Breast Cancer Subtype Classification Using the METABRIC Dataset

Ishan Baweja

Project Repository

The complete codebase, preprocessing scripts, and model notebooks are available at the following GitHub repository:
`https://github.com/Ishan-Baweja/cmse492_project`

Abstract

Breast cancer contains several molecular subtypes that differ in structure and treatment response. The METABRIC study introduced a large collection of clinical and molecular measurements widely used in research. This project builds a complete machine learning workflow to classify tumors into PAM50 and Claudin-low subtypes, which reflect intrinsic biological structure described in earlier clinical studies. The workflow includes exploratory data analysis, missing value handling, encoding, variance filtering, scaling, principal component analysis, model training, model comparison, and interpretation using SHAP values and random forest importance. Logistic Regression produced the strongest performance with an accuracy of approximately 0.729 and a weighted area under the curve (AUC) of approximately 0.917.

1 Background and Motivation

Breast cancer research shows that patients differ across multiple subtypes that are linked to tumor biology and long-term survival outcomes. Identifying these subtypes is important because treatment decisions depend on them. Manual evaluation cannot process a dataset of this scale, which contains thousands of gene-related and clinical measurements. Machine learning supports this process by identifying patterns that connect patient data to intrinsic tumor classes. Previous work has applied linear models, ensemble models, and margin-based algorithms to breast cancer classification.

Accurate subtype prediction is clinically important because treatment decisions, drug response, and long-term survival all depend on subtype assignment. Incorrect or delayed subtype identification can lead to ineffective treatment, higher recurrence rates, and unnecessary toxicity. A reliable automated system reduces diagnostic uncertainty and supports personalized cancer therapy.

Several previous studies have demonstrated that molecular subtypes can be learned from genomic data, and that predictive accuracy improves when combining clinical and gene-expression features. The METABRIC dataset provides benchmark subtype labels widely used for developing machine learning models.

2 Data Description

Data Origin

The METABRIC dataset contains clinical and genomic profiles of more than one thousand breast cancer patients collected using microarray platforms and clinical records.

Rows and Columns

The dataset contains 1,903 samples with a wide range of clinical and genomic features. The target label includes seven tumor subtypes from the PAM50 and Claudin-low classification systems.

Data Types

The dataset includes numeric clinical variables such as tumor size, mutation count, and receptor levels, as well as categorical clinical descriptors including therapy status, histologic grade, and menopausal state.

Missing Values

Missing values appear primarily in clinical variables. Categorical features were imputed using the mode, while numeric variables were imputed using the median.

Class Balance

The seven-class subtype distribution is imbalanced. A stratified train–test split was used to preserve subtype proportions.

3 Exploratory Data Analysis

Exploratory data analysis was performed to understand the distributions, correlations, and overall structure of the dataset before modeling. The mutation count distribution shows strong right skewness. Tumor size includes several noticeable outliers. The correlation heatmap highlights weak to moderate relationships among numeric clinical features. A two-dimensional PCA projection also reveals partial separation between tumor subtypes, suggesting that meaningful structure is present even in a reduced feature space.

