Predictive Modeling and Network Analysis of Gene Expression in Breast Cancer

Eleonora Mesaglio, Margherita Rigato, Francesco Lollato

Abstract

Cancer represents one of the major challenges in medical research, and understanding it at the molecular level is crucial for developing targeted therapies and improving early diagnosis. In this study, the primary goal is to accurately classify tissues as cancer or normal using gene expression data. Additionally, a secondary objective is to analyze whether the correlations between gene expressions vary between normal and cancerous tissues, to provide insights into the molecular mechanisms altered in the tumor context.

1. Dataset

The analyzed dataset was obtained from the *Gene Expression Omnibus (GEO)* platform and includes a total of 289 samples evenly divided into two groups:

- 146 samples of adjacent normal breast tissue (normal);
- 143 samples of breast adenocarcinoma (cancer).

A preliminary analysis of the dataset reveals that the data matrix has a dimension of 289 rows, the analyzed samples, and 35982 columns, the genes for which expression was measured. The original dataset was already balanced, it had already undergone quantile normalization, a fundamental preprocessing step for gene expression datasets, and no missing value was detected.

With the dataset prepared, we proceeded to apply various classification models to distinguish between cancer and normal tissue samples based on their gene expression profiles. The performance of these models was evaluated using the following metrics: accuracy (overall correctness of predictions), sensitivity (ability to identify cancer samples), and specificity (ability to correctly classify normal samples).

2. Lasso Logistic Regression

Lasso Logistic Regression is a type of logistic regression that uses L_1 regularization to perform variable selection by shrinking less relevant predictors to zero, thereby inducing sparsity in the model. The regularization parameter λ controls the strength of this penalization, with higher values leading to more coefficients being eliminated. This makes Lasso particularly effective for high-dimensional datasets, where variable selection is crucial.

In our analysis, we applied Lasso Logistic Regression to select the most relevant predictors. Through cross-validation, we determined the optimal λ value: $\lambda_{min} = 0.01217137$, which minimizes the cross-validation error. The final model was trained using such λ_{min}, retaining 72 features after regularization.

Performance: The model achieved good performance metrics, with a test error rate of 11.1%, sensitivity of 80.0%, and specificity of 100%. This indicates that while the model is very effective in detecting true negatives, it fails to detect a portion of true positives.

3. Lasso Support Vector Machines

Support Vector Machines (SVM) are effective for handling high-dimensional data, separating classes by maximizing the margin between them. We applied SVM models with Lasso regularization for classification, using both squared hinge loss and hinge loss.

For the SVM with **squared hinge loss**, cross-validation identified an optimal regularization parameter of $\lambda_{min} = 0.0586$. The final model, trained with this value, selected 84 variables.

Similarly, for the SVM with **hinge loss**, cross-validation determined $\lambda_{min} = 0.1125$ as the optimal value. The final model, trained with this parameter, selected 117 variables, significantly more than the squared hinge loss model.

Performance: The SVM with **hinge loss** achieved a prediction error of 9.72%, slightly better than the squared hinge loss model. It demonstrated a sensitivity of 82.5% and a specificity of 100%, highlighting its superior ability to correctly identify positive cases without compromising the absence of false positives. Additionally, this model selected 117 variables, indicating a more inclusive feature set.

The SVM with **squared hinge loss** achieved a prediction error of 11.1%, with a sensitivity of 80% and a specificity of 100%. This indicates that the model is less accurate in identifying true positives. Notably, it selected only 84 variables, resulting in a sparser model.

Overall, the hinge loss model offers slightly better classification performance, while the squared hinge loss model emphasizes sparsity.

4. Relaxed Lasso

Recognizing that Lasso, while effective for variable selection, introduces bias by shrinking non-zero coefficients toward zero, we attempted to refine our analysis using a Relaxed Lasso approach. This involved fitting a standard Logistic Regression model with the variables selected by Lasso to obtain unbiased coefficient estimates. However, the presence of strong collinearity among the variables selected by the Lasso prevented the algorithm from converging.

5. Elastic Net Logistic Regression

Since we observed a significant correlation between predictors, we applied an Elastic Net penalization instead of Lasso. The Elastic Net combines Lasso (L_1) and Ridge (L_2) penalties, effectively addressing feature selection and multicollinearity. While Lasso tends to select a small subset of variables, it may struggle with highly correlated predictors, arbitrarily selecting one and shrinking the others toward zero. Ridge, on the other hand, spreads the penalty across correlated variables but does not perform variable selection. Elastic Net overcomes these limitations by combining the L_1 and L_2 penalties, balancing their effects to allow correlated variables to be selected together while still performing effective regularization and variable selection.

Parameters calibration: To optimize the Elastic Net model, the hyperparameters α and λ were tuned using cross-validation.

The parameter α controls the balance between Lasso and Ridge penalties, where $\alpha = 1$ corresponds to pure Lasso regression and $\alpha = 0$ corresponds to pure Ridge regression. A systematic search was performed, starting with a broad range of α values from 0 to 1, identifying $\alpha = 0.8$ as the best. A refined search around 0.8 led to the optimal $\alpha = 0.744$. This value of α balances the Lasso and Ridge penalties, with a slight preference for Lasso. Since the dataset contains a large number of correlated variables, this balance helps in handling multicollinearity.

