of data cleaning is fixing or removing data within a dataset; it may need that for many reasons such as incorrect, duplicate, or irrelevant observations, outliers, corrupted, incorrectly formatted, or incomplete data. The dataset that was customized for this thesis was a need to clean for over reasons. First, duplicate records for the same patient with the same tests (features) were removed, as most patients had duplicate features. Besides processing the missing data, it was processed in two procedures, one, patients who had numerous missing features, and features which have many missing data, and these were removed from the dataset, procedure two was to fill in values for the missing features applying a Probabilistic Principal Component Analysis (PPCA)

algorithm. The tools used for data cleaning the dataset customized for this thesis are Microsoft excel and Microsoft azure ml (classic).

3.5.1.2. Data Transformation

The features that belong to the dataset customized for this thesis have different units that affect prediction models. This is due to the inability of the models to treat these features equally. When features are fed to the model, there is every chance that a feature that has an extensive range will influence the result more due to its more significant value. So the data transformation is needed. One of these that adopted in this thesis, which is Normalization.

Normalization *"is a technique often applied as part of data preparation for machine learning. The goal of Normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information."* [37].

The Technique that is used for Normalization in this thesis is Z-Score. Z-score depends on each feature's mean and standard deviation, as shown in the following equation

$$\text{Z-Score} = \frac{x - mean(x)}{standard\ deviation(x)} \quad (3.1)$$

Where x is feature value.

3.5.1.3. Feature Selection

Feature selection is a dimensionality stenography technique that attempts to select a subset of predictor variables (features) from the raw dataset that provides the best predictive power in modeling a dataset. Feature selection improves the model and avoids common problems, such as no information presenting more than currently selected features provided from redundant or irrelevant features. Whereas the redundant or irrelevant features take extra time in training and testing for the models, it also leads to inaccurate results. Besides, with some ML algorithms, the duplicating information leads to the multicollinearity phenomenon produced from two highly correlated features. Feature selection in this thesis depends in two ways. First, expert physician opinion, he nominated crucial and sensitive features which depend on CKD diagnoses, such as creatinine, gender, and age. Second, using "Filter Based Feature Selection" from Microsoft azure ML Studio (classic) tool, which Identifies the features in a dataset with the greatest predictive power [38]. Depending on 5 methods: Chi-Squared, Pearson Correlation, Mutual Information, Spearman Correlation, and Fisher Score. The result for the order of the features was as the following Table 3.2 below.

Table 3.2 Result of Filter Based Feature Selection

Chi-Squared	**Pearson Correlation**	**Mutual Information**	**Spearman Correlation**	**Fisher Score**
Province	Creatinine, serum	Province	Province	Creatinine, serum
Creatinine, serum	Province	Creatinine, serum	Creatinine, serum	Province
Blood Urea Nitrogen (BUN)	Blood Urea Nitrogen (BUN)	Blood Urea Nitrogen (BUN)	Blood Urea Nitrogen (BUN)	Blood Urea Nitrogen (BUN)
Aspartate	Potassium,	Aspartate	Aspartate	Potassium,

Aminotransferase AST	serum	Aminotransferase (AST)	Aminotransferase (AST)	serum
Potassium, serum	Hemoglobin (HGB)	Potassium, serum	Alanine Aminotransferase ALT	Hemoglobin (HGB)
Alanine Transaminase ALT	Phosphorous	Alanine Aminotransferase ALT	Potassium, serum	Phosphorous
Phosphorous	Calcium, serum	Phosphorous	Phosphorous	Calcium, serum
Calcium, Serum	Aspartate Aminotransferase (AST)	Calcium, serum	Hemoglobin (HGB)	Aspartate Aminotransferase (AST)
Hemoglobin (HGB)	age	Hemoglobin (HGB)	Albumin	age
Uric Acid	Alanine transaminase ALT	Uric Acid	Calcium, serum	Alanine ALT
Age	Gender	Age	Blood Group	Gender
Sodium, Serum	Albumin	Sodium, Serum	Age	Albumin
Albumin	Blood Group	Albumin	Gender	Blood Group
Blood Group	Sodium, Serum	Blood Group	Sodium, Serum	Sodium, Serum
Gender	Uric Acid	Gender	Uric Acid	Uric Acid
Diabetes	Diabetes	Diabetes	Diabetes	Diabetes

After a closer look at this table, diabetes was avoided because it has the lowest correlation in all methods, and Province deposed because the data distribution was not normal, as the study sample contained the largest percentage of chronic kidney patients from the Hebron area, as will be mentioned in chapter 4. Finally, according to specialist opinion, the features selected were (Age, Alanine transaminase ALT, Albumin, Aspartate, aminotransferase AST, Blood Group, Blood Urea Nitrogen (BUN), Calcium,

serum, Creatinine, serum, Gender, Hemoglobin (HGB), Phosphorous, Potassium serum, Province, Sodium Serum, and Uric Acid) were added to the list of 16 features.

3.5.2. Data Partitioning

Before building ML models, the data split into one set that will develop models, preprocess the features, and explore relationships and hidden patterns among the features and the outcome (the training data), and another that will be the final judge of the model performance (the test data). To division, the data, the splitting of the original data set will be done by making random splits.

In this thesis, 70% of the cleaned and processed will be given to the training set, and the remain 30% of data will be given to the testing set. Besides that, the thesis will compare the results between the above way of partitioning and cross validation.

Cross-validation is a practical and reliable way for testing the predicting capability for used methods. Hence, it's necessary for any non-linear models to have a training set and test set in order to verify and to check its validity and estimate the quality of models. Also, Cross-validation is the best choice to avoid minimum overfitting or underfitting. Many styles and varied types of cross-validation can use [39], but a core at the end is similar and in the same meaning.

The simplest kind of cross-validation is the holdout method or sometimes named test sample estimation, the dataset distributes or is divided into two mutually independent subsets, one of them called a training set and the other called a test set. It is widespread to allocate about (0.70) of the data as a partition of the training set and about (0.30) as a

partition of the test set. Approximator employs a function using just the training dataset. Next, the Approximator is carried out to predict the output values for the testing dataset, knowing that this test partition Not seen before.

Another way that can be used is K-fold cross-validation, where the data set is partitioned into k subsets, and the test sample estimation method is repeated k times individually [39]. Every time, we select one of the subsets to be the test set and the other subsets are placed unitedly to be a training set. Thus, each data point captures in a test set precisely. We always look at less expensive computing processes and fewer consumption times more in running so we start to apply the easiest and fastest way to get the desired results, maybe the emergence of problems with regard to (over/underfitting) and therefore we back to other ways such as K-fold technique to be used. In our thesis, we divided data points into two sets randomly, and we selected the portion of each set as a Rate between (70%) for the training set and (30%) for the testing set.

3.5.3. Building Models Phase

Because the tool used is Microsoft Azure ML, the models that are used in predicting or answering a simple question which answers is yes or no, classification which includes: DF, DJ, SVM, NN, LD-SVM, LR, BDT, and BPM.

3.5.3.1. Decision Forest

DF is an ensemble learning method purposed for classification tasks, and ensemble methods are mean do not depend on a single model like a decision trees model, and get better results by creating multiple DT models by separated into subsets and voting the highest number of subset results, as in Figure 3.2 below. the algorithm applied on subsets is the same DT algorithm [33].

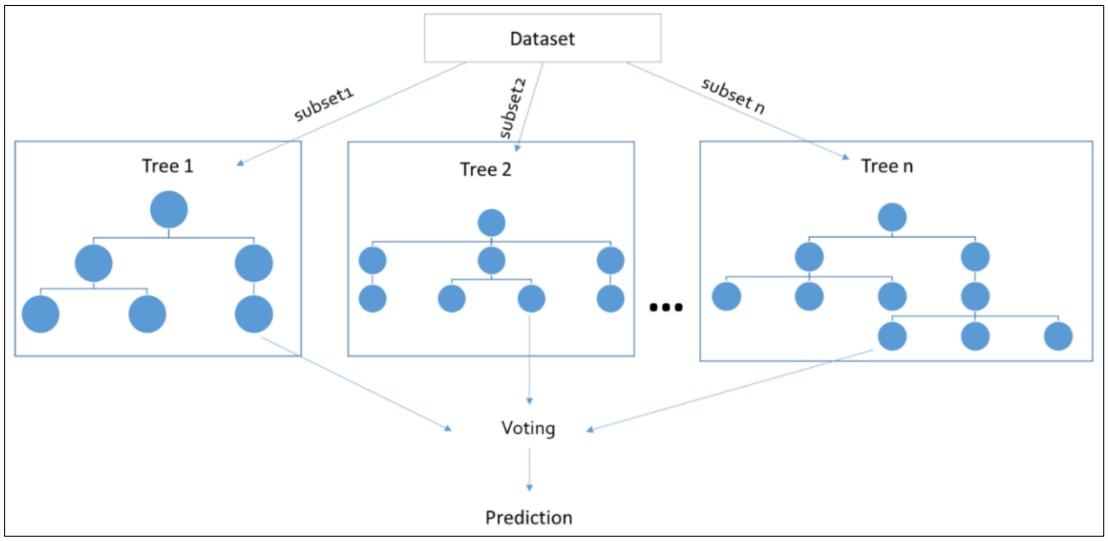


Figure 3.2 Decision Forest

3.5.3.2. Decision Jungle

The DJ module produces an untrained classifier. Then this model could be trained on a labeled training dataset by using Train Model or Tune Model Hyperparameters. Then the trained model ready to use predictions [41]. DJ is a new extension to DF. A DJ contains an ensemble of Decision-Directed Acyclic Graphs (DAGs). By allowing multi-paths from root to each leaf, a decision DAG mostly has a lower memory usage and better performance than a decision tree, but it takes longer training time [41].

3.5.3.3. Support Vector Machine

SVM is one of the most powerful supervised machine learning methods. The classifier that is produced by this built model helps predict two outcomes that depend on continuous or categorical predictor variables [42]. This model is a supervised learning method and so needs a labeled column in the dataset. The model is trained by using the model and the dataset as an input to Train Model or Tune Model Hyperparameters, then can be used to predict the outcome for the new input data. Also, SVM using hyperplanes to help the model by distinguishing the predicted outcome [42].

3.5.3.4. Neural Network

NN is a supervised learning method and consequently needs a tagged dataset that includes a label column [43]. A NN is a set of interconnected layers. It started with the inputs layer and is connected to an output layer by an acyclic graph that included weighted edges and nodes, as shown in Figure *3.3* below [33].

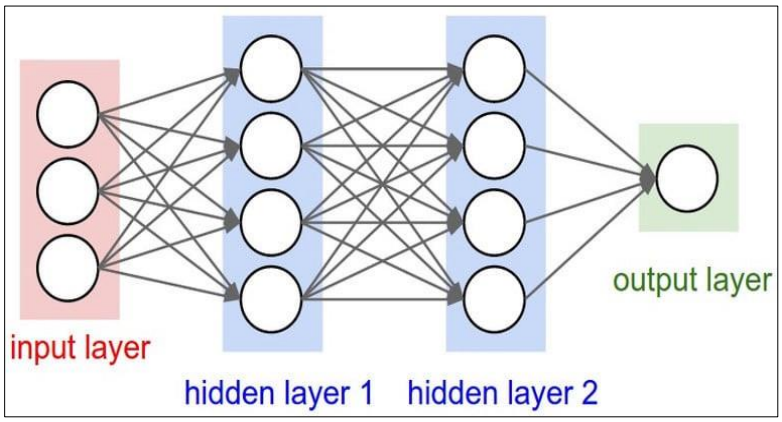


Figure 3.3 Neural Network architecture

In this thesis, the NN model will be 4 tests applied by different hidden nodes, 20, 30,40, and 50. Multiple hidden layers are inserted between the input and output layers, and it could be multiple hidden layers. For predictive models, one or a few hidden layers could be sufficient [43].