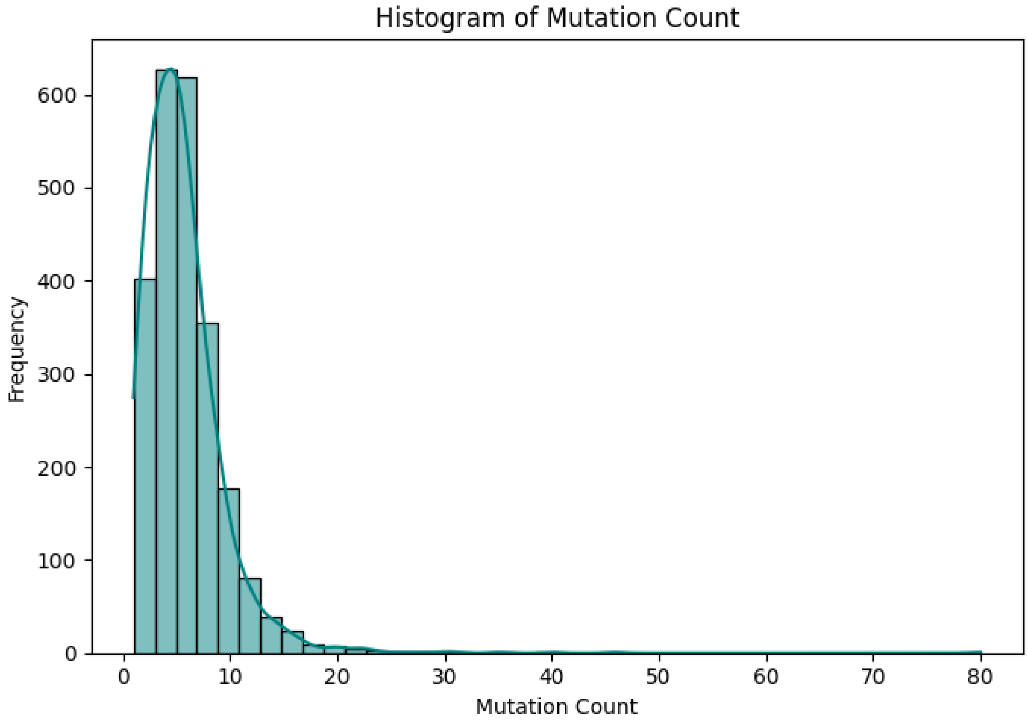

Figure 1: Histogram of mutation count.

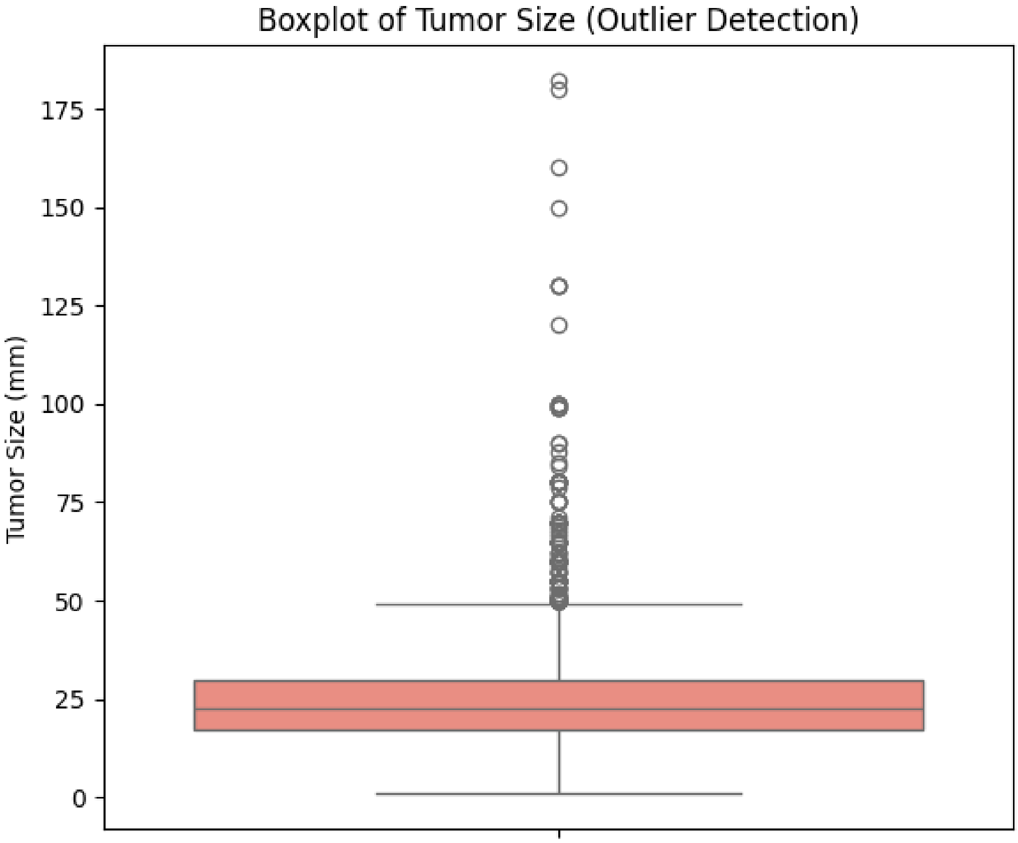

Figure 2: Boxplot of tumor size.

