Arab American University

Faculty of Graduate Studies

Colorectal Cancer Risk Factors' Assessment in Palestine: A framework for prediction tool

By

Mohammad Ali Zedan

Supervisor

Dr. Shahenaz Najjar

Co- Supervisor

Prof. Mohammed Awad

This Thesis was submitted in Partial Fulfillment of the Requirements for the Master's Degree in Health Informatics.

September 2020

Colorectal Cancer Risk Factors' Assessment in Palestine: A Framework for Prediction Tool

By

Mohammad Ali Zedan

This thesis was defended successfully on 26/09/2020 and approved by:

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

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 for them.

Acknowledgements

I feel very grateful for my parents, Reem Daghlas, my daughter (Joud) and Dr. Saba Jarrar who supported me morally. Also, special gratitude to my supervisors Dr. Shahenaz Najjar and Prof. Mohammed Awad who guided me to achieve the goals of the study. In addition, I'd like to thank the National Cancer Registry and Al Quds University for their cooperation. Finally, I won't forget Mr. Yousef Kmail, Mr. Ali Kmail, Miss. Karoleen and Miss. Hiba for their remarkable support.

Abstract

Healthcare is considered as one of the fields that produce tremendous amount of data, and this produced data will be useless if useful patterns are not extracted and used in a proper way. Generally, different types of cancer forms about 14% of mortality in Palestine, and Colorectal Cancer (CRC) prevalence especially scored 15% among men and 14.6% among women of all cancer types. On the other hand, CRC incidence depends relatively on behavioral risk factors that might increase CRC incidence and preventive factors that could decrease CRC incidence. Therefore, this research was carried out in order to spot the behavioral risk factors that affected Palestinian reported CRC cases and to make use of Machine Learning (ML) tools which might be used in CRC prediction, where the use of public CRC classification and prediction tool based on accurate ML tools will help individuals in tuning their behavioral CRC risk factors and enhance their engagement with their own health. In this research, two datasets were collected and analyzed two different CRC datasets, where one obtained from the National Cancer Registry of the Palestine Ministry of Health and the other dataset collected from the database of Al Quds University. The study found that behavioral the most important risk factors to consider are age, past medical history, diet behaviors, physical activity, and obesity. Consequently, different machine learning tools were applied to classify and predict CRC risk factors. In this research, machine learning tools were used to support medical decisions by combining intelligent computational systems with medical datasets. The task is to classify or recognize a different patterns in medical diagnoses to determine medical CRC risk factors. Decision Tree (DT), K-Nearest Neighbor (KNN), Support Victor Machine (SVM) and Artificial Neural Network (ANN) tools were applied. Therefore, this study determines a technique that can be used for diagnosing CRC risk factors. The Artificial Neural Network (ANNs) model outperforms other ML models with better accuracy in the two

collected datasets and it is the best. The ANNs model was applied to the two datasets. Then models' performance was evaluated using different metrics such as the overall accuracy, recall, specificity, and AUC, and others. The obtained results show that the ANNs model outperformed all models applied in this thesis in the classification and prediction of CRC risk factors and their types. Finally, the study also found a crucial need to promote CRC preventive factors such as CRC screening and individuals' awareness about CRC.

Keywords

Colorectal Cancer, CRC, Colon Cancer, Data Mining, Artificial Intelligence, Risk Factors, Palestine, Decision Tree, Support Victor Machine, Artificial Neural Networks, Logistic Regression, K-Nearest Neighbor.

Table of Contents

List of Figures

List of Tables

Table of Abbreviations

No	Abbreviation	Meaning
1	CRC	Colorectal Cancer
2	PA	Physical Activity
3	FBOT	Fecal Occult Blood Test
4	DM	Diabetes-Millets
5	SVM	Support Victor Machine
6	MET	Metabolic Equivalents
7	WHO	World Health Organization
8	NSAID	Nonsteroidal Anti-Inflammatory Drugs
9	BMI	Body Mass Index
10	ML	Machine Learning
11	CA	Cancer
12	AI	Artificial Intelligence
13	ANN	Artificial Neural Network
14	KNN	K-Nearest Neighbor
15	DT	Decision Tree
16	SVM	Support Victor Machine
17	MoH	Ministry of Health
18	AUC	Area Under Curve
19	CVD	Cardiovascular Disease
20	SMO	Sequential Minimal Optimization

21	PSO	Particle Swarm Optimization
22	MRI	Magnetic resonance imaging
23	CT	Computed Tomography
24	PET	Positron Emission Tomography
25	SPECT	Single Photon Emission Computed Tomography
26	MARS	Multivariate Adaptive Regression Splines
27	RF	Random forest
28	ELM	Extreme Learning Machine
29	ROC	Receiver Operating Characteristic
30	TPR	True Positive Rate
31	FPR	False Positive Rate
32	MI	Mutual Information

Chapter One

Introduction

1.1. General Introduction

Generally, cancer is defined as abnormal malignant or tumor cells that grow uncontrollably; these abnormal cells can also infiltrate and sneak into the normal body cells. Further, Cancer and the forming cells are named with the name of the body organ or tissue they are originated from, for example, Colorectal Cancer, Breast Cancer and Leukemia. Colorectal Cancer (CRC) is a term for the cancer that initiates either in the colon of the rectum; therefore, colon cancer and rectal cancer are also names for other names for CRC depending on where the cancer starts. Most CRCs form with Polyps growth on the lining of the rectum of the colon. Consequently, some Polyps may change into cancer over the years, but not all Polyps. The chance of Polyps to be transformed into a cancerous tissue depends on the type of the Polyp itself. (American Cancer Society, 2018)

However, there are two main categories of Polyps; these categories are:

1- Adenomatous Polyps, these Polyps are more likely to change into cancer; thus, it Adenomatous Polyps are also called precancerous condition.
2- Hyperplastic and Inflammatory Polyps, these Polyps are more common rather than Adenomatous Polyps and they are not cancerous.

On the other hand, the probability of Polyps to be developed into cancer also increase when the Polyp is larger than one centimeter, when more than two Polyps exist, and when dysplasia is noticed in the Polyp itself after removal. (American Cancer Society, 2018).

CRC is ranked as the third prevalent cancer as well as the fourth because of cancer related mortality worldwide. However, most CRC cases diagnosed and detected in Western countries and the incidence rate of CRC is increasing from year to year. (Mármol , et al., 2017)

Furthermore, risk factors for suffering from CRC can be categorized into modifiable and non-modifiable risk factors. The modifiable factors can be modified and controlled by individuals such as Obesity, Physical Activity (PA), Diet, Smoking, Alcohol consumption. On the other hand, the non-modifiable risk factors include the factors that cannot be changed by individuals such as Gender, Race, Ethnicity, Family History, Inflammatory Bowel Disease (IBD), Diabetes and Abdominal Radiation. (Rawla, et al., 2019).

On the contrary, some factors that could protect individuals from suffering from Colorectal Cancer and lead to a decreased CRC incidence rate in the future, these factors are named preventive factors such as CRC screening and the Fecal Occult Blood Test (FOBT). However, CRC screening is adopted with a clear policy and guidelines in most European Union (EU) countries and the United States of America. (Qumseya, et al., 2014)

In addition, Qumseya B. et al. (2014) conducted a cross-sectional among Palestinian adults to further analyze CRC screening situation; their study showed that 10% of participants carried FOBT, and only 7% had Colonoscopy screening. Finally, they concluded that the core factors which affect CRC screening level in Palestine are religious, traditional and cultural factors, in addition to lack of CRC screening knowledge.

Consequently, disease risk factors scoring and assessment tools may help individuals to measure and tune the modifiable CRC risk factors; such tools will also motivate individuals to adopt

healthier lifestyles. Also, simple risk scoring and assessment tools provides information that will promote and encourage individuals to adopt preventive CRC risk factors and other factors which may decrease CRC incidence (Miller et al., 2020).

These risk factors can be used to decide by classifying and predicting CRC occurrence and the future behavior of the patients.

On the other hand, Artificial Intelligence (AI) is a term to define the application of pre-programmed machines that will act as humans do, trying to emulate human cognition. AI provides an efficient and accurate result by avoiding human-made errors; thus, AI is considered as one of the valuable advancements of the upcoming technology in cancer studies. (Patel, et al., 2020). Artificial Intelligence plays its vital role in healthcare in generally and in medicine especially, where the AI pre-programmed machines showed their excellent data analysis and accurate pattern identification and recognition, where both are very difficult for a human being to perform (Patel, et al., 2020).

From another side, Artificial Intelligence (AI) played an essential role in healthcare data analysis, data classification, and prediction. AI could be utilized in cancer studies since cancer registries produce huge amount of data. The pattern extraction will lead to rich information to be used further to enhance healthcare services. As shown in figure (1), applications in cancer studies start with data collection, data mining, deep learning, machine learning and of AI outcome for prediction (Marcu, et al., 2019).

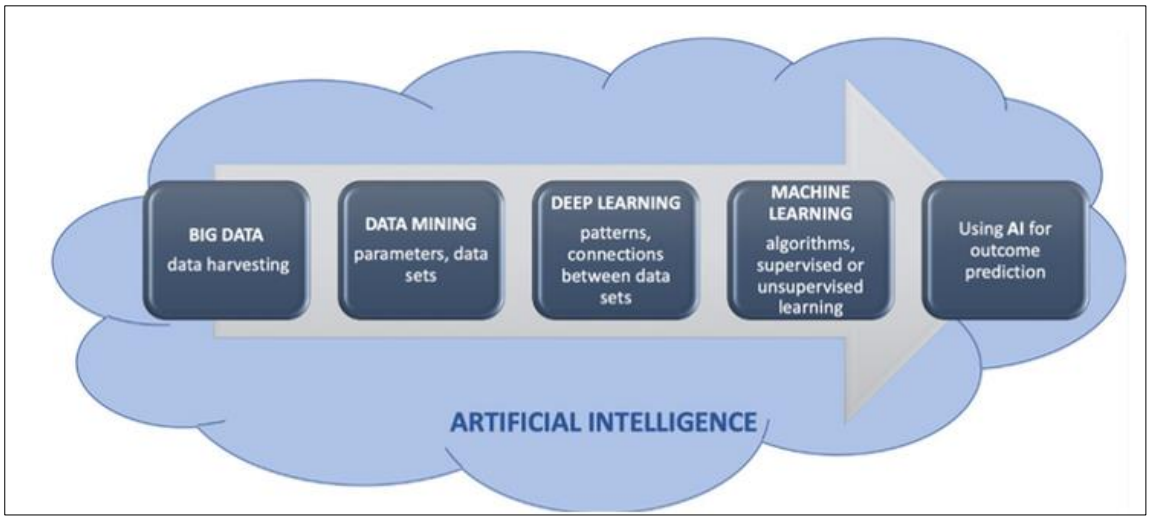


Figure (1): Artificial Intelligence Approach (Marcu, et al. 2019)

Regarding data mining, one success key is selecting the best feature out of the dataset; this selection has a considerable impact on enhancing prediction framework accuracy. Therefore, adopting and applying feature selection methodology is the best data pre-processing approach and will effectively reflect on boosting CRC classification speed and improving the intended prediction efficiency (Rado, et al., 2019).

1.2. Problem Statement

Nationally, cancer is the second cause of death, with 14% Palestine's mortality rate. 10.3% of the total cancer cases are colorectal cancer patients. CRC is the second rank after breast cancer in Palestine (MoH, 2016). While worldwide CRC is one of the most common cancer types in the world among women and men. CRC is not only the second cause of cancer related mortality, but it is also the core cause of gastrointestinal cancer. However, early CRC detection will decrease CRC related mortality rate or at least will reduce the CRC complications and prolong patients' life expectancy rate (Granados-Romero, et al., 2017).

In addition, healthcare in general and cancer registries mainly produce tremendous amount of health related data; the collected data worth to be studied and analyzed. The most powerful data analysis approach is Data Mining, where data mining aims to extract knowledge from the tremendous amount of data to aid healthcare decision-makers through pattern extraction and prediction. The health status indicators will be affected positively decision-makers considered accurate results (Sohail, et al., 2019).

Artificial Intelligence can positively participate in increasing individuals' awareness about CRC risk factors, especially when a prediction tool is published for public. The published tool will allow individuals to tune their risk factor and adopt preventive factors. Artificial Intelligence will help by extracting meaningful patterns from the huge cancer related data obtained.

1.3. Research Objectives

The research aims to assess the risk and preventive factors for CRC in Palestine and contribute to empower Palestinian individuals' awareness for both types of factors (risk and preventive). The research also aims to determine the best data mining approach that provide the highest level of accuracy for CRC risk factors classification and prediction based on the national CRC datasets.

1.3.1. Primary Objectives

1- To identify and analyze CRC behavioral risk factors that could affect Palestinian patients.
2- To study, analyze and compare different data mining methodologies on Palestinian datasets, and then the research will recommend which tool produces the best accuracy to continue from this step toward a national tool for CRC risk factors classification and prediction.

1.3.2. Secondary Objectives

The secondary objective of the study is to promote considering and adopting Colorectal cancer-preventive factors

1.4. Research Expected Outcomes

1- Determining the core CRC risk factors that affect Palestinians.

2- A framework for CRC prediction and classification tool based on the risk factors that affect Palestinian patients.

1.5. Research Significance

Many previous studies worldwide aim to analyze the risk and preventive factors of CRC and their implications. In Palestine, this is the first study that will concentrate on CRC risk and preventive factors, and to formulate a framework that could be a cornerstone for a national CRC prediction. This study is vital to be carried since cancer, in general, is the second cause of mortality among Palestinians, and CRC is achieving the second rank after breast cancer. The results of this study will form the cornerstone of building a national Colorectal cancer prediction system shortly, further; Palestinian citizens will be able to enter the associated parameters to determine their susceptibility for CRC, and they can then optimize and tune their behaviors.

On the other hand, and because early CRC diagnosis increases disease recovery probability, the study will focus on the preventive factors, spot barriers, and promote them. From another side, Machine learning (ML) methods can use in the classification and prediction risk factors in biomedical systems, including Colorectal cancer. However, Artificial Neural Networks (ANNs) and other Machine learning tools are powerful in prediction and classification which is widely used in cancer research, such as Colorectal cancer. The best Machine learning tool will be used

recommended to improve the accuracy of colorectal cancer classification and survival prediction compared to other traditional data analytics methodologies

1.6. Thesis Overview

The arrangement of the thesis will appear as the following:

- Chapter one, includes a general introduction about the research, study objectives, study significance, and problem statement.
- Chapter Two, a literature reviews that will study previous related work and will concentrate on CRC risk factors, CRC preventive factors, Artificial intelligence tools used in classifying and predicting different diseases occurrence and the conceptual framework.
- Chapter three, will present the research methodology including research design, study settings, study tools, data mining tools used, and ethical considerations.
- Chapter four, in which the results and discussion will be presented. This chapter will include descriptive and visualization for the risk factors reported in both datasets. this chapter will also consider presenting and discussing the results of applying AI tools on both datasets.
- Chapter five, finally the conclusion, recommendations, further work, and study limitations will be presented.

Chapter Two

Literature Review

In this chapter we reviewed previous studies and publications related to CRC risk factors, preventive factors and health data prediction.

2.1. CRC Risk Factors

In this section, the core modifiable and non-modifiable factors will be studied.

2.1.1. CRC Modifiable Risk Factors

2.1.1.1. Physical Activity (PA)

Eastern Mediterranean countries have lower rates of Physical Activity (PA), and physical inactivity causes 6% of Coronary Heart Disease, 7% of type 2 Diabetes, 10% of Colorectal Cancer, 10% of Brest Cancer, and 9% of premature mortality. Further, decreasing the physical inactivity rate by 10-25% will help in averting 0.53-1.5 Million deaths yearly, however; eliminating physical inactivity will increase individuals' life expectancy rate (Lee, et al., 2012).

The WHO showed that there is a strong inverse correlation between individuals' PA rate and CRC, breast cancer, diabetes and hypertension, where low PA rate is a key cause for about 21%-25% for both Colorectal and Breast cancers, and 27% for diabetes. Consequently, the WHO recommended that children between 5 and 17 years should obtain one hour of PA daily, while the age ranges (18-64) and (65-above) should accumulate 150 minutes of moderate aerobic PA weekly, or at least 75 minutes of vigorous PA (WHO, 2010).

Recently, physical inactivity is acting as a pandemic public health threat globally, where about 5.3 million deaths worldwide are attributed to low physical inactivity rates among individuals. However, the association between PA and the susceptibility of different types of cancers including CRC and breast cancer is an inverse association. PA is also associated with other chronic diseases

such as diabetes and various cardiovascular diseases. Therefore, individuals are recommended to do deliver 3,000-4,000 MET minutes weekly to decrease the probability of having CRC or other associated diseases with PA by 19% (de Rezende, et al., 2018) (Kyu, et al., 2016) (Simões, et al., 2017).

2.1.1.2. Body Mass Index (BMI) and Diet

Johnson et al., (2013) carried out a meta-analysis study of CRC risk factors. Among the study 12 non-screening attributes were proposed and reviewed comprehensively to assess each risk factor's impact on developing CRC. The study concluded that Inflammatory Bowel Diseases and Colorectal cancer history of an individual's first relative are considered two of the highest risk factors for developing Colorectal cancer. In contrast, other factors such as increasing Body Mass Index (BMI), meat consumption, smoking, physical inactivity, low vegetable intake, and low fruit intake are associated moderately with increasing the risk of developing CRC (Johnson, et al., 2013).

2.1.1.3. Smoking and Alcohol Consumption

On another side, smoking is a well-known risk factor in different types of cancers such as Lung cancer. Every single cigarette contains about 60 carcinogenic compounds that might form DNA adducts by metabolic activation, and DNA adduction may cause mutation in tumor suppressor genes and oncogenes, cell damage and/or tumor development. Since carcinogens from smoking cigarettes may reach the colorectal mucosa, then smoking is considered one of the critical risk factors for colorectal cancer with evidence (Song, et al., 2017).

Recently, Gram et al. (2020) carried out a study that aims to examine whether CRC risk due to smoking differed by gender and anatomical sub-site or not. They found that male smokers have higher left Colon cancer susceptibility, while female smokers have a higher exposure for right

colon cancer. Their study also suggested that male smokers have lower risk of rectal cancer than female smokers do (Gram, et al., 2020).

Besides, cigarettes smoking and alcohol consumption are modestly associated with CRC, significantly increase Colorectal cancer risks, serrated polyps, and adenomatous. In other words, smoking and alcohol consumption are more and strongly associated with colorectal polyps rather than Colorectal cancer development. Furthermore, there is no considerable interaction between smoking and alcohol intake on the multiplicative level. The amount and duration of alcohol consumption and smoking are associated with increasing the risk of CRC in women and men (Fagunwaa, et al., 2017) (Lee, et al., 2019).

2.1.2. CRC None-Modifiable Risk Factors

2.1.2.1. Family History

CRC family history is considered as an independent risk factor, and there is substantial variance in the contribution to the causation of CRC. Recent studies showed that many different genome-wide relation types of research identified different polymorphisms that are associated with CRC risk. The dramatic increase in CRC incidence rates in different Eastern European and Asian countries might be related to the significant shift from traditional lifestyles towards Westernized lifestyles. Thus, adopting a healthier lifestyle will positively reflect on CRC incidence rate and decrease it (Cho, et al., 2018).

Furthermore, healthy lifestyle commitment and adherence are strongly correlated with substantially CRC risk reduction regardless of genetic risk, while genetic risks may be alleviated by adopting appropriate healthy lifestyles (Cho, et al., 2018).

2.1.2.2. Medical History (Diabetes-Mellitus (DM))

In addition, Deng et. al. (2012) carried out a systematic analysis study to analyze Diabetes-Mellitus and the incidence of CRC, they included 24 studies in their research to examine the relationship between Diabetes-Mellitus (DM) and Colorectal cancer incidence, where 8 of the considered studies were case control, and 16 cohort studies. They concluded that there is a strong direct correlation between DM and CRC incidence within men and women. Consequently, and within the same study, they also included 4 studies (1 case-control and 3 cohort studies) in determining whether insulin intake as DM therapy will increase the risk of CRC or not. They also found and supported the existence of a direct correlation between insulin intake and CRC increased risk (Deng , et al., 2012).

2.1.2.3. Medications

Generally, Nonsteroidal Anti-Inflammatory Drugs (NSAIDs) and both non-aspirin and aspirin especially are observed as preventive factors against CRC. For the long term aspirin may decrease CRC incidence rate by 24%, as proved in a twenty years follow up study of five randomized trials. On the contrary, and despite aspirin's positive effects on preventing individuals from CRC, but aspirin is recommended to individuals at risk of cardiovascular disease (CVD) only; due to the

anxiety of gastrointestinal bleeding which and aspirin could negatively affect individuals in this case. (Wang, et al., 2018)

However, non-aspirin NSAIDs consumption is associated with decreasing the risk CRC among people with BMI more than 25, but not those with BMI less or equal 25. Moreover, some case-control studies showed that NSAIDs usage is associated with a more CRC risk reduction among smokers who smoked for more than 40 years compared with non-smokers. On the contrary, some clinical trials carried recently proved that aspirin is significantly affect decreasing the risk of colorectal adenomas among non-smoking individuals (Wang, et al., 2018).

2.2. Preventive Factors (Individuals' Awareness and CRC Screening)

Poor individuals' awareness of cancer symptoms is a predominant cause early cancer symptoms presentation, and leading to low survival rates. However, when cancer symptoms are atypical, poor cancer awareness of early signs and negative (false) beliefs will also delay people seeking medical assistance. Furthermore, educational level of individuals is an essential factor in recognizing their colorectal cancer risk factors. Individual's awareness also will overcome the barriers that prevent individuals from attending screening programs; since screening is one of the critical CRC prevention factors, and screening also helps in early cancer detection and diagnosis, leading to recovery or at least to lessen CRC complications (Al-Azri, et al., 2019).

2.3. Data Mining and Machine Learning

Data mining and machine learning have proven their effectiveness in predicting different types of cancer occurrence. However, efficient as well as accurate classifiers are essential for successful

big data mining and machine learning, where choosing the best classifier will maximize the accuracy of cancer prediction. (Chaurasia & Pal, 2017)

Chaurasia and Pal (2017) carried their study to compare Sequential Minimal Optimization (SMO), K Nearest Neighbor (KNN) and Best First Decision Tree on breast cancer dataset to compare their accuracy and performance. SMO achieved the best performance and accuracy.

2.3.1. Artificial Intelligence (AI) Tools

Data mining is one of the essential tools of AI, aims to extract information from pre-recorded data, and then the information forms knowledge to be stored in the Knowledge Discovery Database (KDD). Figure (2) shows the steps required for ideal knowledge discovery. However, the raw material (data) for data knowledge discovery may be noisy, inconsistent and some data mining fields might be missing. Therefore, any successful data mining process must be starting with a data preprocessing phase, where low data quality will negatively affect the data mining final results (Alasadi & Bhaya, 2017).

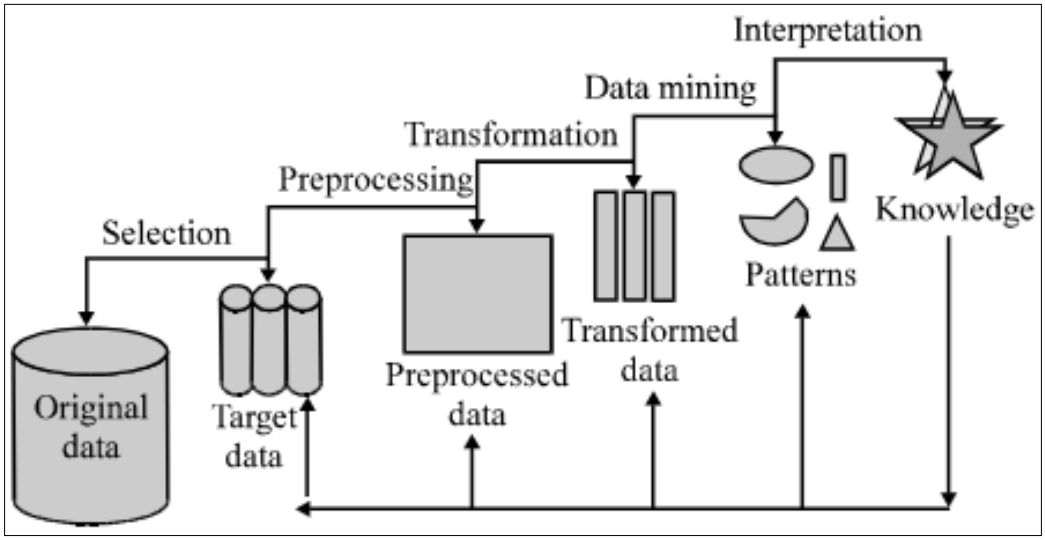


Figure (2): Knowledge Discovery Steps (Alasadi & Bhaya, 2017)

By referring to figure (2), the first and necessary step is data collection. Data selection is performed based on the study needs and requirement, next, data preprocessing methods are applied to enhance data reliability and suitability. The preprocessed data will be transformed into the appropriate form for data mining. After that, the data mining process starts with the selected data mining technique such as regression, clustering and classification, to be able to extract meaningful patterns to be interpreted and then evaluated in the last step (Alasadi & Bhaya, 2017).

Alasadi and Bhaya (2017) mentioned and identified necessary data preprocessing within their study. They strongly recommended that data mining study must carefully select and perform data preprocessing methodologies to enrich the whole data mining process by obtaining a high level of result accuracy. They recommended performing data cleaning to get rid of noisy or unnecessary data and to complete the missing values, also; data integration was recommended if there are different data sources. However, completing the missing values within the data can be performed using one of the following techniques:

- Using the mean of the attribute to fill the missing records.
- Fill the missing value manually, where this approach requires human effort and is not applicable for massive datasets.
- Using a globalized value to fill the missing records.

Artificial Neural Networks (ANNs) emulate neurons in human's brain as a network, where AI is a replica of the network of neurons made to learn from the tasks performed through different algorithms and mechanisms. ANNs work similarly to biological neural networks, where ANNs use from 10 neurons to 10,000, and this number is much less than biological ones. Each ANN is a different computational model that is eligible to apply pattern recognition and machine-learning. Further, there are many other AI methodologies can be used in detecting cancer occurrence as well

as predicting the probabilities of cancer incidence, such as ANNs, Particle Swarm Optimization (PSO), Genetic Algorithms, K-Nearest Neighbor algorithms (K-NNs), Support Victor Machines (SVMs), Linear Regression and Fuzzy Clustering (Patel, et al., 2020).

