Machine Learning–Based Prediction of Pathological Complete Response in Triple-Negative Breast Cancer

Joyce Tang and Amy Xu - Carmel High School, IN

Introduction

- **Triple Negative Breast Cancer** (or TNBC) is a subtype of breast cancer defined by the absence of estrogen receptors (ER), progesterone receptors (PR), and HER2 expression.
- TNBC accounts for **10-15%** of all cancers, and around 42% of patients diagnosed **relapse rapidly** after standard treatment.
- **PAM50 Molecular Subtype**
  A gene-expression–based classification system that categorizes breast tumors into intrinsic subtypes
- **Machine Learning Models**: Logistic Regression, Random Forest, Support Vector, Gradient Boosting, and Neural Network

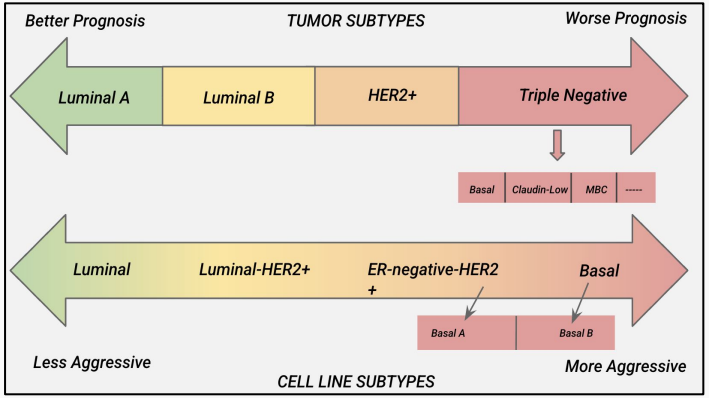


Fig. 1: *BRCA subtypes by prognosis and aggression*

Project Goal:

To evaluate multiple machine learning models that integrate tumor-intrinsic and immune-related features, identify the most effective model for predicting pCR in triple-negative breast cancer, and assess feature importance across models.

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

brca_dldccc_2022_clinical_data (1).tsv

cBioPortal FOR CANCER GENOMICS

Download TNBC datasets from cBioPortal

Encode categorical variables

```
FEATURE_COLUMNS = [
    "Immune Score",
    "Stromal Score ",
    "Tumor Purity",
    "Mutation Count",
    "CD3 Positive IHC (%)",
    "PDL1 Combined Positive Score",
    "xCell Immune Score",
    "xCell Stroma Score"
]
```

Analysis of different models

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

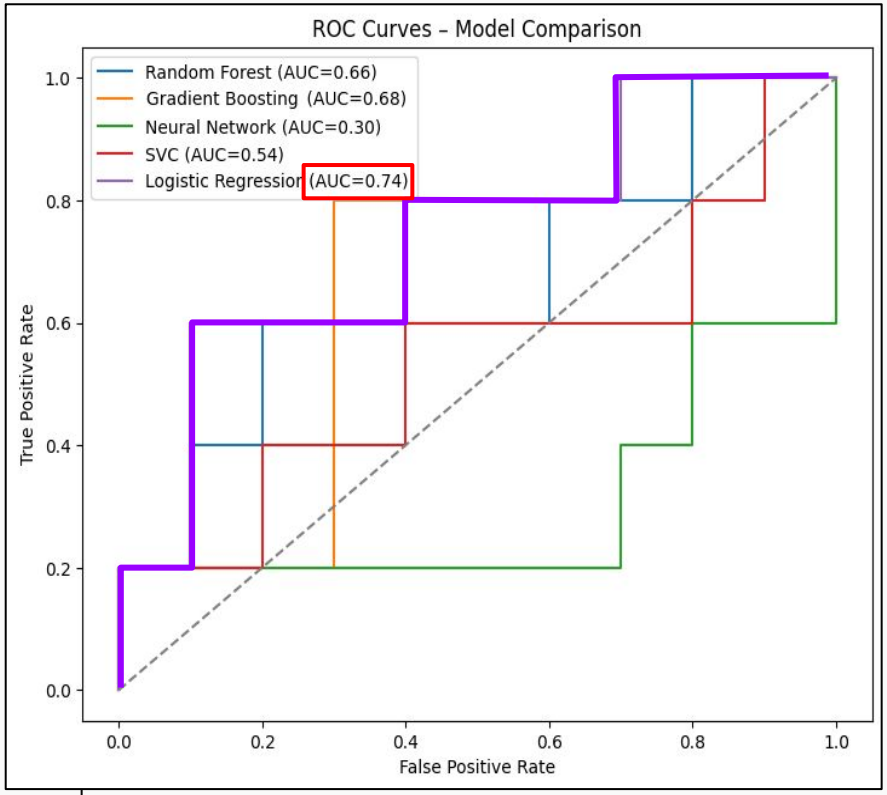
$$\text{Specificity} = \frac{TN}{TN + FP}$$

