Colorectal Cancer Risk Factor Assessment in Palestine Using Machine Learning Models

Mohammad A. Z. Abu zuhri[1], Mohammed Awad[2], Shahnaz Najjar[3], Nuha El Sharif[4], Issa Ghrouz[5]
[1]Department of Nursing, Arab American University, Palestine
Mohammad.zedan@aaup.edu
[2]Department of Computer Systems Engineering, Arab American University, Palestine
Mohammed.awad@aaup.edu
[3]Department of Health Informatics, Graduate Studies, Arab American University, Palestine
Shahenaz.najjar@aaup.edu
[4] Faculty of Public Health, AlQuds University, Palestine
nsharif@staff.alquds.edu
[5]Ministry of Health, Palestine
issa_ikg2006@hotmail.com

Abstract
The healthcare field produces a tremendous amount of data, and this produced data is useless if usage patterns are not extracted and managed properly. Generally, different types of cancers account for about 14% of mortality in Palestine, and Colorectal Cancer (CRC) specifically has a prevalence of 15% among men and 14.6% among women of all cancer types. Therefore, this research was carried out to assess 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 a public CRC classification and prediction tool based on accurate ML tools will help individuals in addressing their behavioral CRC risk factors and enhancing their engagement with their health. In this research, we have collected a local Palestinian dataset that consists of 57 predictors used to diagnose CRC. The dataset consists of 216 instances of CRC in both males and females. Statistical models such as Chi-Square and calculating the P_Value were used to determine the most important features. The study found that the most important risk factors to consider are age, past medical history, diet behaviors, physical activity, and obesity. Consequently, different Machine Learning (ML) models were applied to classify and predict CRC risk factors. The obtained results showed that the Artificial Neural Networks model (ANNs) outperformed all models, with 99.5% accuracy, 100% sensitivity, 99.9% specificity, and 99.9% AUC.

Keywords: Colorectal Cancer, Data Mining, Risk Factors, Machine Learning, Classification.

1. Introduction

Generally, cancer is defined as abnormal malignant or tumor cells that grow uncontrollably, Colorectal Cancer (CRC) is a term for cancer that initiates either in the colon or the rectum. (American Cancer Society, 2018). CRC is ranked as the third-most prevalent cancer as well as the fourth because of cancer-related mortality worldwide. However, most CRC cases are diagnosed and detected in Western countries and the incidence rate of CRC is increasing from year to year. (Mármol , et al., 2017). In Palestine, cancer is the second-most prominent cause of death, accounting for 14% of cancer-related deaths. CRC is the second ranked after breast cancer in Palestine (MoH, 2016). It compromises 10.3% of the total cancer cases. Risk factors for developing a CRC can be categorized into modifiable and non-modifiable risk factors. The modifiable factors are those that can be modified and controlled by individuals such as Obesity, Physical Activity (PA), Diet, Smoking, and Alcohol Consumption. (Rawla, et al., 2019). On the contrary, some factors that could protect individuals from suffering from CRC and lead to a decreased CRC incidence rate in the future, are called preventive factors such as CRC screening and the Fecal Occult Blood Test (FOBT). In Palestine, however, there are core factors which negatively affect the undertaking of CRC screening such as religious, beliefs, traditional practices and attitudes, and cultural factors, in addition to a lack of CRC screening knowledge (Qumseya, et al., 2014). Consequently, disease risk factors scoring and assessment tools may help individuals to measure and tune in to modifiable CRC risk factors; such tools will also motivate individuals to adopt healthier lifestyles. Also, simple risk scoring and assessment tools provide 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). Machine Learning (ML) is a term to define the application of pre-programmed machines that will act as humans do, trying to emulate human cognition. ML provides an efficient and accurate result by avoiding human-made errors; thus, ML is considered valuable advancements in technology in cancer studies (Patel, et al., 2020). ML plays a vital role in healthcare in general and in medicine specifically, where ML pre-programmed machines showed excellent data analysis and accurate pattern identification and recognition, both of which are very difficult for a human being to perform (Patel, et al., 2020). 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 effectively reflects on boosting CRC classification speed and improving the intended prediction efficiency (Rado, et al., 2019).

