














## A Retrospective Data Analysis on the Breast Cancer Patients Epidemiology and Awareness in the Eastern Regions of Saudi Arabia

Mohammed Gollapalli<sup>1</sup>, Ameerah Almahmoudi<sup>1</sup>, Linah Saraireh<sup>2</sup>, Atta Rahman<sup>3\*</sup>,  
Sardar Asad Ali Biabani<sup>4,5</sup>, Tahir Iqbal<sup>2</sup>, Rashad Ahmed<sup>6</sup>, Maqsood Mahmud<sup>7</sup>, Dhiaa Musleh<sup>3</sup>,  
Aghiad Bakry<sup>3</sup>, Dania Alkhulaifi<sup>3</sup>

<sup>1</sup> Department of Computer Information Systems (CIS), College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia

<sup>2</sup> College of Business Administration, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia

<sup>3</sup> Department of Computer Science (CS), College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia

<sup>4</sup> Science and Technology Unit, Umm Al-Qura University, Makkah 21955, Saudi Arabia

<sup>5</sup> Deanship of Postgraduate Studies and Research, Umm Al-Qura University, Makkah 21955, Saudi Arabia

<sup>6</sup> ICS Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

<sup>7</sup> School of Computing, Ulster University, Belfast BT15, United Kingdom

Corresponding Author Email: [aaurrahman@iau.edu.sa](mailto:aaurrahman@iau.edu.sa)

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ijdne.190528>

### ABSTRACT

**Received:** 6 September 2023  
**Revised:** 19 September 2024  
**Accepted:** 14 October 2024  
**Available online:** 29 October 2024

#### Keywords:

*data mining, breast cancer statistics, data analysis, machine learning, disease awareness, sustainable solutions*

According to the Ministry of Health (MoH), breast cancer is the leading cause of death for women in Saudi Arabia. Unfortunately, there is a lack of statistical evidence-based data mining on the population data to help establish clinical centers and programs for breast cancer, especially in the isolated and deserted regions of Saudi Arabia. The main objective of the study is to address this issue and the research gap. In this regard, we have made three major contributions in the current study. First, we illustrate the latest developments in data mining techniques applied to screening data relevant to the region and applicable to patients facing similar socio-economic and socio-cultural challenges. In this case, Naïve Bayes and J48 classifiers were investigated along with K-means clustering. Second, we conducted a retrospective analysis of 10 years of clinical hospital data collected between June 2010 and June 2020 from the breast cancer data repository. Third, we conducted a self-awareness survey in which 1,731 participants responded on various critical reasons for better understanding. Several conclusions were drawn for instance, awareness among the men was significantly low compared to ladies and people are not having a habit of regular checkups. Our contributions aim to provide a sustainable solution to the healthcare section and encourage researchers to develop more sustainable and innovative approaches apart from the techniques used in the study.

## 1. INTRODUCTION

Breast cancer (BC) is a highly concerning type of cancer, especially among women worldwide and in Saudi Arabia [1-5]. According to the Ministry of Health (MoH), BC accounts for 22% of all new cancer cases in women. However, controlling and preventing cancer faces numerous challenges, including a lack of BC screening programs, hospital data registries, and socioeconomic variables for the patients [2-5]. These variables also make it more difficult to follow best practices for BC treatment worldwide. In Saudi Arabia, particularly in rural areas, patients, especially young women, who are suspected of having BC encounter various problems as detailed in [3-6]:

1. Limited cultural knowledge and misunderstanding of the disease's seriousness due to BC patients' reluctance to discuss their experiences, posing a barrier to

spreading awareness.

2. Lack of facilities for Bedouins (nomadic people living in distant rural and abandoned locations of the kingdom) to receive medical treatment, causing financial and physical hardships for the patients when seeking adequate healthcare in the urban areas.
3. Additionally, when women from remote areas seek better care in nearby towns, they require the company of a Mahram (male companion) which potentially hinders the process of seeking appropriate healthcare.

Even with improved access to education, many young women in remote regions of Saudi Arabia seem to be unaware of initiatives aimed at promoting awareness, particularly regarding BC. Although significant research into BC has been conducted in the past five to ten years, there has been limited statistical analysis and data modeling of laboratory-based BC screening data sources. This study focuses on the use of

various data mining approaches to patient repository datasets with similar socio-cultural and socio-economic backgrounds, rather than providing a comprehensive literature review solely on BC. The approaches have been shortlisted based on their effectiveness in handling similar issues in the literature [7-10]. Current research aims to:

1. Present evidence to the government through investigation of data modeling and mining techniques using real information and statistics on BC prevalence in remote regions of Saudi Arabia.
2. Use statistical evidence to prioritize between urban and remote locations in Saudi Arabia based on their needs and the number of cases they face, to assist in the provision of necessary services and funding on a priority basis.
3. Compile statistics from clinical data and display them using classification and clustering charts to show the number of patients, the most affected areas, age ranges, and other information, which can help improve medical services in the areas where they are needed, as well as track the patients' situations.
4. Use survey responses to generate statistical data and graphics regarding community awareness.
5. Increase community awareness programs and establish organizations in the region, particularly through mobile awareness campaigns.
6. Contribute to Saudi Vision 2030 for improved healthcare by digital transformation of the healthcare sector.

The rest of the paper is sectioned as: Section 2 provides a comprehensive review of the literature. Section 3 provides materials and methods. Retrospective analysis is carried out in Section 4. Section 5 provides discussions on the results of analyses while Section 6 concludes the paper.

## 2. LITERATURE REVIEW

In the studies [7-10], authors employed deep learning and transfer learning models equipped with data fusion and fine tuning to predict and detect BC from chest radiographs. The techniques are promising in terms of accuracy and other figures of merit.

Alotaibi et al. [11] investigated the risk factors and incidences of BC in Saudi Arabian patients. The data vaults were acquired from the King Faisal Hospital and Research Centre's Saudi Arabian Cancer Registry, which had information on 8,312 patients (98% female, 1.68% male). Age, gender, tumor grade, stage of disease, laterality, topography, marital status, and residence were the eight factors. Stata, R, and INLA were used to analyze the data. The researchers used chi-squared tests, proportions, and Cox regression models to conduct descriptive and inferential statistics (Frequentist and Bayesian). To provide specific outcomes and evidence, the results were backed up with percentages. The data was a little stale, having been collected between 2004 and 2013, and the results of data mining techniques had to be conducted with additional variables. Their research had some limitations, i.e., since the data was partial and may alter or be reinforced, the researchers eliminated many of their observations. Second, the variables provided were not all-inclusive.

Al Diab et al. [12] conducted study in Saudi Arabia to get expert opinions on several aspects of BC. The data was gathered from a variety of research centers, institutions, and

articles published on the internet, such as Science Direct, PubMed, Data Bases, and Libraries. BC epidemiology, people's knowledge of BC and its investigation, causes of BC, survival rate, and metastasis were all discussed by the writers. The study was quite useful and helped the collection of data from people of various educational levels and conditions, such as those on a diet, those suffering from diabetes and obesity, pregnant women, and those using drugs, while also taking gender and age into account. The study, however, was based on 80 studies conducted more than 30 years ago. The authors also noted the need for more research to add to and support the cumulative information for BC screening, planning, and prevention measures to raise awareness and improve BC management, which was one of their work's shortcomings.

Abdelhadi's [13, 14] proposal was to improve the efficiency of female BC awareness. BC awareness campaigns are scarce, yet the disease's frequency has risen over time. To expand BC programs, researchers are encouraging schoolteachers to join. The sample was gathered from 756 instructors in the eastern province in 2005. The purpose of the survey was to assess knowledge of symptoms, BC examination, BC risk factors, and mammography. The article was quite useful in giving a brief overview of female knowledge in the eastern province based on criteria. Unfortunately, no methodology was used in the research, and no analysis tools or data mining software were used to produce exact results. Furthermore, the research should have been conducted on larger samples to collect more BC cases from both genders and across all age groups.

