STAT432 Fall 2021
Final Project Report
(Option 1)

Zihan Xiong (zihanx3)

Zijun Tan (zijunt2)

I. Project Description and Summary

In this project, we analyzed BRCA Multi-Omics (TCGA) dataset in order to build various kinds of models that can predict distinct properties about subtypes of invasive breast cancer. These properties or targets of our predictions include estrogen receptors (ER), progesterone receptors (PR), HER2 protein overexpression, and identification of invasive breast cancer subtypes, i.e. Invasive/Infiltrating lobular carcinoma (ILC) v.s. Invasive/Infiltrating ductal carcinoma (IDC). Specifically, the former three variables are associated with the typical features of classic ILCs, which are "typically of low histologic grade…[and] express estrogen and progesterone receptors and rarely show HER2 protein overexpression or amplification" (Ciriello, 2015). The dataset was collected from 705 patients and includes 1936 features, including 249 somatic mutations, 860 copy number variations as calculated by gistic, 604 RNA sequencing, and 223 Phospho-protein levels, that are used in our classification.

As for the choices of our classification models, we applied mainly four types of supervised learning models. We chose to implement at least one linear model for each classification task: linear SVM for classifying the status of PR and Logistic Regression with L1 penalized term for differentiating between ILC and IDC. K-Nearest Neighbors (KNN) was modeled to classify the status of PR. Besides, we used Random Forest not only to classify ILC vs IDC but also to select the top 50 important features to be potentially the most representative among all such that they could be used to predict all four outcomes as listed previously.

In order to better evaluate our methods, we used both train-test splits and cross-validation in training and testing the models (we set the seed to be 1 and kept the number of folds 5 in all cross-validation models). Among the four models (linearSVM, linearSVM with cross-validation, KNN, and KNN with cross-validation) for classifying PR status, using classification error of test data as the evaluation criterion, the latter three models have comparable performances, while linearSVM had a slightly higher classification error of about 21%. As for classifying invasive breast cancer subtypes, we compared performances among logistic regression, logistic regression with cross validation, and random forest, and random forest with cross-validation according to their AUC on test dataset, the former two of which showed similar AUCs around 0.93, while the two random forest models had relatively low AUCs of around 0.85.

We measured the importance of each variable from random forest models for each outcome and added them together to extract the top 50 features, which were then fed into the

four types of machine learning models we tried previously: linearSVM, KNN with cross-validation, logistic regression, and random forest. Under the evaluation criterion of AUC and 3-folds cross-validation with seed of 1, logistic regression outperformed the rest, achieving a mean score of 0.9332 over the four response variables, i.e. PR status, ER status, HER2 status, and histological type (ICL vs IDL, the two subtypes of invasive breast cancer).

II. Literature Review

To gain a pre-analysis overview of the distinct properties about subtypes of invasive breast cancer, we conducted a literature review over two relevant research papers: *RNA-Seq-Based Breast Cancer Subtypes Classification Using Machine Learning Approaches,* and *Breast Cancer Subtype Identification Using Machine Learning Techniques*. In this section of the report, a summary of these two selected research papers would be presented followed by a discussion on the prediction methodology they employed and their conclusion.

In the first reference research study, researchers used gene expression information to reliably predict ten breast cancer groups. A hierarchical classification approach is used and builds the classifier concurrently, and the methodology includes Chi2 feature selection and a support vector machine (SVM) classifier. The researchers tried different methods, including solely LibSVM and a combination of Chi-Squared and SVM. The results support that this modified approach to gene selection yields a small subset of genes that can predict these ten subtypes with greater than 95% overall accuracy.

Although the second research paper aimed to examine the molecular regulatory mechanisms of the distinguished subtypes, which might be off topic with our project, we still found this paper valuable for reference as this research paper used machine learning based methodology to conduct the classification tasks between different groups of subtypes. Another reason for using this paper for reference is because of the dataset used in this research, RNA-Seq data (RNA sequencing) which is also one of the primary groups of features in our dataset. In this research, the researchers also encountered problems like imbalanced types of features, and they utilized sampling to lessen the interference of imbalanced data. Random forest and svmRadial (SVM with radial basis kernel) were adopted in this research to train the models and their performance were evaluated in terms of “Sensitivity,” “Specificity,” “Accuracy,” “F1,” and “AUC” metrics. The results came out really well that all the subtypes classifiers have a score

greater than 0.9 for these criteria. The high metric values verify the robustness and effectiveness of both random forest and svmRadial models.

