Integrative Analysis of Breast Cancer Genomic Data for Personalized Treatment

Submitted By: Nabila Noor

Student Number: 501216808

Supervisor's Name: Dr. Ceni Babaoglu

Date of Submission: 17th July, 2023

Ryerson University

Table of Contents

Abstract:

Breast cancer is the most common disease in women, with 2.1 million women affected every year. Accurate prognosis and survival time estimations are the most crucial steps in the clinical decision-making process for cancer patients. Breast cancer patients with the same clinical features and stage of the illness may respond differently to therapy and have a varying overall survival rate.

The primary objectives of this study are to predict the most suitable treatment option based on clinical and genetic data. Alongside, we are aiming to group genes determine the relationship between clinical characteristics and post-treatment survival.

The dataset used for this project is called Breast Cancer Gene Expression Profiles (METABRIC), which includes observations from 1904 patients with breast cancer and has 695 attributes, including 31 clinical attributes, a m-RNA level z-score for 331 genes, and mutation information for 175 genes (Kaggle, nd). Initially, Professor Sam Aparicio of the British Columbia Cancer Centre and Professor Carlos Caldas of the Cambridge Research Institute collected the data, and published it in Nature Communications (Pereira et al., 2016).

To predict suitable treatment options for patients, a subset of the survivors is trained to classify treatment options based on gene expression and clinical data.

For predicting survival, clinical and gene expression data are used as well. The models used for both predictions are Support Vector Machines, Random Forest and Gradient Boosting, SMOTE and ROSE methods.

Clustering techniques like k-means and hierarchical clustering are used to identify gene groups. Our goal is to identify patterns that correlate to overall survival through the examination of the mRNA expression levels of genes, furthering our comprehension of the molecular aspects of the disease.

Finally, survival analysis using statistical methods such as Cox proportional hazards regression and Kaplan-Meier survival curves are performed based on relevant clinical or genetic criteria.

For Data analysis and prediction Python programming languages is used, along with Tableau for data visualization.

This study attempts to improve patient outcomes for breast cancer by analyzing clinical and genomic data. The suggested methods, which include statistical analysis, clustering techniques, and classification algorithms, offer a thorough framework for answering the research questions related to breast cancer prognosis and treatment outcomes. This study may facilitate planning for surgery and personalized treatment options. These results might ultimately lead to more individualized treatment plans and better patient care.

Literature Review:

The original data was collected from a collaborative study between the British Columbia Cancer Agency and the Cancer Research UK Cambridge Institute A thorough analysis of the somatic mutation patterns of 2,433 breast tumors has been published, improving the genomic and transcriptome landscapes of these diseases. The objective of this study was to identify genes associated with the development of breast cancer. It has identified 93 genes that, when mutated, promote the development of breast cancer, and some of these genes make good drug targets (Pereira et al., 2016)

The data from the METABRIC cohort is also used in other studies. In the first study, the METABRIC cohort of breast cancer patients who underwent hormone and/or chemotherapy treatments had their gene expression profiles examined. In order to forecast how patients would respond to various medications and how long they would live, the researchers used a machine learning technique. However, other medications, including methotrexate, tamoxifen, doxorubicin, epirubicin, and 5-fluorouracil, exhibited respectable accuracy. They discovered that gene expression profiles for paclitaxel were the most reliable in predicting survival. The researchers discovered a group of genes, including ABCB11, ABCB1, ABCC1, BAD, ABCC10, BCL2L1, BCL2, BMF , CYP3A4, CYP2C8, MAP4, MAP2, MAPT, SLCO1B3, NR1I2, TUBB1, TTUBB4B, and TUBB4A, that were particularly helpful in predicting survival. They created their models using a variety of machine learning techniques, including support vector machines and random forests. (Rezaeian et al., 2017)

The objective of the second study was to identify integrative clusters (IntClusts) that reflect distinct molecular subtypes. Researchers performed an extensive examination of genomic and transcriptome data from a large cohort of 2000 breast tumors. These IntClusts had distinctive genetic traits and showed variations in responsiveness to various therapies. Furthermore, it was shown that they were connected to particular clinicopathological characteristics such as receptor status, tumor grade, and lymphocytic infiltration.

This study brought to light the drawbacks of correctly classifying breast cancer subtypes purely based on individual clinical characteristics. Even though several clinical traits were correlated with particular IntClusts, no one clinical characteristic was able to accurately classify the wide range of subtypes. These results underline how important it is to use genetic data when deciding how to diagnose and treat breast cancer.

In order to analyze breast cancer samples, the study used genomic and transcriptome profiling technologies, such as Affymetrix SNP 6.0 and Illumina HT-12 v3. Based on clinicopathological factors such as NPI category, tumor size, lymphocytic infiltration, grade, receptor status, histological subtype, and lymph node status, logistic regression models were used to predict molecular subtypes. These models provided insightful information on the relationships between clinical characteristics and molecular subtypes.

The study's findings highlight the intricacy of breast cancer and the shortcomings of conventional categorization techniques. The discovery of integrative clusters (IntClusts) based on genomic and transcriptome profiles has demonstrated the heterogeneity of breast cancer subtypes, with each displaying particular genetic changes, divergent therapeutic responses, and

varying prognoses. the use of genetic information in the detection and treatment of breast cancer. (Mukherjee et al., 2018)

The METABRIC cohort data was used in another study in 2019 to find biomarkers to improve breast cancer detection or prognosis. A data-integration strategy was used to find sub-network biomarkers that might forecast the outcomes of breast cancer therapy, such as disease-free survival and overall survival at five years and over the long term. The prediction ability of gene sub-networks was assessed using gene expression data, and the search for potential gene sub-networks was guided by the protein-protein interaction network. By estimating the predictive power of a group of genes, a score was presented to narrow the search space; as a result, only the candidates with the highest scores were considered by the Support Vector Machine classifier. The biological significance of the sub-networks was further examined using pathway data and cancer-related genes from the literature after the sub-networks with the best classification performance for all seed genes were chosen. The chosen sub-networks produced incredibly precise results and comprise genes linked to several cancer pathways. (Pham et al., 2019)

Olga Zolotareva and colleagues (2022) used “DESMOND 2.0” a biclustering algorithm on the METABRIC cohort along with the TCGA-BRCA cohort to establish the robustness of the bioclusters compared to others. This algorithm looks for submatrices in a two-dimensional sample-gene matrix with a certain pattern, providing a viable alternative to traditional clustering. In order to increase the reliability of the findings, DESMOND leverages interaction networks as constraints. Clustering techniques are typically used for unsupervised patient categorization

based on omics data, however, they may be ineffective for datasets with many patterns that overlap in the rows and columns. (Zolotareva et al., 2022)

In 2023, the METABRIC data set was used to suggest a classifier that is built on a data augmentation pipeline employing deep learning algorithms. Since conventional prognostic tests are only accepted for particular clinical symptoms or disease characteristics, they are limited in their capacity to detect high-risk individuals. Deep learning algorithms have the ability to get over these restrictions, but it's difficult since there are so many variables in omics datasets relative to the number of patients.