3.5.3.5. Locally Deep Support Vector Machine

LD-SVM is a scheme for effective non-linear SVM prediction while preserving classification precision above an acceptable limit [44]. Using a local kernel function allows the model to learn arbitrary local embedding features, including sparse, high-dimensional, and computationally deep features that bring non-linearity into the model [44]. The model employs routines that are effective and primarily infused to optimize the space of local tree-structured embedding features in more than half a million training points for big training sets. LD-SVM model training is exponentially quicker than traditional SVM model training [44].

3.5.3.6. Logistic Regression

LR is a supervised learning technique. It is used for predicting a categorical or discrete value. It can be either Yes or No, 0 or 1, true or false, but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1 [45].

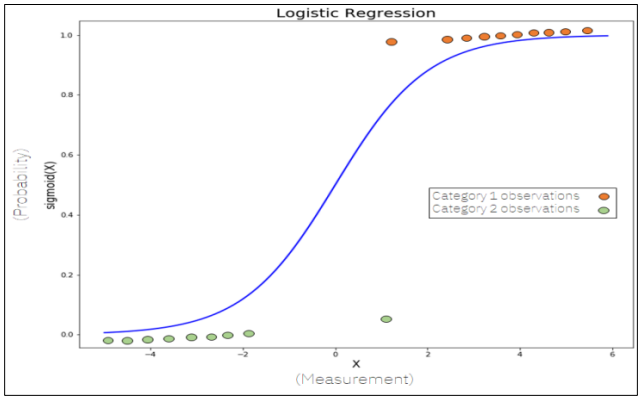


Figure 3.4 Logistic Sigmoid Function

LR transforms its output using the logistic sigmoid function as Figure 3.4 to return a probability value which can then be mapped to two classes [45].

3.5.3.7. Boosted Decision Tree

A BDT is an ensemble learning method in which the second tree corrects the first tree's errors. The third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction [46]. Generally, when adequately configured, boosted decision trees are the easiest methods with which to get top performance on a wide variety of machine learning tasks. However, they are also more memory-intensive learners, and the current implementation holds everything in memory [46]. Therefore, a BDT model might not be able to process the vast datasets that some linear learners can handle [46].

3.5.3.8. Bayes Point Machine

This module's algorithm uses a Bayesian approach to a linear classification called the "Bayes Point Machine". This algorithm efficiently approximates the theoretically optimal Bayesian average of linear classifiers (in terms of generalization performance) by choosing one "average" classifier, the Bayes Point. Because the BPM is a Bayesian classification model, it is not prone to overfitting to the training data [47].

3.5.4. Models Evaluation

After applying the four ML tools on the cleaned dataset customized for this thesis, each tool's performance will be measured based on Accuracy, Recall, Precision, F1 Score, and Area Under the Curve (AUC). Some essential terms have to be explained before start evaluating the CKD classification models, such as:

True Positive TP: Number of records that are correctly predicted the positive class.

False Positive FP: Number of records that are incorrectly predicted the positive class.

True Negative TN: Number of records that are correctly predicted the negative class.

False Negative FN: Number of records that are incorrectly predicted the negative class.

Accuracy: The accuracy of machine learning tools is defined as the percentage of correct test data predictions. It is calculated by the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad 3.2$$

Recall: Recall in machine learning is the percentage of positive cases in test data that were correctly predicted. The recall is also termed sensitivity, and this is calculated based on the following formula:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

Precision: Precision is also termed as Positive predictive value (PPV), and it means the ratio between the TPs and all the Positives. This is calculated based on the following formula:

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{3.4}$$

F1 Score: It is the harmonic mean of precision and sensitivity. This is calculated based on the following formula:

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.5}$$

Area Under Curve (AUC): AUC can be explained as the ratio of correctly classified samples. The shape of the curve gives a hunch for the relationship between TPR (i.e., sensitivity) and FPR (i.e.1-specificity) as a function of the classification threshold or decision boundary.

3.5.5. CKD Web Service Building

After applying all ML models and comparing the results, the best model in terms of results will be selected, and web service for the selected model will be developed through Microsoft Azure ML (classic), to be used in building an interface for make predictions using new data, also on the same tool.

3.6. Ethical consideration

First, data were collected based on a previously signed agreement between the Arab American University and the Palestinian Ministry of Health.

As for the confidentiality of data, the data collected does not contain any patient identifier, no names, no identification numbers, or any other identifier.

3.7. Limitation of the study

Despite this study can be considered is unique, as it is the first of its kind of study in Palestine, but it contains some limitations that were difficult to overcome, such as the inability to obtain a greater number of features that directly affect the diagnosis of chronic kidney disease, and a large amount of missing data. Besides the previous limitations, the dataset has scarce data from other Palestine areas, such as Jerusalem and the Gaza Strip. It can be considered that the dataset was not normally distributed, and

this affected the quality of the results in the descriptive analysis of the data in the Descriptive Statistics and Visualization section.

In addition, the researcher faced logistic barriers, such as, time, as it was considered one of the biggest barriers that faced the researcher. Besides, the dilemma of the Covid 19 epidemic, this obstacle was one of the most difficult obstacles that contributed to the difficulty of working on the thesis, as the researcher could not travel or reach the PMOH to collect data smoothly.

3.8. Summary

In this chapter, the researcher presented study settings, study population and sample size, data management and analysis framework (ML Framework), study tools, data collection methods, data management and analysis framework, ethical considerations, and limitation of the study. Furthermore, the researcher explained how the four ML models developed on five feature selection methods.

Chapter 4: Results and Discussion

4.1. Introduction

In this chapter, the results of the different analysis methods performed on the dataset customized for this thesis will be presented. Also, The ML tools results and general descriptive statistics about the data within the dataset customized for this thesis will be discussed.

4.2. Descriptive Statistics and Visualization

After CKD patients' records were acquired, descriptive analysis and visualization were applied using IBM SPSS to describe the data.

4.2.1. CKD Patients Living Province

As shown in Figure 4.1 and Table 4.1 below, the highest Palestinian Province with CKD patients living in Hebron, where 97.7% of patients and, while Ramallah, Tubas, Tulkarm, and Qalqilya gained the lowest percentage with 0%. In addition, Nablus has the highest percentage of healthy people with 35.1%, while Jerusalem and Gaza gained the lowest percentage with 0%.

Table 4.1 CKD Patients Living Province

		Province										Total
		Hebron	Jerusalem	Nabuls	Bethlehem	Qaza	Jenin	Ramallah	Tubas	Tulkarem	Qalqilya	
CKD	Count	46	0	117	44	0	59	37	5	24	1	333
	% without CKD	13.8%	0.0%	35.1%	13.2%	0.0%	17.7%	11.1%	1.5%	7.2%	0.3%	100.0%
	Count	389	1	1	3	1	3	0	0	0	0	398
	% within CKD	97.7%	0.3%	0.3%	0.8%	0.3%	0.8%	0.0%	0.0%	0.0%	0.0%	100.0%
Total	Count	435	1	118	47	1	62	37	5	24	1	731
		59.5%	0.1%	16.1%	6.4%	0.1%	8.5%	5.1%	0.7%	3.3%	0.1%	100.0%

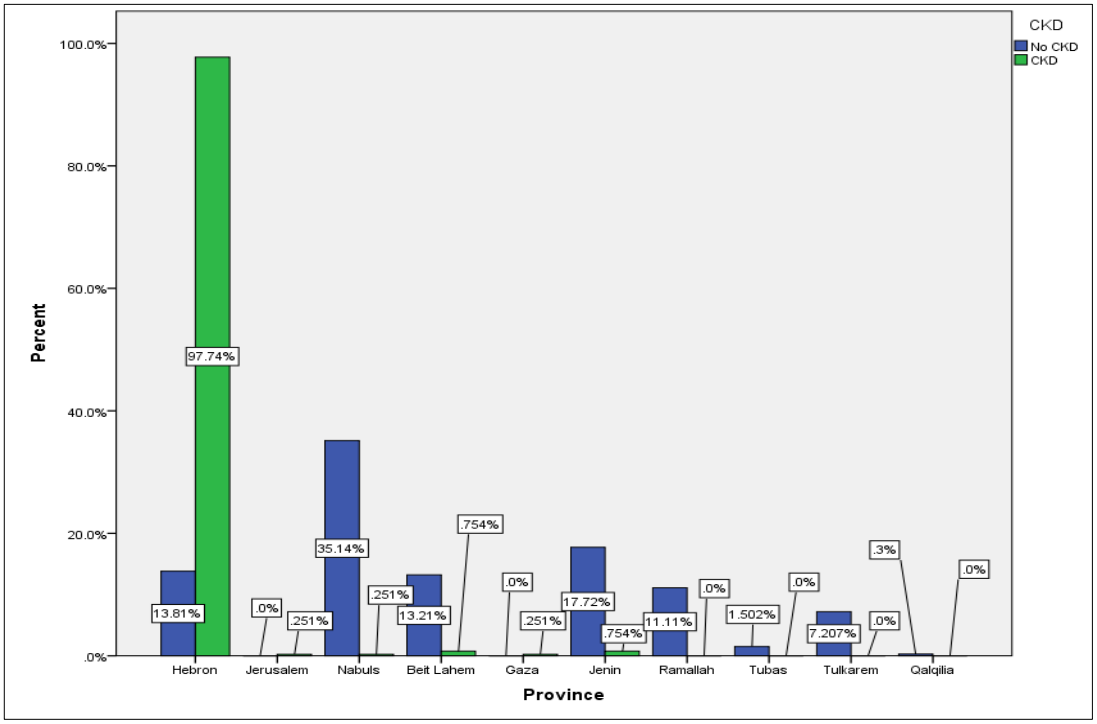


Figure 4.1 CKD Cases Living Province

4.2.2. CKD Patients Ages

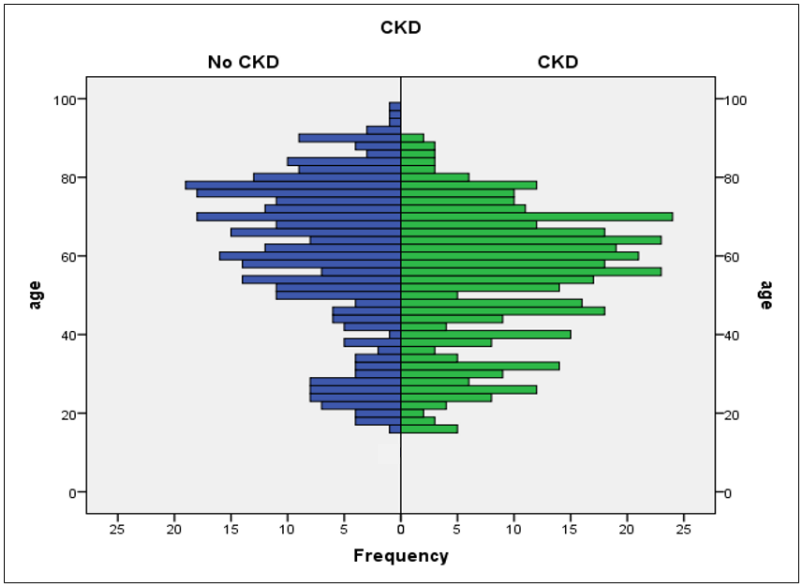


Figure 4.2 below shows the age distribution between 40 and 80 in approximately homogeneous.

We can notice that the number of healthy people increases with increasing age, from 50 to 80 years. While the largest number of patients who have CKD, we can confine them between the ages of 55 and 70.