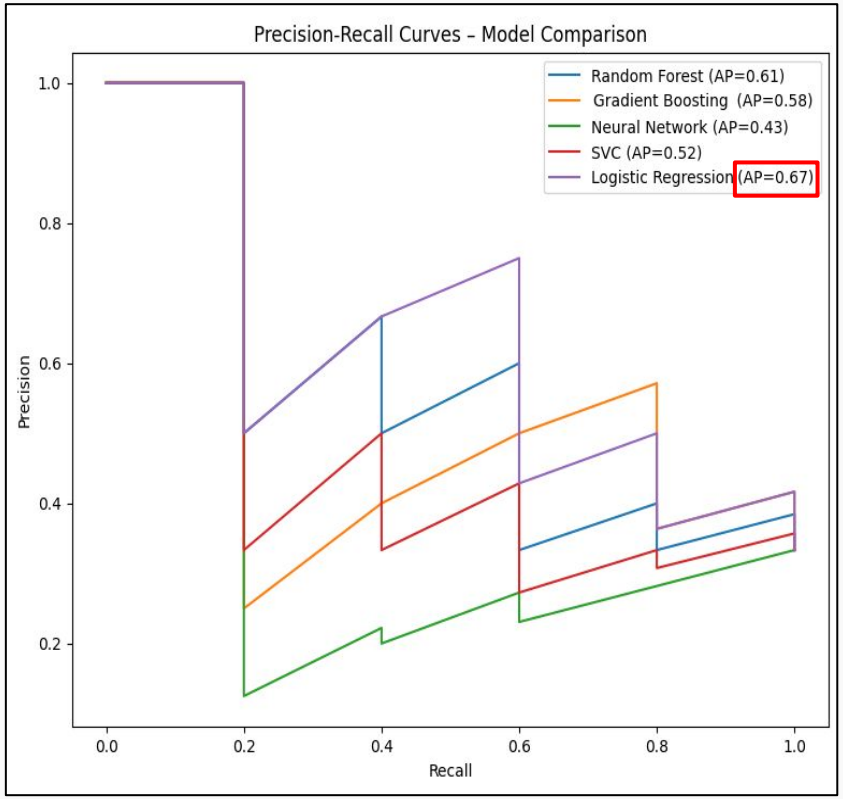
$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Evaluate performance with ROC AUC and threshold dependent metrics

Calculate permutation importance of each feature

Analyze model performance across variables





ROC and Precision–Recall curves comparing model performance in predicting pCR in triple-negative breast cancer, demonstrating that logistic regression achieves the strongest overall discrimination and balance between sensitivity and precision.

Model	Accuracy	Precision	Recall	Specificity	F1 Score
Logistic	**0.8**	**0.75**	**0.6**	**0.9**	**0.66667**
Gradient Boosting	0.73333	1	0.2	1	0.33333
Random Forest	0.73333	0.66667	0.4	0.9	0.5
SVC	0.66667	0	0	1	0
Neural Network	0.53333	0.25	0.2	0.7	0.22222

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

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

Across all models, logistic regression achieved the ***best overall performance***, with the highest F1 score of **0.67**, indicating effective identification of pCR cases.

