Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v18n2p359

Machine Learning-Based Analysis of Cancer Incidence in Jordan (2020–2021): A Decision Tree Approach
By Alsheraideh et al.,

15 October 2025

Electronic Journal of Applied Statistical Analysis
Vol. 18, Issue 2, October 2025, 359-378
DOI: 10.1285/i20705948v18n2p359

Machine Learning-Based Analysis of Cancer Incidence in Jordan (2020–2021): A Decision Tree Approach

Raneem Alsheraideh[a], Mohammed Sh. Alsheraideh[b], Mariam A. Al-Nasser[c], Amjad D. Al-Nasser *[d], Hanan I. Malkawi[d], Abdallah Ma'touq[f], Odai Mohammed[g], Ahamad Eyad[h], Fadia Mayyas[i], Rana Abu-Huwaij[j], and Mohamad Kharashgah[c]

[a] *Jordan University, Faculty of Medicine, Jordan*
[b] *Amman Arab University, Faculty of Applied Medical Sciences, Jordan*
[c] *Yarmouk University, Faculty of Medicine, Jordan*
[d] *Yarmouk University, Faculty of Science, Dept. Statistics, Jordan*
[e] *Yarmouk University, Faculty of Science, Dept. Biological Sciences, Jordan*
[f] *Ministry of Health, Communicable Diseases Dept, Jordan*
[g] *Royal Medical Services, General Surgery Dept, Jordan*
[h] *Al Bashir Hospital, Orthopedic surgery Dept, Jordan*
[i] *Yarmouk University, College of Pharmacy, Jordan*
[j] *Amman Arab University,College of Pharmacy, Jordan*

15 October 2025

This study aimed to classify the cancer types across different regions in Jordan using a machine learning-algorithm and based on the Decision Tree Analysis (DTA) technique. The model employed patients' demographic information—specifically gender, age, and region of residence—as independent variables (IVs) to assess their interaction with cancer types as the dependent variable (DV). The objective was to determine the predictive relationship between these demographic factors and cancer types to support regional cancer profiling and inform targeted public health planning. The research utilized secondary data from the Ministry of Health, the Directorate of Non-Communicable Diseases, and the Jordan Cancer Registry for the years 2020 to 2021. A total of 9,547 cancer cases were analyzed using the DTA model, which effectively identified significant incidence patterns, with the central

*Corresponding authors: amjadn@yu.edu.jo

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

region of Jordan accounting for the highest number of cases (n = 6,815; 71.4%). The model classified cancer into 25 distinct types based on demographic attributes, with breast cancer being the most prevalent, particularly among middle-aged females residing in the central region. The DTA model demonstrated high efficacy in handling and stratifying large-scale medical data, predicting cancer type interactions, categorizing and labeling datasets, and suggesting potential category mergers. These findings have important implications for the development of focused cancer prevention strategies and the efficient allocation of healthcare resources. However, a key limitation of the study is the incomplete characterization of cancer patient attributes across all Jordanian regions.

keywords: Cancer incidence, Machine learning, Risk factors, Predictive modeling, Jordan.

1 Introduction

Cancer is the leading cause of death in Jordan, impacting both men and women across all age groups. As individuals grow older, the risk of developing cancer rises significantly, making it a major public health concern of Health (2021). The increasing incidence of cancer highlights the urgent need for enhanced awareness, early detection, and improved treatment options. Addressing this issue requires a comprehensive approach that includes preventive measures, advancements in medical research, and accessible healthcare services to reduce mortality rates and improve patient outcomes Armonk (2020).

The Hashemite Kingdom of Jordan is a country in the Southern Levant region of West Asia, that is bordered by Syria to the north, Iraq to the east, Saudi Arabia to the south, and the occupied Palestinian West Bank to the west Review (2019). In 2024, the population of Jordan was estimated at 11,654,813. The average number of patients per birth was approximately ±2.8%. According to the 2021 Annual Cancer Registry, the highest incidence rate of cancer cases per 100,000 population was recorded in the central governorates, at 119.6.; Amman (166.8), Balqa (50.8), Madaba (40.9), then Zarqa (32.1). On the other hand, the incidence rate in the northern governorates was 16.7; Ajloun (19.8), Jerash (18.6), Irbid (16.9), then Mafraq (13.2) Department of Statistics - Jordan (2024). This result indicates how much the governorate distribution of cancer types affects the extent of variation incidence rates. The most common types in the various governorates of the Kingdom are affected by several factors as well. Although the specific causes of many types of the cancer are still unknown, about two-thirds of cancer cases diagnosed today are preventable. Therefore, the Jordanian Ministry of Health (MOH) must take specific action to reduce the future burden of cancer. Some of the preventive measures must include eliminating exposure to known risk factors that are related to our environment and lifestyle. This requires a deep search into these factors and finding their classification cancer types reasons and regression analyses and the geographical distribution specifically among males and females in Jordan. Many published research

papers and articles tried to find out the correlations between types of cancer patients that are associated with multivariate including the residential distributions. The latest studies in 2018 shared by the Jordanian Ministry of Health, the Directorate of Non-Diseases, and the Jordan Cancer Registry reported 7094 new cancer cases with a 76.7% Jordanian incidence rate compared to 9248 in total Sung et al. (2021). Global Cancer Statistics in 2020 of cancer incidence and mortality produced by the International Agency for Research on Cancer worldwide was estimated that 19.3 million new cancer cases (18.1 million excluding nonmelanoma skin cancer) and almost 10.0 million cancer deaths (9.9 million excluding nonmelanoma skin cancer) occurred in 2020. Breast-female cancer has overcome lung cancer as the most diagnosed cancer, with an estimated 2.3 million (11.7%) new cases, followed by (11.4%) lung, colorectal (10.0%), prostate (7.3%), and stomach (5.6%) cancers. Lung cancer remained the major cause of cancer death, with an estimated 1.8 million deaths (18%), followed by liver (8.3%), colorectal (9.4%), stomach (7.7%), and female breast (6.9%) cancers. For both sexes, the overall incidence was higher from 2 to 3-fold in transitioned versus transitioning countries, whereas mortality varied less than 2-fold for men and little for women. However, the statistics mentioned that the death rates for cervical cancers (12.4 vs 5.2 per 100,000) and female breast (15.0 vs 12.8 per 100,000), were almost higher in transitioning versus transitioned countries. In 2040 the global cancer burden is expected to be 28.4 million cases, a 47% rise from 2020, with a larger increase in transitioning (64% to 95%) versus transitioned (32% to 56%) countries due to different factors like demographic change. However, this might be further exacerbated by increasing risk factors associated with the growing economy and globalizationBray et al. (2018). however, incidence patterns of cancer and burden in Jordan have never been explored thoroughly, so the study aimed to close this knowledge gap of existing data on cancer incidence, outcomes, and independent variables (IV)World Health Organization (2020). The recent of this article is organized as follows: next section circulate the literature, while section 3, discusses the research methodology, in section 4, the data analysis is introduced, the article ends in section 5 which introduces some conclusions and future research ideas.