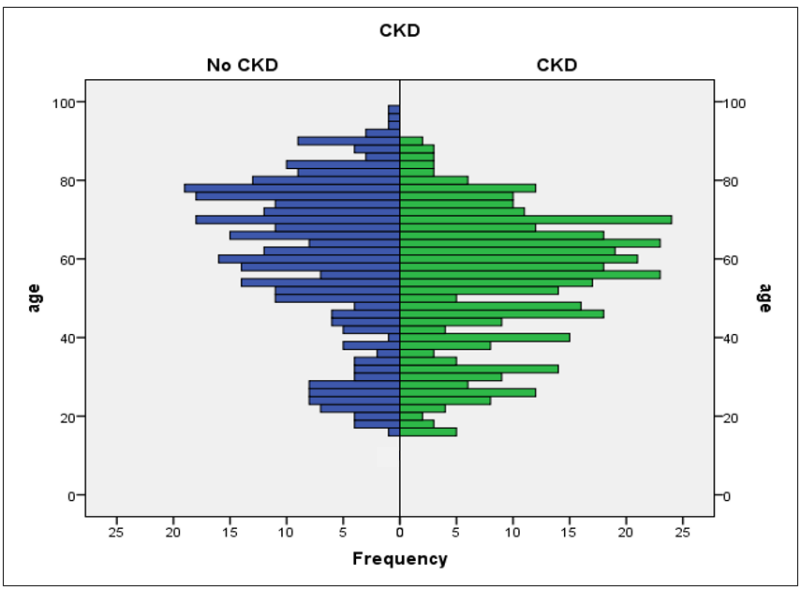


Figure 4.2 CKD Patients Ages

4.2.3. CKD Patients Gender

Figure 4.3 below shows an apparent convergence between the percentage of males and females. However, a slight increase can be observed in the percentage of healthy and patient males than females.

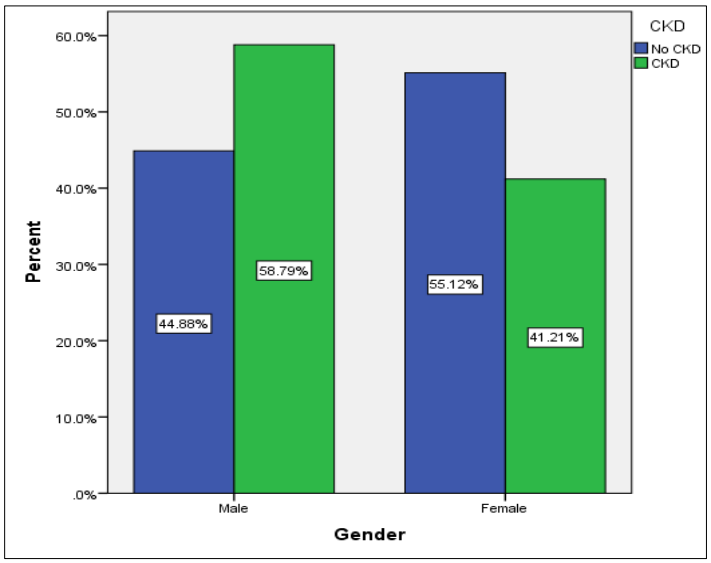


Figure 4.3 CKD Patients Gender

4.2.4. CKD Patients Blood Group

As shown in Figure 4.4 below, the highest blood group with CKD patients is A+, where 37.19% of patients and, while AB- gained the lowest percentage with 0.251%. Also, A+ has the highest percentage of healthy people with 37.65%, while AB- gained the lowest percentage with 0%. Besides that, some records have an Unknown blood group with a 14.82%.

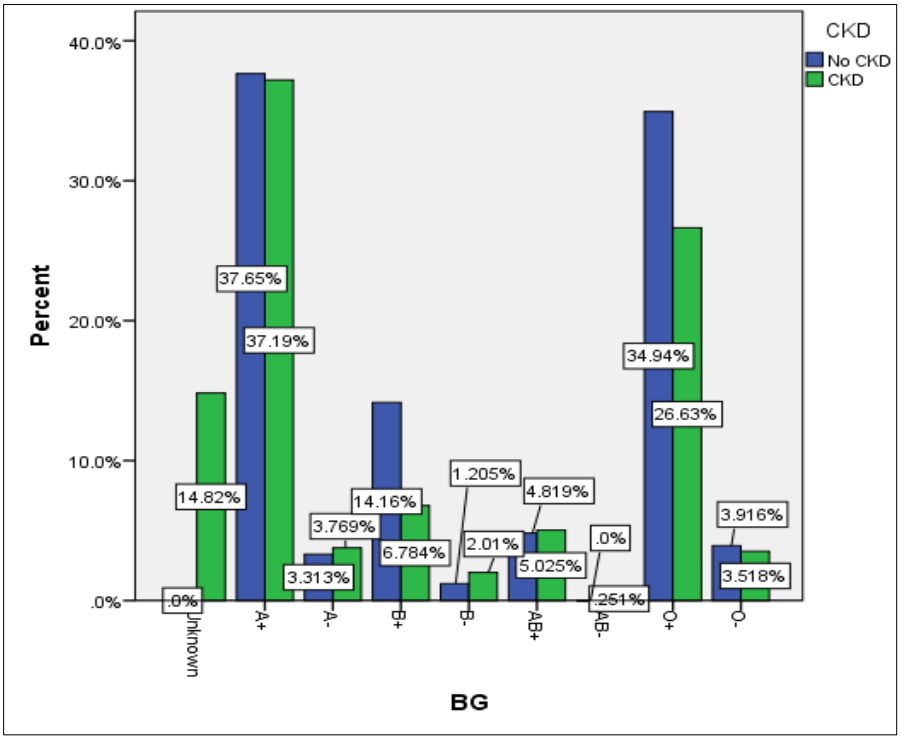


Figure 4.4 CKD Patients Blood Group

4.2.5. ALT & AST Comparing means:

Comparing means using One-way ANOVA between ALT and CKD and between AST and CKD showed that serum AST and ALT levels were significantly lower in CKD patients, as shown in Figure 4.5 and Figure 4.6 below. This result agreed with the studies that investigated the relationship of these two enzymes with CKD [48].

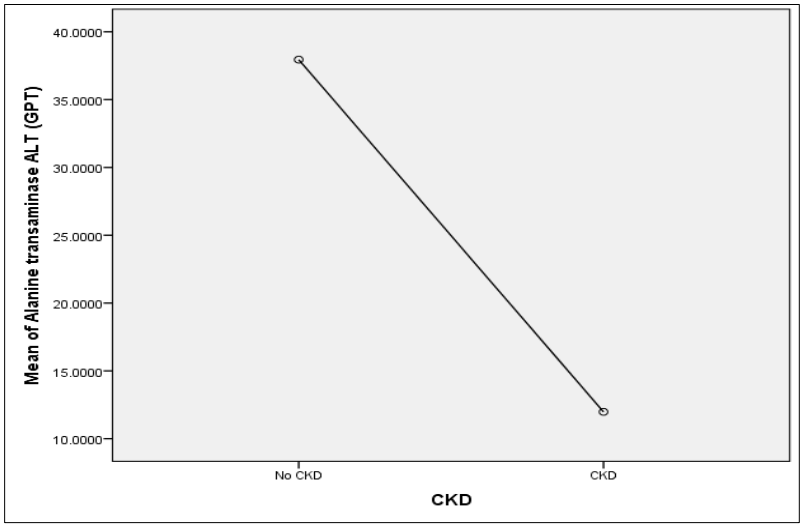


Figure 4.5 Alanine Aminotransferase -CKD Compare Means

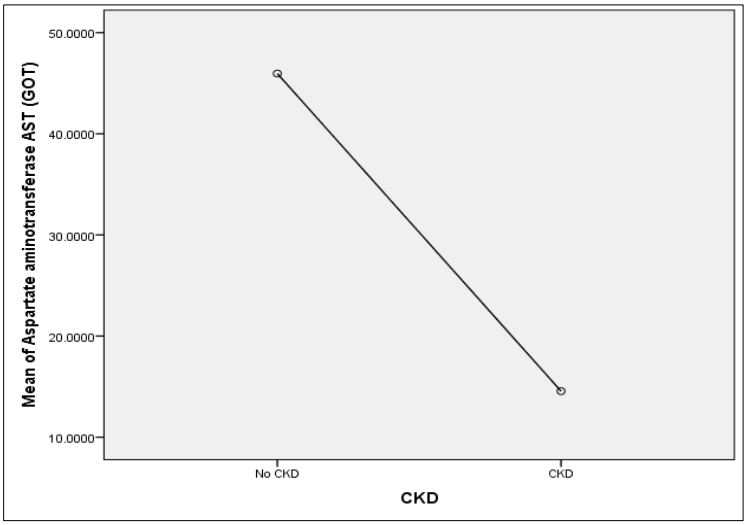


Figure 4.6 Aspartate Aminotransferase-CKD Compare Means

4.2.6. P-Value for Lab Test

On another side, the data analysis for all features that belong to lab test and CKD showed a statistical significance for the positive relationship between an individual's variables and the chances of developing CKD, whereas shown in Table 4.2 below all P-Values for T-Test for the different variables scored less than 0.05, except Sodium Serum and Uric Acid scored more than 0.05, the P-Value was 0.1694 and 0.3195

respectively, that mean Sodium Serum and Uric Acid scored are statistically non-significant.

Table 4.2 P-Value for Lab Test

No.	Variable	P_Value
1	Hemoglobin (HGB)	0.0000
2	Creatinine, serum	0.0000
3	Phosphorous	0.0000
4	Potassium, serum	0.0000
5	Sodium, serum	0.1694
6	Uric Acid	0.3195
7	Alanine Transaminase ALT	0.0004
8	Albumin	0.0004
9	Aspartate Aminotransferase AST	0.0001
10	Blood Urea Nitrogen (BUN)	0.0000
11	Calcium, serum	0.0000

4.3. Experimental Results

After visualizing, describing, and analyzing the cleaned and processed data in the previous section, this section will present the results in detail for applying different ML models on the dataset customized for this thesis by using Microsoft Azure ML (classic). Microsoft ML Studio (classic) is a Microsoft Cloud product, a drag-and-drop tool that can build, test, and deploy ML models. The user interfaces for Microsoft Azure ML (Classic) as shown in Appendix (1) below.

Microsoft Azure ML (classic) publishes models as web services, which can easily be consumed by custom apps or Business Intelligence (BI) tools such as Excel. Microsoft

ML Studio (classic) gives an interactive, visual workspace to build and test a predictive analysis model. It is worked by drag-and-drop datasets and analysis modules onto an interactive canvas, connecting them to form an experiment. When the experiment is ready, it can convert to a predictive experiment and then publish it as a web service, and others can access the model. Programming is not required. Visually connect datasets and modules to construct a predictive analysis model.

4.3.1. Decision Forest

The DF was applied to the dataset that was customized for this thesis. There are some parameters adjusted when applying the model. The number of decision trees and the maximum depth of the decision trees, when applied to the model, were 16 and 32, respectively, that selected after testing the model with various values of these two parameters, such as (8,32), (16,64), (40,69), and the accuracy was 0.932, 0.936 and 0.923, respectively.

From the following Table 4.3 below, we can notice that both training and testing accuracy are shown good results, were 0.998 and 0.936 respectively. Besides, as shown in Figure 4.7 below, an AUC of 0.964 has been obtained, and it is a sane result. In the Confusion Matrix shown in

Figure 4.8 below, the Precision, Recall, and F1 Score calculated were 0.949, 0.933, and 0.941, respectively.

