CSE 283 Final Report: Breast Cancer Recurrence Classification using Time-course and F-value Analysis

Jeff Makings, Harish Rithish

1 Introduction

Breast cancer is one of the most common types of cancer in the US and world. In 2021, the World Health Organization classified breast cancer as the most common cancer globally, accounting for 12% of all new cancer cases annually. Furthermore, recent studies have shown there is a 6.7% chance of local recurrence on average in young breast cancer patients.

With these statistics in mind, it is of great benefit to study the transcriptomic changes and develop classifiers associated with breast cancer and its recurrence. In this paper, we will describe the methods we used to develop various classifiers to identify both breast cancer vs normal samples and recurrent vs non-recurrent breast cancer samples.

2 Dataset

The data used to develop these classifiers was acquired courtesy of Dr. Zhong. It includes the raw counts and transcripts per million from the sequencing of extracellular RNA from approximately 5µl of patient serum samples. The first data set is the data used in Dr. Zhong's PNAS paper (Zhou, 2019) which we will refer to as the training set. This data contains 32 samples from patients without breast cancer, 68 from patients with non-recurrent breast cancer, and 28 from patients with recurrent breast cancer. The second is the "validation" dataset, which we used to test our classifiers. It contained 78 non-cancer samples, 75 non-recurrent cancer samples, and 8 recurrent cancer samples.

The raw counts and tpm for 60,674 transcript reads for each sample are included in this dataset. However, it's important to note this data is *sparse*, and most samples only contain non-zero values for 30,000-41,000 transcripts on average.

3 Classifying Normal vs Cancer Patients

3.1 Feature Selection

With over 60,000 transcripts as well as other information on cancer type, chemotherapy, cancer stage, etc., the biggest challenge with this dataset was narrowing down which features would be most informative in classifying cancer from non-cancer patients.

To reduce the number of features, we used the ANOVA F-value method included in Scikit-learn feature selection's "SelectKBest" method.

$$F = \frac{MSB}{MSE}$$

$$MSB = \frac{Sum\ of\ Squares(Between\ Groups)}{Num\ of\ Groups}$$

$$MSE = \frac{Sum\ of\ Squares(Within\ Groups)}{Groups - Samples}$$

The F value is obtained from the average variability between the groups (Normal vs Cancer in this case) divided by the average variability within groups, for each transcript's TPM counts. So for each transcript, we obtain an F value, and the n highest F values were selected as features in our model.

To test this method, we calculated the Area Under the Reciever Operating Characteristic (ROC AUC) and 3-fold cross-validation accuracy for a Random Forest Classifier using varying number of features on the training dataset. We determined that using the genes with the top 5000 F-values as features was ideal (Figure 1). Both the 3-fold cross validation and the AUC ROC score were 100% in this scenario.

3.2 Results

We next sought to evaluate this model on the testing set. Using a Random Forest Classifier with

Number of Genes Used	3-Fold Accuracy Scores	ROC AUC Score
10	[1.0, 0.97674, 1.0]	1.0
50	[1.0, 1.0, 1.0]	1.0
100	[1.0, 1.0, 1.0]	1.0
200	[1.0, 1.0, 1.0]	1.0
500	[1.0, 1.0, 1.0]	1.0
700	[1.0, 1.0, 1.0]	1.0
1000	[1.0, 1.0, 1.0]	1.0
5000	[1.0, 1.0, 1.0]	1.0
10000	[0.97674, 0.97674, 0.97619]	1.0
20000	[1.0, 0.93023, 0.92857]	0.999349
30000	[0.95349, 0.93023, 0.88095]	0.98763
40000	[0.86047, 0.93023, 0.78571]	0.989095
60674	[0.88372, 0.88372, 0.88095]	0.973633

Figure 1: Table of 3-Fold Classifier Accuracies for increasing numbers of genes as features on the training set. Selecting the top 5000 genes had the most promising accuracy results

100 decision trees and default Sklearn parameters, we first tested the the classifier trained with the top 5000 genes from the training set, as well as varying numbers of genes from 10 to 60,674 (all of the genes). Figure 2 shows the average AUC ROC on the test set of the classifiers using the top 5000 genes, while the the figure on the right shows the test set AUC ROC for varying numbers of features. The classifiers using the top 5000 genes had an average AUC of 0.6668 and a balanced accuracy rate of 0.5648. However, the average best-performing classifier used 30000 genes and achieved an AUC of 0.6838, with a similar balanced accuracy rate of 0.5592.

3.3 Evaluation

As represented by the number of features vs AUC graph, it appears that increasing the number of genes used in a classifier increases AUC until approximately 30,000 genes, where the average AUC levels off. This suggests that there is some variation in the expression level of thousands or tens of thousands of exRNA transcripts when comparing between normal and cancer patients, and generally utilizing more data will be beneficial in differentiating between cancer and normal serum samples.