2 Literature Review

The review study of cancer care in resource-limited countries as an example in Jordan 2024 mentioned that cancer is ranking second in the list of causes of death cancer worldwide, and become a growing health care problem, after cardiovascular diseases. Jordan's national cancer registry still suffers from the problems mostly due to long lag time in reporting, as well as the absence of outcome data, and accurate staging. The number of new patients with cancer diagnosed is increasing at an unexpected rate, fueled by population growth, changing the structure of the population, refugees with the older population, high rate of obesity, smoking, improving life expectancy, and lack of adequate exercise. Also, the study indicated that the age-standardized rate for cancer incidence is significantly lower than the Western societies, while the mortality rate is higher Ministry of Health (2021). Despite the higher efforts done by MOH in Jordan,

all diagnosed cancers are still at more advanced stages and at younger ages. However, the breast cancer program opportunistic screening represents a great example that led to significant downstaging of breast cancer, alongside the feasibility of screening programs that were evaluated for colorectal and lung cancers are underway Ministry of Health (2019). This study of the largest execution of clinical cancer genetics programs not only in Jordan but also in the region, helps to identify patients who are at-risk relatives for hereditary cancers. In addition to a cancer control program that addresses all issues of cancer care starting from screening to early detection, through active cost-effective treatment to assure wider access to palliative care, hospice, and survivorship Abdel-Razeq et al. (2024). Moreover, Ferlay et al. (2010) showed a major public health problem, which is the second cause of death in less developed countries compared with other developed worlds. An estimated 63% of the deaths of 7.6 million cancer deaths occurred in 2008 with occurring in less developed regions, the researcher evaluated many factors like lifestyle, infectious agents, environmental exposures, methods of registry, and differences in the site and the constitution might be considered as a limitation of risk factors which have been attributed to certain types of cancer Ferlay et al. (2010). Other modeling and predicting the spatial dispersion of skin cancer conducted by Masoumi et al. (2018), consider and clarify factors that resulted from their research like; the environmental and sociology-economic, temperature, pressure, interactive effect of bright sunshine hours and solar energy, contact with lead, fruit and vegetable consumption, phosphate fertilizer usage and tobacco consumption might be related or/and affected direct or indirect by geographical factors which can be associated with specific types of cancers .For instance, the incidence study in 2020 of thyroid cancer showed that was on the rise and affected directly or indirectly by geographical factors and these results provided us with additional information relevant to evaluating the tracks of risk, safety, related application of geographical and their associations' factors Kim et al. (2020).

On the other hand, Central and Eastern Europe have the second-highest incidence rate of gastric cancer in the world. On the contrary, the incidence is significantly lower compared with Northern and Western Europe, like that for example in the USA, the reasons for such differences include genetic susceptibility, strains of H.pylori, hygiene, food preparation, and food preservation Xie et al. (2021).

Also, several epigenetic aging, approved that the incidence of cancer is directly related to age, and advanced chronologic age is considered one of the most significant risk factors, besides organismal aging that is associated with changes affected by both genetic and environmental factors at the tissue, cellular, and molecular levels. The specific mechanisms through which these age-associated molecular changes contribute to the increased risk of aging-related diseases, such as cancer, also changes over a lifetime may be part of an "epigenetic aging" process Yu et al. (2020). Also,Bhatia et al. (2022), proves evidence of rural-urban disparities, Rural residence was associated with a higher rate of stage IV colorectal cancer at presentation, one of the causes is health behaviors, the description of the rural-urban differences in the use of tobacco products, diet, physical activity, and obesity Bhatia et al. (2022). In 2023, a research study reported that the central governorates in the occupied West Bank (WB) had the lowest mortality rate for most types of cancer among men and women, whereas lung and prostate cancers were

higher in northern (WB) while showed the lowest percentage in the Jerusalem Governorate. The study also mentioned the reasons behind the higher rates of cancer among males and females among the occupied Palestinian Citizens (OPC), which might be due to factors like unhealthy lifestyles, consumption of processed foods, limiting a healthy Mediterranean diet region, excessive smoking, and lack of physical activity, in addition to the environmental health impacts resulting from air pollution caused by Israeli waste dumped in the occupied WB, which might become from several wrong industrial activities Salem (2023). Moreover McMahon et al. (2023), showed that statistically there were no significant differences regarding race, facility location, or urban/rural residence status and proportions of tumor staging at diagnosis. However, it still needs further studies to recognize the significant effects, in terms of socioeconomic and geographic factors that have an impact on the staging of bone cancer at diagnosis McMahon et al. (2023). Study research on Melanoma and Nonmelanoma skin incidence done in Jordan by 2023, during the 16-year study period, found no significant difference between the various regions in Jordan (North, Middle, and South) in the prevalence of each cancer sub-type Almaani et al. (2023).

This study aimed to classify the cancer type within Jordan regions, according to the identification of patient's general independent variables (IV) information like gender, age, and residency region as they have the strongest interaction with the cancer dependent variable (DV) types, and their outcomes at the level of human health and impact on the governmental treatment and medication of financial resources.

3 Research Methodology

The Decision Tree Analysis (DTA) method closely resembles regression trees; however, instead of predicting a continuous numerical value, it is primarily used for categorizing data into discrete classes. DTA is widely utilized in machine learning and data analysis for classification tasks, helping to structure decision-making processes in an interpretable way James et al. (2013). This section will discuss the idea of the DTA.

