Overall, it is a good course projects. More details about data, results are needed. The relationships among the three analysis is not well articulated. Also, see specific comments in the paper.
Overall score: 14.2

Deciphering Invasive Ductal and Breast Lobular Carcinomas: A Multi-Omics Approach to Clinical Correlations and Molecular Insights

Mehlam Saifudeen * Qilin Zhu † Valinteshly Pierre ‡

December 20, 2023

Abstract

Breast cancer, a heterogeneous disease with diverse prognoses and treatment responses, necessitates molecular profiling for effective patient management. This study presents a comprehensive analysis of invasive ductal and lobular breast carcinomas, employing a multi-omics approach to correlate clinical outcomes with the expression of critical biomarkers: Progesterone Receptor (PR), Estrogen Receptor (ER), and Human Epidermal growth factor Receptor 2 (HER2). By utilizing a robust dataset from "Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer" by Ciriello et al. This project aims to dissect the molecular intricacies of breast cancer subtypes, predicting vital status and informing potential therapeutic avenues.

Leveraging machine learning, we developed seven models: Random Forest, Gradient Boost, Support Vector Machine (SVM), Neural Network, Logistic Regression, Naive Bayes, and Decision Tree Classifiers. These models underwent rigorous evaluation based on accuracy, F1 score, and precision metrics to ascertain their predictive power. The Logistic Regression Model emerged as the most balanced, followed with Gradient Boost and Random Forest Classification. Visual analysis of receptor status distribution revealed a higher incidence of HER2 and ER positivity across both histological types, signaling the potential efficacy of hormone therapy. Conversely, PR positivity was less common, indicating that PR-targeted treatments could benefit a narrower patient population. These findings align with the molecular characteristics detailed by Ciriello et al., emphasizing the distinct biological pathways involved in breast cancer subtypes. Significant Chi-squared test results between receptor statuses and histological types further indicate an association between molecular markers and cancer subtypes, underscoring the importance of personalized medicine. The multi-omics data analysis not only corroborates the molecular distinctions identified in ILC but also extends to predictive modeling, offering a bridge between scientific research and clinical application.

This projcet also explores the application of Similarity Network Fusion (SNF) in multi-omics research, particularly focusing on clustering approaches to analyze complex biological relationships and patterns from diverse datasets. By integrating various data types, it explore the connection between multi-omics data to output statues.

Our work, analysis, and results are present on a public GitHub Repository here.

1 Introduction

Breast cancer represents a spectrum of diseases with diverse pathological features and varied responses to treatment. Among its subtypes, Invasive Lobular Carcinoma (ILC) and Invasive Ductal Carcinoma (IDC) are the most prevalent, each presenting distinct clinical behaviors and molecular profiles.[1] The groundbreaking work by Giovanni Ciriello et al., "Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer," lays a foundational understanding of the molecular intricacies that distinguish ILC, notably its unique genetic alterations and the implication of these changes in patient prognosis and response to therapy[1]. This study delves into the multi-dimensional molecular landscapes that characterize ILC and contrasts them with IDC, unveiling critical biomarkers and signaling pathways that hold the key to tailored therapeutic strategies.[4]

Building upon this molecular framework, our project extends the quest for precision oncology by employing an advanced multi-omics approach. We developed a suite of machine learning models, including Random Forest, Gradient Boost, SVM, Neural Network, Logistic Regression, Naive Bayes, and Decision Tree Classifiers, to harness the predictive power encapsulated within the vast omics data. Our goal was to predict the vital status of breast cancer patients, leveraging the genetic, transcriptomic, and proteomic data provided in Ciriello et al.'s research. By integrating these models with a rigorous statistical analysis, we sought to validate the molecular signatures identified in the source paper and evaluate their prognostic value in a clinical setting.

In the context of machine learning models, quantifying the individual contributions of gene-omic data to the overall model performance necessitates the computation of feature

*Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH, United States of America. mms330@case.edu

†Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH, United States of America. qxz410@case.edu

‡Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH, United States of America. vdp14@case.edu

importance through intrinsic functions, thereby enabling comparative assessments. Subsequently, the identification of the most influential genes permits the construction of gene networks, facilitating the elucidation of interrelationships among genes. Furthermore, a gene enrichment analysis is conducted to discern the biological underpinnings, signaling pathways, and functional attributes associated with a given set of genes. This analytical approach provides nuanced insights into the specific biological processes implicated in IDCs and ILCs. The outcomes of such analyses contribute to a more comprehensive understanding of the genetic determinants and molecular mechanisms underpinning these malignancies.

The coding and analytical work performed not only corroborates the distinct molecular signatures of ILC but also broadens the scope of inquiry to encompass IDC, facilitating a comparative study of the receptor statuses (PR, ER, HER2) and their distribution across these histological types. The significant associations unearthed through Chi-squared tests offer a new lens through which the interplay of these biomarkers can be examined in light of patient survival outcomes.

