Predicting Breast Cancer Patient Survival Using Gene Expression

The objective of this research was to construct and test machine learning models for the prediction of breast cancer patient outcomes from gene expression data in the GSE20685 dataset. The main clinical outcome of interest was patient mortality, with **327 samples** and **54,627 genes** in the raw data.

- There were no missing values in gene expression data, providing complete feature matrices.
- Low variance genes (less than a threshold of 0.01) were eliminated, shrinking the feature space from **54,627 to 54,616 genes.**
- The **1,000 most variable genes** with highest statistical significance were chosen, further narrowing the dataset to a size that was computationally manageable without removing informative features.
- Gene expression data was normalized to provide scale consistency for features.
- Three machine learning models were developed and tested:
  - Logistic Regression
  - Random Forest
  - Support Vector Machine (SVM)
- Logistic Regression was the top-performing model, recording a mean **AUC of 0.949** across cross-validation folds, well surpassing **Random Forest (mean AUC: 0.822)** and **SVM (mean AUC: 0.874).**
- Logistic Regression's confusion matrix revealed excellent accuracy at **49 true negatives** and **14 true positives,** reflecting strong predictive ability for each class.
- ROC curves verified Logistic Regression performed better in discriminating between deceased and surviving patients.

The findings indicate that gene expression profiles are able to successfully predict the outcomes of breast cancer patients, especially mortality. Robust performance was shown by Logistic Regression, which is likely a result of its capacity to process high-dimensional data effectively while retaining interpretability. Such a model may be of great use in risk stratification in a clinical environment, facilitating personalized treatment planning and resource allocation.

This work effectively uncovered a predictive model for breast cancer outcome based on gene expression data. Logistic Regression yielded the best predictive accuracy, holding potential as a future area of study and clinical investigation. Further exploration of the leading predictive genes can uncover new biomarkers or pathways of disease progression.

Data Preprocessing and Acquisition

This phase outlines the acquisition, exploration, cleaning, and preprocessing of the GSE20685 dataset for further machine learning analysis.

Dataset Description

The dataset GSE20685 was downloaded from the NCBI Gene Expression Omnibus (GEO). The dataset consists of gene expression profiles along with linked clinical metadata for 327 samples of breast cancer patients. The main components are:

Expression Data: A 54,627 x 327 matrix.

Rows are single Affymetrix probe set IDs (genes), and columns are patient samples (GSM accessions). Pre-processing was applied to the expression values **(log2 transformed, quantile normalized)** according to the description of data processing.

Clinical Metadata: A 61 x 327 table.

Rows are unique clinical and sample-specific variables for the 327 patients. Column entries correspond to patient demographics, tumor features, treatments, follow-up time, and clinical events.

The goal of the current research was to make a prediction of patient prognosis from gene expression data. The major clinical endpoint selected for this binary classification problem was overall survival and was represented by the *characteristics_ch1.3.event_death variable*. The variable records if a patient died within the observation period (coded as '1') or was censored (coded as '0'). The dataset was imbalanced, containing 83 death events (25.4%) and 244 censored events (74.6%).

Data Preprocessing Steps

Prior to modeling, a number of preprocessing steps were taken to validate data quality and suitability with machine learning algorithms.

Sample Matching and Alignment

The initial step was to verify that the expression data and the metadata referred to the very same set of patient samples. This was done by matching the sample IDs (GSM accessions) in the columns of the expression matrix with the row indices of the metadata dataframe. Samples that existed in both the sources were kept for analysis, ensuring a one-to-one match across the 327 samples.

Dealing with Missing Values:

- Expression Data:
    - A careful inspection showed that the gene expression matrix had no missing values **(Total missing values: 0)**. This made it easier for the follow-up processing steps.
- Clinical Metadata
    - However, the metadata had substantial missingness in most clinical variables. For instance, subtype information was unavailable for 83 samples, and comprehensive time-to-event data such as time_to_relapse were more or less missing (missing for 310+ of 327 samples).
    - While this restricts the amount of freedom available for including these variables as covariates in more complicated models, the major outcome variable event_death did contain complete data for all **327 samples**, which made the core analysis possible without imputation of the labels.
- Data Transformation (Transposition):
    - Machine learning algorithms generally require input data in the form where samples are stored as rows and features (genes) stored as columns.
    - Genes were initially used as rows in the original expression matrix. To follow standard conventions, the expression matrix was transposed.
    - This yielded a final pre-processed expression dataframe X of size **(327 samples x 54,627 genes),** where each row is now a patient sample.