To overcome this difficulty, the researchers created a classifier utilizing a data augmentation pipeline that consists of an embedded auxiliary classifier and a Wasserstein generative adversarial network (GAN) with gradient penalty. T-GAN-D was the method used with 1244 patients from the METABRIC breast cancer cohort. The reiterative GAN-based training approach proved to be a strong classifier capable of reliably stratifying patients based on whole transcriptome data across separate and diverse breast cancer cohorts. (Guttà et al., 2023)

A lot of work has been done in the last decade to aid the decision-making process of oncologists with regard to the detection and management of breast cancer using new bioinformatic techniques and computational tools powered by artificial intelligence. There have been several reported effective uses of machine learning (ML) for image processing, particularly the use of deep neural networks and convolutional neural networks to identify tumor and lymph node locations. In order to grasp molecular-level disease, recent high-throughput molecular quantifications, or quantitative omics approaches, have made it possible to simultaneously

monitor thousands of molecules. Deep learning, network analysis, clustering, and dimension reduction have all been used to analyze this data, which includes gene expression, protein, metabolite, and methylation profiles. The prognosis, metastasis, and treatment outcomes of breast cancer have been predicted using ML, incorporating several clinical-pathological parameters in order to analyze complicated structures of a higher number of variables, becoming an upcoming tool to develop personalized breast cancer treatment. (Sugimoto et al., 2023)

To our current understanding, no studies that used the METABRIC dataset in the past have dealt with the particular research concerns examined in this work. While previous research has mostly concentrated on genomic and transcriptome studies using bioinformatics methods or deep learning approaches, no equivalent work from a data analysis standpoint has been uncovered. The dataset utilized in this study was collected from Kaggle, a website that hosts a number of user-generated notebooks linked to the dataset (Kaggle, nd). Although some initial analyses may be similar, this study's final goals and methodology differ greatly from those of other studies, for example, the use of the SMOTE or ROSE algorithms for unbalanced target variables, clustering genes to identify the aggressiveness of the mutations, etc.

This work offers a novel viewpoint on analyzing the METABRIC dataset by utilizing a distinct set of analytical tools with a focus on data science. This advances our knowledge of breast cancer, offers insightful information for future studies, and enables to examine research problems that have not yet been addressed.

Tentative Methodology:

For the project, the data set is subdivided into three datasets for ease of initial analysis. For initial analysis, univariate, bivariate, and multivariate analyses are done. As a part of the exploratory data analysis, clustering of the genes and mutations based on survival is done. Following that is dimensionality reduction, where Principal Component Analysis and feature selection are done to find out the variables that offer the most variability in the dataset. After that, an experimental design is performed by splitting the test and training data using cross- validation and/or bootstrapping. For the prediction modeling Support Vector Machines, Random Forest, Gradient Boosting, SMOTE, and ROSE methods are used. Finally, the models are evaluated and compared to each other, and any improvements are made if necessary.

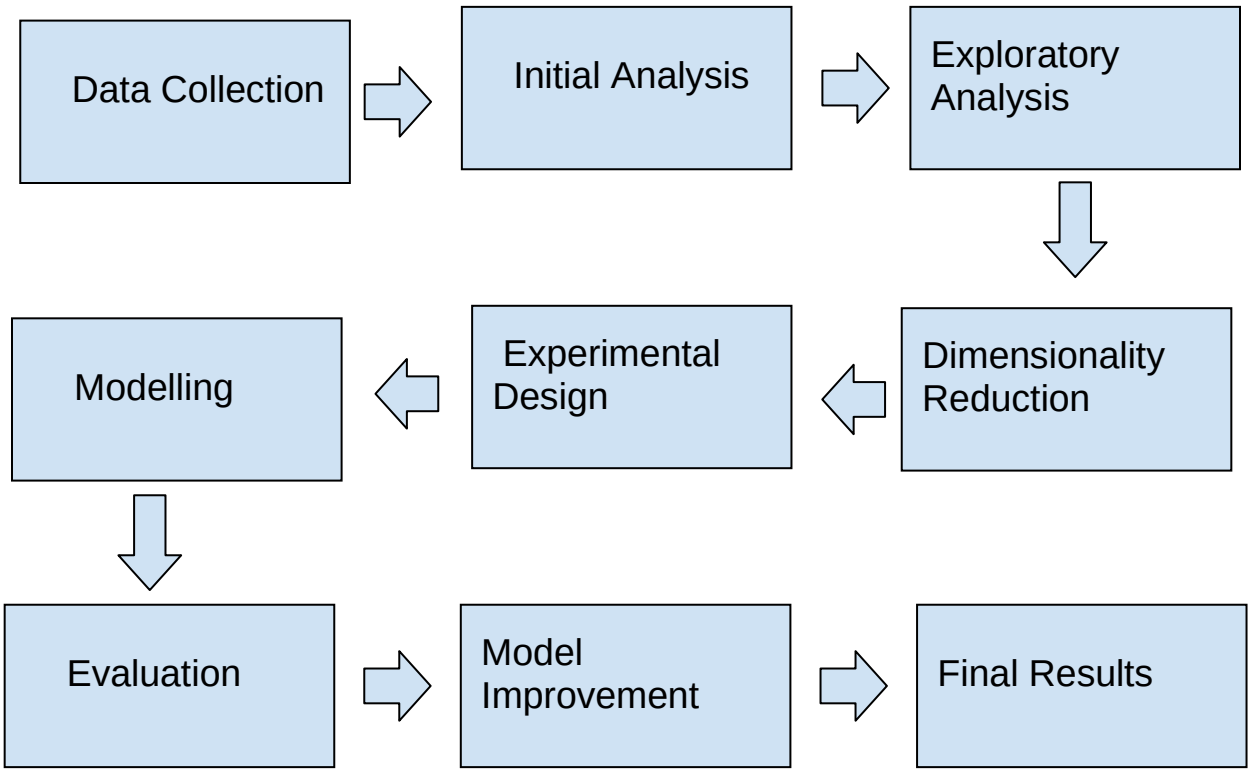


Figure 1: Flow chart for tentative methodology.

Data Description:

The data set has observations of 693 attributes from 1904 patients. The attributes can be divided into three categories:

- Clinical attributes
- Genetic attributes
- Mutations

Depending on the attributes, three subsets of the dataset are created for simplicity of analysis: df_clinical, df_genetic, and df_mutation. The descriptions of the three subset data frames are given below:

Clinical Attributes:

The first 31 columns of the main dataset are related to the clinical observations of patients. A dataframe named df_clinical is created without the patient_id column. The descriptions of the clinical attributes in the dataset are as follows:

Table 1: Description of clinical attributes.

Attribute Name	Data Type	Description
age_at_diagnosis	float	Patient's age at time of diagnosis
type_of_breast_surgery	object	Type of breast cancer surgery: 1. MASTECTOMY, which is a surgical procedure used to cure or prevent breast cancer by removing all breast tissue from the affected breast. 2- BREAST CONSERVING, which describes a procedure in which just the cancerous portion of the breast is removed.
cancer_type	object	Cancer types: 1. Breast Cancer