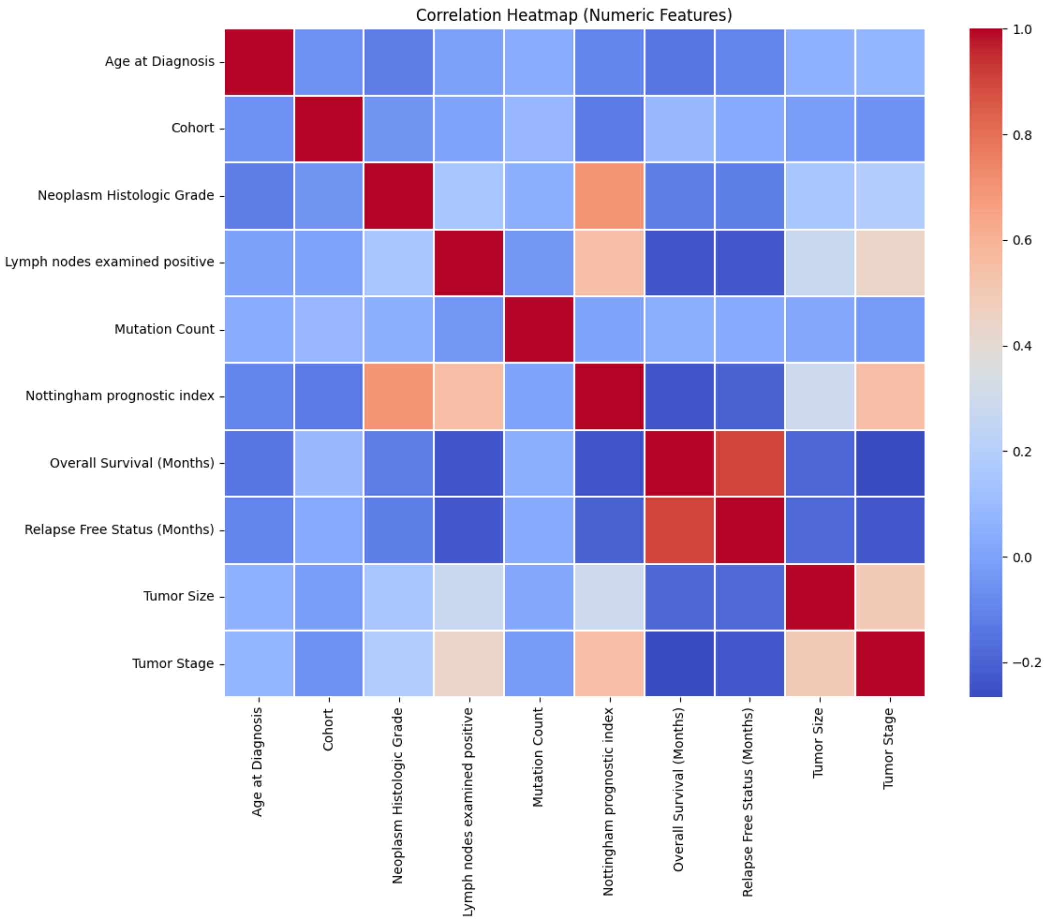

Figure 3: Correlation heatmap of numeric clinical features.

