SEBHA UNIVERSITY JOURNAL OF PURE & APPLIED SCIENCES VOL.21 NO. 4 2022
DOI: 10.51984/JOPAS.V21I4.2126

مجلة جامعة سبها للعلوم البحتة والتطبيقية
Sebha University Journal of Pure & Applied Sciences

Journal homepage: www.sebhau.edu.ly/journal/index.php/jopas

An Exploratory Data Analysis of Breast Cancer Features in South of Libya

*Asma Agaal , Mansour Essgaer

Artificial Intelligence Department, Faculty of Information Technology, Sebha University, Sebha, Libya,

Keywords:
breast cancer
biological markers
tumor markers
routine blood test
Exploratory data analysis
EDA
Sebha oncology center
Libya

ABSTRACT

Exploratory data analysis is a data visualization approach used to extract knowledge from raw data. This approach can be applied to medical data to improve healthcare providers services. In recent years, breast cancer has become more common in women and requires effective procedures to detect it in the early stage. In this context, breast cancer patients' data were collected from the Sebha oncology center through their routine blood tests. The exploratory data analysis technique is used in this study to better analyze the patients' markers. The analysis aims to discover prominent bio and tumor markers that can assist in determining whether a tumor is benign or malignant. Several statistical and visualizations methods are used. The results show that the most effective markers that may be used as cancer predictors are: Cancer Antigen-15.3, Carcinoma Embryonic Antigen, White Blood Cells, Blood platelets, and Albumin. These findings are consistent with the findings of Sebha oncology center specialists. which may eventually aid in their cancer diagnosis.

تحقيق استكشافي لعلامات مرضي سرطان الثدي في جنوب ليبيا

*اسمه اعجال و منصور الصغير

قسم الذكاء الاصطناعي، كلية تقنية المعلومات ، جامعة سبها، سبها ، ليبيا

الكلمات المفتاحية:
مرحلة تحليل البيانات الاستكشافية
مركز علاج الأورام سبها
سرطان الثدي
علامات الورم/ الاورام
علامات بيولوجية
فحص الدم الروتيني
ليبيا

الملخص

تحليل البيانات الاستكشافية هو نهج تصور البيانات المستخدمة لاستنباط المعرفة المدفونة داخل مجموعة البيانات. يمكن تطبيق هذه التقنية على معالجة البيانات الطبية لتحسين خدمات مقدمي الرعاية الصحية. أصبح سرطان الثدي أكثر شيوعًا عند النساء ويتطلب إجراءات فعالة للكشف عن احتمالية حدوثه في مراحله المبكرة من أجل العلاج والشفاء. في هذا السياق، تم جمع بيانات من مركز سبها للأورام من خلال اختبارات الدم الروتينية واستخدمت تقنية التصور مع مجموعات البيانات الحالية للتعلم والتنبؤ. يحاول التحليل في هذه الدراسة اكتشاف العلامات الحيوية ودلالات الأورام البارزة التي يمكن أن تساعد في تحديد ما إذا كان الورم حميدًا أم خبيثًا، من خلال التصورات الإحصائية والاستراتيجيات الرسومية المختلفة المستخدمة لتحقيق هذا الهدف. من المثير للاهتمام نلاحظ أن العلامات البيولوجية الأكثر فاعلية التي يمكن استخدامها للتنبؤ بالسرطان هي: مستضد السرطان، ومستضد السرطان الجنيني، وخلايا الدم البيضاء، والصفائح الدموية، والألبومين. تتوافق هذه النتائج مع استنتاجات خبراء مركز سبها للأورام والتي يمكن أن تساعدهم في النهاية على تحسين تشخيص السرطان لديهم.

Introduction

The practice of analyzing raw data to interpolate meaningful information is known as Exploratory Data Analysis (EDA), which shows data in a visual format, allowing for better comprehension and informed decision making by organizations [1]. ÉDA is a technology that is often used to interpret data in many fields including education, industry, and medicine [2] [3] [4] [5].

In the health sector, the clinical and biological nature of the disease is so closely linked. Therefore, identifying the disease's prognostic factors by EDA is a crucial step, providing information and knowledge by EDA help in decision making, particularly for patients with breast cancer (BC) who are in an early stage [6].

Corresponding author:
E-mail addresses: asma.agaal@sebhau.edu.ly ,(M. Essgaer) man.essgaer@sebhau.edu.ly

Article History : Received 01 May 2022 - Received in revised form 11 June 2022 - Accepted 03 October 2022

Recently, the importance of using EDA in the medical domain has grown significantly, due to how successful this method is in providing valuable knowledge. Furthermore, this reduces the cost of pharmaceuticals and enhances the possibility of conducting additional clinical research [7] [8]. Using the EDA can assist in determining whether cancer is present in the patient. In addition, finding the important features required for constructing a predictive model that could enable early diagnosis in cancer patients [9]

BC is often diagnosed by biopsy, which serves as a prognostic indicator to determine if the tumor is benign or malignant in the advanced phase. However, these tests take a long time, and the patient may die while waiting. Whereas, there are other tests used in the early stages using routine blood tests relying on Biological Markers (BM) such as white blood cells, red blood cells, and Tumor Markers (TM) such as carcinoma embryonic antigen [10, 11]. These tests are less expensive than a biopsy. However, understanding the impact of BM and TM on BC prognosis is still scarce to the researcher's knowledge.

Therefore, the aim of this study is to look into the most crucial BM and TM, relevant to the diagnosis of BC in the southern region of Libya, by using the EDA visualization technique.