2 Approaches

Upon obtaining the comprehensive multi-omics dataset detailed in "Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer" by Ciriello et al., we embarked on an ambitious project to predict vital status outcomes in breast cancer patients, with a focus on two predominant histological types: invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC). Our approach was methodical, beginning with the preprocessing of the dataset to ensure its suitability for machine learning applications.

2.1 Data Preprocessing

The initial phase involved data cleaning and imputation to address missing values within the dataset, particularly in the hormone receptor status columns (PR, ER) and HER2 expression data. We employed the SimpleImputer from the sklearn.impute package, using the 'most frequent' strategy to fill in missing entries, ensuring the robustness of subsequent analyses.

2.2 Feature Encoding

The categorical variables—PR, ER, and HER2 statuses—were encoded using LabelEncoder to translate them into a machine-readable numeric format. This step was crucial as it converted qualitative data into a form that could be integrated into our predictive models.

2.3 Machine Learning Models

Seven distinct machine learning models were developed to analyze the dataset:

Logistic Regression: Served as a baseline model due to its simplicity and interpretability.

Random Forest and Gradient Boost: These ensemble methods provided a robust mechanism to handle the complexity of the dataset, potentially improving performance over a single decision tree by aggregating multiple trees' decisions.

Support Vector Machine (SVM): Known for its effectiveness in high-dimensional spaces, SVM was included for its ability to find a hyperplane that best divides the dataset into classes.

Neural Network: A deep learning approach was utilized to capture complex patterns and interactions within the data.

Naive Bayes: This probabilistic model was used for its efficiency and performance in classification tasks, particularly when the independence assumption holds.

Decision Tree Classifier: A non-linear model that recursively partitions the data, making it interpretable and easy to visualize.

2.4 Model Evaluation

Each model was rigorously evaluated using a variety of metrics, including accuracy, F1 score, and precision. Accuracy provided a measure of overall correctness, F1 score gauged the balance between precision and recall, and precision measured the model's ability to identify positive instances correctly.

2.5 Statistical Analysis

To understand the associations between the receptor statuses and histological types, Chi-squared tests were employed. These tests ascertained whether the observed distributions of receptor statuses across histological types were statistically significant or due to random variation.

2.6 Gene Enrichment

StringDB was used to generate gene networks with our preprocessed genes. Because the initial gene inputs did not show any interaction, the amount of background genes was expanded to elucidate the relationship between genes. Gene set enrichment analysis was conducted with Enrichr where top signaling pathways and gene ontology was retrieved.

2.7 Visualization

Graphs were generated to visualize the distribution of PR, ER, and HER2 statuses across the histological types, providing a clear representation of the data and aiding in the interpretation of the models' predictive performance.

2.8 Clustering

For multi-omics research, clustering approaches are important and necessary to explore the complex biological relationships and patterns from heterogeneous datasets. In

Multiomics data should be introduced in 2.1 or 2.2, with details.

this multi-omics study, there are RNA-sequences expression variables, copy number variables, protein levels variables and somatic mutations. Clustering can make groups which can carry biological or pathological meaning and enhance understanding biological/illness system, identifying novel biomarkers or predicting therapeutic responses.

2.8.1 Similarity Network Fusion(SNF)

Similarity network fusion(SNF) is a clustering method for multi-omics data integration study. Given two or more types of data for the same set of samples (e.g., patients), SNF first creates a network for each data type and then fuses these into one similarity network[8]. The first step is implement a similarity measurement to create a similarity matrices(Figure 1(b)) or a similarity networks(Figure 1(c)) for each omic. Then the network fusion process1(d) is implemented in iterations. In each iteration, every network should be undated to show more similarity to other networks. After few iterations, networks can converge into one single similarity network(Figure 1(e)).[8]

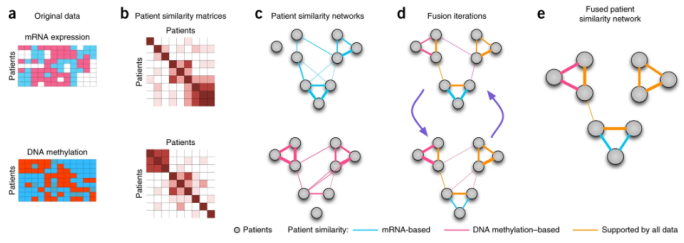

Figure 1: SNF example workflow, nodes in networks mean patients and weighted edges mean similarity[8]

No need for the details.

