Tissue Analysis for Breast Cancer Subtyping: Evaluating ML Models on Full, Reduced, and High-Variance Data

Authors: Yun Ma, Henry Wheeler, Ella Peplinski

Abstract

Breast cancer remains one of the most complex and heterogeneous diseases, with distinct subtypes requiring targeted diagnostic and treatment strategies. This study aims to classify breast tissue samples into six categories - five cancerous subtypes (HER, Basal-like, Luminal A, Luminal B, and Cell Line) and one normal tissue type - using supervised machine learning methods on high-dimensional gene expression data.

The dataset, sourced from Kaggle, contains 151 tissue samples characterized by 54,677 gene features, preprocessed to ensure consistency and reliability. Multiple machine learning models, including Support Vector Machines (SVM), Random Forest, and Gradient Boosting, were implemented across three data representations: the full dataset, a PCA-reduced dataset retaining 99% of the variance, and a high-variance dataset focusing on the most variable genes. Stratified K-Fold Cross-Validation guided model selection, while grid search optimization fine-tuned hyperparameters.

The SVM with an RBF kernel on the full dataset and the XGBoost on the PCA reduced dataset demonstrated the best performance, both achieving a 97% accuracy and high precision and recall for most cancer subtypes. However, they exhibited limitations in detecting normal tissues, underscoring the challenges of class imbalance. Conversely, the Random Forest model excelled at identifying normal tissue samples but struggled with cancer subtype differentiation. The PCA-reduced dataset offered robust performance while reducing computational complexity, highlighting the importance of dimensionality reduction for high-dimensional datasets.

This study illustrates the potential of machine learning for breast cancer subtype classification, providing insights into gene expression patterns and their role in cancer diagnosis. By comparing model performances across varying data representations, the findings pave the way for further exploration of interpretable, scalable solutions in precision oncology.

AI Disclaimer

Artificial Intelligence (AI) tools were utilized during the development of this project to assist with debugging but were not used to generate or write any code. Additionally, AI was employed to perform grammatical checks and improve the clarity of certain sections of this report. All substantive analysis, implementation, and insights are the result of original work.

Member Contribution Statement

All team members contributed equally to every aspect of this project, including coding, presentation development, and final report writing. All ideas and analyses are original, and collaboration was maintained consistently throughout the entire process.

1 Data Overview/Preprocessing

The dataset utilized in this study, sourced from Kaggle, comprises expression profiles from 151 distinct breast tissue samples, each characterized by 54,677 gene features. The data underwent preprocessing and normalization prior to analysis to ensure uniformity across all features, mitigating potential biases arising from variability in measurement scales or experimental conditions.

The samples represent six tissue categories: five cancerous subtypes (HER2-enriched, Basal-like, Cell Line, Luminal A, and Luminal B) and one normal (non-cancerous) tissue type (See Figure 1). Each feature corresponds to a specific gene, with the values representing the expression levels of that gene within a given tissue sample (See Figure 2).

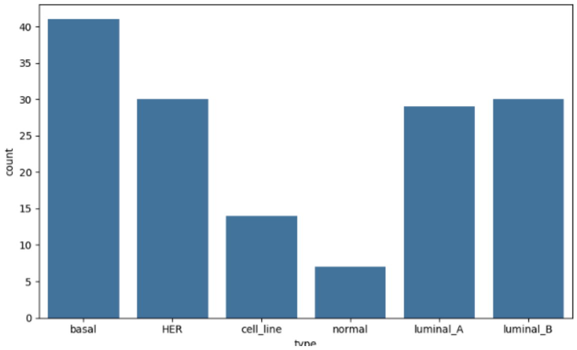
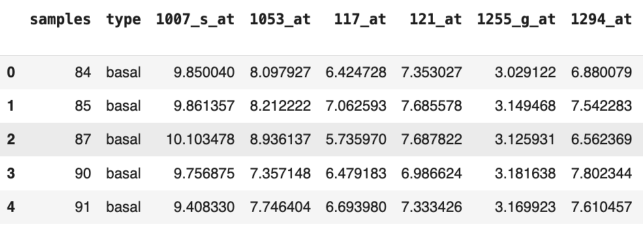


Figure 1. Distribution of Tissue Samples



	samples	type	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at
0	84	basal	9.850040	8.097927	6.424728	7.353027	3.029122	6.880079
1	85	basal	9.861357	8.212222	7.062593	7.685578	3.149468	7.542283
2	87	basal	10.103478	8.936137	5.735970	7.687822	3.125931	6.562369
3	90	basal	9.756875	7.357148	6.479183	6.986624	3.181638	7.802344
4	91	basal	9.408330	7.746404	6.693980	7.333426	3.169923	7.610457

Figure 2. Data Overview

This high-dimensional data presents an ideal challenge for supervised machine learning techniques. The primary objective of this study is to develop a model capable of analyzing gene expression patterns and accurately classifying samples according to their corresponding breast cancer subtype. Such classifications are critical for advancing our understanding of gene expression differences among breast cancer subtypes and could have implications for personalized medicine and treatment strategies.