		2.Breast Sarcoma (Sarcomas are uncommon malignancies that affect the deep skin tissues, muscles, blood vessels, nerves, fat, and fibrous tissues in addition to the bones and soft tissues.)
cancer_type_detailed	object	Breast cancer types in detail: 1- Breast Invasive Ductal Carcinoma 2- Breast Mixed Ductal and Lobular Carcinoma 3- Breast Invasive Lobular Carcinoma 4- Breast Invasive Mixed Mucinous Carcinoma 5- Metaplastic Breast Cancer
cellularity	object	Post chemotherapy cancer

		cellularity , which refers to the amount of tumor cells in the specimen and their arrangement into clusters
chemotherapy	int	If patient has chemotherapy as treatment or not
pam50_+_claudin-low_subtype	object	Pam 50 is a tumor profile test that can help determine if some ER-positive, HER2-negative breast tumors are likely to metastasis, or spread to other organs. Gene expression features, most notably low expression of cell-cell adhesion genes, high expression of epithelial-mesenchymal transition (EMT) genes, and stem cell-like/less differentiated gene expression patterns, identify

		the claudin-low breast cancer subtype.
cohort	float	Group of patents sharing a particular characteristic. Values range from 1-5
er_status_measured_by_ihc	object	A test utilizing immune- histochemistry to assess the expression of estrogen receptors on tumor cells. Results are given as positive or negative.
er_status	object	Presence of estrogen receptors on cancer cells, given as positive /negative.
neoplasm_histologic_grade	int	Determines the aggressiveness of the tumor cells using pathology. Values

		range from 1-3
her2_status_measured_by_snp6	object	Determines the presence of HER2 using advanced next generation sequencing techniques.
her2_status	object	Presence or abscense of HER2.
tumor_other_histologic_subtype	object	Based on microscopic analysis of the cancer tissue, the cancer type is classified as Metaplastic, Mucinous, Medullary,Tubular/cribriform ,Lobular, Ductal/NST, Mixed or Other
hormone_therapy	int	If hormonal therapy is used as treatment

inferred_menopausal_state	object	Menopausal state of the patient, pre or post.
integrative_cluster	object	cancer's molecular subtype based on certain gene expression .It takes a value from "4ER+," "3," "9," "7," "4ER-," "5", "8," "10," "1," "2," "6".
primary_tumor_laterality	object	Presence of cancer on right or left breast
lymph_nodes_examined_positive	float	Number of lymph node samples taken during surgery to evaluate their role in cancer.
mutation_count	float	Gene count with relevant mutations.

nottingham_prognostic_index	float	A calculated index derived from tumor size, amount of positive lymph node samples and tumor grade. It helps to determine the prognosis of cancer after surgery.
oncotree_code	object	The Memorial Sloan Kettering Cancer Centre (MSK) created the OncoTree, an open-source ontology, to standardize cancer type diagnosis from a clinical standpoint by giving each diagnosis a distinct OncoTree code.
overall_survival_months	float	Time in months between intervention and death.
overall_survival	object	Alive or dead.

		Target Variable.
pr_status	object	Progesterone receptors are either present or absent in cancer cells. (positive/negative)
radio_therapy	int	Use of radiotherapy for treatment.
3-gene_classifier_subtype	object	Three subtypes of gene classifier It accepts values from 'ER+/HER2- High Prolif','ER-/HER2- ','ER+/HER2- Low Prolif','HER2+' and nan.
tumor_size	float	Size of tumor measured using imaging techniques.
tumor_stage	float	Depending on the cancer's

		interaction with nearby structures, lymph nodes, and distant dissemination, the cancer's stage is determined
death_from_cancer	int	Death from cancer, or other reasons.

There are some missing values in the clinical data frame. The percentage of missing values in the corresponding columns is presented in the following graph.

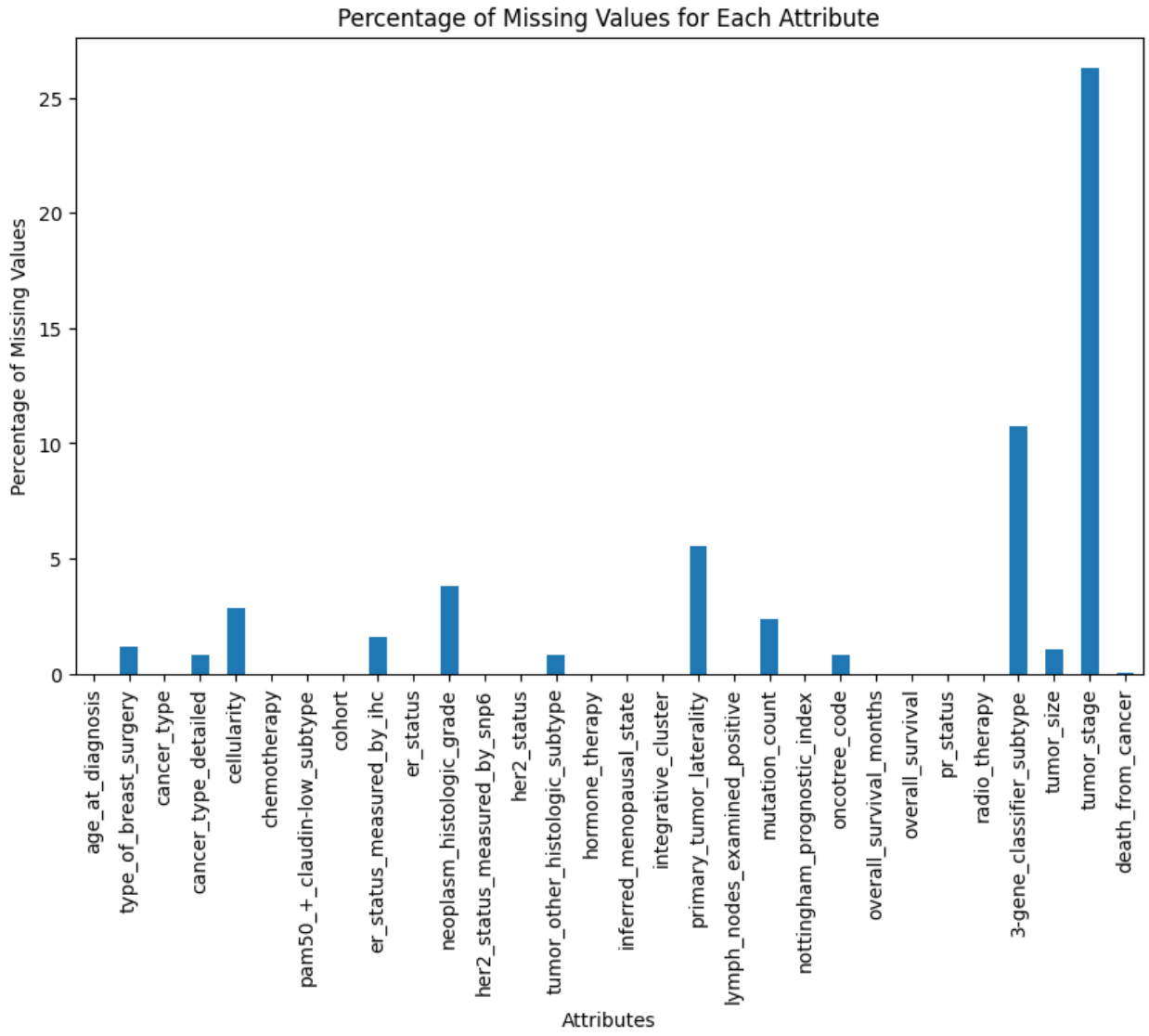


Figure 2: Percentage of missing values

Some of the variables are converted into categorical variables, and their values are counted. Among them, some have an unbalanced distribution. Some of them were removed from the dataset for final modeling. These variables and their counts are given in the table below:

Table 2: Imbalanced categorical variables and their counts

Variable Name	Levels	Count
cancer_type	Breast Cancer	1903
	Breast Sarcoma	1
cancer_type_detailed	Breast Invasive Ductal Carcinoma	1500
	Breast Mixed Ductal and Lobular Carcinoma	207
	Breast Invasive Lobular Carcinoma	142
	Breast Invasive Mixed Mucinous Carcinoma	22

	Breast Metaplastic Breast Cancer	17 1

Other than that, patient_id, cohort, are also removed before final analysis.

Among the numeric variables in the dataset, the mean age of diagnosis is around 61 years with a standard deviation of 12.97, and the maximum and minimum diagnosis ages are 96.29 and 21.93 respectively. It has a normal distribution.

The Nottingham Prognostic Index is a tool to determine cancer prognosis and patient outcomes. It is derived from tumor size, lymph node status, and histological grade. It has a mean of 4.03, maximum of 6.3 and a minimum of 1.00.

