BREAST INVASIVE CARCINOMA MIRSEQ ANALYSIS

Alan Dimitriev – 20062431

Abstract

Breast invasive carcinoma is one of the most commonly diagnosed cancers in Canada, and has a strong genetic basis for its expression. Staging in relation to breast cancer outlines if (and how far) cancer has spread once diagnosed. Of the three main staging features in the TNM system, the N stage represents whether or not the cancer has spread to nearby lymph nodes. In this study we aim to investigate whether miRNAseq reads can be used as discriminating features of a binary pathological N stage classification system (whether or not a given patient's cancer has spread to any nearby lymph nodes or none). To perform this investigation, we applied two statistical tests: Spearman correlation and Mann Whitney U tests in SPSS, and performed both unsupervised and supervised learning. The result of this investigation has concluded that it is unlikely that miRNA can serve as a discriminating feature for pathological N staging.

Introduction/Background/Motivation

Women in the United States have a 1 in 8 chance of developing an invasive form of breast cancer during their lifetime. While certain factors like obesity, dense breasts, or age can put women at a higher risk of developing breast invasive carcinoma, an individual's genetics also play an important role in risk probability. Thus, a prominent method for exploring the relationship between genetics and breast cancer expression is to analyze the association between the expression of certain genes with clinical features. Specifically, microRNA's (MiRNAs) serve as a particular target of interest for uncovering genetic inferences about cancer expression. This is because MiRNAs have been shown to have effects on the hallmarks of cancer when dysregulated. A microRNA is a small single-stranded non-coding RNA molecule that functions in RNA silencing and post-transcriptional regulation of gene expression. An increasing number of studies have identified MiRNAs as potential biomarkers for human cancer diagnosis, making them an ideal starting point for understanding statistical analysis for genetic data.

One area of interest with regards to breast cancer is 'cancer staging', where doctors try to evaluate if (and how far) breast cancer has spread within a patient. There are seven key pieces of information used in modern staging systems (either surgical stage of clinical stage) with the main three being: the size of the primary tumor (pathological T stage), the distant metastasis (pathological M stage) and the spread of the cancer to lymph nodes (pathological N stage). This study focuses on investigating whether microRNAseq read counts are associated with the phenotypic expression of pathological N stage amongst patients with breast invasive carcinoma. Lymph node staging for breast cancer is based on how the nodes look under the microscope, and a deposit of cancer cells must contain atleast 200 cells or be at least 0.2mm across for it to change the N stage (area of cancer spreads smaller than 0.2mm or fewer than 200 cells doesn't change the stage but alters the abbreviation suffixed onto the original stage diagnosis). There are five levels of stage progression when dealing with pathological N stage: NX, N0, N1, N2, and N3. Nx indicates that the nearby lymph nodes cannot be assessed (for example, if they were removed previously). N0 indicates that the cancer has not spread to nearby lymph nodes. N1 indicates that the cancer has spread to 1 to 3 axillary (underarm) lymph node(s), and/or cancer if found in the internal mammary lymph nodes. N2 indicates that the cancer has spread to 4 to 9 lymph nodes under the arm, or cancer has enlarged the internal mammary lymph nodes. A pathological stage of N3 can indicate a variety of outcomes including: cancer spreading to 10 or more axillary lymph nodes, cancer has spread to lymph nodes under the collarbone, cancer has spread to the lymph nodes above the collarbone on the same side of the cancer with at least one area of cancer spread greater than 2mm.

Thus, for the purpose of this project we aim to investigate whether there is a genetic background to the clinical outcome of a patient's pathological N stage. We aim to see if the possibility exists to predict the likelihood of whether or not a patient diagnosed with breast cancer will have that cancer spread to other lymph nodes in their body. We will perform this investigation by performing statistical analysis, as well as supervised and unsupervised learning on miRNAseq data paired with associated clinical staging data.

Methods

Data Procurement and Linking

The data for this project was retrieved from the Broad Institute of MIT and Harvard's firebrowse.org website. Specifically, we retrieved the files from the TCGA data version 2016_01_28 from the BRCA webpage (http://firebrowse.org/?cohort=BRCA&download_dialog=true#). Two specific files were downloaded for the purpose of this investigation:

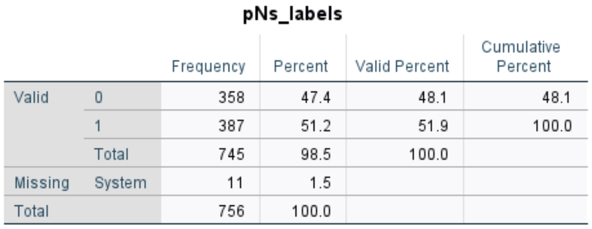
1. BRCA.mirnaseq__illuminahiseq_mirnaseq__bcgsc_ca__Level_3__miR_gene_expression__data.data.txt
2. BRCA.clin.merged.picked.txt

The first file containing the microRNAseq data, the second containing a select number of clinical features. Before we could link the two files, the miRSeq data needed to be filtered to ensure that only samples with a '01A' code present in the fourth section of their hybridization IDs were included. This was done to ensure all samples are from the same tissue type. All values marked as 'NA' in the clinical data were replaced with empty cells, this was done so that when the data is loaded into other programs numerical features are not automatically labelled as 'string' variables due to the presence of the 'NA' string.

Following those preparation steps both data sets were loaded into SPSS and merged using the 'add variable' tool selection. Any miRseq data that did not have accompanying clinical data was removed, and equally any clinical data that did not have accompanying miRseq data was also removed. This left us with 756 samples to work with for our project.

The next steps taken were to bin the clinical feature of 'pathological N stage' into binary groupings. A pathological N stage of 'n0' indicates that cancer has not spread to nearby lymph nodes in a given patient, and thus will be recoded to a binary label of '0'. All other pathological N stages that represent instances where the cancer in a patient has spread to surrounding lymph nodes will be labelled as '1'. This leaves us with 358 samples of label '0' where the cancers being monitored have not spread to surrounding lymph nodes, and 387 samples of label '1' where the cancer has in fact spread.

Table 1. This table displays the frequency of the distribution of our clinical labels of our data samples. A label of 0 indicates pathological N stage of N0, a label of 1 indicates a pathological N stage of N1 or greater.



pNs_labels

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	358	47.4	48.1	48.1
	1	387	51.2	51.9	100.0
	Total	745	98.5	100.0	
Missing	System	11	1.5		
Total		756	100.0		

Data Visualization and Pre-processing

In order to get an understanding of how our data looks and what approach we should take for pre-processing some data visualization was performed to investigate the underlying shape of our dataset.