The remainder of this paper is arranged as follows: Section (2) concerns with literature reviews, section (3) is about the methods and techniques, section (4) is about the results and discussions, Lastly, section (6) the conclusion of the study.

Literature Reviews

BC is a major concern to women, requiring the creation of efficient early detection and treatment strategies. Therefore, many researchers uncover the main traits to detect whether a tumor is benign or malignant. In this regard, a study [12] was conducted to identify the key features of tumors such as Texture, Area, Radius, Fractal Dimension, area-se, concave-points-mean, and Perimeter. Whereby, EDA techniques were employed. The most prominent result of this study was to distinguish between benign and malignant cases, using univariate analysis. In the same way, the bivariate analysis histogram revealed that the mean, with standard deviation, provides the best separations for each class. Furthermore, the correlation plot showed the positive highest correlation between area-se and concave-points-mean features.

Similarly, univariate analysis methods such as swarm plots and pivot tables were used in a study [6] which was done on the Meteoric-BC data set to analyze features such as age at diagnosis, Nottingham Prognostic Index (NPI) score, and survival status. It concluded that the fifth-degree patient has a chance of a life of five or ten years.

On the other hand, in some research, pathological parameters such as: tumor size, tumor grade, number of positive lymph nodes, and hormone receptors, among others have been used to predict BC survivable. This study [7], in contrast, examines the importance of non-clinical prognostic factors such as age, ethnicity, and marital status in determining the prognosis for BC patients and the results indicate that these factors collected from the Surveillance, Epidemiology (SEER) have an impact on the survival rates of BC patients and that Through the use of EDA, survival analysis tools such as Cox proportional hazards and Kaplan-Meier survival curve.

Likewise, one study [13] looked at the effect of marital status on individuals with inflammatory breast cancer (IBC). Overall survival (OS) and breast cancer-specific survival (CSS) rates were compared in a sample of married and unmarried individuals using the SEER database. The comparison was carried out using the Kaplan–Meier technique with a T-test, and multivariate survival analysis of CSS and OS. It has been established that marital status is an independent prognostic factor in IBC patients. It also found that married patients have higher CSS and OS rates than unmarried individuals.

Furthermore, using univariate, bivariate, and multivariate analysis, a study [14] identifies features that aid early screening among women in underdeveloped countries like Kenya. It was discovered that women's geographic location, particularly those living in rural areas, was linked to a reduced likelihood of BC screening. Less educated, poor, and uninsured women are also less likely to get screened for BC than more educated, wealthier, and insured women.

Other sorts of biomarkers and their link with BC, on the other hand, were not discussed in depth by the researchers. Furthermore, the previous studies were limited by small sample sizes. In addition, aspects of the research dataset differed between the proposed study and multiple earlier investigations. As a result, this disparity may pose a challenge in testing the effectiveness of these predictive indicators BM and TM to predict BC using EDA in this investigation.

Material and method

EDA is primarily a strategy to see what the data can express, away from the formal modeling or hypothesis testing task, and assists in the analysis of data sets in order to describe their statistical properties, which include measures of central tendency (the mean, mode, and median), measures of spread (standard deviation and variance), and the shape of the distribution, and the presence of outliers Following the data collecting phase [15] [16]. Fig. 1 represents the essential steps in the exploratory data analysis phase, which are explored in greater detail below:

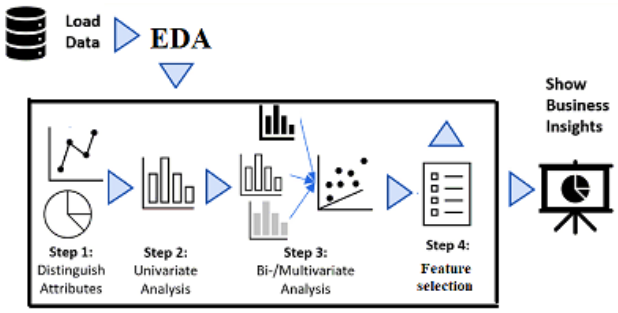


Fig. 1. EDA steps

1. Data collection and description

The BC dataset in this work comes from the Sebha Oncology Center (SOC) in the southern region of Libya. The SOC is a primary hub for all cancer cases in the southern regions, providing follow-up, treatment, and archive services. The dataset contains 2,435 records with 23 features reported between 2015 and 2020 from various southern cities. In this study, only blood test data is analyzed not genomic data. It was obtained manually, and then the raw data set was prepared as a CSV file. TABLE 1 lists the dataset's features, which include 22 features in addition to the target or CLASS which has the values Benign or Malignant.

TABLE 1: dataset features

Feature	Feature description
Sex	The patient's gender
Age	The patient's age
Address	The patient's address
FBS	Blood glucose
Urea	Kidney function test urea
Creatinine	Kidney function test creatinine
ALB	Albumin
T-Ca	Total calcium in the blood
GPT	Liver functions ' gpt'
GOT	Liver functions ' got '
ALP	Alkaline Phosphate
HGB	Hemoglobin
PLT	Blood platelets
ESR	Deposition of blood
LDH	Lactate Dehydrogenizes
Na+	Sodium
K	Potassium
CL-	Chloride acid
CA-15.3	Cancer antigen
CEA	Carcinoma embryonic antigen
WBC	White blood cells
RBC	Red blood cells
CLASS	Benign =0 or Malignant =1

Following data collecting, it is necessary to understand the data and transform it into a useful format. The exploratory step is sufficient to do this.

