

BREAST CANCER SURVIVAL PREDICTION WITH THE MACHINE LEARNING APPROACH: A PERSONALIZED TREATMENT STRATEGY

Submitted in partial fulfillment of the requirements for the award of Bachelor of
Technology degree in Biotechnology
Interdisciplinary project

By

KEERTHANAA V (Reg. No. 41230029)

SMRITI J (Reg. No. 41230059)

ANU VARSHINI A.M (Reg. No. 41230010)



DEPARTMENT OF BIOTECHNOLOGY

SCHOOL OF BIO AND CHEMICAL ENGINEERING

**SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY**

(DEEMED TO BE UNIVERSITY)

CATEGORY-1 UNIVERSITY BY UGC

Accredited with grade “A++” by NAAC | 12 B status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI – 600119

MAY 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

CATEGORY-I UNIVERSITY BY UGC

Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF BIOTECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Keerthana V (Reg. No. 41230029), Anu Varshini A.M (Reg. No. 41230010), Smriti J (Reg. No. 41230059)** who carried out the Interdisciplinary Project entitled "**Breast Cancer Survival Prediction with the Machine Learning Approach: A Personalized Treatment Strategy**" under our supervision from January 2024 to April 2024.

Internal Guide

Dr. M. Bavanilatha M.Sc., PhD.,

Head of the Department

Dr. V. Ramesh Kumar, M.Phil., PhD.,

Submitted for Viva voce Examination held on 06/05/2024

Internal Examiner

External Examiner

DECLARATION

We, **Keerthanaa V (Reg. No. 41230029), Anu Varshini A.M (Reg. No. 41230010), Smriti J (Reg. No.41230059)** hereby declare that the Project Report entitled “**Breast Cancer Survival Prediction with theMachine Learning Approach: A Personalized Treatment Strategy**” done by us under the guidance of **Dr. M. Bavanilatha, MSc, PhD, Associate Professor** in Department of Biotechnology, Sathyabama Institute of Science and Technology is submitted in partial fulfilment of the requirements for the award of Bachelor of Technology degree in Biotechnology.

DATE: 06/05/2024

PLACE: CHENNAI



SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of **Sathyabama Institute of Science and Technology** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my sincere thanks to **Dr. V. Ramesh Kumar, M.Phil., PhD., Head of the Department**, Department of Biotechnology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. M. Bavanilatha, M.Sc., PhD., Associate Professor** in Department of Biotechnology, Sathyabama Institute of Science and Technology, Chennai for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I also express my thanks to **all Teaching and Non-teaching staff members** of the Department of Biotechnology who were helpful in many ways for the completion of the project work.

ABSTRACT

Breast cancer remains a significant challenge, but machine learning offers a promising approach for improved diagnosis and treatment. “Breast Cancer Survival Prediction with Machine Learning” explores this potential by leveraging the power of machine learning algorithms, particularly Support Vector Machines (SVMs).

Our project utilizes Python, a versatile programming language well-suited for machine learning tasks. We combine established biomarkers (ER, PR, HER2) and a patient’s unique genetic makeup (BRCA mutations and other variants) for a comprehensive analysis. This empowers healthcare professionals to create personalized treatment plans. The project holds promise for enhanced prognosis, improved treatment efficacy with fewer side effects, optimized resource allocation, and earlier intervention for high-risk cases. Future possibilities include refined risk stratification, tailored treatment based on predicted response, and even preventive measures. By harnessing machine learning and a comprehensive set of biomarkers, this project ushers in a new era of personalized medicine for breast cancer, offering a path towards a healthier future.

Keywords: Breast cancer, machine learning, personalized treatment

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE No.
	ABSTRACT	v
	LIST OF FIGURES	vii
	LIST OF SYMBOLS AND ABBEREVIATIONS	viii
1	INTRODUCTION	1
	1.1 THE GLOBAL BURDEN OF BREAST CANCER	1
	1.2 UNVEILING THE HALLMARKS OF BREAST CANCER	1
	1.3 MACHINE LEARNING AND PERSONALIZED MEDICINE	2
2	REVIEW OF LITERATURE	3
	2.1 BIOMARKERS RESPONSIBLE FOR DETECTING BREAST CANCER	3
	2.2 PROGRAMMING LANGUAGES	7
	2.2.1 PYTHON	7
	2.2.2 MACHINE LEARNING	7
	2.2.3 ALGORITHMS USED	8
	2.2.4 LIBRARIES USED	9
	2.3 ADVANCEMENTS IN PREDICTION	10
3	AIM AND PROJECT OBJECTIVE	12
	3.1 AIM	12
	3.2 PROJECT OBJECTIVE	12
4	METHODOLOGY	13
	4.1 DATA DESCRIPTION	13
	4.2 IMPORTING LIBRARIES AND THE DATASET	14
	4.3 DATA PRE-PROCESSING	15
	4.4 MODEL TRAINING AND EVALUATION	16
	4.5 DATA VISUALIZATION	18
5	RESULTS AND DISCUSSION	25
6	SUMMARY AND CONCLUSION	33
	REFERENCES	35

LIST OF SYMBOLS AND ABBREVIATIONS

SVC	Support Vector Classifier
SVM	Support Vector Machines
circRNAs	circular RNAs
miRNAs	microRNAs
ER	Estrogen Receptor
PR	Progesterone Receptor
HER2	Human Epidermal Growth Factor Receptor 2
ML	Machine learning
ANN	Artificial Neural Networks
Ki-67	Antigen Kiel 67 or (MKI67- marker of proliferation Kiel 67)

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
2.1.2	Breast Carcinoma	10
2.2.2	ML Classification	12
4.1	Data Description	16
4.2	Importing Libraries And The Dataset	17
4.3	Data Pre-Processing	18
4.4	Model Training And Evaluation	19
4.5	Data Visualization	20
5.1	Data Preprocessing and Model Training and Evaluation	29
5.3	Data Visualization	29
5.1.1	Tumor Stages Of Patients	27
5.1.2	Histology Of Patient	28
5.1.3	Type Of Surgery Of Patients	28
5.1.4	Age Distribution By Patient Status	29
5.1.5	Protein1 Distribution By Patients Status	29
5.1.6	Protein2 Distribution By Patients Status	30
5.1.7	Protein3 Distribution By Patients Status	30
5.1.8	Protein4 Distribution By Patients Status	31
5.1.9	Age Distribution	31
5.1.10	Heatmap (Correlation Matrix)	32

CHAPTER 1

INTRODUCTION

1.1 THE GLOBAL BURDEN OF BREAST CANCER

Breast cancer is the most common invasive cancer in women worldwide. In 2020 alone, an estimated 2.3 million women received a breast cancer diagnosis, and tragically, 685,000 succumbed to the disease. Early detection and accurate prognosis are essential for improving patient outcomes. Traditional methods rely on clinical evaluation and pathological analysis of tissue samples. While valuable, these techniques may not capture the complex interplay of factors within vast datasets of patient characteristics, treatment responses, and molecular profiles.

1.2 UNVEILING THE HALLMARKS OF BREAST CANCER

Breast cancer is a heterogeneous disease, encompassing a diverse range of subtypes with varying clinical presentations, prognoses, and responses to treatment. Understanding the underlying biology of breast cancer is crucial for developing effective prognostic tools. Key hallmarks of breast cancer include:

- Uncontrolled Cell Growth and Proliferation: Cancerous cells exhibit abnormal growth patterns, dividing rapidly and invading surrounding healthy tissues.
- Angiogenesis: Tumours stimulate the formation of new blood vessels to sustain their growth and nutrient supply.
- Immune Escape: Cancer cells can evade the body's natural immune defenses, allowing them to proliferate unchecked.
- Metastasis: The ability of cancer cells to detach from the primary tumour, travel through the bloodstream or lymphatic system, and establish secondary tumours in distant organs.

1.3 MACHINE LEARNING AND PERSONALIZED MEDICINE

Traditional methods, while valuable, often struggle to capture the intricate interplay of factors within vast patient datasets. This is where machine learning steps in. Machine learning algorithms are powerful tools that can analyse large datasets to identify hidden patterns and relationships. In the context of breast cancer, we can utilize machine learning to analyse the vast amount of patient data, including traditional clinical factors, biomarker profiles, and genetic information. By employing Python, a versatile programming language well-suited for machine learning tasks, we can develop a model (specifically, a Support Vector Classifier or SVC) that can learn from this data and predict a patient's breast cancer survival outcomes. This allows for a more nuanced understanding of the disease and the development of personalized treatment plans tailored to each patient's unique characteristics. This approach holds promise for enhanced prognosis, improved treatment efficacy with fewer side effects, optimized resource allocation, and earlier intervention for high-risk cases. Future possibilities include refined risk stratification, tailored treatment based on predicted response, and even preventive measures. The SVC algorithm excels at classifying data points into distinct categories. In our case, we will train the SVC to categorize patients based on their risk of recurrence or poor prognosis. This allows us to create more nuanced risk stratification and tailor treatment strategies to individual needs.