Consequently, Patel et al. (2020) found that deep machine learning with Gaussian model and watershed transform achieved the best accuracy result among different AI methodologies as shown in Figure (3), where this comparison was carried out for a study the accuracy of different AI methods for Liver Cancer detection.

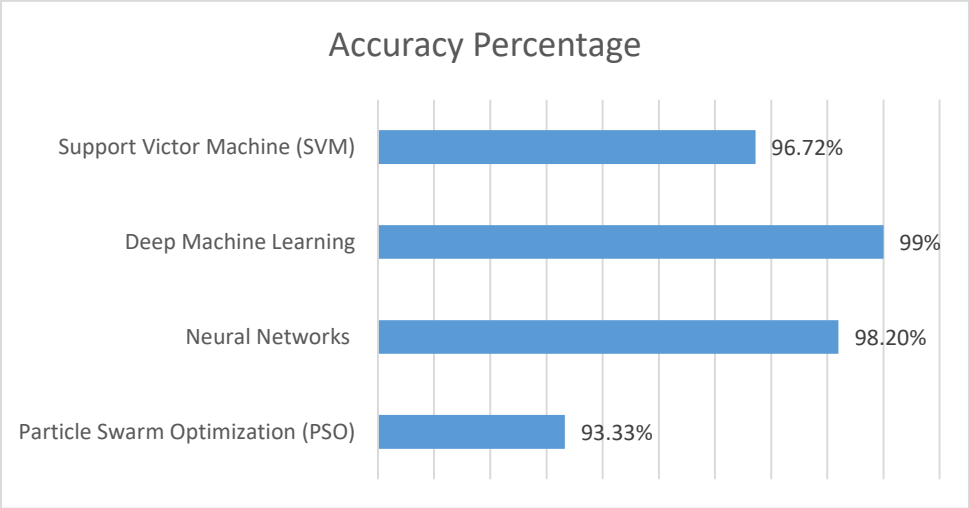


Figure (3): AI Methods' Accuracy Percentage. (Patel, et al., 2020)

Cancer anatomical and functional medical imaging such as Computed Tomography (CT), Magnetic Response Imagining (MRI), Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT) expanded rapidly. Biological screening methodologies such as blood tests, breath tests, proteomics, and genomic. Consequently, wealthy and huge data will be collected and rich information can be generated. This can be employed in big-data application implementation, such as mining tools, ML, and prediction models development. However, adopting different big-data applications will help stratify CA patients based on their diagnosis, lifestyle risk factors, preventive factors, comorbidities and any other

attributes required. Also, AI applications will add positive value to cancer studies through developing predictive methods that can be used to predict the outcome for individuals based on pre-recorded data about different cancer individuals (Marcu, et al., 2019).

2.3.2. Machine learning algorithms

Since early CRC detection increases the probabilities of recovery and decreases the CRC burdens, many data-mining techniques were utilized to strengthen early CRC detection in many primary healthcare institutions. Therefore, Liang et al. (2020) carried out a systematic review study that aims to synthesize different studies that inspects data mining effects on enhancing CRC risk factors prediction. Consequently, they concluded that machine learning algorithms are superior when comparing them with traditional algorithms (Liang, et al., 2020).

Further, Ting et al. (2019) proposed a CRC prediction schema, the proposed schema suggests integrating Multivariate adaptive regression splines (MARS), Random forest (RF), Support vector machine (SVM), Extreme learning machine (ELM) and XGBoost as shown in figure (4).

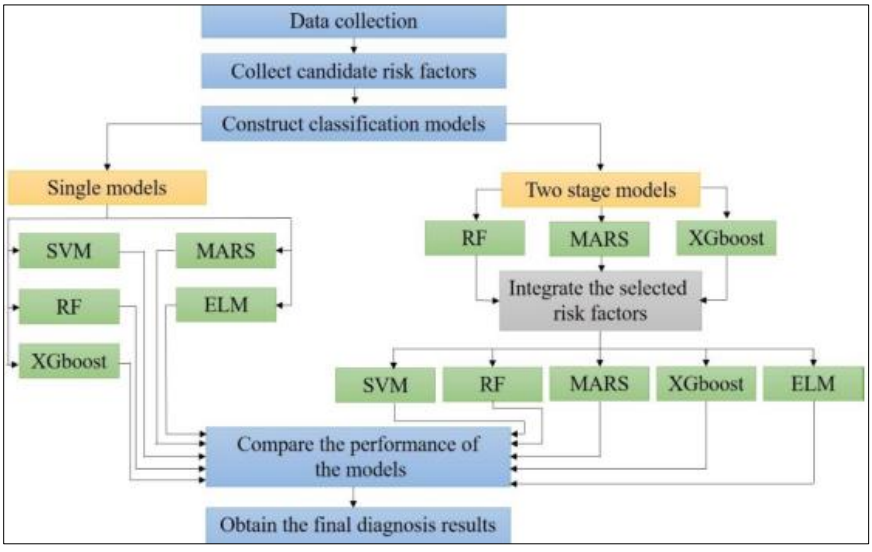


Figure (4)CRC Prediction Scheme (Ting, et al., 2019)

Then the study compared the five data mining algorithms. The comparison was performed to assess each algorithm's score individually by considering accuracy, sensitivity specificity and area under the curve (AUC). After that, the study compared the same scores in two-stage models after integrating different algorithms (Ting, et al., 2019).

The study concluded that the A-XGboost method which is a two-stage and integrated approach achieved the best AUC, sensitivity, and accuracy, where the AUC scored was 71.4%. In comparison, the second-ranked method was S-SVM which is a single method that achieved the second AUC which was 71.1%, and figure (5) shows the AUC score for each technique. Therefore, A-XGboost can be an appropriate method to adopt in CRC prediction, also A-XGboost can be used effectively to select CRC risk factors that influence individuals more (Ting, et al., 2019).

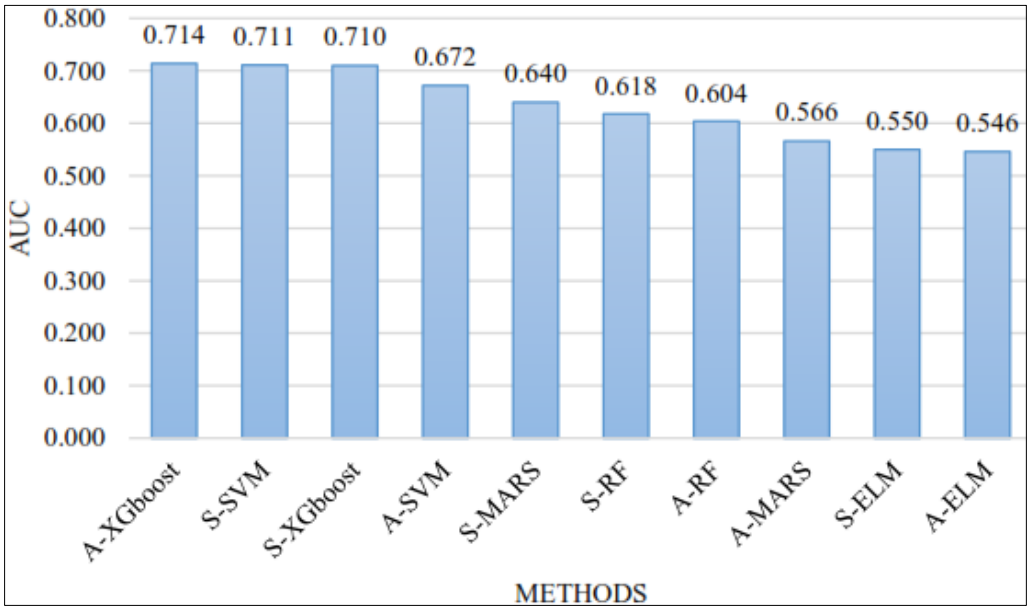


Figure (5): Comparison of the AUC values of the five classifiers with and without using the proposed scheme (Ting, et al., 2019)

Also, Yan et al., (2018) carried out a study aiming to compare the performance of Decision Trees (DTs), ANNs and SVMs data mining methods, for a five years CRC patients' survival prediction. The carried study utilized ANNs, SVMs and DTs within a five-fold cross validation to efficiently compare accuracy. The study also adopted the Random Forest tactic in order to assess and give weight for each variable included.

As shown in figure (6) and figure (7), SVMs were the best method in the five-fold accuracy test with an accuracy average of 75.5%. In comparison, ANNs came after SVMs with an accuracy average of 73.1%, and the last was DT with an accuracy average of 67.6%. Therefore, SVMs and ANNs are more recommended to adopt in data mining methodologies for CRC datasets (Yan, et al., 2018).

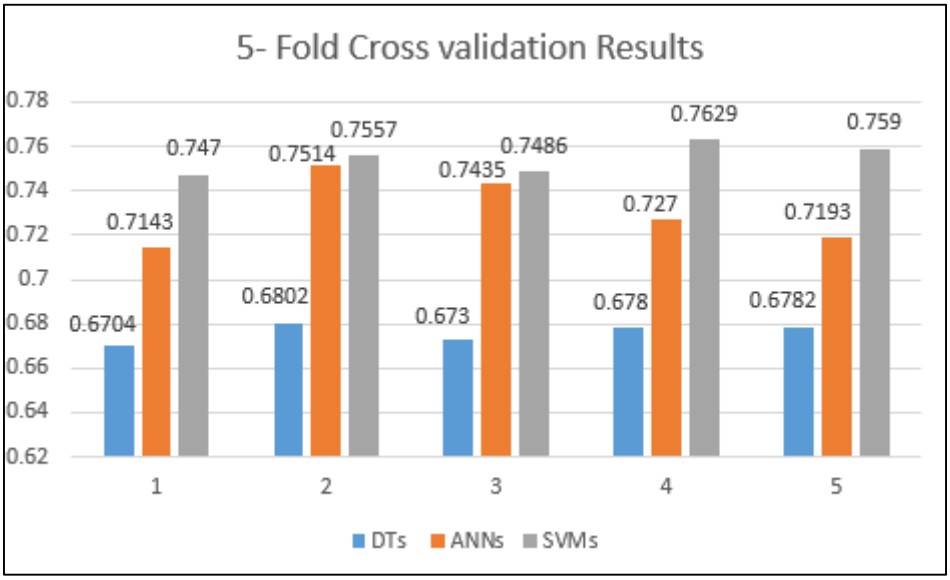


Figure (6): Five-Cross Validation Results (Yan, et al., 2018)

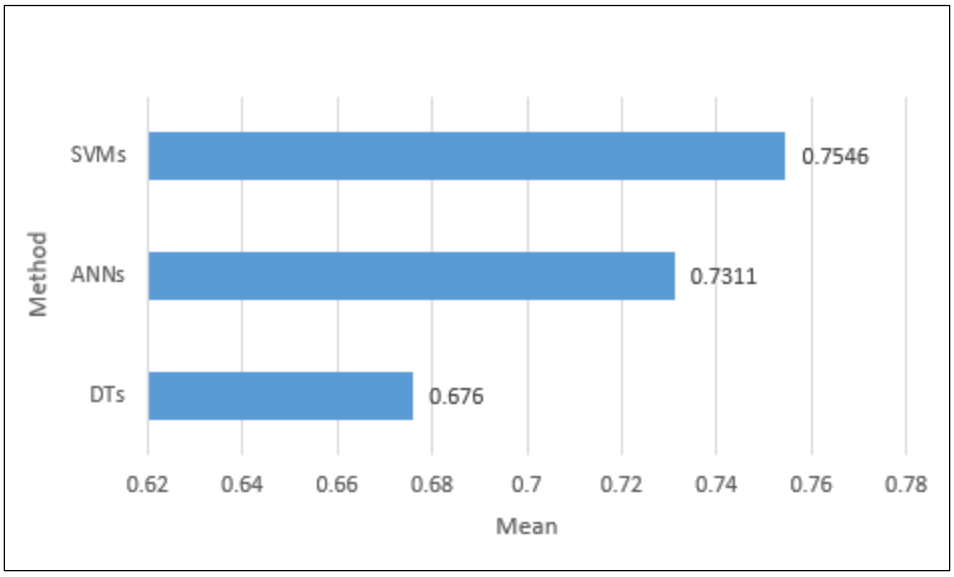


Figure (7): Means of Five-Cross Validation Results (Yan, et al., 2018)

On another side, Asri et al. (2016) carried out a study to evaluate different Machine Learning tools for breast cancer diagnosis and prediction. The considered machine learning tools were SVM, DT, K-NN and Naïve Bayes (NB). The study aimed to evaluate data classification correctness respecting effectiveness and efficiency for each tool in terms of result accuracy using WEKA data mining tool.

K-NN consumed the least time (0.01 S) to build the learning model while SVM consumed (0.07 S), as shown in figure (8). At the same time, SVM scored the best accuracy percentage among all tools with (97.13%), indicating the laziness of K-NN during learning and the superiority of SVM. Also, as shown in figure (9), SVM achieved the best score for the probabilities of having the best classification with Kappa Statistic (93%) and with the least possible error with (MAE 0.02). Consequently, the best machine learning tool to be applied in breast cancer prediction is SVM (Asri, et al., 2016).

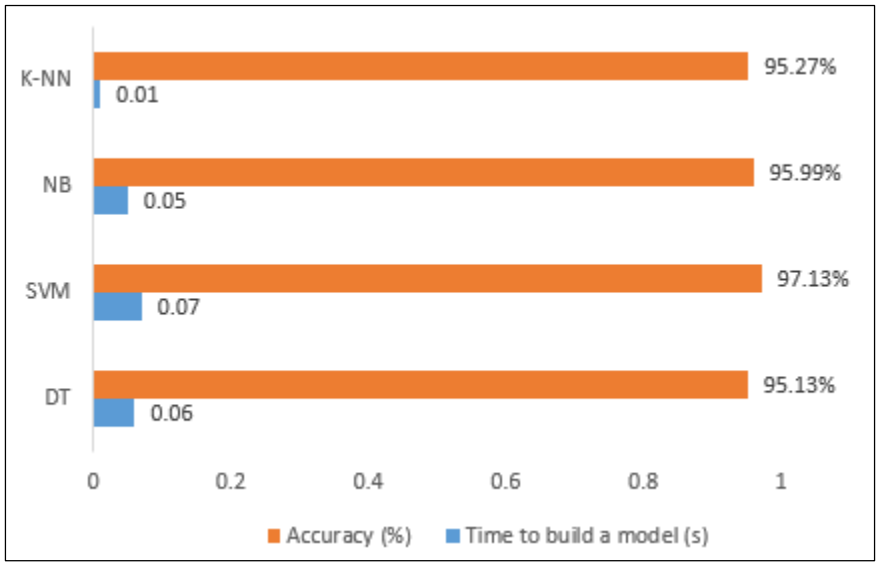


Figure (8): Accuracy and Time to Build a Learning Model Comparison (Asri, et al., 2016)

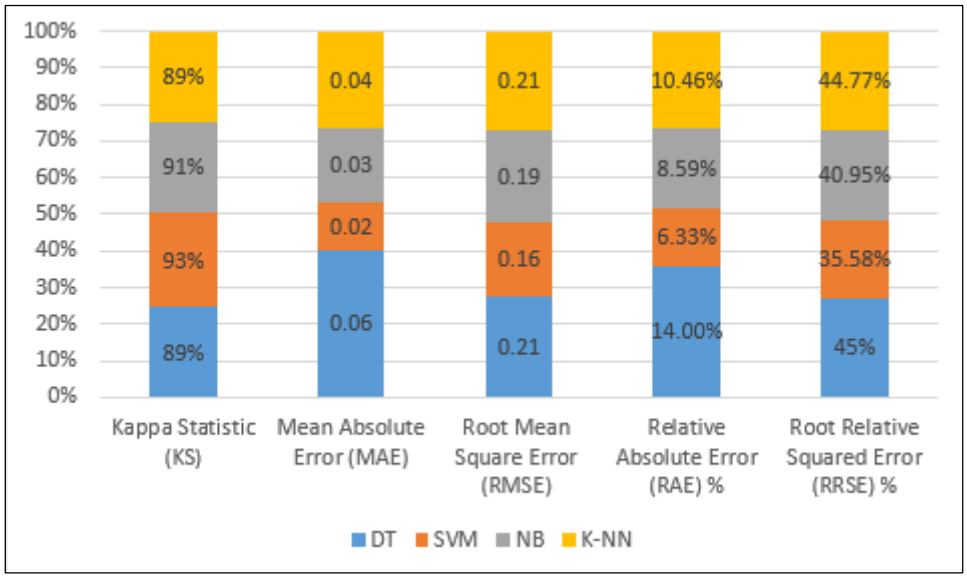


Figure (9): Statistical Analysis of Different Machine Learning Tools (Asri, et al., 2016)

2.4. Conceptual and Theoretical Framework

The study will consider that Colorectal Cancer as a dependent variable that is directly affected by risk factors as independent variables. However, the study also considered intervening variables such as educational level, marital status and living area, where the intervening variables will affect the dependent variable directly or affect the independent variables that will affect CRC in the end, as shown in figure (10).

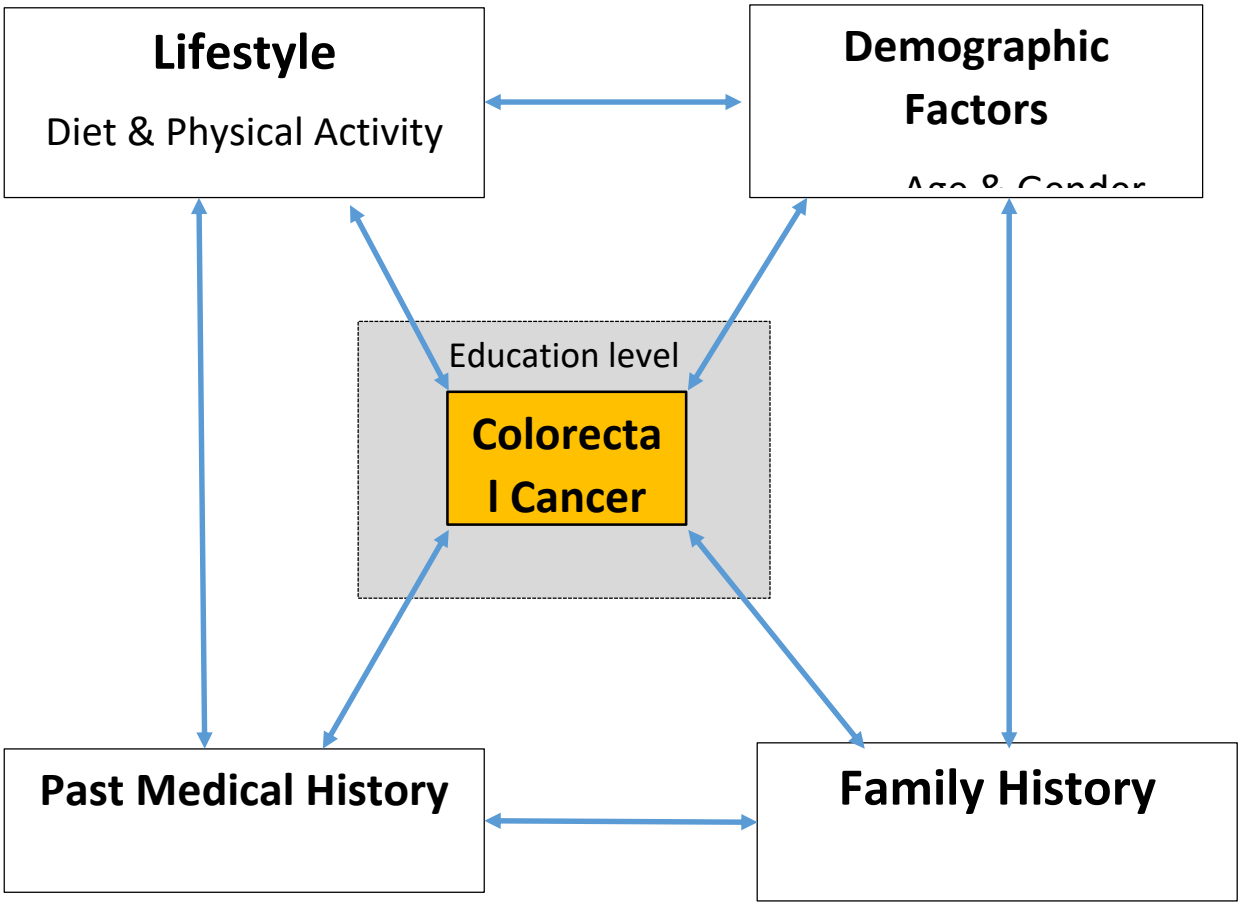


Figure (10): Research Conceptual Framework

The core risk factors included in the study are:

- CRC non-modifiable risk factors:
    - Demographic factors (such as gender and age).
    - Past Medical History (having Diabetes-Millets, Hypertension and Heart Diseases).
    - Family History (a relative with CRC or other cancer).
- CRC modifiable risk factors
    - Smoking (indicates if the individual smoke or no).
    - Lifestyle such as occupation, diet and physical activity rate.

However, educational level, marital status and individuals' living area were considered as intervening variables within the study.

Chapter Three

Research Methodology

In this chapter, study settings, study design, study tools, data gathering methods, data management, data analysis, and ethical considerations will be presented.

3.1. Study Settings

The study was carried out by analyzing two different national datasets for Colorectal Cancer (CRC) patients, the first data set was obtained from the national cancer registry, while the second was obtained from Al-Quds university. The dataset obtained from the cancer registry was for CRC patients only, however; Al-Quds's dataset was based on a group matched case-control study design using comprehensive interviews and records review. All data targeted the CRC patients from the two datasets where included in our study.

3.2. Study Tools and Data Collection

Two national datasets were used to meet our study objectives, one from the national cancer registry and one from Al-Quds university. More details on these datasets can be followed below:

3.2.1. National Cancer Registry Dataset (MoH Dataset)

The dataset which obtained from the national cancer registry contains 243 Colorectal cancer patient records with 21 attributes. However, selecting the important and the crucial attributes as risk factors for CRC, an expert and a specialized physician was consulted, and he reviewed the 21 attributes available within the dataset, then he nominated smoking, past medical history, family history, physical activity rate, alcohol consumption, age, and gender.

On another side, some important variables, such as alcohol consumption, physical activity rate and diet were missing within the provided dataset. Therefore, the physical activity rate was calculated using a formula based on the occupation of the patient since the occupation was one of the

attributes covered within the dataset, while alcohol consumption and diet were excluded from the study.

3.2.2. Al-Quds University Dataset (AQU Dataset)

The dataset obtained from researchers at Al-Quds university contained data collected for a study carried in order to assess CRC determinants within patients who attends Beit-Jala Governmental Hospital. The data within the dataset was collected as primary data for their study since their study design was a group matched case-control study, where the control and case groups were matched by gender and age. The study used a triangulation study design where qualitative and quantitative methods were used. The dataset contained 107 CRC cases and 109 control groups. The variables covered within the study were 58 variables.

3.3. Questionnaire

Using these datasets can't be useful without having case-control groups. As prediction using artificial intelligence tools such as data mining and machine learning requires cases and control groups, so the desired tool can be performing the required training. Al-Quds university dataset had that requirement. However, the national cancer registry has only a case group. Thus, since the cancer registry dataset does not contain healthy people records (people who do not have the disease yet), a questionnaire was designed and validated by researchers to cover the same attributes found in the dataset. The carried questionnaire was distributed and collected electronically using google forms. The questionnaire was designed to collect identical data to the national cancer registry, from 227 individuals participated in the questionnaire. The questionnaire covered living area, family history, age, gender, employment, marital status, past medical history, smoking, alcohol consumption, and physical activity variables, a copy of the carried questionnaire is shown in appendix (3).

3.4. Data Management and Analysis Framework: Data Mining Framework

After extracting data from the two datasets and forming our developed questionnaire, we identified, the study framework to perform data mining processes of different local Palestinian datasets to predict the probabilities of CRC occurrence, and then compare these methods' accuracy results with traditional statistical methodologies. The proposed framework is shown in figure (11).