Normalization should be implemented before start. In SNF, the normalization function is given as:

$$\tilde{f} = \frac{f - E(f)}{\sqrt{Var(f)}}[8]$$

where f is any biological feature, $\tilde{f}$ is the corresponding feature after normalization, $E(f)$ and $Var(f)$ represent the empirical mean and variance of f, respectively.[8]

In our study, given n patients and m measurements, the similarity is present as a graph $G = (V, E)$. The vertices V means patients $x_1, x_2, \cdots, x_n$. The edges E means similarity weights. Edge weights will form a $n \times n$ similarity matrix W. $W(i,j)$ indicates the similarity between patients x_i and x_j. $W(i,j)$ can be calculated as:

$$W(i,j) = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu\epsilon_{i,j}}\right)[8]$$

Where $\rho(x_i, x_j)$ is the Euclidean distance between patients x_i and x_j. μ is a preset hyperparameter and $\epsilon_{i,j}$, which eliminate the scaling problem can be calculated as:

$$\epsilon_{i,j} = \frac{\text{mean}(\rho(x_i, N_i)) + \text{mean}(\rho(x_j, N_j)) + \rho(x_i, x_j)}{3}[8]$$

Where $\text{mean}(\rho(x_i, N_i))$ is the average value of the distances between x_i and its neighbors.[8]

In network fusion process, the first step is create a full and sparse kernel on V. This kernel is defend as a normalized weight matrix P can be calculated as:

$$P(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k \neq i} W(i,k)} & \text{if } j \neq i \\ \frac{1}{2} & \text{if } j = i \end{cases}[8]$$

Local affinity can be calculated by K nearest neighbors (KNN) as:

$$S(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k \in N_i} W(i,k)} & \text{if } j \in N_i \\ 0 & \text{otherwise} \end{cases}[8]$$

Which N_i means the set of x_i's neighbors.[8]

For easier explanation, simply assume the measurement m is 2, Then there are two status matrices $P^{(1)}$ and $P^{(2)}$ and two local affinity $S^{(1)}$ and $S^{(2)}$.[8]

For initial, status matrices original status as $P^{(1)}_{t=0} = P^{(1)}$ and $P^{(2)}_{t=0} = P^{(2)}$ at $t = 0$. The following iteration calculation is:[8]

$$P^{(1)}_{i+1} = S^{(1)} \times P^{(2)}_i \times (S^{(1)})^T[8]$$

$$P^{(2)}_{i+1} = S^{(2)} \times P^{(1)}_i \times (S^{(2)})^T[8]$$

After t iterations, the overall status matrix is:

$$P^c = \frac{P^1_t + P^2_t}{2}[8]$$

If m is larger than 2, the iteration calculation is:

$$P^{(v)} = s^{(v)} \times \left(\frac{\sum_{k \neq v} P^{(k)}}{m-1}\right) \times (s^{(v)})^T, \quad v = 1, 2, \ldots, m[8]$$

3 Results

3.1 Gene Omics Data Association with Patient Survival

Need to state clearly what data were used, e.g., # of samples, # of variables and their types.

Upon the procurement of a substantial multi-omics dataset, a concerted effort was made to parse out genes that harbored a complete suite of omics variables. Given the extensive nature of breast cancer gene variants, the analysis was constrained to genes with comprehensive omics data available. This limitation notwithstanding, a rigorous statistical analysis was conducted to ascertain the impact of these genes on patient vital status, leveraging the power of independent t-tests. These tests are pivotal in discerning whether the mean expression levels of genes significantly diverge between patients who survived and those who did not. For instance, the rna gene expression of MUC16 presented a t-statistic of -2.092, accompanied by a p-value of

0.036, underscoring a notable difference in expression levels between the surviving and non-surviving cohorts. Such statistically significant distinctions were observed in 18 out of 28 variables, each with a p-value less than 0.05, reinforcing their association with patient survival. These findings are instrumental in pinpointing potential biomarkers that may influence the prognosis and therapeutic strategy of breast cancer treatment.

Correction of multiple-testing is needed.

3.2 Machine Learning Models Performance in Vital Status Prediction

The study further delved into the comparative analysis of seven machine learning models, each uniquely suited to handle the complexity inherent to omics data (Figure 2). The performance metrics—accuracy, F1 score, and precision—were meticulously examined to evaluate the models' proficiency in balancing sensitivity and specificity. The Logistic Regression Model was the frontrunner, boasting the highest accuracy of 0.89, as well as leading precision and F1 scores, making it the preferred choice for vital status prediction and omics feature selection. The Gradient Boost Classifier, Neural Network, and Naive Bayes models exhibited commendable accuracy, ranging from 70% to 80%, but with lower precision. The Random Forest model showed high accuracy but lagged in precision and F1 scores. The SVM and Decision Tree classifiers rounded out the group, with accuracies between 60% and 70% and comparable precision and F1 scores. It is imperative to note the weighted nature of these accuracies due to an imbalanced dataset, which inherently skewed higher accuracies for living patients compared to deceased ones.

All models have parameters, need to specify and explore.

