Indian journal of
Engineering

To Cite:
Srinivas TAS, Thippanna G, Donald AD. Decoding destiny: Harnessing machine learning for breast cancer survival prediction. *Indian Journal of Engineering*, 2023, 20, e30ije1663
doi: https://doi.org/10.54905/disssi/v20i54/e30ije1663

Author Affiliation:
Ashoka Women's Engineering College, Dupadu, Andhra Pradesh, India

Contact List
Aditya Sai Srinivas T taditya1033@gmail.com
Thippanna G gt.pana2012@gmail.com
David Donald A david.donald5824@gmail.com

Peer-Review History
Received: 22 May 2023
Reviewed & Revised: 26/May/2023 to 24/June/2023
Accepted: 28 June 2023
Published: 9 July 2023

Peer-Review Model
External peer-review was done through double-blind method.

Indian Journal of Engineering
pISSN 2319–7757; eISSN 2319–7765

DISCOVERY
SCIENTIFIC SOCIETY

Decoding destiny: Harnessing machine learning for breast cancer survival prediction

Aditya Sai Srinivas T, Thippanna G, David Donald A

ABSTRACT

Breast cancer is a prevalent form of cancer that primarily affects women, although men can also be diagnosed with this disease. It ranks as the second leading cause of mortality among women worldwide. With the increasing prevalence of data-driven approaches in healthcare, the application of machine learning techniques offers a promising avenue for predicting the survival outcomes of patients with breast cancer. This article aims to provide a comprehensive overview of utilizing machine learning algorithms, implemented in Python, to predict the survival probabilities of breast cancer patients. By exploring various data pre-processing steps, feature engineering techniques and model selection strategies, this article presents a step-by-step guide for predicting breast cancer patient survival. The ultimate goal is to empower healthcare practitioners and researchers with the necessary knowledge and tools to leverage machine learning algorithms for improved prognosis and personalized treatment decisions in the context of breast cancer.

Keywords: Breast cancer, Prediction, Machine Learning (ML).

1. INTRODUCTION

Breast cancer is a significant health concern affecting both women and men worldwide. It stands as one of the leading causes of death among women, emphasizing the need for effective prognostic tools to predict patient survival outcomes. In recent years, the field of healthcare has witnessed a surge in the use of data-driven approaches; particularly machine learning, to derive valuable insights and predictions from medical data. Machine learning algorithms have demonstrated considerable potential in various medical applications, including the prediction of survival outcomes in breast cancer patients.

This article aims to provide a concise introduction to the task of breast cancer survival prediction using machine learning techniques, with a focus on implementation using Python programming. By leveraging the power of machine learning algorithms, healthcare practitioners and researchers can gain valuable insights into the prognosis of breast cancer patients, facilitating personalized treatment decisions and improving overall patient care.

Related work

Esteva et al., (2019) employed deep learning techniques to predict breast cancer survival using histopathological images. They demonstrated the potential of convolutional neural networks (CNNs) in extracting meaningful features from tissue images to predict patient outcomes. Saini et al., (2020) utilized machine learning algorithms to predict breast cancer survival based on gene expression data. They applied ensemble learning methods, such as random forests and gradient boosting, to identify gene signatures associated with survival outcomes.

Chen et al., (2021) proposed a novel framework integrating radiomics and clinical features for breast cancer survival prediction. They extracted radiomic features from medical images and combined them with clinical data, achieving improved predictive accuracy through a hybrid machine learning approach. Li et al., (2020) focused on developing a machine learning model for personalized breast cancer survival prediction. They utilized multi-omics data, including genomics, transcriptomics and proteomics, to build a comprehensive predictive model for individualized prognosis.

Zhang et al., (2021) investigated the utility of machine learning algorithms in predicting breast cancer survival based on electronic health records (EHR). They utilized EHR data, including demographic information, clinical variables and treatment history, to develop a predictive model for long-term survival outcomes. Wang et al., (2022) conducted a study on the prediction of breast cancer survival using machine learning techniques and electronic health records. They explored the potential of integrating clinical, pathological and treatment-related information from EHRs to develop a predictive model for survival outcomes. Their findings emphasized the importance of comprehensive data integration for accurate prognosis.