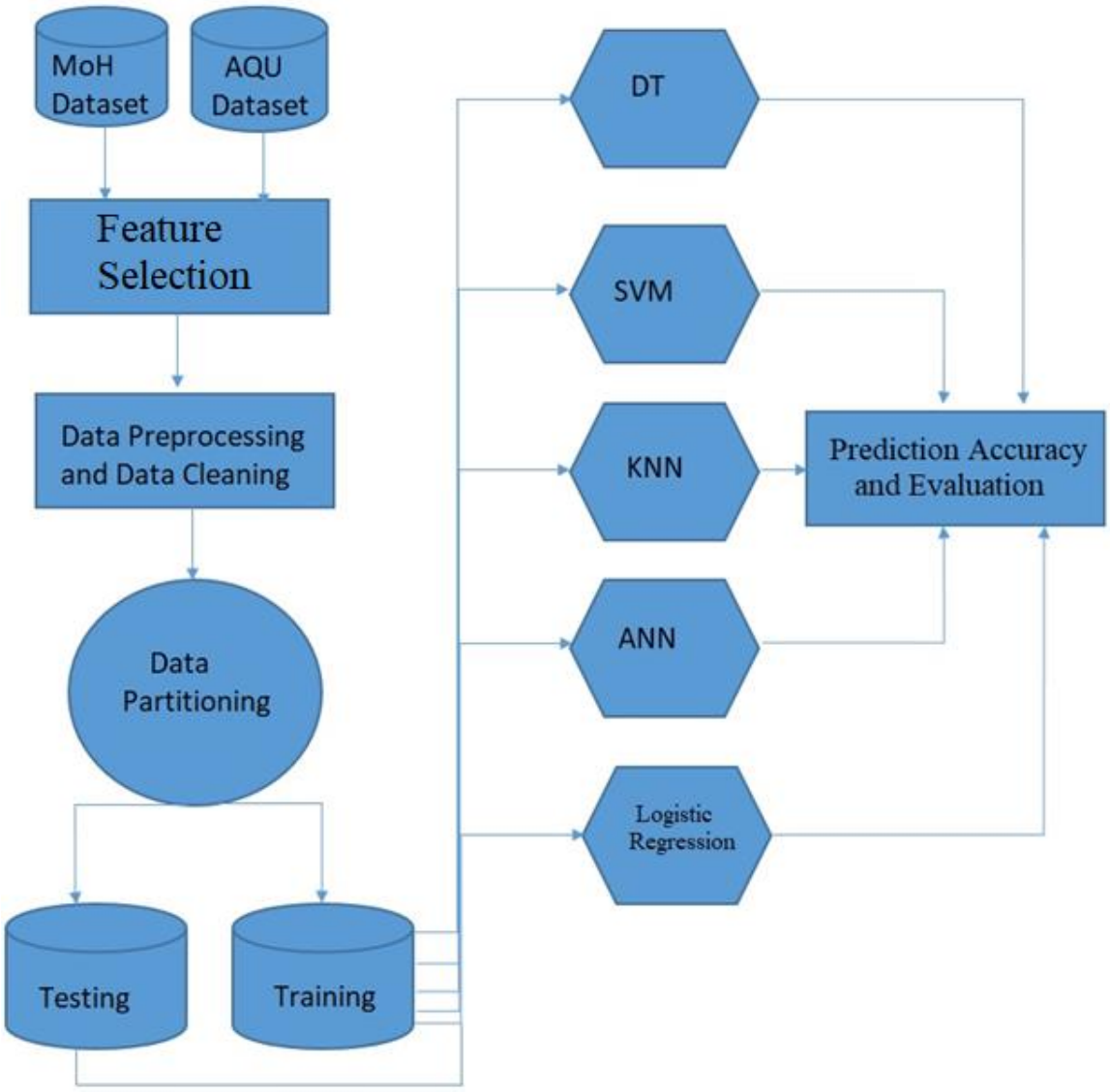


Figure (11): Data Mining Proposed Framework

The following sections will explain the details of each step in the data mining framework:

3.4.1. Feature Selection

An extensive literature review is carried out and an expert in Colorectal cancer field was consulted in order to analyze the profiling methodology using data mining techniques. The irrelative attributes were excluded from the beginning as it cannot be considered the leading risk factors for CRC.

3.4.2. Data Preprocessing

Data preprocessing and data cleaning were performed to prepare both datasets for further processing using artificial intelligence tools. All corrupted records were omitted, the unwanted attributes were excluded, and some were estimated based on some calculations. The data processing process, which was applied for both datasets included the following tasks:

1- Omitting duplicated records.
2- In the national cancer registry, there were three different attributes (PMH1, PMH2 and PMH3) indicating individual's past medical history; thus, they were summarized and merged into one attribute.
3- The invalid entries were treated, the treatment applied by deleting the record without affecting the dataset, and in some cases we changed the value to be unknown.

Consequently, Excel was used in order to produce visualized descriptive information about the datasets and to describe how the data looks. Further, MATLAB was used to run different data mining methodologies on both datasets after data cleaning and preprocessing.

3.4.3. Data Partitioning

After the obtained data well cleaned and processed, the study considered analyzing five different popular artificial intelligence tools, to compare which tool scored the best accuracy, further. This comparison might be the cornerstone of a national Colorectal cancer prediction tool, the study analyzed Decision Tree (DT), Support Victor Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression, and Artificial Neural Network (ANN). More details on the five artificial intelligence tools can be found in the following section.

However, the analysis adopted a 70% training and 30% testing basis, where 70% of each dataset records were dedicated to training the artificial intelligence tool, while 30% of the rest records are dedicated f the tool's output. Then, as shown in figure (12), the carried analysis among the five selected tools adopted a five-fold cross validation, where to get the result validated, wherein each stage the data is divided into testing and training parts, and the training part is divided into four pieces. Also, in each stage the testing data is selected from different place to finally each record will be used for training and testing purposes. Finally, the average for the five stages is calculated as the final accuracy score for each tool.

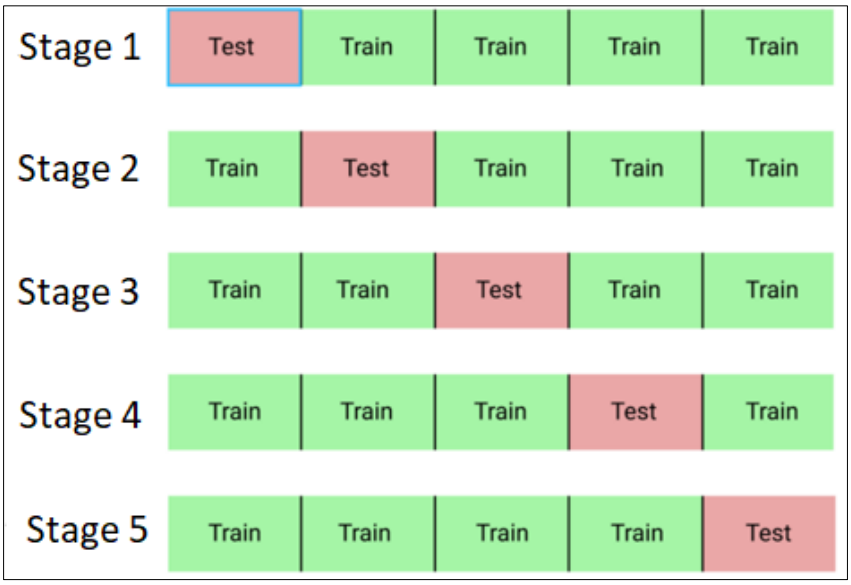


Figure (12): Five Fold Cross Validation

3.4.4. Data Mining Process

In this phase, the most five popular data mining algorithms will be applied on the gathered, selected and preprocessed data. More details about data mining algorithms are shown in the coming sections.

3.4.4.1. Decision Tree (DT)

DTs are considered as one of popular in the data mining field and knowledge extraction. DTs work in systematic sequential structure, where the first node is called root, and the consequent nodes are called internal nodes, while the terminators that shows the result of each decision is called a leaf. (Kretowski, 2019) Decision Trees for work by traversing an object throughout the tree from the root until it reaches one of the leafs. This means that firstly the test which is associated with the root is performed, then when the result is obtained from the first test, the result will be sent to the appropriate branch of the tree, thus; the object will move down reaching the corresponding target

node, this process will keep repeating until the object reaches a leaf, then the value at the final leaf will be the value for the prediction (Kretowski, 2019).

3.4.4.2. Support Victor Machine (SVM)

Support Victor Machine is a supervised learning tool based on different statistical theories, SVM works by separating two data points linearly using hyperplanes to be able to differentiate the class of the predicted object. On the other hand, SVM can also apply non-linear classification using something called kernel trick to correctly map the object to be predicted to the appropriate class (Janardhanan, et al., 2015).

3.4.4.3. K-Nearest Neighbor (KNN)

KNN is one of the popular classification tools which perform classification based on non-parametric methodologies. In other words, KNNs are instance-based classification methods, where objects among the dataset are placed just near to those that have similar attributes (Meneses, et al., 2019).

3.4.4.4. Logistic Regression (LR)

LR is a popular statistical and data mining tool that classifies binary datasets and proportional ones, and sometimes LR is called Logit Regression. LR, as a tool, is used to produce the probability of an object or a class. However, the function logistic function is called Sigmoid function where it's shaping the word "S" that will only accept from (0) to (1) values, and represent the probability of each object's prediction to be in between and with 1 in total for all possibilities (Maalouf, 2011).

3.4.4.5. Artificial Neural Network (ANN)

Artificial Neural Network is a robust mathematical data mining tool that functions as the same as natural neural network. However, ANN is a network of interconnected neurons (processors) with

weight for between each two neurons. ANN's model is usually formed by input, hidden and output levels, where the hidden level represents the neurons (processors) that will be trained on a part of the data which entered by the input level, and then will test (validate) the learning process on the rest of the data to produce the output (Nasser & Abu-Naser, 2019).

3.4.5. Results and Evaluation

After applying the five data mining tools on both datasets, the performance of each tool will be calculated based on accuracy, sensitivity, specificity and Area Under the Curve (AUC). A brief description about each method used to evaluate the tools will be mentioned within the coming section.

3.4.5.1. Accuracy

The accuracy of a data mining tool indicates its correct predictions' percentage. Accuracy is calculated using the formula

$$\frac{\text{True Positive} + \text{True Nagative}}{\text{True Psitive} + \text{False Positive} + \text{True Negative} + \text{False Negative}}$$

Where true positive is the number of positive records that are predicted truly. Similarly, the true negative is the number of records predicted that lie in the negative class and predicted truly.

False positive number indicates the number of records of positive class but predicted incorrectly, and false negative is the number of records that belong to the negative class but the applied tool incorrectly predicts the result.

3.4.5.2. Sensitivity and Specificity

Sensitivity in data mining is used to calculate the proportion of positive cases that are predicted correctly, and it is measured using the formula:

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

On the other hand, specificity in data mining measure the proportion of incorrectly predicted negative class records Specificity is calculated using the formula:

$$\text{Specificity} = \frac{\text{True Negative}}{\text{TTrue Negative+False Positive}}$$

3.4.5.3. Area Under the Curve (AUC)

AUC is the area under the Receiver Operating Characteristic (ROC) curve. ROC curve is a powerful test that evaluates a diagnostic test, in ROC curve the true positive rate (TPR) is calculated in function with false positive rate (FPR) for many predefined parameter thresholds. AUC under the ROC measures the performance and the ability of specific parameter to classify different objects based on their classes.

However, the AUC provides an aggregate measure of performance across all possible classification thresholds. AUC can be interpreted as the probability that the measured model will rank a random positive example higher than a random negative. Simply, AUC evaluates the ability of a model to distinguish between classes (positive and negative).

3.5. Ethical Considerations

First of all, and in order to kick off this research, an approval was obtained from the College of Graduate Studies at the Arab American University – Palestine (AAUP) after the research's

proposal was discussed by the research committee. Then, as shown in Appendix (1), Ministry of Health approval was obtained to collect the required and available data from the national cancer registry for the study.

Also, as shown in appendix (2), an agreement was signed between this study researchers and researchers from Al-Quds University to use the primary data they collected for a previous recent study.

3.6. Chapter Three Summary

This chapter discussed the design and methodology adopted in our study. The following chapter presents the results and discussion of our study.

Chapter Four

Results and Discussion

In this chapter, the results of the different analysis methods performed on the collected data will be presented. In addition, the data mining tools results as well as general descriptive statistics about the data within both datasets will be discussed.

4.1. Descriptive Statistics and Visualization

After CRC patients' records were acquired, descriptive analysis and visualization were performed using excel to describe the data and to prepare it for analysis later using Math-LAB.

4.1.1. National Cancer Registry Descriptive Statistics and Visualization

- **CRC Cases Living Area Distribution**

  As shown in figure (13), the highest Palestinian district with CRC patients living in is Nablus, where 20.58% of patients included in the national cancer registry, while Jerusalem gained the lowest percentage with 2.06%.

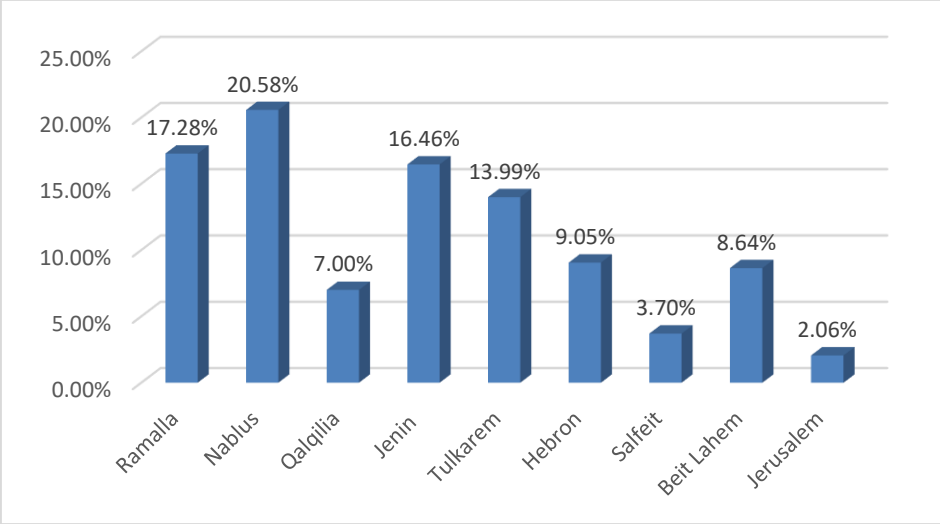


Figure (13):CRC Patients by Living Area (MoH Dataset)

- **CRC Patients' Age Ranges**

Patients' age average on diagnosis is 60.63 years, while the most age repeated (Mode) was 48 years with 30 years old as the youngest case and 91 as the most senior. However, and as shown in figure (14), majority of Palestinian CRC patients lies in the age ranges between 41 years and 80 years.

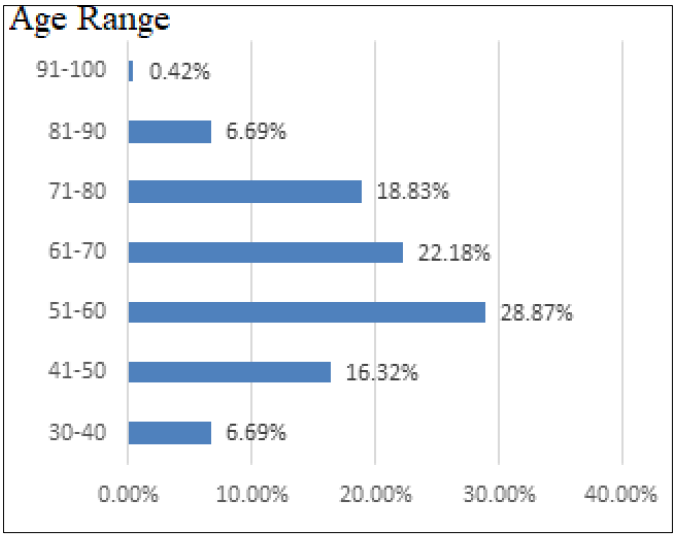


Figure (14): Palestinian CRC Patients' Age Ranges (MoH Dataset)

- **Palestinian CRC Patients' Occupations Distribution**

The study revealed that 45.29% of CRC cases within the national cancer registry data set are housewives, and this is reasonable since female CRC cases about 50% of all CRC cases, and 39.29% are classified as other occupations, while the lowest percentages were achieved by agriculture and nursing workers as shown in figure (15).

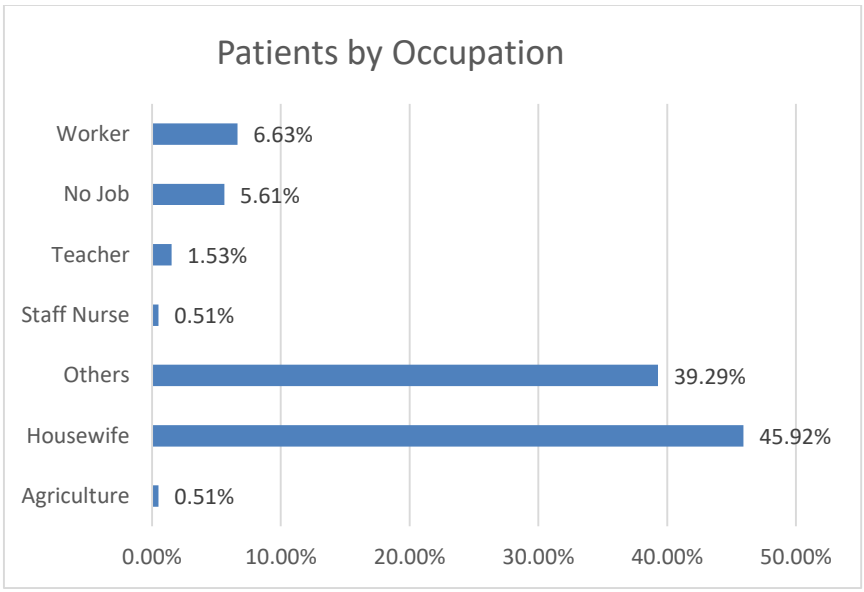


Figure (15): Palestinian CRC Patients by Occupation (MoH Dataset)

- The gender distribution of Palestinian CRC patients reported within the MoH dataset is almost equal, where 50.21% are female patients, and 49.79% are male patients as shown in figure (16).

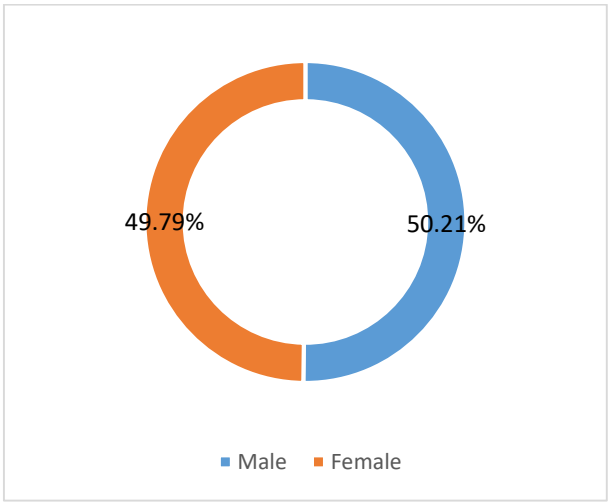


Figure (16): CRC Patients' Gender (MoH Dataset)

- **CRC Patients' Marital Status Distribution**

Most CRC patients (83.80%) are married, and 0.93% are divorced, as shown in figure (17).

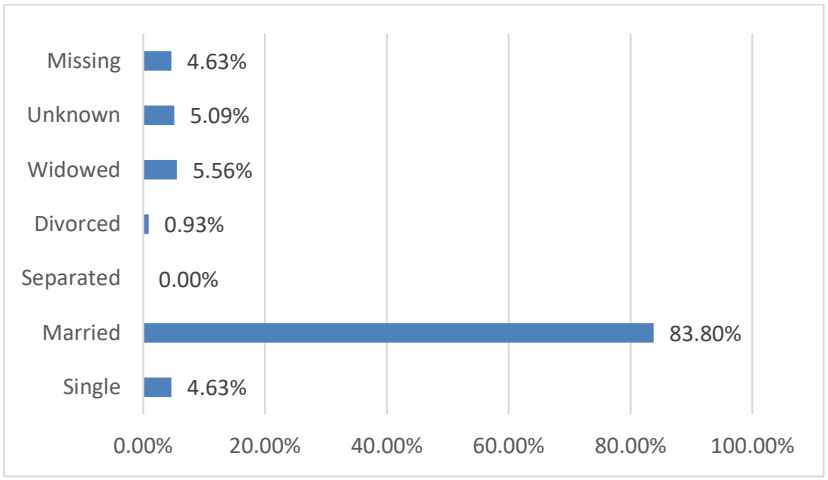


Figure (17): Patients by Marital Status (MoH Dataset)

- **CRC Patients' Past Medical History Distribution**

Regarding past medical history as a risk factor, the study found that 41% of CRC patients are with a Past Medical History (PMH), while 37% are PMH free and 21% unknown as shown in figure (18). While figure (19) shows that 16.05% of patients are diagnosed with Hypertension (HTN) and Diabetes-Mellitus (DM) together as PMH, while 8.64% with HTN only, and 5.35% of the patients with DM only.

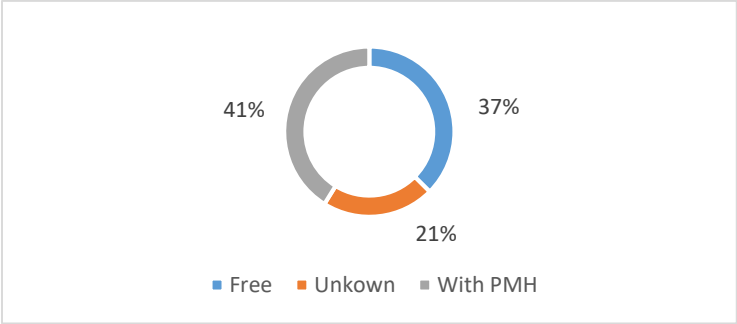


Figure (18): Past Medical History (MoH Dataset)

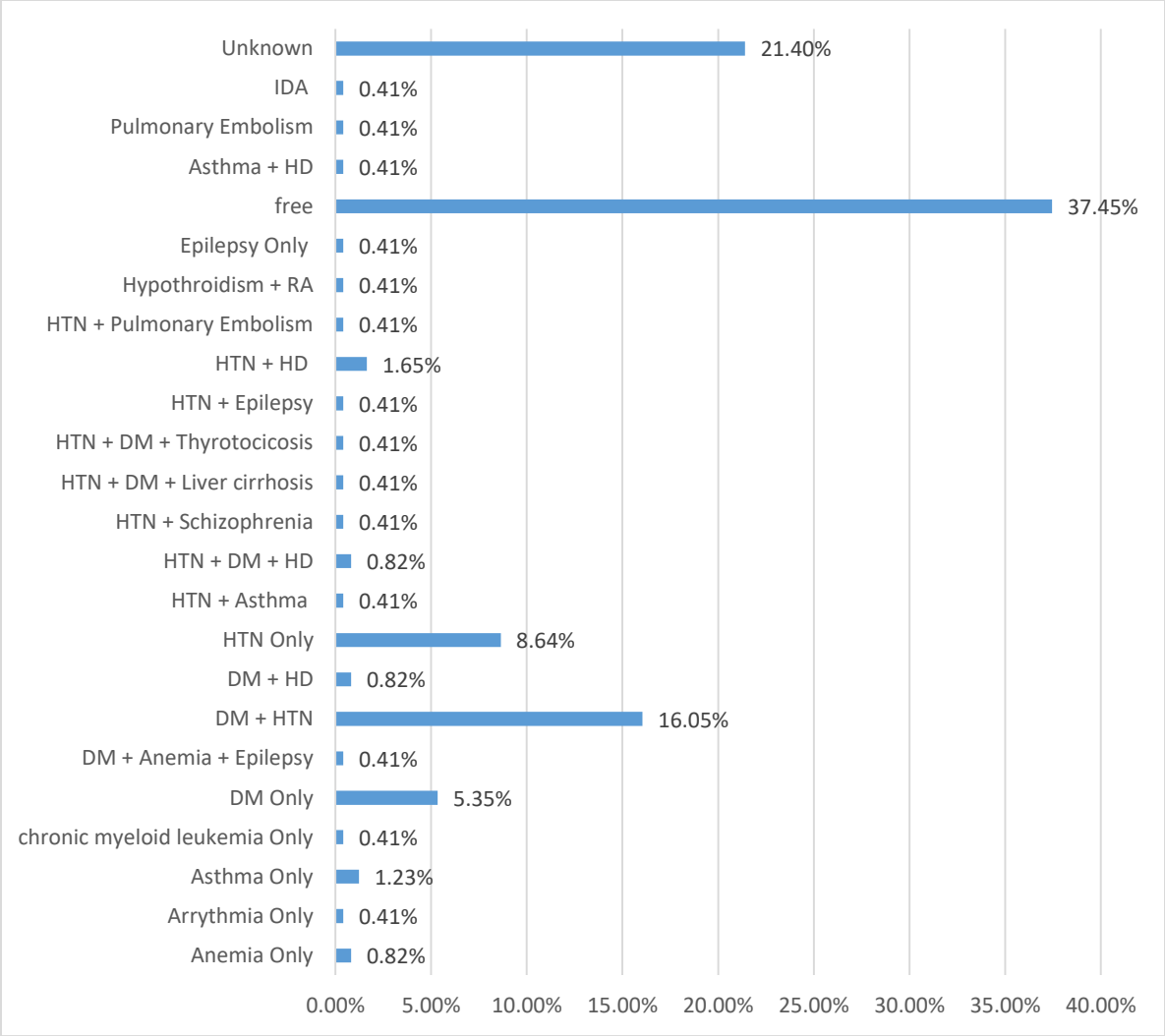


Figure (19): Detailed View of Patients' Medical Past History (MoH Dataset).

Further, since the approved cut-off date approved by the American Cancer Association (2000) is 50 years, then during the study, the researcher tried to describe past medical history based on the age of 50. Consequently, the study found that 21% of CRC cases are 50 years old or older, and they have Diabetes-Mellitus (DM) only or DM with other past medical history. In addition, 27% of the patients are 50 years old or older and they have Hypertension (HTN) only or HTN with other past medical history issues, while only 8% of all patients are 50 years old or older and they

have different Heart Diseases (HDs) only or HDs with other past medical history issues as shown in table (1).

Table (1): DM, HTN and HD Past Medical History with Age (MoH Dataset)

	% OF ALL CRC CASES
Age >=50 With DM	21%
Age <50 With DM	2%
Age >= 50 With HTN	27%
Age <50 With HTN	1%
Age >= 50 With Heart Disease	8%
Age <50 With Heart Disease	1%

4.1.2. National Cancer Registry Dataset Descriptive Statistics and Visualization Summary

First of all, it is clear that a relationship might exist between individual's medical history and CRC occurrence as supported in the literature, where this study revealed that 41% of CRC cases have a medical history of having one chronic disease at least, where majority of the reported medical histories were Diabetes Millets, Heart Diseases and Hypertension. Further, it is clear that there is a relationship between age and past medical history. At the age 50 and more, there is a dramatic increase in CRC cases percentage who have Diabetes-Millets, Hypertension and other different Heart Diseases. From the data visualization, it is clear the individuals who have more than one other chronic disease are more likely to catch CRC, especially if they have Diabetes-Millets and Hypertension at the same time.

From another side, 16.46% of CRC cases reported within the national cancer registry dataset were from Jenin, and this is less than the percentage of Hebron for example. However, when we compare both areas from the population size, it is proved that Hebron has much higher population size. This increase in Jenin's percentage could be because CRC core risk factors are behavioral such as diet and an individual's lifestyle and physical activity rate.

Besides, and as shown in the literature, gender is a risk factor to be studied and analyzed, where the percentage of female CRC cases is slightly less than male percentage, and Arafa and Farhat, (2015) support this. Then found that male individuals' CRC incidence rate is slightly more than female individuals in all Gulf Cooperation Countries except Qatar, wherein Qatar's female CRC incidence is more than males'. On the other hand, individuals' marital status can be considered as a risk factor since there is a statistical significance in married patients, where 83.80% of the cases are married.

