Received: 30 September 2024 | Revised: 2 December 2024 | Accepted: 30 December 2024 | Published online: 22 January 2025

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2025, Vol. 00(00) 1–12
DOI: 10.47852/bonviewJDSIS52024464

BON VIEW PUBLISHING

Breast Cancer Survival Analysis and Mortality Prediction Under Different Treatment Combinations

Parag C. Pendharkar[1,*] **and James A. Rodger**[2]

[1]*School of Business Administration, Pennsylvania State University at Harrisburg, USA*

[2]*Management Information Systems, Slippery Rock University of Pennsylvania, USA*

Abstract: In this paper, a combination of breast cancer treatment procedures is considered, and its impact on breast cancer survival is precisely observed. Both statistical and neural network procedures are used to predict the breast cancer survival time. The results indicate that treatment procedures that use surgical options improve breast cancer survival. In the case of non-surgical options, hormone therapy appears to be the best. Additionally, the results suggest that radiation and chemotherapy combination lead to lower survival rates. The dataset used in this research had limited cases where the chemotherapy option was prescribed. Chemotherapy alone was a confounding cancer treatment option for non-node-positive cancer. For node-positive cancer cases, chemotherapy seems to work best where the surgery option is not considered or is viable. The experiments with neural networks show that neural networks can help predict the event of death, but these techniques could not accurately predict the length of survival.

Keywords: survival analysis, breast cancer, classification, neural networks

1. Introduction

Breast cancer is a significant type of cancer affecting women in the US [1], with an estimated 2.3 million new cases worldwide [2]. Statistics show that nearly one in three female cancer cases are related to breast cancer [3]. Breast cancer is also the second leading cause of female cancer deaths in the US [3]. Breast cancer research studies have primarily focused on statistical methods that use survival analysis [4] and machine learning methods that predict breast cancer occurrence [5] or its intensity (benign or malignant) [6]. Both methods need further improvements.

Survival analysis methods assess the impact of cancer treatment plans on survival outcomes. Prior research on cancer outcomes focused on establishing survival rates under individual treatment plans [4]. While useful, these studies ignore the most realistic situations where doctors often allow patients to undergo different treatment plans to identify a treatment plan that works the best. Under such circumstances, survival outcomes should be measured as treatment combinations instead of an individual treatment plan. These treatment plan combinations could perhaps be sequenced in a certain way to improve the overall efficacy of cancer treatment. Survival analysis under such situations gets challenging, but this study shows that special coding procedures could be used to identify survival curves for such analyses.

Machine learning methods are primarily predictive and cannot be directly compared to statistical survival analysis methods. Statistical survival methods focus on events (cancer death) and survival times until an event is observed. Multiple output machine learning methods such as neural networks allow a researcher to predict two outputs: cancer death and survival time. Such multiple output machine learning methods were not used previously and warrant researchers' attention. The multiple output machine learning methods should also be compared with traditional statistical survival analysis methods to highlight their merits and drawbacks. Treatment plan combinations mentioned earlier can be easily incorporated in machine learning methods as multiple different binary inputs.

Breast cancer treatment procedures can be surgical or non-surgical. Non-surgical breast cancer treatment procedures include hormone therapy, radiotherapy, and chemotherapy [4]. Hormone therapy centers around regulating essential female hormones called estrogen. Estrogen prepares the female body for pregnancy in adulthood and maintains cardiovascular and bone health. Estrogen also helps cancers grow in women because most female organs contain estrogen receptors. One way to reduce breast cancer growth is to reduce the exposure of estrogen to breast cancer cells. This reduction of exposure of estrogen to breast cancer cells is the primary goal of hormone therapy. Given the importance of estrogen to women's overall cardiovascular and bone health, hormone therapy often reduces bone density and increases the risk of heart disease in women. Hormone therapy is not an entirely non-surgical option because surgical options such as the removal of ovaries in premenopausal women can also reduce estrogen in women. While hormone therapy primarily focuses on reducing the growth of breast cancer, radiation therapy

*Corresponding author: Parag C. Pendharkar, School of Business Administration, Pennsylvania State University at Harrisburg, USA. Email: pxp19@psu.edu

kills cancer cells using high-energy X-rays. Side effects of radiation therapy are fatigue and sunburn-like skin damage. These side effects typically go away in several months since radiation treatment. The goal of chemotherapy is to stop the spread of cancer to other parts of the body and to kill fast-growing cancer cells. Chemotherapy is often administered using injections into muscles, veins, or arteries. Chemotherapy's side effects include unwanted damage of cells in other organs (heart, kidney, lungs, etc.,) and temporary hair loss.

Surgical treatment typically focuses on surgical removal of the area containing cancer cells. Two commonly used surgical procedures are lumpectomy [7] and mastectomy [8]. Lumpectomy deals with partial removal of breast, and mastectomy deals with complete removal of the breast.

Breast cancer typically starts in a few cells in breast tissue. Over time, these cells grow and divide and invade other tissues. When cancer growth spreads to lymph nodes, it is ready to metastasize. When breast cancer growth spreads to lymph nodes, it is called node-positive breast cancer. Treatment for such type of breast cancer is very aggressive. Chemotherapy and surgery play a prime role in treating node-positive breast cancers.

This paper investigates two research questions: First, do different breast cancer treatment combinations lead to different survival rates? Second, can multiple output neural networks be used to predict death due to breast cancer and survival time until death? If different breast cancer treatment combinations lead to different survival rates, then optimal treatment procedures can be constructed to improve the survival times of breast cancer patients. Additionally, predicting survival times accurately will lead to smother transition and adjustment for families of breast cancer survival patients.

The paper is organized as follows: Section 2 provides a brief overview of survival analysis methods and reviews the literature on applying machine learning methods for breast cancer diagnosis. Section 3 reports data, experiments, and results. Section 4 concludes this paper with a summary of the results.

2. Overview of Survival Analysis, Methods for Survival Analysis, and Machine Learning Applications in Breast Cancer Research

Survival analysis is used to study the impact of factors that influence the time of a specific event. In cancer survival literature, that event is typically the death of a patient. Classical statistical techniques such as linear regression cannot be used for survival analysis because techniques cannot handle censoring and time to an event is not normally distributed. As a result, survival analysis techniques assume that the event time distribution function, $f(t)$, is non-normally distributed. The cumulative event time distribution function (the probability that an event has happened in time less than or equal to t) is defined as follows:

$$F(t) = P(T \leq t) = \int_0^t f(u)du \tag{2.1}$$

The survival function, $S(t)$, which denotes the probability that an event will happen after a specific time t in the future, can be similarly computed from cumulative event time distribution as follows:

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(u)du \tag{2.2}$$

The hazard function or hazard rate, $h(t)$, is an instantaneous risk of an event happening at time t, given that a subject has survived until time t. This hazard function has the following relationship with probability density function (pdf) $f(t)$ and survival function $S(t)$:

$$h(t) = \frac{f(t)}{S(t)} \tag{2.3}$$

Survival analysis may be parametric or non-parametric depending on whether or not any underlying functional form for pdf is assumed in the analysis. Standard parametric survival analysis assumes any exponential, standard gamma, Weibull, or log-normal pdf. Maximum likelihood estimators are used for both parametric and non-parametric survival analyses. Among the popular approaches for survival analysis are the life table (LT) method, Kaplan-Meier (KM) survival analysis, and Cox proportional hazards regression.