2 Model Selection - Stratified K-Fold CV

To guide model selection, we conducted a Stratified K-Fold Cross-Validation (CV) using 5 folds. This method ensured that the class distribution in the target variable was maintained across all folds, providing a reliable estimate of model performance. For each model, the average cross-validation error was computed to compare their effectiveness. Logistic Regression achieved an average cross-validation error of 0.9696, while the Support Vector Machine (SVM) demonstrated the highest error at 0.9897. Decision Tree and Random Forest classifiers had average errors of 0.9494 and 0.9798, respectively. K-Nearest-Neighbors (KNN) showed the lowest performance with an error of 0.8135, and Gradient Boosting yielded an error of 0.9754.

Based on these results, we selected SVM, Random Forest, and Gradient Boosting for further development, as these models exhibited the highest cross-validation errors, suggesting strong potential for generalization when appropriately tuned. These models were implemented on three-different dataset variations to assess their robustness: the full dataset containing all original features, a PCA-reduced dataset that retained the most significant variance through dimensionality reduction, and a high-variance data set focused solely on predictors with they greatest variance. This approach allowed us to evaluate the adaptability and performance of each model under different data representations.

3 Overview of Methods/Approach

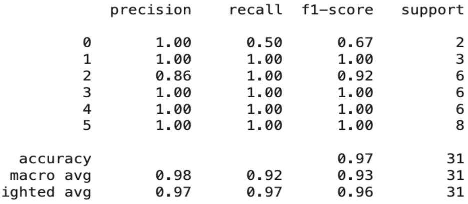
The rationale for running models on the full, PCA-reduced, and high-variance datasets lies in understanding how different data representations impact model performance and generalization. The full dataset serves as a baseline, encompassing all features without reduction, ensuring no potential information is lost. The PCA-reduced dataset emphasizes dimensionality reduction, retaining only the most significant components to mitigate noise and redundancy, which can improve efficiency and reduce the risk of overfitting. Meanwhile, the high-variance data set focuses on features with the greatest variability, under the assumption that such features are most likely to carry meaningful information. By evaluating models across these variations, we aim to identify the optimal data representation for robust and accurate classification.

3.1.1 SVM on Full Dataset

The first model we developed was a multi-class Support Vector Machine (SVM) using the full dataset. To encode the categorical tissue types into numerical values for compatibility with the SVM model, we applied ordinal encoding. The dataset was split into an 80/20 training and testing split, resulting in 120 tissue samples for training and 31 for testing. To address the class imbalance inherent in the dataset, we employed the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic samples for minority classes to ensure more balanced representation during training.
To optimize the SVM's performance, we conducted a grid search over a defined hyperparameter grid. This automated process systematically evaluated all possible combinations of the specified hyperparameters. The grid included kernel types (linear, rbf, and poly), penalty parameter values ("C") ranging from 1 to 5, and kernel coefficient values for the gamma parameter (scale and auto). For each combination, the grid search performed cross-validation, computing the average accuracy as the performance metric. The best-performing configuration was an SVM with an RBF kernel, a C value of 1, and gamma=scale.

The RBF (Radial Basis Function) kernel maps input features into a higher-dimensional space, enabling a linear decision boundary to effectively separate the classes. This capability makes it well-suited for non-linear classification problems, such as breast cancer subtype prediction, where complex relationships exist between features. The parameter C governs the tradeoff between maximizing the decision margin and minimizing classification error, while gamma=scale determines the influence of individual data points on the decision boundary. Together, these settings enabled the model to capture the intricate patterns within the dataset effectively.

After constructing the model, we evaluated its performance on the test dataset. The multi-class SVM demonstrated exceptional performance, achieving an overall accuracy of 97%. The model exhibited perfect precision and recall for the Cell Line, Luminal B, HER, and Basal subtypes, indicating its ability to accurately distinguish between these breast cancer subtypes (see Figure 3).



```
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         2
           1       1.00      1.00      1.00         3
           2       0.86      1.00      0.92         6
           3       1.00      1.00      1.00         6
           4       1.00      1.00      1.00         6
           5       1.00      1.00      1.00         8

    accuracy                           0.97        31
   macro avg       0.98      0.92      0.93        31
weighted avg       0.97      0.97      0.96        31
```

Figure 4. Classification Report on Multi-Class SVM (Full Dataset)

However, it is important to highlight the performance on the normal tissue class, which had only two instances in the test set. The SVM achieved a precision of 1.0 and a recall of 0.5, resulting in an F1-score of 67% for this class. This outcome suggests that while the SVM effectively distinguishes between cancer subtypes, it struggles to correctly identify all instances of normal tissue (see Figure 5). This limitation is critical as it implies the potential misclassification of healthy individuals as having cancer, which could lead to severe clinical consequences.