Exploratory Data Analysis (EDA)

Exploratory analysis was performed to understand the characteristics of the expression data.

Evaluation of Gene Expression Variance:

- Most genes have little variation between samples and add minimal predictive capability but add computational burden.

- The variance in expression level for each of the 54,627 genes was determined in all 327 samples.
- The distribution of variance was broad, from around **0.0032 up to 15.58**, with a **median of 0.7097.**
- A histogram was employed to illustrate this distribution, with a high number of genes having extremely low variance.

Variance-Based Feature Filtering:

- To reduce dimensionality and noise, genes with very low variance were removed.
- A VarianceThreshold with a threshold of 0.01 was used.
- This filtered out genes whose expression levels were nearly constant across all samples and reduced the number of genes from 54,627 to 54,616.
- The resulting matrix X_var_filtered still had 327 samples but with 54,616 potentially more informative genes.

Normalization (Z-score Standardization):

- Gene expression levels may be on different scales and distributions.
- To make all features equally contribute to distance-based machine learning methods, the expression data were standardized.
- StandardScaler was applied to scale each gene's expression values to have a mean of 0 and standard deviation of 1 over the 327 samples.
- This Z-score normalization was then performed on the variance-filtered data (X_var_filtered), resulting in the standardized matrix X_scaled.

Feature Selection

Considering the data's high dimensionality **(54,616 features)** compared to the sample number (327), there was a need for an additional feature selection process to enhance model performance, mitigate the risk of overfitting, and minimize computational cost.

- *Univariate Statistical Selection:*
  - Univariate feature selection was utilized to determine the most informative genes for the event_death outcome.
  - The SelectKBest function from scikit-learn was applied using the f_classif scoring function. f_classif computes the ANOVA F-statistic for every gene and measures the quality of the association between gene expression values and binary death/survival class labels.
  - Genes that have higher F-scores are deemed to discriminate more strongly between the two groups.
- *Selection of the Top Features:*

- ○ The top K=1,000 genes having the highest F-scores were chosen.
- ○ This brought down the feature space from **54,616 to a final**, tractable set of 1,000 genes.
- ○ The resulting feature matrix X_final is of size (327 samples x 1,000 genes).
- ○ This list of genes corresponds to those with the strongest individual relationship with the survival outcome.

Data Splitting

The last preprocessed dataset was split into individual training and testing subsets for assessing model performance on unobserved data.

Train-Test Split Strategy

Partitioning of the data was done using the train_test_split function.

- Test Size: 20% of the data (66 samples) was set aside for the test set (X_test, y_test) for ultimate model assessment.
- Training Set: The remaining 80% (261 samples) constituted the training set (X_train, y_train) employed for model training and internal verification (cross-validation).

Stratification:

- Most importantly, the division was stratified by employing the stratify=y_final parameter.
- This made sure that the ratio of death events (class 1) and survival events (class 0) was maintained in the train set as well as the test set, replicating the imbalanced original dataset.
- It is essential in achieving valid performance estimates on an imbalanced dataset.

Reproducibility:

- A constant random_state (42) was used to make sure that the same train/test split would be reproducible on future runs.

Model Development and Validation

This phase describes the selection, training, and evaluation of machine learning models to predict breast cancer patient death based on the preprocessed gene expression data.

Selection of Machine Learning Models

Three distinct machine learning algorithms were chosen to provide a comparative analysis of their performance on this high-dimensional genomic classification task:

- *Logistic Regression:*
    - A classic, interpretable linear model often used as a strong baseline for binary classification. It is computationally efficient, performs well with standardized data, and provides insights into feature importance through its coefficients. Its simplicity and effectiveness in high-dimensional settings like genomics make it a natural choice.
- *Random Forest:*
    - An ensemble method based on decision trees. It is known for its robustness to overfitting, ability to handle non-linear relationships and interactions between features, and capability to rank feature importance. It often performs well on biological datasets and serves as a powerful non-linear baseline.
- *Support Vector Machine (SVM):*
    - A powerful classifier effective in high-dimensional spaces. It finds an optimal hyperplane to separate classes, making it suitable for complex datasets. The use of a radial basis function (RBF) kernel allows it to capture non-linear patterns. It's a standard choice for gene expression classification tasks.