III. Summary Statistics and Data Processing

In this section, univariate analysis of the BRCA data was conducted to obtain a meaningful insight of these predictors that would be used later for the predictive analysis, with visual elements to give clear demonstration.

Before doing further analysis, we did a processing for the dataset by identifying missing values for the predictors and filtering out observations without the prediction outcomes, and adding binary labels for the categorical outcomes.

To do an all-inclusive univariate analysis, we examined the frequency distribution among different classes for categorical variables and identified outliers for continuous variables, i.e. RNA sequencing (`rs`) and phospho-protein levels (`pp`), followed by trying some transformation of certain variables.

For categorical variables somatic mutations (`mu`) and copy number variation as calculated by gistic (`cn`), 5 from each group are randomly selected and demonstrated in frequency tables. It can be observed that a vast majority of both of these groups of variables are unevenly distributed among each class from the frequency plots as shown in Figure 3.1 and Figure 3.2.

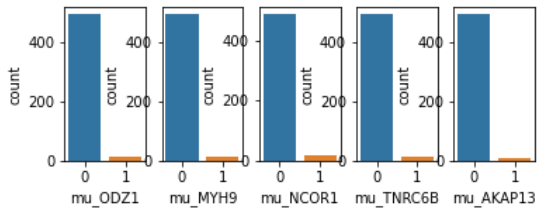

Figure 3.1.
Frequency distribution of Yes/No for Five Randomly Selected `mu` Variables

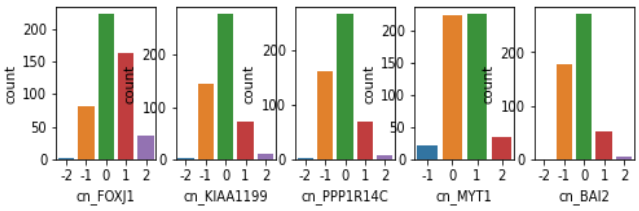

Figure 3.2.
Frequency distribution among 4 classes for Five Randomly Selected `cn` Variables

For `rs` and `pp` continuous variables, we checked the presence of outliers that might affect the prediction accuracy. By using the Interquartile Range technique, we found that half of

the `rs` variables and almost all `pp` variables (220 out of 223) have different numbers of outliers. We still randomly select 5 `rs` variables to see their boxplot and density plots.

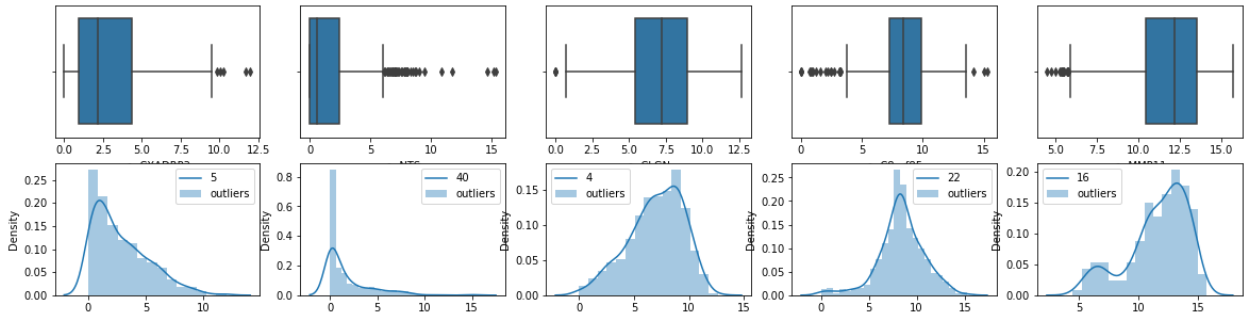

Figure 3.3. Box plot and Density plot for Five Randomly Selected `rs` variables.