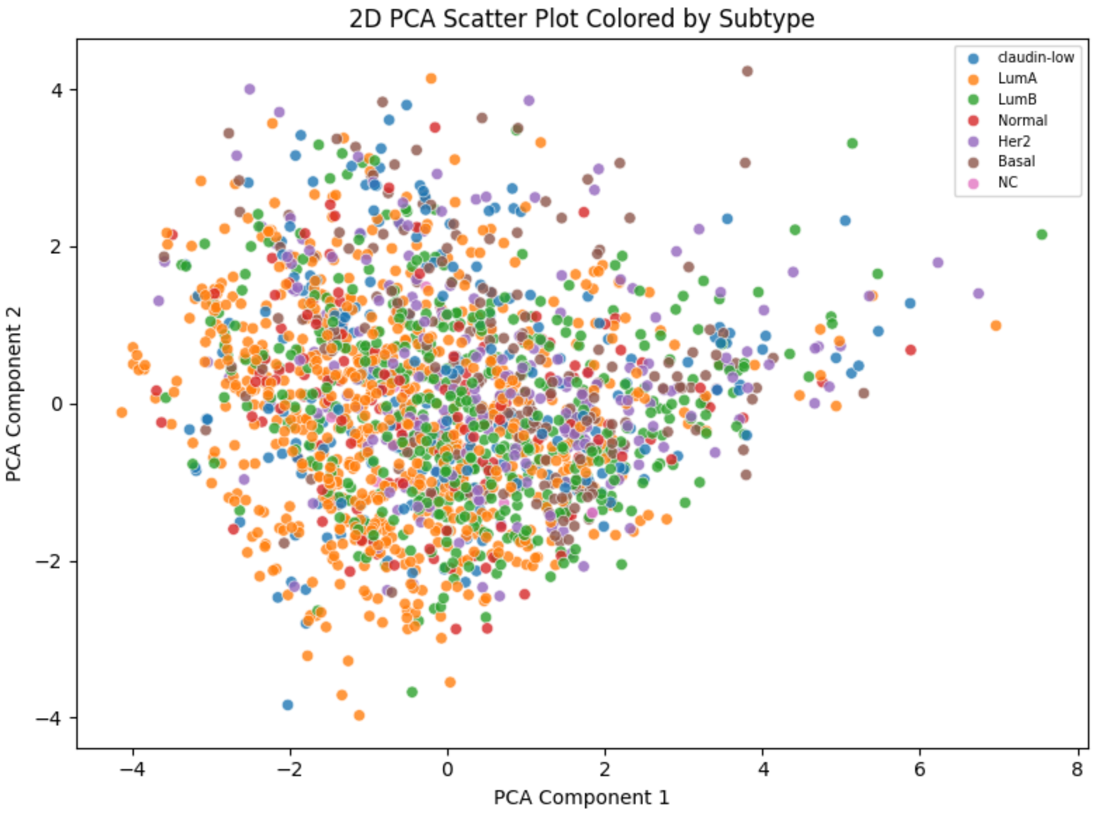

Figure 4: Two-dimensional PCA scatter plot colored by tumor subtype.

4 Preprocessing

The dataset was split into an 80/20 stratified train–test split. Categorical variables were one-hot encoded. A variance filter removed constant features. Standard scaling ensured consistent numerical ranges. Principal Component Analysis (PCA) reduced dimensionality to 50 components, capturing most of the variance.