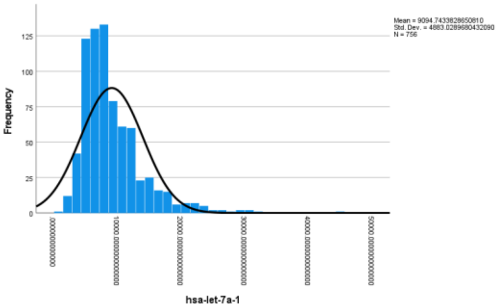


Figure 1. This histogram displays the distribution and normal curve of the read counts per million of the hsa-let-7a-1 gene for all samples in the dataset.

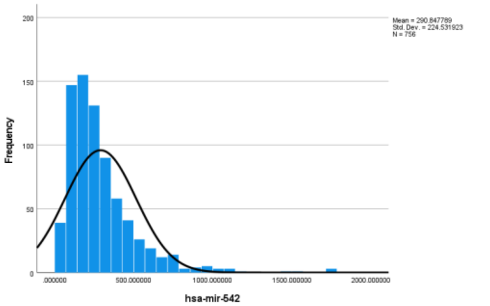


Figure 2. This histogram displays the distribution and normal curve of the read counts per million of the hsa-mir-452 gene for all samples in the dataset.

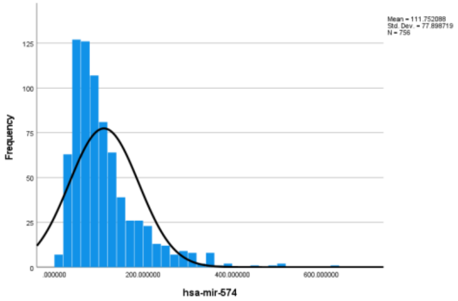


Figure 3. This histogram displays the distribution and normal curve of the read counts per million of the hsa-mir-574 gene for all samples in the dataset.

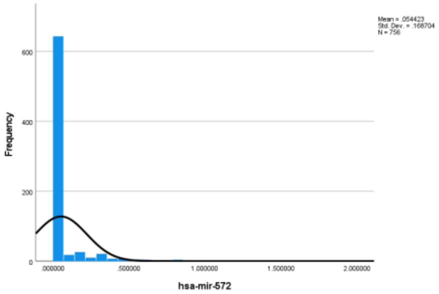
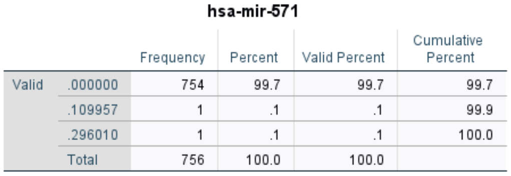


Figure 4. This histogram displays the distribution and normal curve of the read counts per million of the hsa-mir-572 gene for all samples in the dataset.

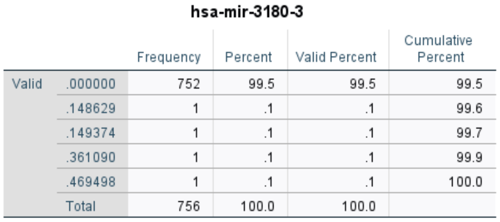
Table 2. This table displays the descriptive frequencies of the values in reads per million of the hsa-mir-571 gene across all samples.



hsa-mir-571

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.000000	754	99.7	99.7	99.7
	.109957	1	.1	.1	99.9
	.296010	1	.1	.1	100.0
	Total	756	100.0	100.0	

Table 3. This table displays the descriptive frequencies of the values in reads per million of the hsa-mir-571 gene across all samples.



hsa-mir-3180-3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.000000	752	99.5	99.5	99.5
	.148629	1	.1	.1	99.6
	.149374	1	.1	.1	99.7
	.361090	1	.1	.1	99.9
	.469498	1	.1	.1	100.0
	Total	756	100.0	100.0	

As can be seen in the histogram visualizations of select genes in Figures 1-3, our data is not normally distributed and thus it will be beneficial for us to subject the data to log2 transformation. Non-normal data distribution is not the only inference we can observe using histograms however, as Figure 4 highlights another problem with our data. Some of the genes are lowly expressed. This is even more noticeable when we look at the descriptive statistics for other genes (Figure 5 and Figure 6) which show that we have genes in our dataset whose expression value of zero makes up a significant percentage of our samples. These features do not provide us with any meaningful information and could hurt any future models we build that takes this data as input. Thus, in addition to transforming the data, we must remove any lowly expressed genes.

The data was ported from SPSS into MATLAB where we removed any genes whose read counts per million had a frequency of being 0 for over 90% of the samples. This lowered the number of genes we aimed to work with to 644. The data was then log2 transformed (normalization methods were not applied as we are working with 'reads per million' data which has already undergone a form of normalization).

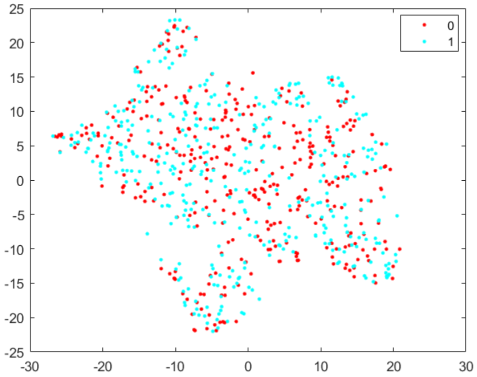


Figure 5. This scatter plot displays the t-distributed stochastic neighbour embedding of the full dataset following log2 transformation. The scatter plot is labeled by class where 0 represent pathological N stage of N0 and 1 represents a pathological N stage equal to or greater than N1.

To get a sense of how our data might be related in a higher dimensional space we plotted the data using t-distributed stochastic neighbor embedding (t-SNE). As can be seen in Figure 5, the data does not seem to visually have any distinct separation between samples labelled '0' (indicating that cancer has not spread to nearby lymph nodes) and samples labelled '1' (indicating samples where cancer has spread to 1 or more nearby lymph nodes). While there lacks a clear visual distinction this isn't necessarily concerning in terms of our investigation as our feature space is still extremely large thus making it difficult to extrapolate relationships from it.

The next data investigation we can perform is to investigate the data inter quartile range (IQR) to see if anything stands out, this will also help us identify any batch effects or outliers.

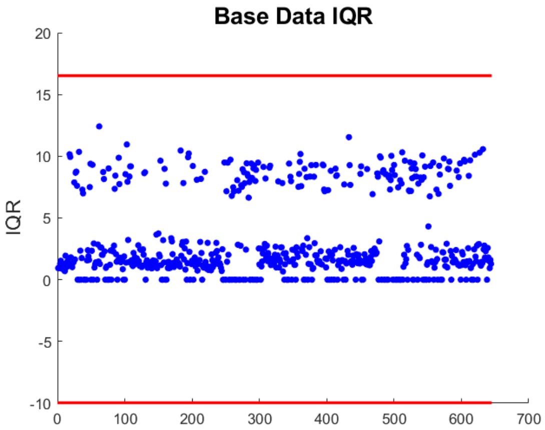


Figure 6. This scatter plot displays the interquartile range of each feature (gene) in the dataset.

