https://doi.org/10.33472/AFJBS.6.6.2024.1021-1037

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Deciphering Breast Cancer Survival: An In-depth Analysis of Predictive Factors using Exploratory Data Analysis

Aditi Kajala[1], Sandeep Jaiswal[1] and Rajesh Kumar [2]

[1] School of Engineering and Technology, Mody University of Science & Technology, Lakshmangarh, India

[2] Electrical Engineering, Malaviya National Institute of Technology, Jaipur, India

Article Info

Volume 6, Issue 6, May 2024

Received: 28 March 2024

Accepted: 30 April 2024

doi: 10.33472/AFJBS.6.6.2024.1021-1037

Abstract:

Aim: The article aims to investigate the influence of vital prognostic variables on breast cancer patients' survival rates, utilizing Exploratory Data Analysis (EDA).

Background: Exploratory Data Analysis (EDA) plays a pivotal role in unraveling patterns and insights within complex datasets, particularly in cancer research. Its value lies in its ability to unveil hidden relationships and inform subsequent hypotheses. Thus, EDA can ultimately lead to better-personalized treatment plans and improved outcomes for patients.

Objective: This study delves into the impact of critical prognostic variables on a breast cancer patient's likelihood of survival. The examined factors include tumor size and stage, age at diagnosis, lymph node involvement, type of treatment undertaken, and the presence of progesterone and estrogen receptors. The analysis is conducted on the Metabric Breast Cancer dataset.

Method: EDA techniques, leveraging Python and employing visualizations such as box plots, scatter plots, and histograms, are utilized to draw conclusions and recognize patterns within the dataset. Subsequently, hypothesis formulation is undertaken to guide the investigation. To enhance the robustness of our findings, a questionnaire was administered to medical experts, seeking their insights and validation of the observed patterns. Statistical analyses, including p-values and chi-square tests, quantify the significance of the relationships identified during the EDA phase.

Result: The comprehensive EDA reveals nuanced associations between the selected variables and breast cancer survival. Expert validation provides additional credibility to the identified patterns.

Conclusion: The study reveals how essential factors interact to determine breast cancer survival rates, highlighting the significance of customized treatment plans for better results.

Keywords: EDA, Metabric Breast Cancer dataset, Breast Cancer, Survival, Hypothesis Formulation, Clinical Factor, Prognostic Variables

1. Introduction

Exploratory Data Analysis (EDA) is a crucial step in the initial stages of data analysis, enabling researchers to identify anomalies, recognize hidden relationships, and formulate hypotheses. Breast cancer, a dynamic and continually evolving field of study, necessitates constant efforts to improve patient outcomes and refine our comprehension of the disease[1]. Effective breast cancer treatment evaluation involves analyzing various options and considering factors such as survival rates, quality of life, and recurrence rates. Clinical trials, observational studies, meta-analyses, and exploratory data analysis all advance our knowledge and refine treatment approaches. By exploring existing data, it is possible to unearth hidden patterns, identify high-risk patient profiles, and evaluate the effectiveness of diverse treatment modalities[1], [2], [3]. The ultimate goal is to craft personalized treatment plans catering to individual patient needs, optimizing outcomes. This study focuses on crucial prognostic variables, including tumor size and stage, age at diagnosis, lymph nodes, type of treatment undertaken, and the presence of Progesterone and Estrogen receptors[4]. These variables collectively form a comprehensive framework for assessing the impact of breast cancer treatment on patient survival and well-being. In tandem with EDA, this study systematically employs hypothesis testing to evaluate conjectures derived from the observed patterns. The integration of statistical analyses, including p-values and chi-square tests, enhances the rigor and precision of our investigation, providing a robust foundation for drawing meaningful conclusions. By focusing on key prognostic variables, the study aims to provide actionable insights into the impact of these factors on patient survival. Further, to enhance the validity of the results, a survey was conducted among experts and medical professionals. Their input strengthens the reliability of the EDA outcomes and consequently concludes, aligning the conclusions with the collective expertise of the field.

The paper is organized as follows: Section 2 provides a comprehensive literature review, delving into earlier research. Section 3 outlines the formulated research questions. Details regarding the materials and procedures employed are presented in Section 4. Section 5 presents the results and subsequent discussion. Finally, Section 6 encapsulates the conclusion and outlines future directions for the study.

2. Literature Review

The incidence of breast cancer cases is on the rise globally, posing significant challenges to public health systems[5], [6]. Early detection through regular screenings and awareness campaigns remains crucial in combating the rising trend of breast cancer cases. Improved diagnostic techniques and advancements in treatment modalities have led to better outcomes for individuals diagnosed with breast cancer despite the increasing number of cases. Machine learning algorithms can analyze mammograms and other medical imaging scans to detect subtle patterns indicative of breast cancer at its early stages, often before they are noticeable to the human eye. These algorithms can assist radiologists in interpreting images more accurately and efficiently, reducing the risk of false negatives and missed diagnoses[7], [8], [9][10] Cancer research heavily relies on the prediction and prognosis of disease development using Machine Learning algorithms, aiming to improve subsequent treatment strategies and overall patient care. The prognosis hinges on cancer type, stage, grade, and individual health factors. The primary factors for predicting survival time can be tumor stage, dimension, and patient age. Integrating two datasets shows good performance[11], [12].[13] Detecting breast cancer early enhances survival rates and reduces mortality. Cancer ranks among the most lethal illnesses, resulting from atypical action in genes governing cell division and growth. Clinical decisions for cancer patients rely on a combination of clinical and genetic information. Bioinformatics tools and algorithms in this setting face substantial problems due to the combination of multimodality and variety.

Researchers and pathologists have employed ML algorithms, like k-nearest Neighbors (k-NN), Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), and Artificial Neural Network for cancer prediction. A scalable and robust pipeline model has been developed to analyze extensive cancer data, enabling real-time prediction of cancerous cells[11][14], [15], [16], [17], [18], [19] An important factor in predicting breast cancer patients' overall survival is the lymph node ratio (LNR), whose prognostic value is determined via estimate. To predict overall survival, LNR was estimated using Bayesian inference networks. The final survival model that included LNR performed better than the other models that were considered[19]. Furthermore, a gradient-boosting algorithm called EXSA was created to forecast how breast cancer would progress.

Additionally, it improved the survival analysis for ties inside the XGBoost framework. EXSA achieved competitively good results for 5-year and 10-year survival [20]. Explainable models help us understand how the model works and what factors influence its decisions. This can help build trust in the model and ensure that the model is not biased or discriminatory[21].[22] [23]. In some cases, the decisions made by machine learning models can have significant consequences. Explainable models can help us understand how these decisions were made and who is responsible for them. Some industries and regulatory bodies require that machine learning models be explainable to comply with regulations and guidelines.

In the healthcare and finance industries, explainable machine learning models are required. Explainable models can provide insights into the data and the underlying patterns that the model is using to make decisions[24] [25] [26] [27].In the landscape of survival prediction studies, a fundamental commonality emerges, uniting both conventional and more contemporary approaches. Regardless of the specific methodology employed, a prerequisite shared by these studies, including those leveraging machine learning and explainable AI, is the cultivation of a robust and comprehensive understanding of the underlying data. This foundational step is a crucial precursor, ensuring that subsequent analyses are anchored in a nuanced awareness of the dataset's intricacies. Within this framework, Exploratory Data Analysis (EDA) emerges as a potent tool, offering a multifaceted advantage. EDA facilitates an intricate comprehension of the data but also serves as a conduit for drawing preliminary insights[4] [28]. Particularly in the context of survival prediction, where the stakes are high, the judicious application of EDA enhances transparency and lays the groundwork for informed decision-making in subsequent stages of the analysis. Table 1 presents the list of current methods used by researchers for breast cancer prognosis.