Table 4.3 Decision Forest Evaluation Results

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
DF	0.998	0.936	0.949	0.933	0.941	0.964

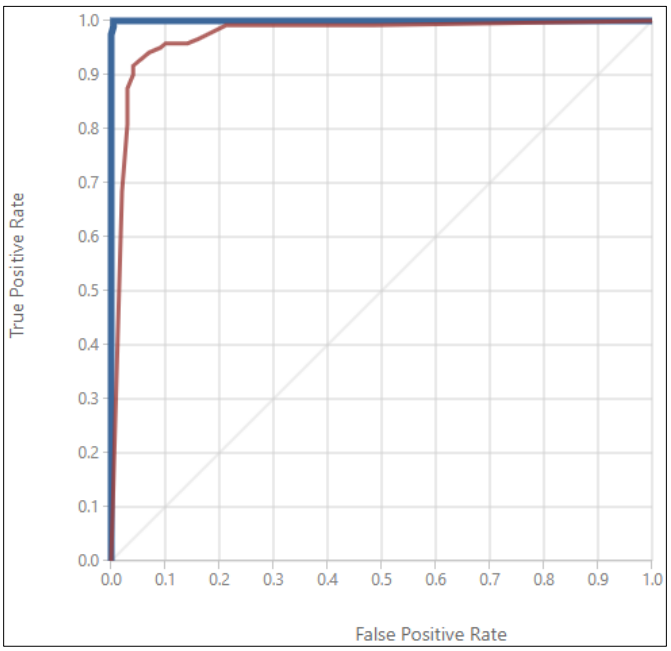


Figure 4.7 Decision Forest Area Under Curve

		Actual Values	
		Positive	Negative
Predicted Values	Positive	111	6
	Negative	9	93

Figure 4.8 Decision Forest Confusion Matrix

4.3.2. Decision Jungle

The DJ was applied to the dataset that was customized for this thesis. There are some parameters adjusted when applying the model. The number of decision DAGs, Maximum depth of the decision DAGs and Maximum width of the decision DAGs when applied to the model were 8, 32, and 128, respectively. From the following Table 4.4, we can notice that both training and testing accuracy are shown good results, which were 0.992 and 0.932, respectively. Besides, an AUC of 0.965 has been obtained, as shown in Figure 4.9, and it is a sane result. In the Confusion Matrix shown in

Figure *4.10* below, the Precision, Recall, and F1 Score calculated were 0.949, 0.925, and 0.937, respectively. By comparing DJ and DF, the results are equal in accuracy, Precision, Recall, and F1 Score, but by comparing the AUC score, that shows DJ better than DF.

Table 4.4 Decision Jungle Evaluation Result

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
DJ	0.992	0.932	0.949	0.925	0.937	0.965

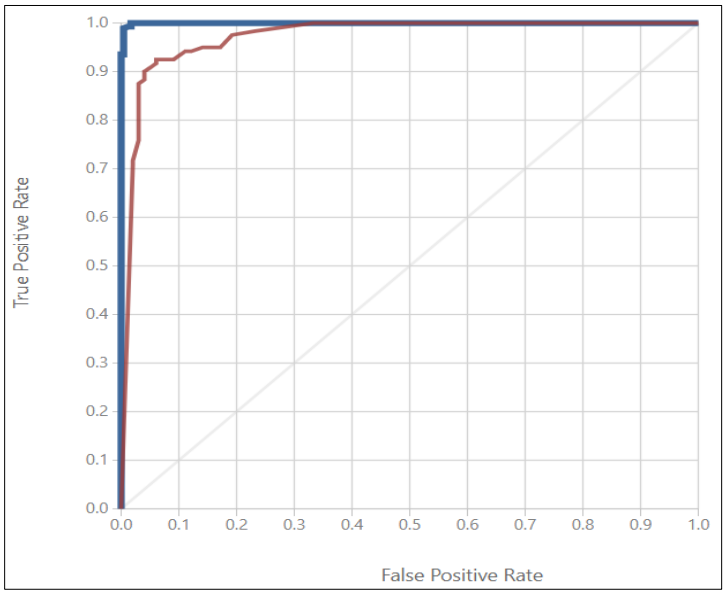


Figure 4.9 Decision Jungle Area Under Curve

		Actual Values	
		Positive	Negative
Predicted Values	Positive	111	6
	Negative	9	93

Figure 4.10 Decision Jungle Confusion Matrix

4.3.3. Support Vector Machine

The SVM was applied to the dataset that was customized for this thesis. From the following Table 4.5, we can notice that both training and testing accuracy are shown

acceptable results, were 0.906 and 0.895 respectively. Also, an AUC of 0.962 has been obtained, as shown in Figure 4.11 below, and it is a sane result. In the Confusion Matrix shown in

Figure 4.12 below, the Precision, Recall, and F1 Score calculated were 0.901, 0.908, and 0.905, respectively. From these results, applying Decision Jungle still better. Whereas the results have shown notable differences.

Table 4.5 SVM Evaluation Results

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
SVM	0.906	0.895	0.901	0.908	0.905	0.962

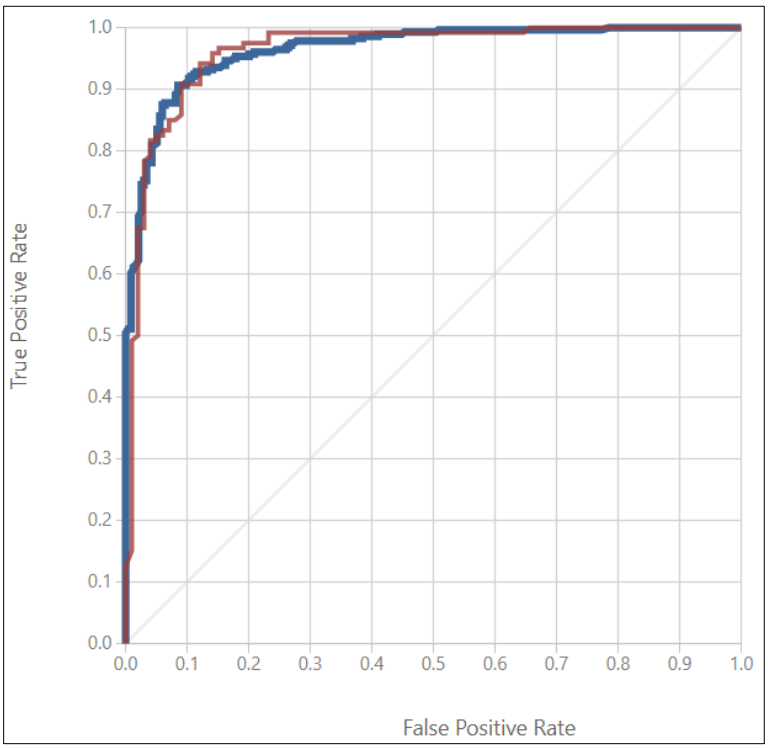


Figure 4.11 Support Vector Machine Area Under Curve

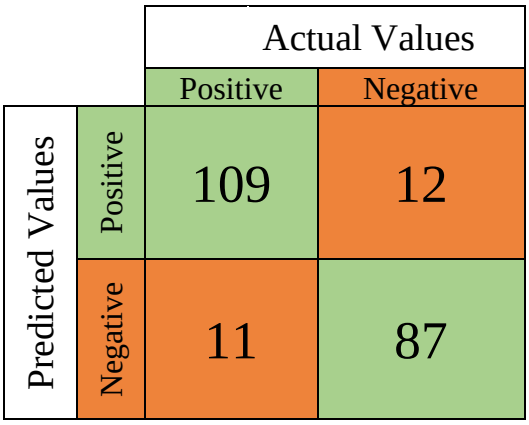


Figure 4.12 Support Vector Machine Confusion Matrix

4.3.4. Locally Deep Support Vector Machine

The LD-SVM was applied to the dataset that was customized for this thesis. There is a parameter adjusted when applying the model, the depth of the tree when applied to the model was 3. From the following Table 4.6, we can notice that both training and testing accuracy are shown good results, were 1 and 0.922 respectively. Also, an AUC of 0.958 has been obtained, as shown in Figure 4.13, and it is a sane result. In the Confusion Matrix shown in Figure 4.14 below, the Precision, Recall, and F1 Score calculated were 0.912, 0.950, and 0.931, respectively.

Also, Decision Jungle is still better in terms of results.

Table 4.6 LD-SVM Evaluation Results

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
LD-SVM	1	0.922	0.912	0.950	0.931	0.958

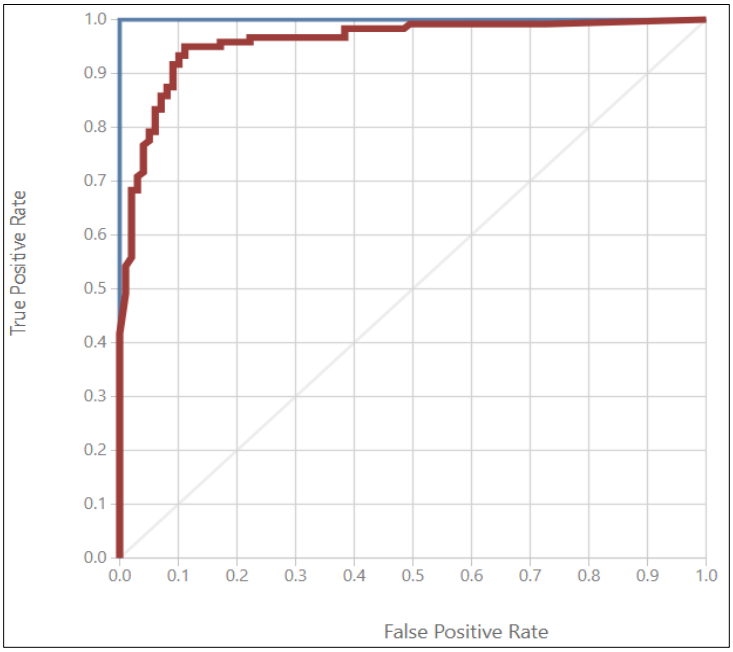


Figure 4.13 Locally Deep Support Vector Machine Area Under Curve

		Actual Values	
		Positive	Negative
Predicted Values	Positive	114	11
	Negative	6	88

Figure 4.14 Locally Deep Support Vector Machine Confusion Matrix

4.3.5. Logistic Regression

The LR applied to the dataset that was customized for this thesis. From the following Table 4.7, we can notice that both training and testing accuracy is shown acceptable results, were 0.922 and 0.918, respectively. Besides, an AUC of 0.964 has been obtained as shown in, and it is a sane result. In the Confusion Matrix shown in Figure

4.16 below, the Precision, Recall, and F1 Score calculated were 0.918, 0.933, and 0.926, respectively. LR only outperformed DJ in Recall because of increasing TP over FN in LR. However, DJ still the best.