3.1 Decision Tree Analysis

At the core of a decision tree, the root node represents the most significant predictor or classifier variable. This variable is chosen based on its ability to maximize information gain, effectively splitting the data set into distinct groups with the highest level of separation. The subsequent internal nodes further partition the population based on additional variables, arranged hierarchically in descending order of importance according to their contribution to information gain. This structured approach ensures that each decision step refines the classification process, leading to more accurate predictions. As an example of a decision tree to predict the cancer type based on patient demographics such as age, gender and cancer type could be obtained as given in Figure 1.

DTA is particularly valuable in applications such as medical diagnosis, risk assessment, and customer segmentation, where clear and logical decision-making is crucial. By identifying the most influential factors in a datasets, decision trees offer a transparent and

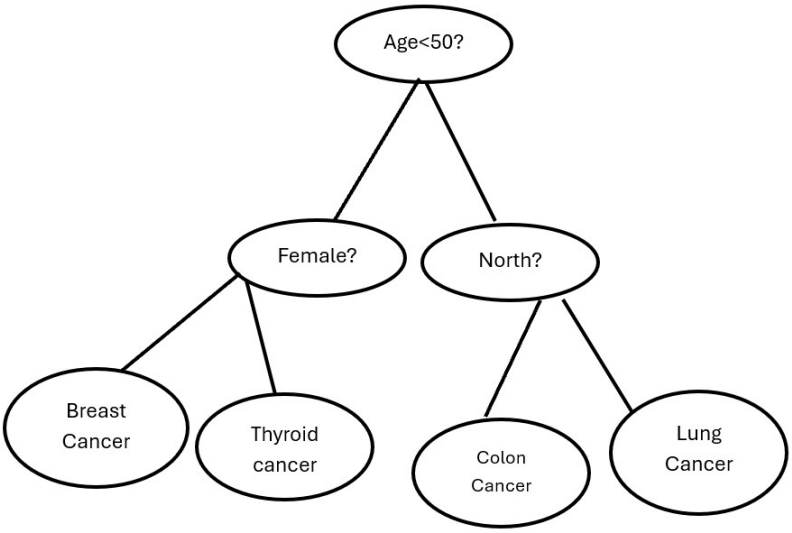

Figure 1: Decision tree example

efficient method for making informed predictions Almaani et al. (2023). Also, the DTA method is considered one of the most interesting statistical methods that can be applied in the field of statistical learning to both classification problems and regression analyses, according to Teli and Kanikar James et al. (2013). Several DTA machine learning models could be employed for data classifications, the choice of any of these models mainly depends on the data types. In this study, the Chi-Square Automatic Interaction Detection (CHAID) will be implemented, because it focuses, generally on assessing the impact of several (IV) like residency of population distribution, age, gender, and (DV) cancer and their major outcomes Tayefi et al. (2017). The CHAID algorithm is a supervised learning method, specifically for the detection of association between the categorical DV and multiple IVs which can be categorical and/or metric; then used the Chi-square test to select the split with the most significant association Teli and Kanikar (2015). The CHAID method requires large amounts of databases that focus on to satisfy the chi-square test assumptions within the split process, because they are at every tree level split into several groups which may become too small for reliable analysis.24 The CHAID is a useful method used to predict the cancer type within the Jordan region model, which also, enables the identification of general independent variables (IV) information such as gender, age, and residency region, that have the strongest interaction with the dependent variable (DV) like cancer types. So, the method setting was used to achieve the study purpose and to answer research questions such as What are the most common types of cancer affecting people in different regions of Jordan, based on general patient factors such as age and gender?

3.2 The SAMPLE

Samples collected depend on the structure of the health system in Jordan, which consists of private, teaching, military, and public sectors that include a total of 118 hospitals . The study was conducted based on the Ministry of Health records on cancer patients. Which provided healthcare services to 27,000 cancer patients in 2022 and received monthly about 800 new cancer patients of Jordanian and non-Jordanian nationality Kass (1980). In general, the data was collected from all governmental, private, Royal Medical Services, and King Hussein Center.

3.3 DATA COLLECTION

Jordan cancer reports were obtained by collecting sealed boxes sent annually from hospitals that deal with treating cancer cases into the Non-communicable Diseases for registration in the Cancer Registry. All printed cancer reports were automated and then computerized case by case separately by experienced physicians with a little statistical background in the program. All acquired data were systematically recorded in a digital Excel spreadsheet for further advanced statistical analysis. A dual approach was adopted to collect comprehensive data, including primary patient demographics data such as age, sex, national ID number, and other important variables such as smoking, status, governorate, and residency. In general, the data was collected from all governmental, private, Royal Medical Services, and King Hussein Center. In this study, a total of 9547 cancer patients were specifically collected and coded, by the highest quality professional doctors, and the highly method using the techniques applied for treatment. In addition to that, some patients come from the Middle East and North Africa region, and other patients from all over the world. Consent Forms and Ethical Approval Documents were obtained from the Research Ethics Committee from the Ministry of Health, Department of Non-Communicable Diseases, which were collected from the National Cancer Registry, noting that all subjects that must be approved and coded with new serial numbers to use it for each one patient, then placed in an electronic file with a secret number using the researchers' computer only, to ensure that the collected data were complete and accurate.