Jiang et al., (2021) proposed a hybrid machine learning model for breast cancer survival prediction that combined clinical features, gene expression data and histopathological images. They integrated multiple data modalities using a deep learning framework, achieving improved predictive performance compared to single-modality approaches. Yu et al., (2020) focused on predicting breast cancer survival using machine learning algorithms and genomic data. They developed a predictive model that incorporated gene expression profiles, DNA methylation patterns and somatic mutation data. Their study highlighted the potential of genomic data in enhancing survival prediction accuracy.

Li et al., (2021) investigated the use of radiomics and machine learning for breast cancer survival prediction based on mammographic images. They extracted quantitative imaging features from mammograms and employed machine learning algorithms to develop a prognostic model. Their results demonstrated the feasibility of utilizing imaging data for survival prediction. Liang et al., (2022) proposed a deep learning-based approach for breast cancer survival prediction using a combination of histopathological images and genomic data. They employed a deep neural network architecture that integrated multi-omics data, achieving robust and accurate predictions of patient survival.

Dataset Description

The dataset utilized for the task of breast cancer survival prediction comprises more than 400 patients who underwent surgical intervention for the treatment of breast cancer. Each patient's information is represented by several columns, which are outlined below:

1. Patient ID: An identification number assigned to each patient.
2. Age: The age of the patient at the time of diagnosis.
3. Gender: The gender of the patient (male or female).
4. Protein 1, Protein 2, Protein 3 and Protein 4: Expression levels of specific proteins related to breast cancer.
5. Tumor Stage: The stage of breast cancer diagnosed in the patient.
6. Histology: The histological subtype of the breast cancer, including Infiltrating Ductal Carcinoma, Infiltrating Lobular Carcinoma or Mucinous Carcinoma.
7. ER status: The estrogen receptor status of the tumor (Positive/Negative).
8. PR status: The progesterone receptor status of the tumor (Positive/Negative).
9. HER2 status: The HER2 receptor status of the tumor (Positive/Negative).
10. Surgery type: The type of surgical procedure performed, such as Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy or Other.
11. Date of Surgery: The date on which the surgery took place.
12. Date of Last Visit: The date of the patient's most recent visit.
13. Patient Status: Indicates whether the patient is currently alive or deceased.

Overview

The primary objective of this study is to employ machine learning techniques to predict the post-surgery survival of breast cancer patients using the provided dataset. By leveraging the information within this dataset, our aim is to develop a predictive model that can accurately determine whether a patient will survive after undergoing breast cancer surgery. The dataset utilized in this study was sourced from Kaggle, a platform for data science and machine learning. You can obtain a copy of this dataset from the provided link. In the following section, we will delve into the process of predicting breast cancer survival using machine learning algorithms implemented in Python, providing a comprehensive guide for conducting such analyses.