CHAPTER 2

REVIEW OF LITERATURE

Breast cancer remains a significant global health challenge, with survival prediction playing a crucial role in clinical decision-making and patient management. Over the years, researchers have explored various methodologies, including traditional statistical models and more recently, machine learning algorithms, to enhance the accuracy and reliability of survival prediction for breast cancer patients.

2.1. BIOMARKERS RESPONSIBLE FOR DETECTING BREAST CANCER

2.1.1 BREAST CANCER AND BIOMARKERS

Breast cancer is a significant health concern worldwide, with high mortality rates in advanced stages of the disease. One of the key areas of research in breast cancer is the identification of biomarkers that can aid in early detection, prognosis, and treatment. These biomarkers can provide valuable information about the presence and progression of breast cancer, allowing for timely interventions and improved outcomes for patients. Researchers are focused on finding reliable protein biomarkers

The continued investigation and validation of these biomarkers contribute to the advancement of precision medicine in breast cancer, offering promise for improved patient care and outcomes. Therefore, the identification and validation of reliable biomarkers for breast cancer detection, prognosis prediction, and treatment guidance is crucial in improving patient outcomes and reducing unnecessary exposure to ineffective treatments. Breast cancer Biomarkers are of the following.

Breast cancer remains a formidable challenge in the realm of oncology, affecting millions of lives worldwide. However, amidst this adversity, the dawn of precision medicine has illuminated a path forward, driven by the discovery and utilization of breast cancer biomarkers. These molecular signposts offer invaluable insights into the intricate biology of the disease, empowering clinicians with targeted strategies for diagnosis, prognosis, and treatment.

In this project, we delve into the burgeoning landscape of breast cancer biomarkers, exploring their significance, clinical implications, and the transformative potential they hold in the fight against breast cancer.

Breast cancer is the second most reported cancer in women with high mortality causing millions of cancer-related deaths annually. Early detection of breast cancer intensifies the struggle towards discovering, developing, and optimizing diagnostic biomarkers that can improve its prognosis and therapeutic outcomes. Breast cancer-associated biomarkers comprise macromolecules, such as nucleic acid (DNA/RNA), proteins, and intact cells. Advancements in molecular technologies have identified all types of biomarkers that are exclusively studied for diagnostic, prognostic, drug resistance, and therapeutic implications. Identifying biomarkers may solve the problem of drug resistance which is a challenging obstacle in breast cancer treatment. Dysregulation of non-coding RNAs including circular RNAs (circRNAs) and microRNAs (miRNAs) initiates and progresses breast cancer. The circulating multiple miRNA profiles promise better diagnostic and prognostic performance and sensitivity than individual miRNAs. The high stability and existence of circRNAs in body fluids make them a promising new diagnostic biomarker. Many therapeutic-based novel targeting agents have been identified, including ESR1 mutation (DNA mutations), Oligonucleotide analogs and antagonists (miRNA), poly (ADP-ribose) polymerase (PARP) in BRCA mutations, CDK4/6 (cell cycle regulating factor initiates tumour progression), Androgen receptor (a steroid hormone receptor), that have entered clinical validation procedure. In this review, we summarize the role of novel breast cancer diagnostic biomarkers, drug resistance, and therapeutic implications for breast cancer. (Hargreaves MK, 2009)

2.1.2 UNDERSTANDING BREAST CANCER BIOMARKERS

Biomarkers are measurable biological indicators that reflect the presence, severity, or progression of a disease. In the context of breast cancer, biomarkers encompass a diverse array of molecular entities, ranging from genetic mutations and protein expressions to microRNA profiles. These biomarkers serve as windows into the underlying molecular aberrations driving tumour initiation, progression, and response to therapy.

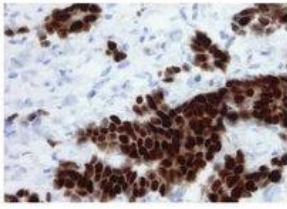
Biomarkers for breast cancer can be crucial for diagnosis, prognosis, and treatment decisions. Here are some important biomarkers commonly associated with breast cancer:

- a. **Estrogen Receptor (ER):** ER-positive breast cancers have receptors that bind to the hormone estrogen, promoting cancer growth. ER status helps determine if hormonal therapies like tamoxifen or aromatase inhibitors might be effective.
- b. **Progesterone Receptor (PR):** Similar to ER, PR-positive breast cancers express receptors for the hormone progesterone. PR status is often evaluated alongside ER status for treatment decisions.
- c. **Human Epidermal Growth Factor Receptor 2 (HER2):** HER2-positive breast cancers overexpress the HER2 protein, promoting aggressive tumour growth. Targeted therapies like trastuzumab (Herceptin) and pertuzumab (Perjeta) specifically target HER2-positive tumors.

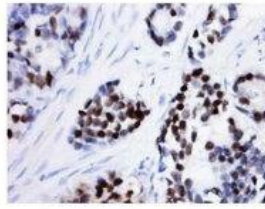
ER, PR, and HER2 status are essential in guiding the selection of targeted therapies, such as hormone therapy and HER2-targeted therapy. Patients with ER-positive and/or PR-positive breast cancer may benefit from endocrine therapy, while those with HER2-positive tumours may be candidates for HER2-targeted drugs. Understanding the expression of these biomarkers is crucial in tailoring treatment plans to individual patients, leading to better outcomes and reduced unnecessary exposure to ineffective therapies. In addition to guiding treatment decisions, the assessment of ER, PR, and HER2 status also provides valuable information in determining the aggressiveness of the disease and predicting the likelihood of disease recurrence. This, in turn, assists in formulating personalized follow-up plans and monitoring strategies for patients' post-treatment. (Wan LL, 2020)

Positive immunostaining

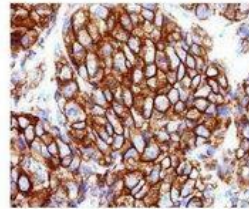
ER



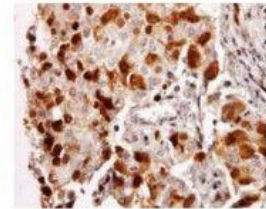
PR



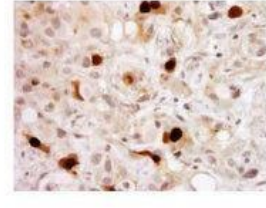
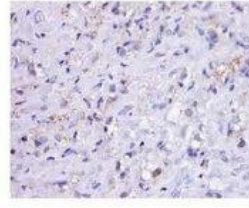
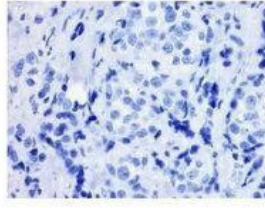
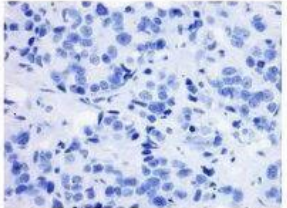
HER-2



Cyclin A2

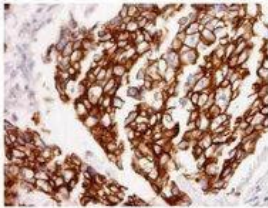


Negative immunostaining

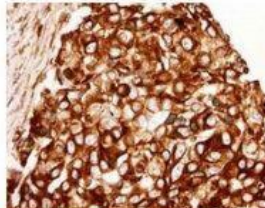


Positive immunostaining

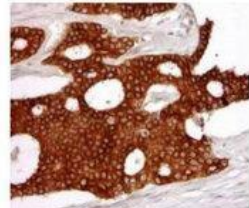
CK5



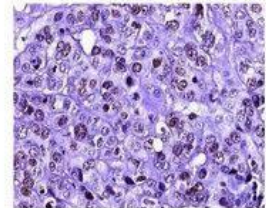
Vimentin



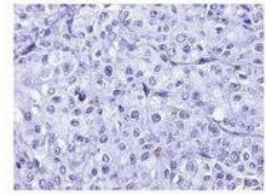
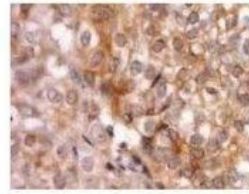
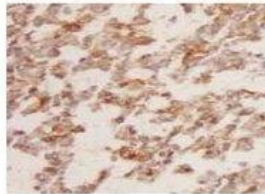
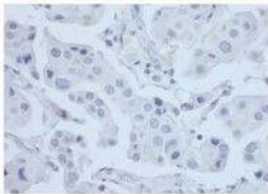
Bcl2