Table 4.7 Logistic Regression Evaluation Results

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
LR	0.922	0.918	0.918	0.933	0.926	0.964

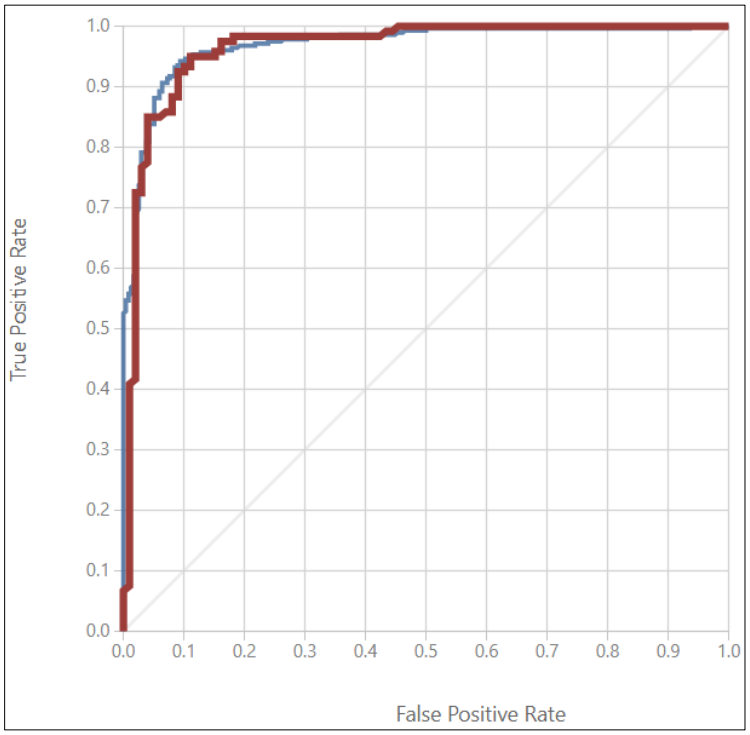


Figure 4.15 Logistic Regression Area Under Curve

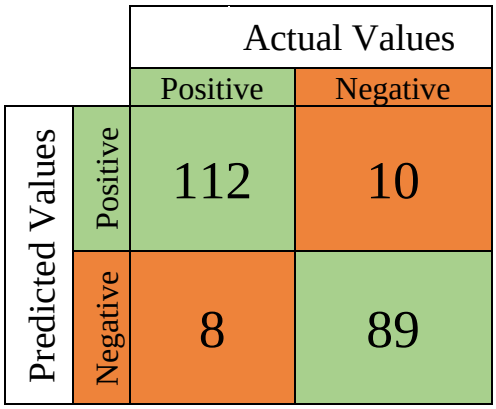


Figure 4.16 Logistic Regression Confusion Matrix

4.3.6. Boosted Decision Tree

The BDT applied to the same dataset. When applying the model, some parameters were adjusted: the maximum number of leaves per tree, and the minimum number of samples per leaf node, when applied to the model, were 20, and 10 respectively, that selected after testing the model with various values of these two parameters, such as (10,5), (30,20), (40,30), and the accuracy was 0.936, 0.941 and 0.936, respectively.

From the following Table 4.8 below, we can notice that both training and testing accuracy has shown good results, were 1 and 0.945, respectively. Besides, an AUC of 0.972 has been obtained, as shown in Figure 4.17, and it is a sane result. From the Confusion Matrix Figure 4.18 shown below, the Precision, Recall, and F1 Score calculated were 0.950, 0.950, and 0.950, respectively. With these results, the BDT outperforms the DJ.

Table 4.8 Boosted Decision Tree Evaluation Results

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
BDT	1	0.945	0.950	0.950	0.950	0.972

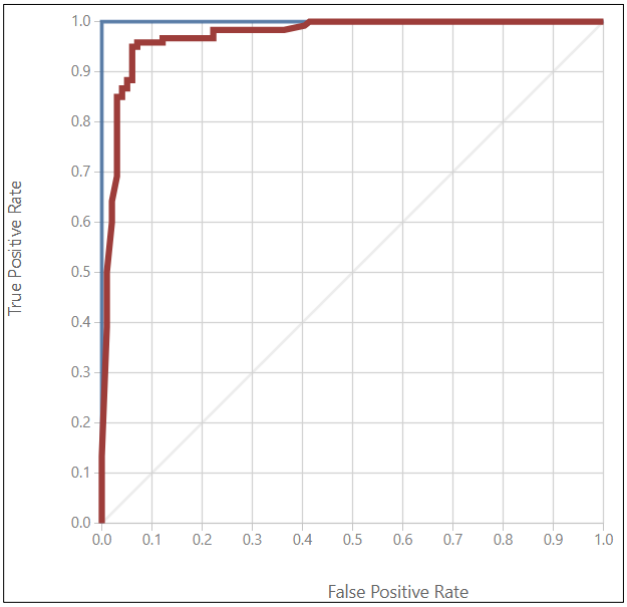


Figure 4.17 Boosted Decision Tree Area Under Curve

		Actual Values	
		Positive	Negative
Predicted Values	Positive	114	6
	Negative	6	93

Figure 4.18 Boosted Decision Tree Confusion Matrix

4.3.7. Bayes Point Machine

The BPM was applied to the same dataset. From the following Table 4.9, we can notice that both training and testing accuracy has shown good results, were 0.939 and 0.918, respectively. Also, an AUC of 0.963 has been obtained, as shown in Figure 4.19, and it

is a sane result. In the Confusion Matrix shown in Figure 4.20 below, the Precision, Recall, and F1 Score calculated were 0.918, 0.933, and 0.926, respectively. Comparing these results with previous results, the BDT results are still better.

Table 4.9 Bayes Point Machine Evaluation Results

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
BPM	0.939	0.918	0.918	0.933	0.926	0.963

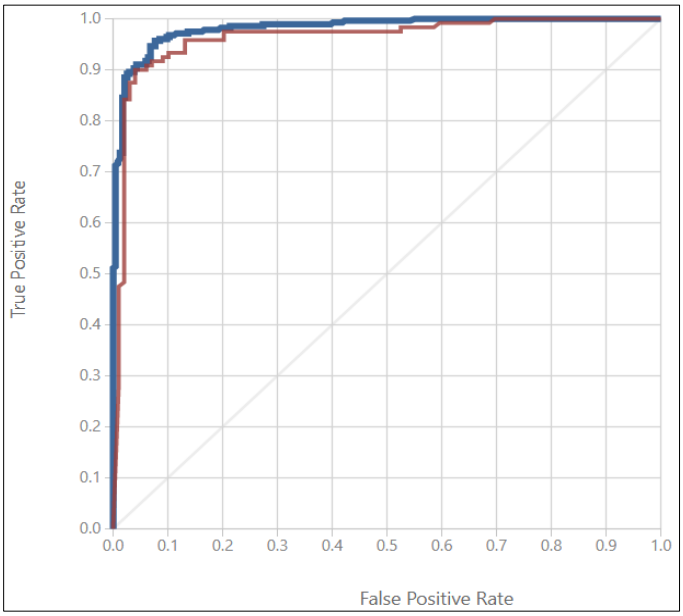


Figure 4.19 Bayes Point Machine Area Under Curve

		Actual Values: Positive	Actual Values: Negative
Predicted Values	Positive	112	10
Predicted Values	Negative	8	89

Figure 4.20 Bayes Point Machine Confusion Matrix

4.3.8. Neural Network

The NN was applied to the dataset that was customized for this thesis. In this model, 4 testing applied by a different number of hidden nodes, 20, 30,40, and 50 nodes. From the following Table 4.10, we can notice that Neural Network with 30 hidden nodes has high testing accuracy, so, for comparison with other ML models, 30 hidden nodes were adopted. Besides, the AUC for selected hidden nodes of 0.965 has been obtained as shown in Figure 4.21, and it is a sane result.

The Confusion Matrix Figure 4.22 shown below, the Precision, Recall, and F1 Score calculated were 0.898, 0.950, and 0.923, respectively.

Table 4.10 Neural Network Evaluation Results

Model	# of hidden nodes	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
neural network	20	0.928	0.904	0.884	0.950	0.916	0.963
	30	0.934	0.913	0.898	0.950	0.923	0.965
	40	0.930	0.904	0.884	0.950	0.916	0.962
	50	0.924	0.904	0.878	0.958	0.916	0.964

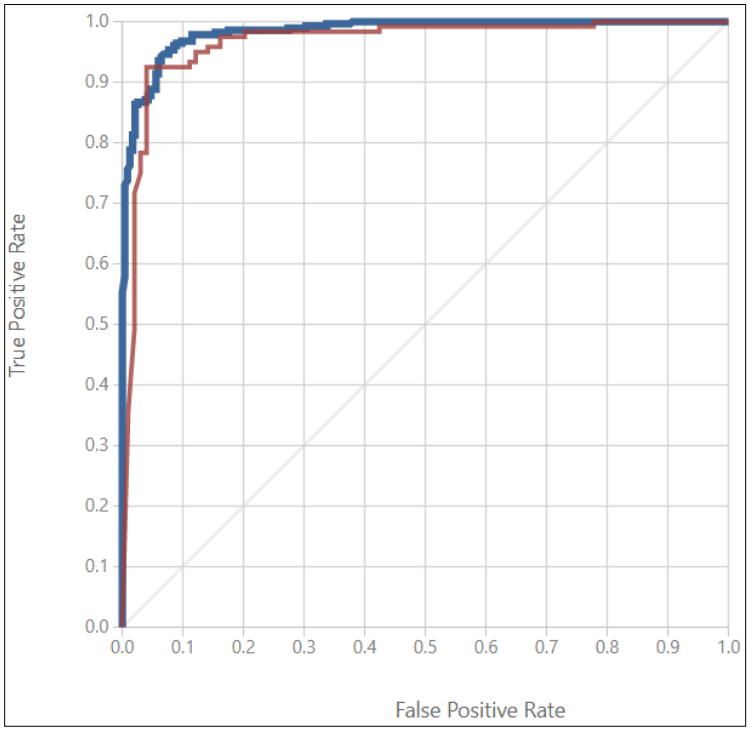


Figure 4.21 Neural Network Area Under Curve

		Actual Values	
		Positive	Negative
Predicted Values	Positive	115	14
	Negative	5	85

Figure 4.22 Neural Network Confusion Matrix

4.4. Cross Validation Result

After trying to test the result using cross validation in selected four ML models and compare the results, we noted the results are approximately close to results at the above section, as shown in the Table 4.11 below.

Table 4.11 Cross Validation Results

	Accuracy	Precision	Recall	F1 Score	AUC
BDT	0.941	0.949	0.942	0.946	0.980
SVM	0.902	0.901	0.920	0.910	0.958
DJ	0.923	0.943	0.915	0.929	0.961
DF	0.937	0.947	0.937	0.942	0.973

and BDT still has the best performance.

4.5. Using UCI Dataset in Four ML Models:

To compare the results between the local dataset and the dataset from UCI, the features must first be compared between them, as it was found that there is a significant difference in the features, the common features between the two datasets are only seven features, namely (Age, Albumin, Blood Urea Nitrogen (BUN), Creatinine serum, Sodium Serum, Potassium serum), which are less than half of the features in local datasets, and therefore it is difficult to compare them with these seven features.

To make the comparison more accurate, these seven features were taken with the addition of nine others based on previous studies.

After applying the UCI dataset in four ML models (BDT, SVM, DJ and DF), the results of the UCI were slightly better than the local dataset, but the UCI results agreed with the local dataset results that the BDT was the best model and the SVM was the lowest model in terms of accuracy, as the Table 4.12 below shows:

Table 4.12 UCI Dataset Results

	Accuracy	Precision	Recall	F1 Score	AUC
BDT	0.975	1.00	0.962	0.981	0.998
SVM	0.917	0.948	0.924	0.936	0.988
DJ	0.967	0.987	0.962	0.974	0.996
DF	0.975	0.987	0.975	0.981	0.997

4.6. CKD Web Service Building

Since the BDT model has the best performance, the web service developed according to that model. So, other models were removed from the project, as the Figure 4.23 below.

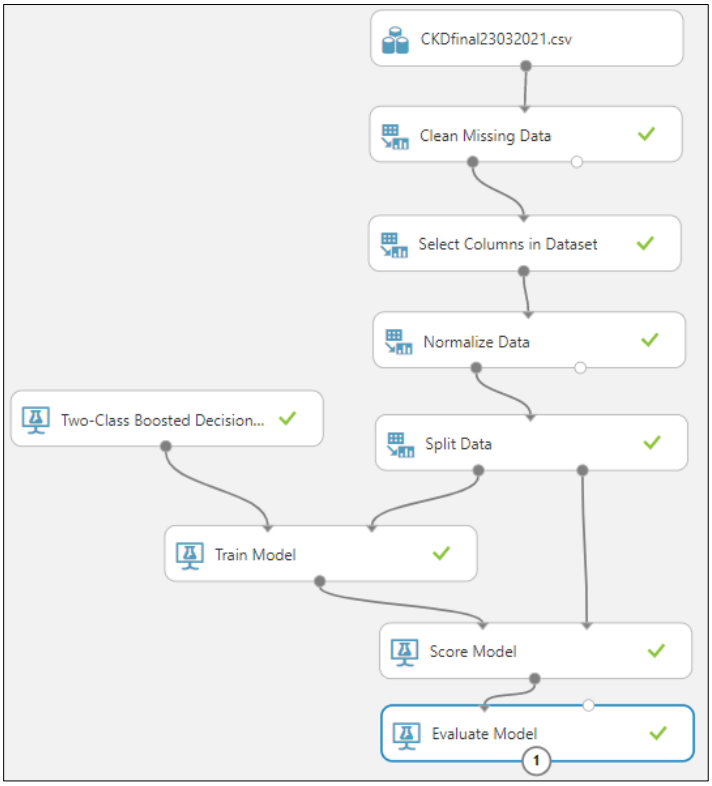


Figure 4.23 Boosted Decision Tree Model

Since the experiment finished training and testing, the selected model's mission is to generate new predictions by scoring the user's input, so some preparation needs to convert this experiment from a training experiment to a predictive experiment.