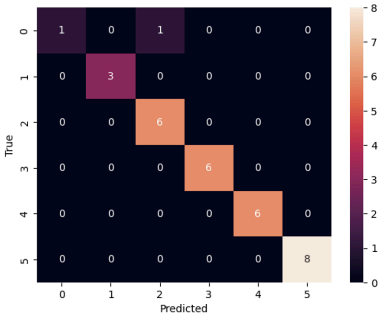


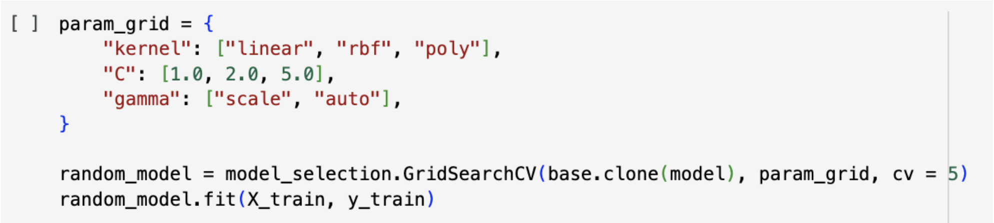
Figure 5. Confusion Matrix for Multi-Class SVM on Full Dataset

To further analyze the impact of preprocessing choices, such as ordinal encoding and oversampling with SMOTE, we proceeded to develop a Random Forest classifier. This model was built without encoding or resampling to evaluate its comparative performance and assess the influence of these preprocessing steps on classification outcomes.

3.1.2 RF on Full Dataset

Following the success of the Multi-Class SVM model on the full dataset, we developed a Random Forest classifier to evaluate its performance under different preprocessing conditions. Unlike the SVM model, this approach did not incorporate ordinal encoding or oversampling techniques such as SMOTE. This decision was made to assess the impact of using raw gene expression features alone, without the potential biases introduced by encoding or resampling. By removing these preprocessing steps, we sought to determine whether the high accuracy achieved by the SVM model was influenced by these modifications.

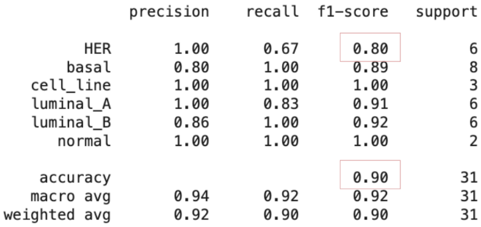
To optimize the Random Forest model, we performed a grid search over a range of hyperparameters (see Figure 6). The search explored various configurations for maximum tree depth, the minimum number of samples required at each leaf, the minimum number of samples needed to split a node, and the total number of trees in the forest. Cross-validation was utilized to evaluate each combination of hyperparameters and identify the configuration that yielded the highest average accuracy. The optimal parameters determined by the grid search included an unrestricted maximum tree depth, requiring a minimum of one sample per leaf and ten samples per split, and constructing a forest with 100 trees. Using these hyperparameters, we trained the Random Forest classifier, which aggregates predictions across multiple decision trees to minimize variance and improve overall performance.



```
[ ]  param_grid = {
         "kernel": ["linear", "rbf", "poly"],
         "C": [1.0, 2.0, 5.0],
         "gamma": ["scale", "auto"],
     }

     random_model = model_selection.GridSearchCV(base.clone(model), param_grid, cv = 5)
     random_model.fit(X_train, y_train)
```

Figure 6. Grid Search for RF (Full Dataset)

The Random Forest model achieved an overall accuracy of 90% (see Figure 7) , which, while still strong, is lower than the performance of the SVM. This difference may be attributed to several factors: the lack of class balancing through SMOTE, the exclusion of encoded features, or inherent differences in the models' ability to handle complex data. SVMs, with their ability to map data into higher-dimensional spaces, may better capture intricate relationships between gene expression levels and tissue types compared to Random Forests.



```
              precision    recall  f1-score   support

         HER       1.00      0.67      0.80         6
       basal       0.80      1.00      0.89         8
   cell_line       1.00      1.00      1.00         3
   luminal_A       1.00      0.83      0.91         6
   luminal_B       0.86      1.00      0.92         6
      normal       1.00      1.00      1.00         2

    accuracy                           0.90        31
   macro avg       0.94      0.92      0.92        31
weighted avg       0.92      0.90      0.90        31
```

Figure 7. Classification Report for Random Forest (Full Dataset)

Notably, the F1-scores for cancerous tissue samples were lower for the Random Forest model relative to the SVM. While the SVM achieved F1-scores close to or at 1.0 across all cancer subtypes, the Random Forest demonstrated greater difficulty distinguishing between these subtypes. This suggests that the Random Forest struggled with the nuanced variations in gene expression levels that differentiate cancer subtypes.