As can be seen in Figure 6, our data has a lot of features whose IQR is 0. This indicates that these features lack any real variance and almost all samples have the same value reported for those specific genes. These genes do not offer much in the way of inference and thus we can remove them.

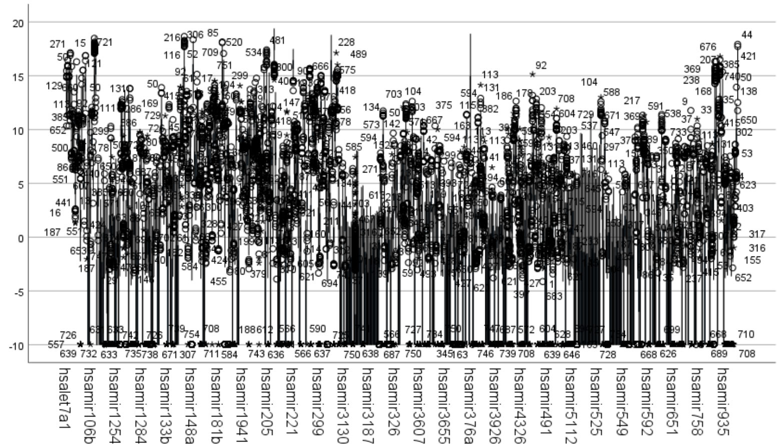


Figure 7. Boxplot of every feature (gene) in the dataset created using SPSS.

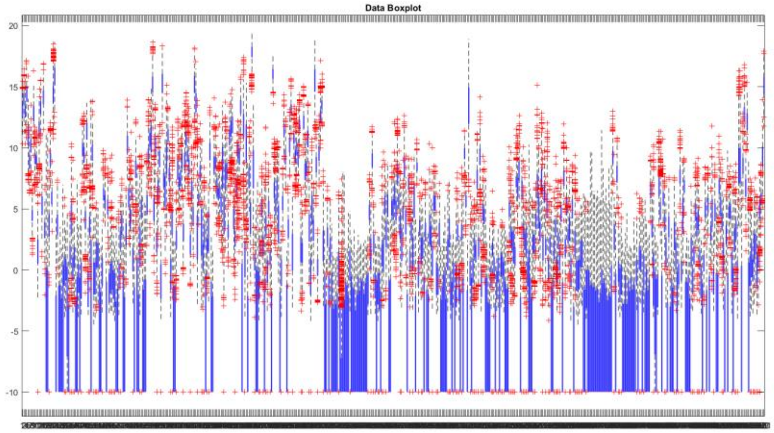


Figure 8. Boxplot of every feature (gene) in the dataset created using MATLAB.

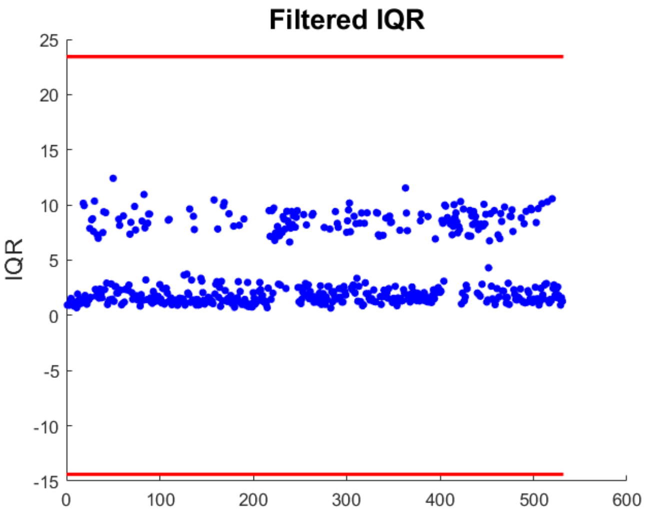


Figure 9. This scatter plot displays the interquartile range of each feature (gene) in the dataset after filtering was applied to remove any features with an IQR of zero.

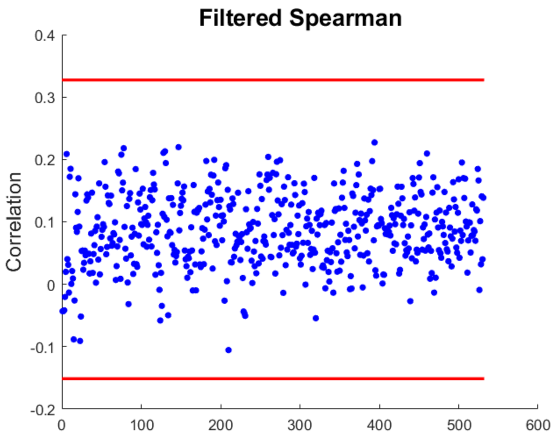


Figure 10. This scatter plot displays the Spearman correlation of all the features (genes) within the dataset. This was calculated using the SampleCorrelation function written in a previous assignment.

Due to the number of features (even after the filtering we've done we still have 531 genes) the boxplots in Figure 7 and Figure 8 aren't exactly easy to decipher. However, they do seem to indicate that there are no severe outliers or noticeable batch effects. The spearman correlation plot and IQR plots, of Figure 9 and Figure 10 respectively, also indicate that there is no visually discernable batch effect and that the reads are consistent throughout the samples. It can be noted however, that in terms of interquartile range there are two distinct levels of variability that exist: that being a 'high variability grouping' whose IQR is above 5, and a 'low variability group' whose IQR is below 5.

Statistical Analysis

After filtering out features with an IQR of zero, we then proceeded to perform two statistical tests. The aim of these test is to see if we can find if there are any genes associated with the clinical outcome of breast cancer pathology N stage. To investigate this, we performed Spearman correlation to see if a monotonic association exists between the genes in our filtered dataset and the binary outcome of n0 or >=n1. The idea being that if the distribution of any genes have a negative or positive correlation with the binary output label we will be able to distinguish them if they have a p value under the 0.05 significance threshold. The second statistical analysis method applied was the Mann Whitney U Test. The aim of this test is to see if the distribution of the samples for each gene is the same across the binary categories of the pathology N stage bins we outlined.

Unsupervised Learning

Following the approach described on the firebrowse page associated with the BCRA data, when investigating unsupervised learning/clustering for our given data set we filtered down to the most variable features. The original firebrowse approach used a magic number cut off of 150, however we filtered according to the IQR distribution differences seen in Figure 11 by taking only the most variable features (those whose IQR was equal or greater to 5) resulting in a working set of 161 features deemed 'most variable'. WE implemented three unsupervised learning approaches post variability filtering: TSN-E and K-Means clustering, and hierarchical clustering.

