Andrew Freitas, Emily Hart, Neha Pai, Chloe Yachanin
Bioengineering 175 Final Project Report

Investigating Features Related to Distinct Breast Cancer Subtypes

Introduction/Motivation:

Breast cancer is among the most common cancers in the US, with invasive ductal carcinoma (IDC) being the most common type and invasive lobular carcinoma (ILC) accounting for 10-15% of cases. As it is more difficult to detect, ILC can result in a worse prognosis, and for that reason, having a means to detect it based on a mutation could be extremely valuable to individualize treatment and provide more efficient healthcare with quicker recognition of cancer development. Identification of the BRCA mutation increases a woman's likelihood of developing breast cancer 5-fold, to around 65%. In this project, we will investigate the effects of four different multi-omics datatypes and the patterns that arise in their correlations to the two different types of tumors that form from the BRCA gene mutation using principal component analysis. This will ideally allow us to identify a relationship between BRCA and the different manifestations of its cancer to help with future prevention.

Problem Definition:

Our selected dataset, the BRCA Multi-Omics (TGCA) data from Kaggle, includes four distinct multi-omics data types: Copy Number Variations, Mutations, Gene Expression, and Protein Levels. This data is sourced from the article by Ciriello et al.: “Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer,” which investigated specific mutations and their correlation to ILC and IDC. The dataset includes measurements of these features in 705 patients, along with their cancer type (ILC or IDC), HER2 expression, PR status, ER status, and vitality.

The goal of our experiment is to use machine learning to learn more about IDC and ILC and their causes; we will explore the relationships between different variables and the patterns that may exist. There are several challenges that may arise due to the nature of this study. Interpreting biological systems is challenging due to their inherent complexity, and this is particularly true when studying multi-omics data. The dataset we are analyzing looks at genetic variations at different levels including the genome, epigenome, and proteome, making it computationally demanding but also providing the opportunity to uncover the root causes of complex problems. However, fitting a model to a large dataset with comparatively small sample sizes can lead to overfitting, which is further complicated by analyzing a dataset with little variability. In such cases, it becomes increasingly difficult to capture a significant amount of variation in the first few principal components, making it crucial to carefully interpret the results.

Our model will utilize PCA to identify the most important features that differentiate between ILC and IDC to explore the relationships between the different variables in the dataset and identify any underlying patterns that exist. PCA can be used as a pre-processing step to reduce the number of variables in the dataset and discover patterns, followed with linear regression to model the relationship between the remaining variables we want to evaluate..

A couple of questions we are interested in:

1. What type of multi-omics prediction model can we build and what does it explain?
2. Where are the strengths/weaknesses of our model?
3. Can we show why it is meaningful to integrate different data types?

Methods:

The model we have chosen to analyze our BRCA dataset is Principal Component Analysis (PCA). PCA is a statistical method commonly used to reduce the dimensionality of a dataset by transforming it into a lower-dimensional space. It can be useful for exploring the underlying structure of complex datasets, identifying patterns, and detecting outliers. We used PCA to investigate the differences between the four multi-omics data feature types in respect to the cancer subtypes and other information. By investigating BRCA datasets using this method, we can help identify potential patterns or correlations among a large number of genetic variables, in this case 1941 features. Analyzing all of these variables can be challenging and time-consuming, and using our method with the PCA algorithm can help simplify this process by identifying the most important features in the dataset or components that explain the majority of the variation in the dataset.

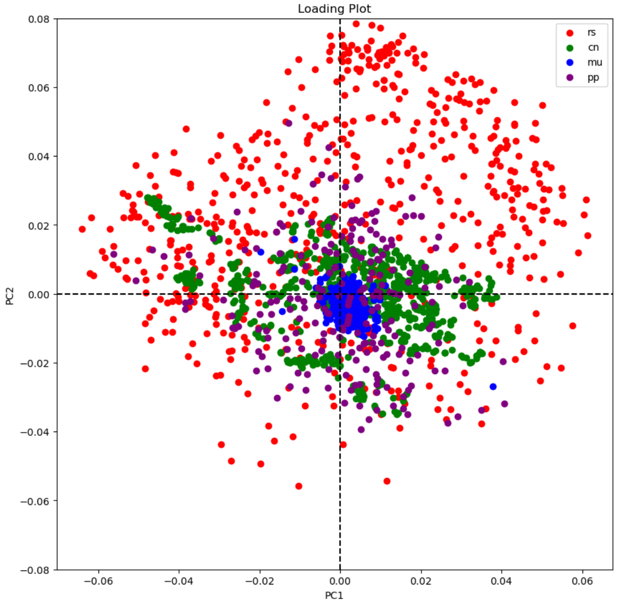
First, we prepared the dataset by cleaning and normalizing it, handling missing values, and encoding any categorical variables. We ran through all of the columns that contained string variables, such as 'PR.Status', 'ER.Status', 'HER2.Final.Status', and 'histological.type', and re-encoded them as integer variables. We filled any empty or missing values with 0 and set up a value map to exchange the variables in the select columns with a number. For instance, 'Positive' became 1, and 'Negative' became 2. We parsed ILC and IDC to be numbers 3 and 4, respectively, while other incomplete data was set to another number. We saved the new dataset under *modified_brca_data_w_subtypes.csv* and initiated the Principle Components Analysis algorithm.