Table 1: current machine learning methods used by researchers for Breast cancer Patient Survival prognosis

S.No.	Methon	Reference
1.	Bayesian Network	[29]
2.	Bayesian Network	[19]
3.	Random Forest, ANN, Logistic regression	[30]
4.	K-NN	[12]
5.	CoxPH	[19], [31], [32]
6.	Ensemble method	[33][20]
7.	XGBoost	[20]

2.1 EDA and Machine Learning

EDA lays the foundation for machine learning prediction by providing insights into the data's characteristics and guiding feature selection, variable transformation, and outlier detection. Machine learning prediction quantifies the predictive power of these variables and develops models for making accurate predictions[1], [2]. EDA is more descriptive and qualitative, focusing on data exploration and visualization, whereas machine learning prediction is more quantitative and predictive, aiming to develop models that optimize predictive performance. The visualization of machine learning models in the medical domain is crucial for gaining trust and acceptance from clinicians and patients.[27], [34]In the exploratory data analysis (EDA) of the Metabaric dataset, the initial focus is on investigating the relationships between various biomarkers and other clinical factors of breast cancer patients and their likelihood of survivability. Following this analysis, specific research questions are formulated based on the observed patterns. Hypotheses regarding the prognostic variables are then crafted. Instead of employing machine learning techniques to test these hypotheses, statistical tests are applied. Additionally, feedback from medical experts is gathered to provide insights and validation for the formulated hypotheses.Validating the results of exploratory data analysis (EDA) through consultation with medical experts offers several benefits. It ensures that the identified patterns and relationships align with medical knowledge and understanding of the disease. It also provides valuable insights into the clinical relevance and significance of the findings, helping prioritize features and variables for predictive modeling. Lastly, expert validation enhances the credibility and applicability of the analysis results, facilitating their integration into clinical practice for improved patient care and outcomes.

3. Research Questions

An EDA is carried out to look into the correlations between particular variables in light of the thorough knowledge of the variables present in the dataset. Taking into account the state of the field's research, questions are put forth to direct the analysis and provide insights into the dataset. The following questions were looked into:

RQ1: What impact does age have on a patient's chance of survival with breast cancer?

RQ2: How much of a patient's chance of survival is influenced by the tumor size at each stage?

RQ3: How do positively checked lymph nodes impact the survival rate?

RQ4: How are the patient's survival months post-diagnosis related to their progesterone receptor status and Estrogen receptor (ER) status?

RQ5: Which treatment modality—chemotherapy, hormone therapy, or radiation therapy is most usually utilized to increase a breast cancer patient's chances of survival?

RQ6: Are there any correlations between the numerical variables of the dataset?

4. Material and Methods

This section will include an overview of the dataset, its attributes, and data preprocessing and cleaning steps. Furthermore, the proposed methodology's workflow will be presented, explaining the sequential steps undertaken to draw valid conclusions from the dataset.

4.1. Dataset-Metabric Breast Cancer dataset:

The Metabric Breast Cancer dataset on Kaggle is a collection of gene expression data and clinical information for breast cancer patients. It includes information on patient age, tumor size, hormone receptor status, and survival outcomes. The dataset is often used for developing machine learning models to predict breast cancer prognosis. It is publically available at the web link:
https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric
In the study, 32 clinical attributes (1 target and 31 others) mentioned in Table 2 are considered with 1904 instances with 1103 of the “Survived” class and 801 instances of the “Dead” class.

Table 2: Description of different attributes of the Metabric Breast Cancer dataset

S.No.	Attribute Name	Description
1	patient_id	patient IDs are unique identifiers assigned to each patient in the dataset

2	age_at_diagnosis	represents the age of the patients when they were diagnosed with breast cancer.
3	type_of_breast_surgery	contains information about the specific type of breast surgery that each patient underwent, such as lumpectomy, mastectomy, or breast reconstruction.
4	cancer_type	is a categorical variable that represents the type of cancer which Breast Cancer for each sample in the dataset.
5	cancer_type_detailed	provides detailed information about the type of cancer, such as Breast Invasive Ductal Carcinoma
6	cellularity	refers to the measure of the proportion of a tissue sample that is composed of cells.
7	chemotherapy	indicates whether a patient received chemotherapy as part of their treatment for breast cancer.
8	pam50_+_claudin-low_subtype	refers to the molecular subtype classification of breast cancer samples. The PAM50 gene signature classifies breast cancer into different molecular subtypes based on the expression of 50 genes. The claudin-low subtype is a specific molecular subtype of breast cancer that is characterized by low expression of cell-cell adhesion genes and high expression of immune response genes.
9	cohort	refers to the different patient cohorts within the dataset. Each cohort represents a group of patients with similar characteristics or from a specific study or clinical trial. The cohort attribute categorizes and organizes the data based on these different groups.
10	er_status_measured_by_ihc	refers to the measurement of the estrogen receptor (ER) status using immunohistochemistry (IHC) in breast cancer samples. This attribute likely contains information about the ER status of the samples
11	er_status	refers to the estrogen receptor status of breast cancer samples. Estrogen receptor (ER) status is an important factor in determining the treatment and prognosis of breast cancer patients.
12	neoplasm_histologic_grade	refers to the histologic grade of the neoplasm, which is a measure of how abnormal the cells in the tumor tissue look under a microscope
13	her2_status_measured_by_snp6	refers to the measurement of the HER2 gene status using SNP6 (single nucleotide polymorphism) technology. HER2 is a gene that can play a role in the development of breast cancer, and its status (whether it is amplified or not) can have implications for treatment decisions.
14	her2_status	refers to a particular sample's human epidermal growth factor receptor 2 (HER2) status. HER2 is a protein that can promote the growth of cancer cells, and its overexpression is associated with a more aggressive form of breast cancer.
15	tumor_other_histologic_subtype	refers to the histologic subtype of the tumor other than the primary subtype. This attribute provides additional information about the specific characteristics of the tumor that may be relevant for research and analysis.
16	hormone_therapy	refers to whether a patient has received hormone therapy as part of their treatment. Hormone therapy is a common treatment for hormone receptor-positive breast cancer, and it works by blocking the effects of hormones or lowering hormone levels in the body to prevent cancer cells from growing.
17	inferred_menopausal_state	refers to the inferred menopausal status of the patients in the dataset. Based on certain clinical and biological factors, this attribute indicates whether a patient is premenopausal or postmenopausal.
18	integrative_cluster	refers to the integration of multiple types of data, such as gene expression, DNA methylation, and clinical information, to cluster breast cancer samples into distinct subgroups based on their molecular characteristics. This integrative clustering approach helps to identify different subtypes of breast cancer with unique molecular features, which can have implications for prognosis and treatment.