In this study, we applied various techniques of machine learning. A local dataset was collected as the primary data 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 controls. The variables covered within the study were 58 variables. In the preprocessing phase, chi-Square and a calculation of the *P*-value were used to determine the most important features, where 5-fold cross-validation has been used to split the datasets into two subsets training and testing. Both two-fold and four-fold cross-validation methods have been applied to validate the performance of the applied models. Confusion matrix, classification accuracy, recall, specificity, AUC were the main measurements used to evaluate the performance of the applied models in this work.

The paper is organized as follows. Section 2 will introduce a set of related work within the same research field. The dataset description will be shown in section 3. Section 4 will illustrate the algorithmic foundations of the applied approaches. Section 5 presents an explanation of the evaluation of the results and results will be presented and discussed in Section 6. Conclusions and future works will be discussed in Section 7.

2. Related Work

Eastern Mediterranean countries have lower rates of Physical Activity (PA), and physical inactivity causes about 10% of CRC. Further, eliminating physical inactivity was shown too 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 a low PA rate is a key cause for about 21%-25% for both Colorectal and Breast cancers, and 27% for diabetes. (WHO, 2010). In addition, there is a strong relationship between age and chances of CRC development, where chances of CRC occurrence become greater after the age of 50 years (Demb, et al., 2019) (American Cancer Society, 2020).

Johnson et al., (2013) carried out a meta-analysis study of CRC risk factors. The study concluded that a history of Inflammatory Bowel Diseases or Colorectal Cancer in an individual's first relative is considered two of the highest risk factors for developing colorectal cancer. 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 to right colon cancer. Their study also suggested that male smokers have a lower risk of rectal

cancer than female smokers do (Gram, et al., 2020). Besides, cigarette smoking and alcohol consumption are modestly associated with the development of CRC but do significantly increase colorectal cancer risks, serrated polyps, and adenomas. In other words, smoking and alcohol consumption are more 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 an increased risk of CRC in women and men (Fagunwaa, et al., 2017) (Lee, et al., 2019). A family history of CRC is considered an independent risk factor, and there is substantial variance in its contribution to the causation of CRC. 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 the 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). Also, Deng et. al. (2012) concluded that there is a strong direct correlation between diabetes mellitus (DM) and CRC incidence within men and women. Consequently, and within the same study, they also included 4 studies (one case-control and three 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). Poor awareness of early cancer symptoms among individuals can lead to late diagnosis and low survival rates. In addition to a lack of awareness, when cancer symptoms are atypical, have negative (false) beliefs seeking medical assistance will also be delayed. Furthermore, the educational level of individuals is an essential factor in recognizing their colorectal cancer risk factors. Individual's awareness can also help them to 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 lessening of CRC complications (Al-Azri, et al., 2019).

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 out their study to compare Sequential Minimal Optimization (SMO), K Nearest Neighbor (KNN), and Best First Decision Tree on breast cancer datasets to compare their accuracy and performance. SMO achieved the best performance and accuracy. Many AI methodologies can be used in detecting cancer occurrence as well as predicting the probabilities of cancer incidence, such as Artificial Neural Networks (ANNs), Particle Swarm Optimization (PSO), Genetic Algorithms, *K*-Nearest Neighbor algorithms (K-NNs), Support Vector Machines (SVMs), Linear Regression and Fuzzy Clustering (Patel, et al., 2020). Consequently, Patel et al. (2020) found that deep machine learning with the Gaussian model and watershed transform achieved the best accuracy result among different AI methodologies.

Further, Ting et al. (2019) proposed a CRC prediction schema, which suggests integrating Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Machine (SVM), Extreme Learning Machine (ELM), and XGBoost. The study concluded that the A-XGboost achieved the best AUC, sensitivity, and accuracy. (Ting, et al., 2019). Additionally, 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 patient survival prediction. SVMs were the best method in the five-fold accuracy test. In comparison, ANNs came after SVMs. Therefore, SVMs and ANNs are highly recommended to adopt in data mining methodologies for CRC datasets (Yan, et al., 2018).On the other hand, Asri et al. (2016) evaluated 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 found that the best machine learning tool to be applied in breast cancer prediction is SVM (Asri, et al., 2016).

In this research, different machine learning models were used to classify CRC. A Palestinian dataset has been collected to be used in this work to evaluate applied models in classifying CRC depending on risk factors. This MLPNN-LM model proves its efficacy in the classification of CRC.

3. Dataset

The study was carried out by analyzing a national Palestinian dataset. The dataset contains a group matched case-control study design using comprehensive interviews and records review. The obtained dataset contains 107 CRC cases and 109 control groups. The variables covered within the study were 58 variables.

