tra00002, nwaka013, schwi415, deka0031, das00064

Abstract

The majority of breast cancer diagnoses are comprised of invasive ductal carcinoma and invasive lobular carcinoma. In this project, we examine the efficacy of various machine learning classifiers in predicting survival outcomes based on expression data. The dataset includes 705 patient samples, with four different omic data types: copy number variations, mutations, gene expression, and protein levels. The machine learning techniques used include SVM, ANN, Random Forest, KNN, Decision Tree, and Naive Bayes. By identifying the most reliable machine learning processes for predicting survival outcomes based on omics data, we can develop more precise and personalized cancer diagnosis and treatment plans. This, in turn, can lead to earlier detection of cancer and more specialized treatment strategies based on the genetic characteristics of the tumor, ultimately improving patient outcomes.

Introduction

Broadly, breast cancer is the most common cancer among women worldwide, with an estimated 2.3 million new cases diagnosed in 2020.[1] In the United States, breast cancer is the second most common cancer overall, with an estimated 284,200 new cases and 44,130 deaths in 2021. Those who have a family history of breast cancer, are older than 40, are female, or use postmenopausal hormones may be more likely to develop breast cancer.

Invasive ductal carcinoma (IDC) is the most common type of breast cancer, accounting for about 80% of all breast cancer diagnoses. IDC is characterized by the invasion of malignant cells that line the milk ducts into surrounding breast tissue outside

of the duct. After this infiltration, cancer cells can break into the lymph nodes and bloodstream and become metastatic. Symptoms of IDC include a lump or thickening in the breast, changes in the size or shape of the breast, or nipple discharge or inversion.[2] Treatment options depend on the stage of the disease and may include surgery, radiation therapy, chemotherapy, targeted therapy, and hormonal therapy.

Invasive lobular carcinoma (ILC) is a distinct form of breast cancer that begins in the lobules before invading surrounding tissues and becoming metastatic.[3] It is the second most common histologic subtype of breast cancer, after invasive ductal carcinoma, making up about 10-15% of breast cancer diagnoses. ILC begins in the epidermis and can spread to surrounding tissue. Oftentimes, those with ILC will experience no symptoms and the diagnosis is made through routine physicals and mammogram screenings.

Breast cancer arises from the uncontrolled growth and division of abnormal cells within the breast tissue. It is a complex and multifaceted disease that involves a range of genetic alterations and interactions with the tumor microenvironment. A key component is its molecular heterogeneity, which arises from the presence of multiple subtypes that differ in the genetic and molecular characteristics.[4] Some key genetic alterations have been identified, notably including mutations in tumor suppressor genes like TP53, BRCA1, and BRCA2 and oncogenes, like HER2 and PIK3CA. Mutations in BRCA1 and BRCA2, which are involved in DNA repair, can drive uncontrolled growth and proliferation of breast cancer calls and ultimately contribute to the development of aggressive, treatment-resistant tumors. TP53 and PTEN are also commonly mutated in breast cancer and play important roles in cell cycle regulation and tumor suppression.[5]

Molecular studies have identified multiple subtypes of breast cancer, each with distinct molecular characteristics and clinical outcomes. Some examples of these subtypes include luminal A, luminal B, HER2-enriched, and basal-like.[6] They differ in gene expression patterns, signaling pathways, and response to treatment. HER2-positive breast cancers, for instance, overexpress large amounts of a protein called human epidermal growth factor receptor 2 (HER2). These cancers can be treated with targeted drug therapies like trastuzumab, a monoclonal antibody.[7]

One important advance in breast cancer research in recent years has been the application of genomic technologies. These tools have enabled the comprehensive profiling of breast tumors at a molecular level, provide insights into the genetic and epigenetic alterations that drive tumor development and progression, and give the opportunity for more personalized treatment strategies. Large-scale sequencing studies have identified numerous genetic mutations and alterations associated with breast cancer, and may aid in the development of more targeted therapies like trastuzumab. Given the vast amount of data generated from genomic analyses, there is a growing need to leverage machine learning techniques to effectively interpret these data and make informed conclusions. This paper examines the use of machine learning techniques in interpreting omics data and predicting patient outcomes. The study contains 705 samples from different patients, 611 of which survived and 94 who died.