4.1.3. Al Quds University Dataset Descriptive Statistics and Visualization

Since age is one of the proven risk factors scientifically, then it is important to study and analyze the age for CRC cases. The majority of CRC cases in Al Quds university dataset range from 40 years to 79 years, with 90% of the cases within this age range as shown in figure (20). While the average age for CRC cases in this dataset is 63 with 24 years old was the youngest CRC patient and 87 as the most senior.

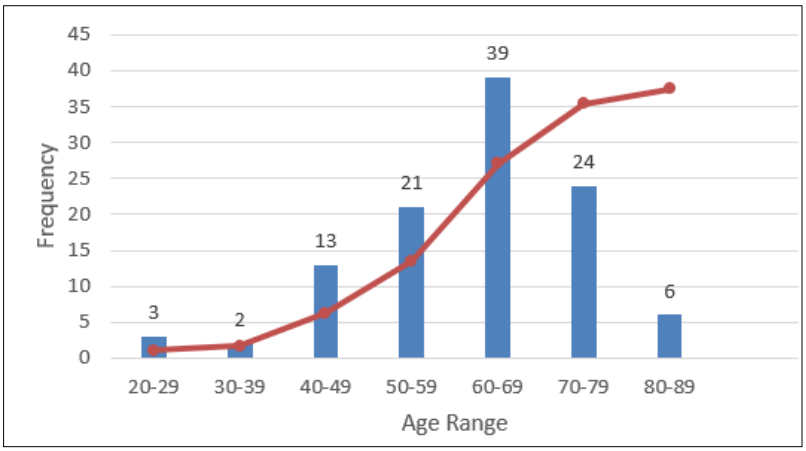


Figure (20): CRC Cases Distribution by Age (AQU Dataset)

When CRC cases' gender was analyzed, we found that 45.37% of the CRC cases in Al Quds university dataset are females, while 54.63% were males, as shown in figure (21). On the other hand, when the study considered the age 50 as a cut-off data as adopted by the American Cancer Association, we found that 92% of all male CRC cases are 50 years or older, while only 73% of all female CRC cases are 50 years old or older as shown in table (2). Thus, this might be an issue for further analysis in future researches.

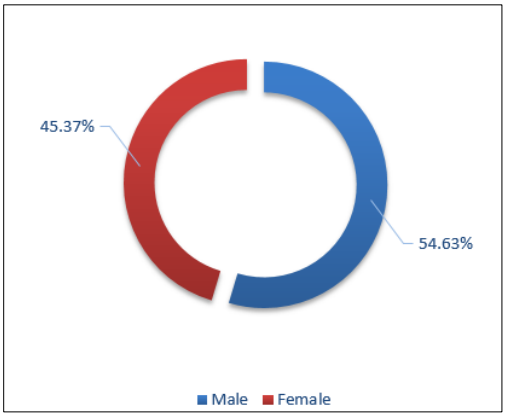


Figure (21): Cases Distribution based on Gender (AQU Dataset)

Table (2): Critical Age Based on Gender (AQU Dataset)

CRITERIA	COUNT	% OF THE SAME GENDER
FEMALE AND AGE >= 50	36	73%
FEMALE AND AGE < 50	13	27%
MALE AND AGE >= 50	54	92%
MALE AND AGE < 50	5	8%

Regarding diet, it was found that the majority of CRC patients eat fruits, red meat and grilled meat a couple of times per week, while they eat vegetables from two to four times a week, on the contrary, the percentage of CRC patients who used to eat fruits more than four times a week is greater than those who eat vegetables four times a week, the percentages for four diet habits are

shown in figure (22) consequently, the study found that 6% only of CRC patients within the study who eat fruits and vegetables together more than four times a week.

Further, Ghrouz & El Sharif (2019) found in their study that fruit, vegetables, red meat and grilled red meat intake as dietary habits are statistically significant to CRC incidence in Palestine, as shown in table (3).

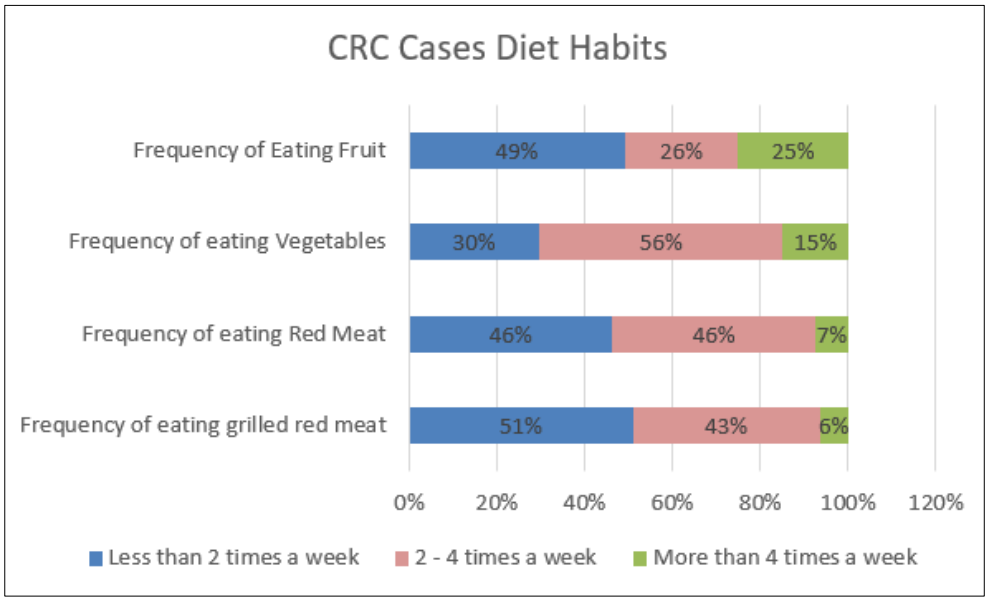


Figure (22): CRC Cases Diet Habits (AQU Dataset)

Table (3): Association between cases and control groups by Diet (Ghrouz & El Sharif , 2019)

variable		Study cases N=105		Control Group N=105		P value of chi Square
		Count	Column N %	Count	Column N %	
Fruit	<2 Weekly	53	50.5%	13	12.4%	0.000
	≥2 Weekly	52	49.5%	92	87.6%	
Vegetables	<2 Weekly	32	30.5%	14	13.3%	0.003
	≥2 Weekly	73	69.5%	91	86.7%	
Red Meat	<2 Weekly	49	46.7%	66	62.9%	0.018
	≥2 Weekly	56	53.3%	39	37.1%	
Grilled Red Meat	<2 Weekly	53	50.5%	73	69.5%	0,005
	≥2 Weekly	52	49.5%	32	30.5%	
Red Meat outside appearance	Lightly browned	71	67.6%	80	76.2%	0.512
	Medium browned	22	21.0%	16	15.2%	
	Heavily browned or blackened	5	4.8%	5	4.8%	
	Don't eat grilled red meat	7	6.7%	4	3.8%	
Red Meat inside appearance	Red	45	42.9%	55	52.4%	0.057
	Pink	9	8.6%	16	15.2%	
	Brown	45	42.9%	27	25.7%	
	I don't eat red meat	6	5.7%	7	6.7%	
Chicken	<2 Weekly	54	51.4%	48	45.7%	0.686
	≥2 Weekly	51	48.6%	57	54.3%	
Grilled Chicken	<2 Weekly	94	89.5%	88	83.8%	0.223
	≥2 Weekly	11	10.5%	17	16.2%	
Chicken outside appearance	Lightly browned	76	72.4%	88	83.8%	0.134
	Medium browned	21	20.0%	12	11.4%	
	Heavily browned or blackened	0	0.0%	0	0.0%	
	Don't eat grilled chicken	8	7.6%	5	4.8%	

The study analyzed the Body Mass Index (BMI) responses for CRC cases. The study found that 3% only of CRC patients are within the normal BMI range, while 59% are classified as overweight, and 38% are obese, as shown in figure (23). However, Johnson (2013) supported a moderate association between abnormal BMI and CRC incidence. Also, Ghrouz and El Sharif (2019) found that there BMI and CRC incidence is statistically insignificant.

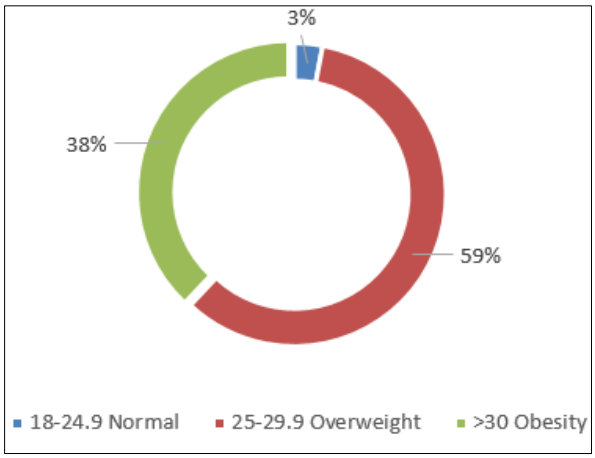


Figure (23): CRC Cases BMI Evaluation (AQU dataset)

The WHO, as well as many other researchers, as shown in the literature review chapter, supported that, Physical Activity (PA) rate of an individual is strongly related to Colorectal cancer prevalence, where the relation between PA and CRC prevalence is an inverse relationship. Therefore, the study analyzed the level of PA among CRC patients included within Al Quds university dataset. The study found that only 4% of CRC patients used to carry out a moderate PA rate with less than 600-1500 Metabolic Equivalent of Task (MET) per Min/Week, while 96% scored low PA rate with less than 600 MET-Min/Week as shown in figure (). In addition, Ghrouz and El Sharif (2019) found that PA rate has statistical significance with CRC incidence.

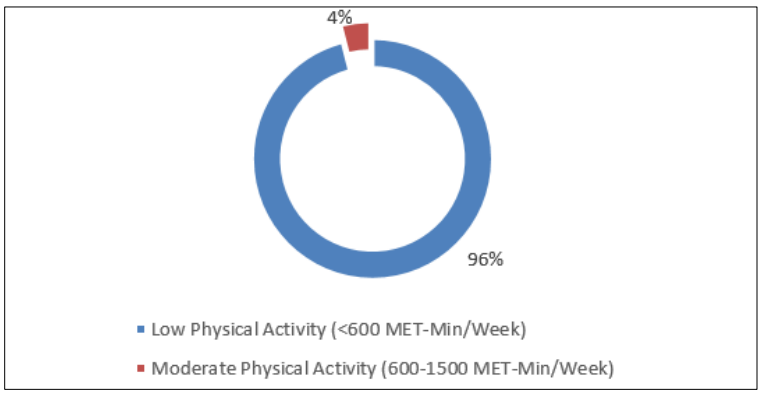


Figure (24): CRC cases Physical Activity Rate Assessment (AQU Dataset)

Regarding smoking, the carried research found that 26% CRC patients are smokers, while 74% of them are non-smokers, as shown in figure (25). Therefore, this leads us to approve that smoking increases the probabilities of CRC occurrence, where Lee et. Al (2019) support that smoking and alcohol consumption are associated with colorectal cancer. In addition, Ghrouz and El Sharif (2019) found that the relationship between smoking and CRC occurrence is statistically significant.

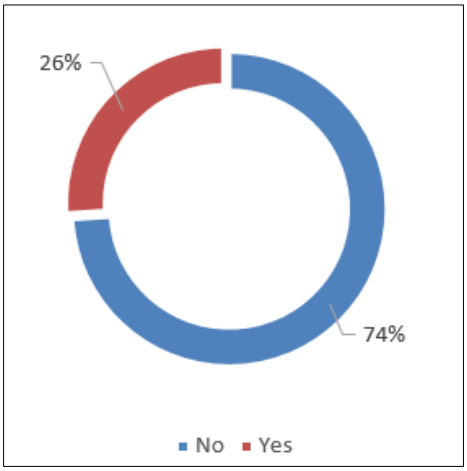


Figure (25): CRC Cases Smoking Percentage (AQU Dataset)

Since gender is one of the risk factors as proved in the literature, and there is a debate on considering smoking as a risk factor for CRC, the study analyzed smoking and gender combined as shown in table (4). Therefore, after both datasets are included in this basic descriptive analysis, there might be no relationship between gender and smoking combined with CRC incidence, where 31% of CRC cases are males and non-smokers, 31% of the CRC cases are females and non-smokers, while 13% all CRC cases are males and smokers, and 14% of them are female cases and smokers. On the other hand, within Al Quds university dataset, 45% of all CRC cases are females and non-smokers, while 29% of the cases are males and non-smokers, also 26% are males and smokers.

Table (4): Results of Combining Smoking and Gender as one risk factor

	MOH DATASET	AQU DATASET
MALE SMOKERS	13%	26%
MALE NON-SMOKERS	31%	29%
MALE BUT SMOKING UNKNOWN	6%	0%
FEMALE SMOKERS	14%	0%
FEMALE NON-SMOKER	31%	45%
FEMALE BUT SMOKING UNKNOWN	5%	0%

From another side, the data analysis for family history and CRC within Al Quds university dataset showed a statistical significance for the positive relationship between an individual's family history and the chances of developing CRC, where as shown in table (5) all *P_value* for Chi-Square for the different family histories scored less than 0.1.

Table (5): P_Value for Chi-Square between Family History and CRC

NO.	VARIABLE	P_VALUE OF CHI-SQUARE
1	Consanguinity	0.000
2	Are you a twin, triplet or other multiple birth siblings?	0.031
3	Relation with family colorectal cancer patient	0.090
4	Family colorectal cancer	0.030
5	Other family cancers	0.009
6	Type of other family cancers	0.029
7	Relationship with family other cancer patients	0.034

4.2. Experimental Results

4.2.1. National Cancer Registry Dataset Experimental Results

After describing and visualizing of the processed data in the previous section, this section will present and discuss a detailed result for applying different Artificial Intelligence tools on the national cancer registry.

4.2.1.1. Decision Tree (MoH Dataset)

The area under the curve or AUC is that which estimates the ability to classify the CRC Patient or not. For the national cancer registry dataset, AUC of 0.97 has been obtained, from which it could be said that the DT models are suitable since it classifies 95.4% CRC risk factors, as shown in figure (26)

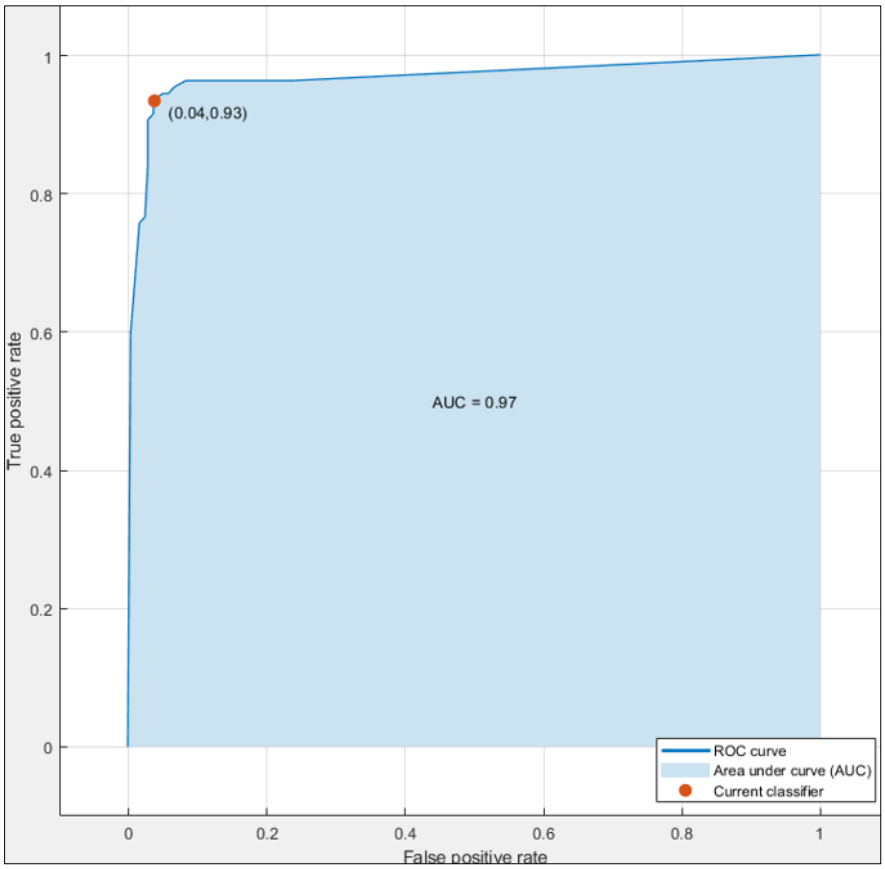


Figure (26): Decision Tree Area Under Curve (MoH Dataset)

Consequently, the true positive for DT scored 93%, while false negative scored 4%. However, the true negative scored 96% and the false-positive scored 7%. Thus, the sensitivity for DT is 0.96, and the specificity score is 0.93, as shown in figure (27), which includes the confusion matrix produced after applying DT on the national cancer registry dataset.

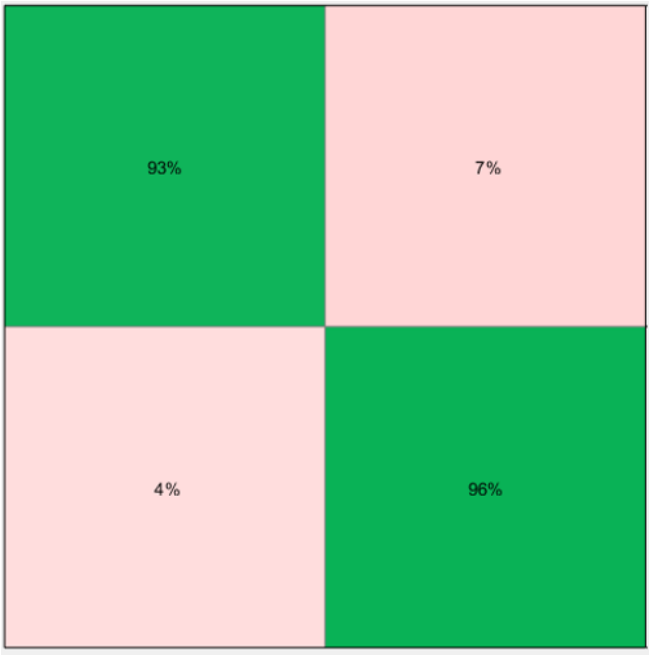


Figure (27): Decision Tree Confusion Matrix (MoH Dataset)

4.2.1.2. Support Victor Machine (MoH Dataset)

By applying the Support Victor Machine (SVM) on the MoH dataset, the AUC of 0.97 has been obtained which indicating the model's ability to classify the individual, where SVM and DT achieved the sane AUC value, shown in figure (28). But SVM classifies 92% of the cases accurately.

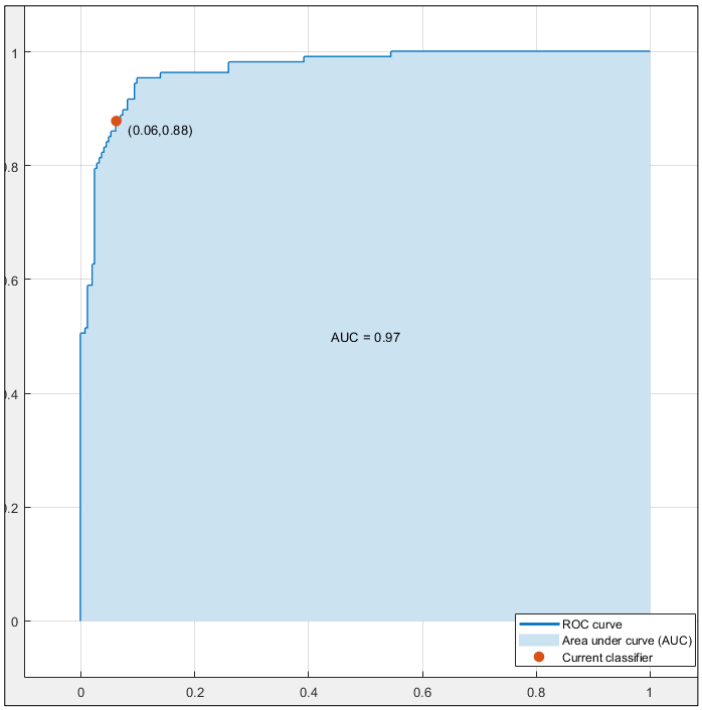


Figure (28): Support Victor Machine AUC (MoH Dataset)

Further, the true positive percent was 88%, while the false-positive percent was 12%. However, the true negative percentage scored 94% and the false-negative scored 6%, as shown in figure (29). Thus, the sensitivity for SVM is 0.94 and the specificity is 0.89.

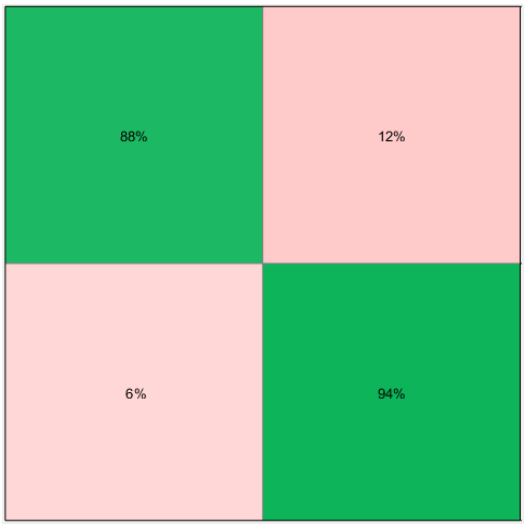


Figure (29): Support Victor Machine Confusion Matrix (MoH Dataset)

4.2.1.3. K-Nearest Neighbor (MoH Dataset)

Using KNN tool revealed a score of (88.8%) accuracy, where the Area Under Curve was 0.89, which is less than DT and SVM as shown in figure (30).

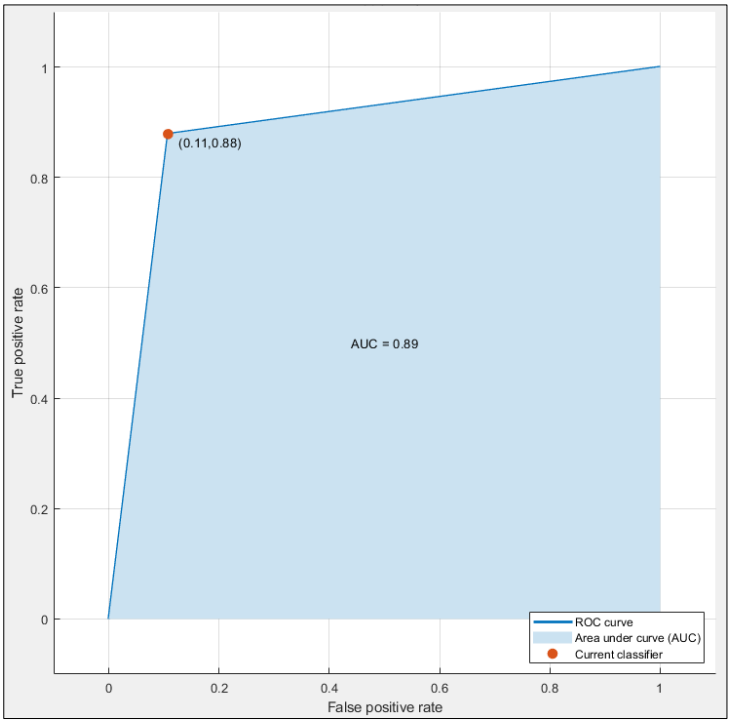


Figure (30): K-Nearest Neighbor AUC (MoH Dataset)

In addition, the true positive was 88% and the false positive 12%, while the true negative 89% and the false negative was 11% as shown in figure (31). This leads us to conclude that the sensitivity for applying KNN on the national cancer registry dataset is 0.89. The specificity of using it is 0.88 with no significant difference between sensitivity and specify.

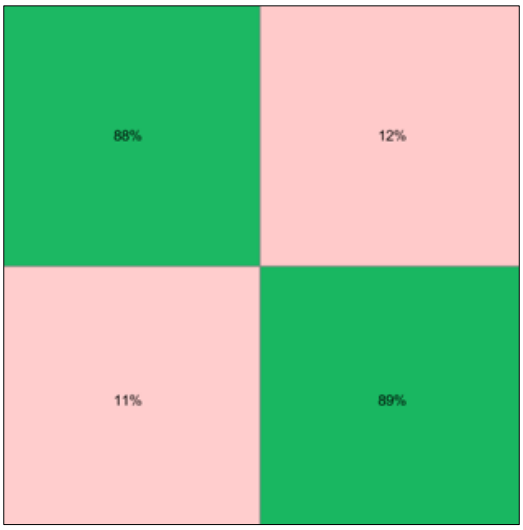


Figure (31): K-Nearest Neighbor Confusion Matrix (MoH Dataset)

4.2.1.4. Logistic Regression (MoH Dataset)

When Logistic regression was applied, it achieved the lowest accuracy score among the other artificial intelligence tools, where the total accuracy rate was 85.8%, with AUC 0.84 as shown in figure (32).

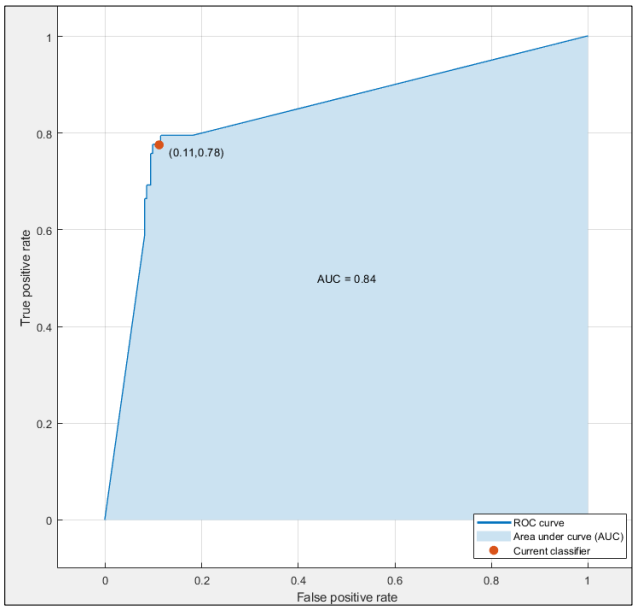


Figure (32): Logistic Regression AUC (MoH Dataset)

After studying the confusion matrix that is shown in figure (33), we can notice that Logistic Regression scored the worst results of true positive and false positive cases' detection with 78% and 11% accordingly, while Logistic Regression score as same as KNN in true negative and false negative cases' detection with 89% and 11% accordingly as well. This leads us to infer that the sensitivity of applying Logistic Regression is 0.88, while the specificity is 0.80.