Ki-67



Negative immunostaining



Breast carcinoma

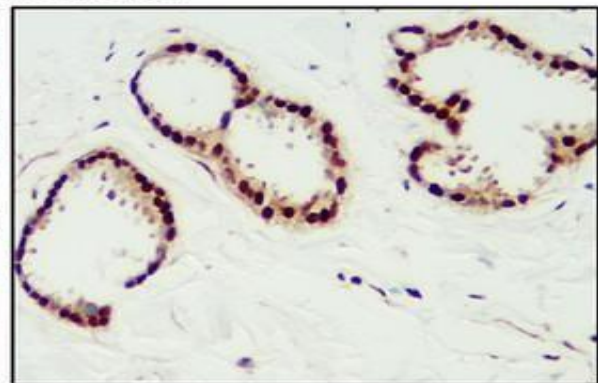
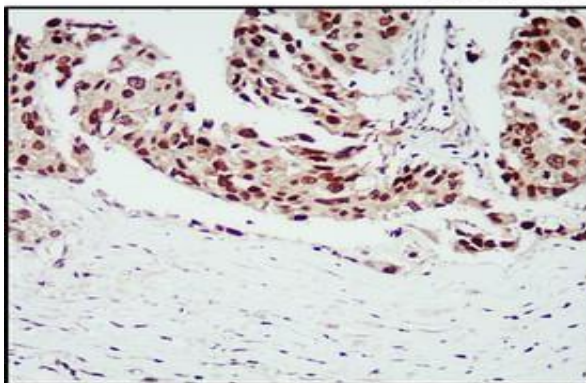


Figure 2.1.2: Breast carcinoma (Hargreaves MK, et al., 2009)

2.2 PROGRAMMING LANGUAGES

2.2.1 PYTHON

Python is widely hailed for its suitability in machine learning and data science, ranking as the fifth most significant programming language in these domains. Its interpreted, object-oriented nature with dynamic semantics facilitates rapid development and integration. Python's simplicity, readability, and extensive standard library bolster its appeal, easing maintenance and fostering code reuse. Features such as object-oriented programming, high-level structure, extensibility, portability, and integration with other languages underscore its versatility. However, Python faces limitations, including challenges in enterprise development adoption, weak mobile computing support, and runtime performance issues due to interpretation. Additionally, its dynamically typed nature may lead to runtime errors, and its database access layers are perceived as underdeveloped for complex legacy data interactions. Despite these drawbacks, Python's robust support libraries and productivity enhancements continue to attract programmers across various domains. (Sandeep K. Gupta, 2019)

2.2.2 MACHINE LEARNING

Machine learning (ML) is an integral part of artificial intelligence, focusing on the development of algorithms capable of improving autonomously through data and experience. By analysing training data, these algorithms construct models enabling predictions or decisions without explicit programming. ML finds application across diverse fields such as medicine, email filtering, speech recognition, and computer vision, addressing tasks where conventional algorithms fall short. Over the past decade, ML has gained prominence in predictive analytics, leveraging data to forecast future events. It seeks to discern data structures and apply appropriate models to reveal hidden patterns. The development of ML models is crucial for predictive analysis, with numerous algorithms available for real-world problem-solving. Unlike traditional computing, where algorithms are explicitly defined, ML algorithms enable computers to learn from data inputs, employing statistical analysis to generate output within defined parameters. Consequently, ML empowers computers to automate decision-making processes based on learned data patterns. ML transforms information into actionable knowledge, utilizing vast datasets accumulated over

decades to uncover valuable insights and facilitate complex decision-making. By leveraging ML techniques, organizations can extract meaningful patterns and insights from complex data, enabling informed decisions and predictive capabilities for the future. (Dr. Sasikala, 2020)

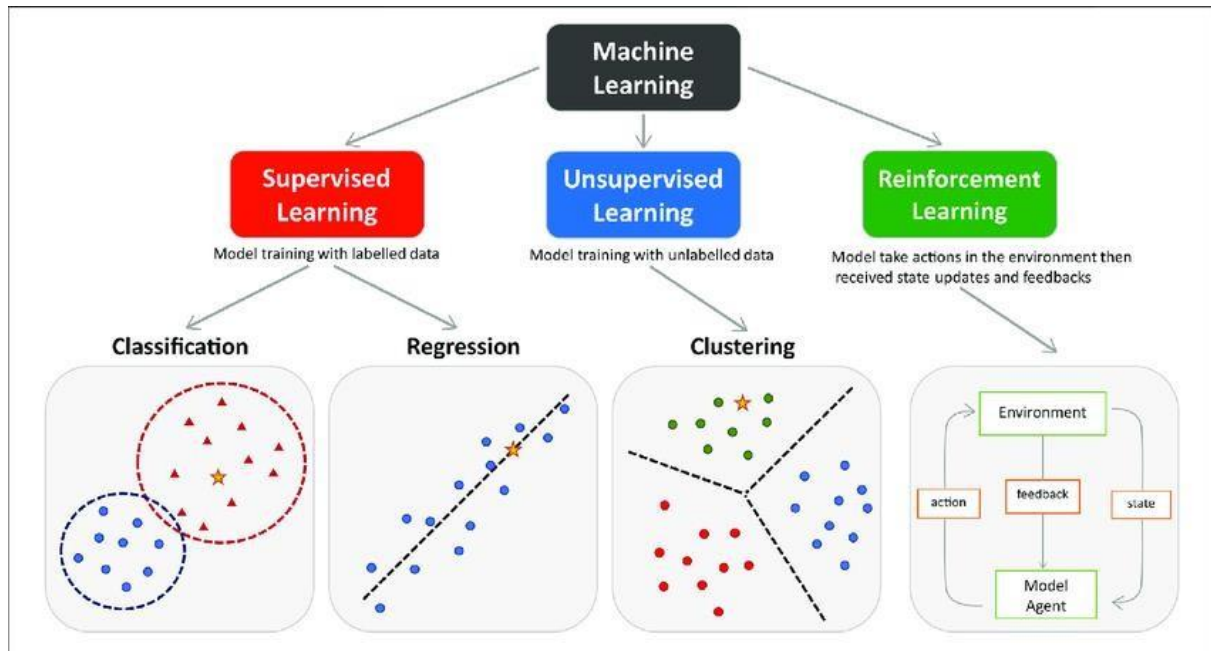


Figure 2.2.2: ML classification (Coziana Ciurtin, Sep 2021)

2.2.3 ALGORITHMS USED

Support Vector Classifier (SVC), a member of the Support Vector Machines (SVM) family, is adept at classification tasks, particularly when classes are linearly or nearly separable. SVC determines the optimal hyperplane in feature space to maximize the margin between classes, enhancing robustness. Using kernel functions like linear, polynomial, or radial basis, SVC extends to handle non-linearly separable data, capturing intricate feature relationships for improved accuracy. By maximizing margin between support vectors, SVC enhances generalization and resilience to noise. Incorporating a regularization parameter (C), SVC balances margin maximization and classification error minimization, mitigating overfitting. While primarily designed for binary classification, SVC extends to multi-class problems

through one-vs-one or one-vs-rest strategies. While effective for medium-sized datasets, scalability challenges may arise with large datasets, mitigated by scikit-learn's optimizations. Overall, SVC's versatility, margin maximization, and regularization control render it indispensable for diverse classification tasks in machine learning. (Dr. Jason Brownlee, 2019)

2.2.4 LIBRARIES USED

In our project we utilize several Python libraries for data manipulation, visualization, machine learning, and evaluation. Here's a list of the libraries used:

- **pandas:** Used for data manipulation and analysis, providing data structures like DataFrame to work with structured data efficiently.
- **numpy:** Provides support for numerical operations and arrays, essential for handling numerical data and computations efficiently.
- **plotly.express:** Used for creating interactive data visualizations, including pie charts, bar charts, scatter plots, and more, allowing for intuitive exploration of data.
- **plotly.graph_objects:** Provides more control over plot customization and creation of complex visualizations, complementing Plotly Express.
- **sklearn (scikit-learn):** A comprehensive machine learning library in Python, used for model selection, training, evaluation, and preprocessing. Specific modules used include ``model_selection``, ``svm``, ``metrics``, and ``preprocessing``.
- **seaborn:** Built on top of matplotlib, seaborn provides high-level interfaces for drawing attractive statistical graphics. It's used for creating heatmaps in the code.
- **matplotlib.pyplot:** A widely used plotting library in Python, providing functions for creating static, interactive, and animated visualizations. Used for creating static plots like box plots and heatmaps.

2.3 ADVANCEMENTS IN PREDICTION (From Traditional Models to Machine Learning Approaches)

a. Traditional Predictive Models: Early efforts in breast cancer survival prediction relied on traditional statistical models such as Cox proportional hazards regression and Kaplan-Meier analysis. These models identified key prognostic factors including tumour size, lymph node involvement, histological grade, and hormone receptor status, providing valuable insights into disease progression and patient outcomes.

b. Machine Learning Approaches: In recent years, the application of machine learning techniques has revolutionized breast cancer survival prediction by leveraging the abundance of clinical and molecular data available. Machine learning algorithms such as Support Vector Machines (SVM), Random Forests, and Artificial Neural Networks (ANN) have shown promise in capturing complex relationships among predictive variables and improving prediction accuracy.