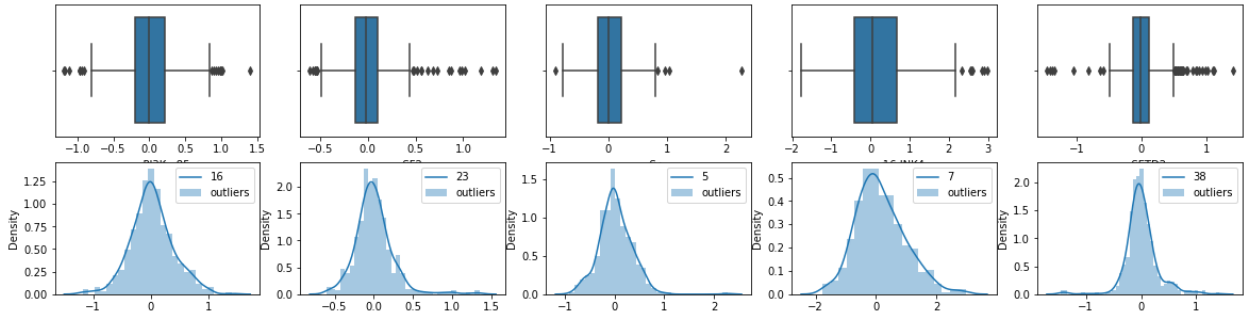

Figure 3.4. Box plot and Density plot for Six Randomly Selected `pp` variables.

We found that lots of variables fail to follow normal distributions (skewed or Multimodal) and tried two kinds of transformations on variables 'rs_CXADRP3' and `rs_NTS`.

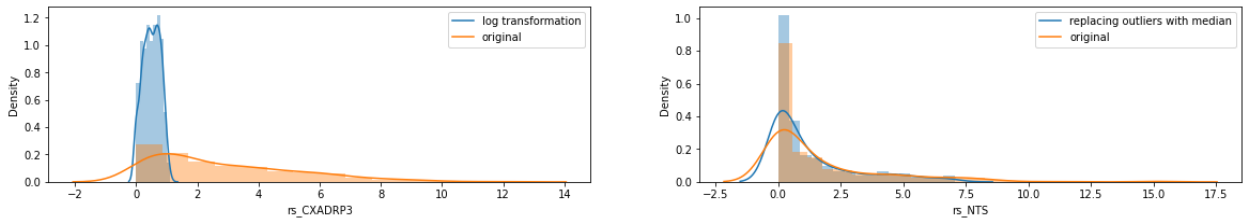

Figure 3.5. Density plot for 'rs_CXADRP3' and `rs_NTS`

For variable `rs_CXADRP3`, right-skewed with 5 outliers, after a log transformation, the distribution is normalized and is more likely to enhance the accuracy for prediction. For variable `rs_NTS`, right-skewed with 40 extreme outliers, after replacement with the median, has a better normalized distribution. However, given the fact that outliers can adversely affect the training

process, some of them might be valuable for considering some particular cases, thus models that are less vulnerable to outliers are preferred for this dataset.

IV. Modeling **PR.Status**

In modeling PR.Status, we implemented two kinds of classification models, selecting only 'Positive' and 'Negative' outcomes from data and encoding them as 1 and 0 respectively.

Support Vector Machine (SVM) is a classical classification model, and even training with data having very high dimension, SVM can still be effective and efficient in training time. We decided to implement the linear SVM with and without 5-folds cross validation to observe whether the input variables are linearly separable. Below in Figure 4.1, we plotted the two most important features ranked by the absolute value of the t-statistics from the linear SVM with 5-folds classification model and colored them by its PR.Status. From the graph, we can see that the two groups are kind of linearly separable, and there are, however, a few points from different groups mixing together. The graph also reflects the good performance of this SVM model, which achieves over 80% accuracy (i.e. around 0.2 classification error).