Next, λ was tuned using cross-validation, with the optimal value found as $\lambda_{min} = 0.0112$, which minimizes the cross-validation error.

The choice of these parameters was crucial in ensuring that the model could handle the complex, high-dimensional nature of the data while preventing overfitting.

Performance: The test error was 9.72%, with a sensitivity of 82.5% and specificity of 100%. This means that the model does not make any false positives, but there are 7 false negatives.

6. Random Forest

Random Forest is an ensemble method that combines predictions from multiple decision trees. It handles high-dimensional datasets effectively, capturing complex relationships while reducing the risk of overfitting compared to individual decision trees through majority voting. Additionally, it provides estimates of feature importance, making it valuable not only for prediction but also for understanding the key drivers of the results.

Parameters calibration: To optimize the Random Forest model, hyperparameters $mtry$ (number of variables selected at each split), $ntrees$ (number of trees), and $node\_size$ (minimum samples per split) were tuned using grid search. The performance was evaluated using the Out-Of-Bag (OOB) error, which estimates accuracy without a separate validation set. The OOB error is computed by making predictions for each observation using only the trees that did not include that observation in their training data (bootstrap sample). In this way, each observation is predicted by the trees that were trained on different subsets of the data, allowing for an unbiased estimate of the model's performance.

After the grid search, the optimal parameters identified based on the minimum OOB error were $mtry = 180$, $ntrees = 400$, and $node\_size = 10$, achieving an OOB error of approximately 10.6%.

Performance: For the Random Forest model, the test error was 8.33%, with a sensitivity of 87.5% and specificity of 96.88%. This means that the model correctly identifies 87.5% of the positive cases, while 12.5% of the positive cases are missed (false negatives). On the other hand, the model achieves a high specificity, correctly classifying 96.88% of the negative cases, with only 3.12% being false positives. Overall, the model demonstrates strong performance, being particularly effective at identifying true positives while maintaining a low rate of false positives.

7. Analysis of the Selected Genes

To study the most important genes influencing the presence or absence of cancer, we selected the top 20 genes deemed most relevant by each of the developed models.

Relevance was determined based on the absolute value of the associated coefficients for Lasso Logistic Regression, Elastic Net Logistic Regression, and both Lasso SVM models, while importance scores were used for the Random Forest. For each identified gene, we examined its associated biological function and analyzed the overlap between the gene sets selected by the five models.

From the top 20 genes selected by each model, the following were identified as common across most models:

- $lincRNA.chr9.34741650.34747975\_F$
- $lincRNA.chr12.76652565.76706444\_F$
- $lincRNA.chr2.102597068.102606493\_R$
- $NM\_001001971$
- $A\_33\_P3296313$
- $A\_33\_P3355694$

These shared genes likely play central roles in the classification task, as they were consistently selected across various model types. Their biological functions primarily relate to gene expression regulation (e.g., $lincRNAs$ and $NM\_001001971$) and cellular metabolism (e.g., $A\_33\_P3296313$ and $A\_33\_P3355694$), highlighting their significance in cancer-related processes.

In a broader perspective, we examined the biological functions associated with all the selected genes. Most of the genes relevant to classification were linked to *gene expression regulation*, *cellular proliferation*, and, to a lesser extent, *cellular metabolism*.

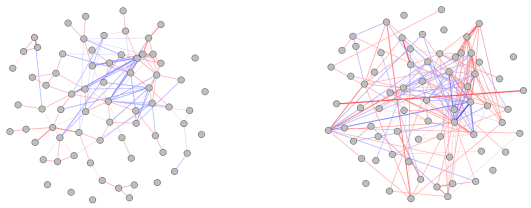
- Gene Expression Regulation: Controls the timing and levels of gene activity. Dysregulation can lead to tumor initiation and progression.
- Cellular Proliferation: Drives cell growth and division. Uncontrolled proliferation is a hallmark of cancer, leading to tumor formation.
- Cellular Metabolism: Provides energy and materials for cellular functions. Cancer cells often alter metabolism to support rapid growth and proliferation.

8. Graphical Lasso

As described in the dataset documentation, there is a significant difference in the relationships between breast tissue genes in normal patients compared to patients with cancer. To uncover and display these differences, we constructed and analyzed graphs that represent the relationships between these variables in both conditions. However, given the high dimensionality of the dataset, we restricted our analysis to the subset of 72 variables previously selected by Logistic Lasso (excluding the intercept). Since the Logistic Lasso model was trained using only the training set, we computed correlations and covariances exclusively on the training set for consistency, dividing it into two subsets: normal and cancer patients.

8.1. Direct Correlation Graphs

As an initial representation, we constructed two graphs to separately capture the relationships between variables for normal and cancer patients using the correlation values - calculated within their respective datasets - as edge weights. To refine the graphs, we removed auto-correlations and we applied a threshold of 0.3 to filter out weak correlations, ensuring that only meaningful relationships were represented. Additionally, we enhanced the visualization of these correlations by incorporating a heatmap, resulting in the graphs displayed below.



