Using radiomics features to predict molecular subtype of breast cancer

Albert Kang[1], Lucy Malmud[2], Ritish Natesan[3], Kate Podrebarac[4]

University of North Carolina at Chapel Hill[1], Johns Hopkins University[2], University of Michigan[3], Wofford College[4]

Introduction

Determining molecular subtype from receptor status is an important step in breast cancer treatment. Previous studies have shown moderate associations between magnetic resonance imaging (MRI) features and molecular subtype. Accurate prediction of molecular cancer characteristics could decrease the need for invasive biopsy procedures, leading to quicker clinical decision making and decreased cost of care for patients.

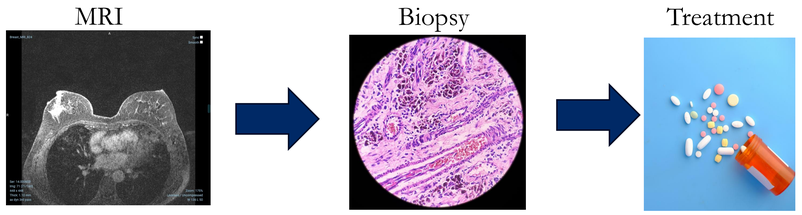


Fig. 1 — Breast cancer diagnosis to treatment pipeline.

Dataset

Our dataset comes from **The Cancer Imaging Archive (TCIA)** and consists of demographics, clinical outcomes, annotated MRIs, and extracted radiomics features for **922 patients.** Our analysis focused on pre-biopsy measurements, which included all 529 columns of the radiomics dataset as well as **four pre-biopsy clinical features:** age, menopausal status, race and ethnicity, and tumor stage. Throughout analysis, we omitted any patients with missing data. Across all **16 selected feature subsets,** the maximum number of rows was 922 (from PCA) and the minimum number of rows remaining was 841 (after merging with clinical features).

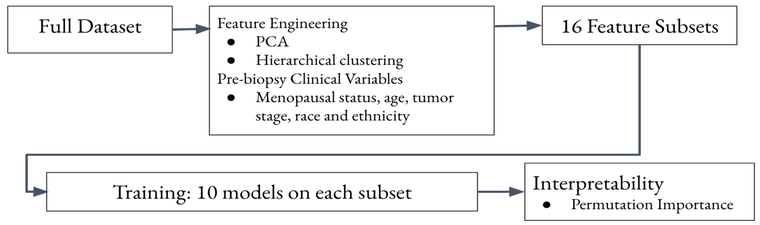
Patient Characteristics	Entire cohort	ER positive	PR positive	HER2 positive
Number of patients	922 (100%)	686 (74%)	598 (65%)	163 (18%)
Median age (age range) in years	52.25 (21.75–89.49)	52.82 (25.7–89.49)	52.42 (25.7–89.49)	48.38 (27.1–79.52)
Race				
White	651 (71%)	510 (74%)	453 (76%)	111 (68%)
Black	203 (22%)	123 (18%)	103 (17%)	36 (22%)
Others*	49 (5%)	38 (6%)	27 (5%)	14 (9%)
Not available	19 (2%)	15 (2%)	15 (2%)	2 (1%)
Menopausal status				
Pre	407 (44%)	293 (43%)	263 (44%)	82 (50%)
Post	499 (54%)	380 (55%)	323 (54%)	79 (49%)
Not available	16 (2%)	13 (2%)	12 (2%)	2 (1%)
Tumor staging (size)				
T1	409 (44%)	327 (48%)	292 (49%)	57 (35%)
T2	395 (43%)	278 (40%)	244 (41%)	83 (51%)
T3	90 (10%)	63 (9%)	47 (8%)	17 (11%)
T4	22 (2%)	14 (2%)	12 (2%)	4 (2%)
Not available	6 (1%)	4 (1%)	3 (<1%)	2 (1.23%)

Table 1. Clinicopathological characteristics by ER, PR, and HER2 status.

Objectives

1. **Predict Biomarker Status:** Predict ER, PR, and HER2 status using 529 DCE-MRI features and assess the need for additional pre-operative clinical features.
2. **Dataset and model comparison:** Determine which combination of model and subset of imaging features is best at classifying each of ER, PR, and HER2 status.
3. **Clustering:** Assess whether there are groups of patients with similar imaging characteristics or groups of features that contain similar information
4. **Feature Importance:** Evaluate which imaging features contribute the most to the model accuracy and which are important across multiple models.

Methods



Principle Component Analysis (PCA)

PCA transforms a high-dimensional dataset into a **different coordinate system** in which the first few coordinates (or components) explain most of the variation.

- Performed PCA on each of the **10 feature groups** which partition all 529 radiomics variables.
- The **loading factor** for a particular covariate represents its correlation with that component.
- For each of the ten feature groups, we selected:
  - (1) the covariates with the highest loading factor.
  - (2) the three covariates with the highest loading factors.
  - (3) enough covariates to explain 90% of the variance within PC1.
- We then selected both the raw data values and the scores for these subsets of covariates.