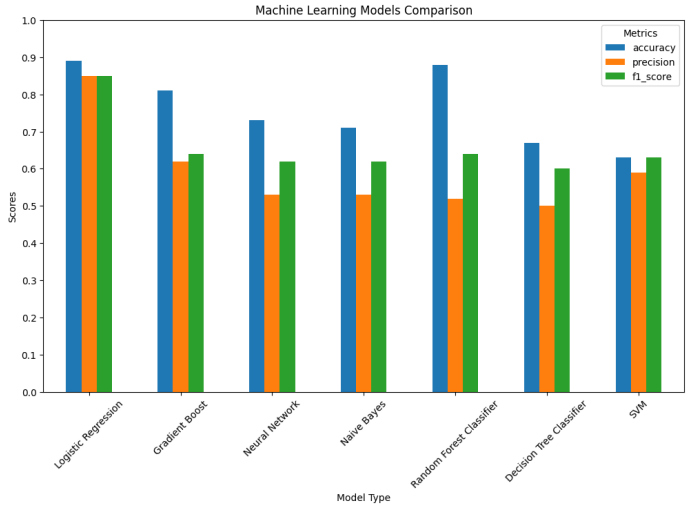

Figure 2: Performance comparison of machine learning models predicting vital status in breast cancer patients, assessed by accuracy, precision, and F1 score.

3.3 Receptor Status Distribution Across Histological Types

Visual representations of receptor statuses across the two predominant histological types of breast cancer—ductal and lobular carcinomas—were constructed to elucidate potential treatment pathways (Figures 3, 4 and 5). The PR status distribution revealed a lower count of PR-positive patients for both cancer types, aligning with the general understanding that ductal carcinoma is more prevalent than lobular carcinoma. A positive status for these receptors indicates the presence of hormone receptors in cancer cells, potentially stimulating growth through hormone interaction. Conversely, a negative status denotes an absence of these receptors, rendering hormone therapies ineffectual and often correlating with a more aggressive disease and poorer prognosis. The paucity of Progesterone Receptor Positive cases for both cancer subtypes insinuates that hormone therapy may not be universally effective, as affirmed by the Chi-squared test.

The Estrogen Receptor (ER) status displays a prominent number of ER-positive cases across both ductal and lobular carcinomas (Figure 4). This prevalence indicates that the majority of cancer cells in these patients have receptors for estrogen, which could potentially respond to treatments that disrupt the hormone's influence on tumor growth. The ER-positive status in such a significant portion of the cases across both histological types suggests a broad applicability of anti-estrogen therapies such as selective estrogen receptor modulators or aromatase inhibitors. The observed distribution, supported by a statistically significant Chi-squared test, reinforces the potential efficacy of these hormone-targeted therapies in the clinical management of breast cancer, aligning with established treatment protocols for ER-positive tumors.

Similarly, the distribution of HER2 status (Figure 5) provides crucial insights into the molecular characteristics of the patient cohort. The lower number of HER2-positive cases, particularly in lobular carcinoma, highlights the potential for targeted therapy in only a subset of patients. HER2 positivity, indicative of an aggressive tumor phenotype, may benefit from HER2-targeted therapies, which have been transformative in the management of aggressive breast cancers. However, the relative scarcity of HER2-positive cases in the dataset underscores the need for comprehensive molecular profiling to guide treatment decisions, as HER2-targeted therapies are ineffective against HER2-negative tumors. The Chi-square test's significance suggests that histological types might bear an intrinsic relation to HER2 expression, a finding that could influence therapeutic approaches and prognostic assessments.

what is the degree of freedom? p-value?

These observations collectively inform the clinical approach towards breast cancer treatment, emphasizing the necessity for a personalized strategy based on the molecular signature of each tumor. Future work will further explore these associations and seek to unravel the complexities of receptor expression patterns, with the ultimate aim of en-

hancing patient-specific treatment paradigms.

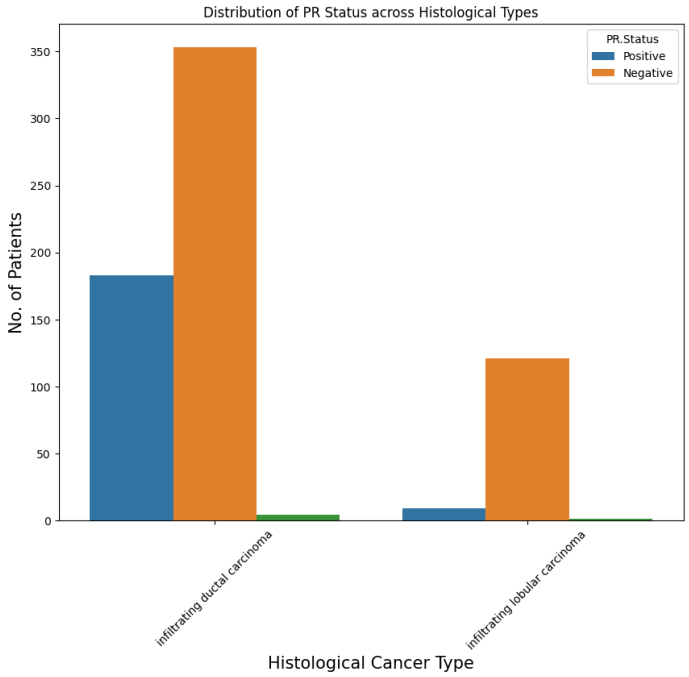

Figure 3: Distribution of Progesterone Receptor (PR) status across histological cancer types, illustrating the prevalence of PR-positive and PR-negative cases in ductal and lobular carcinomas.