By adopting three data mining algorithms and determining the elements affecting survivorship rates, Othoum and Al-Halabi [15] conducted study on Saudi Arabia Data Registry patients to find the best technique in forecasting survivability rates. The data was gathered from 1358 records in various Saudi Arabian hospitals (BCs' survival = 1260, BCs' non surviving = 98) with 9 attributes. The data was cleaned and pre-processed using Weka Data Miner. Weka was then utilized for data mining and using classification models such as Neural Networks, Decision Trees, and Nave Bayes, which displayed the accuracy, sensitivity, and specificity predictions that were used to evaluate the performance. Decision Tree had the best accuracy (97%) and sensitivity (98%) after incorporating these models (98%). In other words, the most accurate predictor was determined to be Decision Tree. However, the data used was skewed because the survival and non-survival rates were not equal or nearly equal, which could have influenced the outcome. In addition, the model required examination and manual comments from specialists to determine whether the generated result was rational. To obtain accurate results, just a limited prediction method was evaluated.

Akbari et al. [16] conducted a retrospective investigation on the state of BC in Iran to determine the most effective early detection methods for enhancing preventative strategies. Between 1998 and 2014, 3010 cases of BC in women in Tehran were examined in this study. The information was gathered from the Tehran Breast Cancer Research Centre (BCRC) and included 32 different features. Following the data cleaning and saving, data mining techniques were used in R to define the problem, prepare the data, develop models, and evaluate the results. In addition, the data was examined using R and descriptive (univariate) analysis. The most important finding of this study was that the BC patients were between the ages of 40 and 50, were generally married with two children, and had no history of smoking or diabetes. Furthermore, most of the patients were diagnosed in the

second stage of BC. We considered this study to be thorough, but the descriptive analysis revealed no connections between the qualities, which made it difficult to make solid evidence-based conclusions.

Salem et al. [17] used data from the Lebanese population to conduct an observational retrospective analysis to determine the link between BC density and occurrence. In this study, 1049 Lebanese women's data was assessed, and screen film mammography or digital mammography was reviewed retrospectively. The breast density was divided into two categories, and the patient's age was divided into six groups; each type of density was evaluated with the age groups and compared to elements that could influence density and BC occurrences. The findings revealed that 76.4% of the patients were between the ages of 30 and 39 and had thick breasts. The timeworn data acquired in their study from 2010 to 2012 was the study's shortcoming.

In the United Arab Emirates (UAE), Elobaid et al. [18] conducted a qualitative study among women to determine the factors that might lead to a delay in presenting BC diagnosis and seeking counsel or treatment following self-discovery. Thematic analysis was employed in this qualitative study to uncover the motivations for delaying and seeking therapeutic assistance. Because the approach included gathering replies to questions that could be considered 'sensitive,' this process required conducting face-to-face interviews with 19 patients. Furthermore, the authors created a model, which is a conceptual framework that depicts the amount of BC patients' interactions and how it influences their decision to seek counsel or therapy. According to the findings of this study, culture was the most important issue to consider when talking to women who had BC. The lack of information among women regarding the symptoms and screening process was the second key reason. Despite the positive results, we believe this was a biased study because only women were interviewed, as well as only individuals with BC in stages II and III who had received formal education [19].

Tchier and Alharbi [20] recommended using a combination of Fuzzy Relational Model and Genetic Algorithm to find the best strategy for early identification and diagnosis of BC in Saudi Arabia. In terms of diagnosis tests and features, the acquired dataset was equivalent to the Wisconsin BC Diagnosis Repository (with 260 cases and 9 features). An expert must first develop a fuzzy based rule, after which the genetic algorithm will determine whether the rule is consequential or requires subset rules to be supported. The result was a benign (malignant) diagnosis with a low confidence value (high). The system guidelines described proved to be successful in early BC diagnosis situations, lowering treatment costs. The disadvantage of this technique was that it had good performance with fewer attributes and rules; otherwise, the genetic algorithm encoding would require a huge dimensional array and many generations. The system's effectiveness would likewise suffer because of this. The combination of tools utilized to construct this technique was also not mentioned by the researchers.

Cakir and Demirel [21] created a "Therapy Assistant" program to advise clinicians about BC treatment options. Based on the Accuracy and Decision Table results, the application applies data mining Classification Algorithms (Decision Tables and Multilayer Perception) on the Ankara Oncology Hospital data repository to determine therapy best practices for patients. The disadvantage of this strategy, we discovered, is that when the number of patients grows, the data

must be manually updated to keep the model up to date, which is time consuming and costly to maintain.

Using data mining approaches, Torrents-Barrena et al. [22] proposed one of the greatest classification accuracies in BC. In their research the authors constructed two stages system that uses multiple data mining classification algorithms (Decision Tables, SVM, radial basis function network (RBFN), GA and Random Forest) on two medical datasets of patients diagnosed with or without BC to help the clinicians. The goal was also to simplify classification, i.e. determining whether the patient was a BC patient. The lack of a mechanism to display the results in an understandable manner was a weakness of this study, making it extremely difficult for clinicians to comprehend and analyze the results.

The researchers also proposed a strategy for detecting malignant tumors in mammographic pictures. To acquire the most accurate findings, they followed several processes that included first identifying texture patterns on mammography pictures, then performed binary classification on the images using data mining classification methods (SVM) and four different types of kernel functions. Before adopting this procedure, it is necessary to clean the pixelated mammography images, which takes time and slows the decision-making process that affects patient cases.

Shrivastavat et al. [23] employed the Decision Tree to provide a quick analysis of BC data. The dataset was obtained from the UCI Machine Learning repository, and the technique gave the basic decision tree model derived findings. As part of their experimental findings, the authors advised comparing the accuracy percentages of at least two algorithms. Their research was limited by the lack of a cleaning method for the dataset before working on it. Furthermore, the dataset only had 699 instances and 10 attributes. Talukdar and Kalita [24] studied BC diagnosis using Weka Data Miner and several data mining methodologies. The authors talked about the prospect of determining the type of BC, whether it's malignant or benign, at an early stage. The experiment was broken down into three stages: data gathering, pre-processing, and classification. The researchers' work had a good side in that they applied two algorithms (ZeroR and J48) in detail, analyzing the findings, comparing them, and then selecting the best, which was J48, based on the right accuracy percentage for ZeroR. ZeroR had a percentage of 63.86%, whereas J48 had a percentage of 95.37%. The authors expressed an interest in expanding their research to include a larger number of patients as well as other organizations and universities.

Chaurasia and Pal [25] presented a study in which participants tested and compared various classification algorithms, as well as their performance, to determine the most accurate algorithm for identifying BC. Using the Wisconsin Datasets Repository's BC dataset, the methods employed were K Nearest Neighbors Classifier (IBK), Best First (BF) Trees, and Sequential Minimal Optimization (SMO). Consequently, SMO was shown to be the most accurate algorithm in their trial, with a prediction accuracy of 96.2%. The disadvantage of this technique was that the properties of the used BC dataset were not viewed as direct indicators of BC patients.

Chaurasia and Pal [26] proposed a BC detection diagnostic system based on three methods: RepTree, radial basis function (RBF) Network, and simple Logistic. Three alternative classification strategies were tested on the survivability rate prediction of BC dataset in this methodology, with the goal of determining the optimum classification technique for predicting survivability rates. With 74.47% accuracy and 0.62

seconds to converge the model, the results demonstrated that the simple logistic classification strategy was the most accurate and helpful in forecasting survivability rates.

Chaurasia et al. [27] carried on their research and used different prediction models constructed using three algorithms: Naive Bayes, RBF Network, and J48, which measured survivorship based on two characteristics: benign and malignant types. The results showed that Naive Bayes was the most accurate algorithm, with a score of 97.36%, and J48 was the least accurate, with a score of 93.41%. The outcomes, on the other hand, are determined by the data's complexity and values. Furthermore, because it is medical data, the outcomes may differ from other types of data, as well as ontologies and semantics for medical short codes.