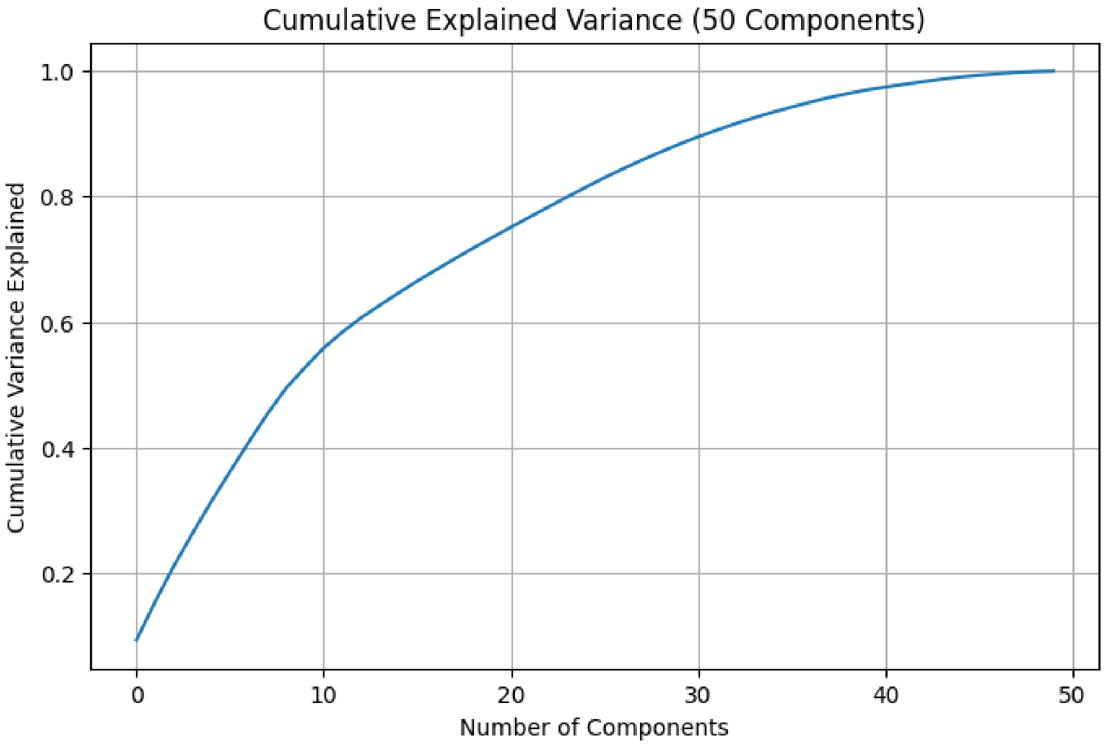

Figure 5: Cumulative variance explained by PCA components.

5 Machine Learning Task and Objective

The goal is supervised multiclass classification using high-dimensional genomic and clinical predictors. Models must learn both linear and nonlinear associations in the PCA-transformed feature space.

6 Models

Three models were evaluated:

- Logistic Regression
- Random Forest
- Support Vector Machine

Logistic Regression provides a linear decision boundary and includes L2 regularization, which penalizes large coefficients to reduce overfitting. Support Vector Machines use a margin-based classifier that implicitly regularizes through the margin parameter C, controlling the balance between maximizing the decision margin and minimizing misclassification. Random Forest models capture nonlinear relationships through an ensemble of decision trees but require careful regularization to avoid overfitting when working with high-dimensional PCA features. Structural regularization in Random Forest is introduced by limiting tree depth and the number of trees.

7 Training Methodology

Hyperparameters were selected using five-fold stratified cross-validation. Learning curves for each model illustrate how training and validation performance change with increasing sample size.

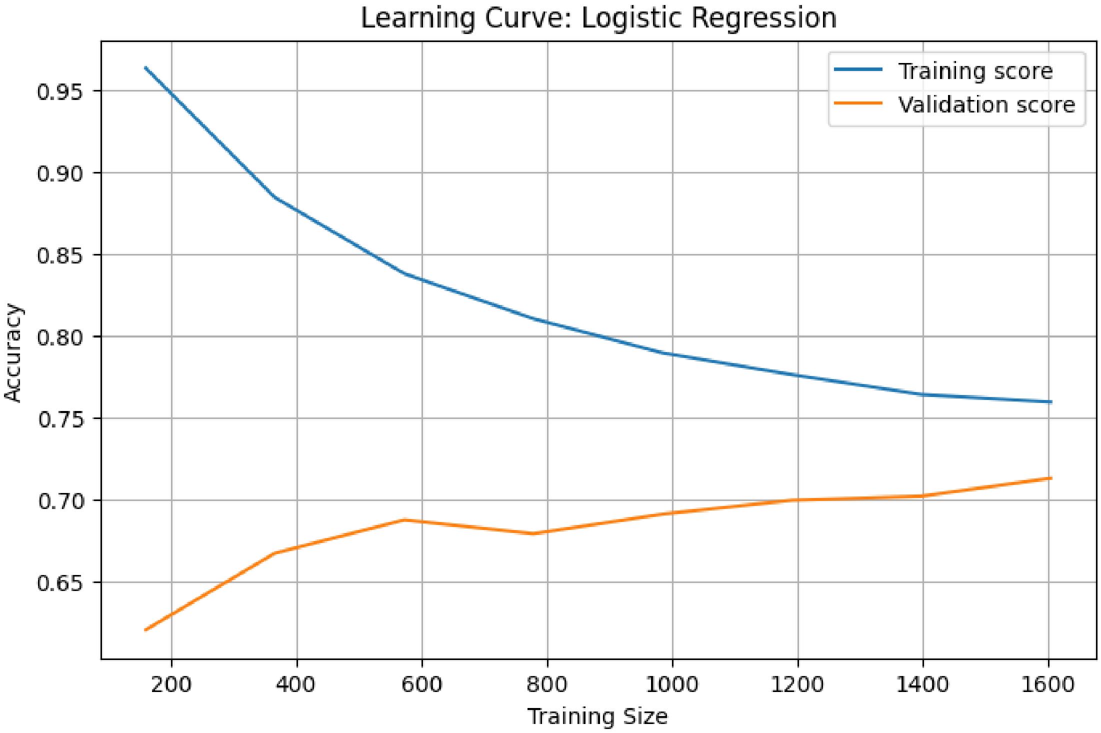

Figure 6: Learning curve for Logistic Regression.