After removing unselected models, the evaluate model must remove, then define where the web service will accept input and where it generates the output, and select features as input, also must using normalization method as a method used in training mode.

The ultimate structure of the experiment is shown in the following Figure 4.24 below

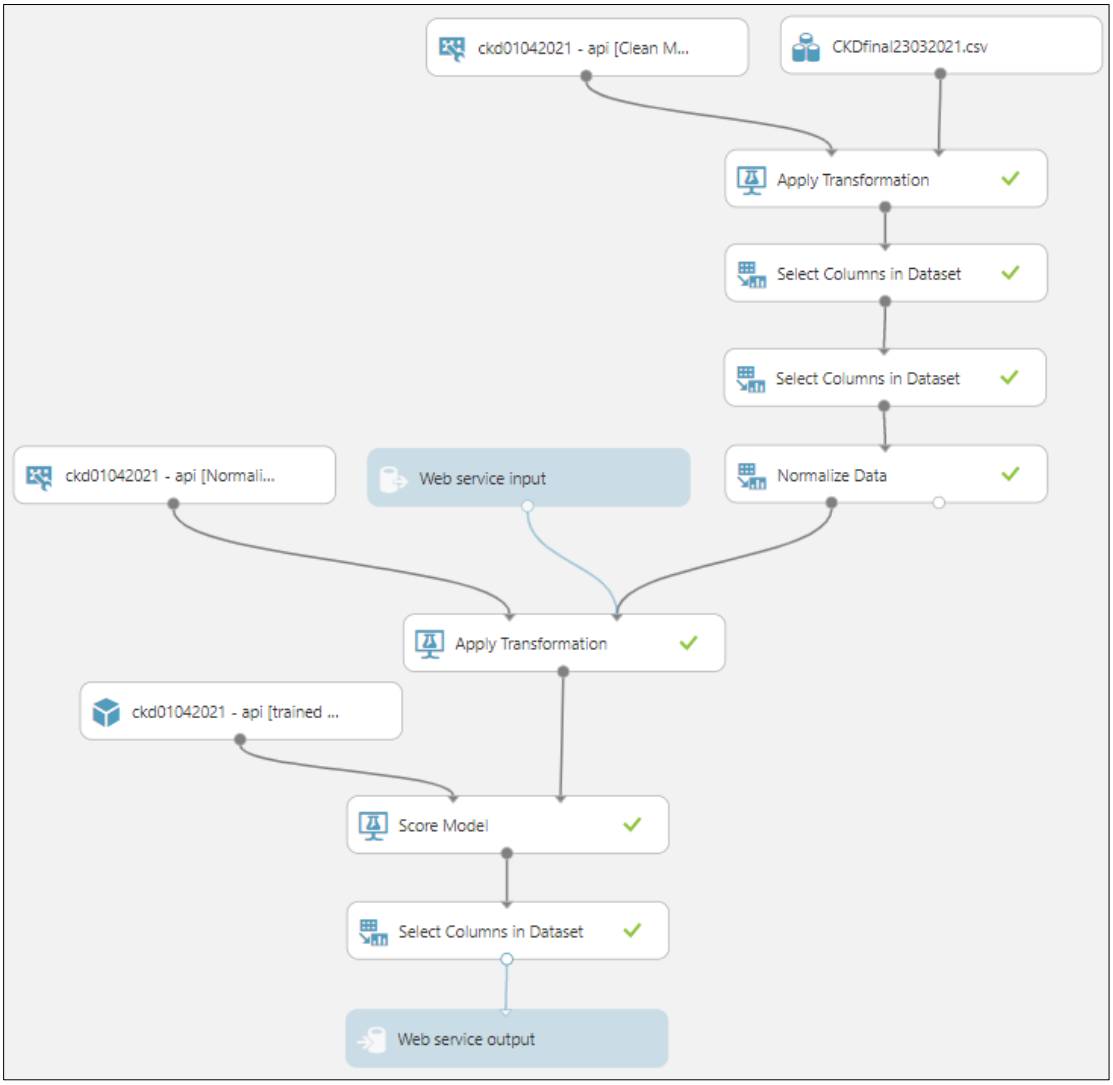


Figure 4.24 Ultimate Structure

After running the experiment, the result shows in Figure 4.25 below

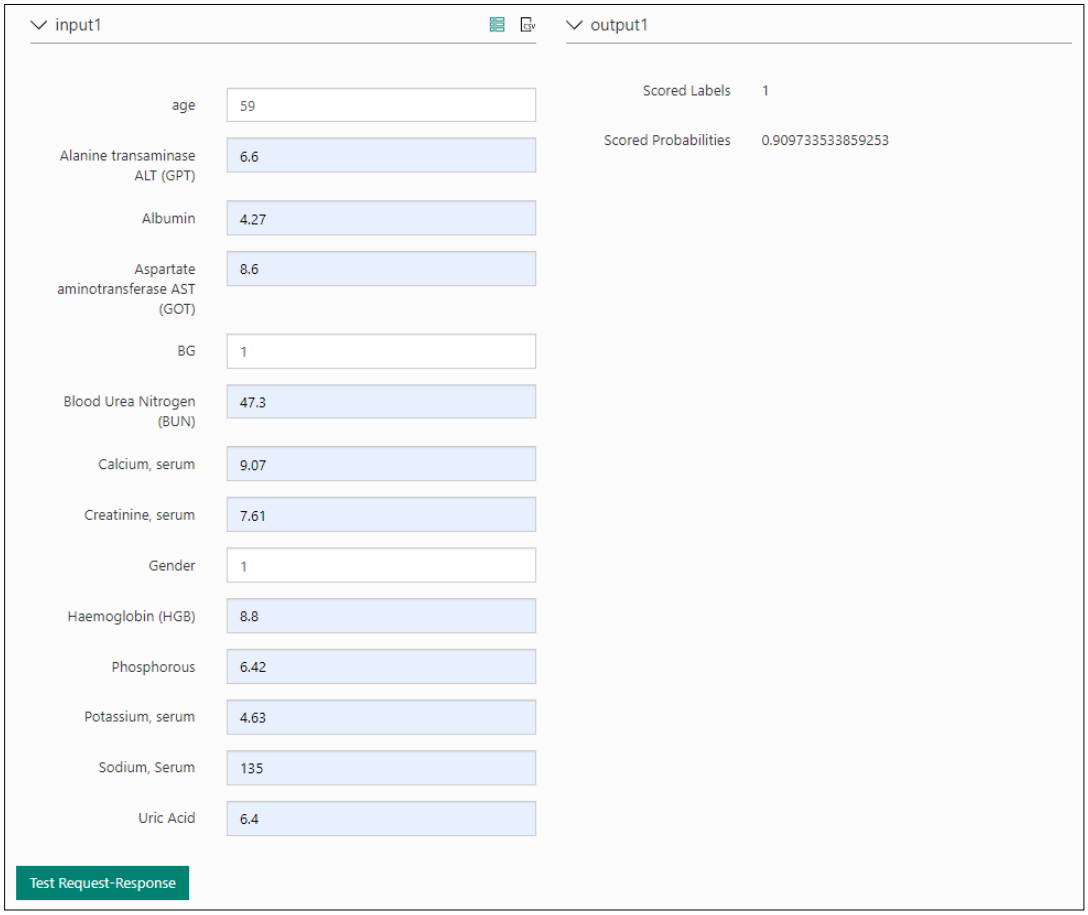


Figure 4.25 Boosted Decision Tree Web Service Form

When 30 untrained patient records were tested on a new experiment, the result was identical with records classes.

4.7. Experimental Result Summary

The above results prove the BDT model outperformed the other models such as NN with its different number of hidden nodes, DF, DJ, SVM, LDSVM, LR, and the BPM models. The BDT model with 20 Maximum number of leaves per tree and 10 Minimum

number of samples per leaf node has achieved an outstanding accuracy of up to 94.5% in predicting CKD and non-CKD cases, and it was the best of AUC, Precision, Recall, and F1 Score were 0.972, 0.950, 0.950, and 0.950 respectively as shown in the Figure 4.26. On the other hand, the worst ML model is SVM.

The above conclusion is reasonable because the BDT model an ensemble learning method, in which the coming tree corrects for the errors of the previous tree, predictions are constructed on the whole ensemble of trees together that makes the prediction.

On the other hand, S. Tekale et al. [25] compared the performance of DT and SVM. They found that SVM has the best accuracy of 96.75%, Charleonnan et al. [26] also agreed with this result, it was the best model in their experiment is SVM, this result was contrary to the results of this thesis, which is the SVM has the worst performance.

Also, all previous studies showed different results, the result was not the same; Snegha et al. [18] concluded Back Propagation neural network is more efficient than RF, with an accuracy of 98.40%, and Pandit et al. [19] concluded Random Forest model is the best performance with an accuracy of 100%.

Each prediction and ML performance depends on the size of the dataset provided (number of features and number of records) and the type of the features basically, that's why there is variation in performance between different studies. Finally, the best ML model to be used on the PMOH dataset for CKD prediction is the BDT.