3.4 DATA ANALYSIS

In this study, three independent variables (IV) are included, namely, gender, age, and residency region, and one dependent variable (DV) is the cancer patients' type. The Statistical Package for Social Sciences IBM SPSS version 25 is used for all types of analyses, including the decision tree analysis (DTA) (Nisbet et al. (2009); Rokach and Maimon (2008); King Hussein Cancer Foundation & Center (2022)). This machine learning model is more precise as it's a non-parametric supervised learning algorithm. The main mechanism of DTA is classifying cases into several homogeneous groups in some categories such as male or female. The final presentation will be a hierarchical tree structure that consists of tree nodes, branches, internal nodes, and leaf nodes. Also, it can be used to predict the values of a response variable based on the values of predictor

variables. Moreover, the DTA can be used for different purposes such as segmentation, stratification, prediction, data reduction, interaction identification category merging, and labels.28-30

Table 1 shows the characteristics of cancer patients IV and DV. The sample consists of 55.1% (n = 5260) females and 44.9% (n = 4286) males. In terms of age, 26.5% (n = 2530) are 40 years or younger, 41.2% (n = 3931) are between 41 and 60 years, and 32.3% (n = 3085) are older than 60 years. Regarding residency, the majority (71.4%, n = 6815) reside in the Middle region, followed by 14.0% (n = 1339) who are non-residents, 9.3% (n = 885) in the North, and 5.3% (n = 507) in the South. It could be noted that there are 25 different cancer types combined with the DV, in which Breast cancer 24.5% (n = 2336) is the most cancer distributed among all other types, followed by Bone cancer 7.8% (n = 747), then Brain cancer 7.3% (n = 693).

Using the CHAID method and applying the Bonferroni correction for type I error, the results in 9 different significant parent nodes for three levels of analysis of the results of cancer type classification in Jordan, based on CHAID methodology are given in Table 2.

Overall, the analysis highlights the substantial influence of residency region, gender, and age on the distribution of cancer types, with Breast Cancer being the most prevalent across different demographics. The most common predicted cancer type across the sample is Breast Cancer, accounting for a majority of the population in several nodes. Residency region is the primary independent variable at the parent node, splitting the sample into: Middle region: 71.4% (Node 1), Non-residents: 14.0% (Node 2), North and South regions: 14.6% (Node 3). Also, Gender is a significant IV within most subgroups. Females predominantly exhibit Breast Cancer, e.g., 40.0% in Node 4 and 7.5% in Node 8. Males are associated with other types, including Lung Cancer (31.4% in Node 5 and 7.1% in Node 9) and Brain Cancer (6.4% in Node 7). However, Age is a critical variable within various nodes: Individuals aged 40 years or less are linked to Breast Cancer (e.g., Node 10: 9.8%) and Bone Cancer (e.g., Node 13: 6.8%). Those aged 41–60 years frequently exhibit Breast Cancer (e.g., Node 11: 18.4%) or Lung Cancer (e.g., Node 14: 11.1%). Individuals over 60 years are associated with a mix of cancer types, such as Breast Cancer (e.g., Node 12: 11.8%), Lung Cancer, and Prostate Cancer. Notable Subtypes and Frequencies: Rare cancer types include Bone Cancer (e.g., Node 19: 2.7%) and Prostate Cancer (e.g., Node 15: 13.5%). Smaller subgroups are observed, especially in lower frequency nodes driven by age and gender interactions.

4 Results

This study consists of 9,546 cancer patients 55.1% (n = 5,260) Females, most of them within the age of 41 years and 60 years 26.5% (n = 3,931). The sample was divided with respect to the patient's region of residency (north, middle, South, and non-residency). Most of the cancer patients are residents in the middle region of Jordan 71.4% (n = 6,815). Moreover, the full decision tree diagram suggests the data could be classified into three levels based on the IVs: - Level One: the most significant IV is the patient's

Table 1: Structure of variables used in CHAID analysis.

Variable	Category	Structure		Type of variable	
		Freq	Per	MS	IV/DV
Gender	Female	5260	55.1	NS	IV
	Male	4286	44.9		
Age	40 yr or less	2530	26.5	OS	IV
	41 yr - 60 yr	3931	41.2		
	More than 60 yr	3085	32.3		
Residency Region	Non-Resident	1339	14.0	NS	IV
	North	885	9.3		
	Middle	6815	71.4		
	South	507	5.3		
Cancer Type	Abdomen	33	.3	NS	DV
	Bladder	399	4.2		
	Bone	747	7.8		
	Brain	693	7.3		
	Breast	2336	24.5		
	Cervix	518	5.4		
	Colon	526	5.5		
	Esophagus	31	.3		
	Eye	80	.8		
	GIT	509	5.3		
	Gland	441	4.6		
	Kidney	279	2.9		
	Liver	78	.8		
	Lung	628	6.6		
	Lymph	504	5.3		
	Mouth	123	1.3		
	Nose	126	1.3		
	Pancreas	174	1.8		
	Prostate	349	3.7		
	Reproductive	257	2.7		
	Respiratory	113	1.2		
	Skin	169	1.8		
	Soft tissue	136	1.4		
	Stomach	217	2.3		
	Others	80	.8		
	Abdomen	33	.3		

MS: Measurement Scale; NS: Nominal Scale; OS: Ordinal Scale; IV: Independent Variable; DV: Dependent Variable; Freq: Frequency; Per: Percent

Table 2: Cancer Type Classification in Jordan: Decision Tree Table Results

	Total Sample		Predicted	Parent	Primary Independent Variable				
Node	N	Per	Category	Node	Variable	*Sig*[a]	Chi-Square	df	Split Values
0	9546	100.0	Breast						
1	6815	71.4	Breast	0	Residency	.000	247.527	48	Middel
2	1339	14.0	Breast						Non-Resident
3	1392	14.6	Breast		Region				North; South
4	3817	40.0	Breast	1	Gender	.000	2776.963	24	Female
5	2998	31.4	Lung						Male
6	728	7.6	Breast	2	Gender	.000	458.557	24	Female
7	611	6.4	Brain						Male
8	715	7.5	Breast	3	Gender	.000	564.237	24	Female
9	677	7.1	Lung						Male
10	931	9.8	Breast	4	Age	.000	654.780	46	40 years or less
11	1757	18.4	Breast						41 yr - 60 yr
12	1129	11.8	Breast						More than 60 yr
13	646	6.8	Bone	5	Age	.000	934.445	46	40 years or less
14	1062	11.1	Lung						41 yr - 60 yr
15	1290	13.5	Prostate						More than 60 yr
16	285	3.0	Breast	6	Age	.000	188.350	46	40 years or less
17	284	3.0	Breast						41 yr - 60 yr
18	159	1.7	Breast						More than 60 yr
19	257	2.7	Bone	7	Age	.000	270.833	46	40 years or less
20	195	2.0	Brain						41 yr - 60 yr
21	159	1.7	Prostate						More than 60 yr
22	222	2.3	Breast	8	Age	.000	89.913	23	40 years or less
23	493	5.2	Breast						41 yr - 60 yr; More than 60 yr
24	189	2.0	Bone	9	Age	.000	213.356	46	40 years or less
25	291	3.0	Lung						41 yr - 60 yr
26	197	2.1	Lung						More than 60 yr

Growing Method: CHAID. Dependent Variable: Cancer Type. Per: Percent. Sig: Significant.
df: Degrees of Freedom
a. Bonferroni adjusted

gender, which means that it is most strongly associated with DV and has the most power in the division of observations into groups. In other words, the first tree level is the IV (Gender), for the observed data this variable has the biggest potential to differentiate and classify patients into 25 groups according to the Cancer types. Figure 2 shows the distribution of the cancer types within the male and female in Jordan.