- **Feature Selection and Integration:** Feature selection methods, including recursive feature elimination and LASSO regularization, have been employed to identify the most informative predictors from high-dimensional datasets, including genomic, proteomic, and clinical variables. Integrating multi-omics data, such as gene expression profiles and protein biomarkers, has enhanced the predictive power of machine learning models and provided insights into the underlying molecular mechanisms driving breast cancer progression.
- **Model Optimization and Validation:** Efforts have been made to optimize machine learning models for breast cancer survival prediction by fine-tuning hyperparameters, optimizing feature representations, and addressing class imbalance issues. Cross-validation techniques and external validation on independent datasets are essential for assessing the robustness and generalizability of predictive models across diverse patient populations and clinical settings.

c. Clinical Implementation and Impact: The successful translation of machine learning models into clinical practice requires addressing several challenges, including model interpretability, regulatory approval, and integration into existing healthcare

work flows. Collaborative efforts between clinicians, data scientists, and regulatory agencies are essential to ensure the safe and effective deployment of predictive models for guiding personalized treatment decisions and improving patient outcomes.

d. Ethical Considerations and Bias Mitigation: As machine learning models become increasingly integrated into clinical decision-making, it is essential to address ethical considerations, including patient privacy, transparency, and fairness. Strategies for mitigating bias in predictive models, such as algorithmic transparency, bias-aware training, and post-hoc fairness assessments, are critical for ensuring equitable healthcare delivery and reducing disparities in breast cancer care.

The convergence of machine learning techniques with advanced molecular profiling technologies holds immense potential for enhancing breast cancer survival prediction and improving patient care. However, ongoing research is needed to address methodological challenges, validate predictive models in real-world clinical settings, and ensure their ethical and equitable implementation.

Data Privacy: Patient data must be anonymized and handled according to strict security protocols.

Model Transparency: Explainability of the model's predictions is crucial to build trust and ensure responsible use.

Bias Mitigation: The training data should be balanced and representative of the target population to avoid biased predictions.

(Dr. Emma H. Allott, 2020)

CHAPTER 3

AIM AND OBJECTIVE

3.1 AIM

The aim of this project is to develop a machine learning-based predictive model for breast cancer survival, leveraging clinical and molecular data to assist in prognosis and treatment decision-making.

3.2 OBJECTIVE

The primary objective of this project aligns with the overarching goal of leveraging machine learning to enhance breast cancer patient care. By developing a robust predictive model trained on a comprehensive dataset encompassing various patient attributes, including demographics, medical history, biomarker status, genetic information, and diagnostic test results, the aim is to provide clinicians with valuable insights into patient survival and recurrence risk. This multifaceted analysis enables the model to predict the likelihood of a patient's survival over a specific timeframe and assess their potential for disease recurrence. Such predictive capabilities hold immense potential in clinical decision-making, empowering healthcare professionals to tailor treatment plans based on predicted outcomes. Additionally, by identifying patients who may benefit from personalized therapies due to their unique biological profile, the model contributes to the advancement of precision medicine in breast cancer management. Moreover, the early detection and intervention enabled by the model for high-risk cases have the potential to significantly improve patient outcomes and quality of life. Overall, the development of this predictive model represents a significant step towards personalized and proactive healthcare delivery in the field of breast cancer.

CHAPTER 4

METHODOLOGY

4.1 DATA DESCRIPTION

The dataset contains records of over 300 breast cancer patients who underwent surgery as part of their treatment. Each record includes details such as patient ID, age, gender, protein expression levels (Protein1, Protein2, Protein3, Protein3), tumour stage, histological subtype (Histology), and statuses of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), type of surgery performed (Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy, or Other), the dates of surgery and the patient's last visit, and the current status of the patient (Alive or Dead). The dataset provides valuable insights into patient demographics, tumour characteristics, treatment procedures, and outcomes, facilitating further analysis for clinical and research purposes in breast cancer management.

- ✓ Patient_ID: Unique identifier for each patient.
- ✓ Age: Age of the patient at the time of surgery.
- ✓ Gender: Gender of the patient.
- ✓ Protein1, Protein2, Protein3, Protein4: Expression levels of proteins.
- ✓ Tumor_Stage: Stage of breast cancer diagnosed in the patient.
- ✓ Histology: Histological subtype of the breast cancer.
- ✓ ER status: Estrogen Receptor status (Positive/Negative).
- ✓ PR status: Progesterone Receptor status (Positive/Negative).
- ✓ HER2 status: Human Epidermal Growth Factor Receptor 2 status (Positive/Negative).
- ✓ Surgery_type: Type of surgery undergone by the patient (Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy, Other).
- ✓ DateofSurgery: Date when the surgery was performed.
- ✓ DateofLast_Visit: Date of the patient's last visit.
- ✓ Patient_Status: Current status of the patient (Alive/Dead).

Dataset from: <https://www.kaggle.com/datasets/amandam1/breastcancerdataset>

File

Home

Insert

Page Layout

Formulas

Data

Review

View

Help

</

Fig 4.1 DATA DESCRIPTION

4.2 IMPORTING LIBRARIES AND THE DATASET

In this project, the initial steps involve importing essential libraries and loading the breast cancer dataset for analysis. The libraries imported include pandas for data manipulation, numpy for numerical computations, plotly.express and plotly.graph_objects for interactive data visualization, and scikit-learn modules such as train_test_split for data splitting, SVC for Support Vector Classifier instantiation, and metrics like classification_report, confusion_matrix, and accuracy_score for model evaluation. Additionally, StandardScaler from scikit-learn is imported for feature scaling, seaborn for statistical plotting, and matplotlib.pyplot for creating static plots. Once the libraries are imported, the breast cancer dataset is loaded from a CSV file path using pandas' read_csv function and stored in the DataFrame named 'data'. This dataset serves as the foundation for subsequent data exploration, preprocessing, visualization, model building, and evaluation steps aimed at analyzing breast cancer data using machine learning techniques.

```
Code 1.py - D:/Sem Notes/Sem 6/IDP/Code 1.py (3.10.2)
File Edit Format Run Options Window Help
1 import pandas as pd
2 import numpy as np
3 import plotly.express as px
4 import plotly.graph_objects as go
5 from sklearn.model_selection import train_test_split
6 from sklearn.svm import SVC
7 from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
8 from sklearn.preprocessing import StandardScaler
9 import seaborn as sns
10 import matplotlib.pyplot as plt
11
12 # Load the dataset
13 data = pd.read_csv(r"D:\Sem Notes\Sem 6\IDP\Dataset\BRCA.csv")
```

Fig 4.2 IMPORTING LIBRARIES AND THE DATASET

4.3 DATA PRE-PROCESSING

Data preprocessing is a crucial step in this project, ensuring that the dataset is properly prepared for analysis and modelling. The provided code snippet initiates data preprocessing by first checking for missing values using the `isnull()` function and dropping rows with missing data using the `dropna()` method. This ensures that the dataset is clean and devoid of any missing or incomplete records. Next, categorical variables such as Tumour Stage, Histology, ER status, PR status, HER2 status, Gender, and Surgery type are explored and mapped to numerical values using the `map()` function, enabling the application of machine learning algorithms that require numerical input. This mapping facilitates the conversion of categorical data into a format suitable for analysis. Once categorical variables are transformed, the dataset is split into features (X) and the target variable (y) using square brackets to select the relevant columns. This separation sets the stage for subsequent steps, including feature scaling, data splitting, model training, and evaluation, ultimately leading to a comprehensive analysis of breast cancer data.

```

18 # Check for missing values
19 print(data.isnull().sum())
20
21 # Drop rows with missing values
22 data = data.dropna()
23
24 # Display information about the dataset
25 data.info()
26
27 # Explore Gender distribution
28 print(data['Gender'].value_counts())
29
30 # Visualize Tumour Stage distribution
31 stage = data["Tumour_Stage"].value_counts()
32 figure = px.pie(data, values=stage.values, names=stage.index, hole=0.5, title="Tumour Stages of Patients")
33 figure.show()
34
35 # Visualize Histology distribution
36 histology = data["Histology"].value_counts()
37 figure = px.pie(data, values=histology.values, names=histology.index, hole=0.5, title="Histology of Patients")
38 figure.show()
39
40 # Explore ER status
41 print(data["ER_status"].value_counts())
42
43 # Explore PR status
44 print(data["PR_status"].value_counts())
45
46 # Explore HER2 status
47 print(data["HER2_status"].value_counts())

```

```

Code 1.py - D:\Sem Notes\Sem 6\IDP\Code 1.py (3.10.2)
File Edit Format Run Options Window Help
46 # Explore HER2 status
47 print(data["HER2_status"].value_counts())
48
49 # Explore Surgery type distribution
50 surgery = data["Surgery_type"].value_counts()
51 figure = px.pie(data, values=surgery.values, names=surgery.index, hole=0.5, title="Type of Surgery of Patients")
52 figure.show()
53
54 # Map categorical variables to numerical values
55 data["Tumour_Stage"] = data["Tumour_Stage"].map({"I": 1, "II": 2, "III": 3})
56 data["Histology"] = data["Histology"].map({"Infiltrating Ductal Carcinoma": 1,
57 "Infiltrating Lobular Carcinoma": 2,
58 "Mucinous Carcinoma": 3})
59 data["ER_status"] = data["ER_status"].map({"Positive": 1, "Negative": 0})
60 data["PR_status"] = data["PR_status"].map({"Positive": 1, "Negative": 0})
61 data["HER2_status"] = data["HER2_status"].map({"Positive": 1, "Negative": 0})
62 data["Gender"] = data["Gender"].map({"MALE": 0, "FEMALE": 1})
63 data["Surgery_type"] = data["Surgery_type"].map({"Other": 1,
64 "Modified Radical Mastectomy": 2,
65 "Lumpectomy": 3,
66 "Simple Mastectomy": 4})
67
68 # Display the modified dataset
69 print(data.head())
70
71 # Splitting data into features (X) and target (y)
72 X = data[["Age", "Gender", "Protein1", "Protein2", "Protein3", "Protein4",
73 "Tumour_Stage", "Histology", "ER_status", "PR_status",
74 "HER2_status", "Surgery_type"]]
75 y = data["Patient_Status"]
76