Aloraini [28] presented a study that tested different machine learning techniques to determine the accuracy of diagnosing benign versus malignant BC. Bayesian Network, Nave Bayes, ADTree, J48, and Multilayer NN were the five algorithms studied in this study (Neural Networks). The tests were performed on the Wisconsin BC Dataset Repository [28], which includes the key features required in laboratory diagnosis. The findings of this experiment revealed that the Bayesian Network method provided the most accurate results. The study has some limitations, such as not considering genes-regularity networks-based methods as a baseline for comparison purposes, even though they are likely to yield the most accurate results. Matsen et al. [29] provided a strategy for gaining a better understanding of young BC women's decision to return for genome sequencing study. Participants in this study were diagnosed with BC at the age of 40 or younger and came from the 'Breast Cancer Program for Young Women' in the US. A total of 1080 people were polled on their preferences for making decisions based on genome sequencing results, as well as the variables that influenced their decision. Multinomial logistic regression was used to examine the results. This study had some limitations; the surveyed group did not differ in criteria such as education, race, or ethnicity, which could have influenced their preferences.

BC feasibility study was conducted by Jonsdottir et al. [30] for the creation of a predictive outcome model. The data was gathered from 257 patients who had survived BC for at least 5 years after being diagnosed. The database was divided into three categories: Base-DS, which had over 150 features picked manually by the author, Med-DS, which had 22 features chosen by the doctor, and Small-DS, which was created from the Base-DS and had only five features. Weka Data Miner was used to pre-process all the databases, and three different classification methods were used: the Naive Bayes (NB) Algorithm, the Decision Tree (J48) Algorithm, and a variety of Meta Algorithms. The results revealed that the number of characteristics has no bearing on the classification process and that the algorithms' performance is unaffected. NB and J48 were not found to be better than the meta-algorithms. The authors admitted in their paper that they didn't pay attention to parts of the classifier's performance, which could have resulted in certain results being overlooked or misconstrued. In addition, the trials should have considered any new features that might have an impact on the results [31, 32].

Based on the comprehensive literature review, it is evident that BC detection and prediction is among the hottest areas of research in medical informatics. In this regard, several studies have been conducted and useful results have been obtained. The datasets utilized in this regard are mainly based on images, clinical data and reviews and prescriptions provided by the

medical experts. Additionally, the datasets collected from the surveys are also found useful in this regard. As far as the schemes are concerned, researchers have investigated deep learning and transfer learning when it comes to the image's dataset. Because such techniques are promising over the image process tasks [33-38]. Nonetheless, when it comes to the clinical data with classification applied already, machine learning has its vital role especially with its ensemble learning counterparts [39-45]. Likewise, when it comes to the survey and statistical data, data mining approaches are promising candidates [46-50].

### 3. MATERIALS AND METHODS

#### 3.1 Clinical data

The study covers a 10-year period of clinical hospital data (June 2010 to June 2020) of Breast Cancer (BC) patients treated at King Fahad University Hospital (KFUH) in the eastern province of Saudi Arabia. The study workflow is illustrated in Figure 1 and involves data cleaning, integration of clinical and pathological data, and data transformation for better quality. Patient-sensitive information such as phone numbers, home location and national identities was excluded, while clinical, laboratory, and pharmacy data needed for analysis were included.

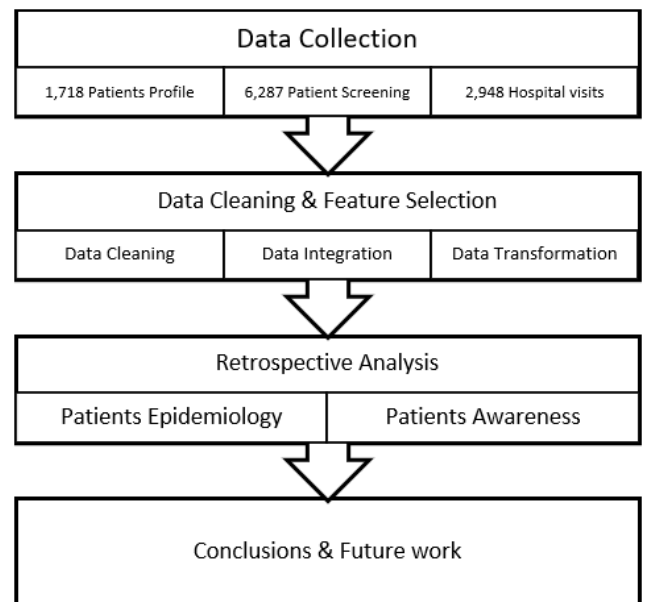


Figure 1. Flowchart on the conducted BC study

In the preprocessing phase, data cleaning was performed including handling missing/duplicate values, and outliers. In the integration phase, clinical and pathological data were integrated. Data transformation including normalization and scaling have been employed for better data quality.

The study included patients of Saudi and non-Saudi nationalities of various ages, mostly from the country's eastern provinces. The clinical dataset comprised 6,287 BC screening records, 2948 hospital visit records, and 1718 BC patient profile records. It included patients of all cancer stages. The primary goal was to better understand these breast cancer patients of both genders using clinical information to identify commonalities among different groups of breast cancer

patients. The screening data's features/variables are more effective for diagnosing BC, while other demographic data is useful for general purposes and public health decision-making (Table 1) [51, 52].

**Table 1.** BC features for retrospective analysis

Variable	Description
<b>Patient Profiles</b>	
Sex	A male/female indicator of patient
Age	Age of the patient when diagnosed
Marital status	Married/Single or unknown
Nationality	Country of origin
Residence	Patient address
Visits count	How many times patient visited the hospital
Year of last visit	Date of the last patient visit to the hospital
Employment status	Patient working for government or non-government
<b>Patient Hospital Visits</b>	
Visit type	Patient visit type (CP, EP, IP, OP, SP)
Description	The doctor's notes written when the patient visited the hospital
ICD 9	International classification of disease, 9 <sup>th</sup> revision
Start date	Time patient spent in the hospital (in days)
ICD 10	International classification of disease, 10 <sup>th</sup> revision
<b>Disease Screening Data</b>	
Procedure	The type of procedure carried out during the screening process
Result date	The date of screening results
Result time	The time of screening results
Line No	The patients screening notes line number
Screening summary	Different sets of screening notes written after investigating the breast cancer patients x-ray screening

### 3.1 Ethical reviews

The institutional review board (IRB) of all participating institutions have provided their permission to this study through the official ethical review process. There were no modifications in the patient's clinical therapy because of this investigation. To ensure the patients' privacy and confidentiality, every information was handled with care.

## 4. RETROSPECTIVE ANALYSIS

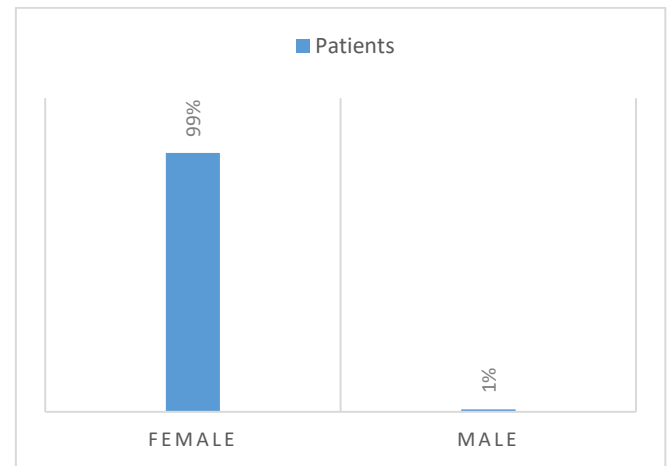
In this section, descriptive statistics are used to describe a retrospective study conducted as the first step in analyzing the data related to breast cancer (BC) patients [6]. The study aimed to gain a better understanding of the determining factors that impact BC patients admitted to the hospital and treated in different departments. The retrospective analysis involved utilizing classification models such as Naïve Bayes and J48 classifiers (a Java implementation of the C4.5 algorithm), as well as clustering using the popular K-Means algorithm with different values of the number of clusters (denoted as k) depending on the data's nature. Generally, three clusters were used to extract various knowledge clusters. This analysis was conducted using the open-source Weka data mining software, which is often used by researchers as evident from the literature review. For the J48 classifier, parameters such as split type (binary), confidence ranging from 50% to 80%, and

minimum number of objects were tuned/utilized. However, for the Naïve Bayes classifier, a default set of parameters was investigated as mentioned in the Weka tool.

