Predicting Pathological Complete Response in Triple Negative Breast Cancer Using Machine Learning Models

Joyce Tang, Amy Xu, Carmel High School, Carmel, IN, United States

Research Goal and Hypothesis

Research Goal

- To develop a machine learning model that combines tumor-intrinsic and immune-related features to accurately predict pathological complete response (pCR) in triple-negative breast cancer patients.

Hypothesis

- A model combining tumor-intrinsic and immune-related features will predict pCR more accurately than models using either feature set alone, due to the joint influence of tumor biology and the immune microenvironment.

Methodology

I. Downloaded publicly available TNBC dataset from cBioPortal with tumor-intrinsic and immune-related features

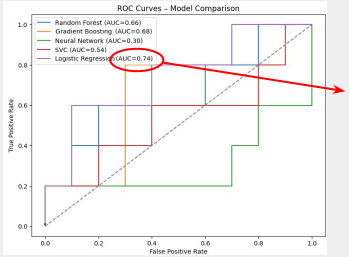
II. Adjusted dataset by scaling numerical variables and encoding categorical variables to remove dataset imbalance

III. Trained logistic regression, random forest, support vector, gradient boosting, and neural network models using same dataset

IV. Compared performance based on ROC AUC and calculated accuracy, precision, recall, specificity, and F1 scores for each model

V. Used permutation importance to identify influence of each feature on pCR predictions

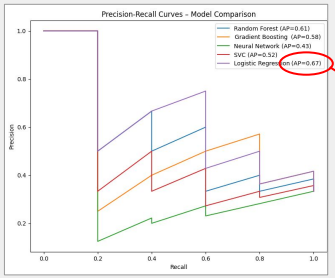
Data Analysis and Results



Lowest false positive and higher true positive rate. (AUC = 0.74)

Feature	Logistic
PDL1 Combined Positive Score	0.086
TNBC Subtype Mesenchymal Stem-Like	0.103
TNBC Subtype Unknown	0.12
PAM50 Subtype HER2	**0.175**

PAM50 Subtype HER2 was the most significant feature in aiding the prediction of pCR.



Fewest false positives and negatives. (AP=0.67)

Measure of how well the model identifies true pCR patients without being misled by class imbalance or high specificity alone.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Model	ROC AUC	Accuracy	Precision	Recall	Specificity	F1 Score
Logistic	0.74	0.8	0.75	0.6	0.9	0.66667

Conclusion and Further Applications

Conclusion

- Logistic regression achieved the highest ROC AUC (0.74), indicating it had the best performance in predicting pCR for this data set
- Based on permutation importance, PAM50 Subtype HER2, CD3 Positive IHC (%), PDL1 Combined Positive Score, and PAM50 Subtype Normal were the more significant features in aiding the prediction of pCR

Further Applications

- Using our machine learning models and feature importances, we can expand our work into developing better systems for diagnosing patients to aid in TNBC prevention and patient care based on identified features.