For lymph nodes examined positive, the maximum is 45 and the minimum is 0 with a mean and standard deviation of 2.002 and 4.07, respectively, with some outliers.

Mutation count has a maximum value of 6.3 and a minimum of 1, with a mean of 4.03. This variable has some missing values. The observations include outliers.

Overall survival months have a minimum of 0 months, a maximum of 355.20 months, and a mean survival of 125.121 months.

Tumor size has a maximum of 182, a minimum of 15.16, and a mean of 26.23, with a few outliers.

For all the numerical values, the distributions between the survivors and dead patients are almost similar.

There is not much correlation between the numerical variables except between the Nottingham prognostic index and lymph nodes examined positively, along with tumor size and histological grade, which are used as factors to determine the index.

Some observations from the dataset are:

- Type of breast surgery is highly correlated with cancer type.
- Cancer type is highly correlated with tumor size, breast surgery, er status measured by ihc, tumor laterality ,histologic subtype and tumor stage.
- Chemotherapy as a treatment is highly correlated with er status measured by ihc.
- Primary tumor laterality is highly overall correlated with cancer type.
- radio_therapy is highly overall correlated with type_of_breast_surgery
- tumor_stage is highly overall correlated with cancer_type

Genetic Attributes:

This subset data frame has an mRNA expression z score of 331 genes.

According to cBioPortal:

The messenger RNA (mRNA) transcripts (sometimes referred to as the transcriptome or the collection of messenger RNA (mRNA) transcripts) that are produced by a set of genes are detected by the DNA molecules attached to each slide. mRNA molecules from an experimental sample and a reference sample are generally gathered to conduct a microarray analysis.

An individual gene and tumor's relative expression to the gene's expression distribution in a reference population is calculated using mRNA expression data. All of the study's samples make up the reference population. The result that was returned (Z-score) represents the number of standard deviations from the expression mean in the reference population. When compared to normal samples or all other tumor samples, this measurement can help identify whether a gene is up- or down-regulated.

```
z = (expression in tumor sample - mean expression in reference sample) / standard deviation of expression in reference sample
```

There is no missing value or duplicate row in this data. The observed z score in most of the patients for most genes is between 5 and -5. The highest observed z score for a patient is 20.39 by cyp3a7, and the lowest observed score is -0.916 by the ttyh1 gene.

Genes with the highest mean expression levels:

- nfkb2 0.000002
- rad51c 0.000002
- pdgfrb 0.000002
- jag1 0.000002
- psen2 0.000002

Genes with the lowest mean expression levels:

- ahnak -0.000003
- cyp17a1 -0.000002
- nfkb1 -0.000002
- gldc -0.000002
- kmt2d -0.000002

Around 47.8% of the genes showed a positive z score, which means they were up-regulated compared to the reference, and 52.2% showed down-regulation or negative z scores.

Mutations:

The mutation subset of the main dataset, or df_mutation, has mutations found in 173 genes. This subset also does not have any missing or duplicated values. The gene with the most unique mutations is tp53 (343), which is a tumor suppressor gene and well associated with cancer prognosis.

Followed by muc16_mut (298), ahnak2_mut (248), kmt2c_mut (222), and syne1_mut (200) .

All the codes and the results can be found at:

https://github.com/nnabila05/nnabila05--Integrative-Analysis-of-Breast-Cancer-Genomic-Data-for-Personalized-Treatment/blob/main/cind_820.ipynb

Results and Discussion:

Feature Selection:

PCA involves the data to be linearly projected into a primary subspace while maintaining their maximum variance. Singular value decomposition (SVD) may be used to calculate the eigenvectors of the data covariance matrix, which represent the primary components. High-dimensional data sets are frequently reduced in dimension using PCA for feature extraction, visualization, and other uses. The variable k, or the number of retained primary components, is a hyperparameter that may be selected.

For feature selection for the project Principle Component Analysis (PCA) is used. From the PCA it is evident that almost 100% of the variance can be explained by first 300 principle components.

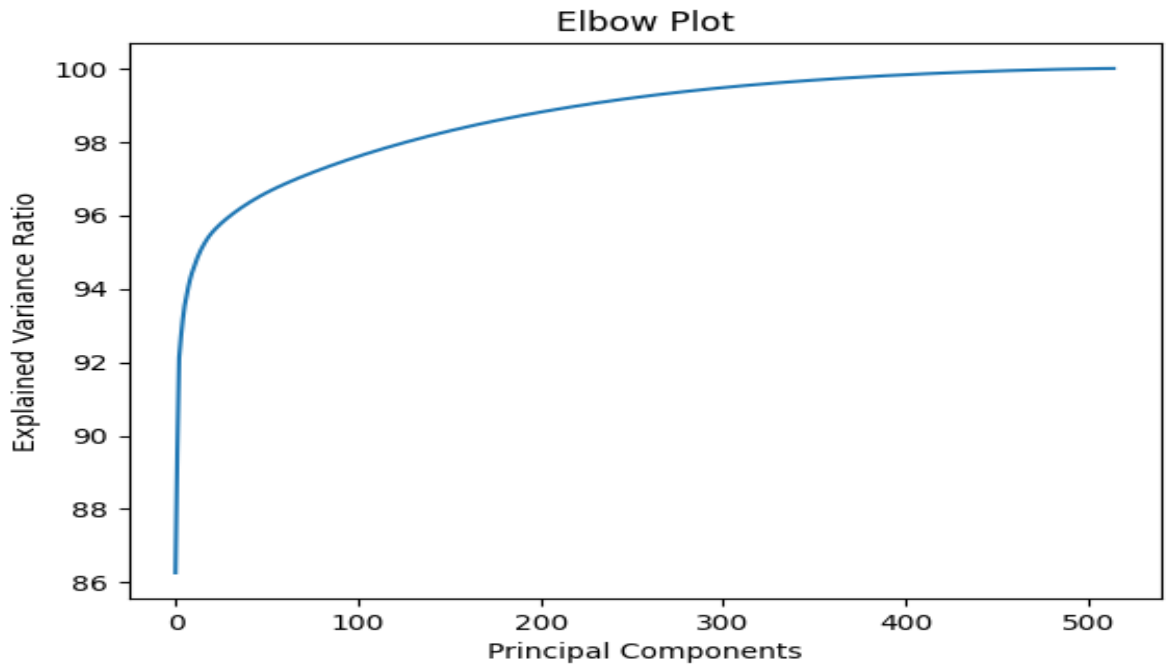


Figure 3: Elbow plot for PCA

Predicting Treatment (chemotherapy):

To find out the best suitable treatment options for the patient, we considered predicting a model using Random Forest, gradient boosting and SVM classifier using 10-fold cross validation. Though in the dataset there are four treatment options, for this project only suitability of chemotherapy is determined, as multioutput multiclassification classification in Scikit learn doesn't have a metric system and deep learning is out of the scope of this project, therefore only suitability of chemotherapy as a treatment option is predicted for this project.

Chemotherapy has imbalanced distribution in the data set which can cause bias in model performance, mislead evaluation metrics and reduce sensitivity of the model.. Therefore, two different resampling algorithms are used before training the models. The sampling algorithms are called SMOTE and ROSE and their results are compared.

SMOTE (synthetic minority oversampling technique) does data augmentation by generating fake data points based on the real ones. SMOTE can be viewed as an improved form of oversampling or as a particular data augmentation procedure. It avoids producing duplicate data points and instead produce synthetic data points that are marginally different from the original data points (Korstanje, 2022).