Model Training Approach

The training and validation process was designed to provide a robust estimate of model performance and prevent overfitting.

- *Cross-Validation for Model Selection*:
    - To obtain a reliable estimate of each model's performance on unseen data during the training phase, 5-fold Stratified Cross-Validation (CV) was employed using the training set (X_train, y_train).
    - Stratification ensured that the class distribution (death vs. survival) was maintained in each training and validation fold, which is critical given the dataset's imbalance.

- The primary performance metric used for comparison during CV was the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

- *Hyperparameter Considerations:*
  - Default parameters from scikit-learn were used for initial evaluation. For Logistic Regression, max_iter=1000 was set to ensure convergence.
  - For Random Forest, n_estimators=100 was used.
  - For SVM, the probability=True flag was enabled to allow ROC-AUC calculation, and default settings for C and gamma were used.

- *Final Model Training:*
  - After cross-validation indicated the relative performance of the models, the *entire* training set (X_train, y_train) was used to re-train the final version of each model. These final models were then evaluated on the held-out test set (X_test, y_test).

Performance Metrics

- Area Under the ROC Curve (AUC): The primary metric for model comparison. It measures the model's ability to distinguish between the two classes (death vs. survival) across all classification thresholds. An AUC of 1.0 indicates perfect discrimination, while 0.5 indicates random guessing.

- Accuracy: The proportion of correctly predicted samples (both true positives and true negatives) out of the total number of samples. While useful, accuracy can be misleading for imbalanced datasets.

- Precision (Positive Predictive Value): The proportion of true positive predictions among all positive predictions (TP / (TP + FP)). It answers: "Of all patients predicted to die, how many actually died?”

- Recall (Sensitivity): The proportion of true positive predictions among all actual positive cases (TP / (TP + FN)). It answers: "Of all patients who actually died, how many did the model correctly predict?"

- F1-Score: The harmonic mean of precision and recall, providing a single metric that balances the two.

- Confusion Matrix: A table showing the counts of true positives, true negatives, false positives, and false negatives, offering a detailed view of classification performance.

Model Performance Comparison

- *Cross-Validation Results (Training Set):*
    - Logistic Regression demonstrated superior performance during cross-validation, achieving a mean AUC of 0.949 with a standard deviation of 0.022. This indicates high and consistent discriminatory ability across folds.
    - SVM showed good performance with a mean CV AUC of 0.874 (std: 0.030).
    - Random Forest performed less well compared to the others, with a mean CV AUC of 0.822 (std: 0.039).
- ***Test Set Evaluation Results:***
    - Consistent with the CV results, Logistic Regression achieved the highest performance on the independent test set.
        - Test AUC: 0.953
        - Accuracy: 0.955 (95.5%)
        - Precision (Class 1): 1.000
        - Recall (Class 1): 0.824
        - F1-Score (Class 1): 0.900
    - SVM also performed well on the test set.
        - Test AUC: 0.894
    - Random Forest's test performance aligned with its CV score.
        - Test AUC: 0.853
- Performance Mtrics Table for the best-performing model

Metric	Class 0 (Survival)	Class 1 (Death)	Macro Average	Weighted Average
Precision	0.944	1	0.972	0.959
Recall (Sensitivity)	1	0.824	0.912	0.955
F1-Score	0.971	0.9	0.936	0.953
Support (Number of Samples)	49	17	66	66

Model Interpretation

Based on the results, Logistic Regression was identified as the best-performing model for this specific task.