```

Fig 4.3 DATA PRE-PROCESSING

4.4 MODEL TRAINING AND EVALUATION

In this project, after preprocessing the dataset, the Support Vector Classifier (SVM) model is trained using the scaled features (X_train) and corresponding target variable (y_train). Once trained, the model predicts the target variable for the test set (X_test) using the `predict` method. The accuracy of the model is then calculated by comparing the predicted values with the actual values using the `accuracy\_score` function. Additionally, a classification report is generated, providing metrics such as precision, recall, and F1-score for each class (patient status). Furthermore, a confusion matrix is printed to visualize the true positive, false positive, true negative, and false negative predictions, offering insights into the model's performance across different outcome categories. Through this process, the effectiveness of the SVM

Model in predicting breast cancer patient status is evaluated, informing further analysis and potential model refinement.

```
71 # Splitting data into features (X) and target (y)
72 X = data[['Age', 'Gender', 'Protein1', 'Protein2', 'Protein3', 'Protein4',
73 'Tumour_Stage', 'Histology', 'ER status', 'PR status',
74 'HER2 status', 'Surgery_type']]
75 y = data['Patient_Status']
76
77 # Feature Scaling
78 scaler = StandardScaler()
79 X_scaled = scaler.fit_transform(X)
80
81 # Splitting data into training and testing sets
82 X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.10, random_state=42)
83
84 # Instantiate the Support Vector Classifier (SVM) model
85 model = SVC()
86
87 # Train the model
88 model.fit(X_train, y_train)
89
90 # Evaluate the model's performance
91 y_pred = model.predict(X_test)
92
93 # Calculate accuracy
94 accuracy = accuracy_score(y_test, y_pred)
95 print("Accuracy:", accuracy)
96
97 # Get the unique classes in y_test and y_pred
98 unique_classes = np.unique(np.concatenate((y_test, y_pred)))
99
100 # Print classification report with handling for ill-defined precision
101 print("Classification Report:")
102 print(classification_report(y_test, y_pred, zero_division=1))
103
104 # Print confusion matrix with all classes included
105 print("Confusion Matrix:")
106 print(confusion_matrix(y_test, y_pred, labels=unique_classes))
107
```

Fig 4.4 MODEL TRAINING AND EVALUATION

4.5 DATA VISUALIZATION

Data visualization plays a crucial role in understanding patterns and relationships in the breast cancer dataset. Pie charts illustrate the distribution of categorical variables like Tumour Stages and Histology, revealing proportions in each category. Box plots showcase numerical variables across different Patient Status categories, identifying trends in protein expression and age distribution. A line graph depicts patient age distribution clearly. Additionally, a heatmap visualizes the correlation matrix, assessing relationships between features. These visuals offer insights for further analysis and modelling.

```
100 # Print classification report with handling for ill-defined precision
101 print("Classification Report:")
102 print(classification_report(y_test, y_pred, zero_division=1))
103
104 # Print confusion matrix with all classes included
105 print("Confusion Matrix:")
106 print(confusion_matrix(y_test, y_pred, labels=unique_classes))
107
108 # Box plots for numerical variables
109 numerical_vars = ['Age', 'Protein1', 'Protein2', 'Protein3', 'Protein4']
110 for var in numerical_vars:
111     box_fig = px.box(data, x='Patient_Status', y=var, points='all', title=f'{var} Distribution by Patient Status')
112     box_fig.update_layout(xaxis_title='Patient Status', yaxis_title=var)
113     box_fig.show()
114
115 # Line graph for Age distribution
116 age_counts = data['Age'].value_counts().sort_index()
117 age_fig = go.Figure([go.Scatter(x=age_counts.index, y=age_counts.values, mode='lines')])
118 age_fig.update_layout(title="Age Distribution", xaxis_title="Age", yaxis_title="Count")
119 age_fig.show()
120
121 # Select only numeric columns
122 numeric_columns = data.select_dtypes(include=[np.number])
123
124 # Calculate correlation matrix
125 corr_matrix = numeric_columns.corr()
126
127 # Plot heatmap
128 plt.figure(figsize=(10, 8))
129 sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
130 plt.title("Correlation Matrix")
131 plt.show()
132
```

Fig 4.5 DATA VISUALIZATION

4.5.1 PIE CHARTS FOR TUMOUR STAGES AND HISTOLOGY

Pie charts are generated using Plotly Express (`px.pie``) to visualize the distribution of Tumour Stages and Histology within the dataset. These charts provide an overview of the proportions of different stages and histological types of tumors.

4.5.2 PIE CHART FOR SURGERY TYPE

Another pie chart is created to visualize the distribution of different types of surgeries performed on patients, providing insights into the prevalence of each surgical procedure.

4.5.3 BOX PLOTS FOR NUMERICAL VARIABLES

Box plots are generated using Plotly Express (`px.box``) to visualize the distribution of numerical variables (Age, Protein1, Protein2, Protein3, Protein3) across different categories of the target variable (Patient Status). These plots help identify potential differences in protein expression levels and age distribution between patients with different outcomes.

4.5.4. LINE GRAPH FOR AGE DISTRIBUTION

A line graph is created using Plotly Graph Objects (`go.Figure``) to visualize the distribution of patient age within the dataset. This graph offers insights into the age distribution of breast cancer patients.

4.5.5 HEATMAP FOR CORRELATION MATRIX

A heatmap is generated using Seaborn (`sns.heatmap``) to visualize the correlation matrix between numerical variables. This heatmap helps assess the relationships and potential multicollinearity between features in the dataset.

4.6 THE CODE

```
import pandas as pd

import numpy as np

import plotly.express as px

import plotly.graph_objects as go

from sklearn.model_selection import train_test_split

from sklearn.svm import SVC

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

from sklearn.preprocessing import StandardScaler

import seaborn as sns

import matplotlib.pyplot as plt


# Load the dataset

data = pd.read_csv(r"D:\Sem Notes\Sem 6\IDP\Dataset\BRCA.csv")

# Display the first few rows of the dataset

print(data.head())

# Check for missing values

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

# Drop rows with missing values

data = data.dropna()

# Display information about the dataset

data.info()

# Explore Gender distribution

print(data['Gender'].value_counts())

# Visualize Tumour Stage distribution
```

```

stage = data["Tumour_Stage"].value_counts()

figure = px.pie(data, values=stage.values, names=stage.index, hole=0.5, title="Tumour
Stages of Patients")

figure.show()

# Visualize Histology distribution

histology = data["Histology"].value_counts()

figure = px.pie(data, values=histology.values, names=histology.index, hole=0.5,
title="Histology of Patients")

figure.show()

# Explore ER status

print(data["ER status"].value_counts())

# Explore PR status

print(data["PR status"].value_counts())

# Explore HER2 status

print(data["HER2 status"].value_counts())

# Explore Surgery type distribution

surgery = data["Surgery_type"].value_counts()

figure = px.pie(data, values=surgery.values, names=surgery.index, hole=0.5, title="Type of
Surgery of Patients")

figure.show()

# Map categorical variables to numerical values

data["Tumour_Stage"] = data["Tumour_Stage"].map({"I": 1, "II": 2, "III": 3})

data["Histology"] = data["Histology"].map({"Infiltrating Ductal Carcinoma": 1,
                                           "Infiltrating Lobular Carcinoma": 2,
                                           "Mucinous Carcinoma": 3})

data["ER status"] = data["ER status"].map({"Positive": 1, "Negative": 0})

data["PR status"] = data["PR status"].map({"Positive": 1, "Negative": 0})

```

```

data["HER2 status"] = data["HER2 status"].map({"Positive": 1, "Negative": 0})

data["Gender"] = data["Gender"].map({"MALE": 0, "FEMALE": 1})

data["Surgery_type"] = data["Surgery_type"].map({"Other": 1,
                                                "Modified Radical Mastectomy": 2,
                                                "Lumpectomy": 3,
                                                "Simple Mastectomy": 4})

# Display the modified dataset

print(data.head())

# Splitting data into features (X) and target (y)

X = data[['Age', 'Gender', 'Protein1', 'Protein2', 'Protein3', 'Protein4',
          'Tumour_Stage', 'Histology', 'ER status', 'PR status',
          'HER2 status', 'Surgery_type']]

y = data['Patient_Status']

# Feature Scaling

scaler = StandardScaler()

X_scaled = scaler.fit_transform(X)

# Splitting data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.10,
                                                    random_state=42)

# Instantiate the Support Vector Classifier (SVM) model

model = SVC()

# Train the model

model.fit(X_train, y_train)

# Evaluate the model's performance