2.1. Life table and Kaplan Meier survival analysis

The LT and KM survival analysis are non-parametric survival analysis methods where the maximum likelihood estimator of survival function $S(t)$ is used. The analysis starts with an ordering of event times in ascending order $t_1 < t_2 < \ldots < t_k$. The primary difference between LT and KM survival analysis is that, in LT analysis, the interval lengths (e.g., $t_2 – t_1 = t_k – t_{(k-1)}$) are fixed, whereas, in KM survival analysis, interval lengths are determined by the events where a death occurs. As a result, the event times in KM are defined by the events, where t_1 may be the shortest time in the database where the event has occurred, and t_k is the longest time in the dataset where the event is observed. The computational method for computing survival function in both methods is the same.

Let for some intermediate time $t \in [t_1, t_k)$, d_j be the number of individuals who have seen the event (died), and n_j be the number of individuals who have not seen the event (survived), then the KM estimator estimates the likelihood of $S(t)$ using the following expression:

$$\hat{s}(t) = \prod\nolimits_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.4}$$

LT and KM analyses can also be extended for hypothesis testing to determine whether the survival curves for different groups (e.g., node-positive cancer vs. non-node-positive cancer) are statistically different. Statistical tests such as log-rank, Wilcoxon, and likelihood ratio statistics are among the popular tests used to test these hypotheses.

2.2. The Cox proportional hazards regression

Let $x = [x_1, \ldots, x_s]^T$ be the set of covariates that impact the hazard rate $h(t)$. Since the hazard rate is always non-negative, assuming parameter vector $\beta = [\beta_1, \ldots, \beta_s]^T$, the hazard function can be written as follows:

$$h(t|x) = exp(\beta_0 + \beta^T x) \tag{2.5}$$

Since the formulation in (2.5) does not model hazard rate dependence on time, an additional multiplier $h_0(t)$, a baseline hazard rate that describes the dependence of hazard rate on time alone, can be added, and formulation (2.5) can be written for some individual "*i*" as follows:

$$h(t|x_i) = h_0(t) exp(\beta^T x_i) \tag{2.6}$$

The term "proportional" comes from the fact that the hazard rate stays constant over time with fixed covariates. For example, for two individuals i and j, the ratio between their hazard rates can be written as follows:

$$\frac{h(t|x_i)}{h(t|x_j)} = exp\left(\beta^T\left(x_i - x_j\right)\right) \tag{2.7}$$

The expression (2.7) illustrates that a subject most at risk at any one time remains most at risk at any other time. A log-linear regression model is obtained using logarithms on both sides of expression (2.6), and maximum likelihood procedures can be used to estimate parameters βs.

2.3. Brief review of literature on machine learning applications for breast cancer prediction

Machine learning techniques are used extensively for breast cancer prediction. Most studies used machine learning to solve traditional classification [9] and clustering breast cancer cases. Ronco [5] used artificial neural networks that used several demographic, hereditary, dietary, and other related variables to identify women who are in high-risk subpopulation groups for contracting breast cancer. Setiono [6] developed a neural network to learn breast cancer prediction rules and proposed a data preprocessing technique to improve classification accuracy further. Other applications of neural networks included the application of self-organization mapping neural networks for clustering breast cancer cases for identifying clinical trends [10], and application of hybrid [11] and evolutionary neural network for the prediction of breast cancer [12]. Some studies used fuzzy classifiers [13], a hybrid fuzzy genetic algorithm [14], and hybrid deep learning techniques [15] for predicting breast cancer.

Non-classification methods used regression, feature selection, and ensemble methods for breast cancer research. Among these studies was the use of support vector regression to predict clinical metastases time for breast cancer patients and rank different genes for their role in breast cancer metastases time [16]; the use of the Mann-Whitney statistical test for feature selection for improved breast cancer prediction [17]; a hybrid decision tree and genetic algorithm-based ensemble for cost-sensitive classification [18]; and a hybrid rough set and support vector machine ensemble for solving traditional breast cancer classification problems.

A few studies have used hybrid statistical and machine learning methods, deep learning, and case-based reasoning methods for breast cancer analysis. These studies addressed problems such as the data imputation procedure to fill in missing values [19], multi-category classification [20], tumor prediction [21], and providing visual reasoning for diagnosis explanations [22].

3. Data, Experiments, and Results

Data on 453 cases were obtained from a large cancer hospital in the northeastern US. The researchers did not formally collect this data as part of the study. Still, it was provided to them, and the researchers of this paper did not control the hospital's data acquisition procedures. The hospital provided all the available data. If some data was excluded, it was due to internal hospital procedures, and researchers had no control over these procedures.

The data were obtained under real-world conditions where not all patients were observed for the same length. Many cases were censored because they may have moved or changed physicians or hospitals. Survival times were computed using the diagnosis data to the date of death. Only the cases with known survival time values were included in the dataset. Death records were matched with the state death files, and deaths were classified as death due to cancer, death for non-cancer-related reasons, and death for unknown reasons. Of these 453 cases, 84 cases belonged to node-positive breast cancer. The overall average age of patients was 63.7 years, with a standard deviation of 12.5 years. The average age of node-positive breast cancer patients was 61.42 years, and the average age of node-negative breast cancer patients was 64.3 years. For the 453 cases in the dataset, the event (cancer-related death) was observed for 374 cases, and the remaining cases were either censored or contained cases due to non-cancer-related deaths. Figure 1 illustrates the procedures performed on these 453 patients. The cancer survival time in months for each case was computed from the date of diagnosis to the date of death or the date of data extraction (when the death event was not observed/censored). 26 patients did not have any procedures performed on them. These patients were either transient patients who visited the hospital for a second opinion before moving on to another care facility or were very near death, where no procedures were performed. These 26 cases were dropped from further data analysis. Two of these dropped cases were node-positive breast cancer cases. The dataset contained other attributes, such as estrogen and progesterone receptor status. These variables were only considered in secondary data analysis because, due to missing values of these variables, using them led to a further reduction in the original dataset size and an increase in binary input combinations.