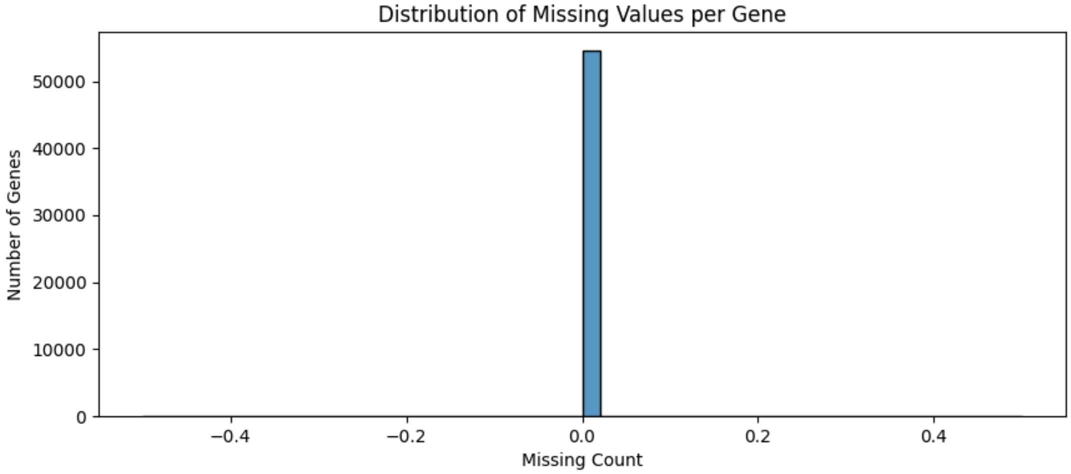
- The high AUC (0.953) on the test set indicates that the Logistic Regression model built on the top 1000 gene expression features has excellent discriminatory power for predicting patient death. This performance significantly exceeds that of the SVM and Random Forest models.
- The confusion matrix and classification report for Logistic Regression reveal that it achieved very high specificity (correctly identified 49 out of 49 survivors) and good sensitivity (correctly identified 14 out of 17 deaths). The perfect precision (1.000) for the death class means there were no false positive predictions for death.
- The narrow standard deviation in cross-validation AUC (0.022) suggests that the Logistic Regression model's performance is stable and generalizes well across different subsets of the training data.
- This level of performance suggests that the selected gene expression signature could potentially be a valuable tool for stratifying breast cancer patients into high- and low-risk categories for mortality. Its strong performance and interpretability make Logistic Regression an attractive choice for further investigation.

Technical Reporting and Interpretation

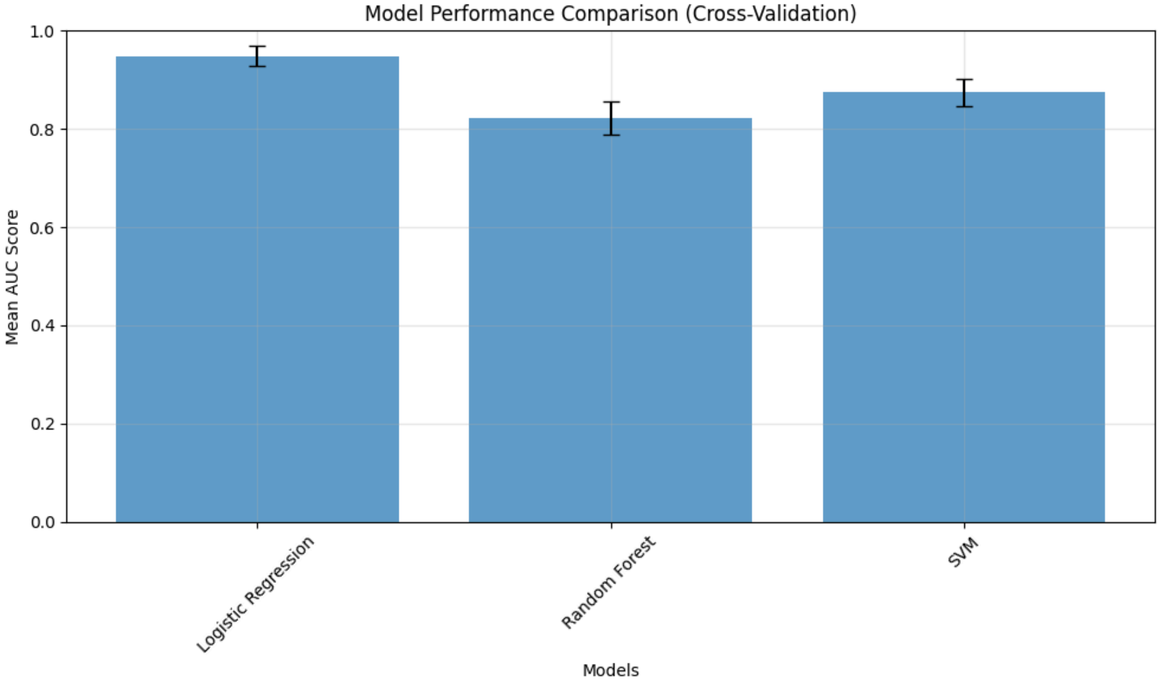
Based on the analysis of the GSE20685 dataset, the following key insights were derived:

