Breast Cancer Vital Status Classification Using Clinical and Genomic Data

DIN Sokheng
RA Veasna

December 15, 2025

1 Introduction

This project analyzes a breast cancer dataset of 1230 patients, integrating clinical variables with high-dimensional mRNA gene expression. The general task is a **binary classification problem**, where the classify is to predict a patient's `vital_status` (`Alive` or `Dead`). The dataset mirrors the structure of the TCGA-BRCA cohort and follows the preprocessing steps described in the associated publication, including log-transformation and variance filtering of gene expression.

1.1 Clinical and Genomic Variables

The clinical variables include demographic factors (age at diagnosis, age at index, initial weight), tumor characteristics (AJCC T stage, morphology, tissue of origin, primary diagnosis), treatment information (prior treatment, disease response), and population descriptors (race, gender, ethnicity). These variables encode patient-level heterogeneity typically used in clinical prognosis.

The genomic component consists of 5000 high-variance mRNA transcripts capturing biological pathways such as immune infiltration, stromal activation, lipid metabolism, and cell proliferation—mechanisms frequently associated with breast cancer progression.

1.2 Exploratory Associations with Vital Status

The target variable is `vital_status`. Exploratory analysis reveals that age-related variables show moderate but statistically significant differences between Alive and Dead groups, while most categorical clinical variables exhibit weak associations. In contrast, differential expression analysis uncovers strong transcriptomic shifts: genes such as **APOB**, **LYVE1**, **LINC01497**, and **AC104211.1** are strongly upregulated in deceased patients, aligning with known biological mechanisms including stromal remodeling, immune suppression, and altered lipid metabolism. Several transcripts display **bimodal expression**, suggesting the presence of molecular subtypes.

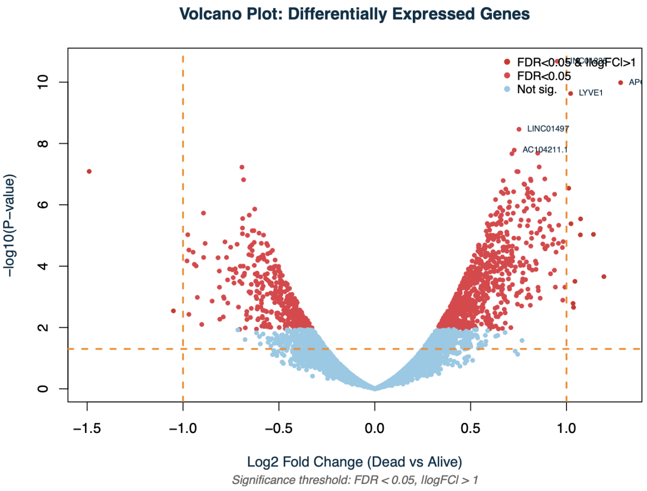

Figure 1: Volcano plot of differential expression (Dead vs Alive). Significantly dysregulated transcripts (FDR $<$ 0.05) show strong fold changes, including APOB, LYVE1, LINC01497 and AC104211.1. This confirms the presence of localized molecular signals relevant for prognosis.

Principal component analysis (PCA) of the most discriminant genes reveals two clear expression-based clusters. However, these clusters do not correspond to survival labels, implying that mortality differences arise from localized gene-expression changes rather than global transcriptomic structure.