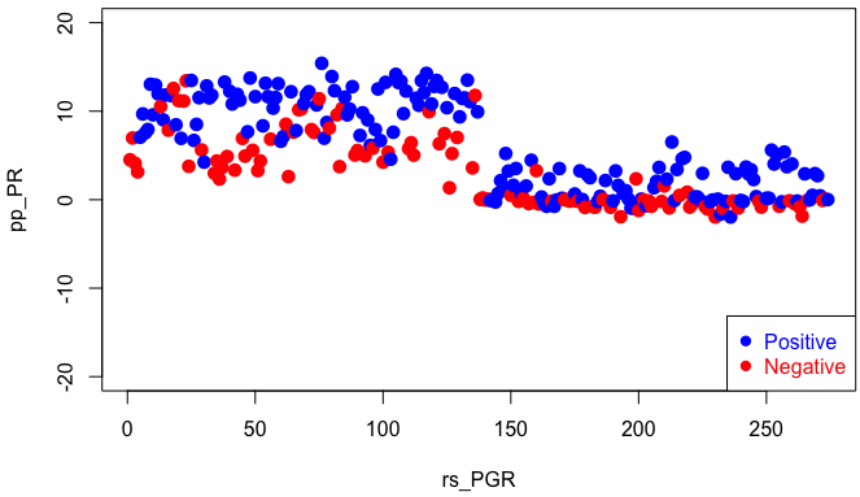

Figure 4.1. Data Visualization of Top 2 Most Important Variables Selected by LinearSVM w/ Cross-Validation

Different from SVM, K-Nearest-Neighbor is a nonparametric method that measures the distances between training data and outputs the label of the test data based on the majority vote of the labels from its k-nearest neighbors. In our implementation, we used Euclidean distance

and selected various values of k (1,5,9,15,25,39,49) in order to select the best k value. The reason why we selected odd values is that we would like the majority vote to have a decisive result rather than a random choice when there are ties. From the two plots in Figure 4.2, we can observe that as the value of k increases, the model tends to first have a better performance until k = 15, and since then the classification errors go up . Even though there are variations of classification errors on the training set, both models selected the same k=15 as the best k. Thus, the two models have the same performance on the test set as shown in Table 1.

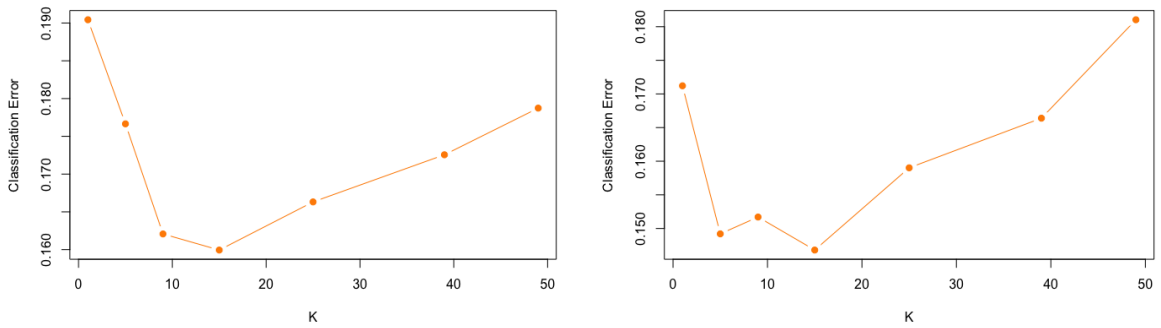

Figure 4.2. KNN Classification Error in Test Set (Left: w/o Cross-Validation, Right: Cross-Validation)

In order to compare across different models, we evaluated every model on a test set that was not involved in the training process of any model and recorded their classification errors as shown in Table 1. Models with cross-validation can have better performance on unseen data than those without. From the statistics, we can see that linearSVM with cross-validation performs slightly better than raw linearSVM. This doesn't happen to KNN models in this case because the two KNN models with and without cross-validation select the same k value of 15 as their best training models to predict the test set.

Model	**LinearSVM**	**LinearSVM w/ Cross-Validation**	**KNN (k = 15)**	**KNN w/ Cross-Validation (k = 15)**
Classification Error of Test Set	0.2117	0.1898	**0.1825**	**0.1825**

Table 1. Classification Error of Test Set in Predicting PR.Status

V. Modeling **histological.type**

In modeling histological.type, we implemented binary classification models, selecting only ‘infiltrating lobular carcinoma’ and ‘infiltrating ductal carcinoma’ outcomes from data and encoding them as 1 and 0 respectively. Different from the previous section where we used classification error as the evaluation criterion, here we used the area under the Receiver Operating Characteristic (ROC) curve, known as AUC. The closer AUC is to 1, the better the model performance is. In other words, AUC tells the model’s ability to separate the positive class and the negative class.