Supervised Learning

We investigated the efficacy of using miRNAseq reads to produce a classification model to identify whether a given sample's pathological N stage was n0 or greater than or equal to a stage of n1. Three separate attempts were made using three different feature selection methods: statistical significance, rank feature importance, and high variability filtering. All investigations were done using MATLAB's classification learner application. For each feature space investigated we constructed a model for each available type of classifier in the application and compared their performance.

Results

Statistical Analysis

Table 4. This table displays the result of the Spearman correlation statistical test, displaying the p value and correlation coefficient of each gene in regards to the binary outcome class of pathological N stage.

Gene	Sig.	Correlation Coefficient
hsa-let-7d	0.004	-0.104
hsa-let-7g	0.004	-0.106
hsa-mir-106a	0.003	-0.109
hsa-mir-106b	0.011	-0.094
hsa-mir-10a	0.007	0.099
hsa-mir-10b	0.019	0.086
hsa-mir-1229	0.026	-0.081
hsa-mir-1248	0.019	-0.086
hsa-mir-1255a	0.007	-0.099
hsa-mir-1258	0.007	0.99
hsa-mir-12702	0.044	-0.074
hsa-mir-1277	0.028	-0.081
hsa-mir-1281	0.031	-0.079
hsa-mir-1282	0.02	-0.085
hsa-mir-1291	0.006	-0.1
hsa-mir-1306	0.019	-0.086
hsa-mir-1307	0.012	-0.092
hsa-mir-130b	0.018	-0.087
hsa-mir-1382	0.012	-0.091
hsa-mir-139	0.013	0.091
hsa-mir-141	<.001	-0.136
hsa-mir-1468	0.036	-0.077
hsa-mir-147b	0.043	-0.074
hsa-mir-148a	0.009	-0.096
hsa-mir-1538	0.005	-0.102
hsa-mir-15a	0.014	-0.09
hsa-mir-15b	0.004	-0.106
hsa-mir-161	<.001	-0.125
hsa-mir-162	<.001	-0.121
hsa-mir-17	0.002	-0.115
hsa-mir-186	0.021	-0.084
hsa-mir-187	0.028	-0.08
hsa-mir-188	0.007	-0.099
hsa-mir-18a	<.001	-0.131
hsa-mir-18b	0.024	-0.083
hsa-mir-191	0.028	-0.081
hsa-mir-1914	0.021	-0.085
hsa-mir-197	0.015	-0.09
hsa-mir-19a	0.004	-0.105
hsa-mir-19b1	0.003	-0.108
hsa-mir-19b2	0.01	-0.095
hsa-mir-200a	0.029	-0.08
hsa-mir-200c	0.032	-0.079
hsa-mir-202	0.011	0.093
hsa-mir-204	0.046	0.073
hsa-mir-20a	0.024	-0.083
hsa-mir-20b	0.023	-0.083
hsa-mir-210	0.045	-0.073
hsa-mir-221	0.004	-0.106
hsa-mir-222	0.002	-0.116
hsa-mir-2277	0.024	-0.083
hsa-mir-242	0.025	-0.082
hsa-mir-27a	0.033	-0.078
hsa-mir-28	0.031	-0.079
hsa-mir-29a	0.043	-0.074
hsa-mir-3074	0.036	-0.077
hsa-mir-30c1	0.012	-0.092
hsa-mir-30c2	0.018	-0.086
hsa-mir-30e	<.001	-0.127
hsa-mir-3138	0.004	-0.106
hsa-mir-31561	0.006	0.101
hsa-mir-31562	0.037	0.076
hsa-mir-31563	0.004	0.105
hsa-mir-31991	0.028	-0.081
hsa-mir-32	0.047	-0.073
hsa-mir-320c1	0.042	-0.075
hsa-mir-320d1	0.029	-0.08
hsa-mir-324	0.023	-0.083
hsa-mir-33a	0.012	-0.092
hsa-mir-345	0.009	-0.095
hsa-mir-3613	<.001	-0.144
hsa-mir-3620	0.037	-0.076
hsa-mir-3622a	0.022	-0.084
hsa-mir-363	0.042	-0.074
hsa-mir-3648	0.025	-0.082
hsa-mir-3676	0.001	-0.12
hsa-mir-3680	0.025	-0.082
hsa-mir-3682	<.001	-0.125
hsa-mir-3687	0.002	-0.115
hsa-mir-378	0.014	-0.09
hsa-mir-378c	0.04	-0.075
hsa-mir-3909	0.025	-0.082
hsa-mir-3922	0.015	-0.089
hsa-mir-429	0.01	-0.094
hsa-mir-455	0.029	-0.08
hsa-mir-505	0.004	-0.106
has-mir-545	0.014	-0.09
hsa-mir-548b	0.031	-0.079
hsa-mir-548d2	0.012	-0.092
hsa-mir-548v	0.049	-0.072
hsa-mir-556	0.004	-0.107
hsa-mir-574	0.049	-0.072
hsa-mir-577	<.001	-0.13
hsa-mir-579	0.026	-0.081
hsa-mir-584	0.049	-0.072
hsa-mir-590	<.001	-0.122
hsa-mir-592	<.001	-0.126
hsa-mir-597	0.029	-0.08
hsa-mir-598	0.009	-0.096
hsa-mir-618	0.012	0.092
hsa-mir-766	0.039	-0.076
hsa-mir-769	0.014	-0.09
hsa-mir-877	0.013	-0.091
hsa-mir-92a2	0.026	-0.082
hsa-mir-939	0.024	-0.083
hsa-mir-940	0.014	-0.09
hsa-mir-942	0.003	-0.11
hsa-mir-95	0.018	-0.087
hsa-mir-98	0.041	-0.075

Table 5. This table displays all the genes deemed ‘statistically significant’ following the application of the Mann Whitney U test and their associated p values.