However, an encouraging result from the Random Forest model was its performance on normal tissue samples. The precision and recall for the normal tissue class were both 1.0, indicating perfect classification of healthy tissue. This marks a significant improvement over the SVM, which struggled with false negatives for the normal class, misclassifying healthy samples as cancerous. These results highlight the trade-offs between the models, where the Random Forest excels in distinguishing between cancerous and healthy tissues but underperforms in differentiating cancer subtypes compared to the SVM.

3.2 Dimensionality Reduction through PCA

Principal component analysis was used to reduce the dimensionality of the breast tissue sample data set. The data set, which originally included nearly 55,000 gene features, was reduced to 144 principal components that explained 99% of the variance in the data. We chose the number of principal components that represents 99% of the data variance because the breast cancer samples involve complex biological data. Using these principal components ensures that key patterns and relationships within the data are not lost. In breast cancer samples, small variations can reveal major distinctions in subtypes. Because PCA reduces the number of features and allows for major reduction in dimensionality, it can help prevent overfitting when using machine learning algorithms.

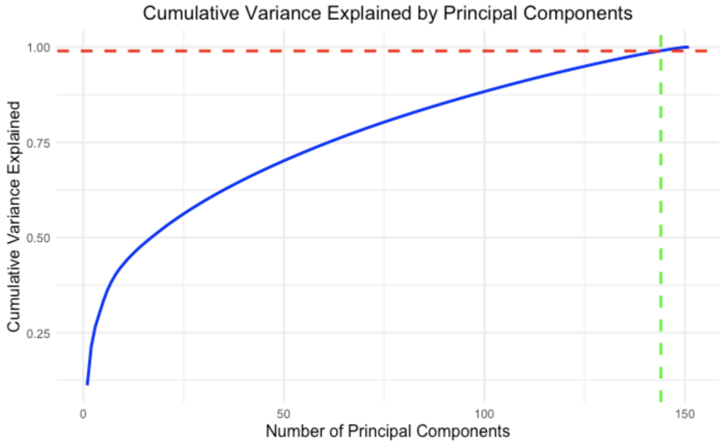


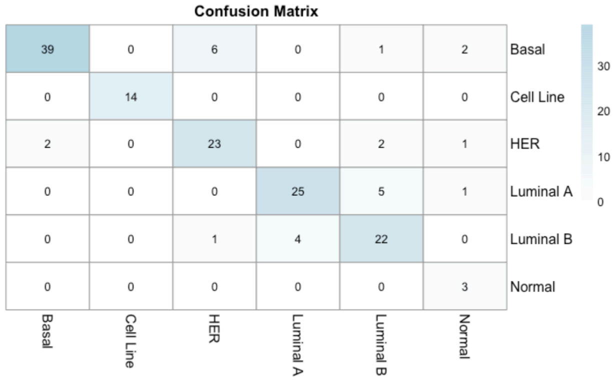
Figure 8. Cumulative Variance Explained by Principal Components

3.2.1 RF On PCA Reduced Dataset

After conducting the principal component analysis, we performed a random forest model on this PCA reduced data set. We wanted to see how this random forest model compared to the one performed on the full data set, given the large dimensionality of the data. Random forest was chosen here because of its ability to handle data with high dimensionality and its robustness to overfitting. This model was performed without resampling and oversampling. Trees were fully grown with no limit on the depth in order to capture complex patterns in the data. The *mtry* value, which determines the number of features considered at each split in the trees, is determined by cross validation. Grid search tests a range of *mtry* values, and cross validation evaluates the model's performance.

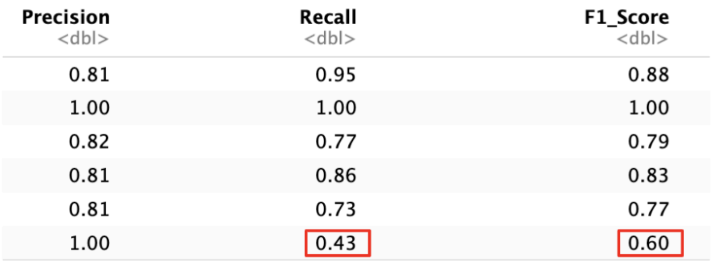
10-fold cross validation determined that the optimal number of features randomly selected at each split is 32. Using this *mtry* value, we performed 10-fold cross validation on the model and determined an overall cross validation accuracy of 83.44%. Additionally, we also tested the model with a 80% train and 20% test split, resulting in 82.14% accuracy. The 80/20 split allowed us to test the model on unseen data and ensure the cross validation accuracy was not due to overfitting. From the confusion matrix (Figure 9), we see that the model classifies the data fairly well overall, with some misclassifications between Luminal A and Luminal B subtypes. Also, a substantial number of Basal subtypes were wrongly classified as HER. The lower classification accuracy for this model is potentially because of class size imbalances, which could be resolved with a technique like SMOTE. Additionally, the lower performance may be because PCA discards the original feature structure and relationships, which random forest uses for splits.