2. Exploratory data analysis (EDA)

In order to determine the most significant BM and TM for use as prognostic markers in BC patients. In this work, EDA is employed as an analysis and data visualization tool. There are two approaches to EDA: graphical and non-graphical analysis [17] [18], described as follows:

2.1 Non-graphical analysis:

a. 5 number summaries

For each continuous feature, the statistical description is used to provide various summary statistics such as the feature's number of records, maximum, minimum values, mean and dispersion of data from the center, as well as its percentage of the overall data set as 25%, 50%, and 75% [19].

b. The Independent T-test

In general, it can be used to: Estimate the difference between two or more categories to determine whether the input feature has a statistically significant association with the output feature (target). Where the statistical tests are based on the null hypothesis, which claims that there is no statistical significance between the data sets, the p-value will be bigger than the population parameter. Otherwise, we state the alternative hypothesis we wish to demonstrate [20].

2.2 Graphical analysis:

a. Univariate analysis tools

It provides a statistical summary for each field in the raw data set or a summary for a single feature. The count plot, pie chart plot, and histogram are examples of this analysis [21]. Several of these *tools* are covered in further detail below:

- ***Count plot*:** literally counts the number of observations per class for a categorical feature and shows the results as a bar chart.
- ***Pie chart plot*:** depicts how a total quantity is distributed between levels of a categorical feature. Each categorical value corresponds to a single slice of the circle, and the size of each slice represents how much of the total each category level takes up [22].
- ***Histogram*:** uses to represent the distribution of numerical (continuous) data and divides the entire value range into a series of intervals, where a frequency distribution is represented by a rectangle with a width representing the class interval [23].

b. *Bivariate analysis tool*s

uses to explore the relationships between each feature in the dataset and the target feature of interest, or two features, and look for connections between them [16]. Some of these tools are covered in further detail below:

- ***Heat map:*** It's a statistical metric that expresses the strength of the relationship and association between two features. Positive and negative correlations are the two basic forms of correlation, and a case where the correlation is zero indicates that no relationship between these two features [24]. It is used to evaluate hypotheses about causing and effect correlations between features.
- ***Scatter plot:*** Its data points are distributed horizontally and vertically to demonstrate how one feature influences the other [25].
- ***Side-by-Side Box plots*:** It is a visual representation of numerical data in the form of quadrants. Often used to detect outliers in a data set. The 25th, 50th, and 75th percentiles are used to summarize sample data in the boxplot. These percentiles are also referred to as the lower, middle, and higher quartiles [26].
- ***Swarm plot:*** It's a form of scatter plot that can represent both continuous and categorical values. All visualized samples are displayed side by side on a single graph, which clarifies the points of separation between samples by avoiding overlapping spots [27]**.**
- ***Mapping:*** A Folium library makes it simple to observe and connect data as geographic locations on an interactive map, at the same shows the spread of a disease on the map [28].

c. Multivariate analysis tools

Uses to discover relationships between more than two features. Examples of these types include pivot tables and bar plots [29]. Some of these are discussed in greater depth below:

- ***Pivot Table*:** It is a data summary in the shape of a table. That may include sums, averages, or other statistics that the pivot table organizes in a usable manner, with any filters and sort orders applied to the data [30].
- ***Bar plot*:** The bar chart depicts comparisons between discrete groups. One axis of the graph shows the specific categories being compared, while the other axis represents the corresponding measured values for those categories [31].

Results and discussions

This section discusses the results of the exploratory phase data analysis, which conducted a series of tests in the Python environment be summarized as follows:

1. Class and Sex distribution

The class label distribution is explored. It's worth noting that the numbers are roughly balanced, with a malignant 44% to a benign ratio of 55%. As shown in Fig**.**1.

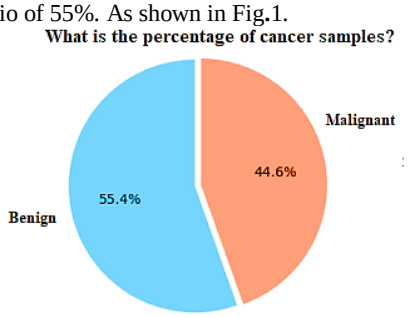


Fig. 1. The class label distribution

The data set, on the other hand, has a sex bias of 98.9 % female to 11% male, as seen in Fig**.**2.

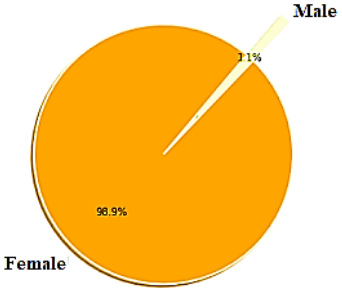


Fig. 2. The gender ratio in the data set

2. 5-number summaries

After looking at Fig**.**3, the statistical description assisted in developing an overview of the data set's features, which is described as follows:

- The data set's average age is 49 years, with the youngest being 23 and the oldest being 90.
- In the data set, 75% of the patients have diabetes as shown from “FBS” feature, while 25% do not. As a result, this feature may have an effect on CLASS.
- 75% of the patients in the data set did not suffer from the kidney as shown from “Urea”, “Creatinine” features or liver disease as shown from “GPT”, “GOT” features, because the rates of these features are in the normal range.
- The tumor markers features such as “CA15”, and “CEA” were in the normal position in 50% of the data set, but were very high in 75%, indicating that these features affect the breast cancer diagnosis.
- Each of the following features “RBC”, as well as “HGB” was in the normal range at 75%. In contrast, the sedimentation level feature was high in 75% of the patients, but it had a normal rate in 25% of the data. When compared to “RBC” and the rate of “HGB” in the blood, the latter has a stronger effect on CLASS.
- The level of salts-related features, such as “CL”, and “Na”’, is lower than the usual rate in 50% of the patients in the data set. The “K” feature, on the other hand, was in the usual range at 75%. We can see from the above that the features “K”, “Na”, and “CL” have no direct effect on CLASS.
- We also observe that majority of the features have extreme values due to the wide gap between 75 % and the max value in each feature, as well as the fact that the mean value of each feature is bigger than the median represented by 50 %. Furthermore, because some features, such as “Na”, “K”, and “CL”, have a significant standard deviation, it may be expected that if these features include big missing values, it is best to delete them.