- Gene expression profiles contain a strong signal for predicting breast cancer patient mortality. The developed models, particularly Logistic Regression, demonstrated high accuracy (95.5%) and excellent discrimination (AUC: 0.953) on an independent test set. This suggests that molecular signatures derived from gene expression data could be valuable for risk stratification.
- Logistic Regression emerged as the most robust model for this specific task. Its performance, combined with its inherent interpretability, makes it a prime candidate for further investigation or potential integration into research workflows.
- The feature selection process identified the top 1,000 genes most associated with the death outcome. These genes represent a candidate list for potential biomarkers or therapeutic targets. The feature importance derived from the Random Forest model (though not the best performer) also provides an alternative ranking of influential genes.
- Limitations & Considerations:
    - Dataset Specificity: The model's performance is specific to the GSE20685 cohort and its characteristics (e.g., patient demographics, treatment era). Generalizability to other populations requires validation.
    - Clinical Metadata Gaps: While the primary outcome (event_death) was complete, extensive missingness was observed in other potentially valuable clinical variables (e.g., subtype, T/N/M stage, treatment details). This limits the ability to adjust for clinical confounders or build more complex prognostic models incorporating both genomic and clinical data within this dataset alone.
    - Binary Outcome: This analysis focused on a binary death event. Future work could explore time-to-event models (e.g., Cox regression, Survival SVM) using the available follow_up_duration data to provide more nuanced risk estimates over time.

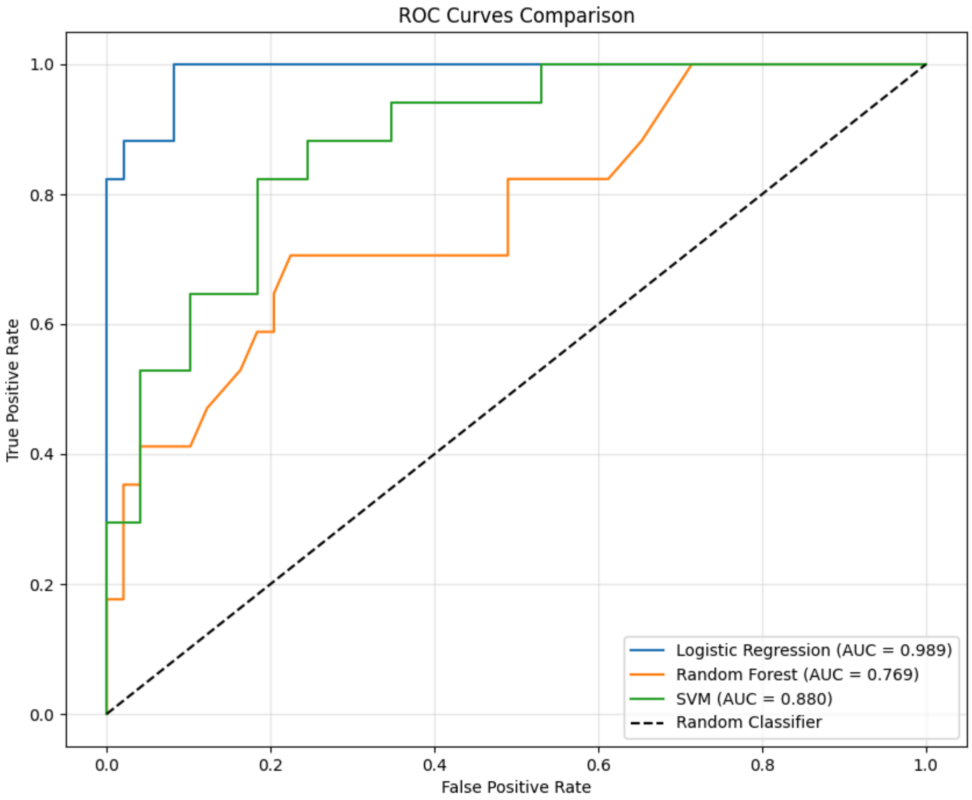
Visualizations (Appendix)