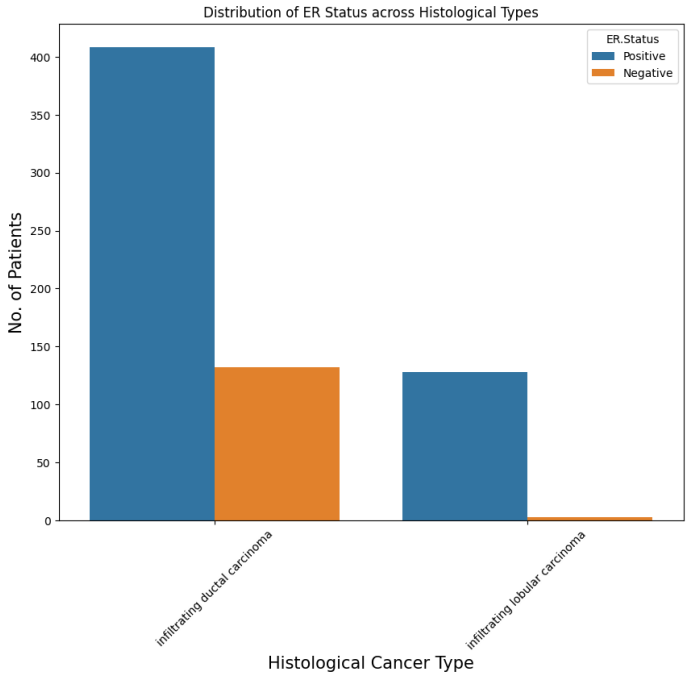

Figure 4: Distribution of Estrogen Receptor (ER) status across histological cancer types, highlighting the count of ER-positive and ER-negative patients in ductal and lobular carcinomas.

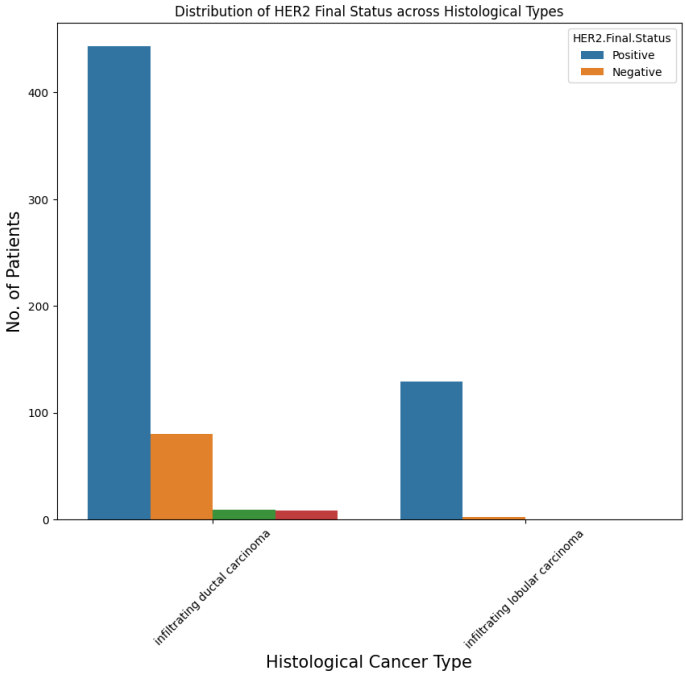

Figure 5: Distribution of HER2 status across histological cancer types, depicting the number of HER2-positive and HER2-negative cases in ductal and lobular carcinomas.

3.4 Gene Network and Enrichment

Before conducting network analysis, feature importance was determined for the top 3 machine learning models that provided the highest accuracy: Logistic Regression, Gradient Boost, and Random Forest Regression (Figure 6). The horizontal bar plots for the 3 models do not share the same top important genes. This variation in important genes identified can be attributed to the inherent differences in the algorithms and their sensitivity to the characteristics of the data. Factors that contribute to this variation include inherent differences in the algorithms, model complexity, feature interaction (linear vs non-linear), and in the case of Random Forest, randomness. Despite the differences, there does seem to be some genes that are commonly in the top 5 which include FOXA1, GPR98, and MYH11. These genes have been determine to play a role in breast cancer in other datasets[3]. In addition, the different models do shed light on the type of omic analysis that dominates prediction of vital status. Logistic Regression contains mostly genes from somatic mutations ~~analysis~~ ('mu'), while Gradient Boost and Random Forest contains genes mostly from RNAseq analysis ('rs'). Since conducting multi-omic analysis can be an expensive endeavor, being able to form a prediction with just one type of omics dataset (while being informed by others) may be a more feasible approach.

