Decoding Breast Cancer Subtypes

Machine Learning Insights into Glycoenzyme Expression Patterns

Cyrus Chun Hong Au Yeung
Department of Biochemistry & Molecular Medicine, George Washington University School of Medicine and Health Sciences

ABSTRACT

In this study, we evaluated the performance of glycosyltransferase (**GT**) gene expression levels in predicting breast cancer subtypes using PAM50 classification labels. Gene expression data comprised of 953 participants and 182 GT predictor genes. We employed nine supervised classification models, using an 80/20 train-test split, with 10-fold cross-validation applied during training. Model selection was based on the kappa statistic to ensure robust performance evaluation. The Support Vector Machine (**SVM**) with a polynomial kernel performed best, with a kappa value of 0.8555 on the test set. Notably, genes involved in N-glycosylation and sialylation processes were identified as key contributors to subtype classification. The findings highlight the predictive potential of GT gene expression in breast cancer subtyping. Future research should focus on exploring the biological mechanisms linking glycosylation pathways to breast cancer progression.

INTRODUCTION

Altered glycosylation is a hallmark of cancer, profoundly impacting tumor progression and metastasis. Glycosylation is a fundamental post-translational modification where sugars are enzymatically attached to proteins and lipids, influencing their structure and function. Glycosyltransferases (**GTs**) are a diverse family of enzymes which catalyzes the transfer of sugar moieties to acceptor molecules, playing crucial roles in biological processes, including cell-cell communication, immune response, and protein folding.

In cancer, aberrant glycosylation can result from changes in the expression levels of GT genes or from genetic mutations like single nucleotide variants (SNVs). Such alterations may lead to modified enzyme activity and the production of atypical glycan structures on cell surfaces and secreted proteins, thereby contributing to oncogenesis and tumor progression.

PAM50 categorize breast cancer into 5 intrinsic subtypes:

Subtype	Description
Luminal A	Low proliferation; best prognosis and responds well to hormonal therapy.
Luminal B	High proliferation; more aggressive than Luminal A.
HER2-enriched	High aggressiveness; but responds to HER2-targeted therapy.
Basal-like	Triple-negative; highly aggressive; poor prognosis.
Normal-like	Similar to normal tissue; less aggressive with more favorable outcomes.

Objective

This study aims to explore the linkages between GT expression altercations and breast invasive carcinoma subtypes. Predictive machine learning models are developed to classify subtypes, and we will identify significant genes as potential biomarkers for prognosis and monitoring.

DATA DESCRIPTION

TCGA Cohort Study

Cancer data was from a cohort study under the TCGA Breast Cancer (BRCA) project. This dataset includes RNA-sequencing data stored at the TCGA Data Coordination Center (DCC) and is publicly accessible through the UCSC Xena platform. Additional patient metadata curated from TCGA's Pan-Cancer Clinical Data Resource are available in the Xena repository.

Data Processing

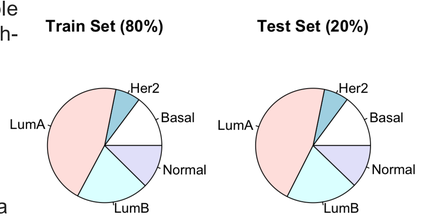
A list of 209 GT genes were mapped to the normalized counts data (n = 1,219, p = 20,530); 182 were successfully mapped. There were no missing values in the counts data. The counts data and metadata were inner joined, and patients without PAM50 annotations (n = 266) were removed.

METHODS

Machine Learning Model Selection

We employed a combination of baseline and ensemble machine learning models specified for handling high-dimensional datasets to classify breast cancer subtypes.

1. **Logistic Regression** (Multinomial) – Baseline model
2. **K-Nearest Neighbors** (kNN) – Non-parametric
3. **Decision Tree** (CART) – Highly interpretable
4. **Random Forest** – Reduces overfitting, robust to counts
5. **Support Vector Machines** (SVM) - High-dimensional data
6. **Boosted Trees** (GBM) – Focuses on wrong predictions
7. **XGBoost** – Regularized GBM, handles class imbalance



Data Analysis Using R (version 4.4.1)

Data Splitting

- The TCGA-BRCA data was split into 80% training and 20% testing sets. The training set was used to train and finetune the models, while the testing set was reserved for evaluating model performances.