ROSE (Random Over-Sampling Examples) is a bootstrap-based technique which aids the task of binary classification in the presence of rare classes. It handles both continuous and categorical data by generating synthetic examples from a conditional density estimate of the two classes (Lunardon et al., 2021) .

The first algorithm used to predict chemotherapy as a suitable treatment option is Random Forest. To create predictions or classifications, it makes use of an ensemble of several decision trees. The random forest method produces a consolidated and more accurate result by integrating the outputs of various trees. By averaging many decision trees, Random Forest lessens overfitting and is less susceptible to noise and outliers in the data. For feature selection and data interpretation, it offers a measure of feature relevance (R, 2023).

Gradient boosting is a type of ensemble technique variations where several weak models are built and integrated them to improve performance overall.

The support vector machine algorithm's goal is to locate a hyperplane in an N-dimensional space (N is the number of features) that clearly categorizes the data points. There are several feasible hyperplanes that might be used to divide the two groups of data points. The biggest margin, or the greatest separation between data points for both classes, is what we are looking for in a plane (Gandhi, 2022).

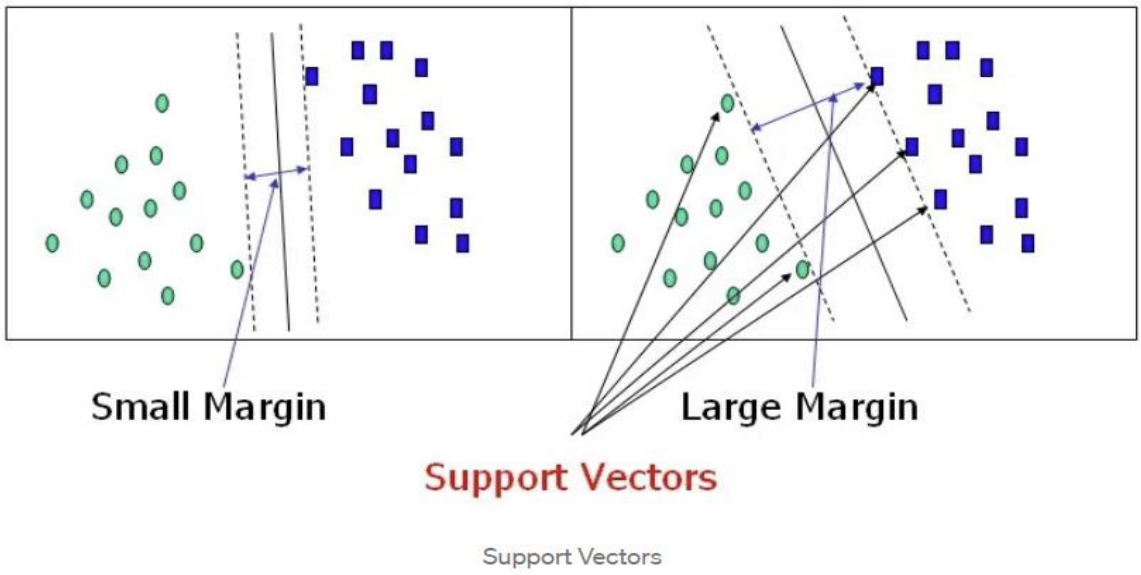


Figure 4: Support vectors with small and large margin (Gandhi, 2022) .

The results from using the three different algorithms with 10-fold cross validation are given below:

Table 3: Classification results after using ROSE

Model Name	Mean Accuracy	Precision	Recall
Random Forest Classifier	99.5	1	3.33
Gradient Boosting Classifier	94.3	65.3	56.6

SVM Classifier	94.8	75.75	62.5

Table 4: Classification results after using SMOTE.

Model Name	Mean Accuracy	Precision	Recall
Random Forest Classifier	96.6	76.1	13.3
Gradient Boosting Classifier	90.8	63.8	57.5
SVM Classifier	95.5	75.3	48.3

Overall, the Random Forest Classifier consistently produces the greatest accuracy among the three algorithms, according to the results with 10-fold cross-validation utilizing the ROSE and SMOTE resampling approaches. Overfitting should be considered, as Random Forests are known to be susceptible.

Given the risk of overfitting, it could be required to adjust the Random Forest hyperparameters. To identify the ideal set of hyperparameters that maximize performance without overfitting the data, methods like grid search or random search can be used. In terms of precision and recall, this model had the lowest balance. Though it could predict most positive instances correctly (high precision), however it could only predict a very low proportion of all positive cases (low recall).

In terms of the resampling methods, ROSE appears to offer somewhat more accurate findings than SMOTE.

Considering all the aspects including accuracy, precision and recall, it can be said that SVM is the most suitable model among the three.

Predicting overall survival:

To predict the overall survival, the same three algorithms are used as before and the results are given below:

Table 5: Evaluation metrics for overall survival prediction

Model Name	Accuracy	Precision	Recall
Random Forest Classifier	0.7373	0.81627	0.6910

Gradient Boosting Classifier	0.8786	0.8844	0.8668
SVM Classifier	0.9873	0.9872	0.9866

In terms of classification accuracy, SVM Classifier has the best accuracy (98.7%), followed by Gradient Boosting Classifier (87.9%) and Random Forest Classifier (73.7%),. This shows that the Random Forest Classifier had the lowest accuracy and the SVM Classifier had the highest amount of properly categorized cases.

The SVM Classifier has the greatest recall (98.7%), followed by the Gradient Boosting Classifier (86.7%), and the Random Forest Classifier (69.1%), which assesses the capacity to properly ide ntify positive cases.

This suggests that the SVM Classifier outperformed the other classifiers in accurately identifying most positive cases in the dataset.

In terms of precision which is the proportion of positives which are actually positive and refers to quality, the SVM Classifier (98.7%) has the best, followed by the Gradient Boosting Classifier (88.4%), and the Random Forest Classifier (81.6%).

The Random Forest Classifier showed a larger percentage of false negatives (139) compared to the other classifiers, which suggests difficulty in accurately detecting positive cases, when the

confusion matrices are analyzed. A better-balanced performance was shown by the Gradient Boosting Classifier and SVM Classifier, which had less false positives and false negatives.

Clustering the Genetic data:

The genetic portion of the data was clustered using k-means clustering algorithm and the adjusted rand index was measured to see if the clustering was random taking overall survival as the dependent variable.

The adjusted rand index is a measurement metric that compares two clustering by taking note of all of the pairings from the n samples and calculates the number of pairs that were allocated to the same or different clusters in the actual and predicted clustering.

Random labelings have an ARI close to 0.0 and 1.0 stands for perfect match. With k =4 the adjusted rand index was 0.02. Which indicates the clustering is random in regard to "death_from_cancer"