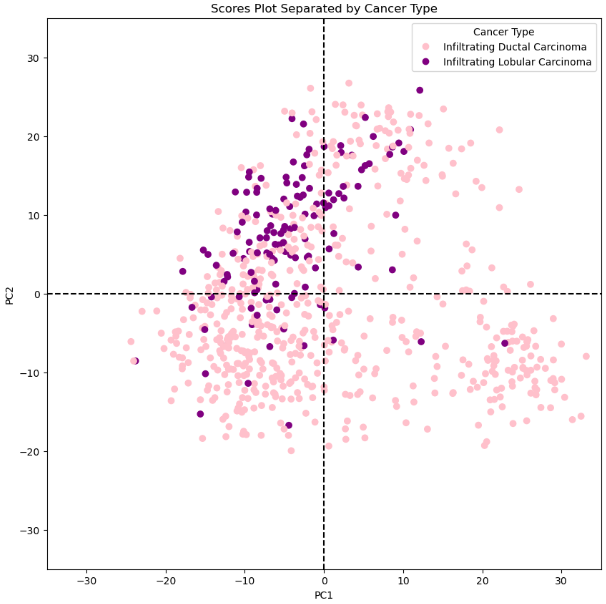
Source code for the algorithm and data processing can be found under *BRCA-PCA-175-Final.ipynb.* The software packages utilized in the development of our algorithm are as followed: StandardScaler, PCA, KFold, LinearRegression, r2_score.

Using the Standard Scalar import, the data was normalized and ready for PCA. We generated loadings and scores using the PCA model and plotted the data using matplotlib to show the relationship between loadings/scores and our principal components. The 'Loadings' plot illustrates all 1941 features of the dataset and color categorizes them by multi-omics data type: gene expression (rs), copy number variation (cn), mutations (mu), and protein levels (pp). This plot shows how strongly each of our variables influences each of the PCs. The five 'Scores' plots illustrate the 705 patients, categorized by specific identifying factors: cancer subtype, ER Status, PR Status, HER2 Status, and Vitality (Live/Dead). Subsequently, the R2X value was calculated to explain the percent variance in our PCs, and Mean R Square Scores were calculated using k-fold cross validation with linear regression to validate the variance in our model with the ‘goodness of fit’ measurement.

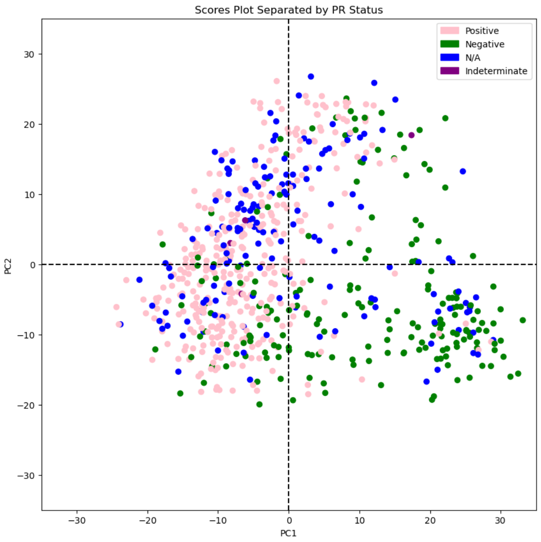
Results:

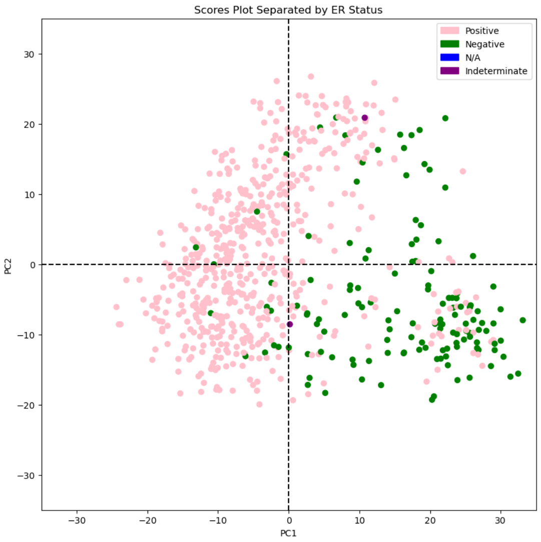
Our model provides key inferences about the relationships between the multi-omic data types and our identifying factors of cancer type, HER2, PR, ER, and vitality. Our loadings plot reveals that there is significant overlap between variables. The data is clustered about zero for PC1 and PC2, with some positive skew in PC2 for gene expression.