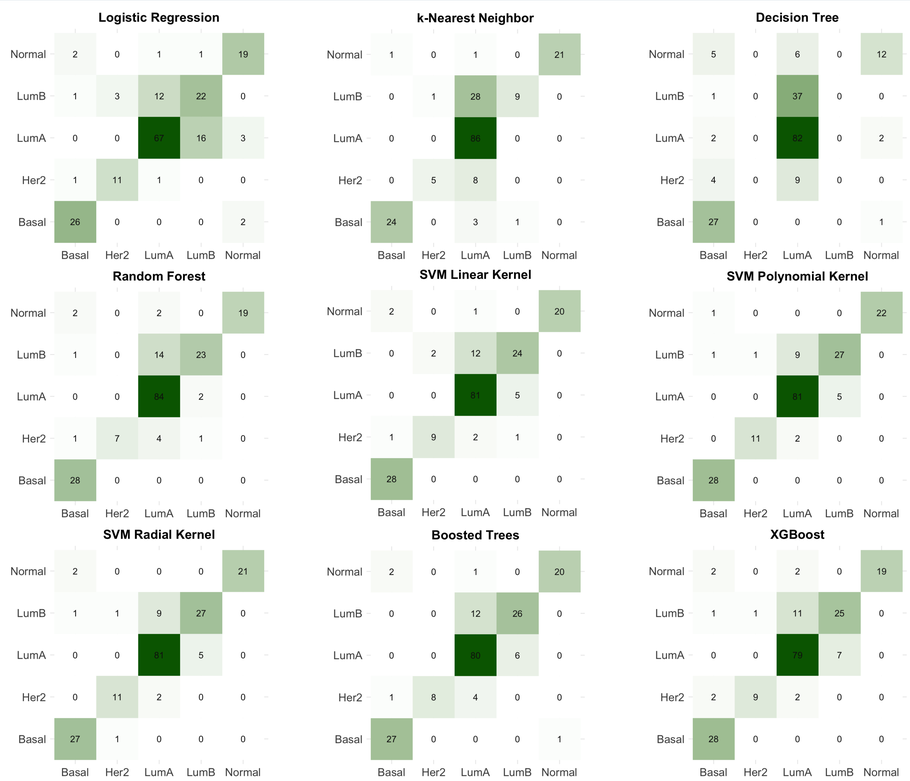
Cross-Validation (CV)

- 10-fold CV was applied to the training set using the `train()` function from the **caret** package
- CV was repeated for all models to identify optimal tuning parameters (i.e. `k` for kNN, `cp` for decision tree, `C`, `degree`, `scale` for SVM, `mtry` for Random Forest, and `n.trees`, `interaction.depth`, `shrinkrate`, `n.minobsinnode` for Boosted Trees)

Assessment of Model Performance

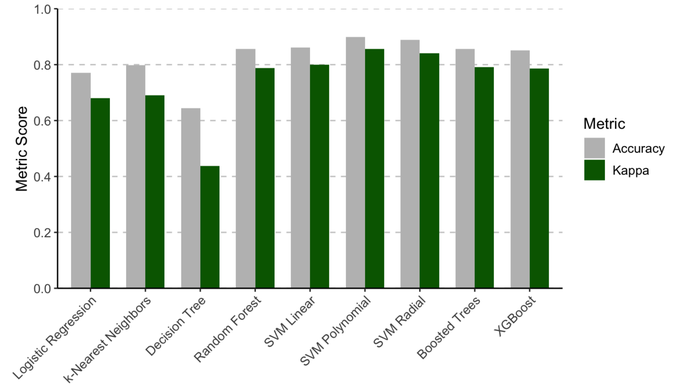
- **Confusion matrices** were generated for each model to evaluate class-level predictions.
- Overall performances were assessed using (1) accuracy and (2) **Kappa Statistic** which was selected as the primary metric for comparisons due to class imbalance, as it adjusts for chance agreement.
- Models were tuned to maximize Kappa to ensure balanced performance across all subtypes, especially for the under-represented HER2-enriched and Normal-like subtypes

RESULTS – CONFUSION MATRICES



RESULTS – PERFORMANCE METRICS

The SVM models performed best, with a 2nd degree polynomial kernel it achieved an accuracy of 0.8989 and Kappa of 0.8555. Random Forest achieved an accuracy of 0.8564 and Kappa of 0.7884.



Predictor variable importance was interpreted from the Random Forest model:

DISCUSSIONS

Best model performance agree with the hypothesis that specific GTs may drive breast cancer subtype-specific behaviors by modifying glycan structures on key cell-surface proteins.

Performance of Models

- **SVM** models, particularly with polynomial and radial kernels achieved the most accurate classifications likely due to their ability to capture non-linear relationships and interactions between GT genes
- **Random Forest** performed well as its ensemble approach effectively reduced overfitting

Top 3 Genes

- **UGT8**: Associated with ceramide glycosylation. Known indicator of tumor aggressiveness and is correlated with the Basal-like subtype
- **B3GNT5**: Involved in GlcNAc addition. Crucial for N-glycosylation
- **ST8SIA1**: Mediates sialylation, which is a prognostic biomarker for cancer progression and metastasis

Functional Analysis of Top 30 Genes by Variable Importance

- **GlcNAc** (N-Acetylglucosamine) - 7 genes
  Indicates altered N-glycosylation, affecting cell signaling
- **Neu5Ac** (Sialic Acid) - 6 genes
- Indicates enhanced sialylation, linked to immune evasion and metastasis

These changes suggest a remodeling of glycan structures, particular in N-glycosylation that is specific to breast cancer subtype.

Video Link: (Manually removed)