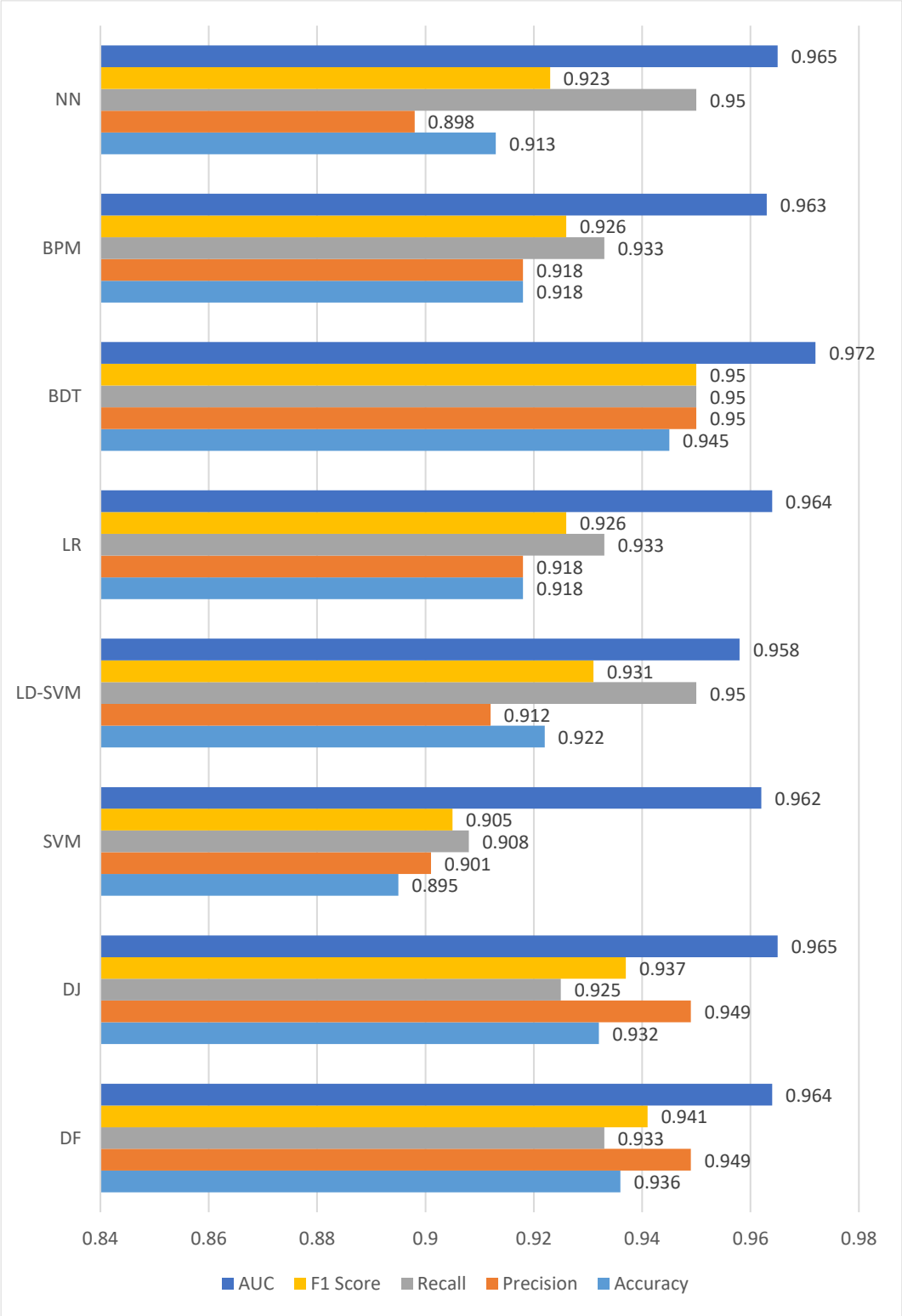


Figure 4.26 Comparison Between Machine Learning Models

4.8. Summary

In this chapter, the system procedures have been applied that have been introduced in the previous chapter. The dataset was preprocessed using the PPCA algorithm and Z-Score. Where the PPCA algorithm has been used to estimate the missing values and filling them. The Z-Score has been used to Normalization the features in the dataset. Descriptive Statistics and Visualization was presented and applied in IBM SPSS. After visualizing, describing, and analyzing the cleaned and processed data, the results were presented in detail after applying different ML models on the dataset customized for this thesis using Microsoft Azure ML. Also, described how to build CKD Web Service. Besides that, the experimental was reported result summary. In short, this chapter appeared the BDT model's ability in CKD predictions where it could outperform other applied models.

Chapter 5: Conclusion and Recommendations

5.1. Introduction

In this chapter, the conclusion and recommendations of the study will be presented. Besides that, further work and intended studies for future work will be presented as well as the strengths of the study.

5.2. Conclusion

Considering all the risks of CKD, the early stages of the disease must detect early. Chronic kidney disease can be controlled once it is detected in its early stages. Correct diagnosis also plays a significant role in reducing disease development, and stages from 3 to 5 are considered dangerous. This type of disease cannot prevent and can only cure with dialysis or kidney transplantation. Therefore, the prognosis for the disease must be very accurate. By focusing on two aspects, the first: prevention and the second: prediction, the problem can limit by reducing budget waste and improving the health status of chronic kidney patients.

This study explored how ML models can be used to classify the possibility of a CKD diagnosis. The research methodology in this study passed through adapted to the context of the problem so that different logically organized stages were taken; Data collection, Preprocessing, Applied ML, Evaluation, and Selection, which allowed the construction of a model capable of classifying the possibility of a CKD diagnosis.

The experiments have been applied some ML techniques for classification and prediction of CKD using Patients' treatment history data taken from the PMOH repository. Moreover, it used eight ML algorithms such as the DF, DJ, SVM, LDSVM, LR, BDT, BPM, NN. These ML models' performances will be compared to decide the best classifier model in predicting CKD for the given dataset. The results of the models that were used showed the superiority of the model BDT. It outperformed other models with the following results: Accuracy=0.945, Precision= 0.95, Recall= 0.95, F1Score= 0.95, and AUC= 0.972. Given the results obtained, the BDT classifier to identify CKD people proposed as a valuable tool to help the CKD specialist detect the disease and confirm their diagnosis. In particular, the system's level of success in diagnosis is very high, depending on the patient's symptoms and test results that are analyzed. Furthermore, many patients can avoid kidney dialysis or search for any donor to perform the kidney transplant. The applications of these ML models aspire to improve the resulting environment of the CKD health system in Palestine, with a methodology adapted to the case study and a production architecture proposal for the model with Microsoft Azure tools in a form that allows satisfying the future scalability of the solution. Furthermore, the methodology that depends on applying ML models could apply to clinical data of other diseases and pathologies in real medical diagnosis.

5.3. Recommendations

The ML models showed good results, but more accurate results can be obtained if we take some recommendations, such as increasing the number of features that directly affect the diagnosis of the disease based on the recommendations of specialists and

increasing the number of records. In addition, the researcher recommends using more various data in terms of place of residence, including the Gaza and Jerusalem area.

It was noticed that there is a high percentage of missing data, and to solve this problem, it is recommended that the PMOH pay more attention to the quality of the data, by adopting more than one strategy, one of which is focusing on developing health personnel by giving them training courses and workshops on topics that show the importance of the data and finding the best ways to enter it on applications used in hospitals.

Finding a department specialized in tracking data, trying to solve data problems in a fast and organized manner, and monitoring the input process to track the sources of defects.

5.4. Future Works

As a short-term future work, we can continue with the expansion of the database would reduce the limited generalization error for the model and at the same time allow the severity of the disease to be detected. This model can be refined as the size and quality of the data increase.

This work can open various studies from other research fields; such as studies economic around the impact of obtaining a diagnosis in less time to treat the disease in early stages, reducing costs in the health care system. This research, in principle, will help interested professionals during their training phase to understand virtual CKD problems. On the other hand, to deliver this service to the most significant number of people,

monitor their health condition, and collect new live data for developing new ML models, the service can be developed for mobile applications.

5.5. Strengths of the Study

In this study, the data have been collected from Palestinian MOH, which means the data is from our population, and the information is updated. Also, the dataset in this study will be 732 records to deliver a more reliable study. The important factor that no research depends on ML to predict and classify CKD in Palestine. ML techniques have been used for the first time to predict and classify CKD local datasets. We aim to produce a classification and prediction result of CKD that can be considered and developed in the future to be a decision support system in this field.

5.6. Summary

In this chapter, the summary of the conclusions and recommendations of the study were presented. Besides that, further work and intended studies for future work were presented as well as the strengths of the study.

Bibliography

[1] "Chronic Kidney Disease (CKD) Symptoms, Treatment, Causes & Prevention - American Kidney Fund (AKF)." https://www.kidneyfund.org/kidney-disease/chronic-kidney-disease-ckd/ (accessed Aug. 01, 2020).

[2] "World Kidney Day: all you need to know! - World Kidney Day." https://www.worldkidneyday.org/world-kidney-day-all-you-need-to-know/ (accessed Jul. 30, 2020).

[3] 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.

[4] "Complications of chronic kidney disease | CKD - American Kidney Fund (AKF)." https://www.kidneyfund.org/kidney-disease/chronic-kidney-disease-ckd/complications/#metabolic-acidosis (accessed Aug. 09, 2020).

[5] "Stages of Chronic Kidney Disease (CKD) - American Kidney Fund (AKF)." https://www.kidneyfund.org/kidney-disease/chronic-kidney-disease-ckd/stages-of-chronic-kidney-disease/ (accessed Aug. 17, 2020).

[6] “Chronic Kidney Disease | Kidney Failure | Stages, Symptoms and Treatment | Patient.” https://patient.info/kidney-urinary-tract/chronic-kidney-disease-leaflet#nav-5 (accessed Oct. 11, 2020).

[7] “Chronic Kidney Disease Basics | Chronic Kidney Disease Initiative | CDC.” https://www.cdc.gov/kidneydisease/basics.html (accessed Oct. 16, 2020).

[8] P. Stenvinkel, D. Fouque, and C. Wanner, “Life/2020-The future of kidney disease,” *Nephrology Dialysis Transplantation*, vol. 35, no. Supplement_2. Oxford University Press, pp. II1–II3, Mar. 01, 2020, doi: 10.1093/ndt/gfaa028.

[9] G. Chironda and B. Bhengu, “Ethical, legal and cultural implications in the management of chronic kidney disease (CKD) patients: a critical review of literature,” *Transactions of the Royal Society of South Africa*, vol. 71, no. 2. Taylor and Francis Ltd., pp. 137–143, May 03, 2016, doi: 10.1080/0035919X.2016.1146928.

[10] Q. Yuan *et al.*, “Role of artificial intelligence in kidney disease,” *Int. J. Med. Sci.*, vol. 17, no. 7, pp. 970–984, 2020, doi: 10.7150/ijms.42078.

[11] A. J. Aljaaf *et al.*, “Early Prediction of Chronic Kidney Disease Using Machine

Learning Supported by Predictive Analytics," Sep. 2018, doi: 10.1109/CEC.2018.8477876.

[12] "التقارير الإحصائية." http://site.moh.ps/index/Books/BookType/2/Language/ar (accessed Nov. 04, 2020).

[13] T. A. V. Pham, "Validation of the salivary urea and creatinine tests as screening methods of chronic kidney disease in Vietnamese patients," *Acta Odontol. Scand.*, vol. 75, no. 8, pp. 551–556, Nov. 2017, doi: 10.1080/00016357.2017.1356467.

[14] "CKD a Leading Cause of Morbidity and Mortality Worldwide - Renal and Urology News." https://www.renalandurologynews.com/home/news/nephrology/chronic-kidney-disease-ckd/ckd-a-leading-cause-of-morbidity-and-mortality-worldwide/ (accessed Apr. 22, 2021).

[15] "Machine Learning in Healthcare: Examples, Tips & Resources | UIC Online." https://healthinformatics.uic.edu/blog/machine-learning-in-healthcare/ (accessed Apr. 21, 2021).

[16] P. Chittora *et al.*, “Prediction of Chronic Kidney Disease -A Machine Learning perspective,” *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3053763.

[17] M. Almasoud and T. E. Ward, “Detection of chronic kidney disease using machine learning algorithms with least number of predictors,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, pp. 89–96, 2019, doi: 10.14569/ijacsa.2019.0100813.

[18] J. Snegha, V. Tharani, S. D. Preetha, R. Charanya, and S. Bhavani, “Chronic Kidney Disease Prediction Using Data Mining,” Feb. 2020, doi: 10.1109/ic-ETITE47903.2020.482.

[19] A. K. Pandit, R. P. Kushwaha, and I. Kumar, “EasyChair Preprint Chronic Kidney Disease Prediction by Machine Learning Chronic Kidney Disease Prediction by Machine Learning,” EasyChair, Jul. 2020.

[20] M. Alloghani, D. Al-Jumeily, A. Hussain, P. Liatsis, and A. J. Aljaaf, “Performance-based prediction of chronic kidney disease using machine learning for high-risk cardiovascular disease patients,” in *Studies in Computational Intelligence*, vol. 855, Springer Verlag, 2020, pp. 187–206.

[21] P. Chittora, G. K. Ameta, P. Chakrabarti, and G. Kumawat, “(PDF) Analysis of Chronic Kidney Disease (CKD) using supervised machine learning classifiers and curve fitting,” *International Journal of Advanced Science and Technology*, 2020. https://www.researchgate.net/publication/344666444_Analysis_of_Chronic_Kidney_Disease_CKD_using_supervised_machine_learning_classifiers_and_curve_fitting (accessed Mar. 08, 2021).