Gene	Sig.
hsa-let-7d	0.005
hsa-let-7g	0.004
hsa-mir-106a	0.003
hsa-mir-106b	0.011
hsa-mir-10a	0.007
hsa-mir-10b	0.019
hsa-mir-1229	0.026
hsa-mir-1248	0.019
hsa-mir-1255a	0.007
hsa-mir-1258	0.007
hsa-mir-12702	0.044
hsa-mir-1277	0.028
hsa-mir-1281	0.031
hsa-mir-1282	0.02
hsa-mir-1291	0.006
hsa-mir-1306	0.02
hsa-mir-1307	0.012
hsa-mir-130b	0.018
hsa-mir-1382	0.013
hsa-mir-139	0.013
hsa-mir-141	<.001
hsa-mir-1468	0.036
hsa-mir-147b	0.043
hsa-mir-148a	0.009
hsa-mir-1538	0.005
hsa-mir-15a	0.014
hsa-mir-15b	0.004
hsa-mir-161	<.001
hsa-mir-162	<.001
hsa-mir-17	0.002
hsa-mir-186	0.021
hsa-mir-187	0.028
hsa-mir-188	0.007
hsa-mir-18a	<.001
hsa-mir-18b	0.024
hsa-mir-191	0.028
hsa-mir-1914	0.021
hsa-mir-197	0.015
hsa-mir-19a	0.004
hsa-mir-19b1	0.003
hsa-mir-19b2	0.01
hsa-mir-200a	0.029
hsa-mir-200c	0.032
hsa-mir-202	0.011
hsa-mir-204	0.046
hsa-mir-20a	0.024
hsa-mir-20b	0.023
hsa-mir-210	0.046
hsa-mir-221	0.004
hsa-mir-222	0.002
hsa-mir-2277	0.024
hsa-mir-242	0.025
hsa-mir-27a	0.033
has-mir-28	0.031
hsa-mir-29a	0.043
hsa-mir-3074	0.036
hsa-mir-30c1	0.012
hsa-mir-30c2	0.019
hsa-mir-30e	<.001
hsa-mir-3138	0.004
hsa-mir-31561	0.006
hsa-mir-31562	0.038
hsa-mir-31563	0.004
hsa-mir-31991	0.028
hsa-mir-32	0.047
hsa-mir-320c1	0.042
hsa-mir-320d1	0.029
hsa-mir-323b	0.023
hsa-mir-33a	0.012
hsa-mir-345	0.009
hsa-mir-3613	<.001
hsa-mir-3620	0.037
hsa-mir-3622a	0.023
hsa-mir-363	0.042
hsa-mir-3648	0.025
hsa-mir-3676	0.001
hsa-mir-3680	0.025
hsa-mir-3682	<.001
hsa-mir-3687	0.002
hsa-mir-378	0.014
hsa-mir-378c	0.04
hsa-mir-3909	0.025
hsa-mir-3922	0.015
hsa-mir-429	0.01
hsa-mir-455	0.029
hsa-mir-505	0.004
hsa-mir-545	0.014
hsa-mir-548b	0.031
hsa-mir-548d2	0.012
hsa-mir-548v	0.049
hsa-mir-556	0.004
hsa-mir-574	0.049
hsa-mir-577	<.001
hsa-mir-579	0.026
hsa-mir-584	0.049
hsa-mir-590	<.001
hsa-mir-592	<.001
hsa-mir-597	0.029
hsa-mir-598	0.009
hsa-mir-618	0.012
hsa-mir-766	0.039
hsa-mir-769	0.014
hsa-mir-877	0.013
hsa-mir-92a2	0.026
hsa-mir-939	0.024
hsa-mir-940	0.014
hsa-mir-942	0.003
hsa-mir-95	0.018
hsa-mir-98	0.041

As can be seen in Table 4, we distinguished 109 genes as being statistically significantly correlated to the binary classification of pathological n stage. As can be seen in Table 5, the Mann Whitney U Test performed in SPSS identified the same 109 genes that were deemed statistically significant by the Spearman test as being significant.

Unsupervised Learning

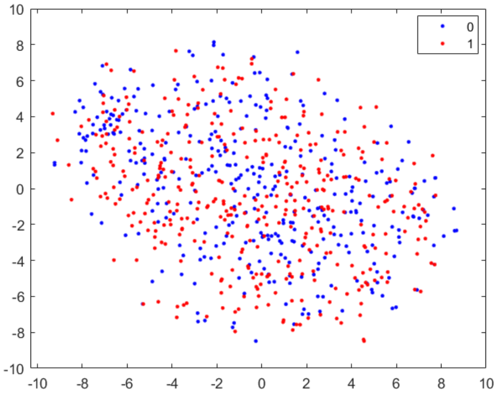


Figure 11. This scatter plot displays the results of t-distributed stochastic neighbour embedding applied to the 161 most variable features in the dataset.

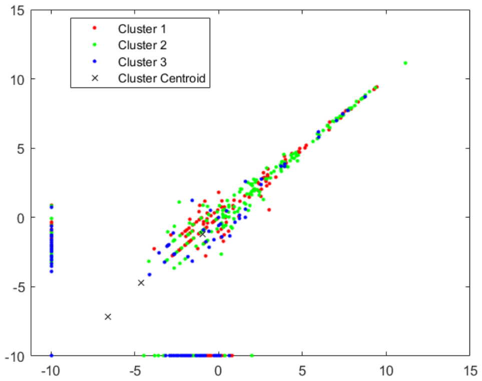


Figure 12. This scatter plot shows the results of k-means clustering with a k of 3 on the 161 most variable features.

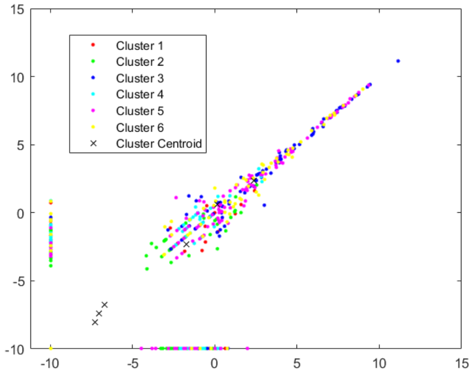


Figure 13. This scatter plot shows the results of k-means clustering with a k of 6 on the 161 most variable features.

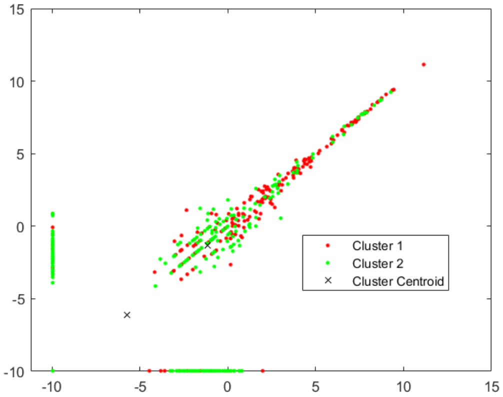


Figure 14. This scatter plot shows the results of k-means clustering with a k of 2 on the 161 most variable features.