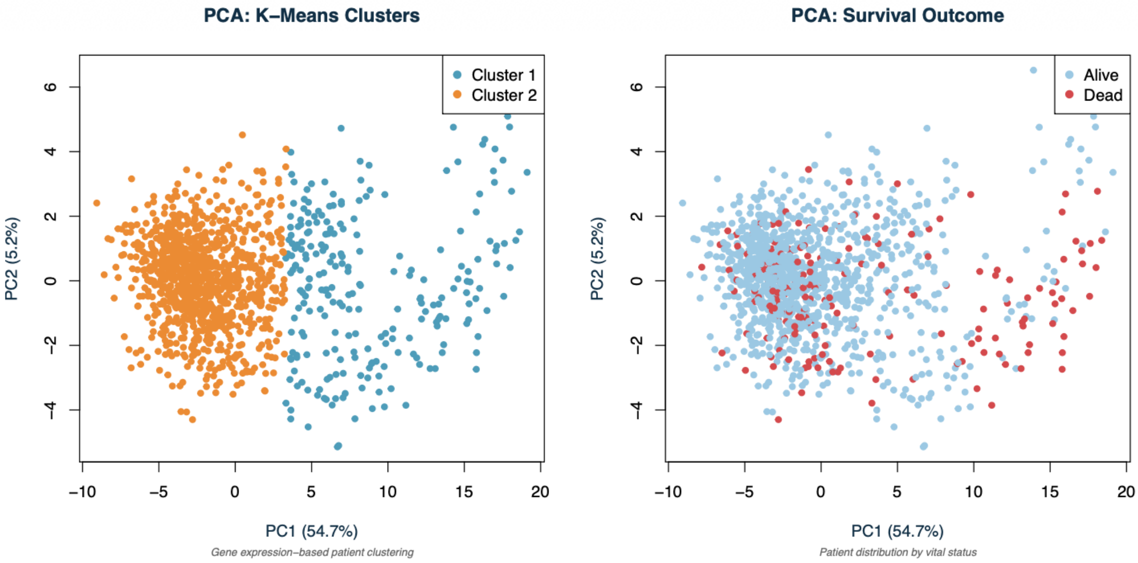

Figure 2: PCA of the top differentially expressed genes. Two expression-based clusters appear clearly, although they are not aligned with survival groups. This illustrates that global transcriptomic variation does not directly separate Alive and Dead patients.

1.3 Global Data Structure and Modeling Implications

The dataset presents substantial class imbalance (`Alive:Dead` $\approx$ 5:1), strong correlations among clinical variables (notably among age-related measures), and structured co-expression blocks among genomic features. These characteristics motivate the use of dimensionality reduction and penalized regression to prevent overfitting.

Overall, the exploratory analysis highlights the biological and statistical complexity of the data, justifying the modeling strategies developed in the following sections.

2 Methodology

Our objective is to select the most appropriate classification model for classifying patient vital status using both clinical variables and high-dimensional gene expression. The methodology follows a structured progression: establishing a baseline, screening genomic features, applying penalized regression, correcting class imbalance, and evaluating models under unified criteria.

2.1 Baseline Model

A logistic regression model using with clinical predictors provides an interpretable reference and allows us to quantify the added value of genomic features and small gene expression feature with **TOP-20**, **TOP-50**, **TOP-100**.

2.2 Gene Screening

Because the dataset contains 5000 transcripts with $p \gg n$, we first rank genes using univariate differential-expression statistics (`limma`). From this ranking we construct nested subsets, TOP-20, TOP-50, TOP-100, TOP-500, TOP-1000, TOP-5000 which enable controlled comparisons across different dimensionalities.

2.3 Penalized Regression Models

To address multicollinearity and prevent overfitting, we fit three classical penalized models:

- **Ridge Regression** (ℓ_2 penalty): stabilizes estimates when features are highly correlated.
- **Lasso Regression** (ℓ_1 penalty): performs embedded feature selection and yields sparse solutions.
- **Elastic Net** (mixed ℓ_1/ℓ_2): handles correlated gene blocks and serves as our extended method.

2.4 Lasso Extensions

We evaluate two lasso refinements to improve stability and interpretability in high-dimensional, correlated genomic data.

- **Adaptive Lasso** (Zou, 2006): applies coefficient-specific weights from a ridge fit, reducing bias and improving variable selection under feature correlation.
- **UniLasso** (Chatterjee, Hastie & Tibshirani, 2025): a two-stage sparse method that stabilizes selection and preserves coefficient signs when signals are weak and imbalanced.

2.5 Handling Class Imbalance

To address the strong Alive:Dead class imbalance (5:1), we apply SMOTE oversampling on the training set. We use standard SMOTE with $k = 5$ nearest neighbours and oversample the minority class from 141 to 705 observations (Alive:Dead $\approx$ 1:1), while keeping the test set untouched. This is expected to improve the models' ability to detect deceased patients.

2.6 Model Evaluation

Models are compared using AUC, accuracy, precision, recall, specificity, F1-score, sparsity, and the train–test AUC gap to assess overfitting. The final selection is based on performance, stability across feature subsets, interpretability, and biological plausibility of selected genes.