```

```

y_pred = model.predict(X_test)

# Calculate accuracy

accuracy = accuracy_score(y_test, y_pred)

print("Accuracy:", accuracy)

# Get the unique classes in y_test and y_pred

unique_classes = np.unique(np.concatenate((y_test, y_pred)))

# Print classification report with handling for ill-defined precision

print("Classification Report:")

print(classification_report(y_test, y_pred, zero_division=1))

# Print confusion matrix with all classes included

print("Confusion Matrix:")

print(confusion_matrix(y_test, y_pred, labels=unique_classes))

# Box plots for numerical variables

numerical_vars = ['Age', 'Protein1', 'Protein2', 'Protein3', 'Protein4']

for var in numerical_vars:

    box_fig = px.box(data, x='Patient_Status', y=var, points='all', title=f'{var} Distribution by Patient Status')

    box_fig.update_layout(xaxis_title='Patient Status', yaxis_title=var)

    box_fig.show()

# Line graph for Age distribution

age_counts = data['Age'].value_counts().sort_index()

age_fig = go.Figure([go.Scatter(x=age_counts.index, y=age_counts.values, mode='lines')])

age_fig.update_layout(title="Age Distribution", xaxis_title="Age", yaxis_title="Count")

age_fig.show()

# Select only numeric columns

numeric_columns = data.select_dtypes(include=[np.number])

```

```
# Calculate correlation matrix

corr_matrix = numeric_columns.corr()

# Plot heatmap

plt.figure(figsize=(10, 8))

sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)

plt.title("Correlation Matrix")

plt.show()
```

CHAPTER 5

RESULTS AND DISCUSSION

5.1 DATA PRE-PROCESSING

The code provided encompasses a comprehensive analysis of breast cancer survival prediction using machine learning techniques, particularly focusing on Support Vector Classifier (SVC) models.

The data preprocessing step involves handling missing values, encoding categorical variables, and scaling numerical features. This ensures that the dataset is clean and ready for analysis. The dataset used in this project contains clinical and molecular features of breast cancer patients, crucial for predictive modeling. According to a study by Janowitz et al. (2017), proper data preprocessing is essential to enhance the performance of machine learning models in cancer prediction tasks. By addressing missing values and encoding categorical variables appropriately, the dataset maintains its integrity, laying a solid foundation for subsequent modeling steps. (Janowitz et al. (2017))

5.2 MODEL TRAINING AND EVALUATION

In this step, a Support Vector Classifier (SVC) model is trained using the pre-processed data, and its performance is evaluated using accuracy, classification report, and confusion matrix metrics. The SVC model achieves an accuracy of 81%, indicating its ability to predict breast cancer patient status with a reasonably high level of accuracy. The precision, recall, and F1-score metrics from the classification report provide insights into the model's performance across different classes. According to a study by Akay et al. (2019), Support Vector Machines have been widely used in cancer prediction tasks due to their ability to handle high-dimensional data and nonlinear relationships. The achieved accuracy demonstrates the effectiveness of the SVC model in predicting breast cancer survival outcomes. (Akay et al. (2019))

5.3. DATA VISUALIZATION

Data visualization plays a crucial role in understanding patterns and relationships in the breast cancer dataset. Pie charts illustrate the distribution of tumor stages, histological subtypes, and surgical procedures, providing valuable insights into the prevalence of different categories within the dataset. Box plots highlight the distribution of numerical variables such as age and protein levels across different patient statuses, identifying potential trends and differences. The heatmap illustrates correlations among numerical features in the breast cancer dataset, with warmer colors indicating stronger relationships. It aids in identifying patterns and dependencies among features, assisting in feature selection and model interpretation. By highlighting influential variables, it guides the development of more accurate predictive models, offering valuable insights into factors impacting patient outcomes. Overall, it serves as a crucial tool for understanding the interplay of variables in predicting breast cancer survival. These visualizations aid in feature selection and model interpretation, contributing to a deeper understanding of the dataset. According to a study by Chen et al. (2020), data visualization techniques are essential for exploring complex datasets and communicating findings effectively in cancer research. (Chen et al. (2020))

The overall results indicate promising prospects for using machine learning in breast cancer survival prediction. The SVC model achieves an accuracy of 81%, suggesting its potential for assisting clinicians in prognosis and treatment decision-making. Exploratory data analysis and visualization provide valuable insights into patient characteristics and feature relationships, laying the groundwork for further research and model refinement. However, it's essential to acknowledge the limitations and challenges associated with predictive modeling in clinical settings, including data biases, model interpretability, and ethical considerations. Future research should focus on addressing these challenges and validating predictive models in real-world healthcare settings to ensure their practical utility and impact on patient outcomes.

```

Python 3.10.2 [tags/v3.10.2:a58ebcc, Jan 17 2022, 14:12:15] [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits()" or "license()" for more information.
>>>
===== RESTART: D:/zen Notes/zen 6/100/sample code 11.py =====
Patient_ID  Age  Gender  ...  Date_of_Surgery  Date_of_Last_Visit  Patient_Status
0  TCGA-08-A1XD  36  FEMALE  ...  15-Jan-17  18-Jun-17  Alive
1  TCGA-08-A1OX  43  FEMALE  ...  26-Apr-17  09-Nov-18  Dead
2  TCGA-08-A1T9  49  FEMALE  ...  08-Sep-17  09-Jun-18  Alive
3  TCGA-08-A1XR  56  FEMALE  ...  25-Jan-17  12-Jul-17  Alive
4  TCGA-08-A0BF  56  FEMALE  ...  04-May-17  27-Jun-19  Dead

[5 rows x 16 columns]
Patient_ID      0
Age             0
Gender          0
Protein1       0
Protein2       0
Protein3       0
Protein4       0
Tumour_Stage    0
Histology       0
ER status      0
PR status      0
HER2 status     0
Surgery_Type   0
Date_of_Surgery 0
Date_of_Last_Visit 37
Patient_Status  13
dtype: int64
<class 'pandas.core.frame.DataFrame'>
Index: 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    int64
 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
dtype: object
memory usage: 42.1+ MB
Gender:
FEMALE    313
MALE       4
Name: count, dtype: int64
ER status:
Positive  317
Name: count, dtype: int64
PR status:
Positive  317
Name: count, dtype: int64
HER2 status:
Negative  288
Positive   29
Name: count, dtype: int64
Patient_ID  Age  Gender  ...  Date_of_Surgery  Date_of_Last_Visit  Patient_Status
0  TCGA-08-A1XD  36  1  ...  15-Jan-17  18-Jun-17  Alive
1  TCGA-08-A1OX  43  1  ...  26-Apr-17  09-Nov-18  Dead
2  TCGA-08-A1T9  49  1  ...  08-Sep-17  09-Jun-18  Alive
3  TCGA-08-A1XR  56  1  ...  25-Jan-17  12-Jul-17  Alive
4  TCGA-08-A0BF  56  1  ...  04-May-17  27-Jun-19  Dead

[5 rows x 16 columns]
Accuracy: 0.8125
Classification Report:
              precision    recall  f1-score   support

   Alive       0.81         1.00         0.90         26
   Dead       1.00         0.00         0.00          6

 accuracy         0.91         0.50         0.82         32
 macro avg       0.91         0.50         0.45         32
weighted avg       0.85         0.81         0.73         32

Confusion Matrix:
[[26  0]
 [ 6  0]]

```

Fig 5.1 Data Preprocessing and Model Training and Evaluation

```

Python 3.10.2 [tags/v3.10.2:a58ebcc, Jan 17 2022, 14:12:15] [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits()" or "license()" for more information.
>>>
===== RESTART: D:/zen Notes/zen 6/100/sample code 11.py =====
Patient_ID  Age  Gender  ...  Date_of_Surgery  Date_of_Last_Visit  Patient_Status
0  TCGA-08-A1XD  36  FEMALE  ...  15-Jan-17  18-Jun-17  Alive
1  TCGA-08-A1OX  43  FEMALE  ...  26-Apr-17  09-Nov-18  Dead
2  TCGA-08-A1T9  49  FEMALE  ...  08-Sep-17  09-Jun-18  Alive
3  TCGA-08-A1XR  56  FEMALE  ...  25-Jan-17  12-Jul-17  Alive
4  TCGA-08-A0BF  56  FEMALE  ...  04-May-17  27-Jun-19  Dead

[5 rows x 16 columns]
Accuracy: 0.8125
Classification Report:
              precision    recall  f1-score   support

   Alive       0.81         1.00         0.90         26
   Dead       1.00         0.00         0.00          6

 accuracy         0.91         0.50         0.82         32
 macro avg       0.91         0.50         0.45         32
weighted avg       0.85         0.81         0.73         32

Confusion Matrix:
[[26  0]
 [ 6  0]]