Figure 1 illustrates that hormone therapy and radiation therapy were the two most frequently used procedures, which were followed by surgery and chemotherapy. Verifying that the overall procedure total exceeds 427 remaining patients is easy. This illustrates that several patients had undergone multiple procedures. The treatment procedure variables in the dataset were assigned binary values, where one indicated that the procedure was performed on the patient and 0 suggested that the procedure was not performed on the patient. Since there are four treatment variables, each taking a binary value, 16 unique treatments exist. Of these 16 unique treatments, 26 cases (mentioned earlier) taking all 0 values for all four procedures were already dropped from the analysis. For the remaining 15 combinations, a mapping that uniquely assigns an integer value for each of the 15 remaining treatment combinations was used. More specifically, the mapping: $f : \{0, 1\}^4 \to \mathbb{Z}$, where integer set $\mathbb{Z} = \{1,2,3,\ldots,15\}$ was used. This mapping was created by assuming that the treatment combinations are defined using a binary component vector $q = [q_1, \ldots, q_4]^{\mathrm{T}}$. The unique value of $\mathbb{Z}$ was then computed using the following expression:

$$\mathbb{Z} = \sum\nolimits_{i=1}^{4} q_i \times 2^{i-1}, \text{ where } \sum\nolimits_{i=1}^{4} q_i > 0 \tag{3.1}$$

Once the $\mathbb{Z}$ values were obtained, samples with sizes of five or fewer cases per $\mathbb{Z}$ value were dropped because computing meaningful survival curves for small sample sizes was challenging to draw and interpret. These samples were for $\mathbb{Z} = 10$ (5 cases), $\mathbb{Z} = 11$ (2 cases), $\mathbb{Z} = 14$ (2 cases) and $\mathbb{Z} = 15$ (1 case). After removing these 10 cases, 417 cases were available for analysis. Table 1 lists the $\mathbb{Z}$ values and number of cases for each category of $\mathbb{Z}$ value. The dataset also provided the size of cancer in centimeters. For the 417 cases, the overall cancer size was 3.03 centimeters, with a standard deviation of 1.87 centimeters. The average cancer size for node-positive breast cancer patients was 2.98 centimeters, and the average cancer size for node-negative breast cancer patients was 3.04 centimeters.

The results of Table 1 indicate that hormone and radiation therapy combination is the most frequently used combination,

Figure 1
Procedure distribution histogram

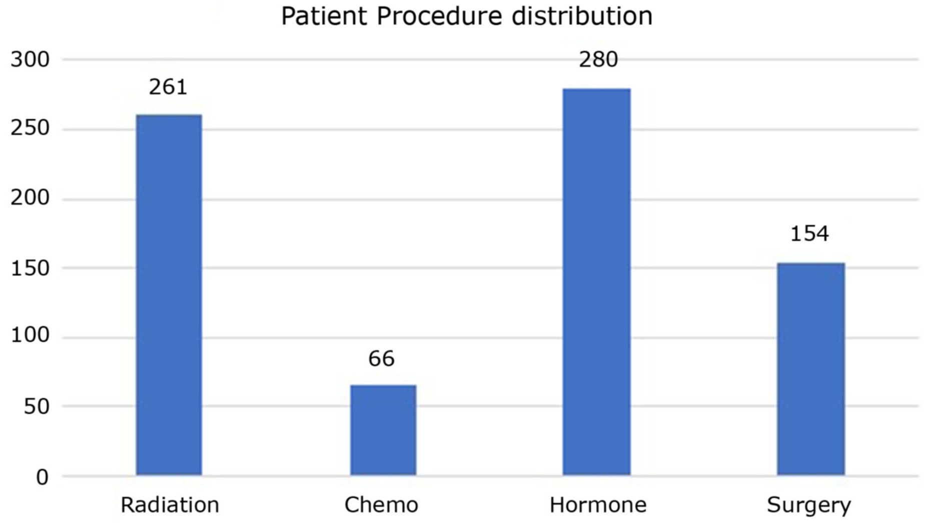


Table 1
Therapy combinations, ℤ Values, and total cases for each ℤ value

Hormone	Chemo.	Radi.	Surgery	ℤ	Total cases
1	0	0	0	1	62
0	1	0	0	2	7
1	1	0	0	3	9
0	0	1	0	4	46
1	0	1	0	5	109
0	1	1	0	6	14
1	1	1	0	7	26
0	0	0	1	8	38
1	0	0	1	9	43
0	0	1	1	12	35
1	0	1	1	13	28

Table 2
Number of cases breakdown by cancer type (Node = 1 is node-positive)

ℤ value	Node = 0	Node = 1
1	52	10
2	1	6
3	3	6
4	15	1
5	93	16
6	9	5
7	16	10
8	33	5
9	34	9
12	32	3
13	23	5

which is followed by hormone-only and radiation-only therapy procedures. Chemotherapy and its combinations are among the least commonly used methods. In the 417 cases used for data analysis, node-positive breast cancer cases were 76 cases, representing about 18.2% of overall cases. The rest of the cases were non-negative breast cancer cases. Table 2 illustrates the breakdown in the number of cases based on cancer type. Radiation and hormone therapy combination remains the most popular treatment option for both types of cancers. However, radiation alone, while a popular procedure for node-negative cancer, is rarely used for node-positive cancers.

Figure 2 illustrates the LT survival analysis and plots of survival functions for node-positive and node-negative breast cancer cases. The SPSS software was used for survival analysis. The median survival times for node-negative and node-positive cancers were 58.57 months and 68.14 months, respectively. The Wilcoxon statistic was 3.262 ($df = 1$), which is non-significant at a 95% statistical level of confidence. The reader needs to notice that node-positive and node-negative breast cancers are not mutually exclusive cancers. Cases with categories labeled as node-positive cancers were previously node-negative breast cancers. A patient with node-positive breast cancer needs to survive as node-negative breast cancer for a while until the cancer spreads and becomes node-positive breast cancer. As a result, care must be exercised in viewing results from Table 2 and Figure 2. Many treatments that show up for node-positive breast cancer in Table 2 may be administered when the patient is in the node-negative breast cancer category. Furthermore, some node-negative breast cancers in the dataset may become node-positive breast cancers in the future.

Given that the survival times for node-negative and node-positive breast cancers were not statistically significant, two different KM survival analyses were performed. In the first analysis, cancer type was ignored, and all cases were treated as overall breast cancer cases. Figure 3 illustrates the survival curves for different treatment plans for this first analysis. Censored cases in Figure 3 are represented by "+" symbol. This censoring label appears in Figures 4 and 5 as well.

Figure 3 and Table 3 illustrate that treatment plans 8, 9, 12, and 13 have the highest survival rates. All of these plans use surgery only or some other treatment in combination with surgery. Chemotherapy