3 Results

3.1 Model Comparison

Six models were evaluated across multiple feature sets, ranging from Clinical-only predictors to Clinical+TOP5000 genes. Table 1 reports the best-performing configuration for each model.

Table 1: Model performance comparison (best feature set per model)

Model	SMOTE	Features	F1	Recall	Precision	AUC
UniLasso	Yes	TOP20	**0.737**	0.700	0.778	0.893
Adaptive Lasso	Yes	TOP20	0.717	0.825	0.635	0.908
ElasticNet	Yes	TOP20	0.717	0.825	0.635	0.905
Lasso	Yes	TOP20	0.696	0.800	0.615	0.905
Logistic	Yes	TOP20	0.711	0.800	0.640	0.902
Ridge	Yes	TOP20	0.653	0.825	0.541	0.896

For imbalanced classification, the F1-score provides the most informative summary of model performance, as it balances precision and recall on the minority class. After applying SMOTE, the **UniLasso + SMOTE** model with **Clinical+TOP20 genes** achieves the highest F1-score (0.737), indicating the best precision–recall trade-off. Notably, all top-performing models rely on Clinical+TOP20 genes, while larger gene sets (TOP100–TOP5000) consistently degrade test performance, indicating overfitting in higher-dimensional settings.

3.2 Effect of SMOTE

SMOTE substantially improves minority-class detection.

Table 2: Average performance: SMOTE vs No SMOTE

Method	**AUC**	**Recall**	**F1-score**
No SMOTE	0.883	0.450	0.573
SMOTE	0.902	0.796	0.705

Without SMOTE, models achieve high accuracy but poor recall, effectively failing to identify deceased patients. SMOTE increases recall to approximately 80% with a moderate reduction in precision.

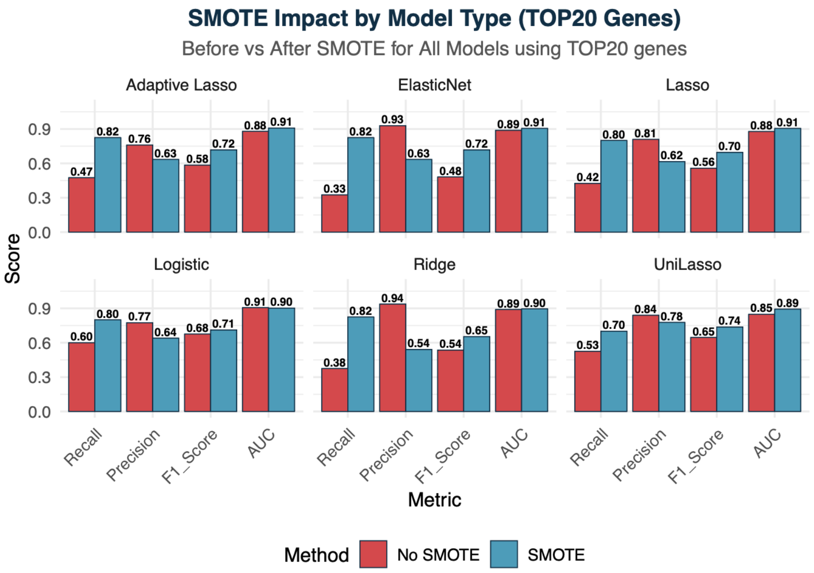

Figure 3: Impact of SMOTE on model performance using Clinical+TOP20 genes. Across all models, SMOTE substantially improves recall and F1-score, while AUC remains stable, indicating better minority-class detection without loss of discrimination.

4 Conclusion

This study addresses breast cancer vital status classification using clinical variables and high-dimensional mRNA expression. Penalized logistic models with feature screening consistently outperformed unpenalized baselines, confirming the need for regularization in $p \gg n$ settings. The **UniLasso + SMOTE** model using **Clinical+TOP20 genes** achieved the best precision–recall balance (F1 = 0.737, AUC = 0.893), providing stable minority-class detection and sparse, interpretable solutions. Clinical variables dominated performance, while genomic features added modest complementary signal; expanding beyond the TOP20 genes consistently degraded test performance due to overfitting, highlighting the importance of sparsity, imbalance handling, and interpretability in clinical classification.