19	primary_tumor_laterality	refers to the location of the primary tumor in breast cancer patients, specifically whether it is located on the left or right side of the body. This attribute can be used to analyze the potential impact of tumor laterality on patient outcomes and treatment strategies.
20	lymph_nodes_examined_positive	refers to the number of lymph nodes examined that tested positive for cancer. This attribute is important in cancer research as it provides information about the spread of cancer in the body.
21	mutation_count	refers to the total number of mutations detected in a specific sample. This attribute provides valuable information about the genomic instability and mutation burden of the sample, which can be important for understanding the underlying genetic factors in breast cancer.
22	nottingham_prognostic_index	The Nottingham Prognostic Index (NPI) is a scoring system used to predict the likelihood of breast cancer recurrence and overall survival. It considers the tumor's size, the number of lymph nodes affected, and the grade of the tumor. The NPI is calculated using the formula: NPI = (0.2 x tumor size in cm) + (1 x lymph node stage) + (1 x tumor grade).
23	oncotree_code	refers to a code used to classify different types of cancer based on the Oncotree classification system. The Oncotree classification system is a standardized ontology for cancer types, which provides a hierarchical structure for organizing and categorizing different cancer subtypes.
24	overall_survival_months	refers to the overall survival time in months for patients with breast cancer. This attribute provides information on the length of time from the initial diagnosis to either the patient's death or the end of the study period.
25	overall_survival	describes the patients' general survival status in the dataset, including their likelihood of survival or death.
26	pr_status	refers to the progesterone receptor status in breast cancer. Unlike other receptor statuses like estrogen (ER) and human epidermal growth factor receptor 2 (HER2), the prediction of PR status might vary in certain analyses
27	radio_therapy	refers to the treatment involving radiation for breast cancer patients.
28	3-gene_classifier_subtype	defines subtypes based on gene expression. This attribute includes values like 'ER-/HER2-', indicating a specific subtype lacking estrogen receptors and HER2 expression, and 'ER+/HER2- High Prolif' representing estrogen receptor-positive tumors with high proliferation rates
29	tumor_size	is a crucial attribute representing the size of breast tumors in mm of patients
30	tumor_stage	refers to the stage of breast cancer
31	death_from_cancer	refers to the patient's state, indicating if they are still alive or if they passed away from breast cancer or another illness.

EDA is applied to this dataset to learn more about the correlations between various factors and the distribution of these variables among patients. EDA is essential to cancer research because it may be used to find patterns, connections, and possible causes that could affect how breast cancer develops and spreads. Through the discovery of novel correlations among variables, EDA can offer significant insights into treatment approaches and facilitate future developments within the domain.

4.1.1 Comprehending and Cleaning Data

- **Missing Values in Different Variables:** Certain variables, such as tumor size and tumor stage, contain missing values. In contrast, variables like age at diagnosis and lymph nodes examined are complete and do not exhibit any missing values. It is depicted in Figure 1.

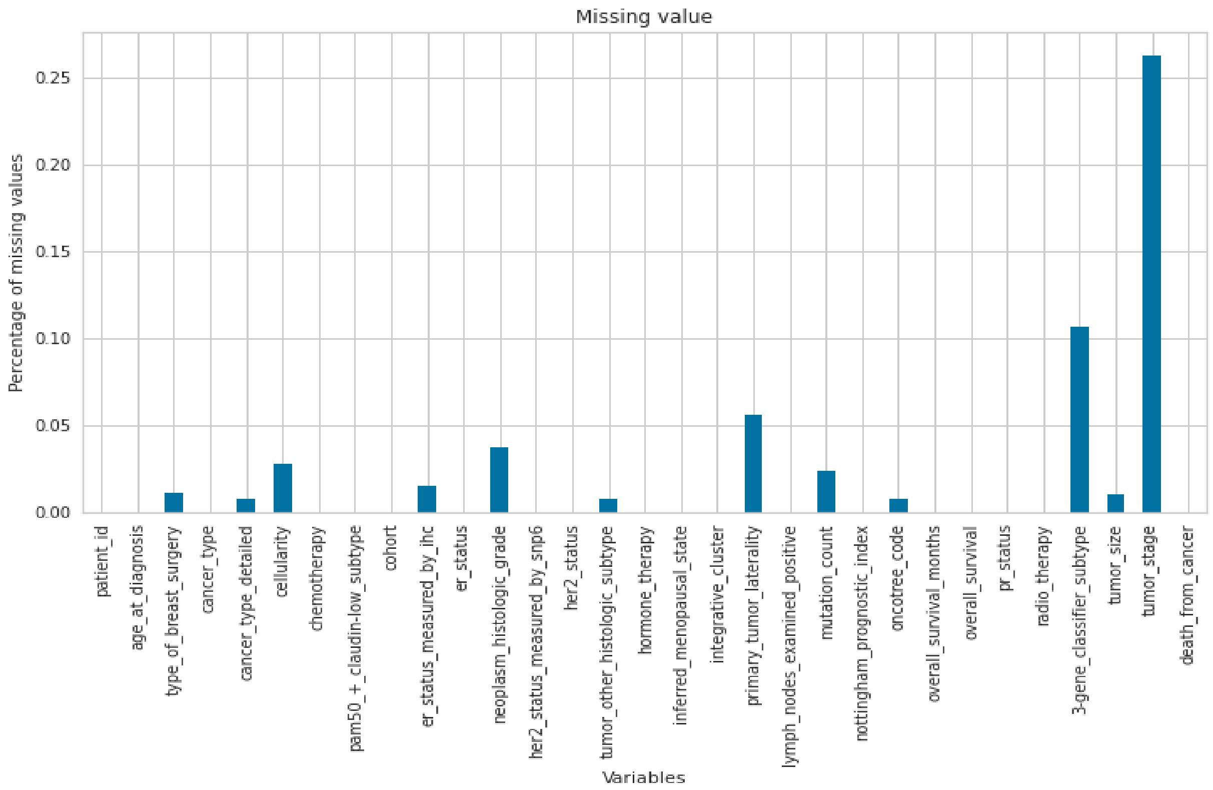


Figure 1: Bar plots showing the %age of missing values in the dataset

4.2. Workflow

The workflow and methodology used for the analysis is depicted in Figure 2.

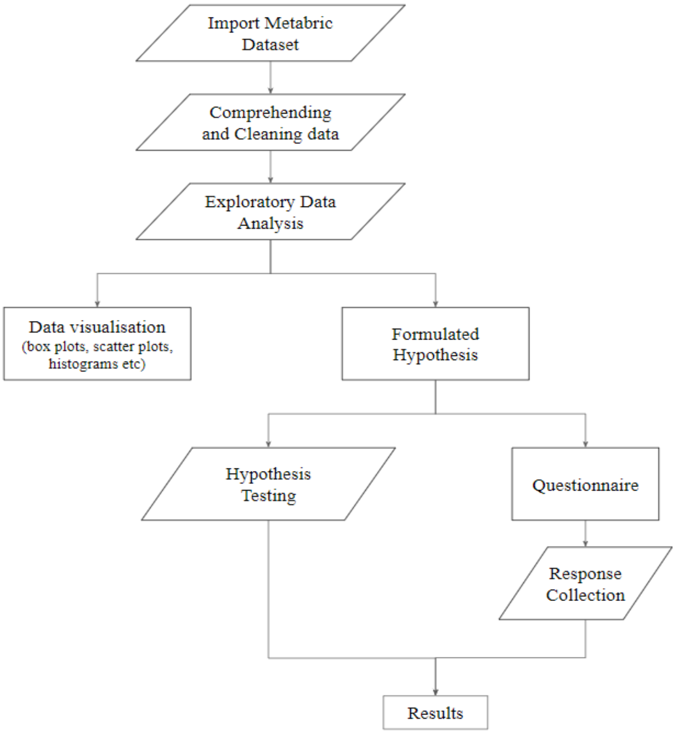

Import Metabric Dataset
Comprehending and Cleaning data
Exploratory Data Analysis
Data visualisation (box plots, scatter plots, histograms etc)
Formulated Hypothesis
Hypothesis Testing
Questionnaire
Response Collection
Results

Figure 2: Workflow of proposed methodology

4.3. Tools

Primarily, Python was used for the data analysis and visualization. Further, statistical tools were used for hypothesis testing and significance analysis.