Figure 2
Survival functions for node-positive and node-negative breast cancers

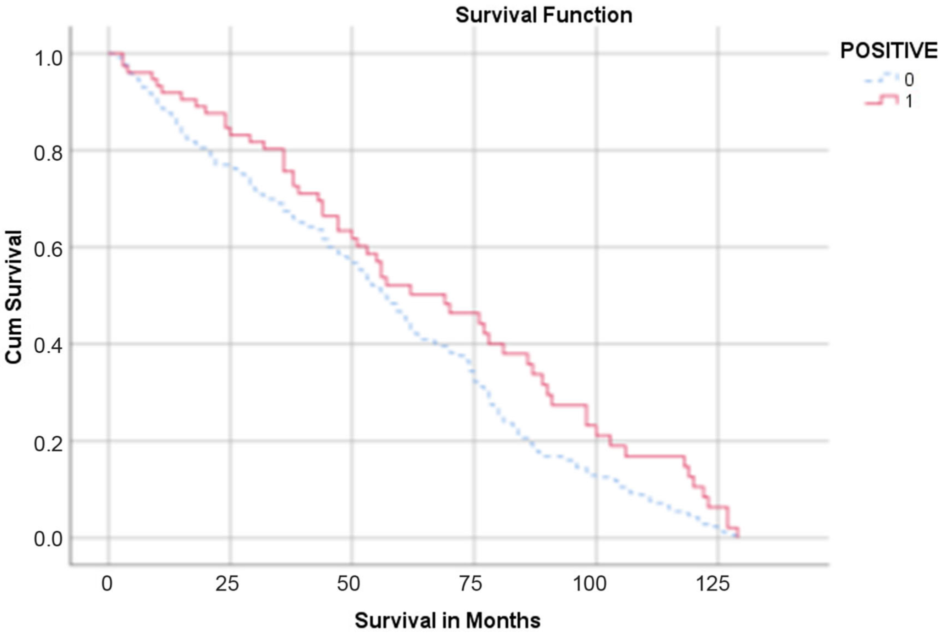
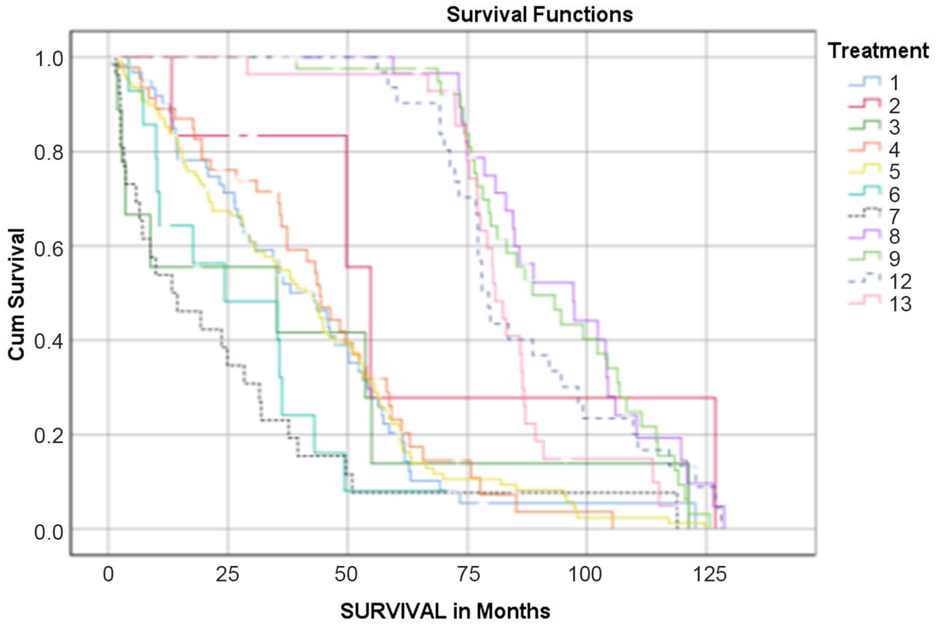


Figure 3
Kaplan-Meier analysis of survival functions



(plan 2) also appears promising, but based on Table 2, our dataset did not have many cases related to this plan. Plan number 7, which was a combination of hormones, chemotherapy, and radiation, appears to be the worst option in terms of survival rates. Plan 5, a combination of radiation and hormone therapy, seems to have average survival rates. There are two options to compute the central tendency of survival times. These two options are median and mean survival times [23]. The median survival time is the smallest survival time for which the survivor function is less than or equal to 0.5. Generally, median survival time computation is desirable when the sample size is large [23]. For smaller sample sizes, some survival functions may not go as far as the value of 0.5. In such cases, median survival time is usually not computed. The mean survival times [24] are computed as an expected value of survival time using the area under the entire survival curve [25]. The mean survival times assume that the longest survival time is the longest

Table 3
Kaplan-Meier analysis mean survival times and confidence intervals

Treatment plan	Mean	Std. Error	95% lower bound	95% upper bound
1	42.12	3.71	34.83	49.40
2	66.48	22.57	22.23	110.72
3	38.71	14.67	9.96	67.46
4	44.87	3.78	37.45	52.29
5	41.85	2.75	36.45	47.25
6	28.20	5.25	17.90	38.51
7	25.39	6.16	13.30	37.48
8	95.42	3.69	88.18	102.66
9	93.07	3.29	86.69	99.53
12	87.49	3.88	79.89	95.10
13	83.85	3.32	77.33	90.37

Table 4
Statistical significance results on equality of survival distributions for different levels of treatments

Method	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	188.122	10	0.000*
Breslow (Generalized Wilcoxon)	216.157	10	0.000*
Tarone-Ware	217.481	10	0.000*

Table 5
Number of node-positive and node-negative cancer cases for each treatment plan

Treatment plan ($\mathbb{Z}$)	Number of cases (Node = 0)	Number of cases (Node = 1)
1	52	10
2	1*	6
3	3*	6
4	45	1*
5	93	16
6	9	5*
7	16	10
8	33	5*
9	34	9
12	32	3*
13	33	5*

*Too few cases. Treatment plans related to these cases are suppressed in survival function plots

for a patient in the dataset [23]. Table 3 provides numeric estimates of mean survival times and 95% confidence levels around these means for different treatment options.

Table 4 illustrates the test results of the null hypothesis that all treatment plans have the same survival distribution. This null hypothesis was rejected, emphasizing that selecting a treatment plan does improve breast cancer survival.

The second KM survival analysis stratified data for computing survival rates using the cancer type variable. Table 5 illustrates the number of cases of each cancer type and treatment plan. To avoid clutter in survival function plots, survival function plots that contained fewer than 6 cases for each cancer type and treatment plan combination were suppressed. These cases with suppressed survival function plots are marked with an asterisk in Table 5.

Figures 4 and 5 illustrate survival function plots for node-negative and node-positive breast cancers. Table 6 provides numeric estimates of mean survival rates and 95% confidence levels around the means. Treatment plans with surgery (plans 8, 9, 12, and 13) all have higher survival rates for node-negative breast cancer cases. Treatment plans containing radiation and chemotherapy (plans 6 and 7) have the lowest survival rates for both node-negative and node-positive breast cancers. Chemotherapy should be generally administered for node-positive breast cancers only. While treatment plan 1 (Hormone therapy only) has low survival rates for node-negative cancers, its pronounced impact in Figure 5 appears to suggest that hormone therapy may work well for certain patients. Perhaps other confounding variables, such as the age of the patient, may play a role in deciding on a hormone-only therapy option. Table 7 illustrates the test results of the null hypothesis that all treatment plans have the same survival distribution adjusted for different breast cancer types. The null hypothesis was rejected, illustrating that the selection of a treatment plan improves the chances of breast cancer survival.