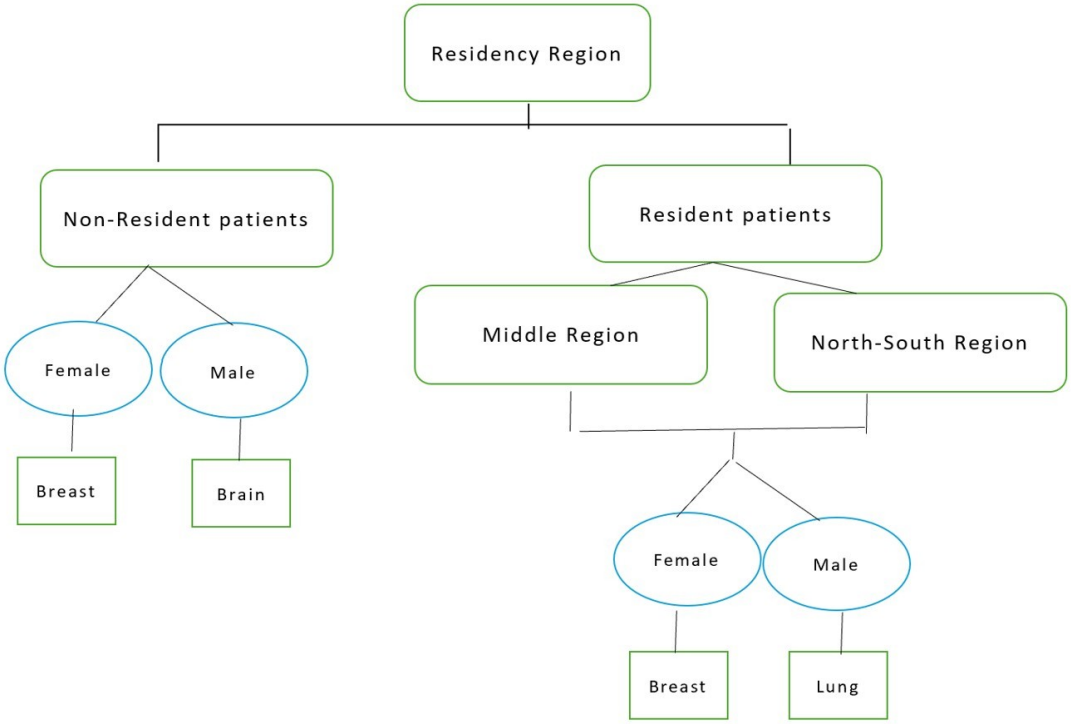

Figure 2: Cancer Type Classification is based on the Residency Region and Patient Gender

Moreover, the statistical significance of gender was found at a significance level of $\alpha = 0.05$ ($\chi^2(24)$ =3761.442, P value=0.000). Figure 3 shows the Cancer Type and classification using the patient's gender. As the first discriminator, it splits the root node, i.e. the sample of 9546 patients, into two groups having different categories of gender presented as node 1 (55.1% female patient (n = 5260)) and node 2 (44.9% male patient (n = 4286)).

The results as given in Figure 3 show the rank of the cancer type based on the patient gender frequency value:

Female : Breast → Cervix → Bone → Brain → Gland ...
Male : Lung → Bone → Bladder → Prostate → Brain ...

When connecting gender with the residency region, the results as given in Table 2 give a new classification with parent nodes 1/2/3 for the three regions with interactions with the patient gender indicating that female patients predict breast cancer as the most