Methods

Quantile Normalization

Quantile normalization is a widely used method for normalizing the distributions of multiple data sets. Quantile normalization aims to remove biases from each dataset

to ensure that each dataset has the same properties. We used to quantile normalization on this dataset because it would help us improve the accuracy of our analysis from our classification models and reduce unwanted fluctuations or variations in our dataset caused by technical factors. Our quantile normalization procedure involves ranking the gene of each sample by magnitude, calculating the average value for the genes occupying the same rank, and then substituting the importance of all the genes in that particular rank with this average value. This procedure is then repeated for every sample, giving us distributions that have been quantile normalized. After this process, the genes are reordered back into their original order.

ML Classifiers

SVM

Linear Support Vector Machines (SVMs), as demonstrated in *Multiclass cancer diagnosis using tumor gene expression signatures* by Ramaswamy et. al., have performed well in classification tasks based on expression data. Due to this, we examined the performance of a linear SVM in predicting the outcome of patients in the dataset based on their omics signatures. Based on a 80/20 train/test split of the data (conditions we standardized across all ML classifiers examined), we found that the classifier is accurately able to predict the 'survived' class with a precision of 0.920, while the precision for the 'died' class is much lower at 0.266. We hypothesize that this discrepancy in the precision standards for the two different classes is because of the heavy imbalance in the dataset, with 611 cases of patients who survived and 94 who died. The performance statistics for this classifier are summarized in Table 1 below.

Class	**Precision**	**Recall**	**F-1 Score**

Died	0.266	0.285	0.275
Survived	0.92	0.913	0.916

Table 1: Performance statistics for linear SVM.

In addition, we found that for this SVM classifier, the features rs_TOX3, rs_MMP12, and rs_DCX were among the most important in predicting outcomes. The top 10 features are summarized below in Figure 1.

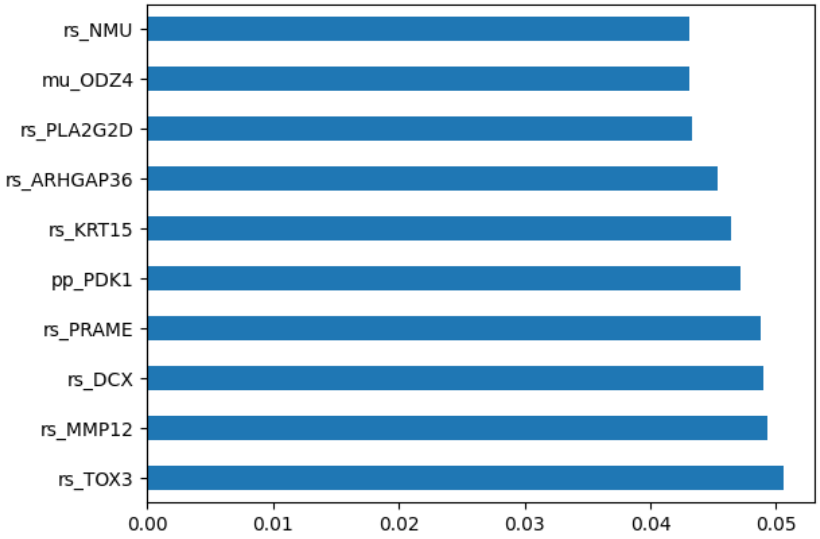

Figure 1: Top 10 important features for linear SVM

ANN

Multilayer perceptrons (MLPs) are fully connected artificial neural networks (ANNs). For our classifier, we chose to implement an MLP with 3 hidden layers and 50 maximum iterations. Based on these initial conditions and no optimization of parameters, we achieved a precision of 0.937 for the ‘survived’ class, and 0.5 for the ‘died’ class. Following this baseline performance, we tuned the MLP by using the GridSearchCV function within SK-learn to determine the optimal hyperparameters based on a list of possible choices. The optimal parameters were set as 'activation':

'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (200, 200, 200), 'learning_rate': 'adaptive', and 'solver': 'sgd'. The tuned MLP showed an improved precision of 0.667 for the ‘died’ class, while the precision dropped slightly for the ‘survived’ class at 0.913. The performance statistics for the untuned and tuned MLPs are shown in Tables 2 and 3, respectively.