Fig. 3. The statistical description of all features

	count	mean	std	min	25%	50%	75%	max
Age	2437.0	49.4	11.3	23.0	41.0	48.0	55.0	90.0
FBS	2048.0	160.7	84.0	11.0	99.0	145.0	200.0	764.0
Urea	2275.0	24.6	20.2	0.5	14.0	21.0	29.0	191.0
Creatinin	2282.0	1.1	1.3	0.1	0.7	0.8	1.0	39.0
ALB	2148.0	4.5	2.5	0.2	3.3	3.8	4.5	38.0
T_Ca	2142.0	8.0	5.5	0.4	7.9	8.5	9.3	99.6
GPT	2140.0	15.6	15.3	1.0	7.0	13.0	19.0	179.0
GOT	2202.0	17.2	14.6	1.0	9.0	15.0	22.0	164.0
ALP	2126.0	140.7	88.1	1.0	95.0	145.0	187.0	867.0
CA15	2435.0	31.0	20.3	0.3	12.8	23.1	50.0	100.0
CEA	2435.0	4.5	3.4	0.0	1.5	3.1	8.1	20.8
WBC	2333.0	7.3	4.6	0.3	4.4	6.3	8.9	88.3
RBC	2369.0	6.5	17.2	0.7	4.1	4.4	5.2	492.0
HGB	2337.0	11.0	6.8	0.3	11.0	12.1	13.0	224.0
PLT	2258.0	283.6	140.8	1.0	197.0	282.0	401.0	947.0
ESR	2049.0	25.4	25.7	1.0	10.0	19.0	34.0	333.0
LDH	2052.0	185.2	100.0	1.0	138.0	185.0	254.0	934.0
Na	2011.0	130.0	48.5	4.1	136.5	139.2	143.0	1377.0
K	2039.0	5.6	22.8	0.4	3.8	4.2	4.5	423.0
CL	1775.0	42.1	42.6	1.0	16.3	18.3	101.3	184.0

3. Determine the features that have an impact on CLASS

In order to identify the features that affect the target feature CLASS, a series of experiments is conducted, which will be described in more depth below:

a. The effect of diabetes on BC

To assess the relationship between the CLASS and the diabetes-related “FBS” feature such as glucose, a barplot figure is used to show this relationship. As shown in Fig.4, found that the number of malignant samples infected with diabetes is more than the number of non-diabetic samples. As a result, this feature has an effect on the target feature.

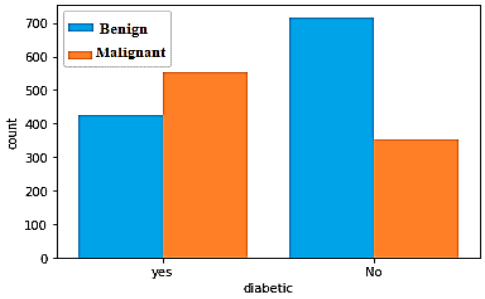


Fig. 4. Glucose levels in benign and malignant samples

b. The effect of kidney disease on BC

Furthermore, Fig.5 demonstrates that the majority of the malignant and benign samples do not have kidney disfunction. On the other hand, the number of malignant samples infected with kidney illness was fewer than the number of those without kidney illness. Thus, the observed results support the initial hypothesis that kidney problems had no effect on the diagnosis of malignant samples.

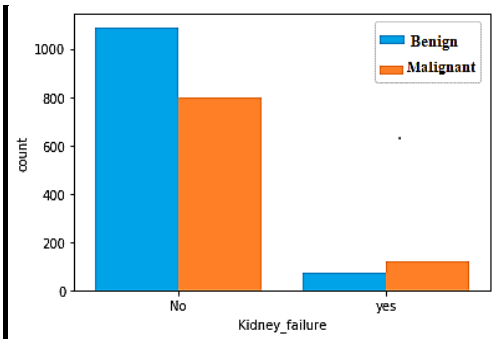


Fig. 5. Effect kidney disease in benign and malignant samples

c. The effect of hepatitis on BC

Similarly, Fig.6 also reveals that the majority of malignant and benign samples are free of liver disease. As a result, this finding supports the initial hypothesis that there is no strong link between liver illness and malignant tumor diagnosis.

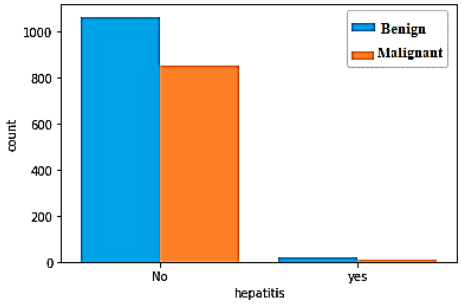


Fig 6. Effect hepatitis in benign and malignant samples

d. *The effect of tumor markers on BC*

Additionally, Fig.7 and Fig.8 confirm the hypotheses obtained from the 5-number summary of the features "CA15" and "CEA" and their influence on the CLASS.