Logistic regression is an example of Generalized Linear Models. The relationship between the variables of data and its label is represented by a linear function of these variables. The original logistic regression doesn’t have a penalized term. In order to deal with the high-dimensional data in our case, we would like to add a penalized term. By comparing the results from adding the Lasso penalty (AUC around 0.93) and the Ridge penalty (AUC around 0.85), we decided to continue with the Lasso penalty or the L1 penalized term and implemented it with and without cross-validation. The two plots in Figure 5.1 show similar trends of ROC curves, and the two models have very close results of AUC on classifying the test set.

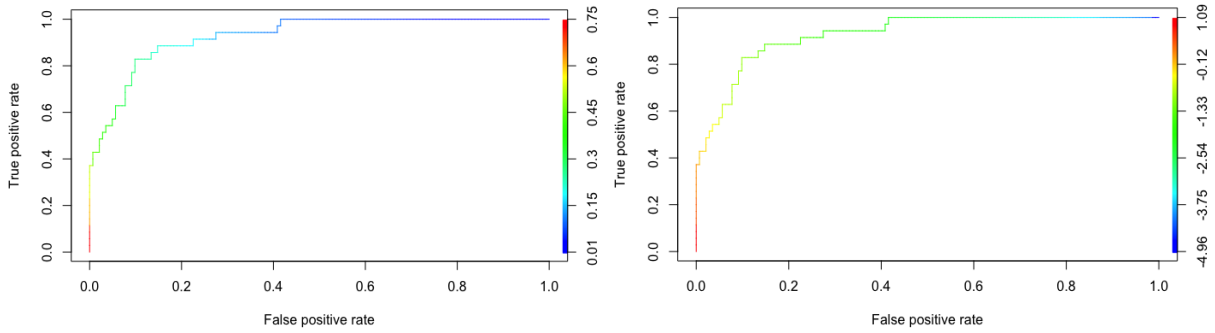

Figure 5.1. Logistic Regression w/ L1-Penalized ROC Curve (Left: w/o Cross-Validation, Right: Cross-Validation)

Random Forest is a collection of decision trees and eventually takes the majority vote of the labels these decision trees select. In our experiment, we modeled a number of 300 decision trees, several values of the number of variables than are randomly sampled as candidates at each split (i.e. 20, 50, 100, 200, 500, 1000), and three node sizes (i.e. 1, 5, and 7). Both Random

Forest with and without Cross-Validation selects 1000 as the best number of variables sampled each time, while the former prefers the node size of 1, and the latter finds the node size of 7 better. Although the model with cross-validation has a slightly higher training accuracy (around 93%) than the other (whose accuracy is about 92.62%), the two models have similar trends of ROC curves and comparable results of AUC.

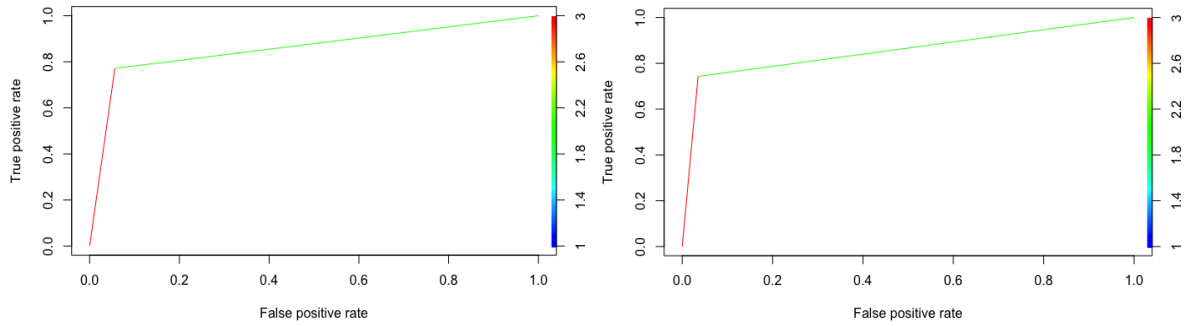

Figure 5.2. Random Forest ROC Curve (Left: w/o Cross-Validation, Right: Cross-Validation)

Among the four models, logistic regression models outperformed the random forest ones according to the statistics in Table 2. Different from modeling PR.Status, in this case cross-validation doesn't make very big changes to the two kinds of models based on their AUC scores.