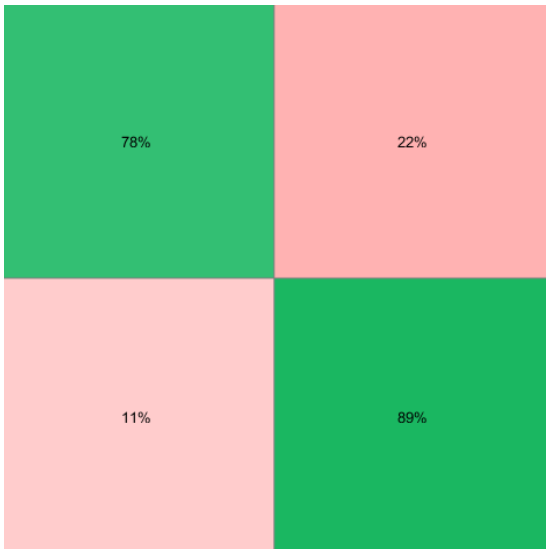


Figure (33): Logistic Regression Confusion Matrix (MoH Dataset)

4.2.1.5. Artificial Neural Network (MoH Dataset)

Since Artificial Neural Network (ANN) works in different way and it relies on neurons (processors) to get trained and then they process the testing procedure, during the study carried four different experiments, where firstly the study carried 5 neurons, secondly 10 neurons, thirdly 15 neurons and finally 20 neurons. Consequently, we compared the carried ANN experiments as summarized in table (6). More details on figures for details related to each single experiment can be found in appendix (3).

After the ANN experiments were carried, we found that the ideal number of neurons used on this dataset is 15. It will be the most efficient, since it will decrease the processors' cost and processing time to produce the most accurate prediction result.

Table (6): ANN Accuracy Table (MoH Dataset)

NO. OF NEURONS	ACCURACY
5	87.8%
10	88.3%
15	94%
20	90%

Further, after analyzing the produced confusion matrix for ANN with 15 neurons as shown in figure (34), we can conclude that the sensitivity score is (96%) while the specificity score is (90%).

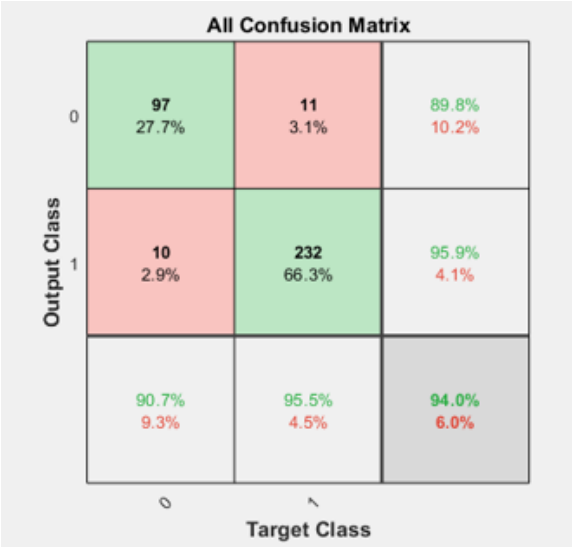


Figure (34): ANN with 15 Neuron Confusion Matrix (MoH Dataset)

4.2.1.6. Automated Feature Selection Results (MoH Dataset)

In order to approve the feature selection process that relied on consulting an expert in the CRC field, the study compared the accuracy score after adopting the features selected by the expert, with the accuracy score for two automated feature selection processes. The processes applied are Chi-Squared and Mutual Information (MI). However, the functions applied on Decision Tree and Artificial Neural Network, because they achieved the best accuracy score among the MoH dataset.

In information and probability theory, MI function aims to measure variables' dependence degree by quantifying the obtained information about a specific variable throughout observing the other randomly selected variable. On the other hand, Chi-Squared is a statistical methodology for feature selection in data mining, where it evaluates every variable by measuring its Chi-Square statistics with corresponding class (Pirooznia, et al., 2008) (Bennasar, et al., 2015).

After MI and Chi-Squared performed on MoH dataset, the feature selection methods arranged the features as shown in table (7).

Table (7): MI and Chi-Squared Feature Selection (MoH Dataset)

CHI SQUARED	MUTUAL INFORMATION
Job	Job
Family_history	Family_History
Age_at_diagnosis	Age_at_Diagnosis
PA_rate	PA_Rate
Marital_status	PMH
PMH	Marital_status
Place_of_residency	Place_of_Residency
Work_yes_no	Work_yes_No
~~Smoking_yes_no~~	~~Smoking_Yes_No~~
~~Rand no~~	~~Rand No~~
~~Sex~~	~~Sex~~

Consequently, after eliminating the last three attributes as shown in table (7), and after performing CRC prediction based on the selected features by MI and Chi-Squared function, the study found that accuracy for DT became 94.3%, where it was 95.4% based on the expert consultation. Also, when the ANN with 15 neurons experiment performed after automated feature selection produced 88.3% accuracy score, while it was 94% when the study relied on the expert consultation to select the features. Different figures about MI and Chi-Squared functions can be found in Appendix (6).

From the above accuracy results, and since the accuracy results are close relatively but better accuracy when the study relied on the expert consultation, we can proof that expert's feature selection adopted in the study is better and more reliable.

4.2.1.7. National Cancer Registry Results' Summary

From the previous results and by referring to figure (35), we can conclude that the best tool to be used for this dataset type is Decision Tree, where it scored the highest accuracy, sensitivity and specificity scores, while the worst one to be used within this type of datasets is Logistic Regression. However, ANN scores for accuracy, sensitivity were relatively close to DT's ones.

The above conclusion is reasonable because the decision tree tool is built to produce a tree of the possible decisions as well as their outcome, while logistic regression is considered as a probability modeling tool to predict the class of an object, which in this case, accepts a choice of two when it can't deal with some attributes within the dataset such as age which is scale.

On the other hand, Patel et. al (2020) compared the accuracy of Neural Networks, SVM, Deep Machine Learning and Particle Swarm Optimization (PSO). They found that Deep Machine Learning achieved the first rank in Liver Cancer detection with accuracy 99% while Neural Networks achieved the second rank with accuracy 98.2%, where Neural Networks were better than SVM that scored accuracy 96.72%.

On the contrary, Yan et. al (2018) carried out a study that aims to compare the performance of Decision Trees (DTs), ANNs, and SVMs data mining methods, for a five years CRC patients' survival prediction. They found that SVM is the best tools and DT is the worst.

Each prediction accuracy and data mining process performance depend on the size of the dataset provided (number of attributes and records) and the type of the attributes basically, that's why there is variation in accuracy scores between different studies. Finally, the best AI tool to be used on the Palestinian National Cancer registry dataset for CRC prediction is the ANN.

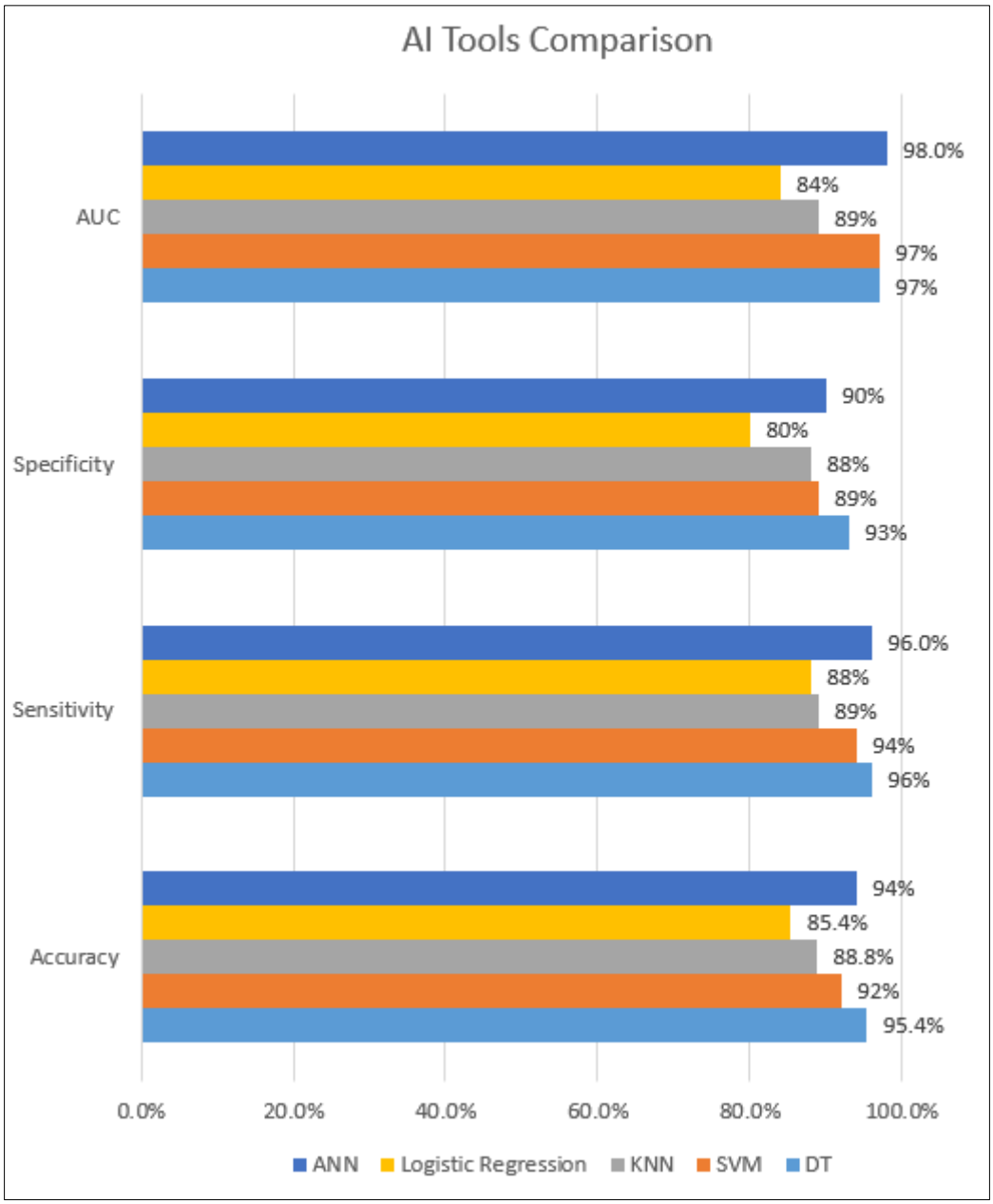


Figure (35): Comparison Between Artificial Intelligence Tools (MoH Dataset)

4.2.2. Al Quds University Dataset Experimental Results

In this section, detailed results for applying different Artificial Intelligence tools on the Al Quds University dataset will be presented and discussed.

4.2.2.1. Decision Tree Results (AQU Dataset)

Decision Tree was applied to Al Quds university dataset, the total accuracy was 94%, where the Area Under Curve (AUC) was (0.97), as shown in figure (36).

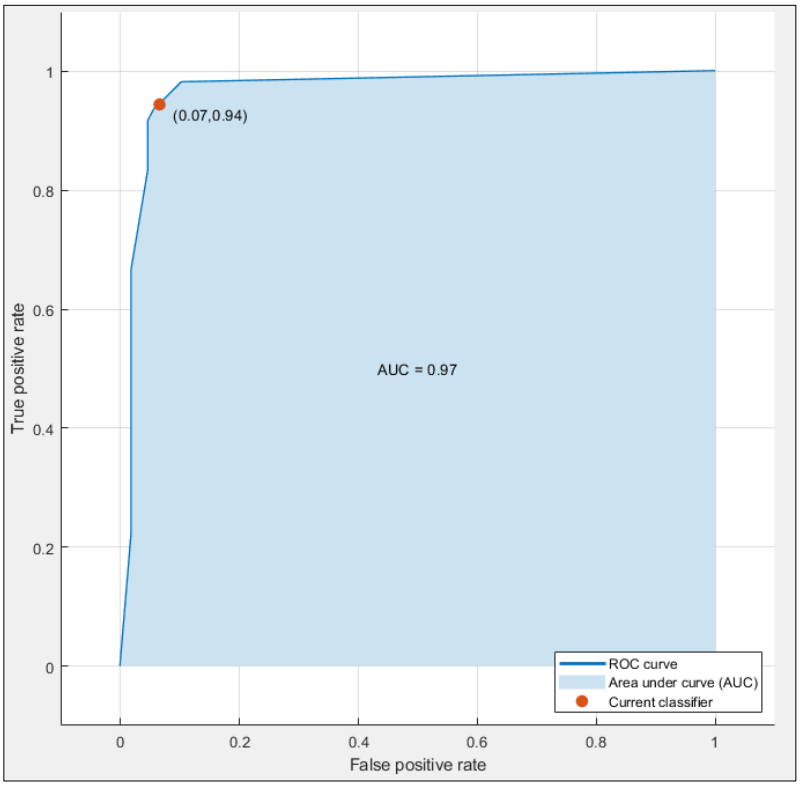


Figure (36): Decision Tree Area Under Curve (AQU Dataset)

The confusion matrix that produced after applying DT on Al Quds university dataset revealed 94% true positive, while false negative scored 7%, as shown in Figure (37). However, the true negative scored 93% and the false positive scored 6%.

Thus, the sensitivity for DT is (0.93), and the specificity score is (0.94) included from the confusion matrix produced after applying DT on AQU dataset.

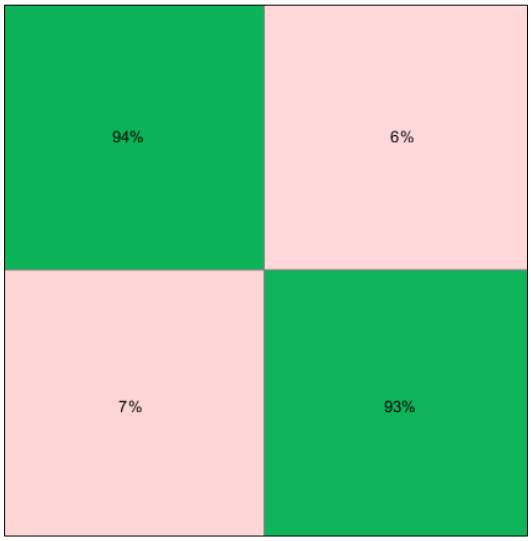


Figure (37): Decision Tree Confusion Matrix (AQU Dataset)

4.2.2.2. Support Victor Machine (AQU Dataset)

Support Victor Machine (SVM) scored total accuracy of 94.9% when applied on AQU dataset, where the AUC was (0.98), as shown in figure (38).

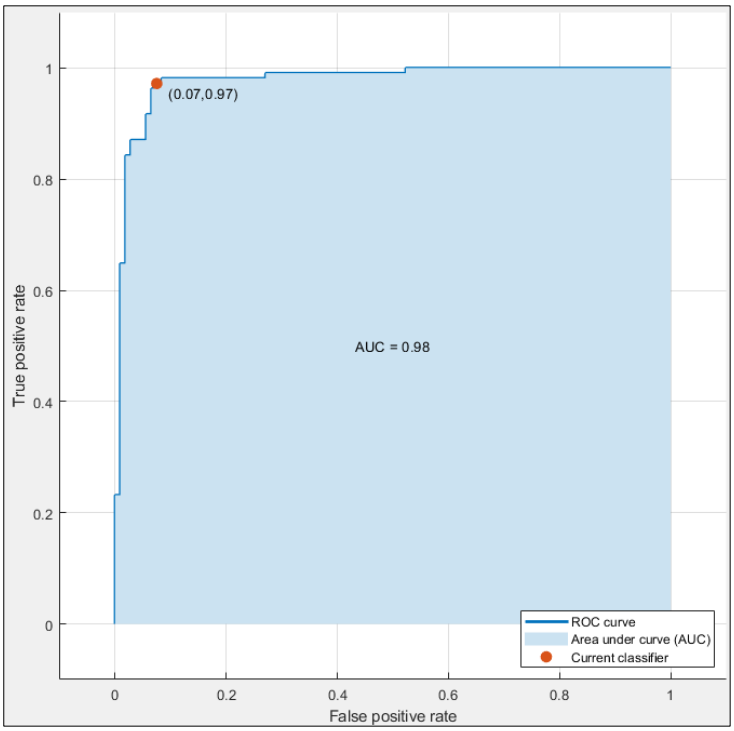


Figure (38): Support Victor Machine AUC (AQU Dataset)

Further, the true positive percent was 97%, while the false positive percent was 3%. However, the true negative percentage scored 93%, and the false negative scored 7%, as shown in figure (39). Thus, the sensitivity for SVM is (0.93) and the specificity is (0.97).

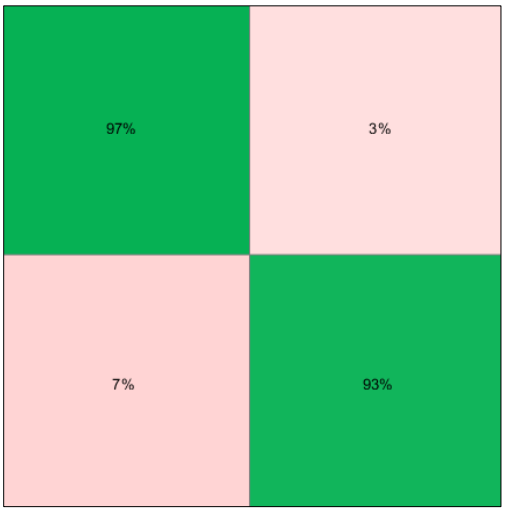


Figure (39): Support Victor Machine Confusion Matrix (AQU Dataset)

4.2.2.3. K-Nearest Neighbor (KNN) Results (AQU Dataset)

KNN tool scored (94%) accuracy, where the Area Under Curve was 0.94 as shown in figure (40). In addition, the true positive was 97%, and the false positive 3%, while the true negative 91%, and the false negative was 9%, as shown in figure (41). This leads us to find that sensitivity for applying KNN on Al Quds university dataset is (0.86), while the specificity of using KNN is (0.97).

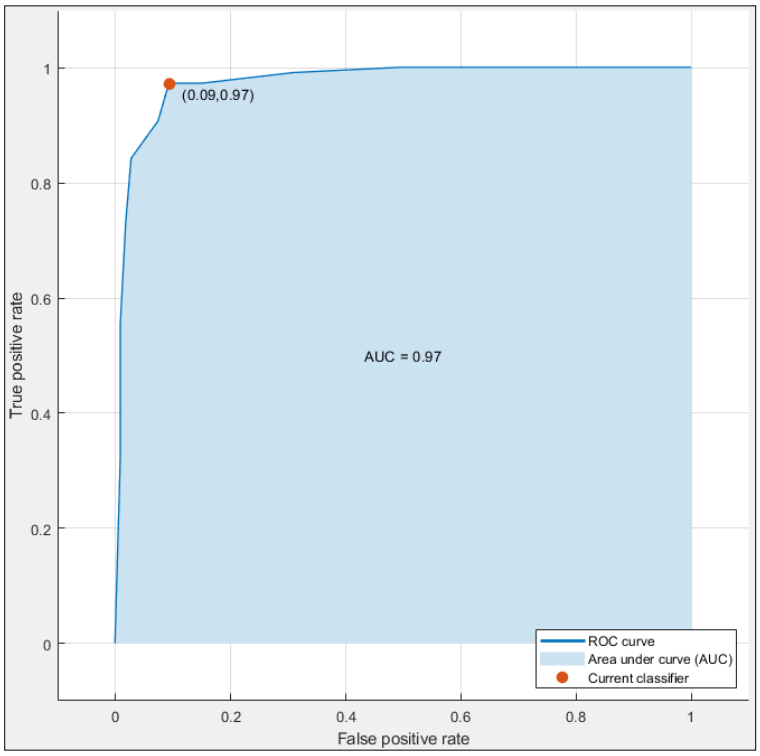


Figure (40): K-Nearest Neighbor AUC (Al Quds University Dataset)

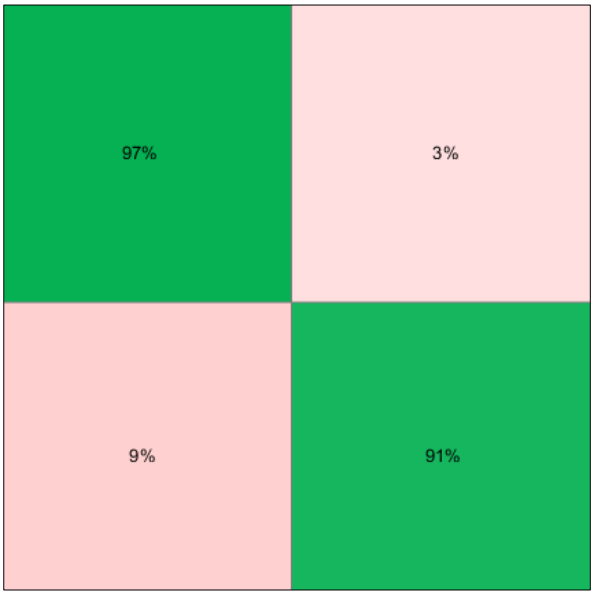


Figure (41): K-Nearest Neighbor Confusion Matrix (AQU Dataset)

4.2.2.4. Logistic Regression (AQU Dataset)

When Logistic regression was applied on the same data set, it achieved the lowest accuracy score among the other artificial intelligence tools. The total accuracy rate was 88.8% with AUC (0.91) as shown in figure (42).

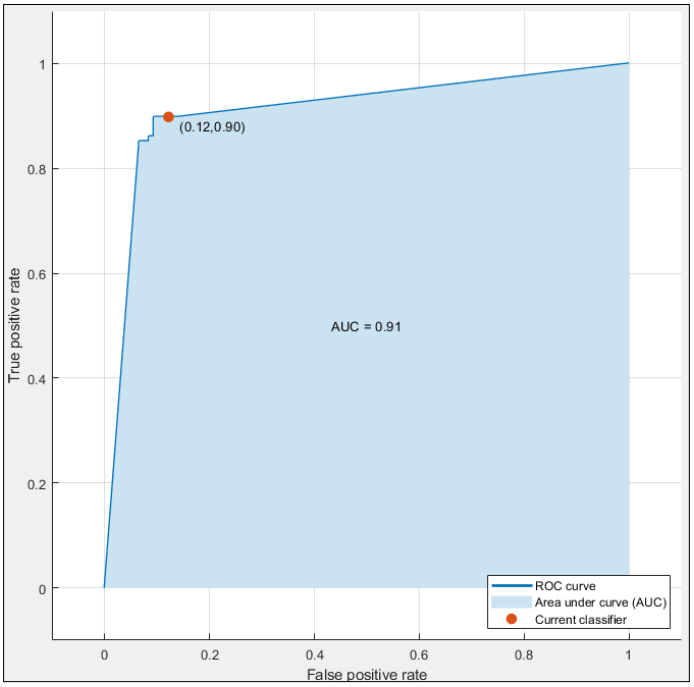


Figure (42): Logistic Regression AUC (AQU Dataset)

In addition, applying the confusion matrix for applying Logistic Regression revealed a true positive of 90% and false positive one with 10%. While the true negative percent was 88%, and the false negative was 12%, as shown in figure (43). This leads us to conclude that the sensitivity for Logistic Regression is (0.88) and the specificity for using it is (0.90)

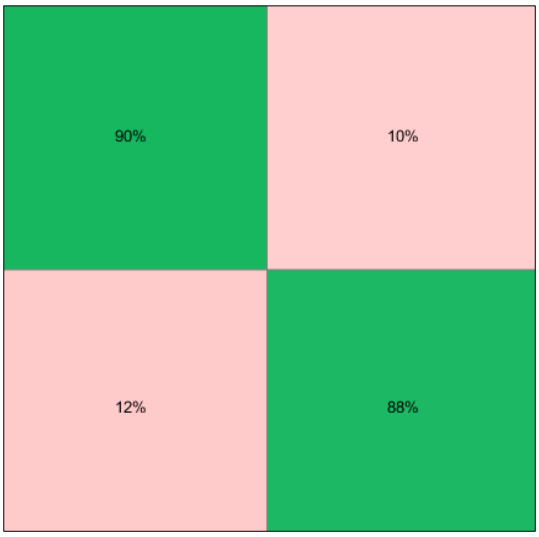


Figure (43): Logistic Regression Confusion Matrix (AQU Dataset)

4.2.2.5. Artificial Neural Networks (ANN) Results (AQU Dataset)

As same as the experiment carried on the national cancer registry, we intended to have the same criteria to assess the performance and accuracy of ANN application on Al Quds university dataset, where the criteria were to use 5 neurons in the first experiment, 10 in the second, 15 in the third and 20 neurons in the fourth experiment. The ANN implementation stopped on the second experiment because the first experiment (with 5 neurons used) achieved 99.5% accuracy and it won't be efficient to use more resources. This excellent accuracy score came out because the dataset is balanced (a case-control), where the number of cases group (CRC cases) and the control group (participants who do not have CRC) are equal.