(a) Normal Correlations (b) Cancer Correlations

As evident in the figures, graph (a) is significantly sparser compared to graph (b). Additionally, in graph (a) positive correlations (red edges) are relatively rare, highlighting a more independent behavior among variables in normal tissues. Conversely, graph (b) shows a higher prevalence of positive correlations, along with an intensification of both positive and negative correlations. This observations suggest that in cancer conditions, the relationships between genes are much stronger and more interconnected than in normal conditions.

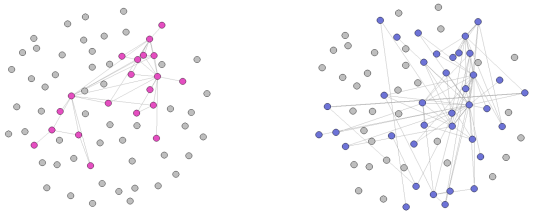
8.2. Graphical Lasso

While the correlation graphs provided a broad overview of gene interactions in normal and cancer patients, they do not distinguish between direct and indirect relationships. To gain deeper insights into the network and identify relevant direct dependencies, we employed Graphical Lasso - a method that estimates sparse precision matrices to focus on conditional dependencies between variables.

The method introduces a regularization parameter, λ, which controls graph sparsity: smaller λ values yield dense graphs, while larger values encourage sparsity by eliminating weaker dependencies.

We applied Graphical Lasso to the covariance matrices for normal and cancer datasets, exploring multiple λ values ($\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$). Based on visual inspection of

the resulting graphs, we selected $\lambda = 0.2$ as a reasonable estimate that balanced interpretability and relevance. At this λ, the graphs effectively highlighted key direct dependencies, revealing important patterns in gene interactions for normal and cancer conditions (see the figure below).



(a) Normal Graphical Lasso (b) Cancer Graphical Lasso

As illustrated in the figure, the Graphical Lasso networks for normal and cancer patients reveal remarkable differences in their structure and connectivity. In the normal network, (a), the graph is relatively sparse, with most nodes isolated or forming small, weakly connected clusters. The most important dependencies are concentrated in the center of the graph, with very few connections extending to the outer nodes. In contrast, the cancer network (b) is much denser, with strong connections not only in the center but also extending to the outer nodes, creating a highly interconnected structure. This suggests that in cancer conditions, gene interactions become more widespread and involve variables that are less significant in normal tissue. Additionally, the variables driving the key dependencies differ between the two networks, reflecting distinct biological processes at play in normal and cancer conditions.

9. Conclusions

9.1. Models Comparison

We conclude by comparing the performance of the analyzed models, evaluating their accuracy, sensitivity, and specificity.

In terms of test set accuracy, the Random Forest model outperforms the others, followed closely by the Logistic Regression with Elastic Net penalization and the SVM with Lasso penalty and Hinge Loss. The small differences in accuracy suggest that all models have competitive performance for this binary classification task.

Regarding sensitivity and specificity, Random Forest shows the highest sensitivity, indicating a better ability to identify positive cases. However, this comes with a slightly lower specificity (96.9%), while all other models achieve perfect specificity. This indicates that Random Forest tends to predict more false positives with respect to the other models.

The trade-off between sensitivity and specificity depends on the task's priorities. If minimizing false negatives is critical (e.g., identifying cancer patients), Random Forest is the best choice due to its higher sensitivity. On the other hand, if avoiding false positives is more important, linear models offer perfect specificity but at the cost of lower sensitivity.

Model	**Test Accuracy**	**Sensitivity**	**Specificity**
Logistic Reg. (Lasso)	0.889	0.8	1
SVM (Lasso, Sq. Hinge Loss)	0.889	0.8	1
SVM (Lasso, Hinge Loss)	0.903	0.825	1
Logistic Reg. (Elastic Net)	0.903	0.825	1
Random Forest	0.917	0.875	0.969

Table 1: Performance comparison between models: accuracy, sensitivity and specificity.

Overall, Random Forest emerges as the best model for balancing sensitivity and specificity, making it ideal for scenarios where both identifying cancer and minimizing misdiagnosis of healthy patients are important. Linear models may be considered when perfect specificity is prioritized, though at the expense of some sensitivity.

In the context of this study, aimed at identifying cancer in patients using gene expression data, maximizing sensitivity is crucial to avoid delays in treatment due to false negatives. However, this should not come at the cost of specificity, as false positives could lead to unnecessary anxiety and testing. Given these priorities, Random Forest is the most suitable choice, with its strong sensitivity and acceptable specificity. Linear models could also be considered if avoiding false positives is the highest priority.

9.2. Final Considerations

In conclusion, our analysis revealed significant differences in gene relationships between cancerous and normal tissues, with distinct variables driving the interactions in each condition. Cancer tissues exhibited stronger and more widespread correlations, while normal tissues showed sparser and more localized connections. Across all models, the most important variables were consistently linked to gene expression regulation, cellular proliferation, and cellular metabolism - critical processes in distinguishing between the two conditions. These findings offer a clearer perspective on the molecular patterns in cancer and normal tissues, highlighting relevant areas for further investigation.