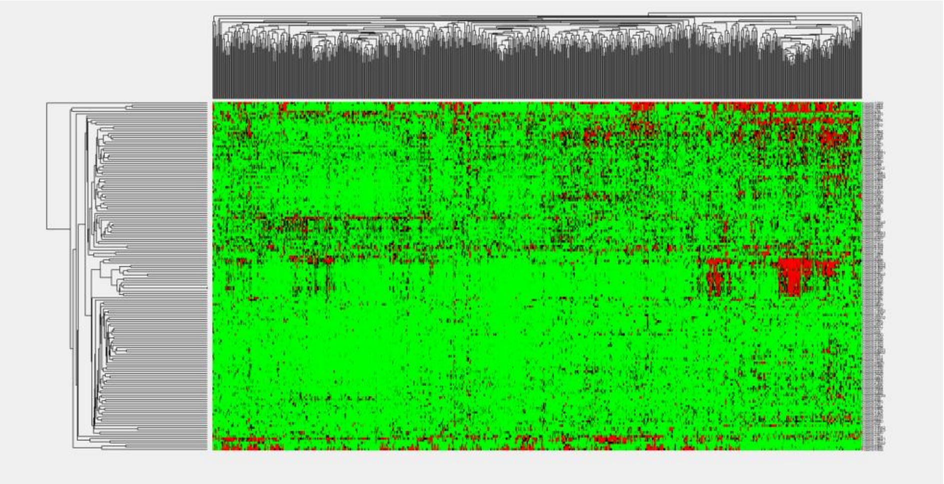


Figure 15. This figure displays the clustergram/dendrogram of the 161 most variable features against the binary classification label of pathological N stage outlined earlier using Spearman correlation.

Figures 11 through 15 display the results of various clustering methods when applied to the 161 most variable features in our dataset.

Supervised Learning

Statistical Significance

For our first attempt we built and trained models using 3-fold cross validation on a feature set containing only genes that were deemed ‘statistically significant’ through statistical analysis done in SPSS.

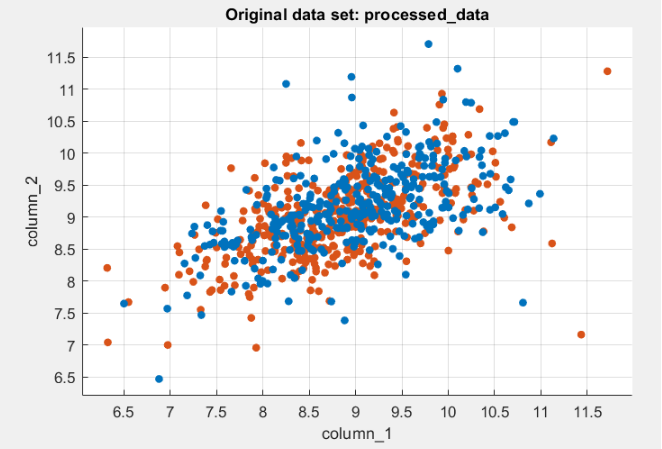
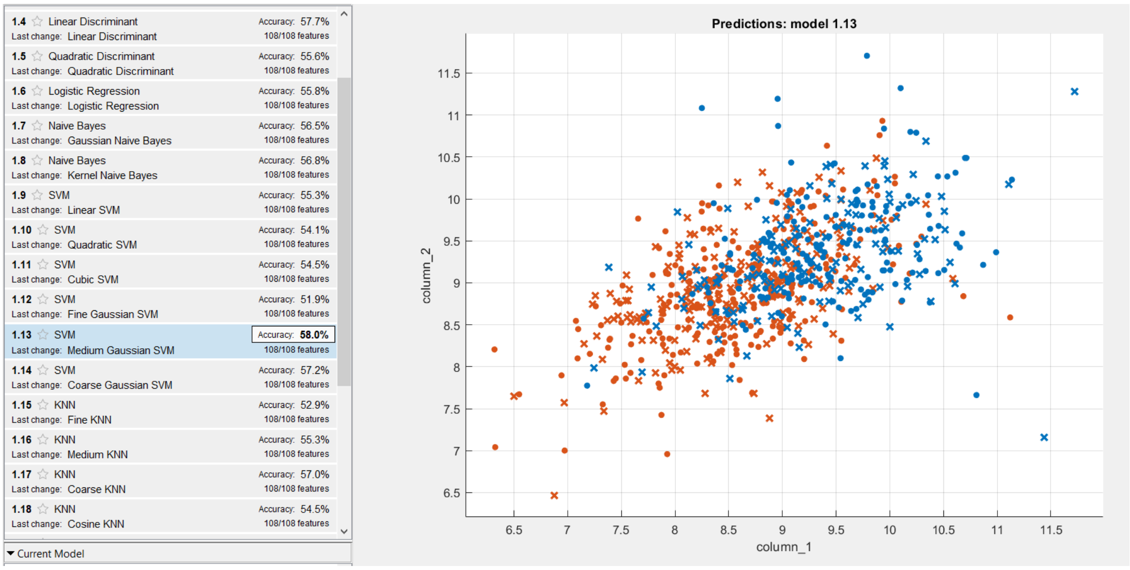


Figure 16. This figure shows the scatterplot of the sampels using a feature space of only the 109 features deemed to be ‘statistically signifcant’ as a result of Spearman and Mann Whitney U Tests in SPSS. Samples of class 0 (pathological N stage N0) are labelled in blue and samples of 1 (pathological N stage equal to or greater than N1) are labelled in orange.



1.4 ☆ Linear Discriminant — Accuracy: 57.7%
Last change: Linear Discriminant — 108/108 features

1.5 ☆ Quadratic Discriminant — Accuracy: 55.6%
Last change: Quadratic Discriminant — 108/108 features

1.6 ☆ Logistic Regression — Accuracy: 55.8%
Last change: Logistic Regression — 108/108 features

1.7 ☆ Naive Bayes — Accuracy: 56.5%
Last change: Gaussian Naive Bayes — 108/108 features

1.8 ☆ Naive Bayes — Accuracy: 56.8%
Last change: Kernel Naive Bayes — 108/108 features

1.9 ☆ SVM — Accuracy: 55.3%
Last change: Linear SVM — 108/108 features

1.10 ☆ SVM — Accuracy: 54.1%
Last change: Quadratic SVM — 108/108 features

1.11 ☆ SVM — Accuracy: 54.5%
Last change: Cubic SVM — 108/108 features

1.12 ☆ SVM — Accuracy: 51.9%
Last change: Fine Gaussian SVM — 108/108 features

1.13 ☆ SVM — Accuracy: **58.0%**
Last change: Medium Gaussian SVM — 108/108 features

1.14 ☆ SVM — Accuracy: 57.2%
Last change: Coarse Gaussian SVM — 108/108 features

1.15 ☆ KNN — Accuracy: 52.9%
Last change: Fine KNN — 108/108 features

1.16 ☆ KNN — Accuracy: 55.3%
Last change: Medium KNN — 108/108 features

1.17 ☆ KNN — Accuracy: 57.0%
Last change: Coarse KNN — 108/108 features

1.18 ☆ KNN — Accuracy: 54.5%
Last change: Cosine KNN — 108/108 features

Figure 17. This figure shows the scatter plot results of the most accurate predictive model trained on the dataset consisting of a feature space of only ‘statistically significant’ genes. Samples of binary class 0 are shown in blue; samples of binary class 1 are shown in orange. Correct predictions are represented as circles and incorrect predictions are represented as x’s.