4 Classifying Recurrent and Non-recurrent Cancer Patients

To create robust recurrent vs non-recurrent breast cancer classifiers, we tested two different methods:

1. Refitting the Normal vs Cancer Classifier to this task, selecting gene features with the highest ANOVA F-value

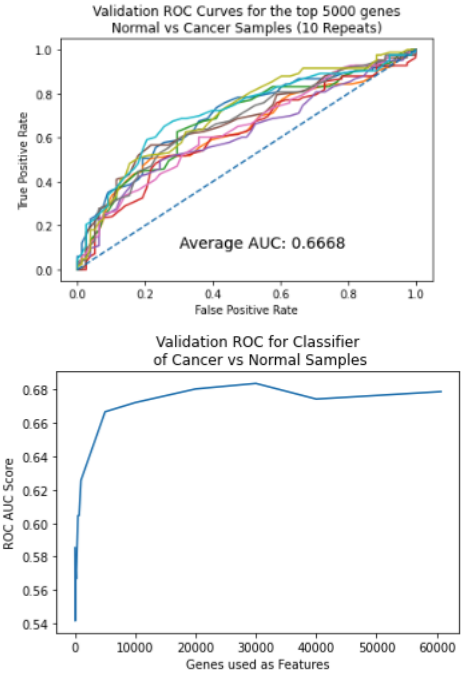

Figure 2: **Top:** Average AUC on the test set of classifiers using top 5000 genes. **Bottom:** Test set AUC for varying numbers of features

2. Taking a new approach selecting genes based on longitudinal gene expression changes after chemotherapy

4.1 Method 1: Refitting Normal vs Cancer to Recurrence Classification

4.1.1 Feature Selection using F-value

Similar to the Cancer vs Normal classifier, we used the ANOVA F-value to select the genes with the most variation between non-recurrent and recurrent breast cancer samples. 3-fold cross-validation was performed on Random Forest classifiers using between 5 and 60,000 genes as features. The accuracy was evaluated via the mean 3-fold cross validation accuracy and the Area under the Receiver Operating Curve.

In Figure 3, the results of this evaluation can be seen. AUC remains above 0.9 between 10 and 1000 features, however 3-fold classification accuracy peaks at 86.46% using just 10 features. Therefore, we decided the ideal number of features to test would be just 10, much fewer than had been used when differentiating cancer and non-cancer.

4.1.2 Results

When testing Random Forest Classifiers with just the top 10 F-value genes differentiating recurrent and non-recurrent samples on the test set, an AUC

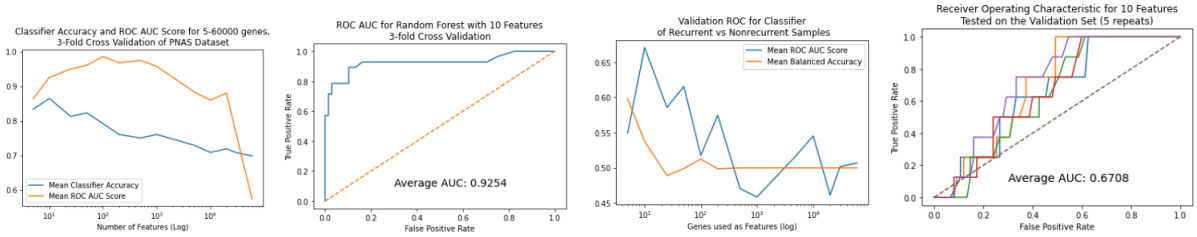

Figure 3: **Far Left:** Mean 3-fold cross validation accuracy peaks at 10 features, while AUC peaks around 100 features in the training set **Center Left:** Average ROC AUC for 3-fold Cross Validation of Random Forest Classifier in training set **Center Right:** Mean AUC and Mean Balanced Accuracy Rate of Random Forest Classifiers on testing set with increasing numbers of top genes used as features. Of note, both decrease as more genes are added **Far Right:** Average AUC for 5 Random Forest classifiers using just the top 10 features, on the testing set.

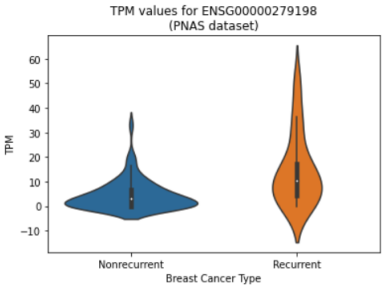

Figure 4: ENSG00000279198, a transcript from Chr19 of unknown biotype, was the transcript upregulated the most in recurrent cancer patients from the PNAS "training" dataset.