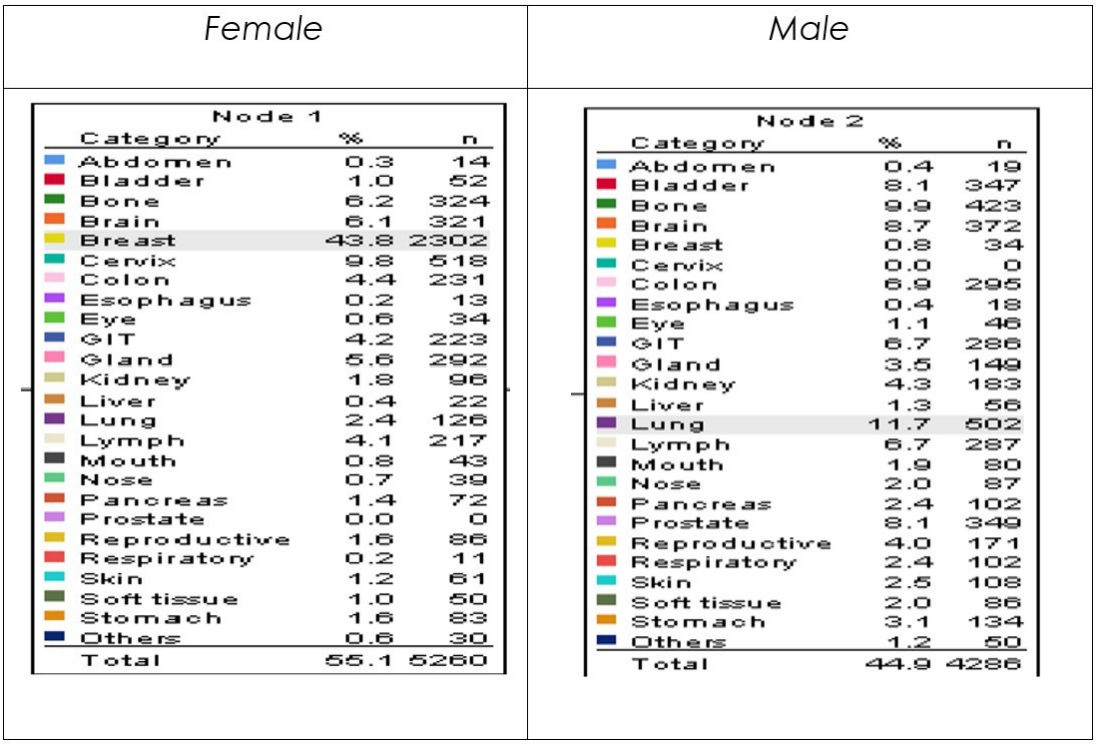

Figure 3: Cancer Type classification using patient's gender

concerned type of cancer with percentages 40%, 7.6% and 7.5% for Middle, Non-Resident and North-South regions, respectively. However, the male patients predict Lung cancer for the Middle and North-South groups, while Brain cancer for the non-resident group. Figure 3 illustrates the predictions of the first level of DTA considering the intersection between the residency region and the patient's gender.

Level two: Within the second level of the tree, the interaction between gender and patient age is statistically significant with the two levels of gender; Female and Male, which splits the Female into three nodes of the age categories 3/4/5 ($\chi_2(46)$ =912.658, P value=.000) and the same for the Male level with three Nodes 6/7/8 ($\chi_2(46)$ =1364.519, P value=.000). Table 3 gives the interaction between the patient's gender and the patient's age for classifying the cancer type. Results show that breast cancer is the most frequent type of cancer among females regardless of age category. On the other hand, in males, the most frequent type of cancer was bone cancer among males less than 40 years old, lung cancer among those who are 41-60 years old, and prostate cancer among patients who are older than 60 years old. Level three: This level is represented by the leaf nods in the last decision tree graph as given in Table.4. The results of the analysis connect the first two IV's (Gender and Age) with the Patient residency. The DTA in this level as given in Table 4, showed the following classification results. The data analysis

accuracy is measured by a percentage of the inaccuracy of the so-called prediction risk value. The estimated risk for this data is 0.684 (SE=0.005).

Also, based on the classification as given in Table 5, the overall accuracy of the DTA model is 31.6%, which means 4753 cases are classified correctly among the total sample in this study. Breast cancer was the most frequent cancer among females regardless of age or residency level. On the other hand, Bone cancer was the most frequent cancer among males who are less than 40 years and regardless of residency level; whereas lung cancer was the most frequent cancer among males who are 41-60 years old and regardless of residency level, while brain cancer was the most frequent cancer in males of the same age category. In addition, prostate cancer was the most frequent cancer among males who were older than 60 years old.

In general, The highest class-wise (Table 6) recall was achieved for Breast cancer (98.5%), reflecting excellent model sensitivity and strong class separability. Prostate cancer followed with a reasonable recall of 78.8%. Conversely, Brain cancer exhibited extremely poor recall (2.7%), with most brain cancer cases misclassified as other types (notably Breast). Bone and Lung cancers also showed limited classification success.

Analysis of off-diagonal values reveals distinct misclassification trends:

- Many Bone and Brain cancer cases were misclassified as Breast cancer (324 and 321 cases, respectively), suggesting potential feature overlap or class imbalance in the training data.
- A notable confusion exists between Lung and Prostate cancers:
  - 272 Prostate cases were predicted as Lung.
  - 59 Lung cases were predicted as Prostate.

The highest precision (Table 7) was observed in Breast cancer (83.2%), indicating reliable classification for this class. Conversely, Brain cancer showed the lowest precision (28.8%), reflecting both low true positives and a high rate of false positives. The performance metrics were computed using standard definitions in multi-class classification (Table 8 and Table 9):

- True Positives (TP): Correct predictions of the class (diagonal entries).
- False Negatives (FN): Actual instances misclassified into other classes (row sum – TP).
- False Positives (FP): Other class instances incorrectly predicted as the class (column sum – TP).
- True Negatives (TN): All other correctly rejected predictions. From these, the following were derived:
- Precision = TP / (TP + FP)
- Recall (Sensitivity) = TP / (TP + FN)
- F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

Table 3: Cancer Type Classification within the interaction between Patient's gender and Patient's Age

Gender	Female						Male					
Age	Less than 40 years		From 40 - 60 Years		More Than 60 years		Less than 40 years		From 40 - 60 Years		More Than 60 years	
Category	%	n	%	n	%	n	%	n	%	n	%	n
Abdomen	0.3	5	0.3	6	0.2	3	1	11	0.5	7	0.3	7
Bladder	0.3	4	0.3	5	2.4	34	1.9	21	9	139	11.4	187
Bone	11.9	171	3.7	87	3.7	52	21.9	239	7.8	121	4	66
Brain	12.1	174	4.7	112	2.4	35	18.9	206	7.5	116	2.7	45
Breast	31.2	449	54.5	1295	56.6	814	0.5	5	1	15	0.9	15
Cervix	5.9	85	10.2	243	13.2	190	0.5	5	1.8	27	1	17
Colon	1.9	28	3	71	7.5	108	2.6	28	7.7	119	7.1	117
Esophagus	0.1	1	0.5	11	0.8	12	0.3	3	0.7	10	0.4	6
Eye	1.4	20	0.5	11	0.2	3	2.7	29	1.1	17	0.5	8
GIT	3.3	48	4.3	103	5	72	4.9	54	3.6	56	5.6	91
Gland	11.9	171	3.5	83	0.8	12	6	66	3.7	58	1.9	31
Kidney	2.2	31	1.6	38	2.6	37	3.4	37	3	47	3.9	64
Liver	0.4	6	2	47	0.7	10	1.1	12	1	16	1.6	27
Lung	0.8	11	2.3	55	4.2	60	1.9	21	13	200	1.5	25
Lymph	8.9	128	2	47	2.9	42	14.1	154	5.9	91	3.3	55
Mouth	0.3	5	0.7	17	0.2	3	0.2	2	2.8	44	0.6	11
Nose	0.8	12	1.5	35	0.3	4	0.5	5	2.4	37	0.3	5
Pancreas	0.5	7	0.5	12	0.6	9	0.3	3	0.4	7	0.4	6
Prostate	0	0	0.2	4	0.1	2	0	0	0.2	3	16.7	275
Reproductive	0.9	13	0.5	12	0.3	4	0.5	5	0.6	9	0.4	7
Respiratory	0.1	2	0.3	7	0.3	4	0.2	2	0.3	5	2.9	48
Skin	1.3	19	0.7	16	0.3	5	4.7	51	2.3	36	3.2	53
Soft tissue	1.7	25	0.5	11	0.4	6	1.7	19	1.9	29	2.5	41
Stomach	1	15	0.5	11	0.2	3	1	11	1.2	18	1.2	20
Others	0.6	8	0.8	20	0.6	9	0.6	7	1.5	23	1.2	20
Total	15.1	1438	25	2383	15.1	1439	11.4	1092	16.2	1548	17.2	1646