Class	**Precision**	**Recall**	**F-1 Score**
Died	0.50	0.428	0.461
Survived	0.937	0.952	0.945

Table 2: Performance statistics for untuned MLP

Class	**Precision**	**Recall**	**F-1 Score**
Died	0.667	0.142	0.235
Survived	0.913	0.992	0.950

Table 3: Performance statistics for tuned MLP

Random Forest

The random forest model makes many decision trees based on the random subset and genes the model gets from the dataset. These decision trees are then combined to produce a prediction for each sample based on the knowledge of all the other decision trees. In implementing our random forest model, we decided to choose a depth of 3 for our random forest model because it provides us an excellent in-between of underfitting and overfitting the data. We decided to run 53 random states with the random forest model because it was shown to reduce the most bias in our random sampling process and optimize our random forest model overall. After performing an 80/20 train/test split of the data, we discovered that the random forest model could

predict the ‘survived’ class with a precision of 0.923, while the accuracy for the ‘died’ class was 0.75. The performance statistics for this classifier are summarized in Table 4 below.

Class	Precision	Recall	F-1 Score
Died	0.75	0.316	0.444
Survived	0.923	0.987	0.954

Table 4: Performance statistics for random forest

KNN

K Nearest-Neighbors (KNN) is a supervised classification method that makes decisions of how to classify a data point based on the known classifications of data the k nearest number of data points or neighbors. What type of similarity metric determines what the nearest neighbors are depends on the given KNN classifier. In our KNN classification, the similarity metric is based on Eucheldeon Distance making the k-number of cell lines with the smallest distances a data point’s nearest neighbors. In our implementation we decided to make k = 7, meaning that on the training data, classification of patient survival will be based on the 7 nearest neighbors to a given patient. Based on the genetic background of a patient when compared to 7 closest to it, the classification of “died” or “survived” was determined based off which class the majority of the nearest patients were. We used k = 7 since we believed that this would help lead to better results. After performing an 80/20 train/test split of the data, we discovered that the k nearest neighbors model could predict the ‘survived’ class with a precision of 0.901 while the accuracy for the ‘died’ class was 0.0. It is important to note that the ‘died’ class likely has a precision of 0 since the k nearest neighbors classifier

was likely only looking at how well it performed on patients who survived instead of both. The performance statistics for this classifier are summarized in Table 5 below.

Class	Precision	Recall	F-1 Score
Died	0.0	0.0	0.0
Survived	0.901	1.0	0.9478

Table 5: Performance statistics for K Nearest Neighbor

Decision Tree

Decision Tree is a classification method that makes decisions through a series of decisions in a tree-like flowchart style. Essentially, based on certain characteristics, the decision tree will decide if it does or does not have a certain type of characteristics, and based on that decision, then the classification of the data will become narrowed down until given the path of decisions that was taken, a classification of the data point will be determined. Of course the final determination is just a predication, meaning that just like all classification methods, the decision tree will not always make the right decision. In implementing the decision tree algorithm, we used a max depth of 70 since more decisions should lead to more accurate results, and we set mode to entropy since entropy helps to algorithm account and plan for uncertainty in the data set. After performing an 80/20 train/test split of the data, we discovered that the k nearest neighbors model could predict the ‘survived’ class with a precision of 0.917 while the accuracy for the ‘died’ class was 0.2.