[22] P.Swathi Baby and T. Panduranga Vital, “Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms,” *Int. J. Eng. Res.*, vol. V4, no. 07, 2015, doi: 10.17577/ijertv4is070234.

[23] M. D. Başar and A. Akan, “Chronic kidney disease prediction with reduced individual classifiers,” *Istanbul Univ. - J. Electr. Electron. Eng.*, vol. 18, no. 2, pp. 249–255, 2018, doi: 10.26650/electrica.2018.99255.

[24] I. A. Pasadana *et al.*, “Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques,” in *Journal of Physics: Conference Series*, Sep. 2019, vol. 1255, no. 1, p. 012024, doi: 10.1088/1742-6596/1255/1/012024.

[25] S. Tekale, P. Shingavi, and S. Wandhekar, “Prediction of Chronic Kidney Disease Using Machine Learning Algorithm,” *IJARCCE*, vol. 7, no. 10, pp. 92–

96, Oct. 2018, doi: 10.17148/IJARCCE.2018.71021.

[26] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, “Predictive analytics for chronic kidney disease using machine learning techniques,” *2016 Manag. Innov. Technol. Int. Conf. MITiCON 2016*, pp. MIT80–MIT83, 2017, doi: 10.1109/MITICON.2016.8025242.

[27] W. H. S. D. Gunarathne, K. D. M. Perera, and K. A. D. C. P. Kahandawaarachchi, “Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for Chronic Kidney Disease (CKD),” in *Proceedings - 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering, BIBE 2017*, Jul. 2017, vol. 2018-January, pp. 291–296, doi: 10.1109/BIBE.2017.00-39.

[28] and A. S. Sahil Sharma, Vinod Sharma, “(PDF) Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis,” *International Journal of Modern Computer Science (IJMCS)*, 2016. https://www.researchgate.net/publication/304642548_Performance_Based_Evaluation_of_Various_Machine_Learning_Classification_Techniques_for_Chronic_Kidney_Disease_Diagnosis (accessed Mar. 09, 2021).

[29] K. R. Anantha Padmanaban and G. Parthiban, “Applying machine learning techniques for predicting the risk of chronic kidney disease,” *Indian J. Sci. Technol.*, vol. 9, no. 29, 2016, doi: 10.17485/ijst/2016/v9i29/93880.

[30] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, “A machine learning methodology for diagnosing chronic kidney disease,” *IEEE Access*, vol. 8, pp. 20991–21002, 2020, doi: 10.1109/ACCESS.2019.2963053.

[31] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, “Predictive analytics for chronic kidney disease using machine learning techniques,” in *2016 Management and Innovation Technology International Conference, MITiCON 2016*, Sep. 2017, pp. MIT80–MIT83, doi: 10.1109/MITICON.2016.8025242.

[32] A. Al Imran, M. N. Amin, and F. T. Johora, “Classification of Chronic Kidney Disease using Logistic Regression, Feedforward Neural Network and Wide Deep Learning,” Mar. 2019, doi: 10.1109/CIET.2018.8660844.

[33] S. Sossi Alaoui, B. Aksasse, and Y. Farhaoui, “Statistical and predictive analytics of chronic kidney disease,” in *Advances in Intelligent Systems and Computing*, Jul. 2019, vol. 914, pp. 27–38, doi: 10.1007/978-3-030-11884-6_3.

[34] Parul Sinha and Poonam Sinha, “Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM,” *Int. J. Eng. Res.*, vol. V4, no. 12, Dec. 2015, doi: 10.17577/ijertv4is120622.

[35] Y. Zhou and Z. Song, “Effectiveness analysis of machine learning in education big data,” in *Journal of Physics: Conference Series*, Nov. 2020, vol. 1651, no. 1, p. 12105, doi: 10.1088/1742-6596/1651/1/012105.

[36] “Liver function tests - Mayo Clinic.” https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595 (accessed Apr. 13, 2021).

[37] “ML Studio (classic): Normalize Data - Azure | Microsoft Docs.” https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/normalize-data (accessed Mar. 13, 2021).

[38] “ML Studio (classic): Filter Based Feature Selection - Azure | Microsoft Docs.” https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/filter-based-feature-selection (accessed Mar. 14, 2021).

[39] “Cross-Validation in Machine Learning | by Prashant Gupta | Towards Data

Science.” https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f (accessed Aug. 02, 2021).

[40] “ML Studio (classic): Initialize Classification Models - Azure | Microsoft Docs.” https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-classification (accessed Mar. 15, 2021).

[41] “ML Studio (classic): Multiclass Decision Jungle - Azure | Microsoft Docs.” https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-jungle (accessed Apr. 11, 2021).

[42] “ML Studio (classic): Two-Class Support Vector Machine - Azure | Microsoft Docs.” https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-support-vector-machine (accessed Apr. 11, 2021).

[43] “ML Studio (classic): Two-Class Neural Network - Azure | Microsoft Docs.” https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-neural-network (accessed Apr. 11, 2021).

[44] “ML Studio (classic): Two-Class Locally Deep SVM - Azure | Microsoft Docs.”

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-locally-deep-support-vector-machine (accessed Apr. 10, 2021).

[45] "ML Studio (classic): Two-Class Logistic Regression - Azure | Microsoft Docs." https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-logistic-regression (accessed Apr. 11, 2021).

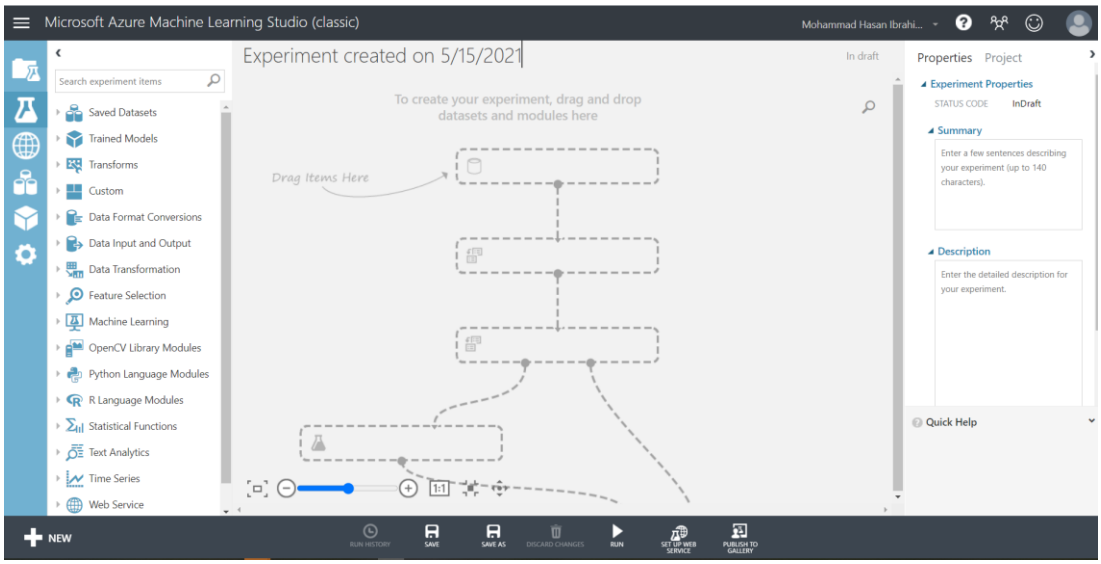
[46] "ML Studio (classic): Two-Class Boosted Decision Tree - Azure | Microsoft Docs." https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-boosted-decision-tree#bkmk_research (accessed Apr. 11, 2021).

[47] "ML Studio (classic): Two-Class Bayes Point Machine - Azure | Microsoft Docs." https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-bayes-point-machine (accessed Apr. 11, 2021).

[48] L. Ray, R. Sarangi, A. Chatterjee, S. Ganguly, and S. Nanda, "A comparative study of serum aminotransferases in chronic kidney disease with and without end-stage renal disease: Need for new reference ranges," *Int. J. Appl. Basic Med. Res.*, vol. 5, no. 1, p. 31, 2015, doi: 10.4103/2229-516x.149232.

Appendices

Appendix (1): Microsoft Azure Machine Learning Studio (classic)



الملخص

تعتبر الأمراض المزمنة من أكثر الأمراض استهلاكا لميزانيات الرعاية الصحية وتتميز بصعوبة علاجها أو استحالة ذلك. مع ارتفاع نسبة المصابين و الوفيات بسبب مرض الكلى المزمن عالميا، اصبح يعتبر هذا المرض مشكلة صحية عالمية. ومن معضلات مرض الكلى المزمن، عدم وجود أعراض واضحة خلال المراحل المبكرة من المرض، وغالبًا لا يلاحظ تطورالمرض الا في المراحل المتأخرة.

مرض الكلى المزمن هو أحد الأمراض التي ليس لها علاج حتى الآن ، وتكلفة العلاج مرتفعة. الوقاية والتنبؤ هما طريقتان يمكن من خلالهما تقليل إهدار الميزانية وتحسين صحة مرضى الكلى المزمن. يساعد الاكتشاف المبكر لمرض الكلى المزمن بتلقي العلاج في الوقت المناسب لمنع تطور المرض الى مراحل متقدمة يصعب علاجها. يمكن لنماذج التعلم الآلي مساعدة المختصين بشكل فعال على تحقيق هذا الهدف نظرًا لأدائهم السريع والدقيق في التعرف على مرضى الكلى المزمن.

في هذه الدراسة ، نقترح منهجية التعلم الآلي لتشخيص مرض الكلى المزمن. تقنيات تعلم الالة المستخدمة لتصنيف مرض الكلى المزمن والتنبؤ به باستخدام بيانات تاريخ علاج المرضى التي تم جمعها من مستودع بيانات وزارة الصحة الفلسطينية . حيث وجد أن عينات مجموعة البيانات كانت غير مهيئة لعمل النماذج ،فعُمل على تمريرها في مرحلة المعالجة المسبقة بثلاث خطوات وهي .cleaning, transformation, and feature selection :بعد الانتهاء من مرحلة المعالجة المسبقة لمجموعة البيانات تم تطبيق ثماني خوارزميات تعلم اله ؛ Decision Forest (DF), Decision Jungle (DJ), Support Vector Machine (SVM), Neural

Network (NN), Locally Deep Support Vector Machine (LD-SVM), Logistic Regression (LR), Boosted Decision Tree (BDT) and Bayes Point Machine (BPM).ستتم مقارنة أداء نماذج تعلم الالة هذه لتحديد أفضل نمـــوذج فـــي توقـــع مرض الكلى المزمن .

تعتمد طريقة البحث على بعض من الخطوات ، بدءًا من جمع البيانات ، والمعالجـــة المســـبقة للبيانات ، وتحديد عوامل التنبؤ والتصنيف ، وسيتم تطبيق مجموعتين بيانات، التدريب والاختبار ، وتنفيذ نماذج تعلم الالة المختلفة باستخدام هذه المجموعتين البيانات لتصنيف مـــرض الكلـــى المزمن والتنبؤ به.

تم تطبيق هذه الخطوات بالكامل على نظام .Microsoft Azure ML Classicأظهرت نتائج النماذج المستخدمة تفوق نموذج BDT على النماذج الأخرى بالنتـــائج التاليـــة: Accuracy = 0.945، Precision = 0.95، Recall = 0.95، F1Score = 0.95، AUC = 0.972 تشير النتائج إلى أن استخدام تقنية ML لبيانات مرضى الكلى المزمنة يعطي ميزة كبيرة للتنبـــؤ بالمرض قبل أن يتطور المرض إلى مراحل متقدمة يصعب التعامل معها ، ويمكن للعديد مـــن المرضى تجنب عملية غسيل الكلى أو البحث عن أي متبرع لإجراء عملية زرع الكلى.