### 4.1 Patients epidemiology

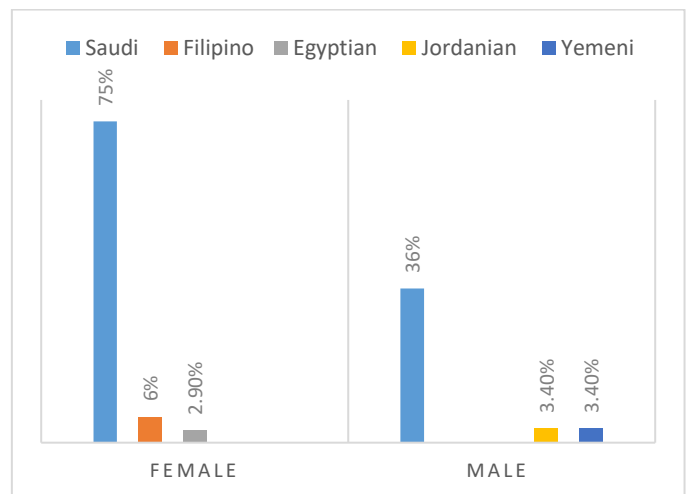
#### 4.1.1 Patients' genders

Firstly, a gender-based analysis was performed. As shown in Figure 2, Female patients accounted for 99% (1,731 patients) of all BC patients in the hospital, much outnumbering male patients, who accounted for only 1% (58 patients). These female BC patients were on average 51 years old, while male BC patients were on average 39 years old.



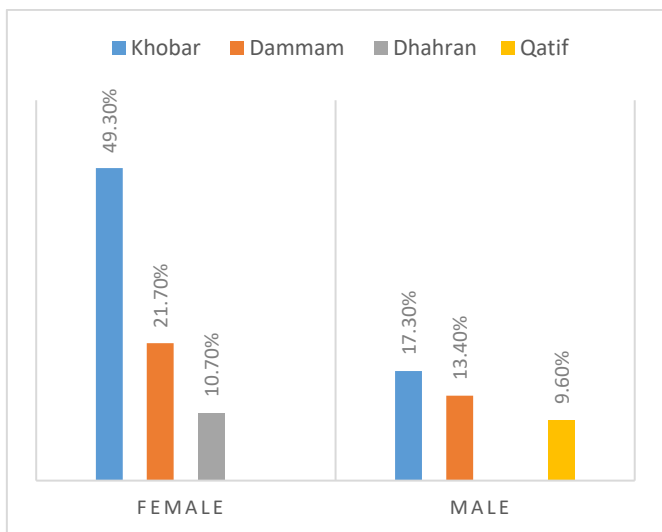
**Figure 2.** BC patients overall gender statistics

As illustrated in Figure 3, the top three nationalities among female BC patients were Saudi citizens (75%), followed by Filipinos (6%), and Egyptians (2.9%). While the top three nationalities of male BC patients were Saudi citizens (36%), followed by the Jordanians (3.4%), and Yemeni (3.4%) citizens.



**Figure 3.** BC patients' nationalities

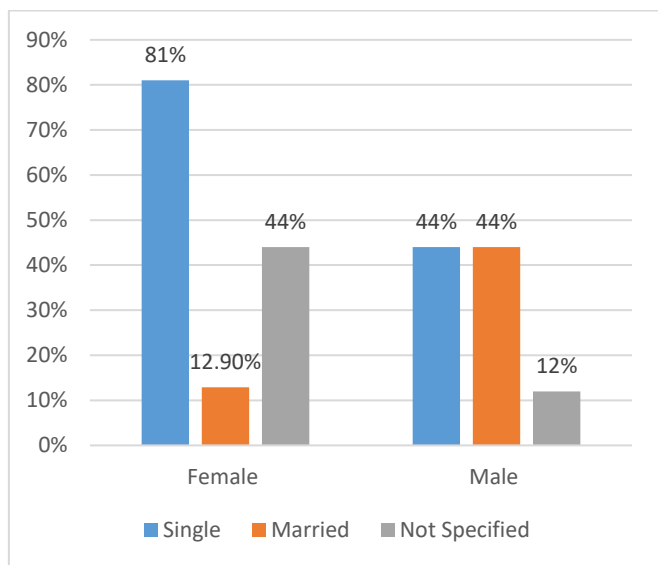
Figure 4 shows the BC patients density across genders. As can be seen, the top three cities of the most female BC patients were in the cities of Khobar (49.3%), followed by the cities of Dammam (21.7%) and Dhahran (10.7%). On the other hand, male BC patients were similarly located in the cities of Khobar (17.3%), followed by Dammam (13.4%) and Qatif (9.6%).



**Figure 4.** BC patients' genders across cities

#### 4.1.2 Marital status

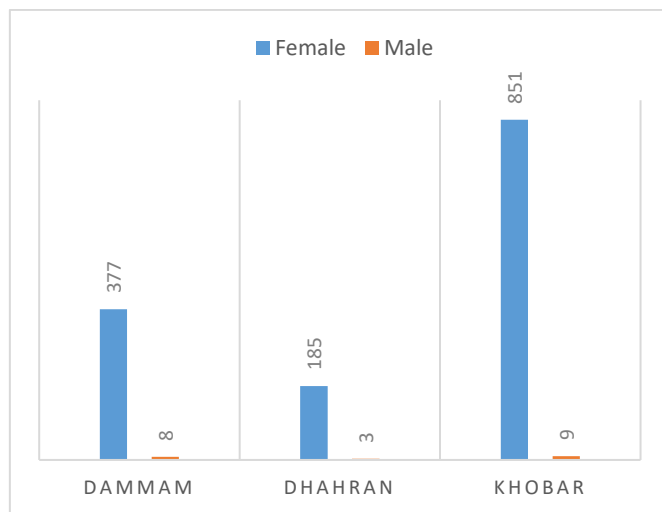
Looking at the marriage status (see Figure 5), majority of the female BC patients (81.21%) were singles, while 220 patients were married (12.9%), and 102 patients did not specify their marriage status (5.8%). Among these women, the average number of visits came to the clinic were 73, while for the males, the average was around 48 visits per BC patient.



**Figure 5.** BC patients' marital status across genders

#### 4.1.3 Top BC patient's residential cities

As illustrated in Figure 6, the top three cities where most BC patients were located across the province were found to be Khobar (860 BC patients, 49%), followed by the cities of Dammam (385 BC patients, 22%), and Dhahran (188 BC patients, 11%). In these cities, the number of female BC patients in the city of Khobar were found to be 851 patients (98.9%), followed by 377 BC patients in the city of Dammam (97.9%), and 185 BC patients in the city of Dhahran (98.4%). While the statistics for the males were 9 BC patients in Khobar (1%), followed by 8 BC male patients in Dammam (2%), and 3 BC male patients in the city of Dhahran (1.59%), respectively.