The most glaring issue of this model is the low recall (0.43) for the Normal subtype (Figure 10). Recall is calculated as $\frac{True\ Positives}{True\ Positives + False\ Negatives}$, meaning it measures the proportion of actual positives correctly predicted by the model. The low recall for the Normal subtype indicates poor detection of true positives for this subtype. However, the high precision (1.00) suggests that when the model classifies a sample as Normal, it is always correct. Although we would like to see better detection of true positives for the Normal subtype, the model is being precise and not misclassifying cancer subtypes as Normal, which is critically important for this data. However, the low recall value for Normal subtypes still indicates that healthy individuals may be incorrectly diagnosed as having cancer, which is also problematic.



	Basal	Cell Line	HER	Luminal A	Luminal B	Normal
Basal	39	0	6	0	1	2
Cell Line	0	14	0	0	0	0
HER	2	0	23	0	2	1
Luminal A	0	0	0	25	5	1
Luminal B	0	0	1	4	22	0
Normal	0	0	0	0	0	3

Figure 9. Confusion Matrix of Subtypes from RF Model



Precision <dbl>	Recall <dbl>	F1_Score <dbl>
0.81	0.95	0.88
1.00	1.00	1.00
0.82	0.77	0.79
0.81	0.86	0.83
0.81	0.73	0.77
1.00	0.43	0.60

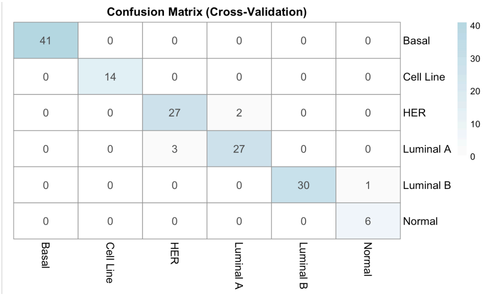
Figure 10: Precision, Recall, and F1-Score for Subtype Classification from RF Model

3.2.2 XGBoost On PCA Reduced Dataset

We also formed a model using XGBoost, which is a highly optimized implementation of gradient boosting. XGBoost performs well with high dimensional data and imbalance sample sizes. For these reasons, and because it performed well initially on the Stratified K-Fold Cross-Validation (CV), we chose to use this model.

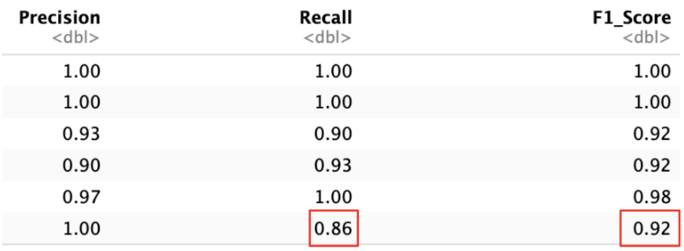
When using this model, 10-fold cross validation was used with the "multi:softmax" objective in order to classify each subtype. The learning rate (eta) of 0.1 and the maxim depth of 6 were used for a balance of model performance and complexity. The learning rate of 0.1 allowed for gradual updates to minimize overfitting, while the maximum depth of 6 ensured the model was able to capture the complexity of the data. Through 10-fold cross validation, the optimal number of boosting rounds was found to be 12. Early stopping determined that 12 rounds minimized error without overfitting the data.

When evaluating this model, we found a cross validation accuracy of 97.42%. This suggests that the XGBoost model generalizes well across folds and is robust in performance. Additionally, we also tested the model with a 80% train and 20% test split, resulting in 96.55% accuracy, further proving the model's ability to generalize to new, unseen data. The precision and recall values are consistently high across all subtypes (Figure 12). This demonstrates that the model performs well in detecting true positives and avoiding false positives. Also, the confusion matrix (Figure 11) shows nearly perfect classification along the diagonal, with a few misclassifications between Luminal A and HER subtypes. Compared to the previous random forest model on the PCA reduced data, we see that the Normal subtypes values for recall and F1-Score have much improved. Because all F1-Scores are above 0.9, this indicates that the model contains a balanced trade off between precision and recall.