5. Results and Discussion

This section discusses the analysis of exploratory data, the creation of hypotheses based on those analyses, and the testing of those hypotheses through the collection of questionnaire responses.

5.1 Data Visualization

To better comprehend the relationships between some of the variables and possibly get insight into how they affect one another, visuals were generated based on the primary questions presented at the beginning of the article.Univariate and multivariate analyses are used to find answers to many questions. Tables and Charts (Boxplot, Histogram, and Barplot) are the primary ways of doing it

RQ1: What impact does age have on a patient's chance of survival with breast cancer?

This can be visualized by a box plot graph illustrating how age at diagnosis affects outcomes. The following things are evident:

- The age at diagnosis column displays the difference in the two distributions, indicating that older individuals **are more likely to die from breast cancer. The duration from the time of intervention until death or the present is longer in patients who make a full recovery.** This suggests that people with breast cancer are either dying young or surviving at an early stage.
- **Individuals who have had breast cancer treatment have a higher chance of living longer than those who have passed away from the disease itself or another cause.** The mortality rate for those with breast cancer is higher than that of people with other illnesses. **An increased chance of dying from another illness exists for patients who are older than 60 at the time of diagnosis**. Breast cancer patients with a younger age are more likely to die from the disease or survive it. This is depicted in Figure 3.

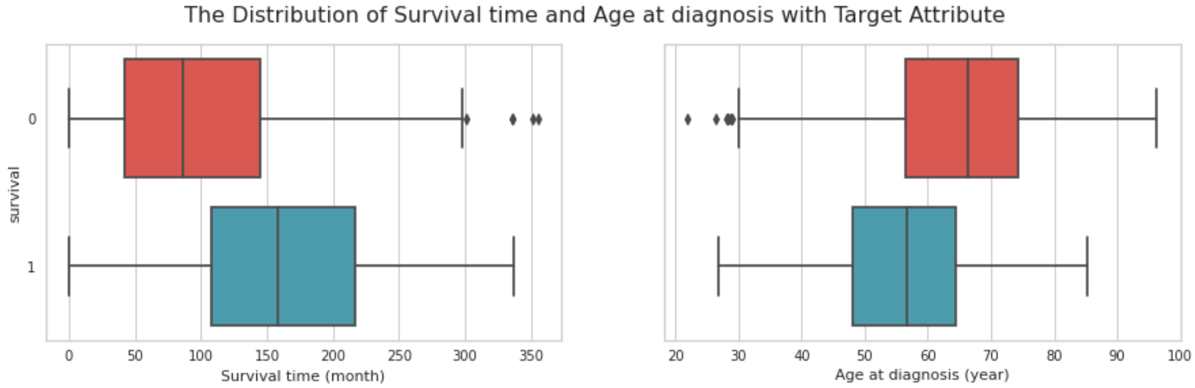


Figure 3: Box plots showing the distribution of Survival time(months) and Age at diagnosis(years) with target attributes

RQ2: How much does the tumor size at each stage affect a patient's chance of survival?

This investigation looks at how tumor size affects survival rate at different stages of breast cancer. The box plot in Figure 4 illustrates the analysis. In this regard, it is important to mention the following considerations**.**

- Stage 0.0: **Deaths had tumors that were roughly 60–70 mm in size**. Tumor sizes in **survivors** range from **15 to 35**. With one exception
- Stage 1.0: With one exception, all patients who died had tumors larger than 150 mm. Patients who survived had tumors distributed rather evenly.
- Stage 2.0 The tumor size distribution of patients who died and those who survived differs somewhat, with about the same median
- Stage 3.0: Tumor sizes in patients who passed away ranged from about 10 to more than 175. Tumor sizes in those who survive range from 20 to 100.
- Stage 4.0: Patients who died have tumor sizes between 20-70. Patients who survived have the same tumor size, around 60-70.

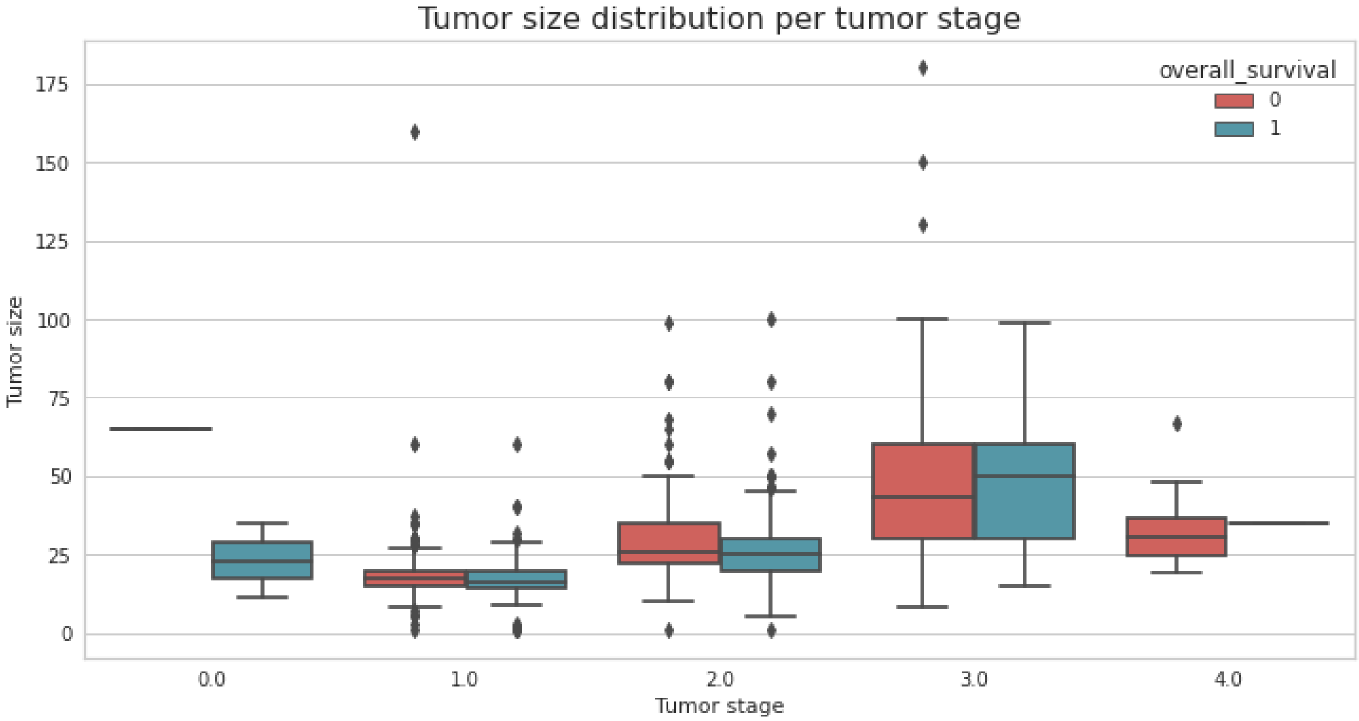


Figure 4: Box plots showing the tumor size distribution per tumor stage

RQ3: What is the impact of positively checked lymph nodes on the survival rate?
Several positive lymph nodes play a more important role in determining the likelihood of survival than tumor size. The survivor class has fewer positive lymph nodes and a smaller median tumor size than the deceased class. Higher numbers of positive lymph nodes correlate with shorter survival durations. This correlation can also be seen from the scatterplot in Figure 5.