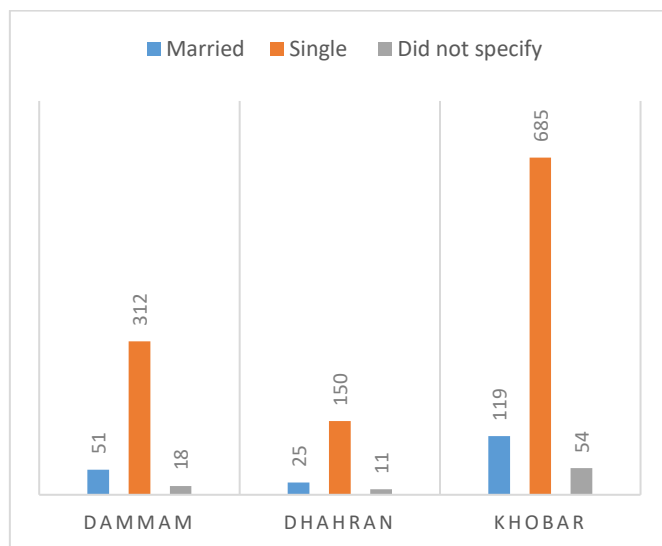
**Figure 6.** BC patients' genders in the top 3 cities

#### 4.1.4 Family history

After identifying the top BC patients located in cities, we look at the nationalities of each of these patients. The top three patient nationalities in the city of Khobar were Saudi citizens (77%), followed by the Egyptians (77%), and Yemeni citizens (2.3%). For the city of Dammam, the top three nationalities were found to be Saudi citizens (64.4%), followed by the Egyptians (5%), and the Yemeni citizens (3.1%). While in the city of Dhahran, the Saudi BC patients were (58.1%), followed by the Egyptians (4%), and Yemeni citizens (2.7%), respectively.

#### 4.1.5 City wise marital statistics

Looking at the marital status across the top cities identified, as illustrated in Figure 7, we see that the marital status of BC patients living in the city of Khobar were 79% singles, 14% were married, and 6% did not specify their marital status. The marital status of BC patients living in the city of Dammam were 81% singles, 13% married, and 4% did not specify their marital status. For the city of Dhahran, around 79% were singles, 13% were married, and 6% did not specify their marital status.



**Figure 7.** BC patients' marital status in the top 3 cities

## 4.2 Patients awareness

This is the second part of our research in which data was collected randomly by questioning (distributing) an awareness survey to the public to measure their awareness of BC and the reasons for their lack of awareness. The researchers used their training and doctor’s help in conducting the survey questions and the survey results were monitored constantly to check the accuracy of the results collected. A total of 1,084 responses were received among which female participants were 1,029 and 55 males participated. The language of the questionnaire was in both English and native Arabic. The Arabic language answers given by the respondents were translated to the English languages by the data collectors. Few questionnaires were excluded during the data analysis because of insufficient information. Descriptive statistics were applied to the collected data using open-source Waikato Environment for Knowledge Analysis (Weka) Data mining software and useful conclusions have been drawn consequently.

### 4.2.1 Family history

The survey question was asked if the participant’s family had a history of BC. As illustrated in Figure 8, the majority responded “No” (435, 90%), indicating that the family had no history of BC, and the remaining (48, 10%) verified that their families had BC patients.

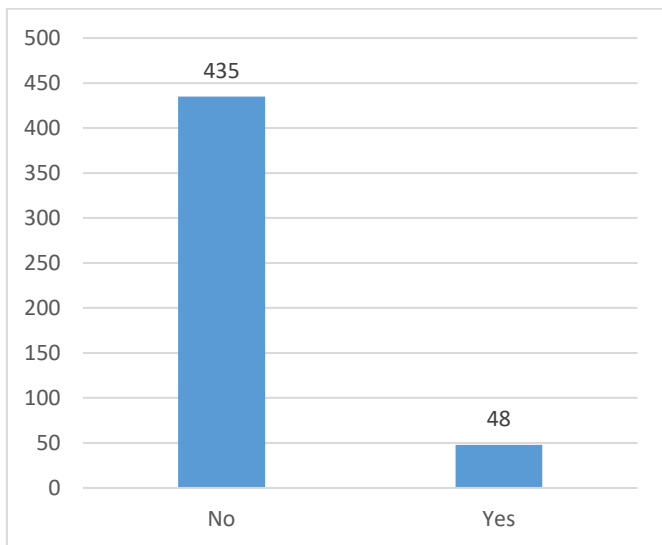


Figure 8. Family history of BC

### 4.2.2 Mammography examination

The survey question was asked if the person had undergone a clinical examination for BC in a hospital setting, which included mammography X-rays to detect any malignancies in the breasts. Most of the people (90%) said “No”, while the rest said they had undergone it at some stage of their life (10%).

### 4.2.3 Non-examination reasons

The following query was asked as to why they didn't do the clinical examination. As illustrated in Figure 9, majority of the participants (62%) said they didn't do the exam because they didn't have any signs of the condition, while others (11%) said they didn't know about mammography, and others had personal reasons including shyness (11%) and dread of the results (16%).

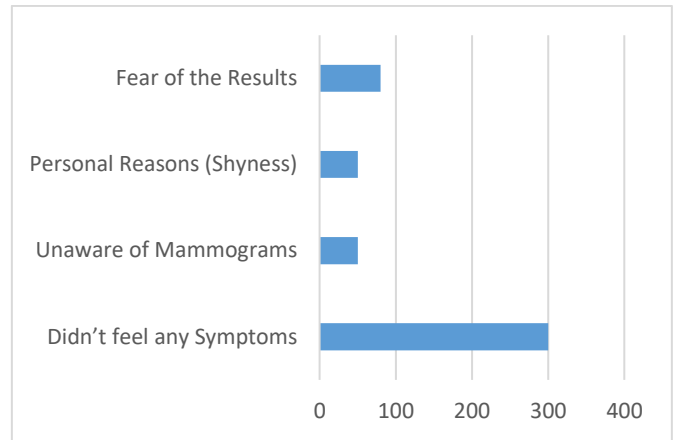


Figure 9. Non-self examination reasons

### 4.2.4 Hosting awareness programs

This question is around whether the participant's city organizes gatherings and events for BC awareness campaigns. As illustrated in Figure 10, most of the participants (74%) verified that their city holds BC awareness activities. Other cities did not host any BC awareness events (7%), and others had no idea whether their towns had these events (17%).

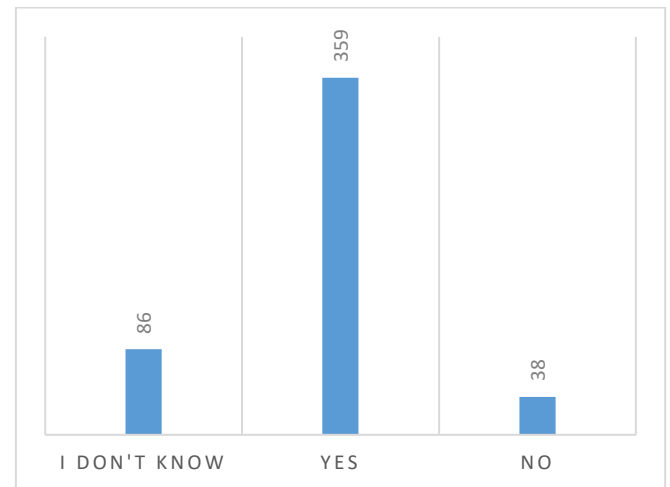


Figure 10. BC awareness campaigns

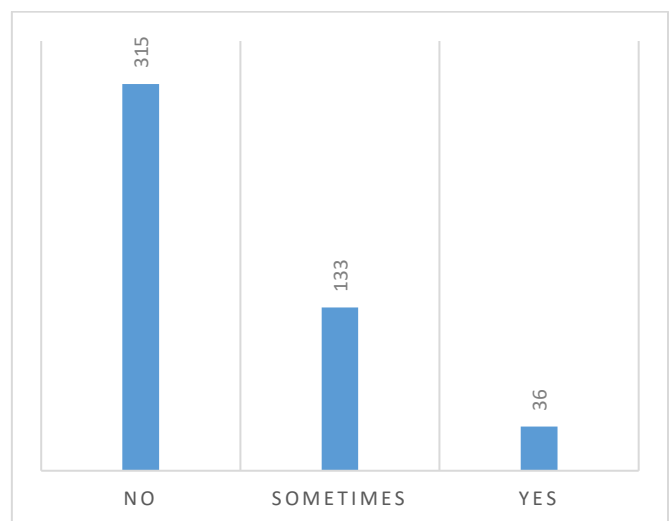


Figure 11. Attending BC programs

The question was asked as to whether the participants are attending the BC awareness activities in their respective cities. As shown in Figure 11, most of the people said they don't go to these programs (315, 65%), while others said they go occasionally (133, 27%), and the rest responded in positive (36, 7%).