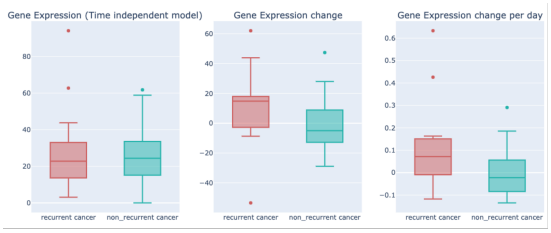

Figure 5: EI3FE gene expression under different computational models

of 0.6708 and balanced accuracy rate of 0.5378 was achieved (Figure 3). Even more surprisingly, classifiers using just the top 5 genes achieved an average AUC of 0.5495 and an even higher balanced accuracy rate of 0.5978. However, using more than 10 genes as features on the test set resulted in significantly decreased accuracy, and when exceeding 50 genes as features, every classifier became trivial, predicting non-recurrent cancer for each test set sample.

4.1.3 Top Transcripts for Classification

Because classifiers using the TPM values of just 10 transcripts or less outperformed other classifiers, we decided to research these transcripts and determine if they had been previously implicated in cancer recurrence. The 10 gene transcripts with the largest F-values from highest to lowest were: ENSG00000279198, ID4, ANXA5, PIGHP1, ENSG00000277692, KCTD12, AMPD2, SKAP2, UBE2FP3, and ENSG00000269931. While many of the top genes such as ID4, ANXA5, and KCTD12 have known associations with breast or other cancers, what we found most interesting were the unnamed transcripts (beginning with ENSG) which were significantly upregulated in recurrent patients. The most striking example was ENSG00000279198 (Figure 5) which had a mean TPM of 7.83 in nonrecurrent patients and 16.92 in recurrent patients. There is very little information on this transcript online, so it would be fascinating to further evaluate this transcript to see if it holds a link to cancer recurrence or if this finding is unique to this dataset.

4.1.4 Evaluation

This data suggests that when classifying the differences between recurrent and non-recurrent cancer using transcription differences, it is crucial to select a set of very specific genes to differentiate the two groups. This differs greatly from the cancer vs normal classifier, where more genes led to a great classification accuracy. Furthermore, more research is necessary to determine if these 10 transcripts play a significant role in breast cancer recurrence.

4.2 Method 2: Longitudinal change in gene expression

Recurrent cancer occurs when cancer cells of the same cancer type or of a different cancer type appear (or reappear) after treatment (chemotherapy). Our model hypothesis stems from that the appear-

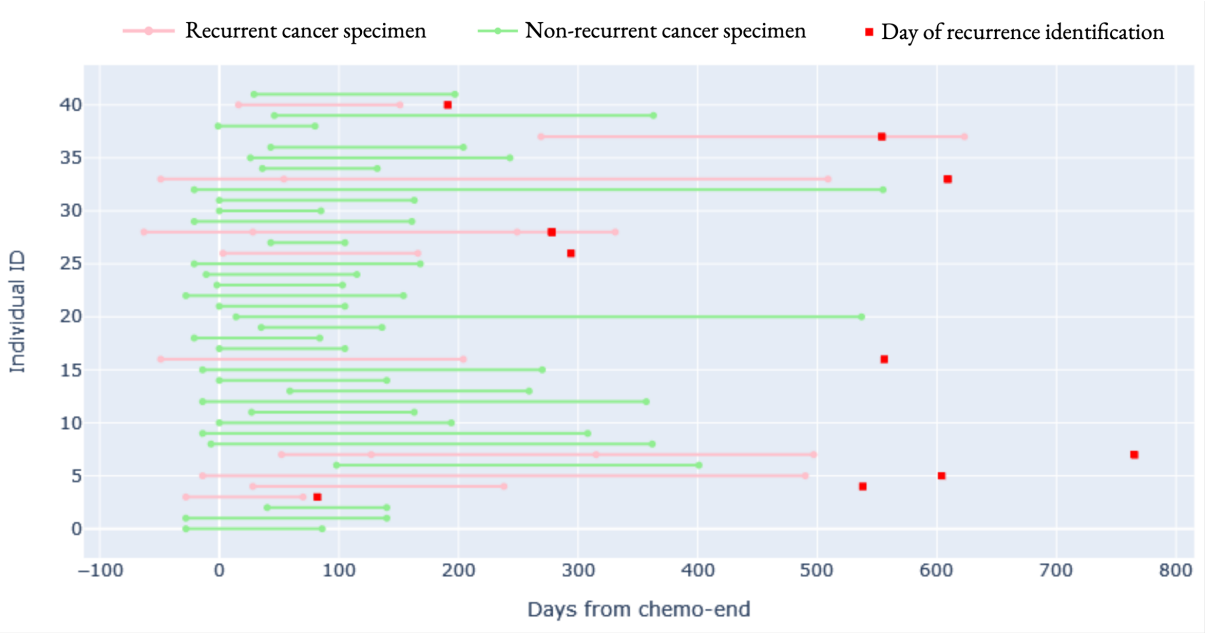

Figure 6: Timeline of specimen collection aligned with chemo-end date

ance of cancer cells is accompanied by changes in gene-expression of yet undetermined oncogenes. As an initial validation of our hypothesis, in Figure 5, we plot the expression levels of the EI3FE gene. When we consider the specimens to be independent (left barplot), we see no difference in gene expression distribution between cancer and non-recurrent cancer. However, when we consider the longitudinal gene expression change post-chemotherapy, we can notice that the distributions of recurrent and non-reccurent cancer individuals start to diverge. This validation provides us motivation to investigate into whether a model based on longitudinal gene expression change could distinguish between recurrent and non-recurrent cancer individuals.