Using ANN with 5 neurons on Al Quds university dataset is shown in figures (44) and (45). Consequently, we conclude that the sensitivity score is (100%) and the specificity is (99.9%) after adopting ANN with 5 neurons and apply it on the intended dataset

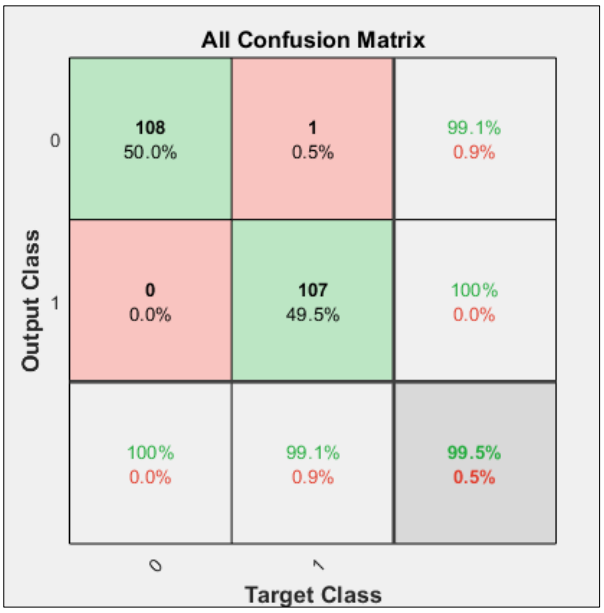


Figure (44): Artificial Neural Network Confusion Matrix Results (AQU Dataset)

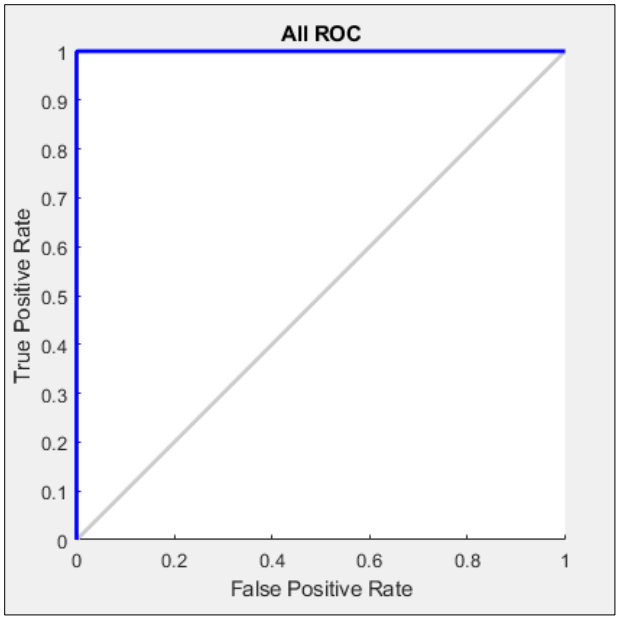


Figure (45): Artificial Neural Network Area Under the ROC Results (AQU Dataset)

4.2.2.6. Automated Feature Selection Results (AQU Dataset)

After MI and Chi-Squared performed on Al Quds university dataset, the feature selection methods arranged the features as shown in table (8).

Table (8): MI and Chi-Squared Feature Selection (AQU Dataset)

CHI SQUARED	MUTUAL INFORMATION
Did Colonoscopy	did Colonoscopy
Did Occult Blood Test	did Occult Blood test
Yearly Income	Yearly Income
Frequency Of Eating Fruits	Frequency of eating Fruits
Type Of Living Area	Type of Living Area
Frequency Of Eating Grilled Red Meat	Frequency of eating Grilled Red Meat
Frequency Of Eating Vegetables	Frequency of eating Vegetables
Smoking	Smoking
Occupation	Frequency of eating Red Meat
Frequency Of Eating Red Meat	Occupation
PA Rate	PA Rate
Age	Frequency of eating Grilled Chicken
Frequency Of Eating Grilled Chicken	Age
BMI	BMI
Do You Suffer From Crohn's Disease	Frequency of eating Chicken
Do You Have Other Cancers ?	do you suffer from Crohn's disease
Frequency Of Eating Chicken	Do you have other cancers ?

~~Educational Level~~	~~Educational Level~~
~~Marital Status~~	~~Marital Status~~
~~Alcohol Consumption~~	~~Do you have Cholesterol~~
~~Do You Have Cholesterol~~	~~Alcohol Consumption~~
~~Work Yes/No~~	~~Work Yes/No~~
~~Did You Remove Gallbladder~~	~~Did you Remove Gallbladder~~
~~Do You Have Diabetes~~	~~Do you have Diabetes~~
~~Gender~~	~~Gender~~

Consequently, after eliminating the last eight attributes as shown in table (8), and after performing CRC prediction based on the selected features by MI and Chi-Squared function, the study found that accuracy for DT became 93.5%, where it was 94% based on the expert consultation. Also, when the ANN with 15 neurons experiment performed after automated feature selection, the experiment resulted in 96.8% accuracy score, while it was 99.5% when the study relied on the expert consultation to select the features. Different figures about MI and Chi-Squared functions can be found in Appendix (6).

From the above accuracy results, and since the accuracy results are close relatively but better accuracy when the study relied on the expert consultation, we can proof that expert's feature selection adopted in the study is better and more reliable.

4.2.2.7. Al Quds University Results Summary

After comparing DT, KNN, ANN, SVM and Linear Regression, we found that the best tool to be applied to Al Quds university data set is the ANN. ANN achieved the best accuracy score among the other four tools. ANN also scored the best sensitivity, specificity and Area Under Curve as shown in figure (46).

The carried experiment showed that, the second ranked tool is SVM with small variation relatively when comparing it with KNN and DT, where DT and SVM, for example, scored the same sensitivity score, while SVM achieved the same specificity score of KNN.

On the other hand, the worst AI tool used in such datasets is Logistic Regression, since this tool is commonly used in order to build models for binary (0/1) dependent variables.

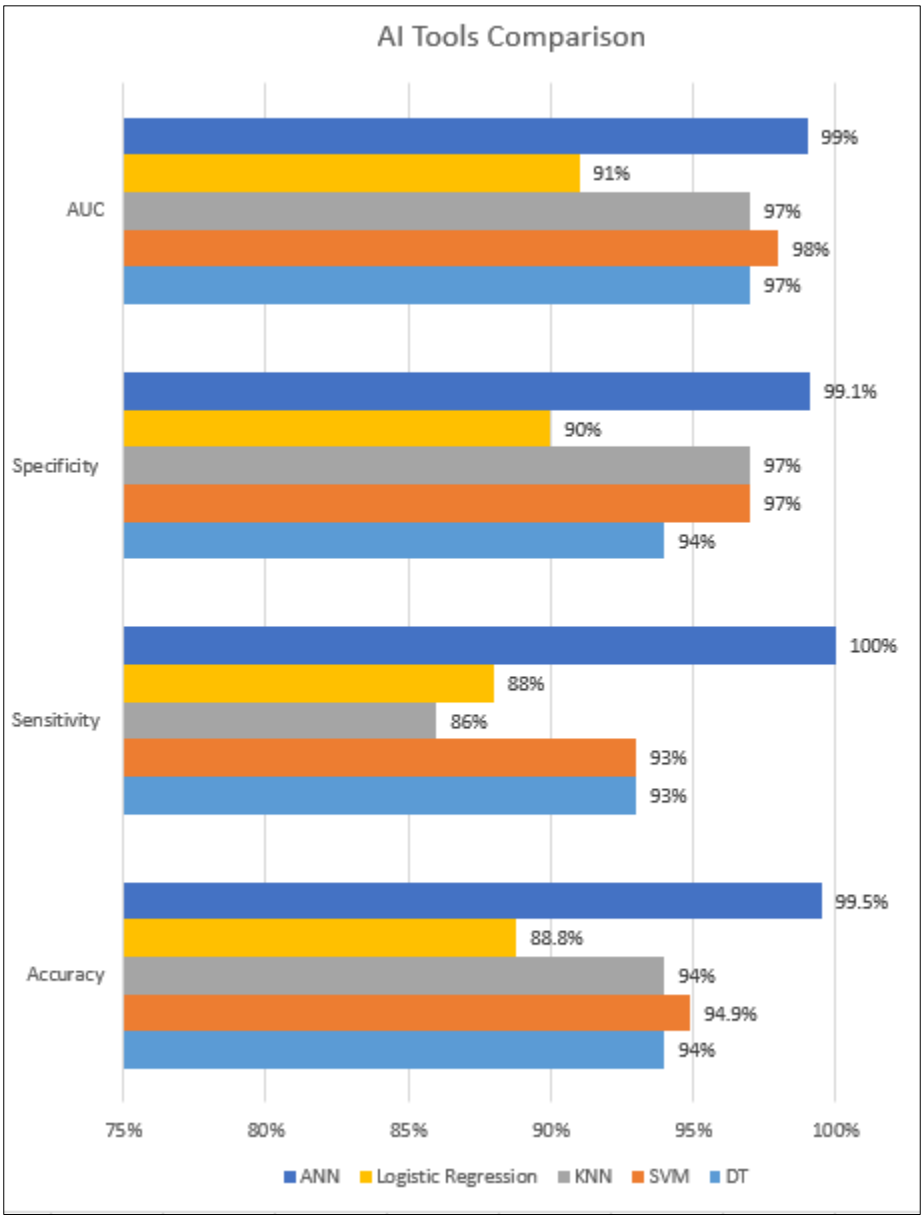


Figure (46): AI Tools Results Comparison (AQU Dataset)

4.3. Experimental Result Summary

When we compare the reported and discussed results in previous sections we found that the best tool to be used in a balanced (matched case-control) dataset for CRC prediction based on risk factors is the Artificial Neural Network (ANN). It will achieve the best accuracy, sensitivity and specificity, rather than a non-balanced (unmatched case-control) dataset.

From another side, if we want to compare Decision Tree (DT) and its performance, we can find that in datasets with fewer attributes DT will produce better accuracy, sensitivity, specificity and Area Under the Curve (AU), where there is a proved inverse relationship between a DT accuracy and the number of dataset attribute, this justify why DT in the national cancer registry scored better accuracy and AUC than Al Quds university dataset.

Chapter Five

Conclusions, Recommendations and Further work

In this chapter, we will present a summary of the conclusions and recommendations of the study. In addition to conclusions and recommendations, further work and intended studies for future work will be presented as well as the strengths and limitations of the of the study.

5.1. Conclusions and Recommendations

Based on our findings, we can conclude that adopting a healthier lifestyle, increasing individuals' engagement with their health and considering preventive factors will decrease Colorectal cancer incidence in Palestine. Adopting healthier lifestyle will enhance the modifiable CRC risk factors such as dietary habits, physical activity, smoking, and obesity.

Also, the study concluded that Artificial Neural Network and Decision Tree are both reliable to be used as methodologies to build a national Colorectal cancer prediction tool, where DT and ANN gained the highest accuracy rates. In this study, and in previous researches, different machine learning algorithms scored between 80% and 99.5% accuracy; thus, these tools are reliable. However, the accuracy score variation between different algorithms is due to dataset size, the dataset organization (balanced or imbalanced), and the nature of the algorithm itself.

In addition, the study concluded that modifiable risk factors such as smoking, physical activity, and diet are CRC risk factors to be considered. From another side, the study also concluded that non-modifiable factors such as family history and past medical history are risk factors for Colorectal cancer occurrence.

Consequently, the following recommendations were formed up after the study was carried:

- Developing a national health policy promotion programs to improve different Palestinian individuals' lifestyle factors, such as Physical Activity will positively affect obesity. For instance, national policy to increasing compulsory sports' and general health education lectures among primary, secondary and higher education institutions within the academic curriculum might be useful. Increasing the physical activity rate will decrease all associated risk factors. This increase might help in swapping students' unhealthy behaviors by healthier ones such as quitting smoking.
- Establishing a well-structured and validated national cancer registry registration records, where special attention should be on gathering a wider variety of data as well as reliable and accurate data about risk factors for Colorectal cancer as well as other types of cancer. Where reliable and accurate data will positively reflect on extracting meaningful and reliable information that will definitely help in improving the health status in Palestine.
- Conducting more scientific researches in order to find the exact reasons behind the high CRC prevalence in Jenin and Tul-Karem areas.
- To carry out a study a comprehensive study aims to assess the impact of different medical history issues as well as diet on Colorectal cancer incidence in Palestine.

5.2. Further Work

By the end of this study, the work to be carried shortly is to design and build a Palestinian Colorectal cancer prediction tool using Artificial Neural Network. It produced the highest accuracy after applying it on two different national datasets.

The lessons learned from this research can be applied to other types of cancer like breast cancer. Future research can be focused on collecting national breast cancer data to make good use of AI and design a national breast cancer prediction tool after the CRC one is built.

Several works can be carried further:

1- To adopt Artificial Neural Networks to build a Palestinian CRC Prediction tool, as shown in figure (47).

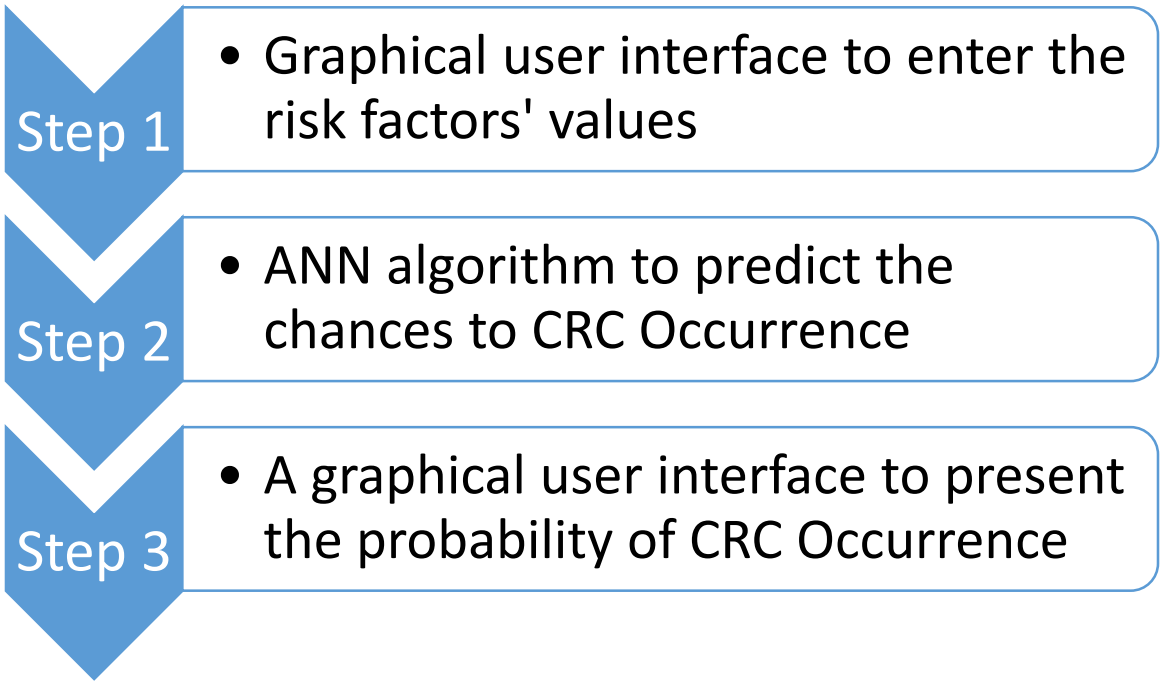


Figure (47): The Intended Palestinian CRC Prediction Tool

The intended tool to be built in the near future will produce a friendly graphical user interface to be used by individuals or healthcare professionals to enter the value of each CRC risk factor, then the system will acquire the entered data and will perform machine learning based on Artificial Neural Network algorithm, in order to present the result as a percentage of CRC occurrence probability.

2- To carry out deeper researches about preventive and risk factors that might affect CRC incidence rate in Palestine.

5.3. Strengths of the Study

Health Informatics is the future of healthcare, and it is also the key to enhance the country's health status if adopted and employed through the right channel. However, Health Informatics is a tool that use different information sciences and other information technology systems in order to produce meaningful information that could be disseminated and utilized among different healthcare settings to improve healthcare services' outcome.

Applying Artificial Intelligence tools are important application of health informatics to serve healthcare, predict patients' health status, and improve decision making.

This study used DT, KNN, SVM, Logistic Regression, and ANN as AI tools to compare which one is the most accurate to predict the probabilities for CRC incidence. Many studies compare two to three tools, but this study compared the most popular five tools.

5.4. Study Limitations

The best data gathering method in such studies is primary data collection. Due to the current situation, the study team did not manage to meet Colorectal cancer patients and gather the primary data to carry out the study.

However, the dataset obtained from the National Cancer Registry contains many irrelevant attributes to the study scope, as well as many missing attributes that are important to be considered as CRC risk factors such as Body Mass Index (BMI) to be able to assess obesity and its impact on CRC in Palestine, also, there were many missing or unknown values for the available attributes. From another side, the records found in both datasets were built based on CRC cases or controls answers, thus the data provided might be inaccurate. In addition, patients MoH's dataset did not include cases from Gaza strip.

From another side, the dataset obtained from Al Quds university only concentrated on Diabetes-Millets and having cholesterol as medical history for the case and control groups. It ignored other medical histories that might increase the risks of developing CRC, such as Hypertension and different heart diseases.

Finally, and in order to allow prediction to be performed by different Artificial Intelligence tools, the dataset should contain records for CRC cases and other records with the same attributes for individuals who do not have CRC, however; the dataset obtained from the National Cancer Registry center contains only CRC cases' records, therefore; the researcher designed a questionnaire to collect data from individuals who do not have the disease, and this consumed effort and time to obtain matched data.

References

- Alasadi, S. & Bhaya, W. S., 2017. Review of Data Preprocessing Techniques in Data Mining. *Medwell Journals*, 12(16), pp. 4102-4107.

- Al-Azri, M. et al., 2019. Awareness of Stomach and Colorectal Cancer Risk Factors, Symptoms and Time Taken to Seek Medical Help Among Public Attending Primary Care Setting in Muscat Governorate, Oman. *Journal of Cancer Education*, Volume 34, pp. 423-434.

- American Cancer Society, 2018. *American Cancer Society.* [Online] Available at: https://www.cancer.org/cancer/colon-rectal-cancer/about/what-is-colorectal-cancer.html
  [Accessed 13 February 2020].

- Arafa, M. A. & Farhat, K., 2015. Colorectal Cancer in the Arab World - Screening Practices and Future Prospects. *Asian Pacific Journal of Cancer Prevention*, 16(17), pp. 7425-7430.

- Asri, H., Mousannif, H., Al Moatassime, H. & Noel, T., 2016. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, Volume 83, pp. 1064-1069.

- Bennasar, M., Hicks, Y. & Setchi, R., 2015. Feature selection using Joint Mutual Information Maximisation. *Expert Systems with Applications*, 42(22), pp. 8520-8532.

- Chaurasia, V. & Pal, S., 2017. A Novel Approach for Breast Cancer Detection using Data Mining Techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(1).

- Chau, R. et al., 2016. Determining the familial risk distribution of colorectal cancer: a data mining approach. *Familial Cancer,* Volume 15, pp. 241-251.

- Cho, Y. et al., 2018. Genetic Risk Score, Combined Lifestyle Factors and Risk of Colorectal Cancer.. *Cancer Research Treatment : Official Journal of Korean Cancer Association.*

- de Rezende, L. F. M. et al., 2018. Physical activity and cancer: an umbrella review of the literature including 22 major anatomical sites and 770000 cancer cases. *British Journal of Sports Medicine ,* Volume 52, pp. 826-833.

- Dekker, E. et al., 2019. Colorectal cancer. *The Lancet,* 394(10207), pp. 1467-1480.

- Deng , L. et al., 2012. Diabetes Mellitus and the Incidence of Colorectal Cancer: An Updated Systematic Review and Meta-Analysis. *Digestive Diseases and Sciences,* Volume 57, pp. 1576-1585.

- Fagunwaa, I. O., Loughreybc , M. B. & Coleman, H. G., 2017. Alcohol, smoking and the risk of premalignant and malignant colorectal neoplasms. *Best Practice & Research Clinical Gastroenterology,* 31(5), pp. 561-568.

- Ghrouz , I. K. & El Sharif , N., 2019. *Determinants of Colorectal Cancer among Patients Attending BietJala Governmental Hospital: A Case-control Study.* Jerusalem: A thesis submitted in partial fulfillment of requirement for the degree of Master of Public Health/School of Public Health/ Al-Quds University .

- Gram, I. T. et al., 2020. Smoking and Risk of Colorectal Cancer may differ by Anatomical Subsite and Sex. *American Journal of Epidemiology.*

- Granados-Romero, J. J. et al., 2017. Colorectal cancer: a review. *International Journal of Research in Medical Sciences,* 5(11), pp. 4667-4676.

- Janardhanan, P., Heena, L. & Sabika, F., 2015. Effectiveness of Support Vector Machines in Medical Data mining. *Journal of Communications Software and Systems,* 11(1), pp. 25-30.

- Johnson, C. M. et al., 2013. Meta-analyses of colorectal cancer risk factors. *Cancer Causes and Control,* Volume 24, p. 1207–1222.

- Kretowski, M., 2019. *Evolutionary Decision Trees in Large-Scale Data Mining.* Swtizerland: Springer.

- Kuipers, E. J. et al., 2015. Colorectal Cancer. *Nat Rev Dis Primers,* 1(15065).

- Kyu, H. H. et al., 2016. Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013. *British Medical Journal (BMJ),* Volume 345.

- Lee, M. et al., 2012. Impact of Physical Inactivity on the World's Major Non-Communicable Diseases. *Lancet,* 380(9838), pp. 219-229.

- Lee, S. et al., 2019. Cigarette smoking, alcohol consumption, and risk of colorectal cancer in South Korea: A case-control study. *Alcohol,* Volume 76, pp. 15-21.

- Liang, H. et al., 2020. Data mining-based model and risk prediction of colorectal cancer by using secondary health data: A systematic review. *Chinese Journal of Cancer Research,* 23(2), p. 242–251.

- Maalouf, M., 2011. Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies,* 3(3), pp. 281-299.

- Marcu, L. G., Boyd, C. & Bezak, E., 2019. Current issues regarding artificial intelligence in cancer and health care. Implications for medical physicists and biomedical engineers. *Health and Technology,* 9(4), pp. 375-381.

- Mármol , I. et al., 2017. Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *International Journal of Molecular Sciences,* 18(197).

- Meneses, J. S., Chavez, . Z. R. & Rodriguez, J. G., 2019. Compressed kNN: K-Nearest Neighbors with Data Compression. *Entropy*, 21(234).

- Miller, C. A., Lafata, J. E. & Thomson, M. D., 2020. The Effects of Personalizing Colorectal Cancer Risk Communication on Risk Perceptions and Health Behavior Intentions: a Randomized Trial of Average-Risk Adults. *Journal of Cancer Education,* pp. 1-9.

- MoH, 2016. *special report on the World Cancer Day,* Ramallah: Palestinian Ministry of Health.

- Nasser, I. M. & Abu-Naser, S. S., 2019. Lung Cancer Detection Using Artificial Neural Network. *International Journal of Engineering and Information Systems (IJEAIS),* 3(3), pp. 17-23.

- Patel, D. et al., 2020. Implementation of Artificial Intelligence Techniques for Cancer. *Augmented Human Research 5*, Volume 6.

- Pirooznia, M., Yang, J. Y., Yang, M. Q. & Deng, Y., 2008. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics,* 9(s13).

- Qumseya, B. et al., 2014. Barriers to colorectal cancer screening in Palestine: a national study in a medically underserved population.. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association,* 12 March, 12(3), pp. 463-469.

- Rado, O. et al., 2019. Performance Analysis of Feature Selection Methods for Classification of Healthcare Datasets. In: K. Arai , R. Bhatia & S. Kapoor, eds. *Intelligent Computing.* Cham: Springer.

- Rawla, P., Sunkara, T. & Barsouk, . A., 2019. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Prz Gastroenterol,* 14(2), pp. 89-103.

- Simões, E. J. et al., 2017. Effectiveness of a scaled up physical activity intervention in Brazil: Anatural experiment. *Preventive Medicine ,* Volume 103, pp. 566-572.

- Sohail, M. N., Jiadong, R., Uba, M. M. & Irshad, M., 2019. A Comprehensive Looks at Data Mining Techniques Contributing to Medical Data Growth: A Survey of Researcher Reviews. In: S. Patnaik & V. Jain , eds. *Recent Developments in Intelligent Computing, Communication and Devices. Advances in Intelligent Systems and Computing.* Singapore: Springer.

- Song, N. et al., 2017. Effects of interactions between common genetic variants and smoking on colorectal cancer. *BMC Cancer,* 17(869).

- Ting, W.-C., Chang, H.-R., Chang, C.-C. & Lu, C.-J., 2019. A Novel Prediction Scheme for Risk Factors of Second Colorectal Cancer in Patients with Colorectal Cancer. *PrePrints* , Volume 1.

- Wang, X. et al., 2018. Influence of Smoking, Body Mass Index, and Other Factors on the Preventive Effect of Nonsteroidal Anti-Inflammatory Drugs on Colorectal Cancer Risk. *American Association for Cancer Research (AACR),* 78(16).

- WHO, 2010. *Global Recommendations on Physical Activity for Health,* Switzerland: WHO.

- Yan, L. et al., 2018. Comparison of three data mining methods in predicting 5-year survival of colorectal cancer patients. *The Journal of China Universities of Posts and Telecommunications,* 25(6), pp. 65-73.

Appendices

Appendix (1): Ministry of Health Approval

State of Palestine
Ministry of Health
Minister's Office

دولـــة فـلـسـطين
وزارة الصحة
مكتب الوزير

20/905/16
62/4/29

الأخ الاستاذ الدكتور علي زيدان أبو زهري حفظه الله
رئيس الجامعة العربية الامريكية

تحية طيبة وبعد،،،

الموضوع: التعاون البحثي بين الجامعة العربية الأمريكية وسجل تسجيل حالات السرطان

تهـديكم وزارة الصـحـة أطيـب التحيـات، وإشـارة الـى كتـابكم بتـاريخ 2020/3/25، والمتعلـق بالموضـوع المذكور أعلاه، يُرجـى العلم أنـه لا مـانع لدينـا من تواصـل الطالب محمد أبو زهري مـع الإدارة العامـة للسياسـات الصحيـة للحصـول علـى مـا يلـزم من معلـومات ، الـدكتورة علا العكر موبايـل رقم: 0562-402217، على أن يزودنا بما آلت اليه الدراسة.