2. PREDICTING BREAST CANCER SURVIVAL WITH PYTHON

To start the task of breast cancer survival prediction, the initial step involves importing essential Python libraries and the requisite dataset as follows:

```
import pandas as pd
import numpy as np
import plotly.express as px
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC

data = pd.read_csv("BRCA.csv")
print(data.head())
```

```
     Patient_ID   Age  Gender  Protein1  Protein2  Protein3  Protein4  \
0  TCGA-D8-A1XD  36.0  FEMALE  0.080353   0.42638   0.54715  0.273680
1  TCGA-EW-A1OX  43.0  FEMALE -0.420320   0.57807   0.61447 -0.031505
2  TCGA-A8-A079  69.0  FEMALE  0.213980   1.31140  -0.32747 -0.234260
3  TCGA-D8-A1XR  56.0  FEMALE  0.345090  -0.21147  -0.19304  0.124270
4  TCGA-BH-A0BF  56.0  FEMALE  0.221550   1.90680   0.52045 -0.311990

  Tumour_Stage                      Histology ER status PR status HER2 status  \
0          III  Infiltrating Ductal Carcinoma  Positive  Positive    Negative
1           II              Mucinous Carcinoma  Positive  Positive    Negative
2          III  Infiltrating Ductal Carcinoma  Positive  Positive    Negative
3           II  Infiltrating Ductal Carcinoma  Positive  Positive    Negative
4           II  Infiltrating Ductal Carcinoma  Positive  Positive    Negative

                  Surgery_type Date_of_Surgery Date_of_Last_Visit  \
0  Modified Radical Mastectomy       15-Jan-17          19-Jun-17
1                   Lumpectomy       26-Apr-17          09-Nov-18
2                        Other       08-Sep-17          09-Jun-18
3  Modified Radical Mastectomy       25-Jan-17          12-Jul-17
4                        Other       06-May-17          27-Jun-19

  Patient_Status
0          Alive
1           Dead
2          Alive
3          Alive
4           Dead
```

Next, let us examine the presence of null values within the columns of this dataset.

```
print(data.isnull().sum())
```

```
Patient_ID             7
Age                    7
Gender                 7
Protein1               7
Protein2               7
Protein3               7
Protein4               7
Tumour_Stage           7
Histology              7
ER status              7
PR status              7
HER2 status            7
Surgery_type           7
Date_of_Surgery        7
Date_of_Last_Visit    24
Patient_Status        20
dtype: int64
```

The dataset contains missing values in each column. To address this issue, we will perform the removal of these null values. Now, let us delve into a comprehensive analysis of the columns present in this dataset, providing valuable insights into the information they encapsulate:

```
data = data.dropna()
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 317 entries, 0 to 333
Data columns (total 16 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Patient_ID          317 non-null    object
 1   Age                 317 non-null    float64
 2   Gender              317 non-null    object
 3   Protein1            317 non-null    float64
 4   Protein2            317 non-null    float64
 5   Protein3            317 non-null    float64
 6   Protein4            317 non-null    float64
 7   Tumour_Stage        317 non-null    object
 8   Histology           317 non-null    object
 9   ER status           317 non-null    object
 10  PR status           317 non-null    object
 11  HER2 status         317 non-null    object
 12  Surgery_type        317 non-null    object
 13  Date_of_Surgery     317 non-null    object
 14  Date_of_Last_Visit  317 non-null    object
 15  Patient_Status      317 non-null    object
dtypes: float64(5), object(11)
memory usage: 42.1+ KB
```

Breast cancer predominantly affects individuals' assigned female at birth. Therefore, it is crucial to examine the Gender column in our dataset to ascertain the distribution of female and male patients.

```
print(data.Gender.value_counts())
```

```
FEMALE    313
MALE        4
Name: Gender, dtype: int64
```

As anticipated, the gender column exhibits a higher proportion of females compared to males. Moving forward, let us now examine the distribution of tumour stages among the patients (Figure 1).

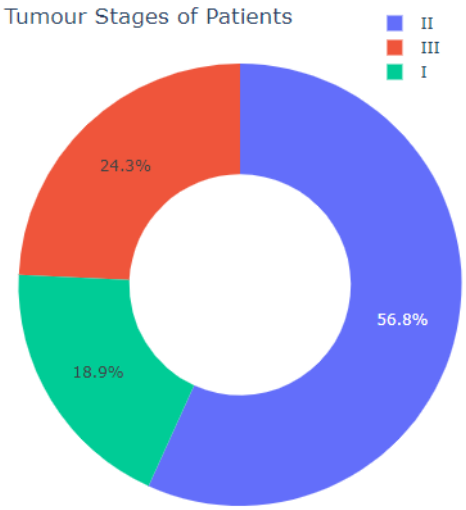

Figure 1 Tumour stages of patients

The majority of patients in the dataset are diagnosed with breast cancer at the second stage, indicating a considerable presence of this stage among the recorded cases. To gain further insights into the characteristics of the breast cancer patients, an examination of histology is warranted. Histology refers to the descriptive analysis of a tumour based on the level of abnormality exhibited by cancer cells and tissues when observed under a microscope (Figure 2). It also provides valuable information regarding the rate at which the cancer can proliferate and metastasize.