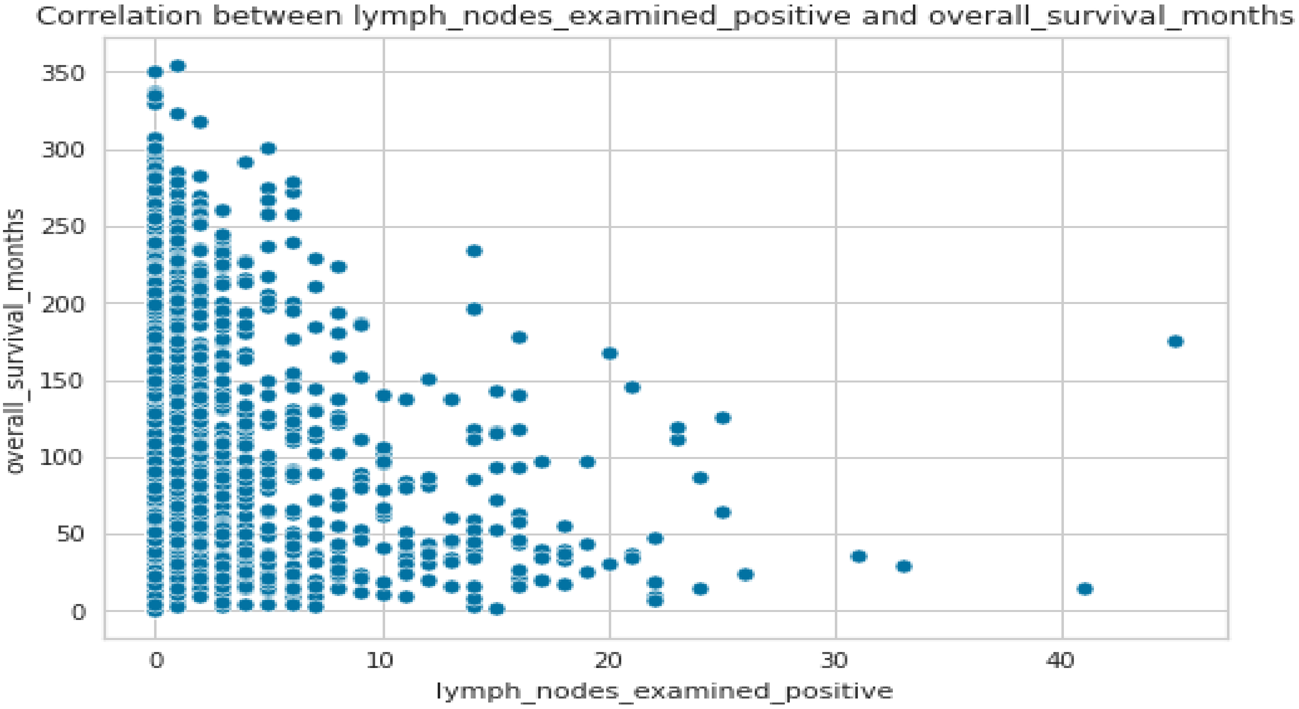


Figure 5: Scatter plot showing the correlation between positive lymph nodes and the number of survival months post-diagnosis.

Table 3 provided herein offers descriptive statistics derived from the dataset concerning tumor size and the count of positive lymph nodes. Observationally, a distinct pattern is absent in the relationship between tumor stage and tumor size. Conversely, a consistent correlation between the number of positive lymph nodes and the tumor stage is evident. Notably, this association is particularly conspicuous in the mean column. The mean tumor size displays no discernible trend, exhibiting a lack of consistent variation across stages. In contrast, the mean count of lymph nodes portrays a consistent increasing trend across stages, signifying a progressive relationship between lymph node count and the advancing tumor stage.

Table 3: 5 number summary statistics of Tumor Size and Positive Lymph Node attributes from the dataset.

Tumor Stage	Minimum	First Quartile	Mean	Third Quartile	Maximum
Tumor Size (diameter in mm)					
0	11	23	35	50	65
1	1	15	17	20	160
2	1	21	25	30	100
3	8	30	45	60	180
4	19	25	31	35	67
Number of Positive Lymph Nodes					
0	0	0	0	0	0
1	0	0	0	0	22
2	0	0	1	2	41
3	0	2	6	10	23
4	0	0	6	7	25

RQ4: How are the patient's survival months post-diagnosis related to their progesterone receptor (PR) status and Estrogen receptor (ER) status?

The presence of positive Estrogen Receptors (ER) and Progesterone Receptors (PR) is identified as a favorable prognostic factor for breast cancer survival, indicative of a potentially slower and less aggressive cancer growth. The visual representation of this observation is depicted through box plots in Figure 6, revealing distinct distributions between positive and negative receptor statuses. Notably, the box plots consistently illustrate that the interquartile range (Q1-Q3) for cases with positive ER and PR attributes tends to be associated with a greater number of survival months in comparison to cases with negative receptors.

The analysis reveals a notable difference in the range between Q1 and Q3 for negative receptors, signifying a more varied distribution, while positive receptor statuses exhibit a more normalized distribution. These findings underscore the significance of ER and PR status in influencing breast cancer outcomes, providing valuable insights into the potential impact of hormone receptor expression on the trajectory of the disease. The 5-number summary in Table 4 numerically denotes the conclusion, as shown in the box plots .

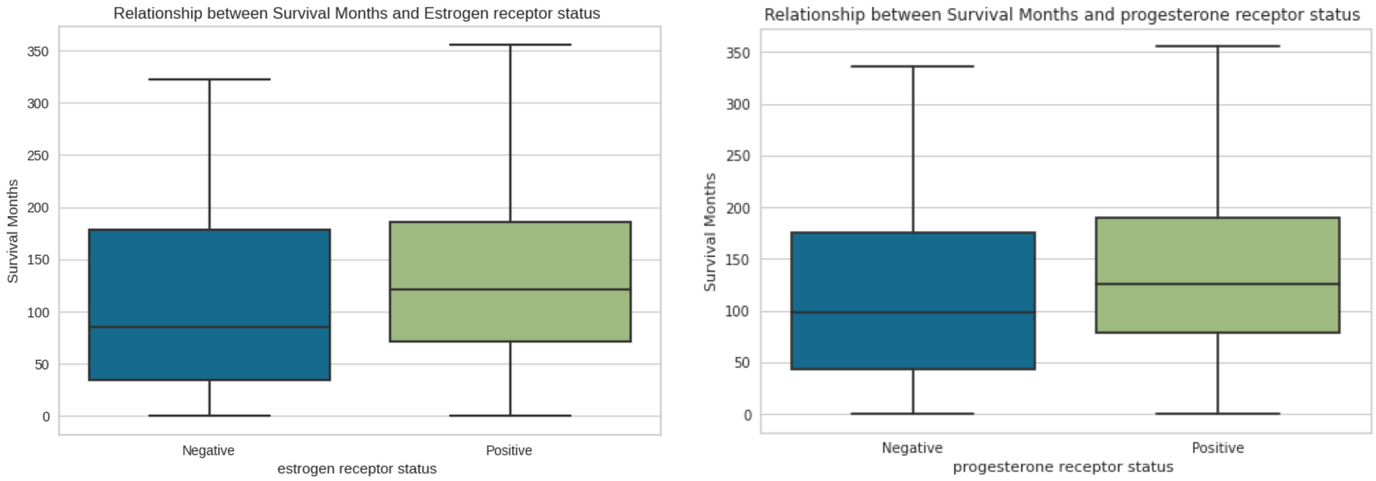


Figure 6: Box plots showing the relationship between survival months post-diagnosis related to their ER status and PR status