- In the same context. The “CA15” features value in the benign class spans from 3 to 30, while it ranges from 36 to 80 in the malignant group, as illustrated in Fig.7. This suggests that the benign and malignant samples are clearly distinguished. Thus, it supports the initial hypothesis regarding the effect of “CA15” on the CLASS.

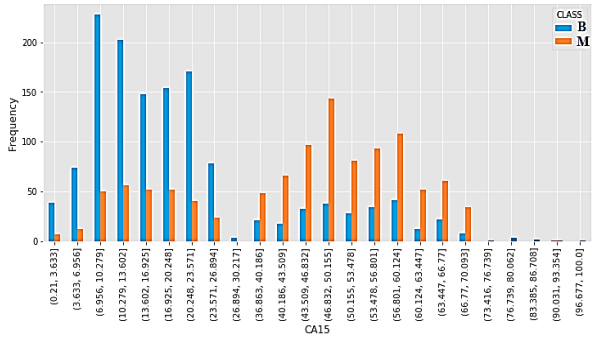


Fig.7. Effect CA15 in benign and malignant samples

- Likewise the "CEC" has an effect on the CLASS. Because there is a clear distinction between benign and malignant sample, illustrated in Fig.8, with values of the benign sample ranging from 0.7 to 4.1 and values of the malignant sample ranging from 5.5 to 11.

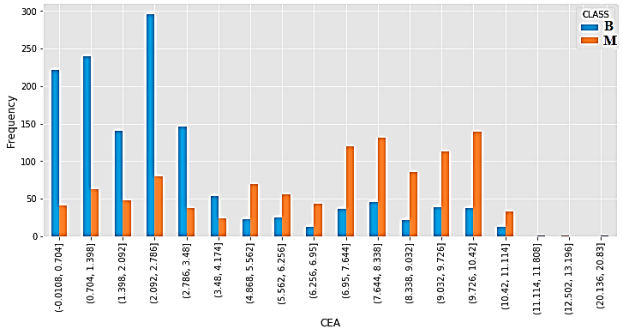


Fig. 8. Effect CA15 in benign and malignant samples

4. *Top 10 features by pearson correlation analysis*

Another experiment regarding the correlation with the target CLASS and between the features each other is conducted. Fig.9, shows the best 10 CLASS-correlated features in descending order, where the yellow colors represents a positive correlation and the blue color represent a negative correlation. Using Pearson correlation analysis, "CA15" was the best feature linked with CLASS, accounting for (52%), followed by "CEA" (51%), "WBC" (28%), "PLT" (21%), and "ALB" (20%) respectively. The "ALP" feature had the poorest relationship with CLASS, with a rate of (0.09%). This result of correlated features matches the SOC doctor's expectation. As a result, the features that are highly connected with the target will be chosen for further investigation and testing.

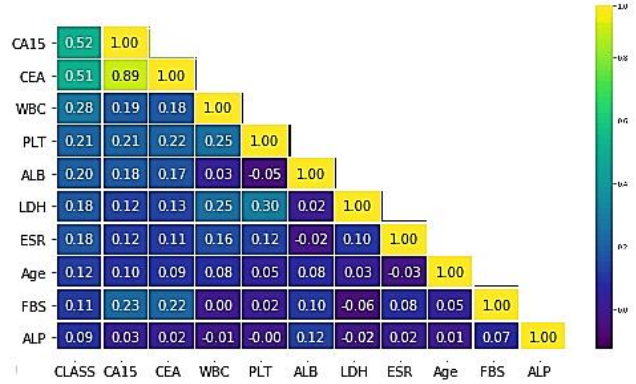


Fig. 9. The best 10 CLASS-correlated features in descending order

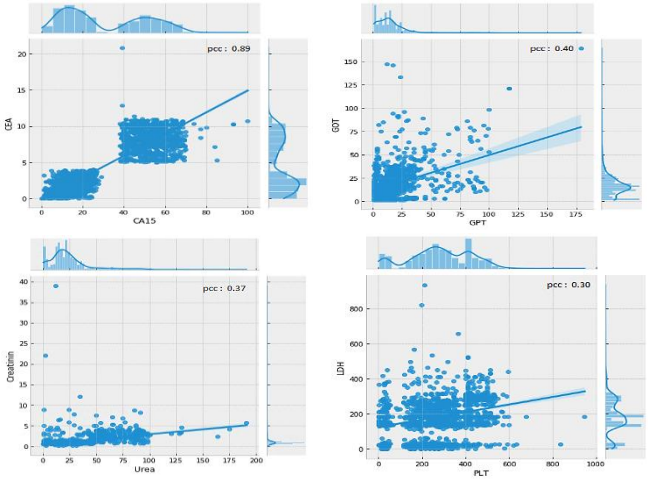
when we investigate the linear correlations between the features, we notice that the strongest liner relationships, according scatter plot, were between "CA15" and "CEA" at (89%), "GPT" with "GOT" at (40%), "Urea" with "Creatinine" at (37%), and "PLT" with "LDH" at (30%) as shown in Fig.10.

Fig. 10. A scatter plot that shows the strongest relationship between two features in the dataset

5. *The separation between benign and malignant samples*

We will examine the potential of the effect in greater depth using the box plot, swarm plot, and T-test. In this analysis, only some filtered features from Pearson correlation analysis were chosen, their influence on the CLASS will be inspected, and their separation of benign and malignant samples explained below:

a. Fig.11 shows that the "CA15", "CEA", "Age", "FBS", and" LDH", have a clear separation of both benign and malignant samples. While other features such as "UREA", "ALB", and "GOT" has no clear separation. Moreover, some features could not be examined against the CLASS due to outliers.