## 5. DISCUSSIONS

In our analysis of gender attributes, we found that 95% more female participants than male participants took part, with the majority falling between the ages of 20 and 30. Regardless of gender, most participants indicated that their families have no history of breast cancer (82.05% female, 92.85% male). Additionally, most of the participants (25.6% female, 12.6% male) were from the city of Dammam. The analysis also revealed that 70.99% of females never have a clinical examination, with 65.81% of them stating that they never notice any obvious changes or symptoms of breast cancer in their breasts. In contrast, 94.8% of male participants have never performed a self-examination. However, 35.7% of females express more concern about self-examination and observation.

Regarding clinical examinations, our findings indicate that only 28% of participants have undergone a clinical examination for breast cancer, with the majority falling between the ages of 41 and 50. On the other hand, 72% of the participants have never had a clinical examination, and most of them were between the ages of 20 and 30. It's worth noting that the majority (72.93%) of participants who have never had a clinical examination also never or rarely examine their breasts for any signs or symptoms of breast cancer. A clinical examination is crucial for early detection, and those who have never undergone it should consider performing regular self-checks.

Moreover, only 25.82% of individuals with a family history of breast cancer (BC) have undergone a clinical examination. This rate is lower than expected given their higher risk of developing BC compared to the general population. Several factors hinder people from seeking clinical examinations for BC; for instance, 10% of women are unaware of mammography examinations. Additionally, 8% of individuals cited fear of examination results due to social factors such as shyness, and 8.7% reported being afraid of the results. Furthermore, only 16.17% of those attending BC awareness seminars have undergone clinical breast examinations.

In the analysis of BC treatment, it was found that many participants never visit a doctor because they do not notice any visible changes in their breasts (53.8% female, 56.1% male), and 74.22% of females prefer to have a female doctor conduct the examination. However, 44.47% of females stated that they are occasionally afraid of visiting the doctor, while 42.53% claimed they are always afraid. Additionally, 35.07% mentioned being very busy, and 40.01% had faced trouble in securing a practitioner's/doctor's appointment for a clinical examination. Finally, it was discovered that only 11.43% of females and just 5.2% of males participate in BC awareness programs in their localities, which is a significantly low number for both genders.

We found that 83% of the participants were aware of breast cancer awareness programs in their city, with 96.77% of them being female. Among these females, 38.37% were between the ages of 20 and 30, 35.83% were between the ages of 41 and

above, and 4.4% were between the ages of 10 and 19. Additionally, 5% stated that their city had no awareness programs, while 12% were unsure if their community had any. The cities with the highest awareness of breast cancer programs were Dammam (26%), Khobar (14.27%), Riyadh (9.93%), Dhahran (6.34%), and Jeddah (3.91%). On the other hand, the cities with the lowest awareness were Safwa (0.105%), Arar (0.21%), Hafr-Al-Batin (0.21%), Jouf (0.21%), and Khafji (0.21%). Despite having awareness programs across the country, only 13.11% of people attended these programs regularly, while 29.66% attended occasionally. This highlights the immediate need to increase awareness program activities and encourage regular attendance. Regarding the study's limitations, the dataset was collected from a single hospital. For a more comprehensive analysis, it is recommended to have diverse and augmented datasets from various sources. This issue can be addressed in future research. Furthermore, exploring hybrid intelligence-based approaches for improved analyses is also suggested [53-56].

## 6. CONCLUSIONS

Breast cancer is a significant health concern in Saudi Arabia, especially among women in rural areas. We began by reviewing existing literature on data mining related to regions with similar socio-cultural and socio-economic factors, as well as global research on breast cancer data mining, in order to understand its relevance to Saudi Arabia. We also discussed the advantages and disadvantages of data mining to determine the best approach for extracting evidence-based outcomes for preventing breast cancer, particularly among young women.

In the next stage of our work, we conducted a retrospective data analysis using a 10-year (June 2010 to June 2020) breast cancer data repository from King Fahd University Hospital in the eastern province region of Saudi Arabia, where breast cancer cases have been on the rise over the past decade. For the mammography screening laboratory data, we utilized the Weka open-source Data Miner software. We employed descriptive statistics for retrospective analysis and used classification models (Naïve Bayes and J48 classifiers) as well as clustering (K-Means clustering) to extract various knowledge clusters. Data cleaning pre-processing techniques were applied, including data discretization and NominalToBinary methods. As a future step to gain deeper insights into patients' behavior, we plan to use Python Visualization to create knowledge discovery domains, such as Plot Matrixes.

In the third stage of our research, we conducted a self-awareness survey with 1,084 participants. The survey revealed that the majority of participants were female (1,029 females and 55 males). Most participants were aged 20-30 and lived in the city of Dammam, as anticipated. Data analysis suggests that in the coming years, females aged 40-50 are at the highest risk of developing breast cancer. When asked about family history of breast cancer, most participants believed that having a close relative with the disease does not increase their own risk. The data also indicated that many participants feel uncomfortable discussing symptoms with male doctors and avoid hospital visits unless the doctor is female. Additionally, most participants do not perform self-checks and refrain from visiting doctors due to lack of symptoms and time constraints. Despite awareness programs being available in their cities, many participants do not attend.



Our research aims to help the government, clinical centers, and hospitals by using data mining and visualization to detect breast cancer in its early stages, reduce the number of cases, and provide proper treatment to as many people as possible. This involves building more qualified hospitals, spreading awareness programs, and establishing new specialist centers in highly affected areas. While our focus is on breast cancer patients in the Kingdom, we believe our contributions will also be helpful in other countries, especially in regions with similar cultural and socioeconomic backgrounds where women face similar obstacles.