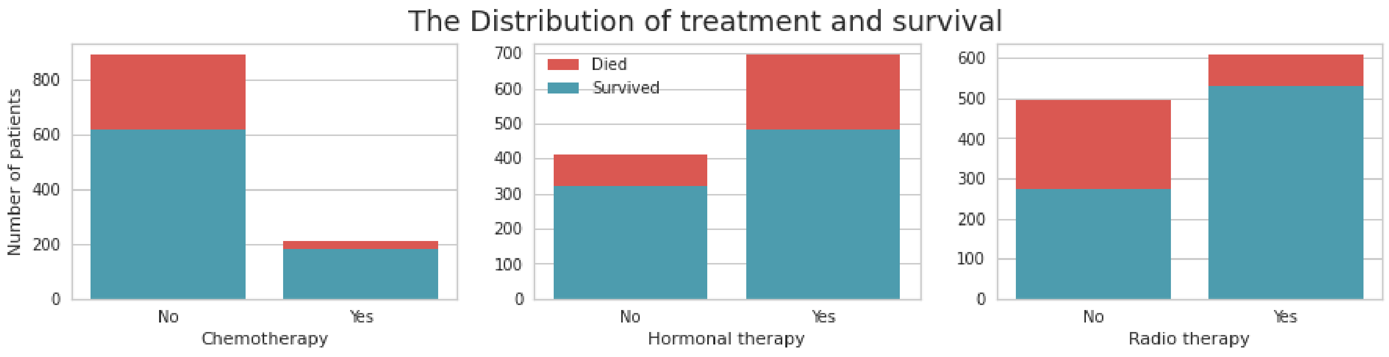
Table 4: 5 number summary statistics of the distribution of survival months categorized by ER and PR status

	Minimum	**First Quartile**	**Mean**	**Third Quartile**	**Maximum**
ER Status	**Survival Months**				
Negative	0	34.7667	85.3667	178.4	322.833
Positive	0.1	71.4833	121.533	186.1667	355.2
PR Status	**Survival Months**				
Negative	0	43.2833	98.7	175.8667	335.733
Positive	0.1	78.6	125.6	190.1	355.2

RQ5: Which treatment modality—chemotherapy, hormone therapy, or radiation therapy—is most usually utilized to increase breast cancer patient's chances of survival?

- **Chemotherapy treatment:** There is a substantial difference between the patients in the dead class and the survivors class when compared to those who did not get chemotherapy. In the chemotherapy-free death class, there are roughly 800 patients. In contrast, there are only a little over 200 patients in the chemotherapy-dead class. This can be seen in Figure 7 (a).

- **Hormonal therapy:** There is a discernible difference between the patient classes who died and those who survived when comparing those on hormone therapy to those who did not. There are only about 700 patients in the chemotherapy death class, which is a pretty high amount. There are little over 400 people in the chemotherapy death class, which is a rather small amount. This can be seen in Figure 7 (b).
- **Radiotherapy:** When radiation recipients are compared to non-receivers, there is a noticeable difference between the patient classes that died and those that lived. In comparison to other treatment modalities, the class in which the greatest number of patients died both with and without radiation. This can be seen in Figure 7 (c).
- Most patients first take chemotherapy and hormone therapy, and then radiotherapy and chemotherapy. Radiation therapy is the most often used treatment when it comes to single people. There are no patients who underwent radiation and hormone therapy at the same time.



(a) (b) (c)

Figure 7: Bar plots depicting the comparison of three treatment therapies

RQ6: Are there any correlations between the numerical variables of the dataset?

The attributes that are negatively and positively correlated with the target class are displayed in Table 5.

Table 5: Attributes sorted by their correlation to overall survival.

S.No.	Name of attribute	value
1	overall_survival	1.000000
2	overall_survival_months	0.384467
3	type_of_breast_surgery_BREAST CONSERVING	0.187856
4	inferred_menopausal_state_Pre	0.170915
5	radio_therapy	0.112083
6	3-gene_classifier_subtype_ER+/HER2- Low Prolif	0.094463
7	pam50_+_claudin-low_subtype_claudin-low	0.091397
8	integrative_cluster_10	0.076256
9	pam50_+_claudin-low_subtype_LumA	0.065186
10	3-gene_classifier_subtype_ER-/HER2-	0.065135
11	lymph_nodes_examined_positive	-0.164498
12	inferred_menopausal_state_Post	-0.170915
13	type_of_breast_surgery_MASTECTOMY	-0.184259
14	tumor_stage	-0.188790
15	age_at_diagnosis	-0.303666

5.2 Formulation of Hypotheses:

After the data analysis, hypotheses were formulated based on observed patterns. Subsequently, a chi-square significant test and t-tests were performed to identify statistically significant hypotheses, refining our understanding of key variables and their impact on the outcomes. Table 6 illustrates the results derived from chi-square and t-tests. With consistently low p-values across all hypotheses(except 2 and 7), we reject the null hypothesis, signifying a substantial association between variables such as age at diagnosis, tumor size, lymph nodes, and biomarkers (ER, PR) with patient survival. Additionally, our analysis highlights the efficacy of post-surgery radiotherapy in mitigating recurrence rates and reducing breast cancer-related fatalities.

Table 6: Shortlisted hypothesis with Chi-Square and T-test p-values.

S.No.	**Hypothesis**	**p-value (chi-square)**	**p-value (t-test)**
1.	Older patients face a higher risk of mortality. Patients over 60 at diagnosis face increased risks due to other illnesses.	2.34e-52	5.97e-62
2.	Age at diagnosis and the potential for distant metastasis significantly influence outcomes.	3.21e-22	0.067
3.	Tumor size, particularly beyond 150 mm, is associated with increased fatalities. Survivors tend to have smaller and more evenly distributed tumor sizes.	1.12e-21	4.54e-16
4.	Positive lymph node counts are decisive in predicting survival outcomes, even more so than tumor size.	-	-
5.	Lymph node involvement significantly impacts overall survival rates.	1.75e-22	3.40e-23
6.	Understanding factors such as age, tumor size, estrogen receptor status, progesterone receptor status, and lymph node examined involvement aids in prognosis and treatment decisions for metastatic cancers.	-	-
7.	Post-surgery radiotherapy reduces recurrence and breast cancer-related deaths for specific groups.	4.28e-22	0.068

5.3 Creating a questionnaire:

Built based on our hypotheses, a methodologically crafted questionnaire was distributed to medical experts, constituting a vital step in our research methodology. This process serves as a means of human validation, essential for verifying the reliability of our findings. By involving medical professionals, we seek authoritative perspectives that complement and authenticate our quantitative analyses, enhancing our research outcomes' overall rigor and validity.