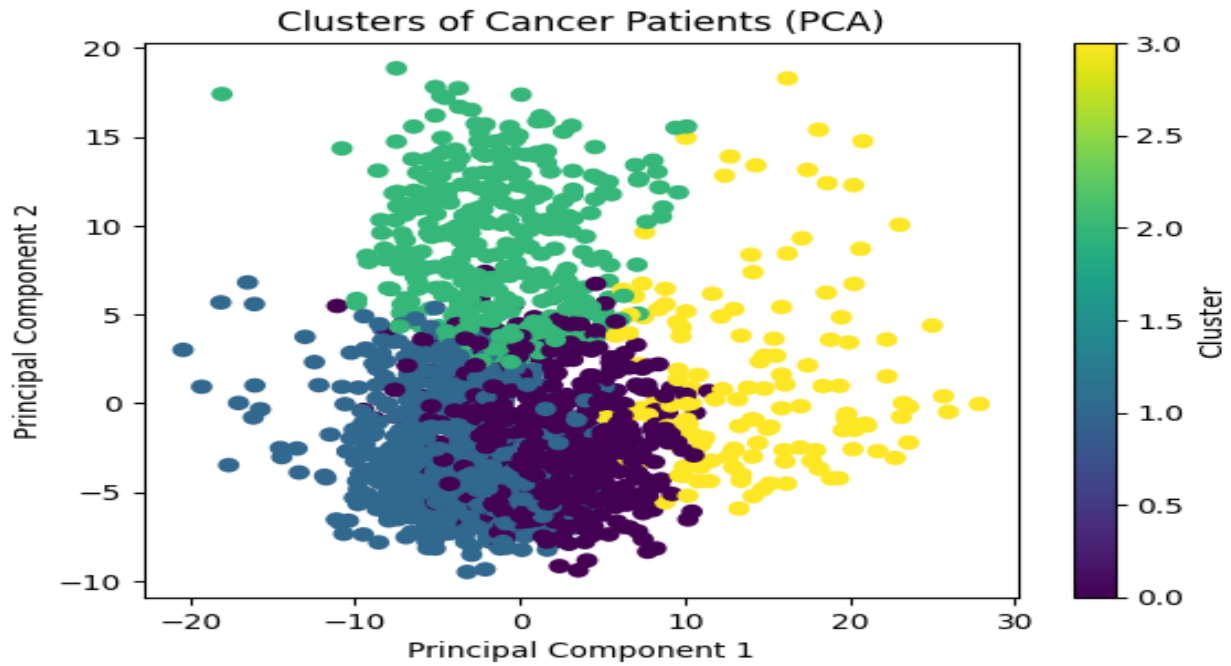


Figure 5: Cluster of patients.

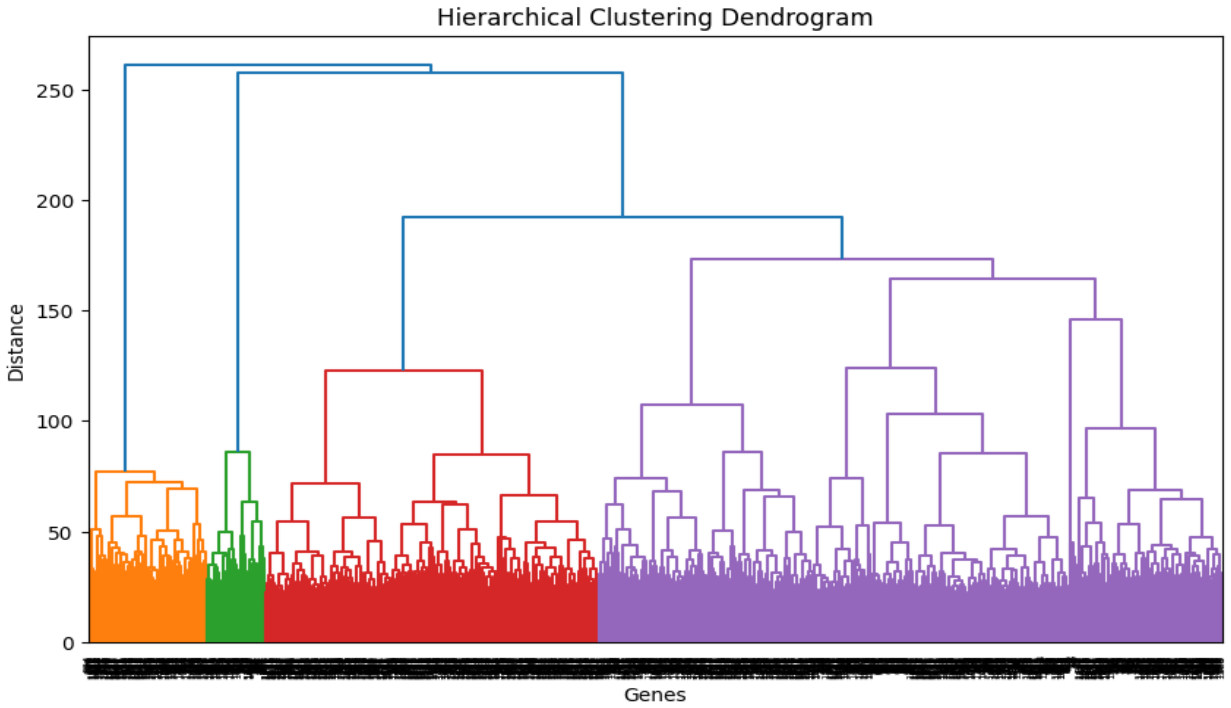


Figure 6: Hierarchical Clustering of genetic data.

Survival Analysis:

Survival analysis is done to get more information of the dataset. The survival function is estimated using the Kaplan-Meier estimator. The Kaplan-Meier curve, a visual depiction of this function, indicates the likelihood of an event (such as survival) occurring at a specific time period. The curve ought to resemble the genuine survival function for the population under study if the sample size is sufficient (Van Paemel, 2021).

The predicted survival probabilities as time passes for the population under study are shown in the Kaplan-Meier survival curve (Figure 7). The curve depicts the probability of survival at different time points. The curve's shape and trend reveal information on the patterns of survival.