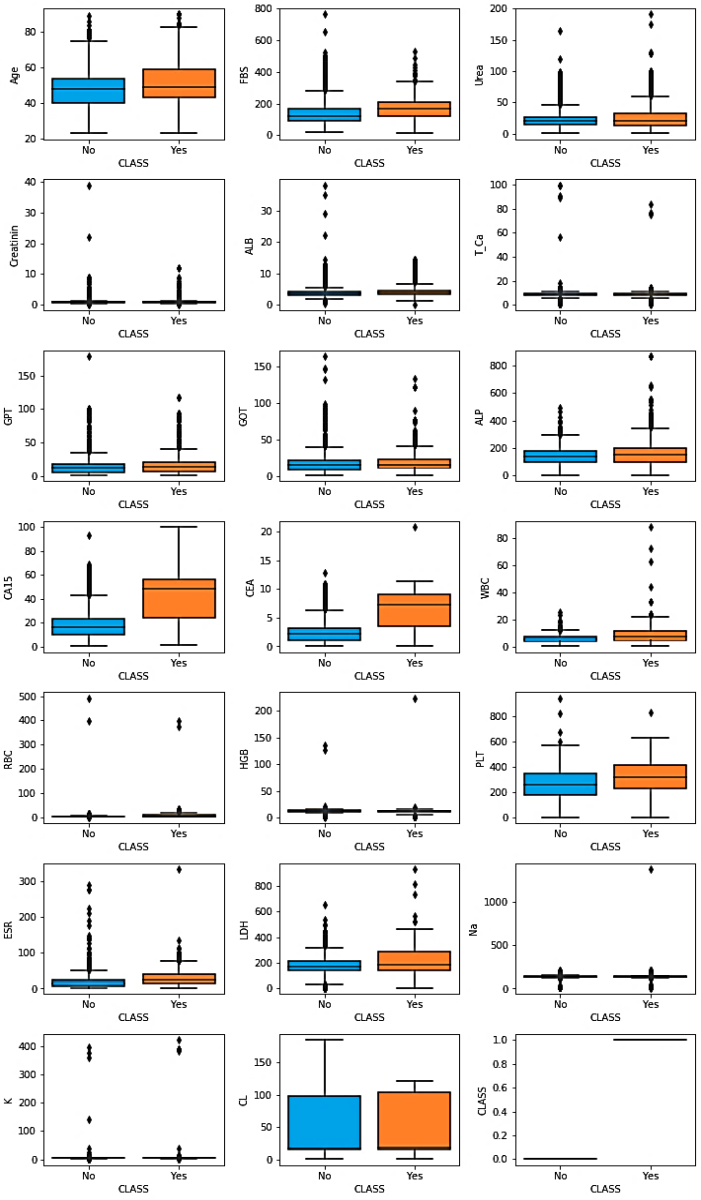


Fig. 11. A box plot showing the ability of features to separate

b. Moreover, a swarm plot to find the most relevant features is further used. from Fig.12 the separation of benign and malignant is clear in some features such as “CA15”, “CEA”, “WBC”, “ESR”, “LDH”, “ALB”, “RBC”. On the other hand, the separation is not clear in “CL”, “Na”, “K”, and Sex compared to the rest of the features.