5.3.1 Participant Recruitment and Questionnaire Administration:

What is your current role in the Medical Field?
37 responses

Years of Experience
37 responses

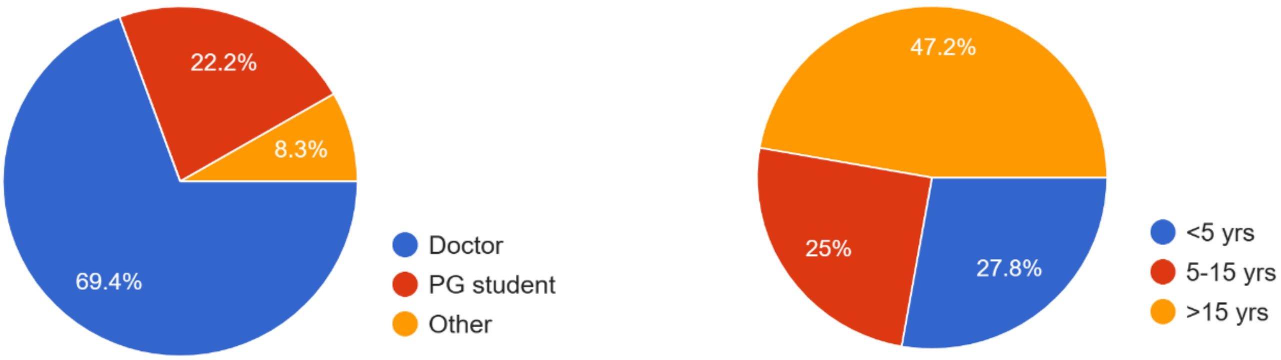


Figure 8: Pie charts depicting the distribution of professional and years of experience of the participants.

5.3.2 Questionnaire Design and Observations

The questionnaire employed in this study comprised four sections: "Effect of Age", "Effect of Tumor Size", "Effect of Lymph Nodes" and "Effect of Cancer Treatment" following the EDA on these specific prognostic variables. Each section contained a list of relevant hypotheses, allowing participants to select multiple statements aligning with their perspectives. The decision to employ a multiple-choice format rather than a binary yes/no was made to accommodate the complexity of breast cancer outcomes and to recognize that responses might encapsulate multifaceted perspectives. However, for further analysis, the responses were treated as binary. This structured approach facilitated a comprehensive exploration of factors influencing breast cancer outcomes.

Due to the binary nature of the responses, the small sample size of the data set, and to assess individual hypotheses with a categorical and nominal nature, a Binomial test was performed to estimate the statistical significance of the responses using p-values. Table 7 presents the results of questionnaire responses alongside binomial test p-values. It is evident that hypotheses 1, 4, 5, and 6 align closely with the findings of the exploratory data analysis (EDA)

Table 7: %age agreement from questionnaire responses with Binomial test p-values to test statistical significance.

S.No.	**Hypothesis**	**Agreement %**	**P-value**
1.	Older patients face a higher risk of mortality. Patients over 60 at diagnosis face increased risks due to other illnesses.	47.2%, 75%	0.2592
2.	Age at diagnosis and the potential for distant metastasis significantly influence outcomes.	41.7%	**0.0886**
3.	Tumor size, particularly beyond 150 mm, is associated with increased fatalities. Survivors tend to have smaller and more evenly distributed tumor sizes.	47.2%	0.6173
4.	Positive lymph node counts are decisive in predicting survival outcomes, even more so than tumor size.	58.3%	0.1332
5.	Lymph node involvement significantly impacts overall survival rates.	75%	**5.636 e^-7**
6.	Understanding factors such as age, tumor size, estrogen receptor status, progesterone receptor status, and lymph node examined involvement aids in prognosis and treatment decisions for metastatic cancers.	61%	**0.0352**
7.	Post-surgery radiotherapy reduces recurrence and breast cancer-related deaths for specific groups.	55.6%	0.3682

6. Conclusion

In conclusion, the study systematically evaluated several hypotheses supported by analyzed data to unravel critical factors influencing breast cancer outcomes presented in Tables 6 and 7. It is evident that certain factors significantly influence cancer prognosis and treatment outcomes.

Older patients, particularly those over 60 at diagnosis (Hypothesis 1), face a significantly higher risk of mortality, with statistical tests yielding extremely low p-values ($p < 0.001$). Expert consensus also strongly supports this notion, with 75% agreement among experts. Tumor size and lymph node involvement emerge as critical determinants of survival rates. Statistical analyses demonstrate significant associations between tumor size beyond 150 mm and increased fatalities ($p < 0.01$), as well as between lymph node involvement and overall survival rates ($p < 0.00001$) (Hypothesis 3). Expert opinions further confirm the importance of these factors, with 47.2% to 75% agreement among experts. Furthermore, post-surgery radiotherapy appears to be a promising intervention for reducing recurrence and cancer-related deaths in specific patient groups. While the statistical significance varies ($p < 0.01$ to $p > 0.05$), expert consensus suggests its potential efficacy, with agreement percentages of 55.6%. The borderline significance in some cases (Hypothesis 2 & 7) suggests nuanced associations that may warrant further investigation or consideration in a larger cohort.

The integration of a questionnaire-based survey provided a valuable human validation layer to our quantitative analyses. Despite some discrepancies between statistical findings and expert opinions on certain hypotheses, such as the influence of distant metastasis on outcomes, there is generally a high level of alignment regarding the impact of key factors like lymph node involvement.

While the survey's modest sample size of 37 remains a limitation, it sets the stage for future endeavors. The study's findings underscore the need for larger-scale investigations to validate and refine our observations. Moreover, addressing the survey's shortcomings, such as the potential for response bias, could enhance the robustness of future analyses.

References

[1] "Exploratory Data Analysis in Machine Learning for Data Science." Accessed: Dec. 29, 2023. [Online]. Available: https://www.linkedin.com/pulse/exploratory-data-analysis-machine-learning-science-aritra-pain

[2] "Role of Exploratory Data Analysis in Machine Learning - The Talent500 Blog." Accessed: Dec. 29, 2023. [Online]. Available: https://talent500.co/blog/role-of-exploratory-data-analysis-in-machine-learning/

[3] "The Ultimate Guide to Machine Learning: Exploratory Data Analysis (EDA) — Part -1 | by Simranjeet Singh | Medium." Accessed: Dec. 29, 2023. [Online]. Available: https://medium.com/@simranjeetsingh1497/the-ultimate-guide-to-machine-learning-from-eda-to-model-deployment-part-1-46083efb2d3c

[4] E. J. Sweetlin and S. Saudia, "Exploratory data analysis on breast cancer dataset about survivability and recurrence," *2021 3rd International Conference on Signal Processing and Communication, ICPSC 2021*, pp. 304–308, May 2021, doi: 10.1109/ICSPC51351.2021.9451811.

[5] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA Cancer J Clin*, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/CAAC.21763.

[6] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA Cancer J Clin*, vol. 72, no. 1, pp. 7–33, Jan. 2022, doi: 10.3322/CAAC.21708.