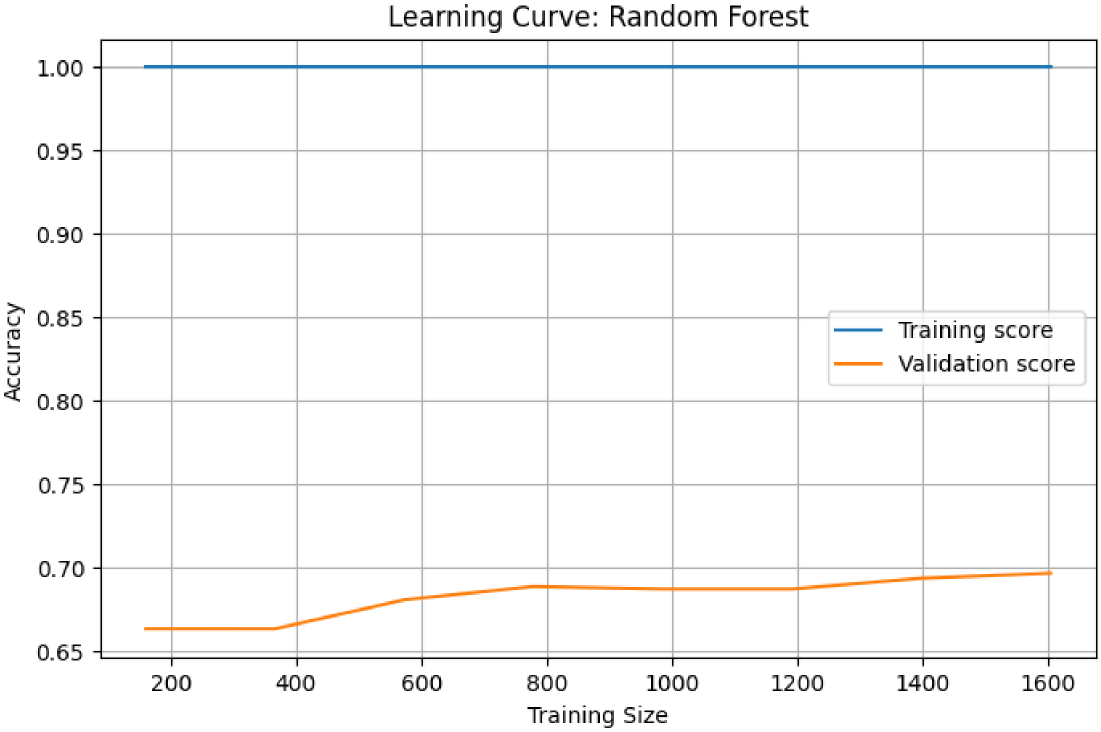

Figure 7: Learning curve for Random Forest.