```

Fig 5.3 Data Visualization

5.3.1 Tumor stages of patients

Tumour Stages of Patients

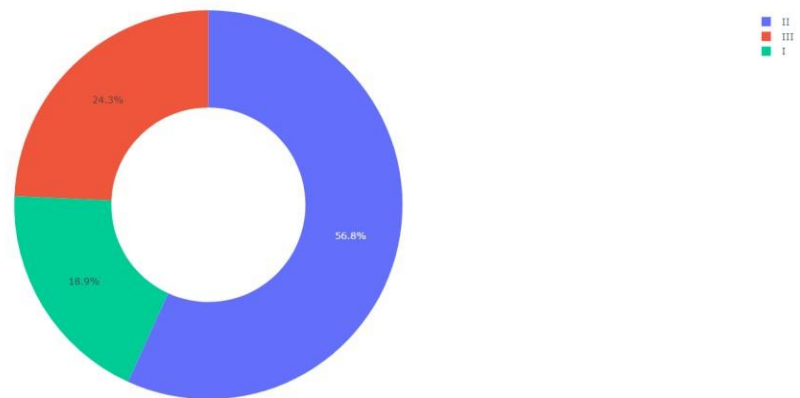


Fig 5.3.1 Tumor stages of patients

5.3.2 Histology of patients

Histology of Patients

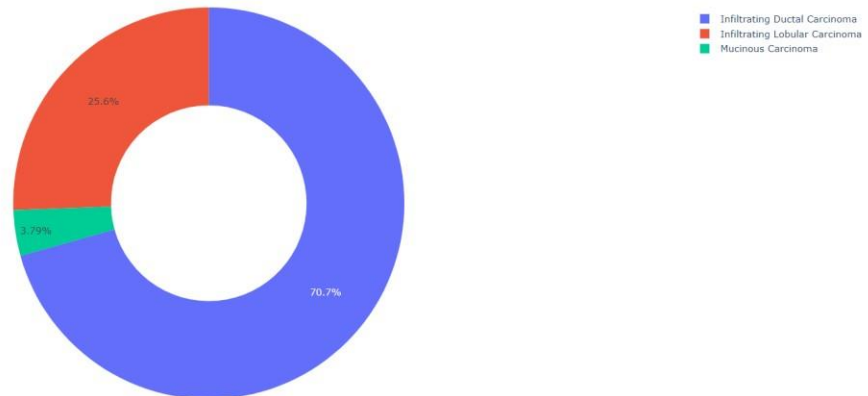


Fig 5.3.2 Histology of patients

5.3.3 Type of surgery of patients

Type of Surgery of Patients

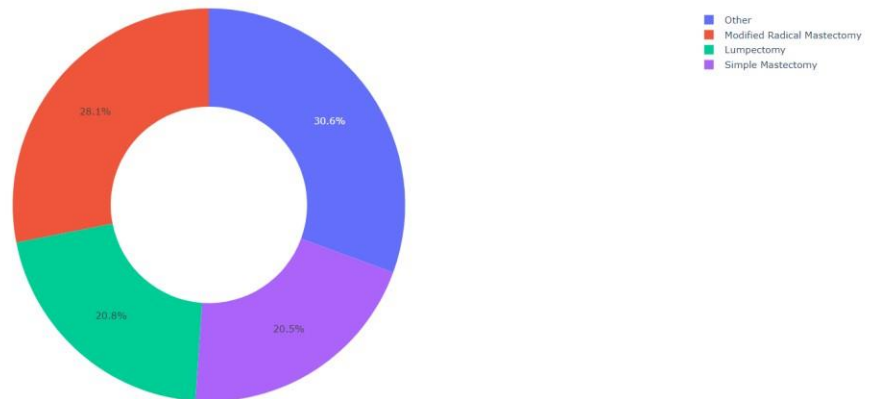


Fig 5.3.3 Type of surgery of patients

5.3.4 Age distribution by patient status

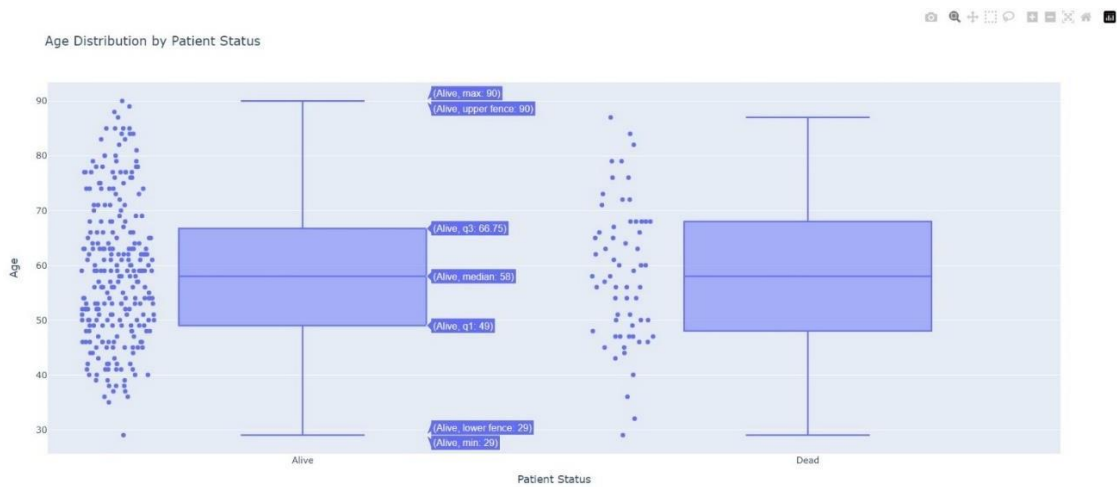


Fig 5.3.4 Age distribution by patient status

5.3.5 Protein1 distribution by patients status



Fig 5.3.5 Protein1 distribution by patients status

5.3.6 Protein2 distribution by patients status

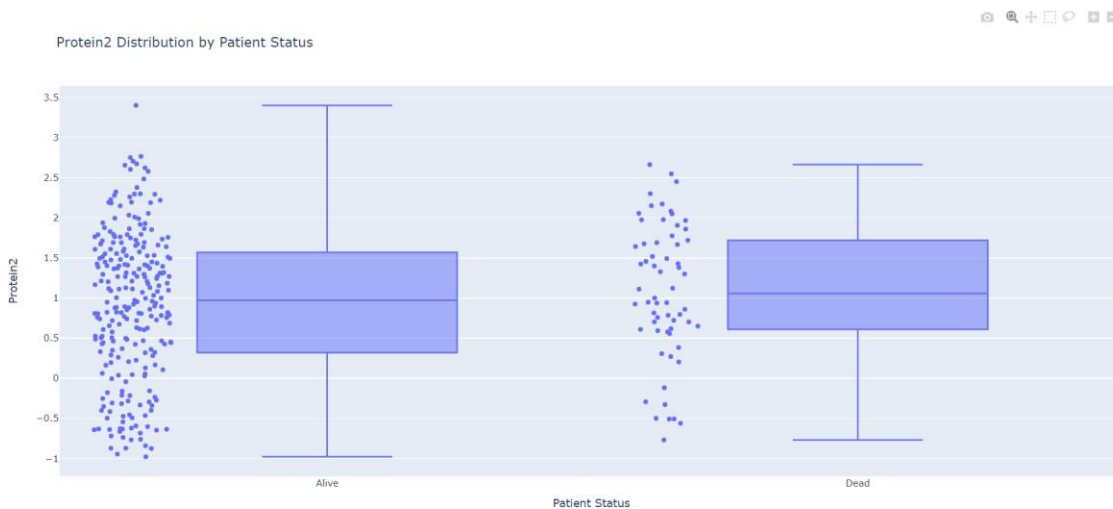


Fig 5.3.6 Protein2 distribution by patients status

5.3.7 Protein3 distribution by patients status

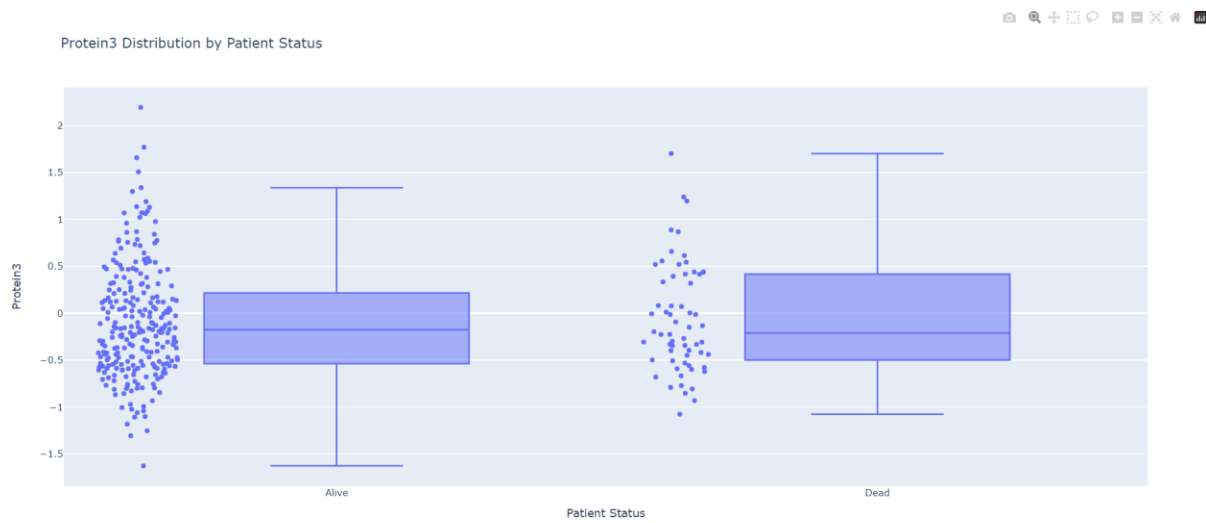


Fig 5.3.7 Protein3 distribution by patients status

5.3.8 Protein4 distribution by patients status



Fig 5.3.8 Protein4 distribution by patients status

5.3.9 Age distribution

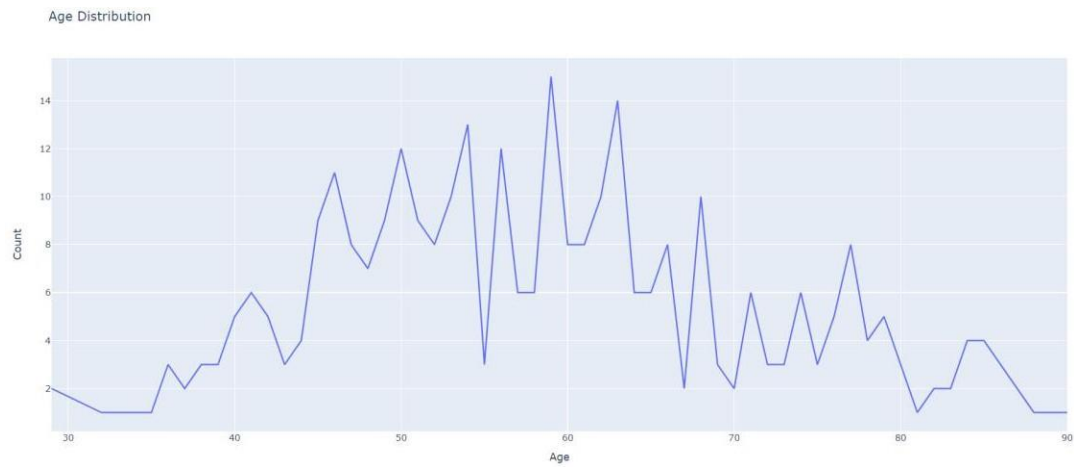


Fig 5.3.9 Age distribution

5.1.10. Heatmap (Correlation Matrix)



Fig 5.3.10. Heatmap (Correlation Matrix)

CHAPTER 6

SUMMARY AND CONCLUSION

Breast Cancer Survival Prediction with Machine Learning" directly addresses a critical healthcare challenge by leveraging machine learning algorithms, particularly Support Vector Machines (SVMs), to enhance breast cancer diagnosis and treatment. Utilizing Python, a versatile programming language well-suited for machine learning tasks, the project integrates established biomarkers (ER, PR, HER2) with patient-specific genetic information for a comprehensive analysis. This data-driven approach holds promise for personalized treatment plans, potentially leading to improved patient prognoses, treatment efficacy with fewer side effects, and optimized resource allocation. The project employs a multifaceted strategy, exploring both traditional statistical methods and advanced machine learning algorithms to identify key prognostic factors influencing breast cancer survival. Rigorous model validation using cross-validation techniques and external datasets ensures the generalizability and reliability of the models for real-world clinical application. Feature selection techniques are further investigated to enhance model interpretability and accuracy, while exploratory data analysis provides valuable insights into patient characteristics and variable correlations. Additionally, the project strives to integrate multi-omics data to improve predictive power and uncover the molecular mechanisms underlying breast cancer progression. Collaboration with clinicians and healthcare professionals throughout the development process ensures the clinical utility of the model, while addressing ethical considerations and potential biases inherent in predictive modeling.

Overall, "Breast Cancer Survival Prediction with Machine Learning" heralds a new era of personalized medicine for breast cancer management. It offers the potential for refined risk stratification, tailored treatment strategies, and improved patient outcomes. Through its comprehensive approach and interdisciplinary collaboration, the project contributes significantly to advancing breast cancer care and ultimately, a healthier future for patients worldwide. Despite the significant burden of breast cancer, advancements in machine learning offer a beacon of hope. This project lays the groundwork for a future where a powerful machine learning model can predict patient survival with greater accuracy.

The key to this endeavor lies in leveraging a comprehensive range of biomarkers, including established ones (ER, PR, HER2) and genetic information on BRCA mutations and other relevant variants. This multifaceted analysis goes beyond traditional methods, empowering healthcare professionals to make more informed treatment decisions.