Table 4: Chi-Square Test Results by Gender, Age, and Residency

Gender	**Age Group**	**Chi-Square**	**p-value**	**Residency Level**	**Parent Node**	**Predicted Cancer Type**
Female	Less than 40 years	$\chi^2(23) = 78.138$	0.013	Middle-North	9	Breast
				Non-Resident	10	Breast
	41–60 years	$\chi^2(23) = 44.091$	0.036	South	11	Breast
				Middle	12	Breast
				Non-Res., North, South	13	Breast
	More than 60 years	All Residency regions equally likely			5	Breast
Male	Less than 40 years	$\chi^2(46) = 88.514$	0.000	Middle	14	Bone
				Non-Resident	15	Bone
	41–60 years	$\chi^2(46) = 93.871$	0.000	North-South	16	Bone
				Middle, South	17	Lung
				Non-Resident	18	Brain
				North	19	Lung
	More than 60 years	All Residency regions equally likely			8	Prostate

Table 5: Classification Matrix

	Bone	Brain	Breast	Lung	Prostate	Percent Correct
Bone	239	13	324	105	66	32%
Brain	206	19	321	102	45	2.7%
Breast	5	1	2302	14	14	98.5%
Lung	28	18	126	184	272	29.3%
Prostate	0	15	0	59	275	78.8%
Overall Percentage	11.4%	2%	55.1%	14.2%	17.2%	31.6%

Table 6: Class-wise Accuracy (Recall)

Cancer Type	**Correct Predictions**	**Total Actual Cases**	**Recall (%)**
Bone	239	747	32.0%
Brain	19	693	2.7%
Breast	2302	2336	98.5%
Lung	184	628	29.3%
Prostate	275	349	78.8%

Table 7: Estimated Precision by Predicted Class

Predicted Class	**Correct Predictions**	**Total Predicted Cases**	**Precision (%)**
Bone	239	478	50.0%
Brain	19	66	28.8%
Breast	2302	2766	83.2%
Lung	184	464	39.7%
Prostate	275	400	68.8%

Table 8: Confusion Matrix Counts for Each Cancer Type

Cancer Type	**TP**	**FN**	**FP**	**TN**
Bone	239	508	239	4275
Brain	19	674	47	4521
Breast	2302	34	492	433
Lung	184	444	280	2353
Prostate	275	74	397	2515

Table 9: Classification Performance Metrics for Each Cancer Type

Cancer Type	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Bone	50.0	32.0	39.0
Brain	28.8	2.7	5.0
Breast	83.2	98.5	90.2
Lung	39.7	29.3	33.7
Prostate	69.3	78.8	73.7

5 Concluding Remarks

In this study, the classification performance of a decision tree analysis (DTA) model was assessed for five cancer types: Bone, Brain, Breast, Lung, and Prostate. The model achieved an overall accuracy of 31.6%, though its effectiveness varied significantly across the different classes. The confusion matrix indicated that Breast cancer was the most accurately predicted type, with a recall of 98.5% and a precision of 83.2%, resulting in a strong F1 score of 90.2%. Prostate cancer followed with an F1 score of 73.7%, supported by high recall (78.8%) and precision (69.3%). In contrast, the model performed poorly in detecting Brain cancer, which had a recall of only 2.7% and an F1 score of 5.0%, indicating significant misclassification. Bone and Lung cancers also showed limited predictive success, with moderate precision (50.0% and 39.7%, respectively) but low recall, leading to F1 scores of 39.0% and 33.7%. These patterns are visualized in the confusion matrix heatmap (Figure 4), which illustrates frequent misclassification—particularly the over-prediction of Breast cancer and the confusion between Lung and Prostate cases.

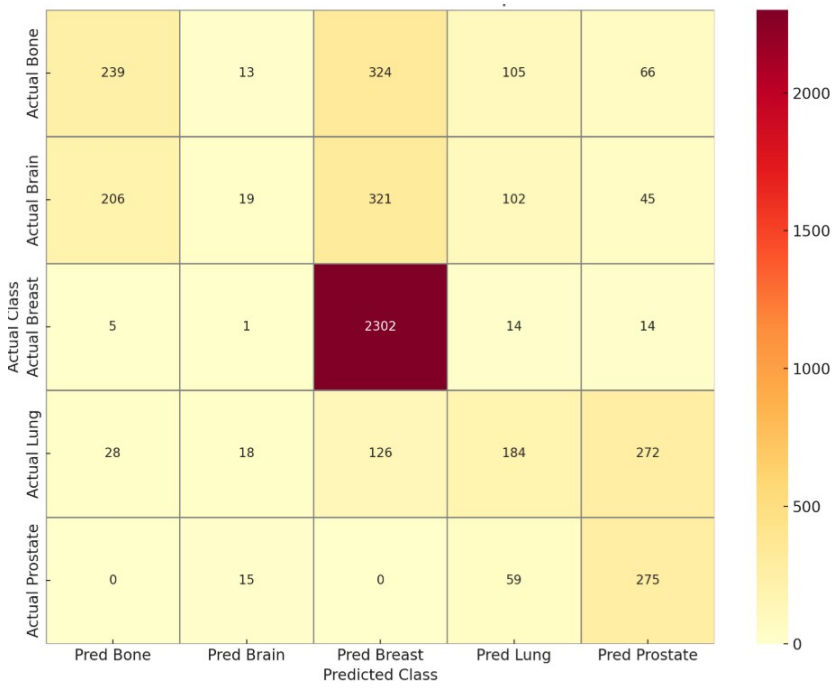

Figure 4: confusion matrix heatmap

The Breast cancer class had the highest classification performance, while Brain cancer was the most misclassified. Bone and Lung cancers showed a conservative model behavior with moderate precision but low recall, indicating hesitation in predicting those classes. The study collected data from multiple hospitals across Jordan, including the King Hussein Cancer Center, using a convenience sampling method. Jordan is recognized as a leading country in cancer treatment among developing nations, ranking first