3.1 Data Preprocessing

Data preprocessing and data cleaning were performed to prepare the dataset for further processing. All corrupted

records were omitted, the unwanted attributes were excluded, and some were estimated based on chi-Square. The feature selection method was based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data.

3.2 Feature Selection

An extensive literature review was carried out and an expert in the colorectal cancer field was consulted to analyze the profiling methodology using ML techniques. The irrelative attributes were excluded from the beginning as they cannot be considered the leading risk factors for CRC. After MI was performed on the dataset, the feature selection methods arranged the features as; did Colonoscopy, did Occult Blood test, Yearly Income, Frequency of eating Fruits, Type of Living Area, Frequency of eating Grilled Red Meat, Frequency of eating Vegetables, Smoking, Frequency of eating Red Meat, Occupation, PA Rate, Frequency of eating Grilled Chicken, Age, BMI, Frequency of eating Chicken, having Crohn's disease, and having other cancers.

3.3 Data Partitioning

After the obtained data were cleaned and processed, the analysis adopted a 70% training and 30% testing basis as shown in figure 1, where 70% of each dataset records were dedicated to training the ML models, while 30% of the records were dedicated to the tool's output. The carried out analysis among the applied ML tools adopted five-fold cross-validation. To get the result validated, at each stage, the data were divided into testing and training parts, and the training part was divided into four pieces. Also, in each stage, the testing data was selected from a different place and finally, each record was used for training and testing purposes. Finally, the average of the five stages was calculated as the final accuracy score for each tool.

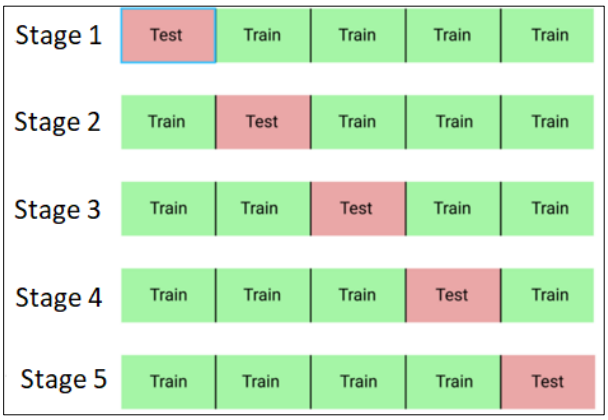

Figure 1: Five-fold Cross-Validation Process

4. Methodology

Five popular ML methods were applied to the gathered, selected, and preprocessed data, where first of all a feature selection process was carried out to select the attributes to be used, then data cleaning and data preprocessing took place by omitting the corrupted or duplicated records. Consequently, the processed and clean data were divided into training and testing sets to train and test the Decision tree (DT), Support vector machine (SVM), K nearest neighbor (KNN), Artificial Neural Networks (ANNs), and Logistic Regression (LR) and evaluate the best suites the Palestinian dataset, as shown in Figure 2.

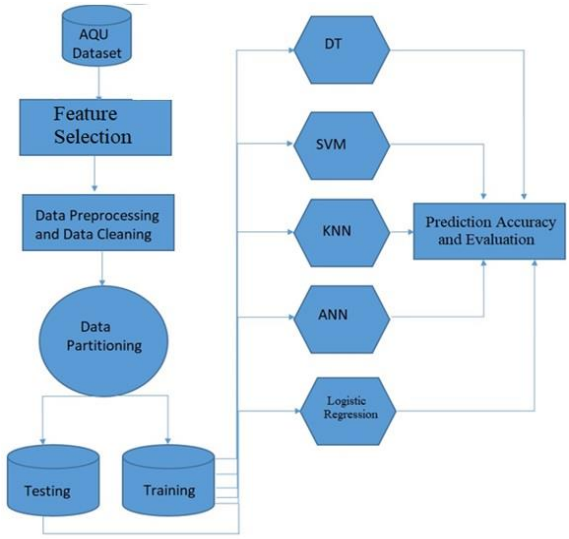

Figure 2: Data Mining Proposed Framework

5. Results and Evaluation

After applying the five ML tools to the dataset, the performance of each tool was calculated based on accuracy, sensitivity, specificity, and Area Under the Curve (AUC).