وتفضلوا بقبول فائق الاحترام والتقدير،،،

الدكتورة مي سالم الكيلة
وزير الصحة

نسخة:
/ ق أ بأعمال مدير عام الإدارة العامة للسياسات الصحية والتخطيط المحترمة

Ministry of Health - Nablus- Tel.: 09/2384771/6 - Fax : 09/2384777
Ministry of Health -Ramallah- Behind Palestine Medical Complex
Tel.: 02/2964183 - Fax : 02-2964182
Ministry of Health - Gaza- Tel.: 08/2866949 - Fax : 08/2866965

وزارة الصحة - نابلس - تلفون : 09/2384771/6، فاكس : 09/2384777
وزارة الصحة - رام الله - خلف مجمع فلسطين الطبي
تلفون : 02/2964183 فاكس : 02/2964182

Appendix (2): The Agreement with Al Quds University Researchers

الجامعـــة العربيـــة الأمريكيـــة

كلية الدراسات العليا

جامعـــــة القـــــدس

كليـــــة الصحـــــة العامة

اتفاقية مشاركة قاعدة بيانات لأغراض البحث العلمي

الطرف الأول:د. شاهيناز نجار عضو هيئة التدريس / الجامعة العربية الامريكية
أ.د. محمد عوض عضو هيئة التدريس / الجامعة العربية الامريكية
و يمثلون المشرف الرئيسي والمشرف المشارك على رسالة ماجستير الطالب **محمد أبو زهري**(ماجستير معلوماتية صحية) المنوي استخدام قاعدة بيانات لأغراض البحث فيها.

الطرف الثاني:الدكتورة نهى الشريف/منسقة برنامج الصحة العامة /جامعة القدس
السيد عيسى عروز /ماجستير صحة عامة

حيث سيقوم الطرف الثاني بتزويد الطرف الأول بقاعدة بيانات تتعلق بمرضى سرطان القولون والمستقيم والتي تم جمعها سابقا عن طريق الطالب عيسى عروز لرسالته الماجستير بأشراف الدكتورة نهى الشريف. وتكونت هذه الدراسة من ١٠٥ حالة و١٠٥ضابط. وتتعلق البيانات بمعلومات ديموغرافية عن المريض، النشاط البدني، التدخين،استهلاك الكحول، الحالة التغذوية (استهلاك اللحوم والخضار وغيرها) ، الوزن و الطول (مؤشر كتلة الجسم)، عوامل وراثية، امراض مزمنة والحالة الصحية للمريض (ادوية معينه، امراض والتهابات لها علاقة في الامعاء).

سيقوم الطرف الأول باستخدام هذه البيانات لأغراض البحث العلمي مع المحافظة على سرية البيانات وعدم استخدامها في أي أغراض أخرى. هذا ويحفظ الطرف الأول احقية البيانات للطرف الثاني على ان يستخدمها الطرف الأول بالشروط التالية:

1. تستخدم البيانات للإجابة على أسئلة بحثية لرسالة ماجستير للطالب محمد زيدان أبو زهري فقط تحت عنوان " تحليل عوامل الخطر والحماية لسرطان القولون و المستقيم باستخدام طرق الذكاء الاصطناعي للتنبوء بالمرض"والذي تم الموافقة عليه من قبل وزارة الصحة الفلسطينية (كتاب رقم ١٢٠/٩٠٥/١٦)
2. يتم ذكر مصدر البيانات في رسالة الطالب والية جمعها والتأكيد من انها جمعت بموافقة وزارة الصحية ويوثق الموافقة في الرسالة (كتاب رقم ١٢٠/٩٠٥/١٦).
3. الورقة البحثية التي ستنشر نتاجا للعمل على قاعدة البيانات المزودة من الطرف الثاني سيكون ترتيب الباحثين من جامعة القدس (الدكتورة نهى الشريف والاستاذ عيسى عروز) على الورقة البحثية بعد اسم طالب الماجستير محمد زيدان أبو زهري ومشرفيه على ان تظهر أسماء الجامعتين اللتين يعمل بهما الطرف الأول والطرف الثاني.

Jenin Tel: +970-4-2418888 Ext.:1471,1472Fax: +970-4-2510810 P.O. Box:240
RamallahTel: +970-2-2941999Fax: +970-2-2941979 Abu Qash - Near Alrehan
E-mail: FGS@aaup.edu ; PGS@aaup.eduWebsite:www.aaup.edu

4. في حال لم يقم الطالب بالنشر خلال عام من تاريخ مناقشة الرسالة, سيقوم المشرفين على رسالة الماجستير في الطرف الأول والمزودين لقاعدة البيانات من الطرف الثاني بالنشر دون الرجوع للطالب.
5. تنتهي صلاحية هذه الاتفاقية بانتهاء هذا البحث ولا يتم استخدام هذه البيانات لاي دراسة أخرى الا بعد الحصول على موافقة خطية جديدة من الطرف الثاني.
6. يحق للطرف الثاني ايقاف النشر في حالة لم يلتزم الطرف الاول بشروط الاتفاقية.
7. تنتهي فاعلية هذه الاتفاقية عند الانتهاء من نشر النتائج ولا يحق للطرف الاول استخدامها لاحقا او مشاركتها مع اي طرف اخر.
8. يتم اتلاف البيانات عند الانتهاء من استخدامها بعد النشر.

د. نهى الشريف

د. شاهيناز نجار

أ. عيسى عروز

أ.د. محمد عوض

أ.محمد أبو زهري

التاريخ:

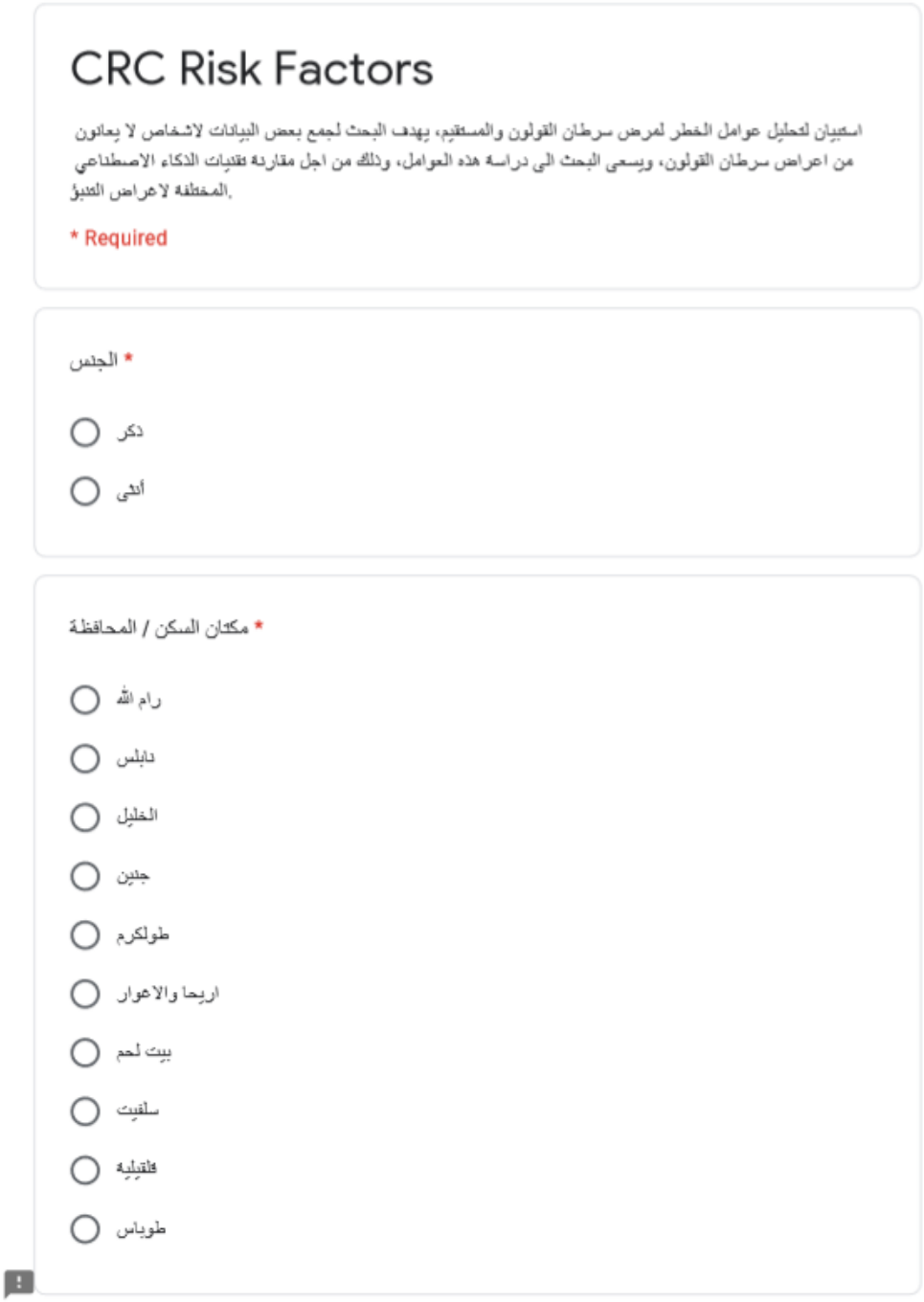
التاريخ:9/7/2020

نسخة عميد كلية الصحة العامة د. اسمى الامام المحترمة

نسخة عميد الدراسات العليا د. عبد الرحمن ابوليده المحترم

Jenin Tel: +970-4-2418888 Ext.:1471,1472Fax: +970-4-2510810 P.O. Box:240
RamallahTel: +970-2-2941999Fax: +970-2-2941979 Abu Qash - Near Alrehan
E-mail: FGS@aaup.edu ; PGS@aaup.eduWebsite:www.aaup.edu

Appendix (3): Questionnaire



CRC Risk Factors

استبيان لتحليل عوامل الخطر لمرض سرطان القولون والمستقيم، بهدف البحث لجمع بعض البيانات لاشخاص لا يعانون من اعراض سرطان القولون، ويسعى البحث الى دراسة هذه العوامل، وذلك من اجل مقارنة تقنيات الذكاء الاصطناعي المختلفة لاعراض التنبؤ.

* Required

الجنس *

ذكر

أنثى

مكان السكن / المحافظة *

رام الله

نابلس

الخليل

جنين

طولكرم

اريحا والاغوار

بيت لحم

سلفيت

قلقيلية

طوباس

* الجنس

- ذكر
- أنثى

* تاريخ الولادة

MM DD YYYY

/ /

* ما هو مستوى التحصيل العلمي ؟

- المرحلة الابتدائية
- المرحلة الثانوية أو توجيهي
- دبلوم
- بكالوريوس
- ماجستير
- دكتوراه

* مجال العمل (الوظيفة)

- بدون عمل
- ربة منزل
- موظف حكومي
- موظف قطاع خاص
- عامل
- مزارع
- طبيب
- ممرض
- مدرّس
- Other:

* هل لديك اي امراض مزمنة (مثل سكري، ضغط الدم .. الخ)

- نعم
- لا

في حال كان لديك مرض مزمن، ما هو ؟

- سكري
- امراض قلب
- ضغط الدم
- Other:

* هل لديك اقارب يعانون من مرض سرطان القولون ؟

- نعم
- لا

في حال كان لديك اقرب يعانون من سرطان القولون، ما هي صلة القرابة ؟

- قرابة درجة اولى (أب، أم، أخ، أخت، ابن)
- قرابة درجة ثانية (أجداد او احفاد)
- قرابة درجة ثالثة (عم، عمّة، خال، خالة، ابن الاخ، ابن الاخت)
- قرابة درجة رابعة (ابن العم، ابن العمّة، ابن الخال، ابن الخالة)

* التدخين

- مدخن
- غير مدخن

في حال كنت مدخن، ما هو معدل استهلاك السجائر / سيجارة باليوم

Your answer

* شرب الكحول - هل تستهلك الكحول ؟

- نعم
- لا

في حال كنت تشرب الكحول، ما هو معدل استهلاكك الكحول

- بكثرة
- بشكل منتظم ولكن بكميات قليلة
- احيانا او في المناسبات

* هل قمت بعمل منظار للقولون والمستقيم من قبل ؟

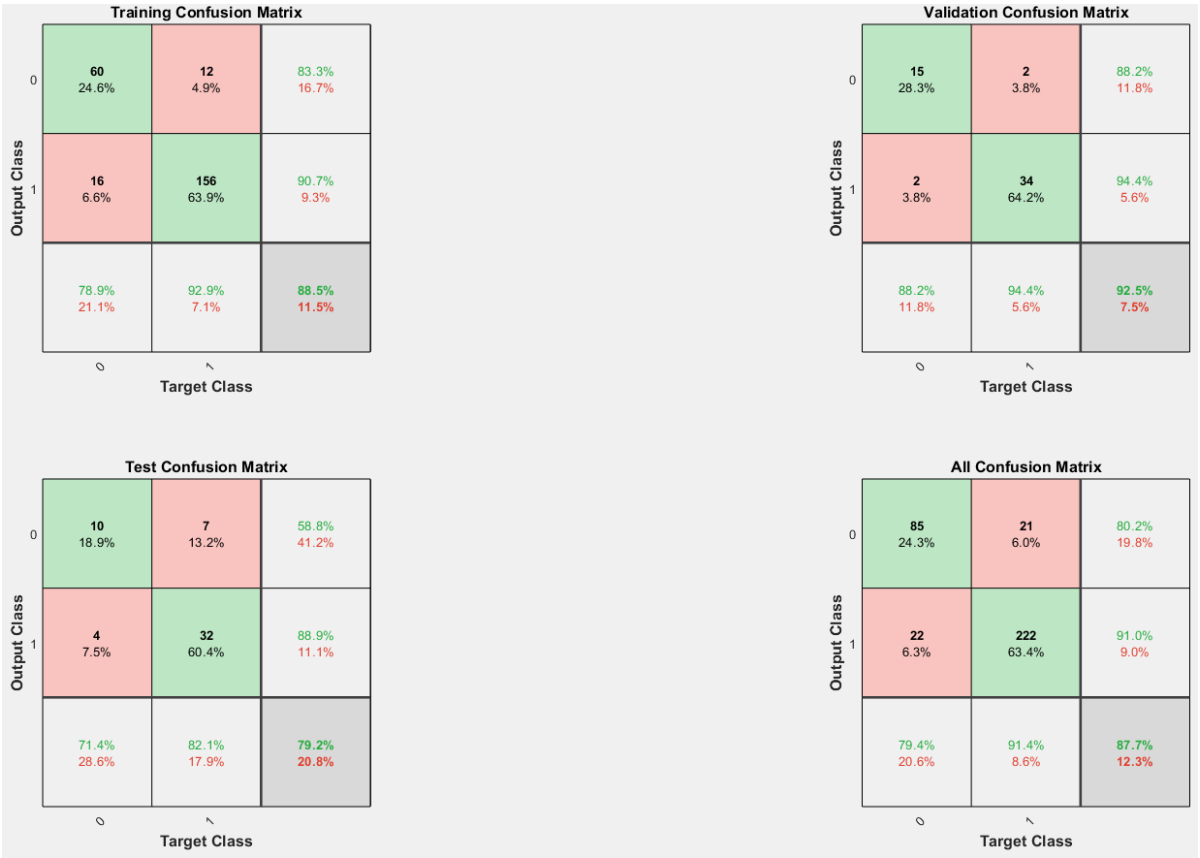
- نعم
- لا

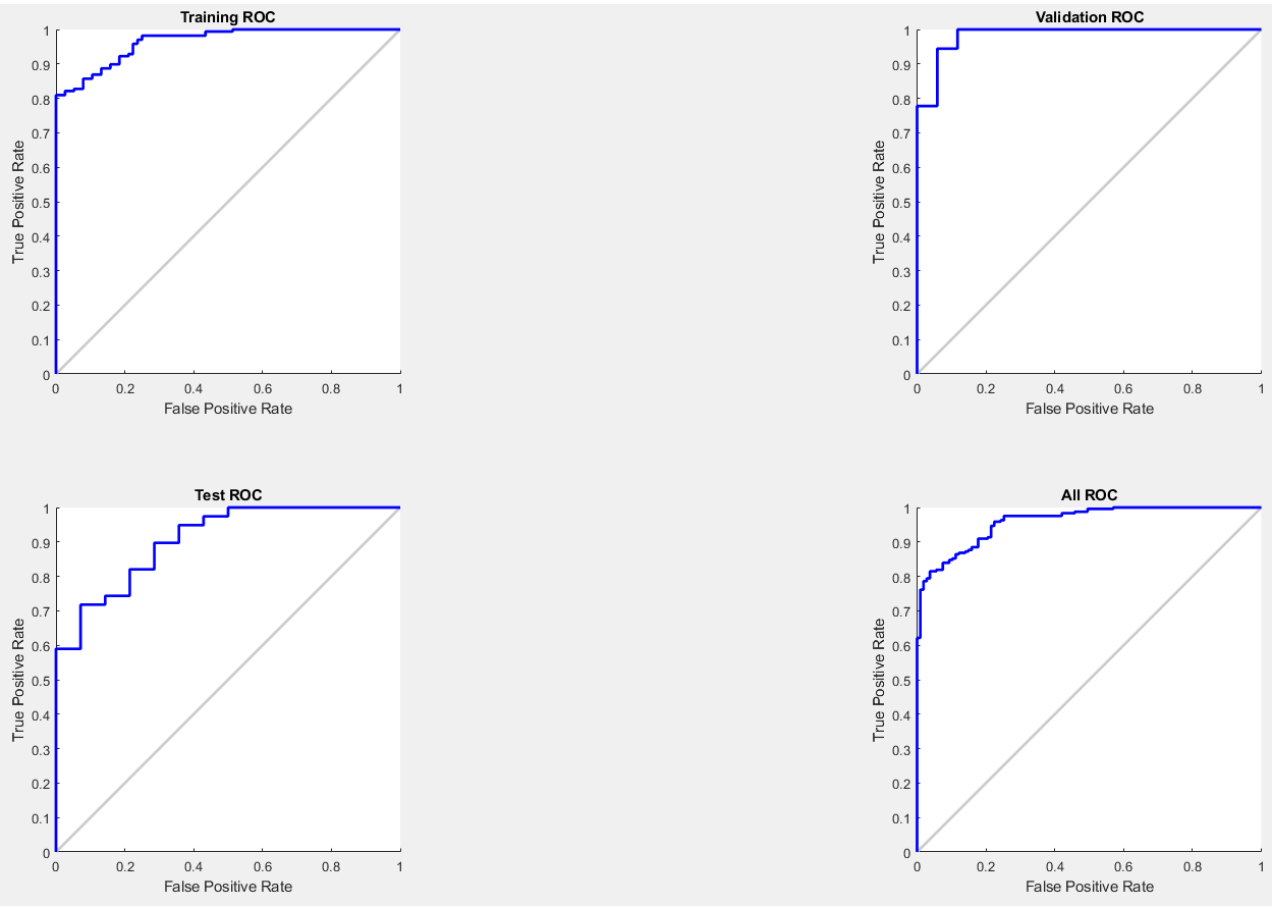
Submit

Never submit passwords through Google Forms.

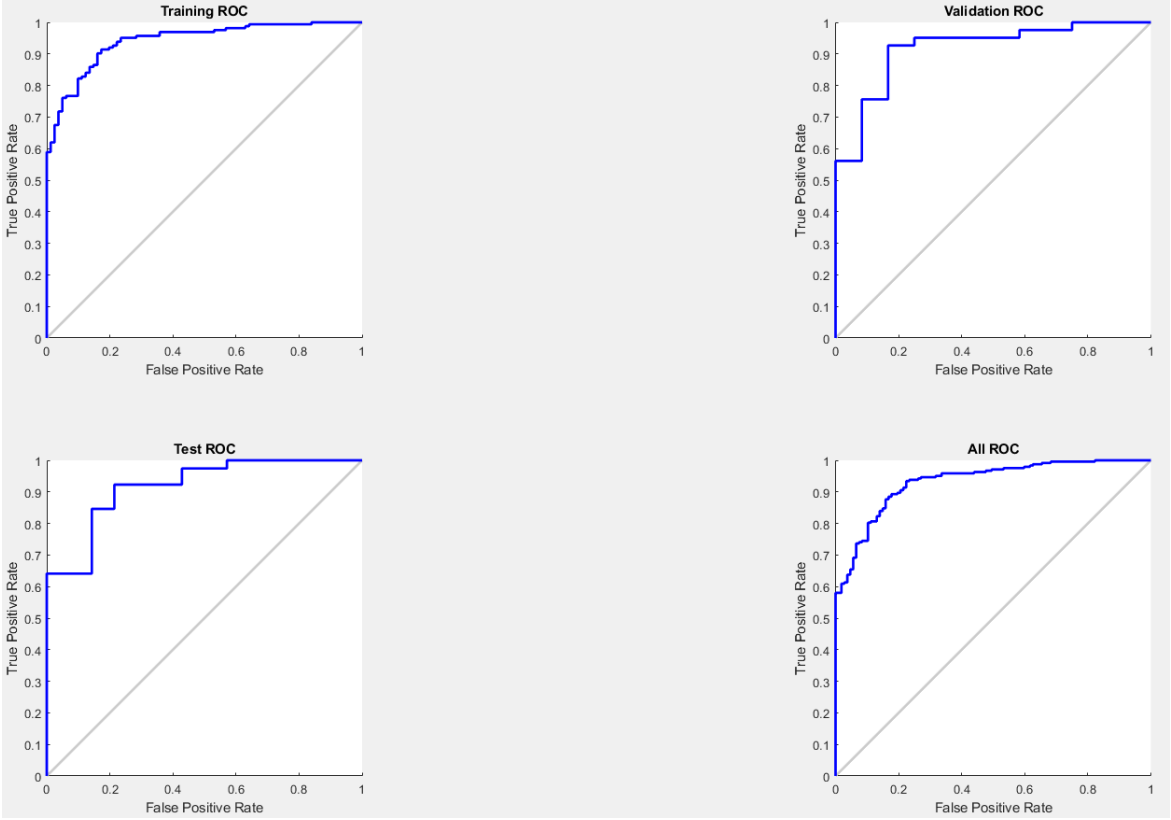
Appendix (4): Artificial Neural Network Experiment Details (MoH Dataset)

- (5) Neurons Experiment

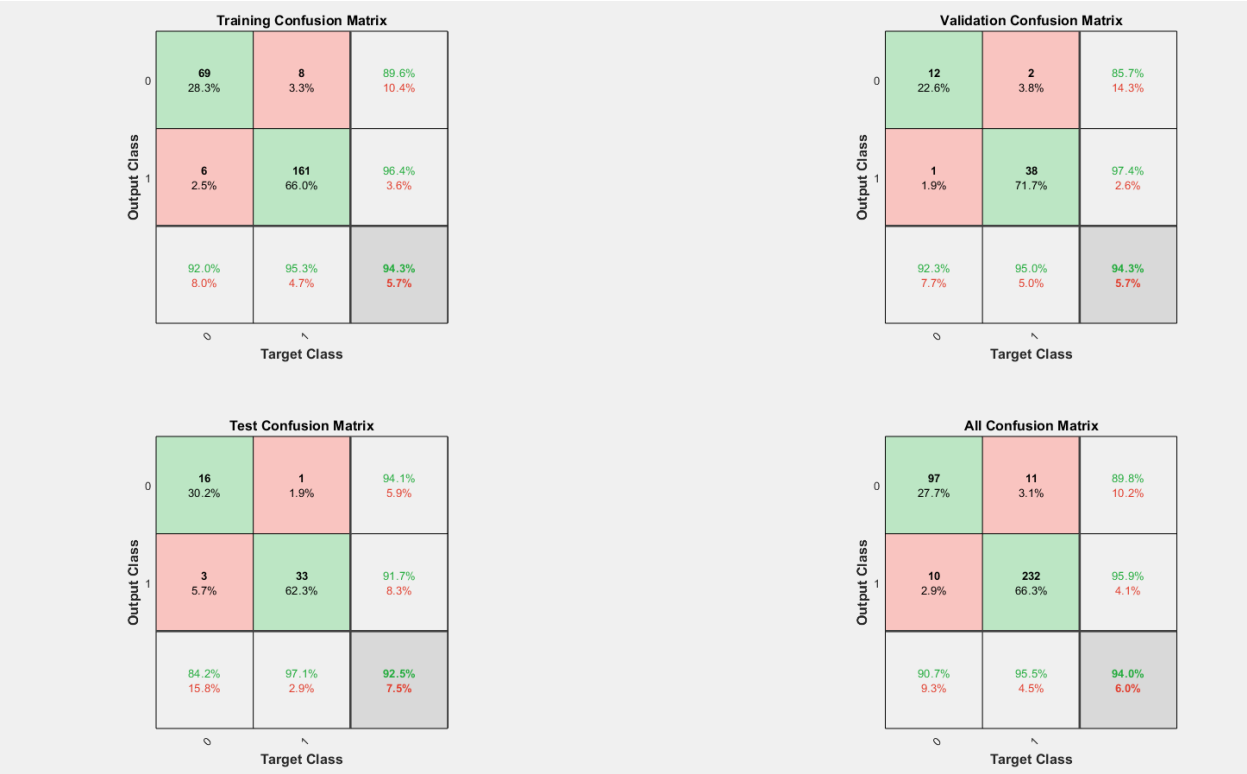


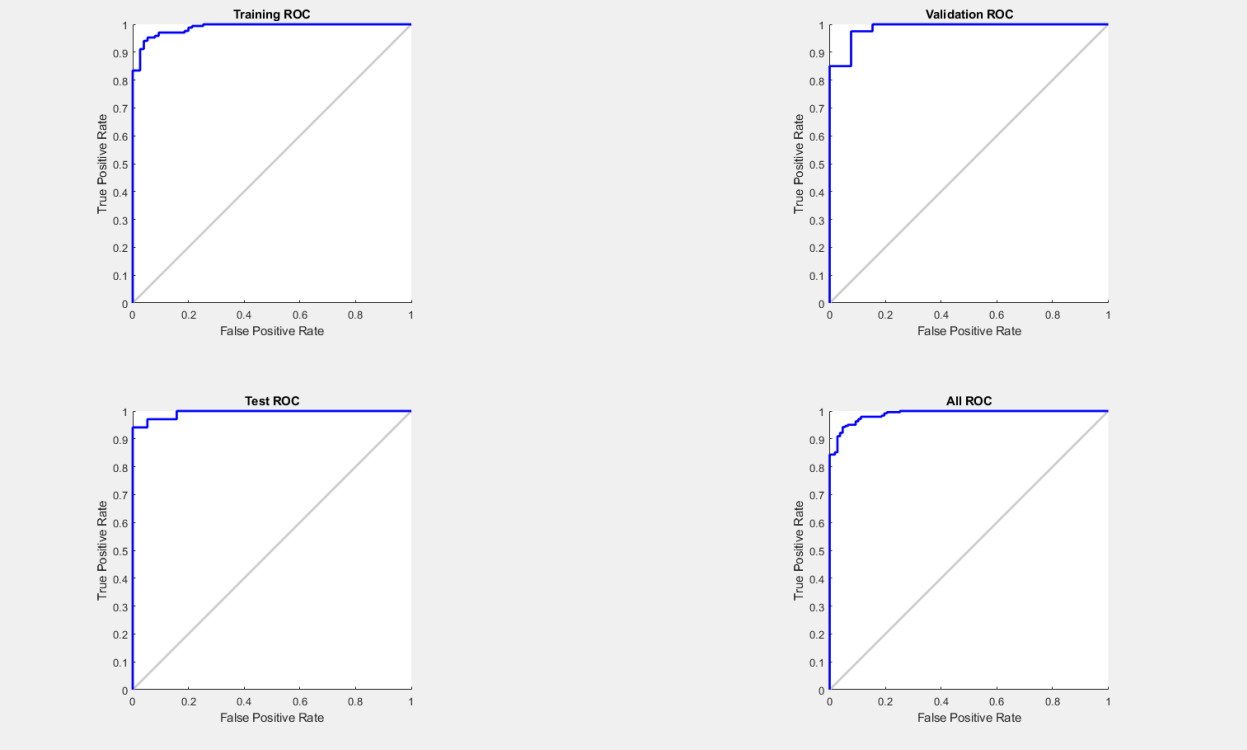

Training ROC
True Positive Rate
False Positive Rate
Validation ROC
True Positive Rate
False Positive Rate
Test ROC
True Positive Rate
False Positive Rate
All ROC
True Positive Rate
False Positive Rate