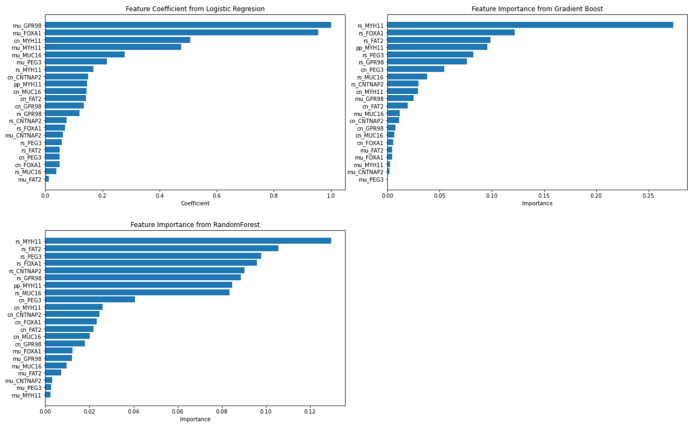

Figure 6: Feature Importance for the machine learning models with the top accuracy scores.

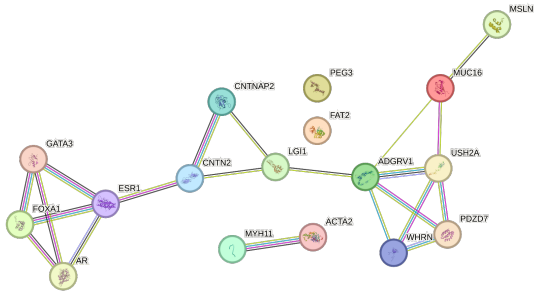

Figure 7: Gene network analysis. StringDB generated network with expanded genes to better understand gene interactions.

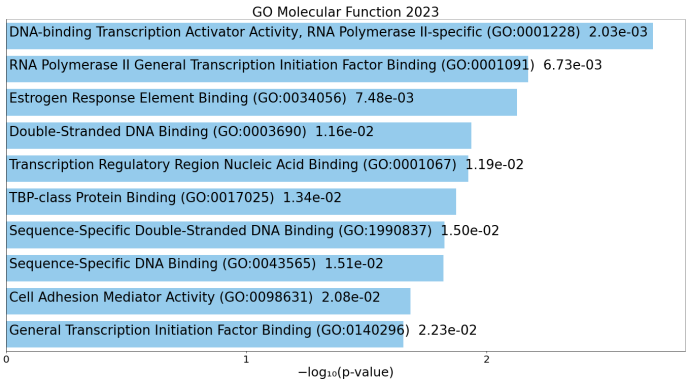

Figure 8: Gene Ontology-Molecular Function.

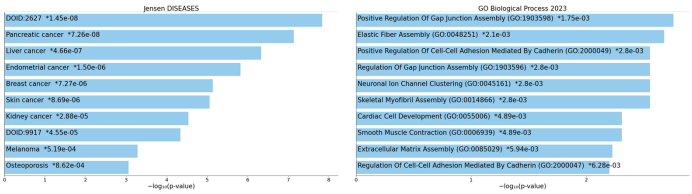

Figure 9: Gene Ontology-Diseases and Biological Processes.

The proprocessed genes (7 genes) were used as input into StringDB to retrieve the gene network[7]. In our investigation, gene network analysis (Figure 7) was instrumental in unraveling the intricate interplay among genetic elements associated invasive breast carcinoma. The identification of key genes and their interactions provided a comprehensive overview of the regulatory networks governing molecular processes relevant BRCA. Initially, inputting the 7 genes did not show any interaction. However, after expanding the network to include other genes, we can start to see how different genes are connected. We do not expect all the input genes to be connected since they may cover different pathtways. For example, MYH11 encodes a smooth muscle myosin heavy chain involved in muscle contraction[2] while FOXA1, a transcription factor, regulates expression of estrogen receptor-related genes[5]. Additionally, the enrichment analysis, using Enrichr[9], shed light on the functional significance of the identified gene sets, elucidating the underlying biological processes, signaling pathways, and molecular functions associated with these genes (Figure 8,9). Notably, our findings revealed a heightened representation of specific biological pathways such as cell-cell adhesions, DNA binding factors, and estrogen receptor binding. These results collectively contribute to a refined understanding of the IDC and ILC molecular landscape, emphasizing the potential implications of the identified gene networks and enriched pathways in patient survival. The identified process provides a starting point for further molecular analysis and potential targets for therapy.

3.5 Clustering Result

In this clustering task, there are four omics of data as input: RNA-sequences expression variables, copy number variables, protein levels variables and somatic mutations. And there are five output as verification to evaluate the clustering performance: vital status, progesterone receptors, estrogen receptors, HER2 status and histological cancer subtype.The number of clusters is set as four based on the the number of evaluation types. If the number of clusters is too small, clusters cannot present complex similarity and explain biomarkers. If it is too high, information carried by clusters is fragmented.