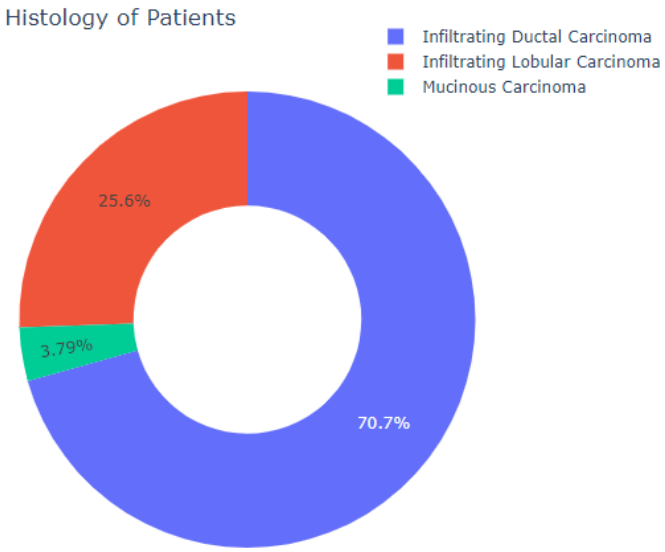

Figure 2 Histology of patients

Next, we will examine the values associated with the ER status, PR status and HER2 status variables for the patients in our dataset.

```
Positive    317
Name: ER status, dtype: int64
Positive    317
Name: PR status, dtype: int64
Negative    288
Positive     29
Name: HER2 status, dtype: int64
```

Now, let us examine the various surgical procedures performed on the patients in the dataset (Figure 3):

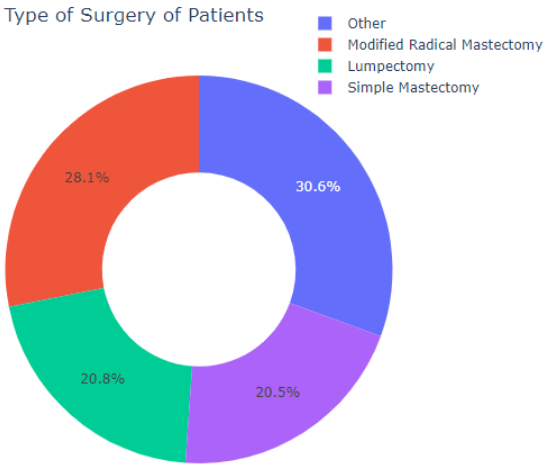

Figure 3 Type of surgery of patients

Upon examining the dataset, it was evident that numerous categorical features were present. To facilitate the training of a machine learning model using this data, it became imperative to transform the values within all the categorical columns. The subsequent section outlines the approach for transforming the categorical features in the dataset, enabling their utilization in the machine learning model.

	Patient_ID	Age	Gender	Protein1	Protein2	Protein3	Protein4	\
0	TCGA-D8-A1XD	36.0	1	0.080353	0.42638	0.54715	0.273680	
1	TCGA-EW-A1OX	43.0	1	-0.420320	0.57807	0.61447	-0.031505	
2	TCGA-A8-A079	69.0	1	0.213980	1.31140	-0.32747	-0.234260	
3	TCGA-D8-A1XR	56.0	1	0.345090	-0.21147	-0.19304	0.124270	
4	TCGA-BH-A0BF	56.0	1	0.221550	1.90680	0.52045	-0.311990	