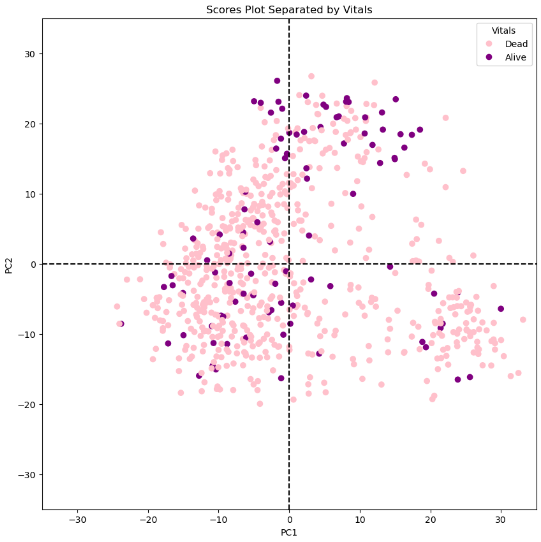


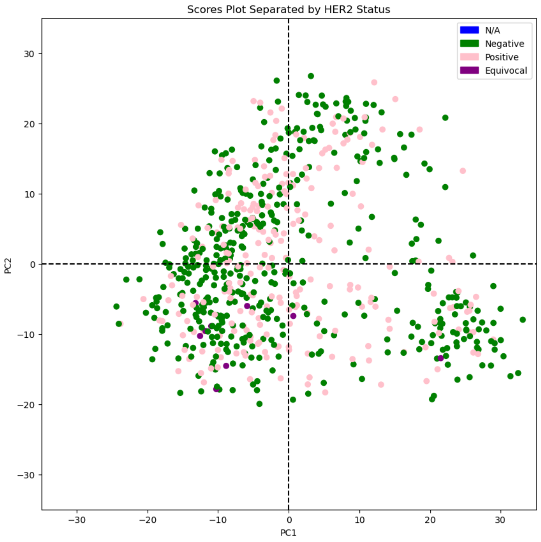
Our scores plots exhibit similar behavior in the overlap between variables. Additionally, scores plots with values that are positive in PC2 can be related to increased gene expression from our loadings. For cancer type, ILC appears to be negative in PC1 and positive in PC2, which can correlate to gene expression in the loadings plot. For PR status, PR-positive is primarily negative in PC1 and positive in PC2. Similar ER status, ER-positive is primarily negative in PC1 and positive in PC2. Therefore, we can discern that ILC is more closely associated with being PR positive and ER positive.





The scores plots for vitality and HER2 status exhibit too much overlap to make out any discernible patterns. ILC cannot be correlated to HER2 or Vitals. There is a general lack of variability in the data when plotted against the first two principal components given by a calculated R2X value of 0.08, indicating that 8% of the variation is explained by our PCs, and therefore our PCA model is not able to explain large proportions of variability. Our model succeeds in visualizing the spread of our data sets and its variables, but it is difficult to draw concrete conclusions about their relationships with the principal components.





From calculated the Mean R Squared Score for each of the scores plot, the results were as followed:

Cancer Subtype: 0.616

ER Status: 0.544

PR Status: 0.4795
HER2 Status: 0.534
Vitality: 0.208

It was noticed that all values were around 0.5-0.6, except vitality with a score of 0.2. This indicates that our model had some predictive power, around 50%-60% of the variance is explained, but there still is a substantial amount of variability unaccounted for in the data.

With our multi-omics dataset, we were able to build a predictive model using PCA to better understand key elements that differentiate ILC and IDC. We determined that our model is able to stratify the data by enabling us to look at the different data types on their own. This is especially helpful when looking at a dataset with a multitude of variables; without it distinct relationships like those exhibited in 'ER Status' or ‘PR Status’ would not be visible. While PCA is the only model we looked into for this experiment, future directions could include testing out different models and evaluating the strengths and weaknesses compared to the model we developed, or isolating specific genes from our correlated loadings to analyze with our scores.

Citations:

1. Demharter, Sam. “BRCA Multi-Omics (TCGA).” *Kaggle*, 15 Feb. 2022, https://www.kaggle.com/datasets/samdemharter/brca-multiomics-tcga
2. Ciriello et al. “Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer - Cell.” Cell.com, 8 Oct. 2015, https://www.cell.com/cell/fulltext/S0092-8674(15)01195-2.
3. Github Repository