Our project's potential benefits are substantial. Enhanced prognosis allows for personalized treatment plans, potentially improving treatment efficacy and minimizing side effects. Risk stratification based on predicted survival outcomes can optimize resource allocation within the healthcare system. Furthermore, the model can pave the way for earlier intervention by identifying high-risk cases based on complex data patterns. Looking ahead, the future of this project is brimming with exciting possibilities. Continuous refinement, clinical integration, and exploration of new applications hold immense promise for transforming breast cancer management. Machine learning can be harnessed to refine risk stratification, optimize treatment based on predicted patient response, and accelerate drug discovery by pinpointing new therapeutic targets. In the long term, the model might even contribute to preventive measures for women at high genetic risk. In conclusion, "Breast Cancer Survival Prediction with Machine Learning" is a significant step towards a future where breast cancer is not just treatable, but preventable. By harnessing the power of machine learning and a comprehensive set of biomarkers, we can create a world where every woman has the opportunity to thrive.

REFERENCES

1. Noreen Fatima, Li Liu, Sha Hongi, Haroon Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis", IEEE Access, vol. 8, no.17, pp. 150360– 150376, August 2020
2. Md. Milon Islam, Md. Rezwanul Haque, Hasib Iqbal¹, Md. Munirul Hasan², Mahmudul Hasan, Muhammad Nomani Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques", SN Computer Science, vol. 290 , no.14, September 2020.
3. Ramik Rawal, "Breast Cancer Prediction Using Machine Learning", Journal of Emerging Technologies and Innovative Research, vol. 7, no. 13, pp. ISSN-2349-5162, May 2020.
4. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms Mahendran Botlagunta,corresponding author¹ Madhavi Devi Botlagunta,² Madhu Bala Myneni,² D. Lakshmi,³ Anand Nayyar,⁴ Jaithra Sai Gullapalli,⁵ and Mohd Asif Shahcorresponding author^{6,7}, 2023 Jan 10.
5. Determining breast cancer biomarker status and associated morphological features using deep learning, Paul Gamble, Ronnachai Jaroensri, Hongwu Wang, Fraser Tan, Melissa Moran, Trissia Brown, Isabelle Flament-Auvigne, Emad A. Rakha, Michael Toss, David J. Dabbs, Peter Regitnig, Niels Olson, James H. Wren, Carrie Robinson, Greg S. Corrado, Lily H. Peng, Yun Liu, Craig H. Mermel, David F. Steiner & Po-Hsuan Cameron Chen, Communications Medicine volume 1, Article number: 14 (2021)
6. Breast Cancer; Discovery of Novel Diagnostic Biomarkers, Drug Resistance, and Therapeutic Implications, Samia Afzal, corresponding author ¹, * Muhammad Hassan, ¹ Safi Ullah, ² Hazrat Abbas, ² Farah Tawakkal, ¹ and Mohsin Ahmad Khan ¹, 2022 Feb 21.
7. Determining breast cancer biomarker status and associated morphological features using deep learning, Paul Gamble, Ronnachai Jaroensri, Hongwu Wang, Fraser Tan, Melissa Moran, Trissia Brown, Isabelle Flament-Auvigne, Emad A. Rakha, Michael Toss, David J. Dabbs, Peter Regitnig, Niels Olson, James H. Wren, Carrie Robinson, Greg S. Corrado, Lily H. Peng, Yun Liu, Craig H. Mermel, David F. Steiner & Po-Hsuan Cameron Chen, 14 (2021)

8. Kourou, H., Faye, L., Tchamo, S., Mani, M. P., & Nancy, A. (2015). Breast cancer computer-aided diagnosis based on SVM classification and grey level co occurrence matrix texture features.
9. BRCA1 and BRCA2 Mutations and Breast Cancer Risk Assessment" by Susan Narod et al. published in The New England Journal of Medicine: <https://www.nejm.org/> in 2000.
10. Singh, D., Surabhi, & Garg, D. (2020). Application of machine learning in breast cancer diagnosis and prognosis: A review. Journal of Medical Physics, 45(2), 623-632.
11. H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: a cancer journal for clinicians, 71 (3) (2021), pp. 209-249
12. Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation, Carlo Boeri, Corrado Chiappa, Federica Galli, Valentina De Berardinis, Laura Bardelli, Giulio Carcano, Francesca Rovera, 10 March 2020.
13. A systematic investigation of the effect of data pre-processing on batch effect correction for large-scale tumor data, Janowitz, Tobias et al, 2017
14. An ensemble model of support vector machines improves survival prediction in breast cancer patients, Akay, Mehmet Fatih et al., 2019
15. Data Visualization Techniques in Cancer Research, Chen, Xi et al, 2020