Accuracy: The accuracy of a ML tool is indicated by the percentage of correct predictions. Accuracy is calculated using the formula

$$Accuracy + \frac{TP + TN}{TP + FP + TN + FN}$$

Where TP is the true positive, i.e. the number of positive records that are predicted truly. Similarly, the TN is the true negative, i.e. the number of records predicted that lie in the negative class and are predicted truly. The FP is the number of false positives and indicates the number of records in the positive class but predicted incorrectly, and FN is a false negative, which is the number of records that belong to the negative class but the applied tool incorrectly predicted the result.

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

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

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

$$\text{Specificity} = \frac{TN}{TN+FP}$$

The area under the curve (AUC): is the area under the receiver Operating Characteristic (ROC) curve. The ROC curve is a powerful test that evaluates a diagnostic test. In the ROC curve, the true positive rate (TPR) is calculated in function with the false-positive rate (FPR) for many predefined parameter thresholds. AUC under the ROC measures the performance and the ability of specific parameters 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).

6. Study Results

6.1 Descriptive Statistics and Visualization

After CRC patient records were acquired, descriptive analysis and visualization were performed to describe the data and to prepare it for analysis later using MATLAB software. Since age is one of the scientifically proven risk factors for CRC, it is important to study and analyze the age of CRC cases. The majority of CRC cases in the dataset range from 40 years to 79 years of age, with 90% of the cases within this age range as shown in Figure 3.When CRC cases' gender was analyzed, we found that 45.37% of the CRC cases in the dataset are females, while 54.63% are males. On the other hand, when the study considered the age of 50 as a cut-off 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 1. Thus, this might be an issue for further analysis in future researches.

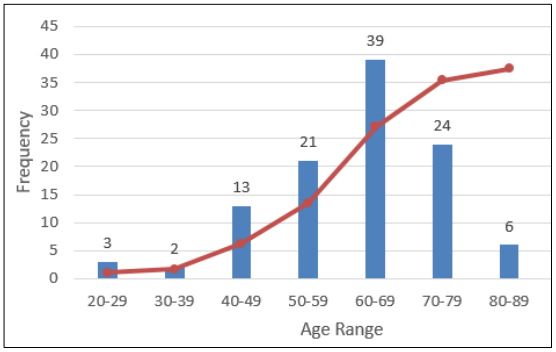

Figure 3: CRC Cases Distribution by Age

Table 1: Critical Age Based on Gender

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 eat fruits more than four times a week is greater than those who eat vegetables four times a week (data not shown).

Johnson (2013) supported a moderate association between abnormal BMI and CRC incidence. However, Ghrouz & El Sharif (2019) found that BMI and CRC incidence is statistically insignificant.

The WHO, as well as many other researchers, as shown in the literature review chapter, supported that, the Physical Activity (PA) rate of an individual is strongly related to CRC prevalence, where the relation between PA and CRC prevalence is an inverse relationship. Therefore, the study analyzed the level of PA among CRC patients. Ghrouz & El Sharif (2019) found that PA rate has statistical significance with CRC incidence.

Ghrouz & El Sharif (2019) found that the relationship between smoking and CRC occurrence is statistically significant. 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 CRC.

From another side, the data analysis for family history and CRC showed statistical significance for the positive relationship between an individual's family history and the chances of developing CRC.

6.2 Machine Learning Results

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

- **Decision Tree Results**

A decision Tree was applied to the dataset and, the total accuracy was 94%, where the Area Under Curve (AUC) was 0.97. The confusion matrix that was produced after applying DT revealed that 94% were true positive, while false negative scored 7%. 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 0.94included from the confusion matrix produced after applying DT to the dataset.

- **Support Vector Machine Results**

Support Vector Machine (SVM) scored a total accuracy of 94.9% when applied to the dataset, where the AUC was 0.98. 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%. Thus, the sensitivity for SVM is 0.93 and the specificity is 0.97.

- **K-Nearest Neighbor (KNN) Results**

The KNN tool had 94% accuracy, where the Area Under Curve was 0.94. In addition, the true positive was 97%, and the false-positive 3%, while the true negative was 91%, and the false-negative was 9%. This leads us to find that the sensitivity for applying KNN on the dataset is 0.86, while the specificity of using KNN is 0.97.

- **Logistic Regression Results**

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 an AUC of 0.91.

In addition, applying the confusion matrix for the Logistic Regression revealed a true positive of 90% and a false positive of 10%, while the true negative percent was 88%, and the false-negative was 12%. This leads us to conclude that the sensitivity for Logistic Regression is 0.88 and the specificity for using it is 0.90.