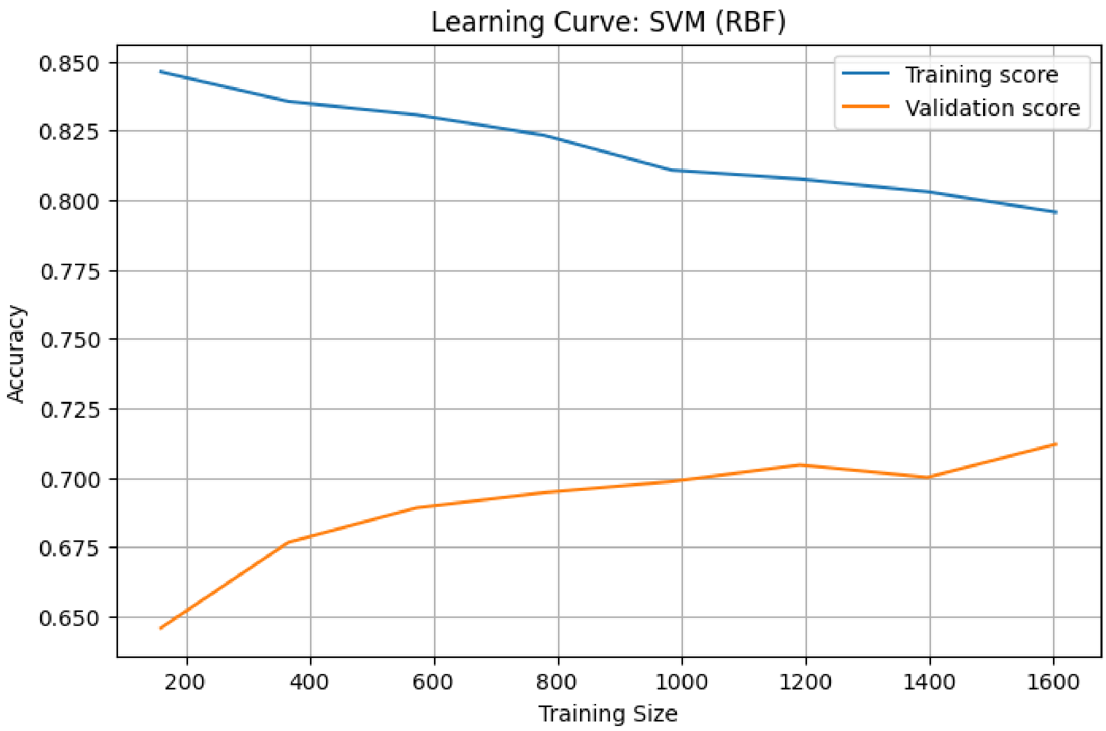

Figure 8: Learning curve for Support Vector Machine.

8 Metrics

Accuracy, precision, recall, F1-score, and AUC were used to evaluate each model. Weighted metrics were computed to address class imbalance, ensuring fair evaluation across all seven subtypes.

9 Results and Model Comparison

Logistic Regression performed best because the PCA transformation produced linearly separable components that favor linear classifiers. Random Forest performed slightly worse due to the difficulty of modeling very high-dimensional PCA features, which require more samples to avoid overfitting. Support Vector Machine showed the weakest performance because the RBF kernel is sensitive to class imbalance and requires extensive hyperparameter tuning with many classes.

These differences explain the performance gaps observed across accuracy and AUC metrics.

Performance Table

Model	Accuracy	Precision (weighted)	Recall (weighted)	F1 (weighted)	AUC
Logistic Regression	0.729	0.722	0.729	0.708	0.917
Random Forest	0.717	0.667	0.717	0.685	0.912
Support Vector Machine	0.655	0.615	0.655	0.607	0.907

Table 1: Performance of all models.

Training and Inference Time Table

Model	Training Time (s)	Inference Time (s)
Logistic Regression	0.06849	0.00058
Random Forest	2.71626	0.01984
Support Vector Machine	0.91161	0.04627

Table 2: Training and inference times for all models.

10 Conclusion

This project developed a complete machine learning workflow for breast cancer subtype classification using the METABRIC dataset. Logistic Regression achieved the strongest performance, while Random Forest captured more nonlinear structure and provided stronger interpretability. Support Vector Machine performance was sensitive to subtype imbalance. PCA improved model stability and reduced dimensionality. Future work includes experimenting with class-balancing techniques such as SMOTE, feature engineering, and deep learning models for genomic data.

References

[1] Curtis et al. The molecular taxonomy of breast cancer international consortium. *Nature*, 2012.

[2] Parker et al. Pam50 gene expression assay for intrinsic subtyping of breast cancer. *Journal of Clinical Oncology*, 2009.

[3] Breiman. Random forests. *Machine Learning*, 2001.

[4] Cortes and Vapnik. Support vector machines. *Machine Learning*, 1995.

[5] Lundberg and Lee. A unified approach to interpreting model predictions. *NeurIPS*, 2017.

[6] Pedregosa et al. Scikit learn machine learning library, 2011.