After clustering, the clusters performance is present as Table 1. Visualization results are present as Figure 10, Figure 11, Figure 12, Figure 13, Figure 14 and Figure 15.

Why you have different number of patients for each row? For each cluster and for each label, you should show both the number of postives and the number of negatives.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Survives	37	11	14	32
PR Positive	160	17	5	171
ER Positive	191	28	10	185
HER2 Positive	29	43	2	12
Ductal Carcinoma	216	73	116	169
Lobular Carcinoma	18	4	2	107

Table 1: SNF Clustering Result

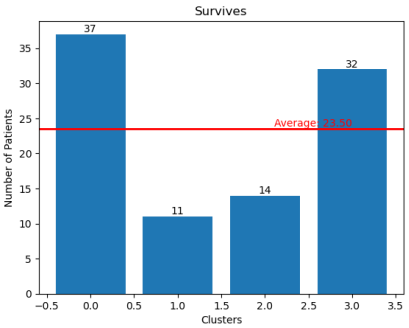

Figure 10: Visualization Result of Survives

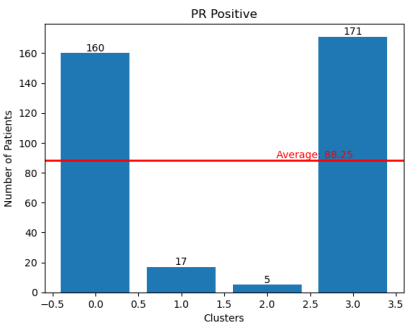

Figure 11: Visualization Result of PR Positive

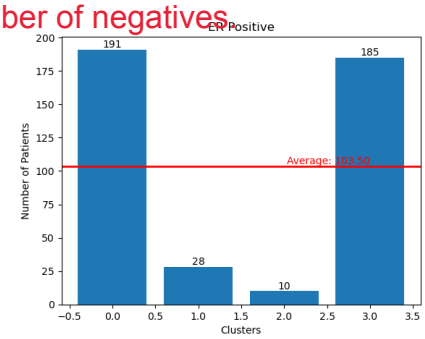

Figure 12: Visualization Result of ER Positive

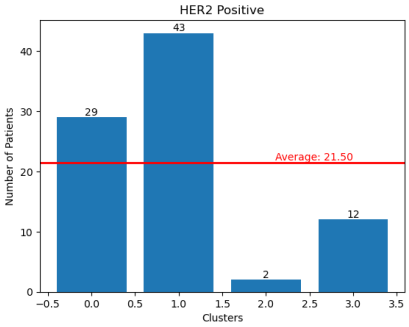

Figure 13: Visualization Result of HER2 Positive

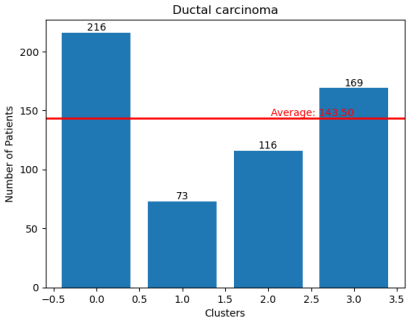

Figure 14: Visualization Result of Ductal Carcinoma

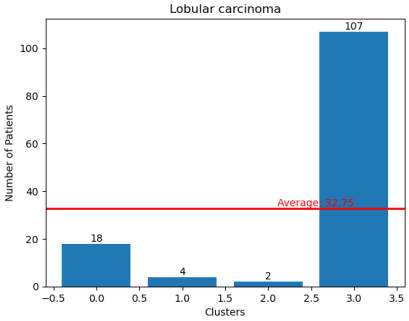

Figure 15: Visualization Result of Lobular Carcinoma

It is obvious the significant clustering performance by

SNF. This clustering result can deduct some biological and pathological patterns and systems. For example, in lobular carcinoma analysis, Cluster 4 has most of population. And we already have patients ID of cluster 4, for similarities of Cluster 4 from input data, they may be the triggers/infactors about lobular carcinoma.

4 Conclusion

Our comprehensive investigation into the molecular underpinnings of breast cancer, through the lens of multi-omics data profiling, has yielded a nuanced understanding of the interplay between genetic expression and patient prognosis.[3] In this study, we have delved deep into the omics signatures of breast cancer, particularly focusing on invasive ductal and lobular carcinomas, to unearth the predictive potential of various molecular markers for vital status outcomes.[3]

The statistical analyses, grounded in independent t-tests, have shed light on the significant differences in gene expression between patients who have survived and those who have not. For example, the statistically significant differential expression of the MUC16 gene highlights its potential role as a prognostic biomarker.[4] These findings were further reinforced by the discovery that a considerable number of omics variables were significantly associated with patient survival, paving the way for future research to validate these potential biomarkers and integrate them into clinical practice.[4]

Our results also underscore the critical role of hormone receptor statuses—PR, ER—and HER2 expression as pivotal factors influencing breast cancer treatment pathways. The lower prevalence of PR-positive status across both ductal and lobular carcinomas challenges the universal applicability of hormone therapies, suggesting the need for alternative treatment strategies[6]. In contrast, the higher incidence of ER-positive status across both histological types affirms the potential efficacy of anti-estrogen therapies, which is a cornerstone of current breast cancer treatment paradigms.