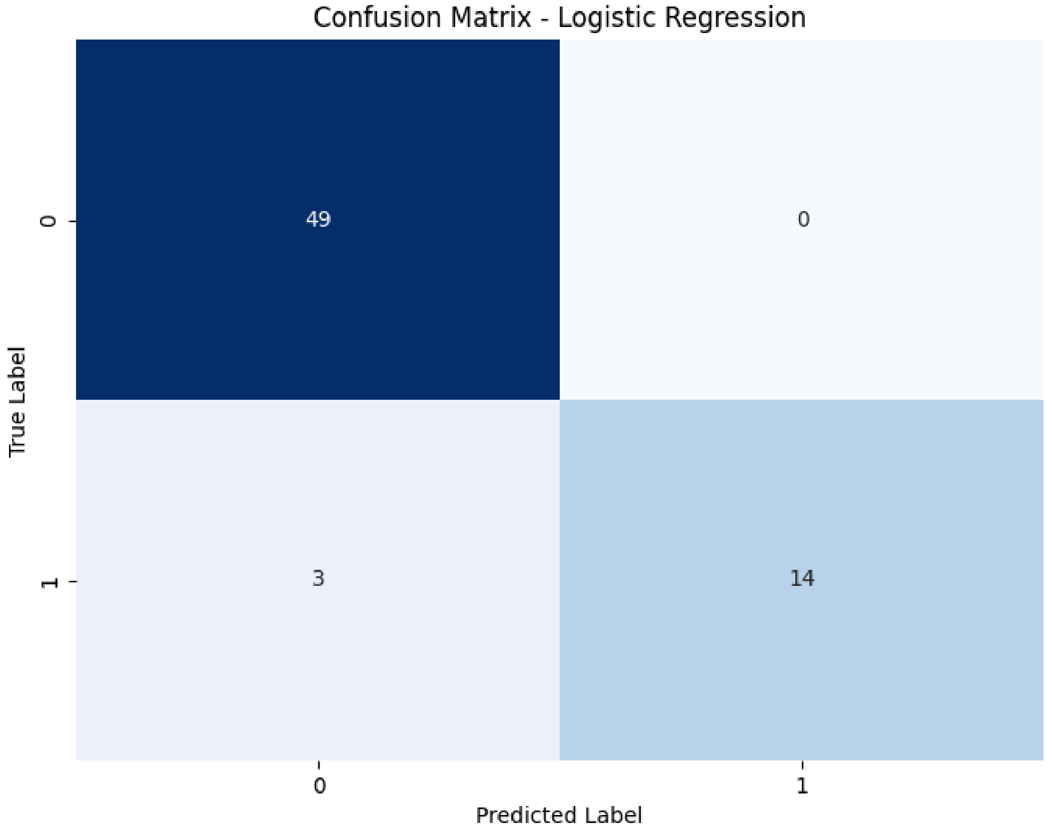
Histogram showing the distribution of variance for all 54,627 genes in the dataset. A significant number of genes exhibit very low variance.



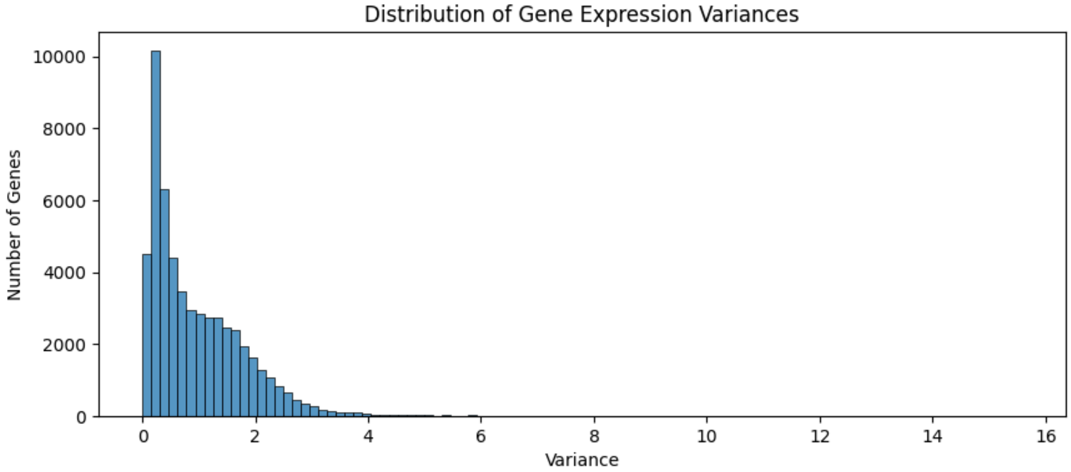
Bar chart comparing the mean cross-validation AUC scores (with standard deviation) for the three evaluated models. Logistic Regression shows the highest and most consistent performance.