	Tumour_Stage	Histology	ER status	PR status	HER2 status	Surgery_type	\
0	3	1	1	1	2	2	
1	2	3	1	1	2	3	
2	3	1	1	1	2	1	
3	2	1	1	1	2	2	
4	2	1	1	1	2	1	

	Date_of_Surgery	Date_of_Last_Visit	Patient_Status
0	15-Jan-17	19-Jun-17	Alive
1	26-Apr-17	09-Nov-18	Dead
2	08-Sep-17	09-Jun-18	Alive
3	25-Jan-17	12-Jul-17	Alive
4	06-May-17	27-Jun-19	Dead

3. MODEL FOR PREDICTING BREAST CANCER SURVIVAL

Proceeding towards training a machine learning model for predicting breast cancer patient survival, it is imperative to partition the data into distinct sets for training and testing purposes. This segregation aids in evaluating the model's performance on unseen data and prevents over fitting. Therefore, the next step involves the division of the dataset into training and test sets.

Now, let us proceed with the methodology of training a machine learning model in the following manner:

```
model = SVC()
model.fit(xtrain, ytrain)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning:

A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
```

```
SVC()
```

Now, let us input the comprehensive set of features utilized for training this machine learning model and subsequently predict the patient's likelihood of survival from breast cancer.

```
features = np.array([[36.0, 1, 0.080353, 0.42638, 0.54715, 0.273680, 3, 1, 1, 1, 2, 2,]])
print(model.predict(features))
```

```
['Alive']
```

4. CONCLUSION

Breast cancer is a significant global health issue and accurate prediction of patient survival outcomes is crucial for informed treatment decisions and personalized care. In this article, we explored the task of breast cancer survival prediction using machine learning techniques implemented in Python. By leveraging a dataset comprising over 400 breast cancer patients who underwent surgery, we demonstrated the potential of machine learning algorithms in predicting patient survival. We considered various features such as age, gender, protein expression levels, tumour stage, histology, receptor status, surgery type and relevant dates to develop a predictive model.

Ethical issues

Not applicable.

Informed consent

Not applicable.

Funding

This study has not received any external funding.

Conflict of Interest

The author declares that there are no conflicts of interests.

Data and materials availability

All data associated with this study are present in the paper.

REFERENCES AND NOTES

1. Chen L, Li Y, Li Z, Zhao J, Yang R. Integrating radiomics and clinical features for breast cancer survival prediction. IEEE Trans Med Imaging 2021; 40(3):749-758.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2019; 542(7639):115-118.
3. Jiang Y, Chen H, Lu J, Zhang Y, Zhu F, Wu FX, Wang J. Multi-modal deep learning for breast cancer survival prediction. Front Genet 2021; 12:703076.
4. Li C, Wang L, Zhu L, Huang Z, Li H, Zhou Y, Zhou J. Multi-omics prediction of survival in patients with breast cancer. Front Genet 2020; 11:226.

5. Li X, Yang L, Song Y, Zhang R, Liu L, Qiu T, Wang S. Predicting breast cancer survival using mammographic images: A systematic review and meta-analysis. Comput Biol Med 2021; 132:104366.
6. Liang M, Li Z, Chen T, Zeng J, Zeng X, Zhang L. Breast cancer survival prediction using histopathological images and multi-omics data with deep learning. Comput Biol Med 2022; 140:105009.
7. Saini HK, Griffith OL, Griffith M. Predictive modeling of breast cancer survival using integrated clinical and omics data. iScience 2020; 23(11):101719.
8. Wang X, Qin L, Huang T, Yang Z, Chen L, He L. A deep learning model based on electronic health records for breast cancer survival prediction. IEEE J Biomed Health Inform 2022; 26(1):206-215.
9. Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, Snyder M. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2020; 11(1):1-10.
10. Zhang Y, Liao X, Zhang L, Chen Z, Zhang S, Xu S. Predicting long-term survival of breast cancer patients based on electronic health records using deep neural network. Comput Biol Med 2021; 136:104689.