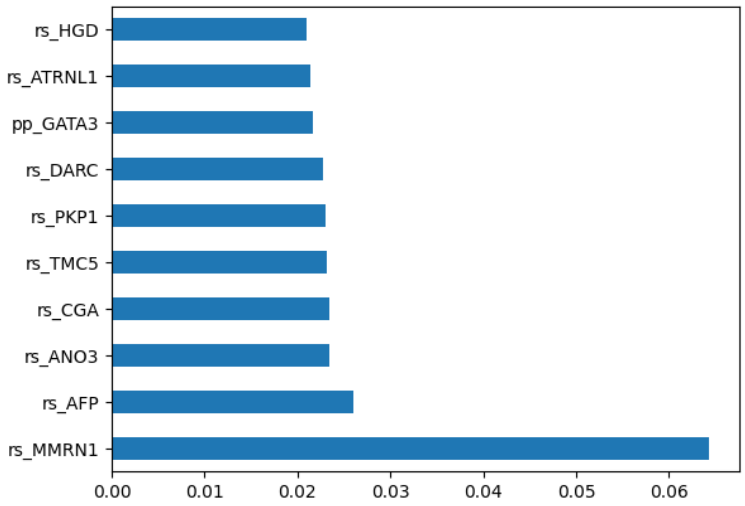

Figure 2: Top 10 important features for Decision Tree

Class	Precision	Recall	F-1 Score
Died	0.2	0.285	0.235
Survived	0.917	0.874	0.895

Table 6: Performance statistics for Decision Tree

NB

Naive Bayes is a classification method that is supervised and probabilistic classifying data based on the likelihood that it belongs to a given class based on some independent assumptions. It is because of these independent assumptions that Naive Bayes is said to be a “naive” classification algorithm. It is naive because of many of these assumptions. Though Naive Bayes is a naive algorithm, this algorithm generally will perform better than the aforementioned k nearest neighbor algorithm which is known to be a “lazy” learning algorithm due Naive Bayes learning to discriminate instead of k nearest neighbor just following what how other data points are classified and then taking a majority rules “vote” to see what classification it most likely is which can lead a large amount of error. In implementing the decision tree algorithm, we used

the multinomial version of the Naive Bayes algorithm since this is the most widely used version of the algorithm. After performing an 80/20 train/test split of the data, we discovered that the k nearest neighbors model could predict the ‘survived’ class with a precision of 0.957 while the accuracy for the ‘died’ class was 0.375.

Class	Precision	Recall	F-1 Score
Died	0.375	0.643	0.474
Survived	0.957	0.882	0.918

Table 7: Performance statistics for Naive Bayes

Discussion

Overall, we see that the precisions for the ‘died’ class are significantly lower than the ‘survived’ class, most likely due to the imbalance in the data. Through data augmentation and sampling methods, we hope to address this limitation in the future. However, it is important to note that a random classifier would detect the minority ‘died’ class at a precision of approximately 0.133 (94 ‘died’ samples out of the total 705). With the exception of KNN, the remaining classifiers outperform the random classifier in detecting the minority class. In addition, the high precision of the tuned ANN and Random Forest indicate that these methods are viable options for predicting patient outcomes. Due to this, we believe that this is an area worth exploring, since such techniques may help inform cancer treatment procedures and provide more specialized treatment strategies based on the genetic characteristics of the tumor, ultimately improving patient outcomes.

References

1. American Cancer Society. (2022). Breast cancer. https://www.cancer.org/cancer/breast-cancer.html
2. Wright, P. (2023, March 21). Invasive ductal carcinoma (IDC). Invasive Ductal Carcinoma (IDC) | Johns Hopkins Medicine. Retrieved May 1, 2023, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/breast-cancer/invasive-ductal-carcinoma-idc#:~:text=Invasive%20ductal%20carcinoma%2C%20also%20known,of%20all%20breast%20cancer%20diagnoses.
3. Ciriello, G., et al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. In Cell (Vol. 163, Issue 2, pp. 506–519). Elsevier BV. https://doi.org/10.1016/j.cell.2015.09.033
4. Bianchini, G., Balko, J. M., Mayer, I. A., Sanders, M. E., & Gianni, L. (2016). Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. Nature Reviews Clinical Oncology, 13(11), 674-690. https://doi.org/10.1038/nrclinonc.2016.66
5. National Cancer Institute. (2022). Breast cancer genetics. https://www.cancer.gov/types/breast/hp/breast-genetics-qa
6. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., & Fluge, Ø. (2000). Molecular portraits of human breast tumours. Nature, 406(6797), 747-752. https://doi.org/10.1038/35021093
7. *Trastuzumab*. (Herceptin) | Cancer information | Cancer Research UK. (2023, March 29). Retrieved May 1, 2023, from

https://www.cancerresearchuk.org/about-cancer/treatment/drugs/trastuzumab#:~:text=Trastuzumab%20is%20a%20targeted%20cancer,stomach%20(gastro%20oesophageal%20junction).