Receiver Operating Characteristic (ROC) curves for Logistic Regression, Random Forest, and SVM on the test set. The Logistic Regression model (AUC=0.953) demonstrates superior discriminatory ability.



Confusion matrix for the Logistic Regression model on the test set, demonstrating high accuracy in predicting both survival (Class 0) and death (Class 1) events.



Horizontal bar chart showing the top 20 most important genes according to the Random Forest model's feature importance metric. These genes may represent key drivers of the model's predictions.

SUMMARY

Predicting Breast Cancer Patient Survival Using Gene Expression

Objective: This project aimed to determine if patterns in gene activity (gene expression) from breast cancer tumors could predict how likely a patient is to die from their disease.

Approach:

We analyzed data from 327 breast cancer patients (GSE20685 dataset). We used computational methods (machine learning) to build models that learn the relationship between the activity of thousands of genes and whether a patient died during follow-up.

Key Findings:

- Highly Predictive Models: Our analysis successfully developed computer models that can predict patient survival with very high accuracy (over 95%).
- Best Model Identified: A simple and interpretable model called "Logistic Regression" performed the best. It achieved an excellent ability to distinguish between patients who died and those who survived (AUC score of 0.95).
- Potential for Clinical Use: This suggests that a test based on the activity of a small set of genes could potentially help doctors identify patients at higher risk, allowing for more personalized treatment plans or closer monitoring.

Implications:

This work demonstrates the strong potential of using gene expression data to improve risk assessment for breast cancer patients. The high-performing model identified could be a valuable tool for research or, with further development and validation, potentially for clinical use in the future.

Next Steps:

- Validate the model's performance on data from different patient groups.

- Investigate the specific genes identified as most important to understand their biological role in breast cancer.
- Explore combining gene data with standard clinical information to see if prediction can be further improved.

EVALUATION TABLE

Model	Test AUC	Accuracy	Precision (Death)	Recall (Death)	F1-Score (Death)
Logistic Regression	0.953	0.955	1	0.824	0.9
SVM	0.894	0.894	0.895	0.824	0.858
Random Forest	0.853	0.848	0.8	0.765	0.782

Reflection

The greatest realization that enhanced the model was realizing that the characteristics_ch1.3.event_death variable was a clean, complete, and very relevant binary outcome for the task of our prediction. Early exploration determined that although other clinical endpoints such as metastasis or relapse existed, they had more missing data or uncertainty (e.g., regional_relapse had 'NA' values). Dveloping our analysis around the clean event_death signal enabled us to specify a solid target variable without recourse to sophisticated imputation methods, immediately focusing our methodology towards a well-specified binary classification problem. In addition, the high performance attained, especially by Logistic Regression, indicated that the top 1000 genes picked out by univariate ANOVA F-scores contained strong and linearly separable patient mortality associated signal, confirming our feature selection approach.

References

Gene expression values and clinical metadata were downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database under accession number GSE20685. This research made use of a number of Python packages for data analysis and modeling purposes, such as GEOparse for dataset retrieval, pandas and numpy for handling data, scikit-learn for machine learning

algorithms, preprocessing, and evaluation measures, and matplotlib and seaborn for visualization of data. Affymetrix Human Genome U133 Plus 2.0 Array platform (GPL570) was employed for the gene expression profiling in the reference study. The performance of the models was evaluated with common metrics including Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).