in the region and sixth globally. The dataset was compiled in collaboration with the Ministry of Health and the Directorate of Non-communicable Diseases, based on official Cancer Registry records. However, lifestyle data was not included, which limits the model's ability to associate cancer incidence with behavioral or environmental risk factors. The DTA model effectively predicted cancer types based on demographic features such as age, gender, and residency region. The results indicated that the central region of Jordan accounted for the highest cancer incidence, especially breast cancer among middle-aged women. Notably, 71.4% of patients (n = 6,815) resided in the central region. These findings hold significant implications for targeted public health strategies, suggesting the need for enhanced early detection, education, and prevention programs in that region.

In conclusion, the integration of machine learning methods like decision tree analysis can provide valuable insights for healthcare policymakers. However, future research should aim to incorporate lifestyle, environmental, and clinical variables to build more comprehensive and predictive cancer models. Additionally, further studies are recommended to explore healthcare-seeking behavior Corallo et al. (2020), premedication, and preventive strategies associated with demographic risk factors. Expanding focus on pre-public healthcare planning and lifestyle-related interventions is essential to reduce cancer incidence and improve early diagnosis and management across Jordan.

References

Abdel-Razeq, H., Al-Ibraheem, A., Al-Rabi, K., Shamiah, O., Al-Husaini, M., and Mansour, A. (2024). Cancer care in resource-limited countries: Jordan as an example. *JCO Global Oncology*, 10:e2400237.

Almaani, N., Juweid, M., Alduraidi, H., Ganem, N., Abu-Tayeh, F., Alrawi, R., et al. (2023). Incidence trends of melanoma and nonmelanoma skin cancers in jordan from 2000 to 2016. *JCO Global Oncology*, 9:e2200338.

Armonk (2020). Ibm spss statistics for windows, version 25.0.

Bhatia, S., Landier, W., Paskett, E. D., Peters, K. B., Merrill, J. K., Phillips, J., and Osarogiagbon, R. U. (2022). Rural-urban disparities in cancer outcomes: Opportunities for future research. *Journal of the National Cancer Institute*, 114(7):940–952.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

Corallo, A., Fortunato, L., Massafra, A., Pasca, P., Angelelli, M., Hobbs, M., Al-Nasser, A. D., Al-Omari, A. I., and Ciavolino, E. (2020). Sentiment analysis of expectation and perception of milano expo2015 in twitter data: a generalized cross entropy approach. *Soft Computing*, 24(18):13597–13607.

Department of Statistics - Jordan (2024). Department of statistics - jordan. `https://dosweb.dos.gov.jo`.

Ferlay, J., Shin, H., Bray, F., Forman, D., Mathers, C., and Parkin, D. (2010). Estimates of worldwide burden of cancer in 2008: Globocan 2008. *International Journal of Cancer*, 127(12):2893–2917.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Springer.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2):119–127.

Kim, J., Gosnell, J. E., and Roman, S. A. (2020). Geographic influences in the global rise of thyroid cancer. *Nature Reviews Endocrinology*, 16(1):17–29.

King Hussein Cancer Foundation & Center (2022). Khcf & khcc in numbers 2022. `https://www.khcc.jo/en/news/khcf-khcc-in-numbers-`.

Masoumi, Z. V., Genderen, J. L., and Mesgari, M. S. (2018). Modeling and predicting the spatial dispersion of skin cancer considering environmental and socio-economic factors using a digital earth approach. *International Journal of Digital Earth*, 13(6):661–682.

McMahon, K., Eaton, V., Srikanth, K., Tupper, C., Merwin, M., Morris, M., et al. (2023). Odds of stage iv bone cancer diagnosis based on socioeconomic and geographical factors: a national cancer database (ncdb) review. *Cureus*, 15(2):e34819.

Ministry of Health (2019). *Annual Statistical Book 2019*. Ministry of Health, Jordan.

Ministry of Health (2021). Annual report of registered cancer incidents in jordan for 2021. `https://www.moh.gov.jo`.

Nisbet, R., Elder, J., and Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Elsevier Inc.

of Health, M. (2021). Annual report of registered cancer incidents in jordan for 2021.

Review, W. P. (2019). Jordan population 2017 (demographics, maps, graphs.

Rokach, L. and Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications*. World Scientific Publishing.

Salem, H. S. (2023). Cancer status in the occupied palestinian territories: types; incidence; mortality; sex, age, and geography distribution; and possible causes. *Journal of Cancer Research and Clinical Oncology*, 149(8):5139–5163.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

Tayefi, M., Esmaeili, H., Saberi, K., Amirabadi, Z., Ebrahimi, M., Safarian, M., et al. (2017). The application of a decision tree to establish the parameters associated with hypertension. *Computer Methods and Programs in Biomedicine*, 139:83–91.

Teli, S. and Kanikar, P. (2015). A survey on decision tree based approaches in data mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(4):613–617.

World Health Organization (2020). Global health estimates 2020: Deaths by cause, age,

sex, by country and by region, 2000-2019. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe. Retrieved December 11, 2020.

Xie, Y., Shi, L., He, X., and Luo, Y. (2021). Gastrointestinal cancers in china, the usa, and europe. *Gastroenterology Report*, 9(2):91–104.

Yu, M., Hazelton, W. D., Luebeck, G. E., and Grady, W. M. (2020). Epigenetic aging: More than just a clock when it comes to cancer. *Cancer Research*, 80(3):367–374.