Estrogen and progesterone receptor status were included in the dataset for additional analysis. Including these two variables reduced the database size to 367 cases for analysis. This data size reduction was due to eliminating cases containing missing values for estrogen and progesterone receptor status variables. A selection choice between estrogen and progesterone receptor status variables was made to avoid a further reduction in sample sizes of cases belonging to different combinations of estrogen and progesterone receptor status variables and to improve the generalizability of the results. The status variable that best splits the database using the entropy criterion [26] was selected. This best separating variable and associated rule separating 114 cases in the database was:

$$\begin{aligned}&\mathit{IF\ Estrogen\ Receptor\ Status} = \mathit{Positive\ Then\ Treatment}\\&\mathit{Plan} = 5(98\ \mathit{total\ cases})\\&\mathit{Else}\\&\mathit{Treatment\ Plan} = 4(16\ \mathit{cases}).\end{aligned}$$

Since radiation was administered in both plan 4 and 5 and hormone therapy was only administered in plan 5, it appears that estrogen receptor status is the key determining factor in hormone

Figure 4
Survival functions for node-negative breast cancer

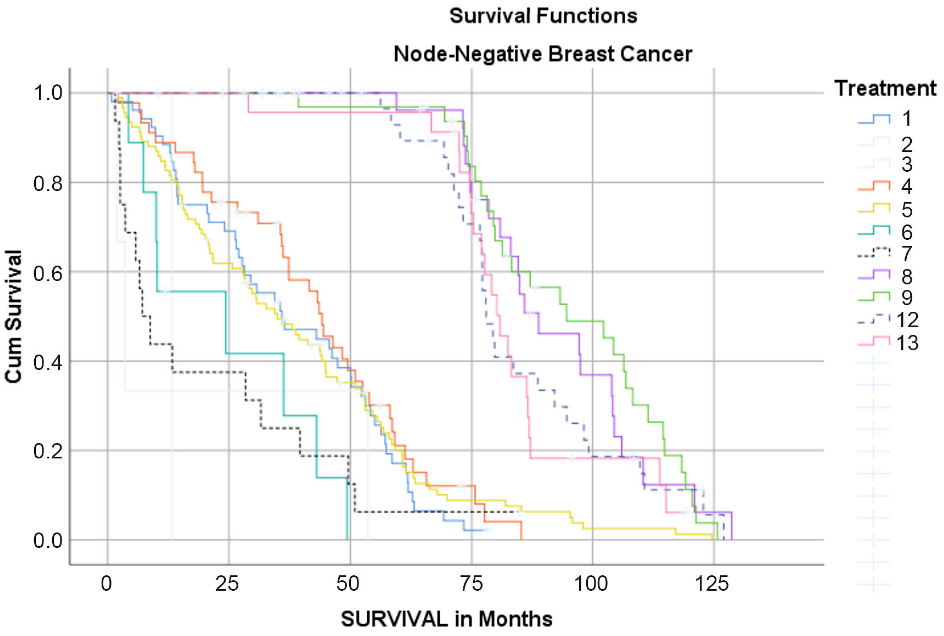
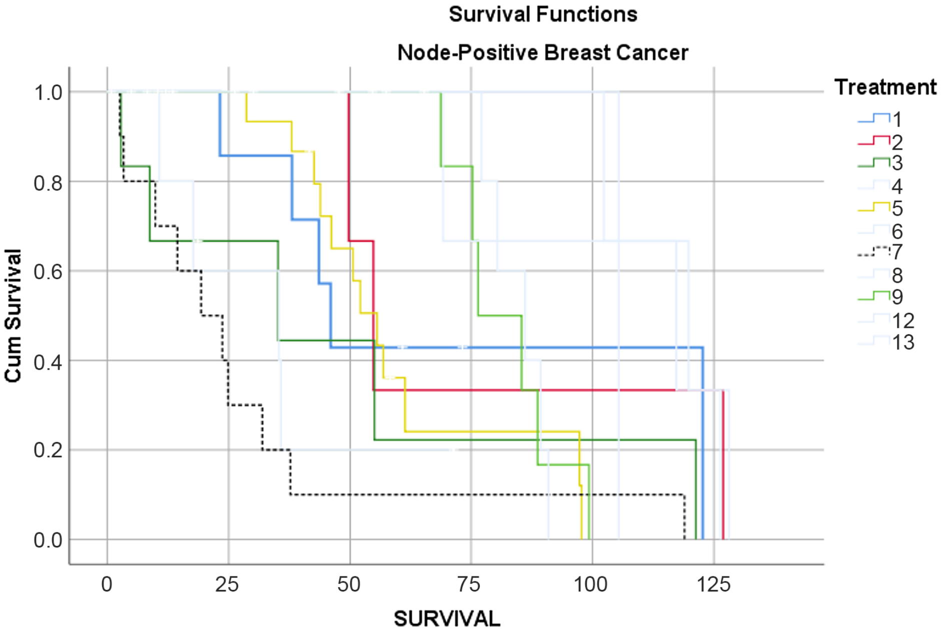


Figure 5
Survival functions for node-positive breast cancer



therapy treatment. Table 8 illustrates average survival rates. The table demonstrates that positive estrogen status only improves survival rates for treatment plan 13. For all other treatment plans, higher average survival rates were observed when estrogen receptor status was negative. The analysis indicates that using estrogen receptor status to separate different treatment plans may be beneficial. Except for treatment plans 4 and 5, our sample sizes for other treatment plans were too small to draw reliable conclusions. The estrogen receptor status variable certainly has merit in determining treatment plans.

A Cox regression analysis with independent variables of age, cancer size in centimeters, breast cancer type, and 11 different treatment plans, with the dependent variable of survival time, was conducted in months. Table 9 illustrates the results of variable significance for Cox regression analysis. The base-level treatment plan was $\mathbb{Z} = 13$. The overall Omnibus test of the model was significant with a −2 log-likelihood value of 3319.50 (chi-square: 210.26 at df = 13), which was significant at a 99% level of statistical significance. The results indicate that cancer size, type, and treatment plan are important in cancer hazard rates. An inverse relationship exists between cancer size and cancer type with hazard rate. Larger-size cancers and node-positive cancers reduce the hazard rate. The relationship between lower hazard rates for node-positive breast cancer is consistent with higher survival rates observed in Figure 2. The relationship between cancer size and hazard rate may not be very intuitive, but many factors may play a role in explaining this relationship. For example, it appears that large cancers may be easy to detect, treatment plans for large cancers may be more aggressive, and surgical treatment plans may be recommended for large-size cancers.

Table 10 illustrates Helmert contrast statistics between different treatment plans. Helmert contrasts the hazard rates of each treatment plan with the average effects of previous categories. For example, if u denotes an index for a treatment plan with u taking a value of 1 when $\mathbb{Z} = 1$ and u taking a value of 11 when $\mathbb{Z} = 13$ and μ_u denotes mean survival rate for a treatment plan with index u, then Helmert contrasts test following ten null hypotheses: $\left(1-\frac{1}{u}\right)\mu_u - \left(\frac{1}{u}\sum_{q=1}^{u-1}\mu_q\right) = 0$, where $u = 2, \ldots, 11$. The last column in Table 10 represents the hazard rates. Generally, lower values represent lower hazard rates for the treatment plan. The best treatment plan has a u-value of 8 and $\mathbb{Z} = 8$. All treatment plans with surgery as an option have statistically significantly lower hazard rates than average combinations of non-surgical and/or surgical and non-surgical options. For the non-surgical option, hormone therapy is best. Chemotherapy and radiation therapy options seem to be the worst options.