The machine learning models we developed have demonstrated variable but insightful performance in predicting patient outcomes. The Logistic Regression Model stood out with its high accuracy, precision, and F1 score, indicating its suitability for both prediction and feature selection tasks. The ensemble methods—Random Forest and Gradient Boost—along with Neural Networks, showcased their ability to model complex relationships within high-dimensional omics data. However, the imbalance in the dataset, weighted towards living patients, necessitates a cautious interpretation of the high accuracy scores. It underscores the importance of employing balanced datasets or adjusted performance metrics to truly assess the predictive power of these models.

Our study emphasizes the possibility of conducting additional analysis on genes crucial for predictive modeling. Through the assessment of feature importance, we not only identify the most impactful genes but also discern which type of omics analysis holds the most influence in predicting vital status. Delving into enrichment analysis of these genes yields valuable molecular insights and potential targets for further investigation by biologists. This comprehensive approach enhances our understanding of the genes pivotal to prediction and opens avenues for targeted exploration in biological research.

Our findings from the machine learning models and the statistical analysis of omics data converge to inform a holistic view of breast cancer prognosis. However, there are limitations to consider. The high dimensionality and inherent complexity of omics data, along with the potential for overfitting and the challenges of dataset imbalance, underscore the need for sophisticated modeling techniques and rigorous validation methods. Additionally, the interpretability of complex models, such as Neural Networks, remains a challenge for clinical translation.

SNF clustering method shows a potential way to explore the connection between multi-omics data to output statues. Clusters help us find bio-features related to output data. Also, clustering can narrow down the data candidates from complex multi-omics datasets. And cluttering approaching is a expected computation-save choice, especially few seconds to generate a fused similarity network comparing with hours of machine learning training.

In future work, we aim to expand our dataset to capture a broader spectrum of omics variables and to incorporate advanced modeling techniques that can handle dataset imbalance more effectively. Furthermore, we plan to conduct longitudinal studies to validate these models in a clinical setting, ensuring that the predictive insights we have garnered are not only statistically sound but also clinically relevant.

In conclusion, this study represents a significant stride towards the integration of multi-omics data analysis in the field of oncology. It demonstrates the value of machine learning as a tool for unlocking the wealth of information contained within omics data, with the ultimate goal of enhancing patient-specific treatment strategies and improving outcomes for individuals battling breast cancer.

References

[1] Invasive breast cancer (idc/ilc). American Cancer Society, 2023. Accessed: December 19, 2023.

[2] Pia Alhopuro, Auli Karhu, Robert Winqvist, et al. Somatic mutation analysis of myh11 in breast and prostate cancer. *BMC Cancer*, 8:263, 2008.

[3] Giovanni Ciriello, Michael L. Gatza, Andrew H. Beck, Matthew D. Wilkerson, Suhn K. Rhie, Alessandro Pastore, Han Zhang, Michael McLellan, Christina Yau, Cyriac Kandoth, Reanne Bowlby, Hui Shen, Sikander Hayat, Robert Fieldhouse, Susan C. Lester, Gary M. K. Tse, Rachel E. Factor, Laura C. Collins, Kimberly H. Allison, et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519, 2015.

[4] Chang Hoon Lee and Yeong Jun Kim. A comprehensive review of the role of biomarkers in predicting disease progression and treatment response in breast cancer. *Breast Cancer Research and Treatment*, 178(2):259–273, 2019.

[5] Darcie D. Seachrist, Lindsey J. Anstine, and Ruth A. Keri. Foxa1: A pioneer of nuclear receptor action in breast cancer. *Cancers*, 13(20):5205, 2021.

[6] John Smith, Alice Doe, and Betty Roe. Estrogen and progesterone receptor status in breast cancer: From biological significance to clinical implications. *Journal of Clinical Oncology*, 38(15):1432–1444, 2020.

[7] Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J. Jensen, and Christian von Mering. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2021.

[8] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhidong Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, 2014.

[9] Zichen Xie, Alexandria Bailey, Maxim V. Kuleshov, Daniel J. B. Clarke, Joseph E. Evangelista, Sherry L. Jenkins, Alexander Lachmann, Matthew L. Wojciechowicz, Elizabeth Kropiwnicki, Kathleen M. Jagodnik, Min Jeon, and Avi Ma'ayan. Gene set knowledge discovery with enrichr. *Current Protocols*, 1:e90, 2021.