Model	**Logistic Regression w/ L1-penalized**	**Logistic Regression w/ L1-penalized & Cross-Validation**	**Random Forest**	**Random Forest w/ Cross-Validation**
AUC	0.9286	**0.9290**	0.8575	0.8538

Table 2. AUC of Test Set in Predicting ICL vs IDL (histological.type)

VI. Variable Selection for All Outcomes

In order to select the most representative 50 variables that can predict all four outcomes: PR status, ER status, HER2 final status, and histological.type, we measured the importance of variables through random forest models. In Figure 6, we present a bar plot of the importance of the top 10 features ranked by the random forest model by the mean decrease in Gini coefficient,

which is “a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest” (Martinez-Taboada, 2020). From there, we chose the top 50 features and fed input data with only these 50 feature columns into four models implemented previously that have relatively good performances.

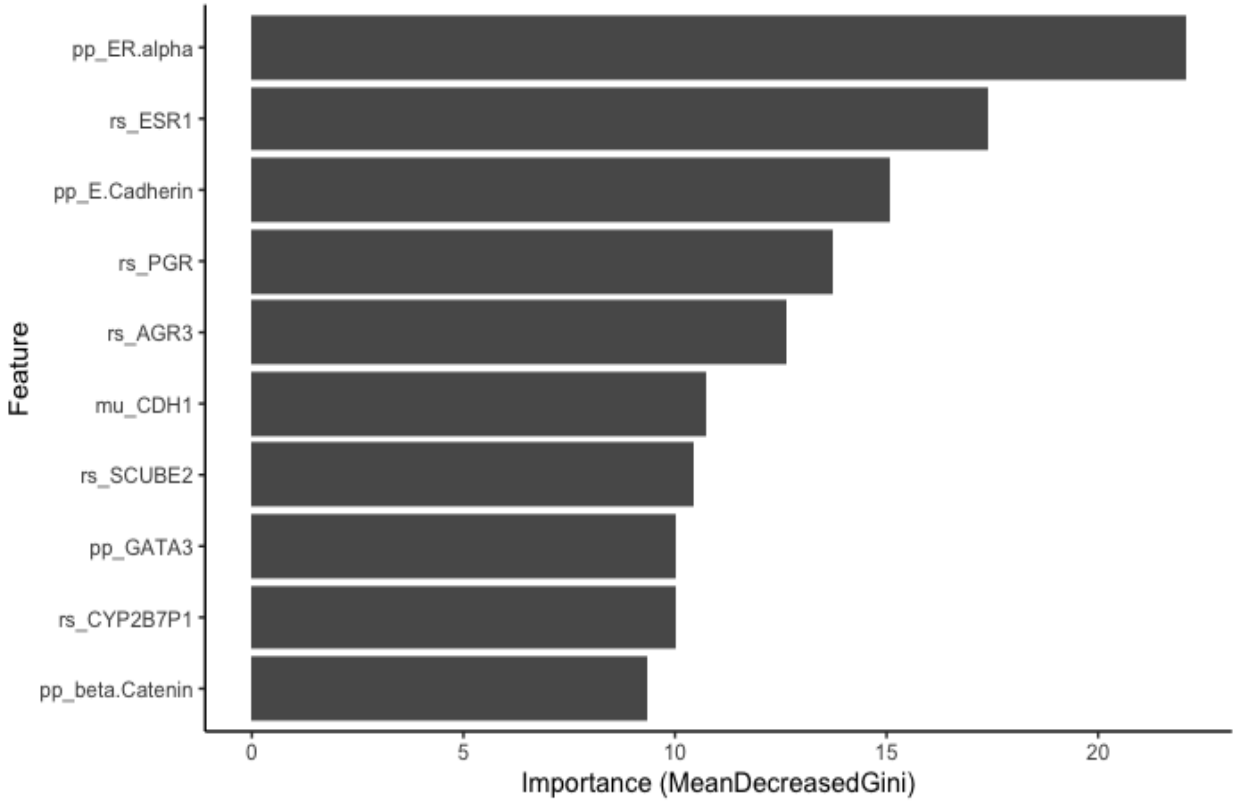

Figure 6. Top 10 Cumulative Feature Importance from Four Random Forest Models

As listed in Table 3, the top 50 features include 36 RNA sequencing or gene expression features, 9 phospho-protein levels features, 4 copy number variation as calculated by gistic features, and only 1 somatic mutation feature. In Figure 6, we can see that even though there is only one variable from the somatic mutation group, it is very important to predict the outcomes, as it is the sixth most important in the random forest importance ranking.