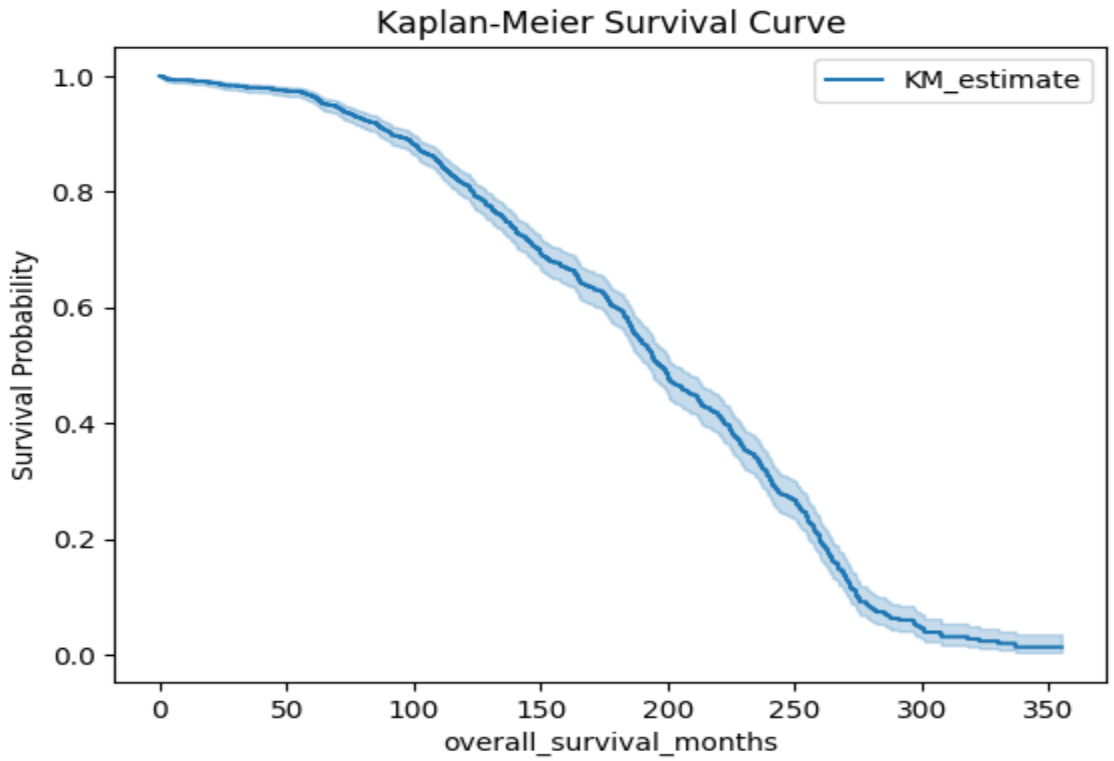
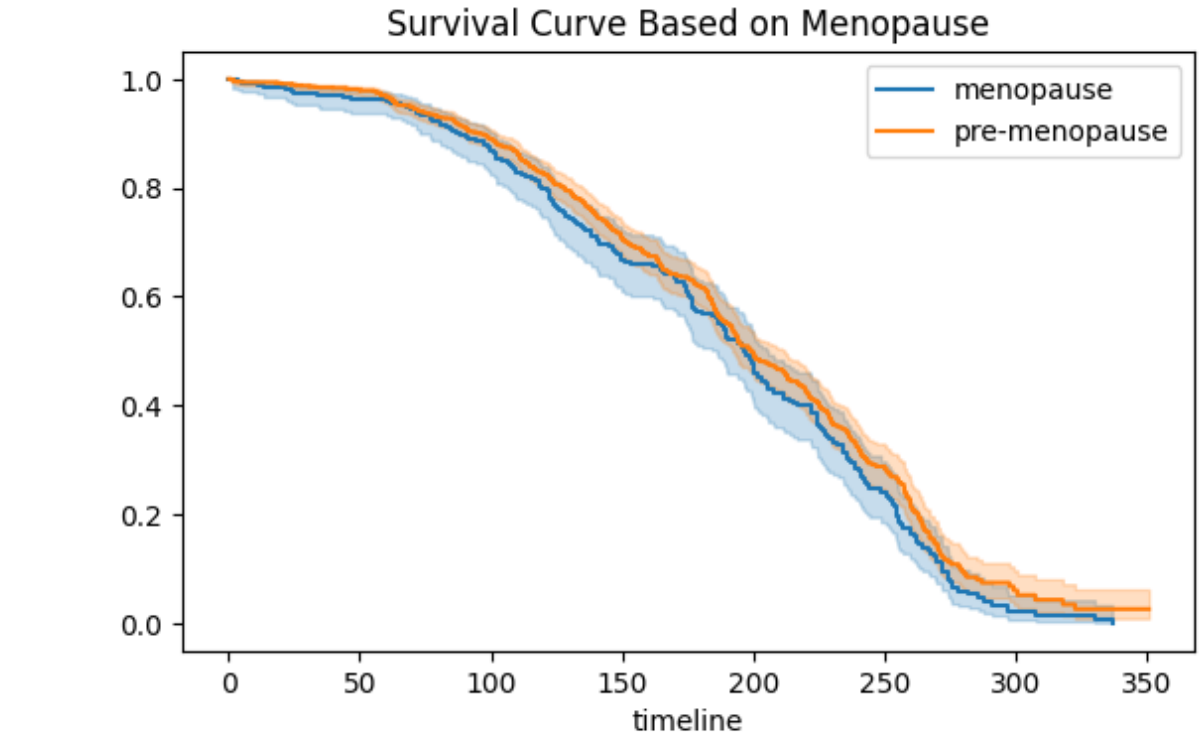


Figure 7: Kaplan Meirer Survival Curve.

At 95% confidence interval the median survival is 199.96 months, this suggests that around 50% of the patient has encountered the death during the midpoint of the recorded survival durations. The upper limit is 211.77 and lower limit is 193.966 months. As the range between the upper and lower limit is quite narrow, that represents enough confidence in the results and sufficient patient enrollment in the study.



pre-menopause								
At risk	1444	1156	821	493	249	91	8	1
Censored	0	261	501	685	806	881	907	910
Events	0	27	122	266	389	472	529	533

Figure 8 : Survival curve based on menopause

Log rank test was also done to see if there is any significant difference in survival probabilities between menopause and pre-menopausal patients and patients who are her-1 positive and negative.

For the menopausal and pre-menopausal group, a small difference in survival probability is observed. However, the p value from the log rank test is 0.06 which is higher than 0.05. Therefore, we can say there is no significant difference in survival probabilities between the pre-menopausal and menopausal patients.

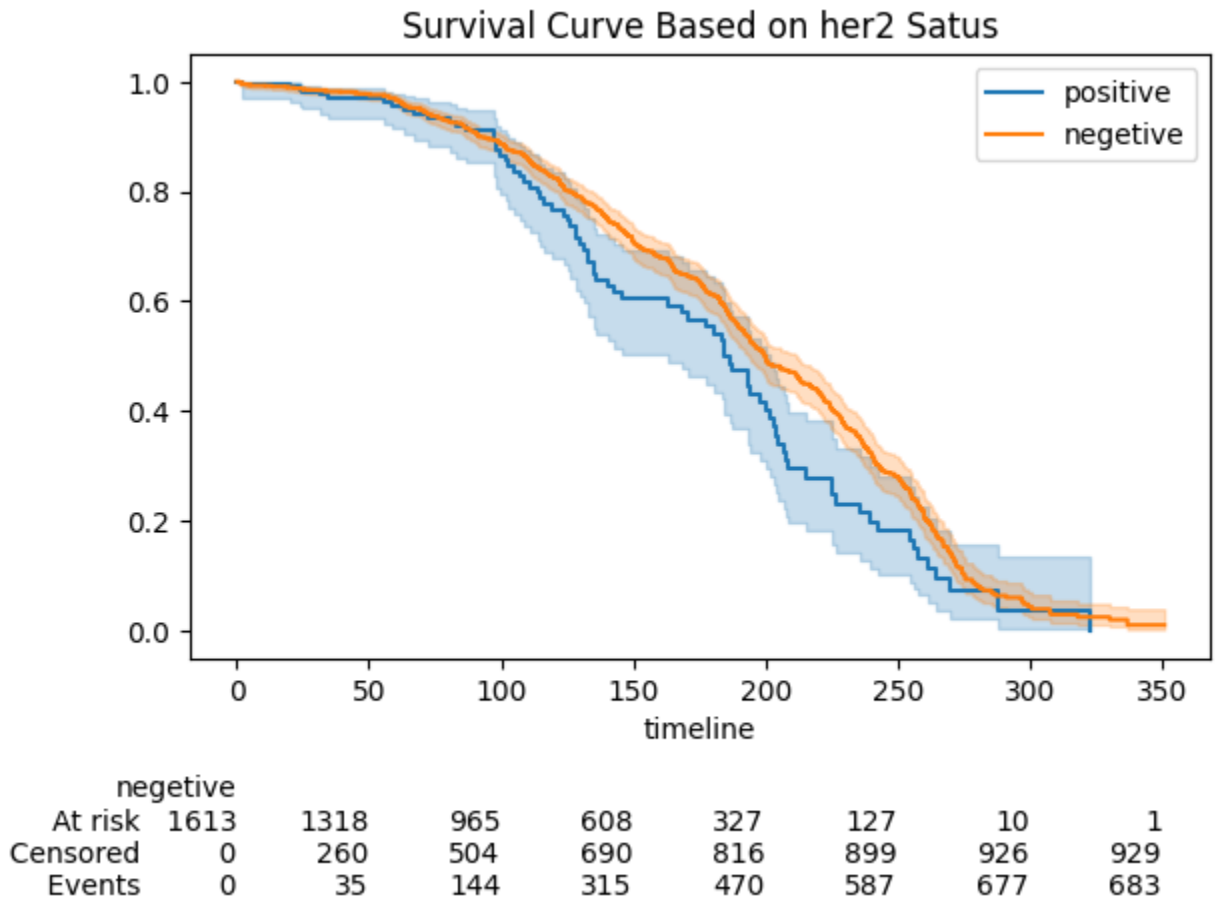


Figure 9 : Survival curve based on her-2 status

From the above figure it is also evident that there is some difference in survival probabilities based on her 2 statuses, patients with negative her 2 status have better survival probabilities. As the p value from the log rank test is 0.02 which is less than 0.05, we can reject the null

hypothesis and conclude there is a significance difference in survival probability based on her 2 status.