Machine learning techniques have limited usefulness in determining the efficacy of breast cancer treatment procedures, but these techniques can still help predict death and cancer survival times (durations) [3]. Among different available methods, neural networks are an attractive option for testing the performance of a machine learning technique in predicting death and survival duration. A C++ program and IBM SPSS Modeler were used for this part of the research. The C++ program was used to generate samples for V-fold validation, and IBM SPSS software was used to run neural network experiments on these samples.

Table 6
Kaplan-Meier analysis mean survival times and confidence intervals

Treatment plan	Mean	Std. Error	95% lower bound	95% upper bound
Node-Negative Breast Cancer				
1	37.69	2.90	31.99	43.38
2	13.27	0	13.27	13.27
3	19.71	16.96	0	52.97
4	43.02	3.39	36.36	49.67
5	38.88	2.93	33.13	44.62
6	24.78	6.05	12.91	36.65
7	21.21	5.88	9.68	32.74
8	92.71	3.74	85.37	100.04
9	95.26	3.79	87.81	102.70
12	85.53	3.78	78.12	92.95
13	83.67	4.03	75.75	91.58
Node-Positive Breast Cancer				
1	74.13	17.99	38.86	109.40
2	77.12	24.93	28.25	125.98
3	48.91	21.10	7.56	90.27
4	105.40	0	105.40	105.40
5	60.40	6.67	47.32	73.49
6	34.19	9.40	15.75	52.63
7	28.68	10.67	7.77	49.58
8	116.15	7.16	102.10	130.20
9	82.31	4.49	73.50	91.11
12	104.85	18.08	69.41	140.30
13	84.78	2.63	79.63	89.82

Table 7
Statistical significant results on equality of survival distributions for different levels of treatments (adjusted for breast cancer type)

Method	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	209.60	10	0.000*
Breslow (Generalized Wilcoxon)	219.40	10	0.000*
Tarone-Ware	238.64	10	0.000*

*Significant at 99% statistical level of significance

Table 8
Survival rates based on estrogen receptor status

$\mathbb{Z}$ value	Negative	Positive
1	44.75	34.53
2	34.83	0.83
3	88.13	12.43
4	44.75	41.60
5	47.84	37.46
6	31.68	20.85
7	22.05	21.96
8	80.09	70.31
9	93.98	83.41
12	93.13	76.94
13	64.85	83.73

Table 9
Variable significance test results in Cox regression analysis

Variable	Beta value	Std. Error	Wald	df	Sig.
Age	0.004	0.005	0.73	1	0.391
Size	−0.079	0.031	6.64	1	0.010**
Cancer Type	−0.796	0.168	22.37	1	0.000*
Treatment			182.0	10	0.000*
Treatment 1	1.448	0.253	32.82	1	0.000*
Treatment 2	0.639	0.579	1.21	1	0.270
Treatment 3	1.585	0.438	13.12	1	0.000*
Treatment 4	1.226	0.264	21.55	1	0.000*
Treatment 5	1.272	0.229	30.77	1	0.000*
Treatment 6	2.307	0.379	37.00	1	0.000*
Treatment 7	2.137	0.306	48.79	1	0.000*
Treatment 8	−0.521	0.290	3.23	1	0.072
Treatment 9	−0.331	0.265	1.56	1	0.212
Treatment 12	−0.210	0.279	0.56	1	0.452

*Significant at 99% statistical level of significance; **Significant at 95% statistical level of significance

Table 10
Helmert contrast statistics for different treatment plans

u-index	Beta value	Std. Error	Wald stat.	df	Sig.	Exp (beta)
2	−0.81	0.558	2.105	1	0.147	0.445
3	0.542	0.450	1.447	1	0.229	1.719
4	0.002	0.287	0.000	1	0.995	1.002
5	0.047	0.208	0.052	1	0.820	1.049
6	1.073	0.329	10.639	1	0.001*	2.925
7	0.724	0.239	9.218	1	0.002*	2.063
8	−2.03	0.254	64.382	1	0.000*	0.130
9	−1.59	0.217	53.727	1	0.000*	0.203
10	−1.29	0.215	36.140	1	0.000*	0.274
11	−0.95	0.227	17.699	1	0.000*	0.385

*Significant at 99% statistical level of significance

Figure 6
A 7-hidden node neural network architecture for predicting events and survival time

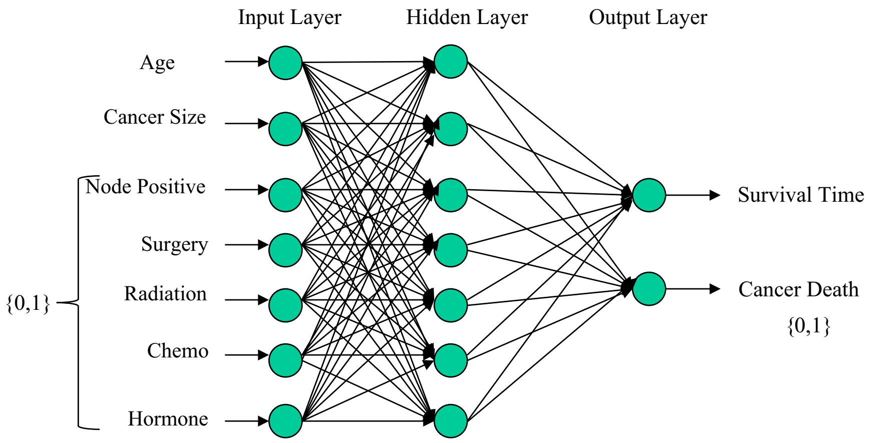


Figure 6 illustrates a neural network used in this research. The inputs in the neural network were patient age, cancer size, cancer type (node-positive or node-negative), and different treatment procedures. The outputs were cancer survival time and cancer death. Figure 6 indicates the variables assigned binary values using label {0,1}. All cases where the treatments suggested by inputs were administered, cancer was node-positive, or an event (death) was observed were assigned a value of 1. Otherwise, a value of 0 was assigned to the variables. Cancer death was assigned a value of 1 only when death was attributed to breast cancer. For all other possibilities, such as deaths unrelated to cancer, censored cases, or when the event of death was not observed, a value of 0 was assigned to the cancer death variable.

Among the factors that impact the performance of a neural network are the design of a neural network and data bias [26]. The neural network design factors consist of the learning algorithm used to learn connection weights, the architecture of a neural network, and the neural network learning parameters (learning rate and stopping criterion) [27]. Three-layer neural network architectures with either seven hidden nodes or 14 hidden nodes and two output nodes were chosen for the experiments. The learning algorithm used in the research was error backpropagation with a learning rate of 0.01 and a stopping criterion of 5,000 learning iterations or convergence of error values. Figure 6 illustrates seven hidden node neural network architectures used in current research.