- (10) Neurons Experiment

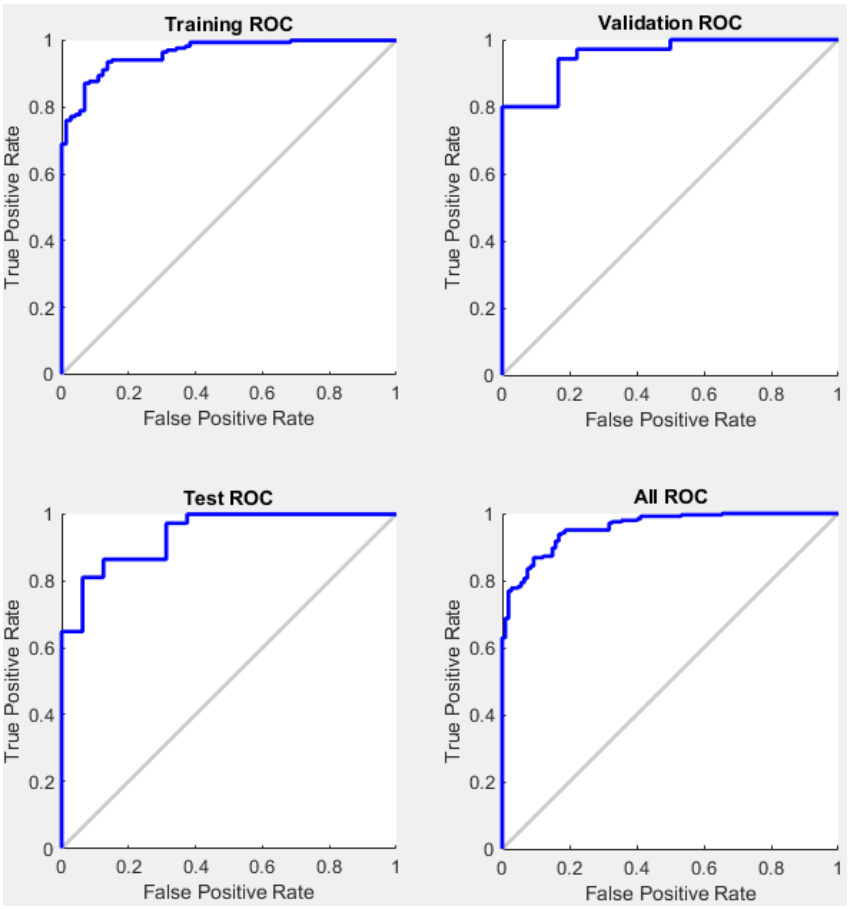

Training ROC
True Positive Rate
False Positive Rate
Validation ROC
True Positive Rate
False Positive Rate
Test ROC
True Positive Rate
False Positive Rate
All ROC
True Positive Rate
False Positive Rate

- (15) Neurons Experiment

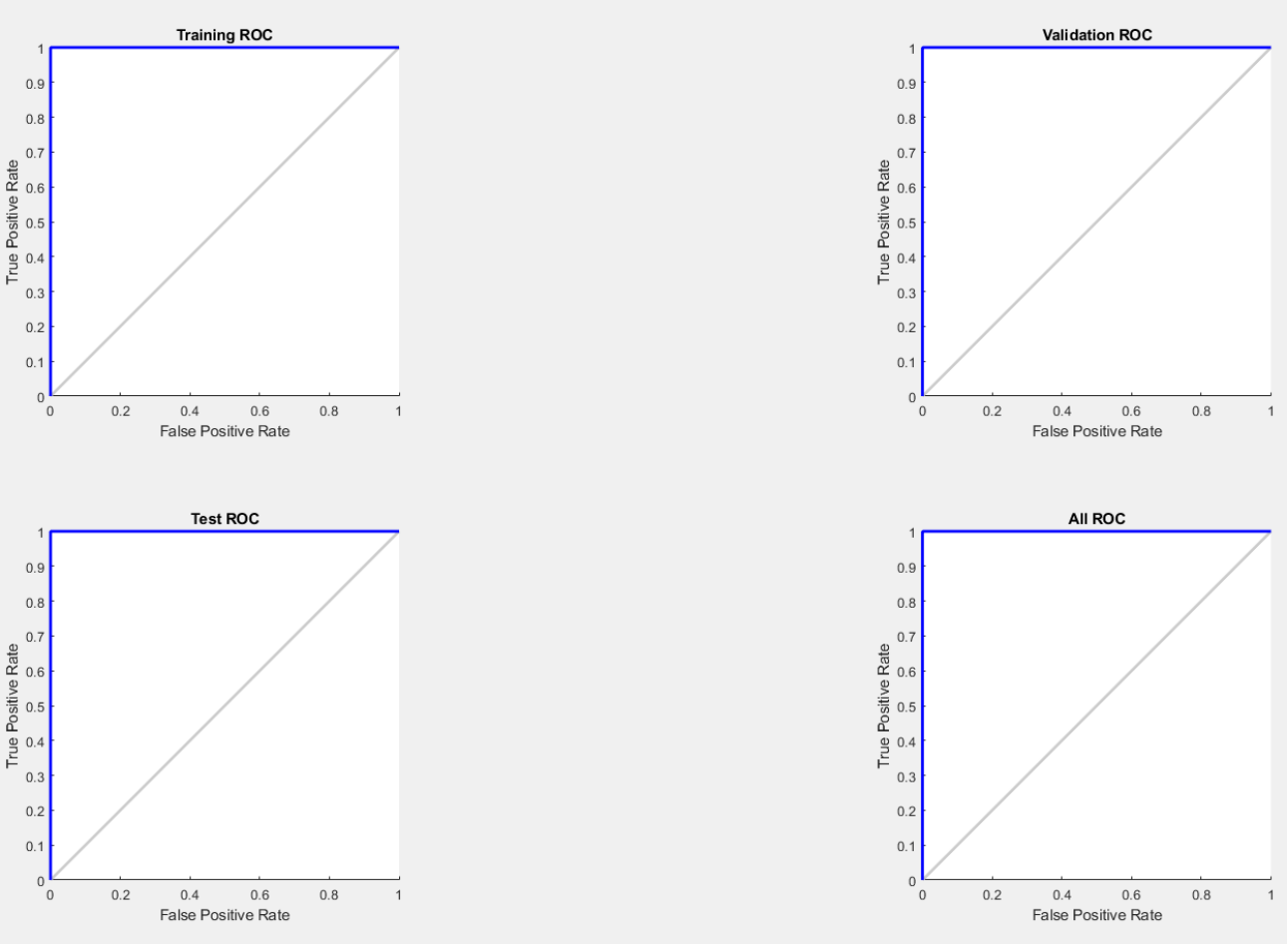




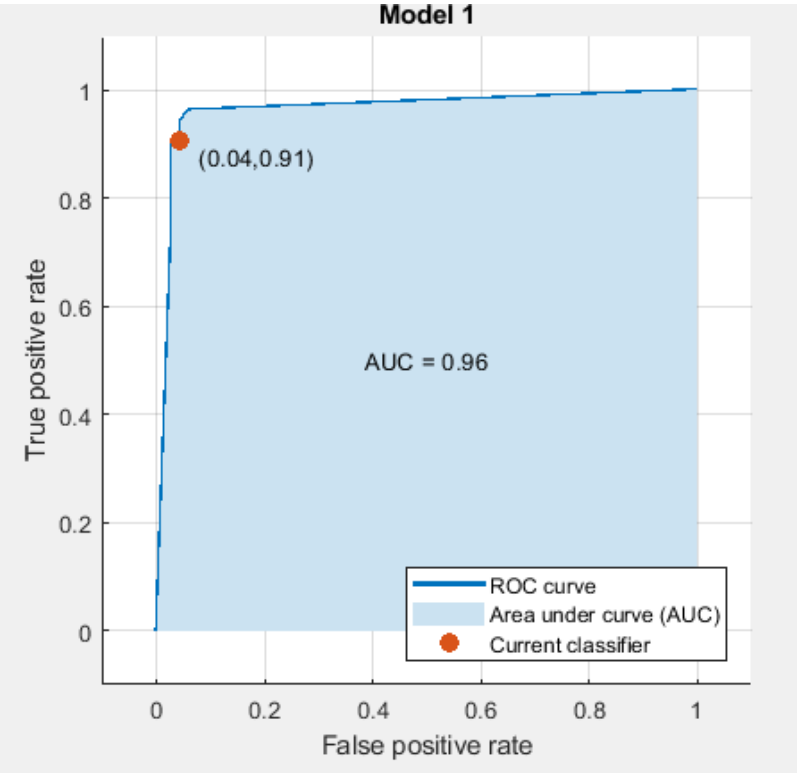
- (20) Neurons Experiment


Training ROC
Validation ROC
Test ROC
All ROC
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1

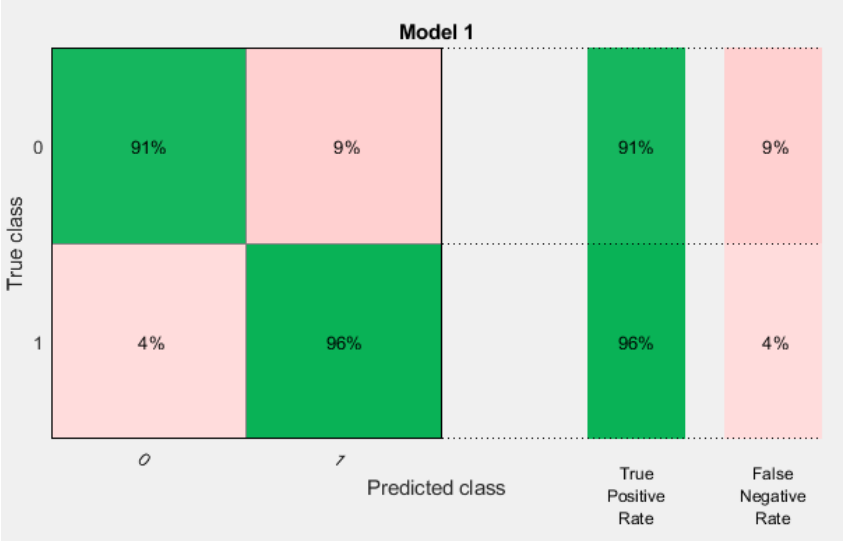
Appendix (5): Artificial Neural Network Experiment Details (AQU Dataset)


Training ROC
True Positive Rate
False Positive Rate
Validation ROC
True Positive Rate
False Positive Rate
Test ROC
True Positive Rate
False Positive Rate
All ROC
True Positive Rate
False Positive Rate

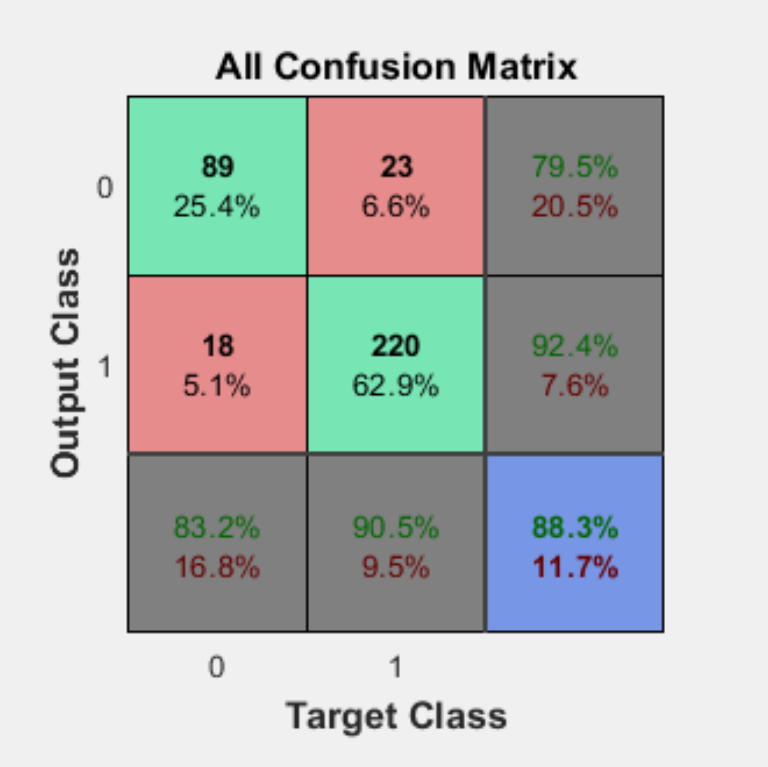
Appendix (6): DT and ANN Results after MI and Chi-Squared Feature Selection



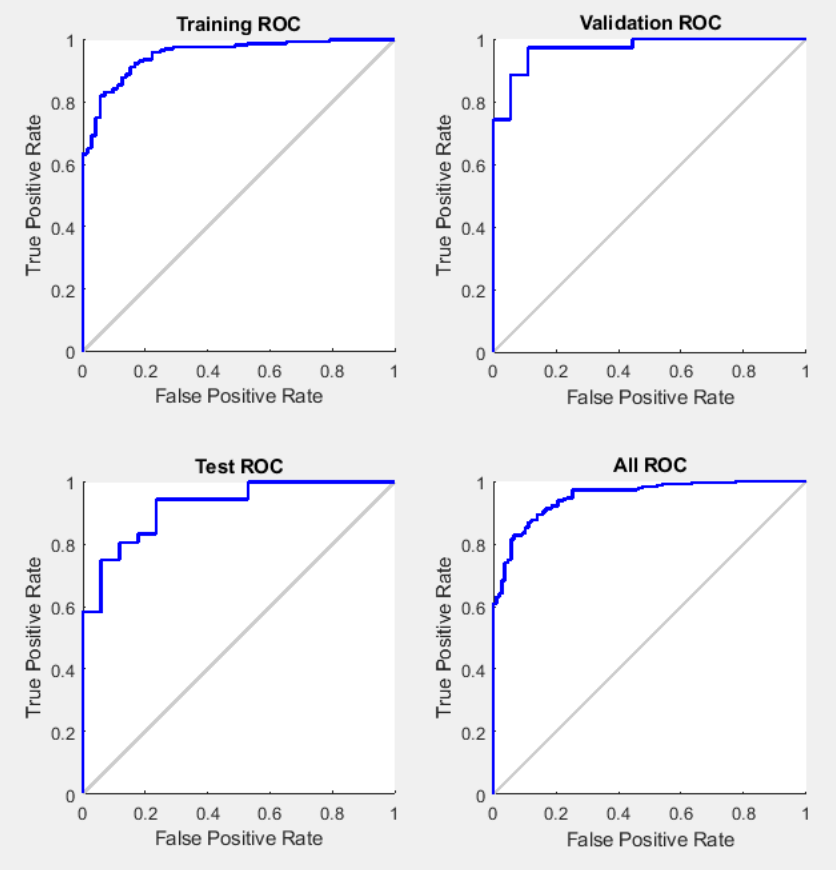
DT AUC After MI and Chi-Square (MoH Dataset)



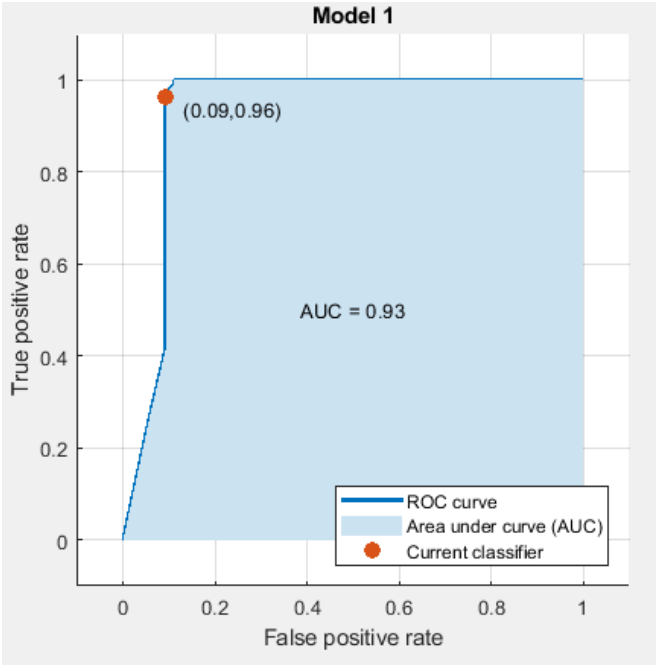
Confusion Matrix for DT After MI and Chi-Square (MoH Dataset)



Confusion Matrix for ANN After MI and Chi-Square (MoH Dataset)

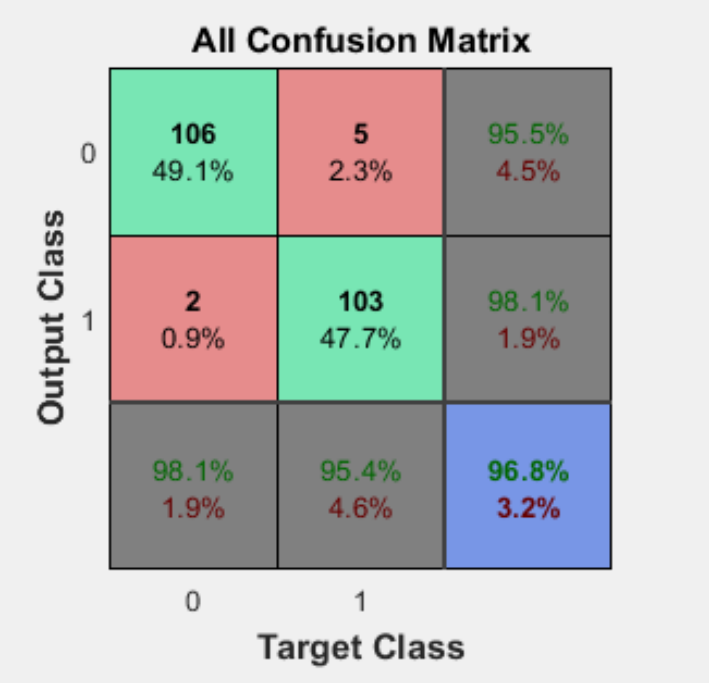


ANN AUCs After MI and Chi-Square (MoH Dataset)

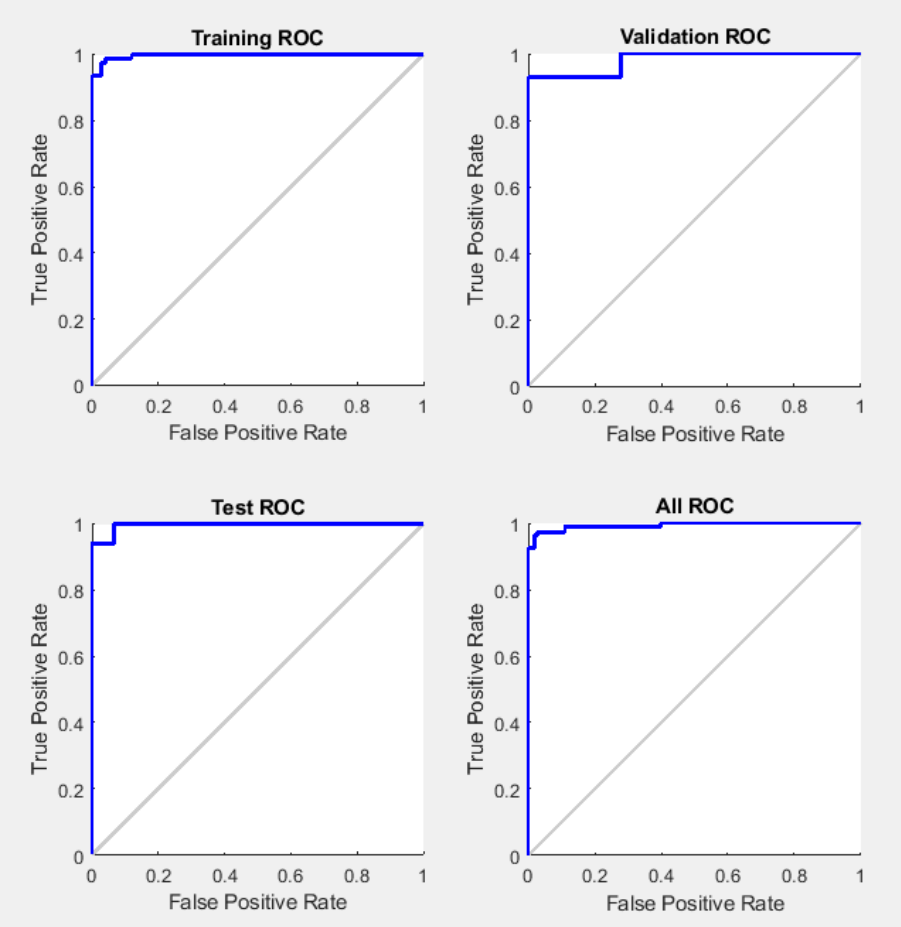


DT AUC After MI and Chi-Square (AQU Dataset)

Confusion Matrix for DT After MI and Chi-Square (AQU Dataset)



Confusion Matrix for ANN After MI and Chi-Square (AQU Dataset)



ANN AUCs After MI and Chi-Square (AQU Dataset)

الملخص

تعتبر الرعاية الصـــحية أحد المجالات التي تنتج كمية هائلة من البيانات ، وســتكون هذه البيانات الناتجة عديمة الفائدة إذا لم يتم استخراج الأنماط المفيدة واستخدامها بطريقة مناسبة. بشكل عام ، تشكل أنواع السرطان المختلفة حوالي 14٪ من الوفيات في فلسطين ، وبلغ معدل انتشار سرطان القولون والمستقيم 15٪ بين الرجال و 14.6٪ بين النساء من جميع أنواع السرطان.

من ناحية أخرى ، يعتمد حدوث ســرطان القولون والمســتقيم نسـبيًا على عوامل الخطر السـلوكية التي قد تزيد من حدوثه، إلى جانب العوامل الوقائية التي يمكن أن تقلل من حدوث ســـرطان القولون والمســـتقيم. لذلك ، تم إجراء هذا البحث من أجل تحديد عوامل الخطر السلوكية التي أثرت على حالات مرضى سرطان القولون والمستقيم المبلغ عنها في فلسطين ، وتم الاستفادة من أدوات التعلم الآلي (Machine Learning) التي يمكن اســتخدامها في عمليتي التصـــنيف والتنبؤبحدوث المرض، حيث من الممكن ان يتم اســتخدام التصــنيف والتنبؤ من قبل المواطنين بالمســتقبل لكي يقومو بتعديل عوامل الخطر الســلوكية المرتبطة بمرض سرطان القولون والمستقيم، اضافة الى تعديل وضبط عوامل الخطر السلوكية، اتاحة التنبؤ بالمرض للعامة سوف يساهم ايجابياً في تعزيز اهتمام المواطنين بصـــحتهم الشـــخصـــية. في هذا البحث ، تم جمع وتحليل مجموعتي بيانات مختلفتين من مجموعة بيانات سرطان القولون والمستقيم ، حيث تم الحصول على إحداهما من السجل الوطني للسرطان في وزارة الصحة الفلسطينية، ومجموعة البيانات الأخرى التي تم استيرادها من جامعة القدس (أبو ديس). ووجدت الدراسة أن أهم عوامل الخطر السلوكية التي يجب مراعاتها هي العمر، التاريخ الطبي السابق، النظام الغذائي، النشاط البدني والسمنة.

وعلى ما ســبق، تم تطبيق أدوات مختلفة للتعلم الآلي لتصــنيف عوامل الخطر والتنبؤ بها. في هذا البحث ، تم اســتخدام أدوات التعلم الآلي لدعم القرارات الطبية من خلال الجمع بين الأنظمة الحســـابية الذكية ومجموعات البيانات الطبية. تتمثل المهمة في تصنيف أو التعرف على نمط مختلف يستخدم في التشخيصات الطبية لتحديد عوامل الخطر الطبية لسرطان القولون والمستقيم، وتم تطبيق ومقارنة دقة خمس من أدوات التعلم الالي والذكاء الاصـــطناعي الاكثر شـــهرةً وشـــيوعاً في المجالات الصـــحية (Decision Tree, Support Victor Machine, K-Nearest Neighbour, Artificial Neural Network و Linear Regression) ، على مجموعتي البيانات اللتان تم جمعهما. ثم تم تقييم أداء الادوات الخمس المختلفة باســتخدام مقاييس مختلفة مثل الدقة الشــاملة ، والاســترجاع ، والخصــوصــية ، و المســاحة تحت المنحنى ، وغيرها. أظهرت النتائج أن (Artificial Neural Network) هي افضل اداة ممكن ان يتم استخامها. أخيرًا ، وجدت الدراسة أيضًا أن هناك حاجة ماسة لتعزيز العوامل الوقائية اجراء الفحص الدوري وزيادة وعي المواطنين لمسببات سرطان القولون والمستقيم.