4.2.1 Dataset

We construct the dataset by aligning the specimens with respect to the chemo end-date, as indicated in Figure 6. As we are interested in gene-expression changes after chemotherapy, we remove specimens that were collected 60 days before end of chemo-therapy and 800 days after end of chemo-therapy. After applying this filter, we have 42 individuals - 10 with recurrent cancer and 32 with non-recurrent cancer. Separately, since the dataset is sparse, we replace genes with zero expression level with their corresponding mean expression level. We also filter out genes for which we have fewer than 8 recurrent cancer individuals or fewer than 8 non-recurrent cancer individuals. After applying this filter to all the genes, we are left with 15897 genes.

Figure 7: Violin plot for identifying the top genes for our longitudinal gene expression change model; p-value is calculated using the student's t-test.

4.2.2 Feature Selection

We consider two models - gene expression change and gene expression change per day. For each of these models, we identify the top genes that can help distinguish recurrent cancer from non-recurrent cancer. For identifying the top genes, we use the student's t-test for finding the probability that the two cancer distributions are equal. We visualize the logarithm of the mean fold-change versus 1-p-value as a violin plot in Figure 7. We only consider the genes that have a mean-fold change of at-least 2 and p-value less than 0.05. Using this criteria, we obtain 88 genes for the gene-expression change model and 136 genes for the gene-expression change per-day model.

Figure 8: Longitudinal model results using 3-fold cross validation; (**left**) mean AUC scores for gene expression change model with varying number of genes; (**center**)mean AUC scores for gene expression change per-day model with varying number of genes; (**right**) ROC for gene experssion change model with 20 features;

4.2.3 Results

For classifying the specimens as recurrent or non-recurrent, we take the expression levels of the top genes shortlisted above and pass them to the Random Forest classifier. We use a 3-fold cross validation on the (Zhou, 2019) dataset for evaluating our model. As shown in Figure 8, the AUC saturates at around 0.75 for both the gene-expression change and gene-expression change per-day models. These results are encouraging as they indicate that longitudinal change in gene-expression shows promise for distinguishing recurrent from non-recurrent cancer specimens. Another interesting insight from these experiments is the lack of difference between gene-expression change and gene-expression change per-day model. This suggests that the absolute change in gene-expression between specimens matters the most for distinguishing recurrent cancer from non-recurrent cancer. However, we note that with the current model construction, we are not able to completely separate the recurrent specimens from non-recurrent specimens. As future work, we would like to explore whether gene-expression changes can be used as features for distinguishing recurrent cancer and non-recurrent cancer.

5 Author Contributions

Jeff: Built normal vs cancer classifier from front to back, as well as the F-value classifier for recurrent vs nonrecurrent breast cancer. Performed research and analysis on the transcripts that accounted for the greatest difference between recurrent and non-recurrent samples. Created AUC and violin plots for this data. Wrote and presented about the information in Sections 1-4.1 in this report.

Harish: Formulated the recurrent vs non-recurrent cancer problem as a longitudinal gene expression change model. Constructed an initial validation for this model's hypothesis. Upon initial validation, re-constructed the provided dataset (Zhou, 2019) by aligning to suit the model construction and filtered out outliers. Provided a mechanism for handling data sparsity. Performed feature selection using student's t-test statistic. Evaluated the proposed model using cross-validation for different number of features. Wrote and presented about information in Section 4.2 in this report.

6 References

1. "Breast Cancer Facts and Statistics." Breast Cancer Facts and Statistics, 10 Mar. 2022, https://www.breastcancer.org/facts-statistics.

2. Li, Y., Lu, S., Zhang, Y. et al. Loco-regional recurrence trend and prognosis in young women with breast cancer according to molecular subtypes: analysis of 1099 cases. World J Surg Onc 19, 113 (2021). https://doi.org/10.1186/s12957-021-02214-5

3. Zhou, Zixu, et al. "Extracellular RNA in a Single Droplet of Human Serum Reflects Physiologic and Disease States." Proceedings of the National Academy of Sciences, vol. 116, no. 38, 2019, pp. 19200–19208., https://doi.org/10.1073/pnas.1908252116.

4. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

5. https://www.cancer.gov/types/recurrent-cancer