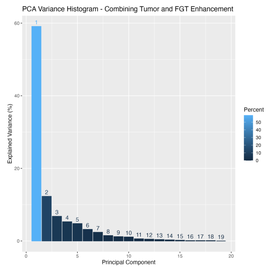


Fig. 2 Scree plot for one feature group.

Correlation Analysis

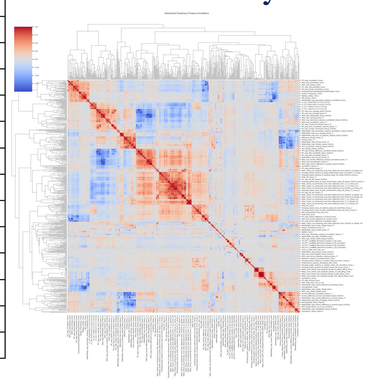


Fig. 3 Clustered correlation map on all imaging features.

We hypothesized many features within the radiomics set were highly correlated.

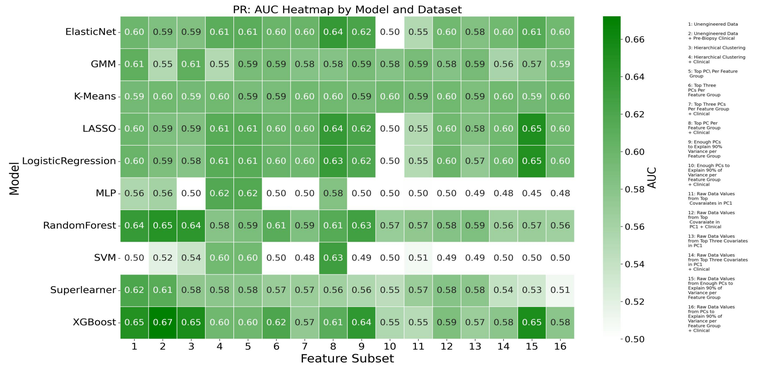
- Applied correlation-based **hierarchical clustering**.
- Computed absolute value of the **Pearson correlation coefficient** and transformed to a distance metric to perform agglomerative clustering.
- Dendrogram cut corresponding to a **correlation of 0.9** and then selected one representative from each cluster.
- After applying clustering, the number of features was reduced from **529 to 251.**

Results

Table 2. Best-performing model and dataset combinations with highest mean AUC across 5-fold cross-validation.

Target	Model	Dataset	AUC	Reference AUC
ER	XGBoost	Hierarchical Clustering + Clinical	0.661	0.649
PR	XGBoost	All Imaging + Clinical	0.672	0.622
HER2	Random Forest	Raw Data Values from PCs to Explain 90% of Variance per Feature Group + Clinical	0.655	0.500

Fig. 4 For one target, PR, the heatmap below shows performance across models (y-axis) and feature subsets (x-axis). We note that **adding clinical features** (2) outperforms the raw data (1) in some cases. We also note that using **the score from PC1** (8) outperforms the raw data value from the covariate with the highest loading factor in PC1 (11)



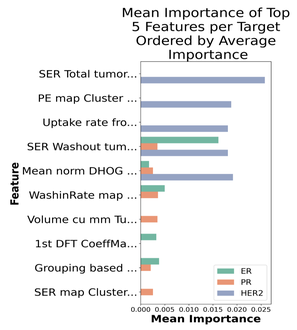


Fig. 5 Feature importance is determined by selecting one input feature, shuffling the values, and recording the change in AUC. Average importance per target was computed across all datasets. Consistently, features about **dye washin** were ranked in the top ten.

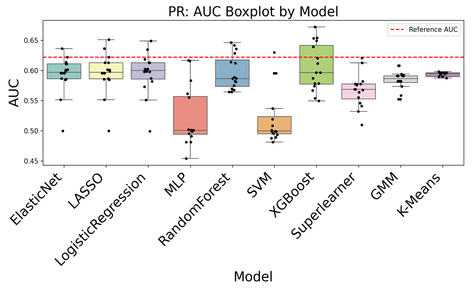


Fig. 6 Prediction performance on PR across models. Each dot represents a feature subset and model combination.

Conclusion

Summary

- Performing dimensionality reduction was able to increase the predictive performance in some settings
- Adding pre-biopsy clinical features can increase performance
- Training multiple models helped discover the best model and dataset for each target

Future Work

- Improve performance of MLP and SVM models by using raw imaging data
- Predict molecular subtype as a multinomial outcome

Thank you to our incredible advisors, Krithika Suresh, Nicholas Hartman, and Grant Carr. Thank you to Matt Zawistowski, Sabrina Olsson, Neo Kok, and Hannah Venera for making BDSI possible.

References

1. Saha, A., Harowicz, M. R., Grimm, L. J., Weng, J., Cain, E. H., Kim, C. E., Ghate, S. V., Walsh, R., & Mazurowski, M. A. (2021). **Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [Data set]**. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.e3sv-re93
2. Saha A, Harowicz MR, Grimm LJ, Kim CE, Ghate SV, Walsh R, Mazurowski MA. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. Br J Cancer. 2018 Aug;119(4):508-516. doi: 10.1038/s41416-018-0185-8. Epub 2018 Jul 23. PMID: 30033447; PMCID: PMC6134102.