- **Artificial Neural Networks (ANN) Results**

To adhere to 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 the 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 was not efficient to use more resources. This excellent accuracy score came out because the dataset is balanced (a case-control), where the number of cases in the case group (CRC cases) and the control group (participants who do not have CRC) are equal. The confusion matrix for ANN with 5 neurons on the dataset is shown in Figure 1. Consequently, we conclude that the sensitivity score is 100% and the specificity is 99.9% after adopting ANN with 5 neurons and applying it to the intended dataset.

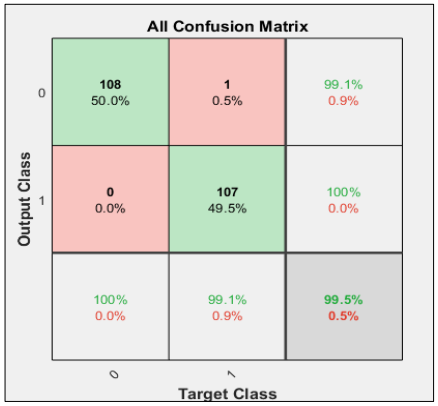

Figure 4: ANN Confusion Matrix Results

6.3. Results Summary

After comparing DT, KNN, ANNs, SVM, and Linear Regression, we found that the best tool to be applied to the dataset is the ANN. ANN achieved the best accuracy score among the other four tools. ANNs also scored the best sensitivity, specificity, and Area Under Curve as shown in Figure 5. They experimented showed that the second-best ranked tool is SVM with a relatively small variation 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 to build models for binary (0/1) dependent variables.

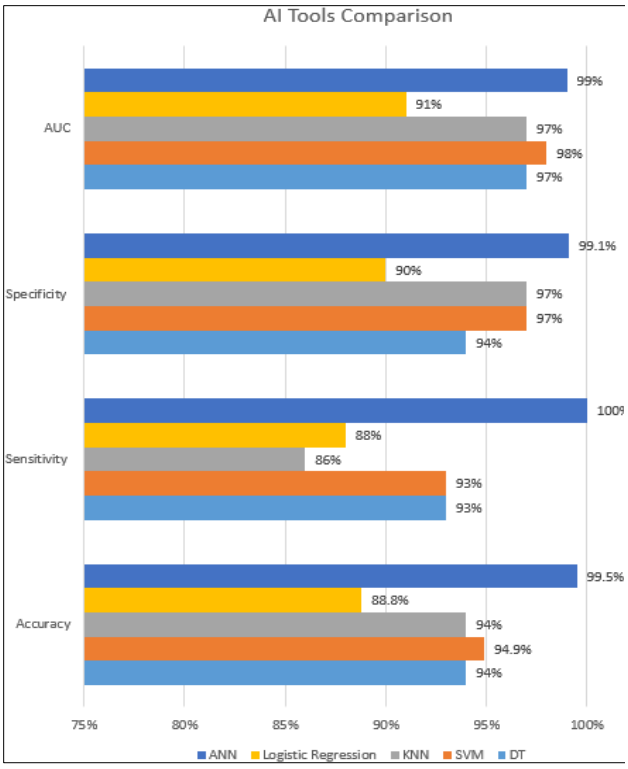

Figure 6: AI Tools Results Comparison

7. 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 CRC incidence in Palestine. Adopting healthier lifestyles 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 CRC prediction tool, where DT and ANNs gained the highest accuracy rates. In this study, and 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 CRC occurrence. The following recommendation was formed based on these conclusions: Develop national health policy promotion programs to improve Palestinians lifestyles, such as promoting physical activity, which will positively affect obesity. For instance, a national policy to increase 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 for healthier ones such as quitting smoking. It is also recommended to carry out a comprehensive study that aims to assess the impact of different medical history issues as well as diet on CRC incidence in Palestine.

References

- 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].

- American Cancer Society, 2020. *Cancer.org.* [Online] Available at: https://www.cancer.org/content/dam/CRC/PDF/Public/8605.00.pdf [Accessed 13 January 2021].

- 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.

- 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).

- 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.*

- Demb, J. et al., 2019. Risk factors for colorectal cancer significantly vary by anatomic site. *BMJ Open Gastroenterology* .

- 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.

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

- 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.

- 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.

- 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).

- 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.
- 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.
- 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.
- 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.