Rank Feature Importance

For our second attempt we used a feature space consisting only of the top 50 features (picked from the pool of 531 features after filtering out genes with an IQR of zero) deemed 'most important' to predicting the binary label using the fscmrmr (minimum redundancy maximum relevance algorithm) feature rank method in MATLAB. The top 50 were chosen as this provided a minimum ranks score for each feature of at least 0.005.

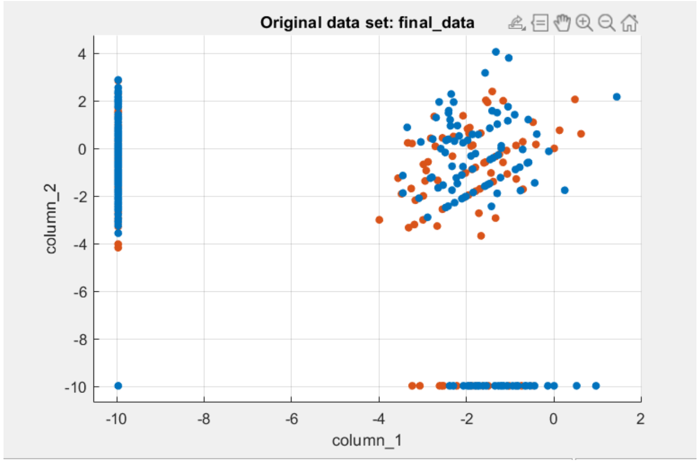
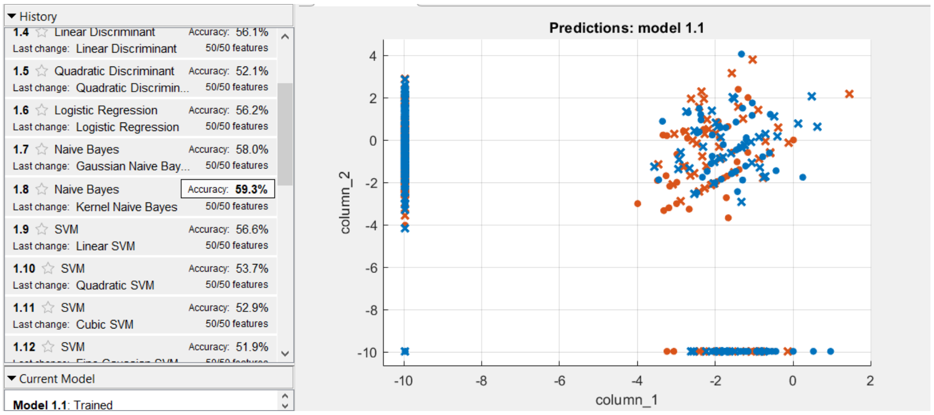


Figure 18. This figure shows the scatterplot of the sampels using a feature space of only the top 50 features deemed to be most important to label prediction. Samples of class 0 (pathological N stage N0) are labelled in blue and samples of 1 (pathological N stage equal to or greater than N1) are labelled in orange.



1.4 ☆ Linear Discriminant Accuracy: 56.1%
Last change: Linear Discriminant 50/50 features

1.5 ☆ Quadratic Discriminant Accuracy: 52.1%
Last change: Quadratic Discrimin... 50/50 features

1.6 ☆ Logistic Regression Accuracy: 56.2%
Last change: Logistic Regression 50/50 features

1.7 ☆ Naive Bayes Accuracy: 58.0%
Last change: Gaussian Naive Bay... 50/50 features

1.8 ☆ Naive Bayes Accuracy: **59.3%**
Last change: Kernel Naive Bayes 50/50 features

1.9 ☆ SVM Accuracy: 56.6%
Last change: Linear SVM 50/50 features

1.10 ☆ SVM Accuracy: 53.7%
Last change: Quadratic SVM 50/50 features

1.11 ☆ SVM Accuracy: 52.9%
Last change: Cubic SVM 50/50 features

1.12 ☆ SVM Accuracy: 51.9%

Model 1.1: Trained

Figure 19. This figure shows the scatter plot results of the most accurate predictive model trained on the dataset consisting of a feature space of the top 50 most impactful genes in terms of outcome prediciton. Samples of binary class 0 are shown in blue; samples of binary class 1 are shown in orange. Correct predictions are represented as circles and incorrect predictions are represented as x's.

High Variance Only

The final feature set we experimented with when trying to make our predictive model was to only use the most variable features (IQR greater than or equal to five) as our starting pool and then select based off rank importance the features from that pool to train our model on. Based on the chart of feature importance (Figure 20) we made a cutoff of 17 features as there was a slight distinct uptick in feature importance at that range.

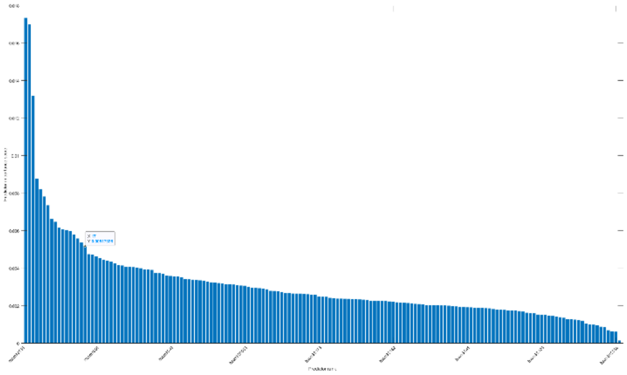
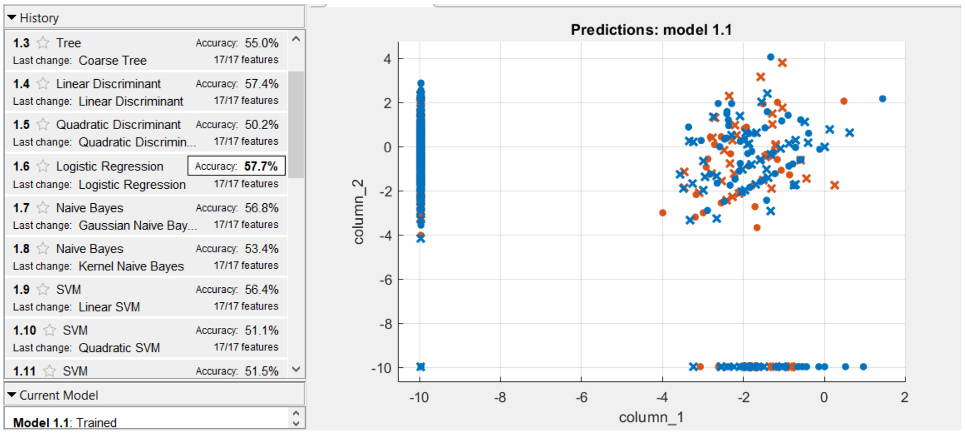


Figure 20. This chart displays the feature importance of the high variability genes.