Confusion Matrix (Cross-Validation)

	Basal	Cell Line	HER	Luminal A	Luminal B	Normal
Basal	41	0	0	0	0	0
Cell Line	0	14	0	0	0	0
HER	0	0	27	2	0	0
Luminal A	0	0	3	27	0	0
Luminal B	0	0	0	0	30	1
Normal	0	0	0	0	0	6



Precision <dbl>	Recall <dbl>	F1_Score <dbl>
1.00	1.00	1.00
1.00	1.00	1.00
0.93	0.90	0.92
0.90	0.93	0.92
0.97	1.00	0.98
1.00	0.86	0.92

Figure 11. Confusion Matrix of Subtypes from XGBoost Model

Figure 12. Precision, Recall, and F1-Score for Subtype Classification from XGBoost Model

3.3 Variable Selection Based on Variance

The rationale for selecting variables with the highest variance is based on the assumption that variation in gene expression differentiates normal cells from cancerous ones. For instance, healthy cells are expected to exhibit distinct gene expression patterns compared to cancerous cells, where these differences are reflected in genes with high variance (Figure 13). This approach is comparable to PCA, as both methods focus on maximizing variance within the dataset. However, unlike PCA, this method evaluates each gene independently, thereby improving the interpretability of the results. By applying this method, we aim not only to train a model with high predictive accuracy for various cancer types, but also to identify important genes that serve as signatures for specific tissue types.

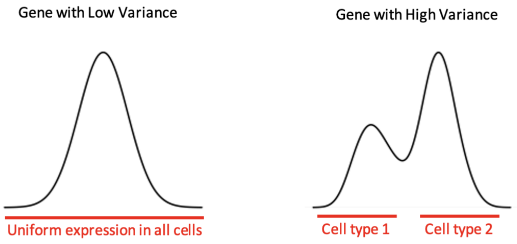
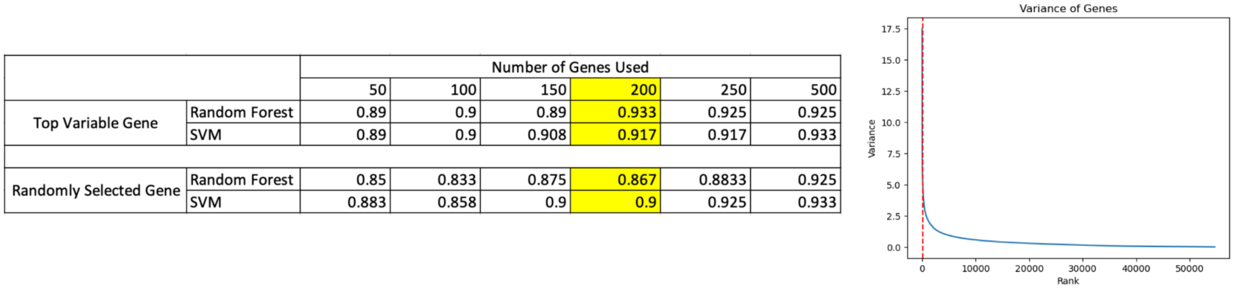


Figure 13. Variance in Gene Expression Drive Different Cell Labeling

Notably, a key limitation of this method compared to PCA is that it does not guarantee orthogonality in the variable space. As a result, the model is more prone to overfitting and multicollinearity issues. To address this, we conducted a kth-fold CV to determine the optimal number of genes to include in our models. To further assess the efficiency of our selection method and prevent overfitting, we randomly selected the same number of predictors for each group as a control and recorded their corresponding grid search CV scores. The expectation is that when overfitting occurs, both the random group and the high-variance group will exhibit similarly high performance in CV. Our goal is to identify predictor set sizes that maximize CV scores while maintaining a clear distinction between the performance of the high-variance group and the random group. This ensures that our modeling results are meaningful and not a result of overfitting. Based on this method, we determined the optimal variable size as the top 200 variably expressed genes (Figure 14).



		Number of Genes Used					
		50	100	150	200	250	500
Top Variable Gene	Random Forest	0.89	0.9	0.89	0.933	0.925	0.925
	SVM	0.89	0.9	0.908	0.917	0.917	0.933
Randomly Selected Gene	Random Forest	0.85	0.833	0.875	0.867	0.8833	0.925
	SVM	0.883	0.858	0.9	0.9	0.925	0.933

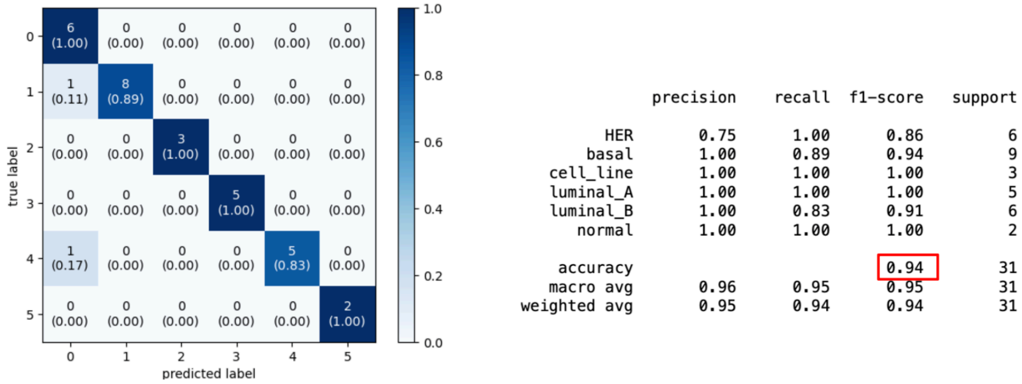
Figure 14. CV Selection of 200 Top Variable Genes