A brief overview of neural network learning is as follows. The input neurons are assigned actual case values of inputs. For each non-input layer of neurons, assume that there are *i* neurons in the previous layer and k neurons in the current layer. For any of the hidden layer or output layer neurons, say n_k, their output (o_k) is represented using the following expression: $o_k = f\left(\left(\sum_i w_{ik} p_i\right) + w_{i+1,k}\right)$,

$$\text{where } f(x) = \frac{1}{1+e^{-x}} \tag{3.1}$$

At each iteration, the connection weights between layers *i* and *k* are adjusted using the following formula:

$$w_{ik}^{new} = w_{ik}^{old} - \eta \delta_k \, p_i \tag{3.2}$$

The weights w_{ik}^{new} are the new values of weights and w_{ik}^{old} are old values from the previous iteration. The value of δ_k is determined as follows:

$$\delta_k = \begin{cases} o_k(1-o_k)(y_k - o_k) \ \textit{if } n_k \textit{ is an output neuron} \\ o_k(1-o_k) \ \sum_j w_{kj}\delta_j \ \textit{if } n_k \textit{ is hidden layer neuron.} \end{cases} \tag{3.3}$$

The subscript *j* denotes the number of neurons in the output layer, and δ_j is computed only for the output neurons using the first part of the Equation (3.3). The variables y_k are actual values of outputs that a neural network attempts to learn by applying weight adjustments.

The parameter η is the learning rate. The weights are assigned random values at the beginning of the learning procedure. The procedure terminates when weights converge or some predetermined number of iterations are completed (5,000 in our case).

A V-fold holdout sampling approach was used, where the original data set of 417 cases was broken into five nearly equal-sized samples [26]. Five different holdout experiments were conducted using each of the five approximately equal-sized samples as holdout datasets and the remaining four training samples as the training dataset. For two different types of outputs, two different performance metrics were used. The first metric used for survival time output was the root mean square (RMS) error. Several performance metrics were necessary to account for dataset bias for the cancer death output. Out of a total of 417 cases, 352 cases contained the event of cancer death. That is approximately 85% of cases where the event of death was observed. Thus, if a neural network predicted the event of death for all 100% of cases, then it would be approximately accurate 85% of the time. To deal with this data bias issue, several other performance metrics are necessary to assess the predictive performance of a neural network appropriately. These performance metrics were computed using the confusion matrix shown in Figure 7. A confusion matrix is created by comparing the prediction for cancer deaths from a neural network with their known actual values from the dataset. The variables *a*, *b*, *c*, and *d* are integer values. For example, the variable a represents total positives, where the neural network correctly predicts an event death for holdout sample cases, etc.

Table 11 illustrates the results of neural network experiments. For cancer death output and its metrics, the accuracies of the two neural network architectures look similar. Sometimes, the 7-hidden nodes neural network performs slightly better, and other times, the 14-hidden nodes neural network performs better. The F-measure that combines precision and recall into one measure slightly appears to favor 7-hidden node architecture. Assuming that decision-makers are trying to predict survival from breast cancer, the specificity metric is the most critical. For this measure, the 7-hidden node architecture performs slightly better in a higher average value for the specificity metric. The 7-hidden node architecture consistently provides lower RMS errors for the survival time output. So overall, the 7-hidden node architecture is slightly better than the 14-node architecture.

Table 11 also includes a reliability column. This reliability column only looks at the ratio between training and test accuracy (reported in the 3rd column of Table 10). Since the training dataset contains more examples than the test dataset, training dataset accuracies were consistently lower than test datasets but higher than 85%. We do not report training dataset accuracies because they can be computed by multiplying the second column of Table 11 with the third column. The ideal value of the reliability column will be 1, where both training and test dataset accuracies are precisely equal. A general expectation of reliability is that test

Figure 7
The cancer death class performance metrics

		Predicted Class	
		1	0
Actual Class	1	a	b
	0	c	d

$$\boldsymbol{Accuracy} = \frac{a+d}{a+b+c+d}$$

$$\boldsymbol{Specificity} = \frac{d}{d+c}$$

$$\boldsymbol{Precision\ (p)} = \frac{a}{a+c}$$

$$\boldsymbol{Recall\ (r)} = \frac{a}{a+b}$$

$$\boldsymbol{F\text{ - }measure\ (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

Table 11
Summary of neural network experiments

Fold	Rel.	Acc. %	Recall %	Spec. %	Prec.%	RMS Error
7-Hidden Node Architecture						
1	0.86	98.77	98.55	0	85.00	27.64
2	0.95	90.48	91.67	16.67	86.84	31.29
3	0.94	44.05	93.94	5.56	78.48	29.14
4	0.86	97.62	98.67	11.11	90.24	26.37
5	0.95	90.48	92.86	21.43	85.53	30.11
14-Hidden Node Architecture						
1	0.93	95.08	95.65	8.33	85.71	30.27
2	0.97	90.48	90.28	8.33	85.53	33.60
3	0.93	96.43	95.45	0	77.78	29.33
4	0.89	96.43	97.33	11.11	90.12	27.96
5	0.96	91.67	92.86	14.29	84.42	32.48

performance will also worsen if training accuracies deteriorate. The reliability measure indicates that the 14-node architecture is slightly better. For large dataset sizes, decision-makers may prefer this architecture because it will slightly adjust well for large datasets.

The results of experiments with neural networks indicate that these techniques add marginal value because the prediction accuracies for death output are higher than 90%, which is higher than the average death cases in the dataset at nearly 85%. The survival time predictions have high RMS errors. The errors are so high that most decision-makers may not want to use these predictions in their patient recommendations. Using these predictions can have an error of two or more years.

While this research used traditional neural networks, advanced hybrids that combine global and local search can improve predictive performance [27]. When available, incorporating additional relevant variables in the dataset may also improve predictive performance.

4. Summary, Discussion, and Conclusions

Given that it is rare for breast treatment plans to consider only a single treatment, this study investigated survival analysis for different treatment plane combinations for breast cancer treatment. A simple binary function was introduced to map different treatment plans into a unique category integer value. Statistical survival analysis was then conducted on other treatment plans, and it was found that surgery and its combination with hormone therapy led to the highest survival rates. For node-positive cancers, radiation was not the best option to consider.

Chemotherapy was a confounding cancer treatment option. At first, it did not appear to improve survival rates. However, for node-positive cancer cases, chemotherapy may be considered to improve survival, albeit only after a surgical option is considered. Chemotherapy was also the least prescribed option in the dataset (Figure 1). The small sample size for chemotherapy patients reduces the significance of conclusions that may be drawn for this treatment option. The results of the study also suggest that the estrogen receptor status variable may play a role in determining if a combination of hormonal therapy and radiation therapy should be used with a surgery option.

The experiments with neural networks suggest the importance of survival analysis for predicting breast cancer survival rates. Neural networks helped predict the event of death, but these techniques could not accurately predict the length of survival. It is possible that increasing the number of outputs in a neural network leads to a sacrifice in accuracy for some outputs. Part of the reason may be that the weights for the first layer of connections are learned by considering both outputs. Perhaps two separate networks, each with one output, may provide for marginal improvement in prediction accuracy. Future research is needed to address this issue.

The current research only used demographics, cancer size, and cancer treatment variables. Cancer diagnostic procedures have improved substantially. In particular, minimal invasion biopsies are very common nowadays. These biopsies provide information on tissue trauma, tissue inflammation, genetic alteration, and chromosomal instability information [28]. Such additional information may be used to improve optimal treatment procedure recommendations. When biopsy information is available, researchers may be overwhelmed with much available information. In such situations, dimensionality reduction and special sampling approaches helped screen samples and essential variables for cancer prediction [29]. Ensemble techniques are also beneficial in cancer research [30] to improve prediction accuracies. The use of dimensionality reduction and ensemble techniques represents another possible extension of the current study.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this work.