## REFERENCES

- [1] Alshahrani, S.M., Fayi, K.A., Alshahrani, S.H., Alahmari, D.S., Al Bejadi, K.M., Alahmari, D.M., Alshahrani, T.M., Alsharif, M.N. (2019). Assessment awareness of public about breast cancer and its screening measurements in Asir Region, KSA. *Indian Journal of Surgical Oncology*, 10: 357-363. <https://doi.org/10.1007/s13193-019-00899-5>
- [2] Fayed, R., Hamza, D., Abdallah, H., Kelany, M., Tahseen, A., Aref, A.T. (2017). Do we need regional guidelines for breast cancer management in the MENA region? *MENA Breast Cancer Guidelines project. Ecancermedicalscience*, 11: 783. <https://doi.org/10.3332%2Fecancer.2017.783>
- [3] Abdel-Aziz, S.B., Amin, T.T., Al-Gadeeb, M.B., Alhassar, A.I., Al-Ramadan, A., Al-Helal, M., Bu-Mejdad, M., Al-Hamad, L.A., Alkhalaf, E. H. (2017). Perceived barriers to breast cancer screening among Saudi women at primary care setting. *Asian Pacific Journal of Cancer Prevention*, 18(9): 2409-2417. <https://doi.org/10.22034%2FAPJCP.2017.18.9.2409>
- [4] Balqis, A., Mazen, H., Areej, B., Abdulaziz, A., Hana, A. (2018). Age distribution and outcomes in patients undergoing breast cancer resection in Saudi Arabia. *Saudi Medical Journal*, 39(5): 564-469. <https://doi.org/10.15537/smj.2018.5.21993>
- [5] Ahmed, H.G., Ashankyty, I.M., Alrashidi, A.G., Alshammeri, K.J., Alrasheidi, S.A., Alshammari, M.B., Research, C. (2017). Assessment of breast cancer awareness level among Saudi medical students. *Journal of Cancer Prevention & Current Research*, 7(4): 00241. <https://doi.org/10.15406/jcpcr.2017.07.00241>
- [6] Gollapalli, M., Al-Jaber, E., Selham, J., Al-awazem, W., Al-Sayoud, Z., Al-Qassab, S. (2015). Saudi rural breast cancer prevention framework. In 2015 International Conference on Cloud Computing (ICCC), Riyadh, Saudi Arabia, pp. 1-8. <https://doi.org/10.1109/CLOUDCOMP.2015.7149653>
- [7] Arooj, S., Khan, M.F., Shahzad, T., Khan, M.A., Nasir, M.U., Zubair, M., Ouahada, K. (2023). Data fusion architecture empowered with deep learning for breast cancer classification. *CMC-Computers, Materials & Continua*, 77(3): 2813-2831. <https://doi.org/10.32604/cmc.2023.043013>
- [8] Khan, M.B.S., Nawaz, M.S., Ahmed, R., Khan, M.A., Mosavi, A. (2022). Intelligent breast cancer diagnostic system empowered by deep extreme gradient descent optimization. *Mathematical Biosciences and Engineering*, 19(8): 7978-8002. <https://doi.org/10.3934/mbe.2022373>
- [9] Arooj, S., Zubair, M., Khan, M.F., Alissa, K., Khan, M.A., Mosavi, A. (2022). Breast cancer detection and classification empowered with transfer learning. *Frontiers in Public Health*, 10: 924432. <https://doi.org/10.3389/fpubh.2022.924432>
- [10] Nasir, M.U., Ghazal, T.M., Khan, M.A., Zubair, M., Rahman, A.U., Ahmed, R., Al Hamadi, H., Yeun, C.Y. (2022). Breast cancer prediction empowered with fine-Tuning. *Computational Intelligence and Neuroscience*, 2022(1): 5918686. <https://doi.org/10.1155/2022/5918686>
- [11] Alotaibi, R.M., Rezk, H.R., Juliana, C.I., Guure, C. (2018). Breast cancer mortality in Saudi Arabia: Modelling observed and unobserved factors. *PLoS One*, 13(10): e0206148. <https://doi.org/10.1371/journal.pone.0206148>
- [12] Al Diab, A., Qureshi, S., Al Saleh, K.A., Al Qahtani, F.H., Aleem, A., Algamdi, M. (2013). Review on breast cancer in the Kingdom of Saudi Arabia. *Middle-East Journal of Scientific Research*, 14(4): 532-543. <https://doi.org/10.5829/idosi.mejsr.2013.14.4.7327>
- [13] Abdelhadi, M.S. (2008). Breast cancer management delay-time for improvement: A reflection from the eastern province of Saudi Arabia. *Journal of Family and Community Medicine*, 15(3): 117-122.
- [14] AbdelHadi, M.S. (2006). Breast cancer awareness campaign: Will it make a difference? *Journal of Family & Community Medicine*, 13(3): 115-118.
- [15] Othoum, G., Al-Halabi, W. (2011). Predicting breast cancer survivability rates: For data collected from Saudi Arabia Registries. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*.
- [16] Akbari, M.E., Sayad, S., Sayad, S., Khayamzadeh, M., Shojaee, L., Shormeji, Z., Amiri, M. (2017). Breast cancer status in Iran: Statistical analysis of 3010 cases between 1998 and 2014. *International journal of breast cancer*, 2017(1): 2481021. <https://doi.org/10.1155/2017/2481021>
- [17] Salem, C., Atallah, D., Safi, J., Chahine, G., Haddad, A., El Kassis, N., Maalouly, L.M., Moubarak, M., Did, M., Ghossain, M. (2017). Breast density and breast cancer incidence in the Lebanese population: Results from a retrospective multicenter study. *BioMed Research International*, 2017(1): 7594953. <https://doi.org/10.1155/2017/7594953>
- [18] Elobaid, Y., Aw, T.C., Lim, J.N., Hamid, S., Grivna, M. (2016). Breast cancer presentation delays among Arab and national women in the UAE: A qualitative study. *SSM-Population Health*, 2: 155-163. <https://doi.org/10.1016/j.ssmph.2016.02.007>
- [19] Sweileh, W.M., Zyoud, S.H., Al-Jabi, S.W., Sawalha, A.F. (2015). Contribution of Arab countries to breast cancer research: Comparison with non-Arab Middle Eastern countries. *BMC Women's Health*, 15: 1-7. <https://doi.org/10.1186/s12905-015-0184-3>
- [20] Tchier, F., Alharbi, A. (2016). Fuzzy relational model and genetic algorithms for early detection and diagnosis of breast cancer in Saudi Arabia. *Filomat*, 30(3): 547-556. <https://doi.org/10.2298/FIL1603547T>
- [21] Cakir A., Demirel, B. (2010). A software tool for determination of breast cancer treatment methods using data mining approach. *Journal of Medical Systems*, 35(6): 1503-1511. <https://doi.org/10.1007/s10916-009-9427-x>