3.3.1 SVM on High-Variance Genes

After reducing our feature space to the 200 genes with the highest variance, we trained an SVM model and compared its performance to the SVM trained on the full dataset (Section 3.1.1). A grid search with CV was used for parameter tuning, without resampling the data. Model selection identified an SVM with a radial basis function kernel and a margin penalty parameter c=1, indicating a relatively large margin for the separation hyperplane.

The SVM model trained on the reduced feature set achieved an overall accuracy of 94% on test data, which is slightly lower than the accuracy of the SVM trained on the full dataset. However, the recall and precision for the normal class were both 1, indicating that the SVM trained on the high-variance data set performed better at distinguishing normal tissue from cancerous tissue compared to the model trained on the full dataset.

On the other hand, the SVM trained on the high-variance data set struggled to distinguish between certain cancer subtypes, particularly laminal_b and basal from HER cancer types (Figure 15). This limitation could have serious clinical implications, as different cancer subtypes may require distinct treatment approaches.



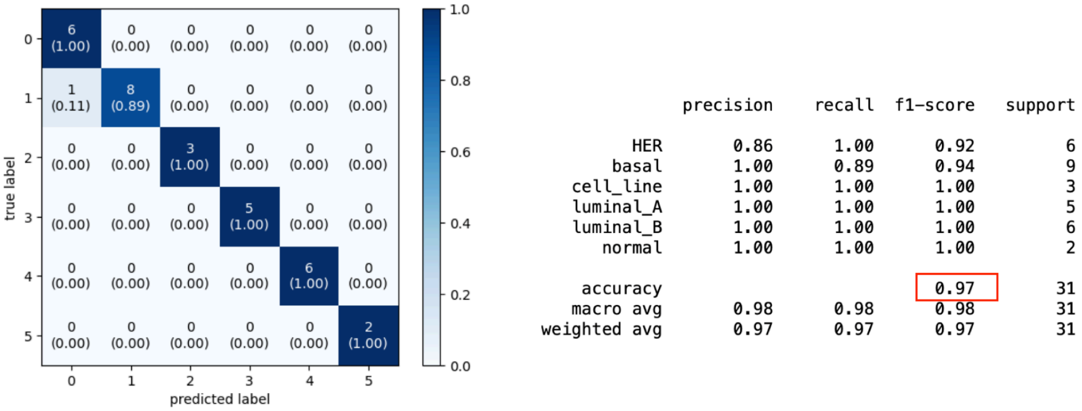
```
              precision    recall  f1-score   support

         HER       0.75      1.00      0.86         6
       basal       1.00      0.89      0.94         9
   cell_line       1.00      1.00      1.00         3
   luminal_A       1.00      1.00      1.00         5
   luminal_B       1.00      0.83      0.91         6
      normal       1.00      1.00      1.00         2

    accuracy                           0.94        31
   macro avg       0.96      0.95      0.95        31
weighted avg       0.95      0.94      0.94        31
```

Figure 15. SVM on Test Data

3.3.2 RF on High-Variance Genes

Building on our SVM results, we sought to improve prediction accuracy using a Random Forest model. Random Forest offers the advantage of greater interpretability compared to SVM and is robust in handling potential multicollinearity issues, which is important since the features selected by our method are not orthogonal. As with the SVM model, we tuned the Random Forest model using a grid search with cross-validation (CV), without resampling the data. Model selection identified a Random Forest configuration with no limitation on tree depth, a minimum of one sample per leaf, a minimum of two samples per split, and 300 trees. This model achieved an overall prediction accuracy of 97% on the test data, representing the highest accuracy among all Random Forest models we trained (Figure 16).

Compared to the SVM model in 3.3.1, Random Forest successfully distinguished basal from HER cancer types. However, analysis of the confusion matrix revealed that Random Forest, like SVM, struggled to distinguish between luminal B and HER cancer types (Figure 16). Upon further investigation, we found that the misclassification between luminal B and HER cancer types in both the SVM and Random Forest models involved the same instance. Interestingly, this misclassification was not observed in an XGBoost model trained on features derived from PCA reduction (Section 3.2), suggesting that our variable selection method may miss critical features to classify certain edge cases.