A cox proportional hazard model was also created using the dataset which is a semi parametric model. According to the model age at diagnosis, mutation count, tumor stage, chemotherapy and hormonal therapy have significant effect on overall survival and have p values less than 0.05.

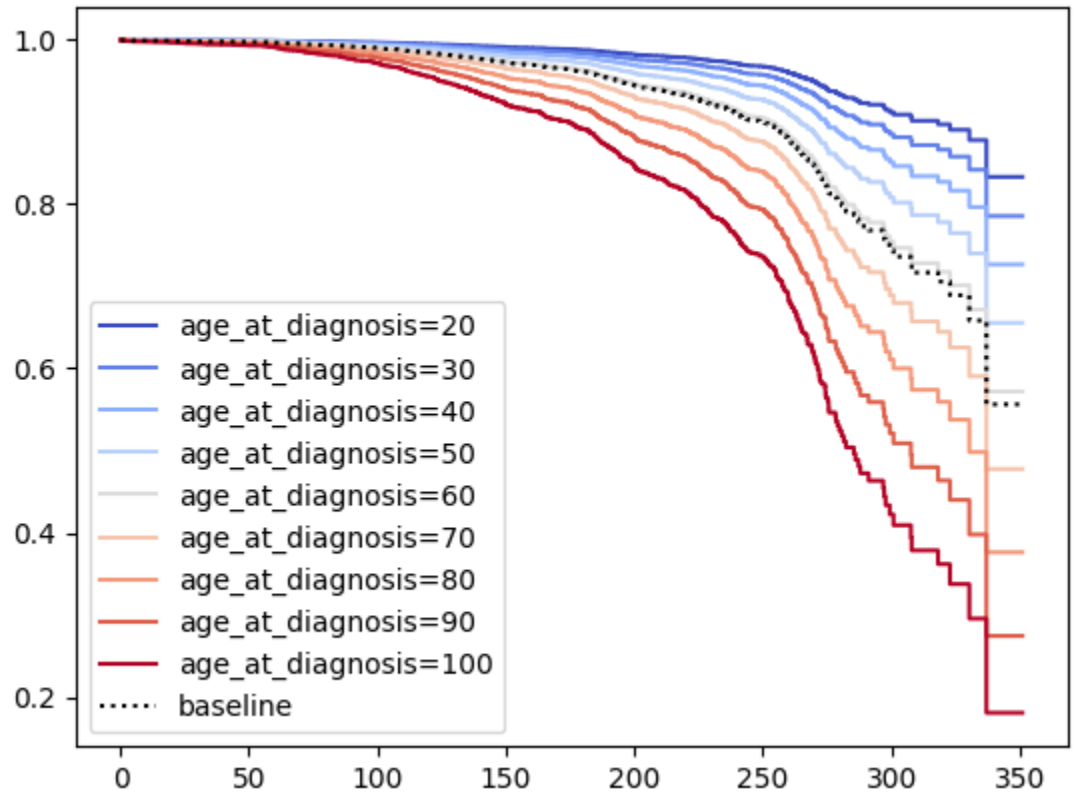


Figure 10: Effect of age on baseline hazard

From the above figure it is evident that as the age at diagnosis increases the probability of survival decreases.

Limitations and future scope:

Prior research on the dataset has mostly concentrated on applying deep learning methods and bioinformatic algorithms. As a result, it might not be possible to compare the project's findings directly to earlier studies. However, several actions may be made to improve the project's findings. First off, adding new evaluation measures on top of the ones already in place would result in a more thorough examination. A larger variety of methods and strategies might also produce more insightful outcomes. Concentrating on the most crucial traits and examining how they affect survival might produce insightful results.

Conclusion:

It is important to recognize that the inability to directly compare results with earlier research is one of the current work's limitations. The project may also have room for development in terms of adopting more sophisticated methodologies and investigating other subsets of attribute sets for analysis.

Increasing the scope of the study to incorporate additional assessment metrics and statistical methods will enable it to capture a more comprehensive perspective of the data and produce findings that are more reliable. Which will help future work to expand on the present research, get over challenges, and improve the body of understanding in the subject.

Reference:

- Breast cancer gene expression profiles (METABRIC) Kaggle. (n.d.). https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric

- Gandhi, R. (2022, November 14). Support Vector Machine — Introduction to Machine Learning Algorithms. Medium. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

- Guttà, C., Morhard, C., & Rehm, M. (2023, April 3). Applying a GAN-based classifier to improve transcriptome-based prognostication in breast cancer. PLOS Computational Biology, 19(4), e1011035. https://doi.org/10.1371/journal.pcbi.1011035

- Korstanje, J. (2022, January 5). *SMOTE | Towards Data Science*. Medium. https://towardsdatascience.com/smote-fdce2f605729

- Lunardon, Menardi, & Torelli. (2021, June 14). *Random Over-Sampling Examples*. https://cran.r-project.org/. https://cran.r-project.org/web/packages/ROSE/ROSE.pdf

- Mukherjee, A., Russell, R., Chin, S. F., Liu, B., Rueda, O. M., Ali, H. R., Turashvili, G., Mahler-Araujo, B., Ellis, I. O., Aparicio, S., Caldas, C., & Provenzano, E. (2018, March 7). Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort. Npj Breast Cancer, 4(1). https://doi.org/10.1038/s41523-018-0056-8

- Pereira, B., Chin, S. F., Rueda, O. M., Vollan, H. K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S. J., Tsui, D. W. Y., Liu, B., Dawson, S. J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R., . . . Caldas, C. (2016, June 6). Erratum: The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nature Communications, 7(1). https://doi.org/10.1038/ncomms11908

- Pham, guba, Porter, Gawanmeh, & Ngom. (2019, September). A Network-based Machine Learning Approach for Identifying Biomarkers of Breast Cancer Survivability. Association for Computing Machinery, 639–644. https://doi.org/10.1145/3307339.3343480

- R, S. E. (2023, July 5). *Understand Random Forest Algorithms With Examples (Updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

- Rezaeian, I., Mucaki, E. J., Baranova, K., Pham, H. Q., Angelov, D., Ngom, A., Rueda, L., & Rogan, P. K. (2017, January 27). Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Biochemically-inspired Machine Learning. F1000Research, 5, 2124. https://doi.org/10.12688/f1000research.9417.2

- Sugimoto, M., Hikichi, S., Takada, M., & Toi, M. (2023, March). Machine learning techniques for breast cancer diagnosis and treatment: a narrative review. Annals of Breast Surgery, 7, 7–7. https://doi.org/10.21037/abs-21-63

- Van Paemel, R. (2021, December 9). Kaplan Meier curves: an introduction - Towards Data Science. Medium. https://towardsdatascience.com/kaplan-meier-curves-c5768e349479

- Zolotareva, Isaeva, Hartung, Maier, Chaves, Kaufmann, Savchik, Chervontseva, Probul, Abisheva, Abisheva, Tsoy, Blumenthal, Ester, & Baumbach. (2022, August 26). DESMOND 2.0: Identification of differentially expressed biclusters for unsupervised patient stratification. ScienceOpen. https://doi.org/10.14293/S2199-1006.1.SOR-.PPPSLHRB.v1