[7] R. Rabiei, S. M. Ayyoubzadeh, S. Sohrabei, M. Esmaeili, and A. Atashi, "Prediction of Breast Cancer using Machine Learning Approaches," *J Biomed Phys Eng*, vol. 12, no. 3, p. 297, Jun. 2022, doi: 10.31661/JBPE.V0I0.2109-1403.

[8] C. G. Yedjou, S. S. Tchounwou, R. A. Aló, R. Elhag, B. Mochona, and L. Latinwo, "Application of Machine Learning Algorithms in Breast Cancer Diagnosis and Classification.," *International Journal of Science Academic Research*, vol. 2, no. 1, pp. 3081–3086, Jan. 2021, Accessed: Dec. 23, 2023. [Online]. Available: https://europepmc.org/articles/PMC8612371

[9] R. Rabiei, S. M. Ayyoubzadeh, S. Sohrabei, M. Esmaeili, and A. Atashi, "Prediction of Breast Cancer using Machine Learning Approaches," *J Biomed Phys Eng*, vol. 12, no. 3, p. 297, Jun. 2022, doi: 10.31661/JBPE.V0I0.2109-1403.

[10] M. A. Wakili *et al.*, "Classification of Breast Cancer Histopathological Images Using DenseNet and Transfer Learning," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/8904768.

[11] I. Mihaylov, M. Nisheva, and D. Vassilev, "Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies," *Inf.*, vol. 10, no. 3, 2019, doi: 10.3390/INFO10030093.

[12] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 16, no. 3, pp. 841–850, May 2018, doi: 10.1109/TCBB.2018.2806438.

[13] J. Keyl *et al.*, “Multimodal survival prediction in advanced pancreatic cancer using machine learning,” *ESMO Open*, vol. 7, no. 5, Oct. 2022, doi: 10.1016/j.esmoop.2022.100555.

[14] A. Spooner *et al.*, “A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction,” *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-77220-w.

[15] W. Tizi and A. Berrado, “Machine learning for survival analysis in cancer research: A comparative study,” *Sci Afr*, vol. 21, p. e01880, Sep. 2023, doi: 10.1016/J.SCIAF.2023.E01880.

[16] W. Zhu, L. Xie, J. Han, and X. Guo, “The Application of Deep Learning in Cancer Prognosis Prediction,” *Cancers (Basel)*, vol. 12, no. 3, Mar. 2020, doi: 10.3390/CANCERS12030603.

[17] N. Arya, A. Mathur, S. Saha, and S. Saha, “Proposal of SVM Utility Kernel for Breast Cancer Survival Estimation.,” *IEEE ACM Trans. Comput. Biol. Bioinform.*, vol. 20, no. 2, pp. 1372–1383, Mar. 2023, doi: 10.1109/TCBB.2022.3198879.

[18] H. Bai, Y. Wang, H. Liu, and J. Lu, “Development of a Four-mRNA Expression-Based Prognostic Signature for Cutaneous Melanoma,” *Front Genet*, vol. 12, Jul. 2021, doi: 10.3389/FGENE.2021.680617/PDF.

[19] J. Teng *et al.*, “Bayesian Inference of Lymph Node Ratio Estimation and Survival Prognosis for Breast Cancer Patients,” *IEEE J Biomed Health Inform*, vol. 24, no. 2, pp. 354–364, Feb. 2020, doi: 10.1109/JBHI.2019.2943401.

[20] P. Liu, B. Fu, S. X. Yang, L. Deng, X. Zhong, and H. Zheng, “Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer,” *IEEE Trans Biomed Eng*, vol. 68, no. 1, pp. 148–160, Jan. 2021, doi: 10.1109/TBME.2020.2993278.

[21] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” *Electronics (Switzerland)*, vol. 8, no. 8, Aug. 2019, doi: 10.3390/ELECTRONICS8080832.

[22] S. C. Lu, C. L. Swisher, C. Chung, D. Jaffray, and C. Sidey-Gibbons, “On the importance of interpretable machine learning predictions to inform clinical decision making in oncology,” *Front Oncol*, vol. 13, p. 1129380, Feb. 2023, doi: 10.3389/FONC.2023.1129380/BIBTEX.

[23] E. Onose, “Explainability and Auditability in ML: Definitions, Techniques, and Tools,” neptune.ai. [Online]. Available: https://neptune.ai/blog/explainability-auditability-ml-definitions-techniques-tools

[24] A. Sarica, F. Aracri, M. G. Bianco, F. Arcuri, A. Quattrone, and A. Quattrone, “Explainability of random survival forests in predicting conversion risk from mild cognitive impairment to Alzheimer’s disease,” *Brain Inform*, vol. 10, no. 1, p. 31, Dec. 2023, doi: 10.1186/s40708-023-00211-w.

[25] F. Di Martino and F. Delmastro, “Explainable AI for clinical and remote health applications: a survey on tabular and time series data,” *Artif Intell Rev*, vol. 56, no. 6, pp. 5261–5315, Jun. 2023, doi: 10.1007/s10462-022-10304-3.

[26] T. Suresh, T. A. Assegie, S. Ganesan, R. L. Tulasi, R. Mothukuri, and A. O. Salau, “Explainable extreme boosting model for breast cancer diagnosis,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5764–5769, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5764-5769.

[27] “Improve explainability of ML models to meet regulatory requirements - Amazon Science.” Accessed: Dec. 24, 2023. [Online]. Available: https://www.amazon.science/latest-news/remars-revisited-improve-explainability-of-ml-models-to-meet-regulatory-requirements

[28] M. Stewart *et al.*, “An Exploratory Analysis of Real-World End Points for Assessing Outcomes Among Immunotherapy-Treated Patients With Advanced Non–Small-Cell Lung Cancer,” *JCO Clin Cancer Inform*, no. 3, pp. 1–15, Dec. 2019, doi: 10.1200/cci.18.00155.

[29] P. Shahmirzalou, M. J. Khaledi, M. Khayamzadeh, and A. Rasekhi, “Survival analysis of recurrent breast cancer patients using mix Bayesian network,” *Heliyon*, vol. 9, no. 10, p. e20360, Oct. 2023, doi: 10.1016/J.HELIYON.2023.E20360.

[30] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Med Inform Decis Mak*, vol. 19, no. 1, Mar. 2019, doi: 10.1186/s12911-019-0801-4.

[31] Z. Zhang, H. Chai, Y. Wang, Z. Pan, and Y. Yang, "Cancer survival prognosis with Deep Bayesian Perturbation Cox Network," *Comput Biol Med*, vol. 141, Feb. 2021, doi: 10.1016/J.COMPBIOMED.2021.105012.

[32] C. H. Yang, S. H. Moi, F. Ou-Yang, L. Y. Chuang, M. F. Hou, and Y. Da Lin, "Identifying Risk Stratification Associated with a Cancer for Overall Survival by Deep Learning-Based CoxPH," *IEEE Access*, vol. 7, pp. 67708–67717, 2019, doi: 10.1109/ACCESS.2019.2916586.

[33] B. Fu, P. Liu, J. Lin, L. Deng, K. Hu, and H. Zheng, "Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data," *IEEE Trans Biomed Eng*, vol. 66, no. 7, pp. 2053–2064, Jul. 2019, doi: 10.1109/TBME.2018.2882867.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 1135–1144, Aug. 2016, doi: 10.1145/2939672.2939778.