Group	Variables
RNA Sequencing or Gene Expression (`rs`)	`rs_ESR1, rs_AGR3, rs_PGR, rs_CYP2B7P1, rs_SCUBE2, rs_GFRA1, rs_RGS22, rs_TFF1,`

	rs_A2ML1, rs_ERBB4, rs_CLSTN2, rs_ANKRD43, rs_FLJ45983, rs_SERPINA11, rs_DEGS2, rs_GRPR, rs_TFF3, rs_SBSN, rs_TUBA3E, rs_AGR2, rs_PPP1R14C, rs_C1orf64, rs_FSIP1, rs_ABCC8, rs_SYT9, rs_NAT1, rs_SOX11, rs_TTC36, rs_ADAMTS15, rs_C2orf54, rs_NEK10, rs_AFF3, rs_TRH, rs_FOXA1, rs_TPSG1, rs_LOC389033
Phospho-Protein Levels (`pp`)	pp_ER.alpha, pp_E.Cadherin, pp_beta.Catenin, pp_GATA3, pp_HER2.pY1248, pp_PR, pp_HER2, pp_EGFR.pY1068, pp_INPP4B
Copy Number Variation as Calculated by Gistic (`cn`)	cn_PPP1R1B, cn_PNMT, cn_IKZF3, cn_STAC2
Somatic Mutation (`mu`)	mu_CDH1

Table 3. Top 50 Important Variables Used in Predicting All Four Outcomes

In order to evaluate the performance of different models, we set a 3-folds cross-validation to the input data and feed in qualified data, i.e. outcomes falling in targeted categories, to set up independent models on four outcomes. The final AUC score is averaged from all AUCs from the four independent models of each outcome and used for comparison. The mean AUC score is listed in Table 4. According to the statistics in Table 4, we can observe that LinearSVM and Random Forest have comparable performances or mean AUC scores over 3-folds cross-validation of the four outcomes of qualified input data. Logistic regression with L1-penalized term outperformed the other three, with a mean AUC of 0.9332.

Moreover, the table suggests that the four outcomes are closely related to each other or can be predicted from the same group of variables with relatively high AUCs. The four types of model have very similar performances on the four outcomes. This also corresponds to the fact stated previously in the project description that ICL is closely related to low histologic grade, express estrogen and progesterone receptors (ER and PR) and rarely show HER2 protein overexpression or amplification (Ciriello, 2015). Indeed, based on the AUC scores of the four models, especially that of Logistic Regression, our selection of 50 features through random forest importance rank is successful.

	LinearSVM	KNN	Logistic Regression w/ L1-penalized	Random Forest
PR.Status	0.8560	0.7958	**0.9235**	0.8508
ER.Status	0.8533	0.8757	**0.9608**	0.8986
HER2.Final.Status	0.8172	0.6428	**0.9173**	0.8528
histological.type	0.8331	0.6936	**0.9313**	0.8305
Mean AUC	0.8399	0.7519	**0.9332**	0.8582

Table 4. AUCs of 3-Folds Cross-Validation Using Selected 50 Variables

VII. References

Ciriello, G., et al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, *163*(2), 506–519. https://doi.org/10.1016/j.cell.2015.09.033

Firoozbakht, F., Rezaeian, I., Porter, L., & Rueda, L. (2014). Breast cancer subtype identification using machine learning techniques. *2014 IEEE 4th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), Computational Advances in Bio and Medical Sciences (ICCABS), 2014 IEEE 4th International Conference On*, 1–2. https://doi-org.proxy2.library.illinois.edu/10.1109/ICCABS.2014.6863912

Martinez-Taboada, F., & Redondo, J. I.. (2020). *Variable importance plot (mean decrease accuracy and mean decrease Gini).* (Version 1). PLOS ONE. https://doi.org/10.1371/journal.pone.0230799.g002

Yu, Z., Wang, Z., Yu, X., & Zhang, Z. (2020). RNA-Seq-Based Breast Cancer Subtypes Classification Using Machine Learning Approaches. *Computational Intelligence & Neuroscience*, 1–13. https://doi-org.proxy2.library.illinois.edu/10.1155/2020/4737969