- [22] Torrents-Barrena, J., Puig, D., Melendez, J., Valls, A. (2016). Computer-aided diagnosis of breast cancer via Gabor wavelet bank and binary-class SVM in mammographic images. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(1-2): 295-311. <https://doi.org/10.1080/0952813X.2015.1024491>
- [23] Shrivastavat, S.S., Sant, A., Aharwal, R.P. (2013). An overview on data mining approach on breast cancer data. *International Journal of Advanced Computer Research*, 3(4): 256-262.
- [24] Talukdar, J., Kalita, S.K. (2015). Detection of breast cancer using data mining tool (weka). *International Journal of Scientific & Engineering Research*, 6(11).
- [25] Chaurasia, D.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).
- [26] Chaurasia, D.V., Pal, S. (2014). Data mining techniques: To predict and resolve breast cancer survivability. *International Journal of Computer Science and Mobile Computing*, 3(1): 10-22.
- [27] Chaurasia, V., Pal, S., Tiwari, B.B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology*, 12(2): 119-126. <https://doi.org/10.1177/1748301818756225>
- [28] Aloraini, A. (2012). Different machine learning algorithms for breast cancer diagnosis. *International Journal of Artificial Intelligence & Applications*, 3(6): 21-30.
- [29] Matsen, C.B., Lyons, S., Goodman, M.S., Biesecker, B.B., Kaphingst, K.A. (2019). Decision role preferences for return of results from genome sequencing amongst young breast cancer patients. *Patient Education and Counseling*, 102(1): 155-161. <https://doi.org/10.1016/j.pec.2018.08.004>
- [30] Jonsdottir, T., Hvanngberg, E.T., Sigurdsson, H., Sigurdsson, S. (2008). The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining. *Expert Systems with Applications*, 34(1): 108-118. <https://doi.org/10.1016/j.eswa.2006.08.029>
- [31] Aličković, E., Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Computing and Applications*, 28: 753-763. <https://doi.org/10.1007/s00521-015-2103-9>
- [32] Jan, F., Rahman, A., Busaleh, R., et al. (2023). Assessing acetabular index angle in infants: A deep learning-based novel approach. *Journal of Imaging*, 9(11): 242. <https://doi.org/10.3390/jimaging9110242>
- [33] Ahmed, M.I.B., Saraireh, L., Rahman, A., et al. (2023). Personal protective equipment detection: A deep-learning-based sustainable approach. *Sustainability*, 15(18): 13990. <https://doi.org/10.3390/su151813990>
- [34] Ahmed, M.I.B., Alabdulkarem, H., Alomair, F., Aldossary, D., Alahmari, M., Alhumaidan, M., Alrassan, S., Rahman, A., Youldash, M., Zaman, G. (2023). A deep-learning approach to driver drowsiness detection. *Safety*, 9(3): 65. <https://doi.org/10.3390/safety9030065>
- [35] Ahmed, M.S., Rahman, A., AlGhamdi, F., AlDakheel, S., Hakami, H., AlJumah, A., AlIbrahim, Z., Youldash, M., Khan, M.A.A., Basheer Ahmed, M.I. (2023). Joint diagnosis of pneumonia, COVID-19, and tuberculosis from chest X-ray images: A deep learning approach. *Diagnostics*, 13(15): 2562. <https://doi.org/10.3390/diagnostics13152562>
- [36] Ahmed, M.I.B., Alotaibi, R.B., Al-Qahtani, R.A., Al-Qahtani, R.S., Al-Hetela, S.S., Al-Matar, K.A., Al-Saqer, N.K., Rahman, A., Saraireh, L., Youldash, M., Krishnasamy, G. (2023). Deep learning approach to recyclable products classification: Towards sustainable waste management. *Sustainability*, 15(14): 11138. <https://doi.org/10.3390/su151411138>
- [37] Farooqui, M., Rahman, A.U., Alorefan, R., Alqusser, M., Alzaid, L., Alnajim, S., Althobaiti, A., Ahmed, M.S. (2023). Food classification using deep learning: Presenting a new food segmentation dataset. *Mathematical Modelling of Engineering Problems*, 10(3): 1017-1024. <https://doi.org/10.18280/mmep.100336>
- [38] Ibrahim, N.M., Gabr, D.G., Rahman, A., Musleh, D., AlKhulaifi, D., AlKharraa, M. (2023). Transfer learning approach to seed taxonomy: A wild plant case study. *Big Data and Cognitive Computing*, 7(3): 128. <https://doi.org/10.3390/bdcc7030128>
- [39] Gollapalli, M., Rahman, A., Kudos, S.A., Foula, M.S., Alkhalifa, A.M., Albisher, H.M., Al-Hariri, M.T., Mohammad, N. (2024). Appendicitis diagnosis: Ensemble machine learning and explainable artificial intelligence-based comprehensive approach. *Big Data and Cognitive Computing*, 8(9): 108. <https://doi.org/10.3390/bdcc8090108>
- [40] Rahman, A., Youldash, M., Alshammari, G., et al. (2024). Diabetic retinopathy detection: A hybrid intelligent approach. *Computers, Materials & Continua*, 80(3).
- [41] Singh, K.R., Dash, S. (2022). Detection of arrhythmia from ECG Signal Using bat algorithm-based deep neural network. In *International Conference on Advanced Computing and Intelligent Engineering*, pp. 83-95.
- [42] Alabbad, D.A., Ajibi, S.Y., Alotaibi, R.B., Alsqer, N.K., Alqahtani, R.A., Felemban, N.M., Rahman, A., Aljameel, S.S., Ahmed, M.I.B., Youldash, M.M. (2024). Birthweight range prediction and classification: A machine learning-based sustainable approach. *Machine Learning and Knowledge Extraction*, 6(2): 770-788. <https://doi.org/10.3390/make6020036>
- [43] Mukhtar, M., Yunus, F., Li, J., Mahmood, T., Ali, Y. A. A. (2023). Future prospects and challenges of on-demand mobility management solutions. *IEEE Access*, 11: 114864-114879. <https://doi.org/10.1109/ACCESS.2023.3324297>
- [44] Gollapalli, M., Rahman, A.U., Youldash, M., et al. (2023). Machine learning approach to users' age prediction: A telecom company case study in Saudi Arabia. *Mathematical Modelling of Engineering Problems*, 10(5): 1619. <https://doi.org/10.18280/mmep.100512>
- [45] Sajid, N.A., Rahman, A., Ahmad, M., Musleh, D., Basheer Ahmed, M.I., Alassaf, R., Chabani, S., Ahmed, M.S., Salam, A.A., AlKhulaifi, D. (2023). Single vs. multi-label: The issues, challenges and insights of contemporary classification schemes. *Applied Sciences*, 13(11): 6804. <https://doi.org/10.3390/app13116804>
- [46] Gollapalli, M., Rahman, A., Alkharraa, M., et al. (2023). SUNFIT: A machine learning-based sustainable university field training framework for higher education. *Sustainability*, 15(10): 8057.

- <https://doi.org/10.3390/su15108057>
- [47] Khan, T.A., Fatima, A., Shahzad, T., Alissa, K., Ghazal, T.M., Al-Sakhnini, M.M., Abbas, S., Khan, M.A., Ahmed, A. (2023). Secure IoMT for disease prediction empowered with transfer learning in healthcare 5.0, the concept and case study. *IEEE Access*, 11: 39418-39430. <https://doi.org/10.1109/ACCESS.2023.3266156>
- [48] Ahmed, M.I.B., Zaghoud, R.A., Ahmed, M.S., Ahmed, M.S., Alrabeea, M., Alsuwaiti, A., Alzaid, N., Alyousef, A., Khan, M.A.A., Rahman, A., Chabani, S., Krishnasamy, G., Alturkey, A. (2023). Intelligent directional survey data analysis to improve directional data acquisition. *Mathematical Modelling of Engineering Problems*, 10(2): 482-490. <https://doi.org/10.18280/mmep.100214>
- [49] Alghamdi, A.S., Rahman, A. (2023). Data mining approach to predict success of secondary school students: A Saudi Arabian case study. *Education Sciences*, 13(3): 293. <https://doi.org/10.3390/educsci13030293>
- [50] Ahmed, M.I.B., Zaghoud, R.A., Al-Abdulqader, M., Kurdi, M., Altamimi, R., Alshammari, A., Noaman, A., Ahmed, M.S., Alshamrani, R., Alkharraa, M., Rahman, A., Krishnasamy, G. (2023). Ensemble machine learning based identification of adult epilepsy. *Mathematical Modelling of Engineering Problems*, 10(1): 94-92. <https://doi.org/10.18280/mmep.100110>
- [51] Hantom, W.H., Rahman, A. (2024). Arabic spam tweets classification: A comprehensive machine learning approach. *AI*, 5(3): 1049-1065. <https://doi.org/10.3390/ai5030052>
- [52] ur Rahman, A. (2022). Geo-spatial disease clustering for public health decision making. *Informatica*, 46(6). <https://doi.org/10.31449/inf.v46i6.3827>
- [53] Dash, S., Luhach, A.K., Chilamkurti, N., Baek, S., Nam, Y. (2019). A Neuro-fuzzy approach for user behaviour classification and prediction. *Journal of Cloud Computing*, 8(1): 1-15. <https://doi.org/10.1186/s13677-019-0144-9>
- [54] Rahman, A.U., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., Khan, M.A., Mosavi, A. (2022). Rainfall prediction system using machine learning fusion for smart cities. *Sensors*, 22(9): 3504. <https://doi.org/10.3390/s22093504>
- [55] Rahman, A.U., Alqahtani, A., Aldhafferi, N., Nasir, M.U., Khan, M.F., Khan, M.A., Mosavi, A. (2022). Histopathologic oral cancer prediction using oral squamous cell carcinoma biopsy empowered with transfer learning. *Sensors*, 22(10): 3833. <https://doi.org/10.3390/s22103833>
- [56] Ghazal, T.M., Al Hamadi, H., Umar Nasir, M., Gollapalli, M., Zubair, M., Adnan Khan, M., Yeob Yeun, C. (2022). Supervised machine learning empowered multifactorial genetic inheritance disorder prediction. *Computational Intelligence and Neuroscience*, 2022(1): 1051388. <https://doi.org/10.1155/2022/1051388>

## NOMENCLATURE

BC	Breast cancer
BCD	Breast cancer disease
DT	Decision trees
SVM	Support vector machine
SMO	Sequential minimal optimization
RBF	Radial basis function
AI	Artificial Intelligence
NB	Naïve Bayes
ML	Machine Learning