	0	1	2	3	4	5
0	6 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
1	1 (0.11)	8 (0.89)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
2	0 (0.00)	0 (0.00)	3 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
3	0 (0.00)	0 (0.00)	0 (0.00)	5 (1.00)	0 (0.00)	0 (0.00)
4	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	6 (1.00)	0 (0.00)
5	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	2 (1.00)

```
              precision    recall  f1-score   support

         HER       0.86      1.00      0.92         6
       basal       1.00      0.89      0.94         9
   cell_line       1.00      1.00      1.00         3
   luminal_A       1.00      1.00      1.00         5
   luminal_B       1.00      1.00      1.00         6
      normal       1.00      1.00      1.00         2

    accuracy                           0.97        31
   macro avg       0.98      0.98      0.98        31
weighted avg       0.97      0.97      0.97        31
```

Figure 16. Random Forest on Test Data

Another advantage of Random Forest is its ability to identify important features. By cross-validating the features deemed important by the Random Forest model with existing biological knowledge, we can evaluate whether the predictors used in our model are biologically meaningful (Figure 17).

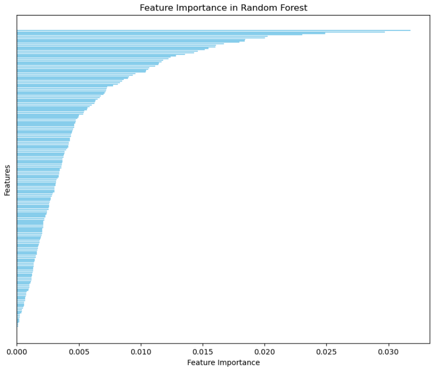


Figure 17. Random Forest Feature Importance

4 Discussion and Future Directions

Summary and Comparison

In this study, we employed multiple machine learning techniques to classify six tissue types related to breast cancer (one healthy control and five cancer subtypes) using 151 tissue samples and 54,677 gene features. We assessed model accuracy across various approaches, including SVM, Random Forest, and XGBoost, and evaluated the impact of different dimensionality reduction methods: no reduction, PCA, and high-variance gene selection. Among these methods, the full dataset (no reduction) produced the best results for SVM, achieving an accuracy of 97%. PCA reduction yielded strong performance with XGBoost, achieving an accuracy of 96.55%, while high-variance gene selection resulted in the best Random Forest performance, also achieving 97% accuracy.

Overall, the full dataset demonstrated superior ability in distinguishing between cancer subtypes, while dimensionality reduction methods were more effective at differentiating between cancerous and healthy tissues. Specifically, PCA and high-variance gene selection methods are better at distinguishing cancer from healthy tissues but struggle in differentiating closely related subtypes, such as luminal A and B, and HER cancer types.

Limitations

Our dataset was pre-scaled and normalized, which likely contributed to the high performance of the models. Future studies should explore how these models perform on real-world datasets processed similarly to validate their generalizability. Additionally, while our high-variance gene selection method proved effective, it could be refined further. For instance, cross-validation comparisons between the top n variable genes and n randomly selected genes could benefit from a finer grid search of n. Furthermore, developing standardized metrics to assess the difference in CV scores between randomized and high-variance variable sets would help enhance the rigor of the evaluation process.

Future Directions

Although Random Forest provides insights into which genes are important for classification, it does not pinpoint the specific genes that signify each cluster. Identifying subsets of genes which over- or under-expression distinguishes cancerous cells from healthy ones, or differentiates between cancer subtypes, could be a direction of future study. Addressing this issue has two major benefits: first, validating the features identified in this study against existing biological knowledge would strengthen the reliability of our approach. Second, discovering genes associated with specific cancer subtypes could provide valuable insights for future clinical research and therapeutic strategies.

Unsupervised learning techniques offer a promising pathway for this type of analysis. For example, performing PCA on the dataset and isolating PCs that differentiate clusters could reveal key patterns. While PCs are not directly interpretable, examining the loading coefficients of significant PCs can identify the genes contributing most to these separations. This approach has the potential to deepen our understanding of the biological mechanisms underlying cancer development.

Bibliography

B. Grisci, “Breast Cancer Gene Expression (CUMIDA),” *Kaggle*, 2020. [Online]. Available: https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida.[Accessed: 03-Dec-2024].

M. H. Rahman, J.M. Franco Cortés, and M. Ghosh, “Applications of Support Vector Machines (SVM) Learning in Cancer Genomics,” PLOS ONE, vol. 13, no. 2, e0191901, Feb. 2018. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5822181/. [Accessed: 03-Dec-2024]

P. Meesala, “Breast Cancer Classification 1.0.” *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/code/prasadmeesala/breast-cancer-classification-1-0/notebook. [Accessed: 03-Dec-2024].