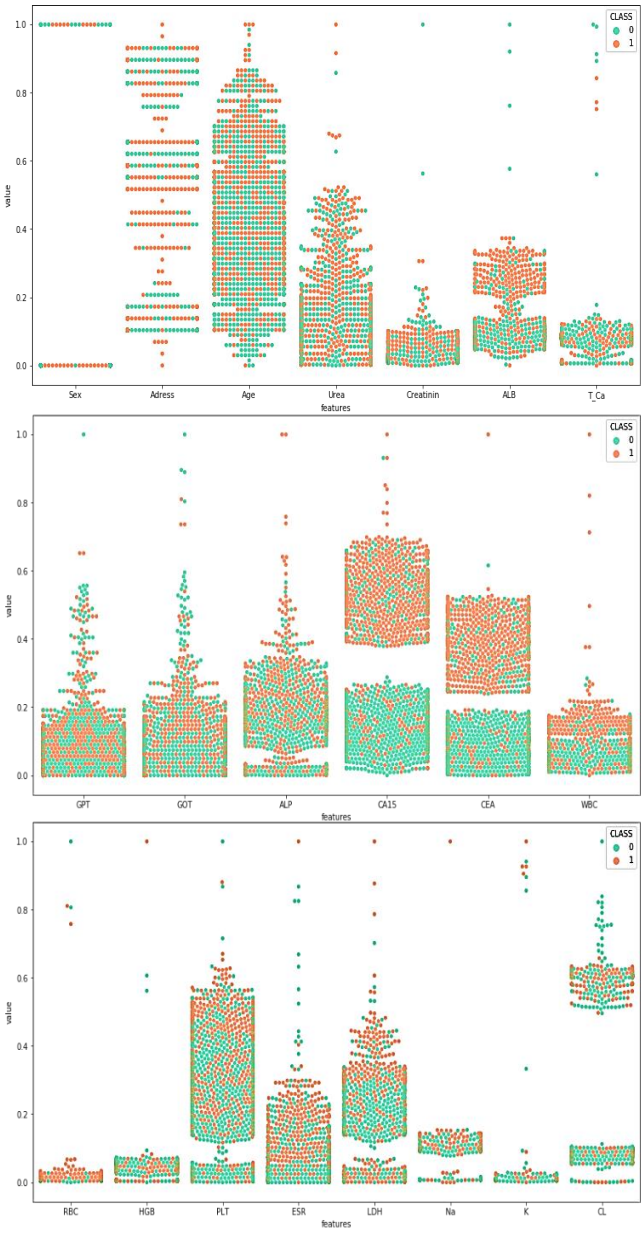


Fig. 12. A swarm plot showing the features ability to separate

c. Moreover, a T-test is conducted to show if there is a difference between the observed features. Since the p-value is greater than the critical level for the following features: “Sex”, “Address”, “Urea”, “Creatinine, “T-Ca”, “GPT”, “GOT”, “HGB”, “Na”, “K”, and “CL”. This means that there is no difference between the average of the benign and malignant categories, and so these findings confirm the null hypothesis, which states that there is no statistical significance of those features on the CLASS. Otherwise, the rest of the features exhibit statistical significance, indicating that they have an effect on CLASS.

TABLE 2: T-test values for the attributes of the dataset

Features	Statistic population	p-value
Sex	-1.02	0.30
Address	-5.56	2.85
Age	5.83	4.12
FBS	5.06	4.35
Urea	394	8.24
Creatinine	2.91	10.00
ALB	9.56	2.54
TCa	-0.85	0.39
GPT	2.36	3.01
GOT	1.26	4.20
ALP	4.37	1.25
CA15	29.66	2.21
CEA	29.55	2.63
WBC	13.87	3.26
RBC	3.86	0.00
HGB	-3.72	0.00
PLT	10.34	1.44
ESR	8.10	8.49
LDH	8.36	1.03
Na	2.09	3.03
k	0.60	1.54
CL	0.34	0.73

e. As described below, according to Fig.13, to learn more about the best predictive features that have a link with "CLASS", based on the above results:

- Males have higher "CA15" and "CEA" levels in malignant samples than females.
- Males who are older have a higher risk of acquiring breast cancer, both benign and malignant.
- Similarly, males have higher levels of "ESR" in benign samples than females, and females have higher levels of "ESR" in malignant samples than males.
- The "FBS" of malignant samples in males has no effect on the aim feature, whereas the "FBS" of malignant samples in females does.
- the level of "LDH" in males belonging to the benign samples is lower than the normal range. Furthermore, "PLT" is higher in females and males in malignant samples than in benign samples.

CLASS	Sex	ALB	Age	CA15	CEA	ESR	FBS	LDH	PLT	WBC
Benine	Femal	3.9	48.1	21.6	2.9	20.8	150.3	168.4	257.9	6.1
	meal	3.8	50.1	18.5	2.3	68.7	300.0	132.0	169.2	6.0
Mealigent	Femal	5.0	50.7	42.5	6.5	30.5	171.1	204.4	317.6	8.6
	meal	4.6	70.4	53.7	7.8	25.3	178.5	194.0	260.7	8.5

Fig. 13. The Pivot Table plot shows the average of the top 9 features relative to sex attribute and the CLASS

6. Distribution of patients on the Libya map

Fig.14 displays the distribution of BC on the Libyan map from 2015 to 2020 which was collected from SOC, with the Sebha region having the most patients with (1246) samples, followed by Al Shati (399) sample, Ubari (231) samples, and Marzouk (100) sample. Moreover, the number of patients ranges between one and seventeen from other Libya cities.

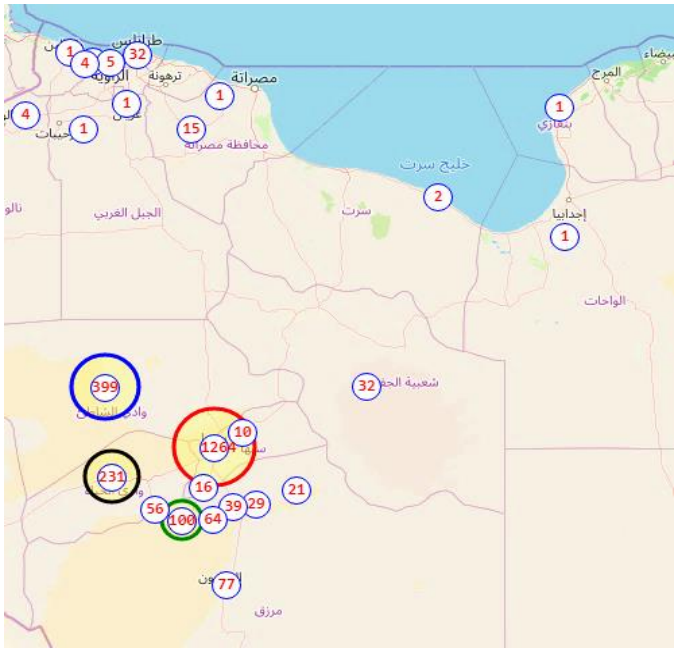


Fig. 14. determine the extent of the disease spread on the map of Libya, according to the pioneers of the SCO.

Conclusion and Recommendations

The study makes an interesting contribution to finding the most relevant predictive features of an early-stage breast cancer diagnosis. The exploratory stage was applied and used for analyzing routine blood data collected from SOC, such as biomarkers and tumor markers of 2435 BC patients. The results demonstrated that "CA-15.3", "CEA", "WBC", "PLT", "ALB", "LDH", "ESR", "Age", "FBS", and "ALP" are the 10 most important biomarkers in diagnosing BC, by using univariate, bivariate, and multivariate analysis. In contrast, the t-test showed that some features such as "sex", "Address", "Urea", "Creatinine", "TC-a", "GPT"," GOT"," HGB", "N"', "K", and "CL" have no effect on the diagnosis.