- *While gradient boosting reached perfect precision (1.0) and specificity (1.0), its very low recall (0.2) suggests it failed to identify most true pCR patients,* ***limiting its clinical usefulness***

Table 1: Feature Importance Comparison Across Models

Feature	Logistic	Random Forest	Gradient Boosting	SVC	Neural Network
CD3 Positive IHC (%)	-0.001	**0.167**	0.077	0.024	0.001
Immune Score	0.005	0.031	-0.0335	-0.031	-0.095
Mutation Count	-0.012	-0.099	0	-0.064	-0.066
PAM50 Subtype HER2	**0.175**	0.006	0.018	0.015	**0.058**
PAM50 Subtype LumA	0	0	0	0	0
PAM50 Subtype LumB	0	0	0	0	0
PAM50 Subtype Normal	-0.014	0	0	**0.092**	0.035
PAM50 Subtype Unknown	0.049	0.003	0	0.008	0.032
PDL1 Combined Positive Score	0.086	0.127	**0.169**	0.042	0.007
Stromal Score	0.012	0.003	-0.034	-0.054	0.026
TNBC Subtype Basal-Like 2	0.015	0.001	0	-0.008	-0.034
TNBC Subtype Immunomodulatory	-0.053	-0.023	-0.018	-0.056	-0.002
TNBC Subtype Luminal Androgen Receptor	0.008	-0.008	0	-0.086	-0.053
TNBC Subtype Mesenchymal	-0.052	0.007	0	-0.107	-0.104
TNBC Subtype Mesenchymal Stem-Like	0.103	0.023	0	0.026	-0.005
TNBC Subtype Unknown	0.12	0.005	0	-0.031	0.043
TNBC Subtype Unspecified	-0.013	0.024	0	-0.044	-0.103
Treatment Status Pre-Treatment	0	-0.004	-0.023	-0.081	0.005
Tumor Purity	0.002	0.006	-0.006	-0.058	-0.081

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

Conclusion

We demonstrate that pCR results in TNBC react positively to tumor subtype and microenvironmental context, with logistic regression providing the best accuracy and precision in predicting pCR based on the ROC AUC among 5 separate models. Multiple factors influenced stronger feature contributions relating to the prediction of pCR (PAM50 Subtype HER2). These factors, paired with our analytical models, further TNBC pCR research and earlier diagnosis to aid in prevention and patient care.

References

- Anurag, M., Jaehnig, E. J., Krug, K., Lei, J. T., Bergstrom, E. J., Kim, B. J., Vashist, T. D., Huynh, A. M. T., Dou, Y., Gou, X., Huang, C., Shi, Z., Wen, B., Korchina, V., Gibbs, R. A., Muzny, D. M., Doddapaneni, H., Dobrolecki, L. E., Rodriguez, H., Robles, A. I., ... Ellis, M. J. (2022). Proteogenomic Markers of Chemotherapy Resistance and Response in Triple-Negative Breast Cancer. Cancer discovery, 12(11), 2586–2605. https://doi.org/10.1158/2159-8290.CD-22-0200
- Dell'Aquila, K., Vadlamani, A., Maldjian, T., Fineberg, S., Eligulashvili, A., Chung, J., Adam, R., Hodges, L., Hou, W., Makower, D., & Duong, T. Q. (2024). Machine learning prediction of pathological complete response and overall survival of breast cancer patients in an underserved inner-city population. Breast cancer research : BCR, 26(1), 7. https://doi.org/10.1186/s13058-023-01762-w
- Jiang, C., Zhang, X., Qu, T., Yang, X., Xiu, Y., Yu, X., Zhang, S., Qiao, K., Meng, H., Li, X., & Huang, Y. (2024). The prediction of pCR and chemosensitivity for breast cancer patients using DLG3, RADL and Pathomics signatures based on machine learning and deep learning. Translational oncology, 46, 101985. https://doi.org/10.1016/j.tranon.2024.101985
- SEER*Explorer: An interactive website for SEER cancer statistics [Internet]. Surveillance Research Program, National Cancer Institute; 2025 Apr 16. [updated: 2026 Jan 8; cited 2026 Feb 11]. Available from: https://seer.cancer.gov/statistics-network/explorer/. Data source(s): SEER Incidence Data, November 2024 Submission (1975-2022), SEER 21 registries (excluding Illinois). Expected Survival Life Tables by Socio-Economic Standards.
- Wang, H., & Yee, D. (2019). I-SPY 2: a Neoadjuvant Adaptive Clinical Trial Designed to Improve Outcomes in High-Risk Breast Cancer. Current breast cancer reports, 11(4), 303–310. https://doi.org/10.1007/s12609-019-00334-2