Data Availability Statement

The data that support the findings of the study are openly available in GitHub at https://github.com/Parag8219/BreastCancerSurvival/blob/main/GHdata.csv.

Author Contribution Statement

Parag C. Pendharkar: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **James A. Rodger:** Investigation, Resources, Data curation.

References

[1] Bilani, N., Zabor, E. C., Elson, L., Elimimian, E. B., & Nahleh, Z. (2020). Breast cancer in the United States: A cross-sectional overview. *Journal of Cancer Epidemiology*, *2020*(1), 6387378.

[2] Sharma, A., Debik, J., Naume, B., Ohnstad, H. O., Bathen, T. F., & Gisjeodegard, G. F. (2024). Comprehensive multi-omics analysis of breast cancer reveals distinct long-term prognostic subtypes. *Oncogenesis*, *13*(1), 1–13.

[3] Potta, M., Narayanan, B., & Balmuri, K. R. (2024). A review of breast cancer prediction using machine learning and deep learning techniques. *Turkish Journal of Computer and Mathematics Education*, *15*(3), 1–20.

[4] Abadi, A., Yavari, P., Dehghani-Arani, M., Alavi-Majd, H., Ghasemi, E., Amanpour, F., & Bajdik, C. (2014). Cox models survival analysis based on breast cancer treatments. *Iranian Journal of Cancer Prevention*, *7*(3), 124–129.

[5] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, *8*, 150360–150376.

[6] Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in Medicine*, *18*(3), 205–219.

[7] Clark, R. M., Whelan, T., Levine, M., Roberts, R., Willan, A., McCulloch, P., . . . , & Mahoney, L. J. (1996). Randomized clinical trial of breast irradiation following lumpectomy and axillary dissection for node-negative breast cancer: An update. *Journal of National Cancer Institute*, *88*(22), 1659–1664.

[8] Elkefi, S., & Matthews, A. K. (2024). Exploring health information—Seeking behavior and information source preferences among a diverse sample of cancer survivors: Implications for patient education. *Journal of Cancer Education*, *39*(6), 650–662.

[9] Ponnaganti, N. D., & Anitha, R. (2023). Enhancement of classification and prediction accuracy for breast cancer

detection using fast convolution neural network with ensemble algorithm. *International Journal of Computational Science and Engineering*, *26*(2), 171–181.

[10] Markey, M. A., Lo, J. Y., Tourassi, G. D., & Floyd, C. E. (2003). Self-organizing map for cluster analysis of a breast cancer database. *Artificial Intelligence in Medicine*, *27*(2), 113–127.

[11] Chander, G. P., & Das, S. (2025). A hybrid decision support system in medical emergencies using artificial neural network and hyperbolic secant grey wolf optimization techniques. *Cluster Computing*, *28*(1), 43.

[12] Mahapatra, A. K., Panda, N., Mahapatra, M., Jena, T., & Mohanty, A. K. (2025). A fast-flying partical swarm optimization for resolving constrained optimization and feature selection problems. *Cluster Computing*, *28*(2), 91.

[13] Barno, S. (2024). A fuzzy approach to breast cancer diagnosis in poorly formalized processes. *AIP Conference Proceedings*, *3244*(1), 030067.

[14] Pena-Reyes, C. A., & Sipper, M. (1999). A fuzzy-genetic approach to breast cancer diagnosis. *Artificial Intelligence in Medicine*, *17*(2), 131–155.

[15] Tagnamas, J., Ramadan, H., Yahyaouy, A., & Tairi, H. (2024). Multi-task approach based on combined CNN-transformer for effcient segmentation and classification of breast tumors in ultrasound images. *Visual Computing for Industry, Biomedicine, and Art*, *7*(1), 2.

[16] Chiu, S. H., Chen, C. C., & Lin, T. H. (2008). Using support vector regression to model the correlation between the clinical metastases time and gene expression profile for breast cancer. *Artificial Intelligence in Medicine*, *44*(3), 221–231.

[17] Karim, S. A., Mohamad, U. H., & Nohuddin, P. N. (2023). Feature selection techniques on breast cancer classification using fine needle aspiration features: A comparative study. In *International Visual Informatics Conference*, 568–582.

[18] Krawczyk, B., Schaefer, G., & Wozniak, M. (2015). A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. *Artificial Intelligence in Medicine*, *65*(3), 219–227.

[19] Jerez, J. M., Molina, I., Garcia-Laencina, P. J., Alba, E., Ribelles, N., Martin, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), 105–115.

[20] Gandomkar, Z., Brennan, P. C., & Mello-Thoms, C. (2018). MuDeRN: Multi-category classification of breast histopathological image using deep residual networks. *Artificial Intelligence in Medicine*, *88*, 14–24.

[21] Gu, D., Liang, C., & Zhao, H. (2017). A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis. *Artificial Intelligence in Medicine*, *77*, 31–47.

[22] Lamy, J. B., Sekar, B., Guezennec, G., Bouaud, J., & Seroussi, B. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, *94*, 42–53.

[23] Hanson, R. K. (2022). Kaplan-Meier survival analysis. In R. K. Hanson (Ed.), *Prediction statistics for psychological assessment* (pp. 81–97). American Psychological Association.

[24] Mohammad, P., & Khan, A. L. (2024). Bayesian extension of the Weibull AFT shared fraility model with generalized family of distributions for enhanced survival analysis using censored data. *Journal of Applied Statistics*, *51*(15), 3125–3153.

[25] Hanson, R. K. (2022). Discrete-time survival analysis. In R. K. Hanson (Ed.), *Prediction statistics for psychological assessment* (pp. 63–80). American Psychological Association.

[26] Jamsa, K. (2021). *Introduction to data mining and analytics*. USA: Jones & Barlett Learning.

[27] Pendharkar, P. C. (2001). An empirical study of design and testing of hybrid evolutionary-neural approach for classification. *Omega: An International Journal of Management Science*, *29*(4), 361–374.

[28] Franzen, B., Auer, G., & Lewensohn, R. (2024). Minimally invasive biopsy-based diagnostics in support of precision cancer medicine. *Molecular Oncology*, *18*(11), 2612–2628.

[29] Jopek, M. A., Pastuszak, K., Sieczczynski, M., Cygert, S., Zaczek, A. J., Rondina, M. T., & Supernat, A. (2024). Improving platelet-RNA-based diagnostics: A comparative analysis of machine learning models for cancer detection and multiclass classification. *Molecular Oncology*, *18*(11), 2743–2754.

[30] Tang, P., Li, B., Zhou, Z., Wang, H., Ma, M., Gong, L., . . . , & Zhang, H. (2024). Integrated machine learning developed a prognosis-related gene signature to predict prognosis in oesophageal squamous cell carcinoma. *Journal of Cellular and Molecular Medicine*, *28*(21), e70171.

How to Cite: Pendharkar, P. C., & Rodger, J. A. (2025). Breast Cancer Survival Analysis and Mortality Prediction Under Different Treatment Combinations. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS52024464