Many questions remain unanswered at present, so further investigations with more biomarkers and tumor markers, as well as the application of feature selection algorithms to identify BC disease-relevant markers and compare them with the results of this study, is highly recommended.

References

[1]- Velleman, P.F. and D.C. Hoaglin, *Exploratory data analysis.* 2012.

[2]- Sarker, I.H.J.S.C.S., *Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective.* 2021. **2**(5): p. 1-22.

[3]- Wongsuphasawat, K., Y. Liu, and J.J.a.p.a. Heer, *Goals, process, and challenges of exploratory data analysis: an interview study.* 2019.

[4]- Sarker, I.H.J.S.C.S., *Machine learning: Algorithms, real-world applications and research directions.* 2021. **2**(3): p. 1-21.

[5]- Eken, S.J.J.o.A.I. and H. Computing, *An exploratory teaching program in big data analysis for undergraduate students.* 2020. **11**(10): p. 4285-4304.

[6]- Sweetlin, E.J. and S. Saudia. *Exploratory Data Analysis on Breast cancer dataset about Survivability and Recurrence*. in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*. 2021. IEEE.

[7]- Kanaan, Y.M., et al., *Exploratory Data Analysis on Breast Cancer Prognosis*, in *Encyclopedia of Information Science and Technology, Fourth Edition*. 2018, IGI Global. p. 1794-1805.

[8]- Kumar, D., J.C.J.I.S. Bezdek, Man,, and C. Magazine, *Visual approaches for exploratory data analysis: A survey of the visual assessment of clustering tendency (vat) family of algorithms.* 2020. **6**(2): p. 10-48.

[9]- Morrow, A.K., et al., *Mango: Exploratory Data Analysis for*

Large-Scale Sequencing Datasets. 2019. **9**(6): p. 609-613. e3.
[10]- Di Gioia, D., et al., *Tumor markers in the early detection of tumor recurrence in breast cancer patients: CA 125, CYFRA 21-1, HER2 shed antigen, LDH and CRP in combination with CEA and CA 15-3.* 2016. **461**: p. 1-7.
[11]- Nicolini, A., P. Ferrari, and M.J. Duffy. *Prognostic and predictive biomarkers in breast cancer: past, present and future*. in *Seminars in cancer biology*. 2018. Elsevier.
[12]- Khan, S.A. and S.S. Velan. *Application of exploratory data analysis to generate inferences on the occurrence of breast cancer using a sample dataset*. in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*. 2020. IEEE.
[13]- Liu, Y.-l., et al., *Marital status is an independent prognostic factor in inflammatory breast cancer patients: an analysis of the surveillance, epidemiology, and end results database.* 2019. **178**(2): p. 379-388.
[14]- Antabe, R., et al., *Utilization of breast cancer screening in Kenya: what are the determinants?* 2020. **20**(1): p. 1-9.
[15]- Purohit, K.J.I.J.o.D.S. and Analysis, *Separation of Data Cleansing Concept from EDA.* 2021. **7**(3): p. 89.
[16]- Sahoo, K., et al., *Exploratory data analysis using Python.* 2019. **8**(12): p. 2019.
[17]- Komorowski, M., et al., *Exploratory data analysis.* 2016: p. 185-203.
[18]- Bisong, E., *Matplotlib and seaborn*, in *Building machine learning and deep learning models on google cloud platform*. 2019, Springer. p. 151-165.
[19]- Čižmešija, M., *Five-Number Summaries*, in *International Encyclopedia of Statistical Science*. 2011, Springer. p. 526-527.
[20]- Wang, Z., et al., *Preoperative prediction of axillary lymph node metastasis in breast cancer using CNN based on multiparametric MRI.* 2022.
[21]- Cox, V., *Exploratory data analysis*, in *Translating Statistics to Make Decisions*. 2017, Springer. p. 47-74.
[22]- Godbole, M. and A. Agarwal, *Efficacy Analysis of Technology Approaches Toward Auto-assignment of Clinical Codes to the US Patient Medical Record*, in *Advanced Computing Technologies and Applications*. 2020, Springer. p. 423-440.
[23]- Salem, N., H. Malik, and A.J.P.C.S. Shams, *Medical image enhancement based on histogram algorithms.* 2019. **163**: p. 300-311.
[24]- Guo, H., et al., *Heat map visualization for electrocardiogram data analysis.* 2020. **20**(1): p. 1-8.
[25]- Lamy, J.-B., et al., *Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach.* 2019. **94**: p. 42-53.
[26]- Menger, R., et al., *Health Care Lobbying, Political Action Committees, and Spine Surgery.* 2020. **45**(24): p. 1736-1742.
[27]- Faris, H., et al., *Medical speciality classification system based on binary particle swarms and ensemble of one vs. rest support vector machines.* 2020. **109**: p. 103525.
[28]- Jeanne, L., et al., *Economic globalization and the COVID-19 pandemic: global spread and inequalities.* 2022: p. 1-8.
[29]- Fathoni, M.I.A., F. Adi-Kusumo, and S.H. Hutajulu. *Survival analysis of breast cancer patients in Yogyakarta*. in *Journal of Physics: Conference Series*. 2021. IOP Publishing.
[30]- Henley, A. and D.J.L.i.C. Wolf, *Learn Data Analysis with Python*. 2018.
[31]- Embarak, D.O., Embarak, and Karkal, *Data analysis and visualization using python*. 2018: Springer.