▼ History

1.3 ☆ Tree Accuracy: 55.0%
Last change: Coarse Tree 17/17 features

1.4 ☆ Linear Discriminant Accuracy: 57.4%
Last change: Linear Discriminant 17/17 features

1.5 ☆ Quadratic Discriminant Accuracy: 50.2%
Last change: Quadratic Discrimin... 17/17 features

1.6 ☆ Logistic Regression Accuracy: 57.7%
Last change: Logistic Regression 17/17 features

1.7 ☆ Naive Bayes Accuracy: 56.8%
Last change: Gaussian Naive Bay... 17/17 features

1.8 ☆ Naive Bayes Accuracy: 53.4%
Last change: Kernel Naive Bayes 17/17 features

1.9 ☆ SVM Accuracy: 56.4%
Last change: Linear SVM 17/17 features

1.10 ☆ SVM Accuracy: 51.1%
Last change: Quadratic SVM 17/17 features

1.11 ☆ SVM Accuracy: 51.5%

▼ Current Model

Model 1.1: Trained

Figure 21. This figure shows the scatter plot results of the most accurate predictive model trained on the dataset consisting of a feature space of the top 17 most important features for outcome prediction pulled from a feature set that only included features whose IQR was above 5. Samples of binary class 0 are shown in blue; samples of binary class 1 are shown in orange. Correct predictions are represented as circles and incorrect predictions are represented as x's.

Discussion

While Table 4 and Table 5 display a large number of genes that are technically 'statistically significant' in terms of being both correlated to the distribution of the binary class pathological N stage labels and being non-equally distributed in across either 0 (pathological N stage N0) or 1 (pathological N stage greater than or equal to N1), this does not necessarily mean however that these are valid or reliable associations. This is highlighted specifically in Table 4, which displays the correlation coefficient associated with each gene and the binary label outcome. A perfect Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotone function of the other, the closer to +1 or -1 a coefficient is the more representative the feature is of the associated outcome. In our case, not one feature has a correlation above an absolute value of 0.2, with most having an absolute value less than 0.1. This implies that while in our correlation test the features, we have listed in the tables might be 'statistically significant' they are still not actually that strongly correlated with the output class. Less can be inferred from the results of the Mann Whitney U test as the significance and genes identified by the test are equal to the results of the Spearman correlation. However, these preliminary statistical results do offer us some insight that it might be difficult to produce an accurate predictive model or meaningful clustering on our given dataset as the correlation between the distribution of miRNAseq reads and the binary outcome class labels does not seem to be strongly correlated.

With that in mind we can shift our focus to the results of our unsupervised learning. Figure 11 displays the t-SNE plot of the 161 most variable features in our dataset. We reduced to the most variable features for unsupervised learning following similar methods taken by the investigation posted on the firebrowse webpage associated with our dataset. As can be seen in the figure, the plot is non-differentiable and the data (even after filtering out a large portion of the feature space) is not easily clustered. While this aligns with the results of our statistical tests, that being that there might not be solid associations/differences in the miRNAseq reads between the binary class labels, it also implies that the task of accurately predicting our binary class labels from the data might not be achievable. Figures 12, 13, and 14, all show the results of K-means clustering using K values: 2, 3, and 6. What's interesting is that because K-means clustering (by nature if being unsupervised) does not use labels when generating its clusters. The original investigation published by the firebrowse team on the associated webpage claimed that the ideal cluster number for this data was 6 (it should be noted that they used a feature space of 150 and consensus NMF clustering on 1078 samples), however we can visually infer from these figures that not only are there no well-defined clusters at any K value, there is also no indication of which K value would be preferable to the others or if this problem can be clustered at all. Figure 15 displays the resulting clustergram and dendogram of our data and serves to further the point that there is no clear way to cluster our data.

Once our data was loaded into MATALB's classifier learner app this problem of our data lacking clear distinguishable clusters became apparent again, as seen in Figures 16 and 18. Both class labels are not easily distinguishable from each other, and even in cases like in Figure 18 where there a large regions of higher dimension space dominated by one class label (in this case the line shaped groupings near the borders of the scatter plot are populated heavily by blue labels of class 0) there still exists outliers of the opposite class. Figures 17, 19, and 21 all showcase that it is indeed nearly impossible to make an accurate predictive model for our binary class output regardless of how we approach feature selection. The highest accuracy we recorded was around 59%, which for a binary predictive model is not ideal.

We can attribute the poor performance of our models to the lack of any differentiable clustering in our data. The question then shifts to why does our data lack such distinction and instead have this intermixed blob like structure highlighted in Figure 11? One potential reason is that pathological N stage is subjective in a sense. While there are clear delineations between stages exist, it is up to the discretion of the pathologist analyzing the tissue samples to label them as a given stage. While this shouldn't be a problem for our investigation as there is a distinct difference between a label of N0 and stages greater than or equal to N1, the interesting problem is that just because a patient's cancer has not currently spread to surrounding tissue at the time of reporting does not mean it will never spread. This leads to the potential issue of us looking at pathology N stage from a genetic angle where a patient labelled N0 could potentially have cancer that one day could spread to other lymph nodes. If this was the case then we would have patients in our N0 group who (if we are assuming that N stage is influenced by genetic forces) have a genetic makeup more inline with those in the N1 or greater class.

That being said, the most likely answer to why our predictive models performed poorly and why we failed to generate meaningful clusters is that miRNAs do not serve as discriminating features for pathological N stage. This was forewarned in our early-stage statistical analysis when our 'statistically significant' genes all had low absolute values of their correlation coefficients. This was an early indicator that miRNAs were not ideal discriminating features for the classification problem at hand. Figure 22 showcases this issue as when we look at the 20 top ranking miRNAs determined by fscmrmr we can see that there is no discrimination between the two classes.

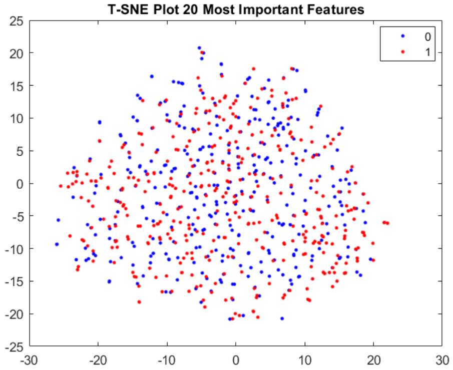


Figure 22. This scatter plot displays the results of t-distributed stochastic neighbour embedding applied to the 20 features deemed the most important by rank for outcome prediction using fscmrmr in MATLAB.

In conclusion it is likely that miRNAs do not serve as discriminating features for pathological N stage diagnosis in breast cancer patients. While this does not rule out the possibility of a genetic linkage between miRNA expression and pathological N stage outcomes, future research done with the aim of predicting pathological N stage diagnosis or likelihood in breast cancer patients should focus on other features with the hopes that they may provide a higher discriminative ability.