Cancer Data Science

Thursday, July 24th, 2025

Background: Breast Cancer

- Very prevalent cancer
- High burden of disease
- Patients can experience recurrence or death
- Many variables can impact outcomes

24,240 will **die** from Breast Cancer in 2025

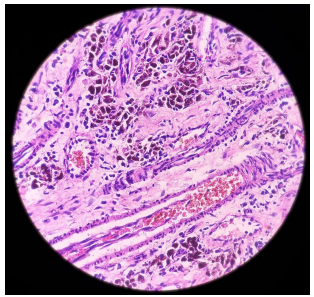
Source: https://www.nationalbreastcancer.org/breast-cancer-facts/

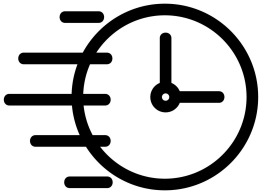
The Breast Cancer Care Process

(1) Using MRI data to predict molecular subtype

(2) Survival Analysis of Recurrence
(3) Assessing recurrence risk

MRI → Biopsy → Treatment → Follow-Up

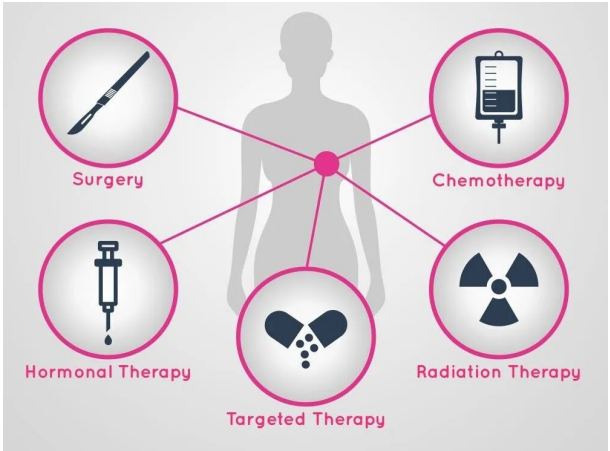




Nuances and Acknowledgements

- Our dataset and projects represent only a small snapshot of the journey patients go through.

- Treatment is not a “one size fits all”
  - Neoadjuvant treatments: before surgery
  - Adjuvant treatments: after surgery
  - Many factors go into deciding treatment such as cancer type, clinician expertise, and patient preferences.



Source: https://nhcancerclinics.com/cancer-types/breast-cancer/

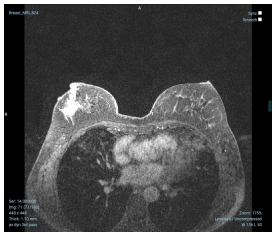
Background: Dynamic contrast-enhanced MRIs of breast cancer patients

- The Cancer Imaging Archive: Duke Breast Cancer MRI
- Dataset: “single-institutional, retrospective collection of 922 biopsy-confirmed invasive breast cancer patients, over a decade”
- Combination of 96 clinical and 529 imaging features

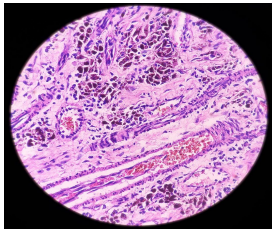
Using radiomics features to predict molecular subtype of breast cancer

Albert Kang, Lucy Malmud, Ritish Natesan, Kate Podrebarac

MRI

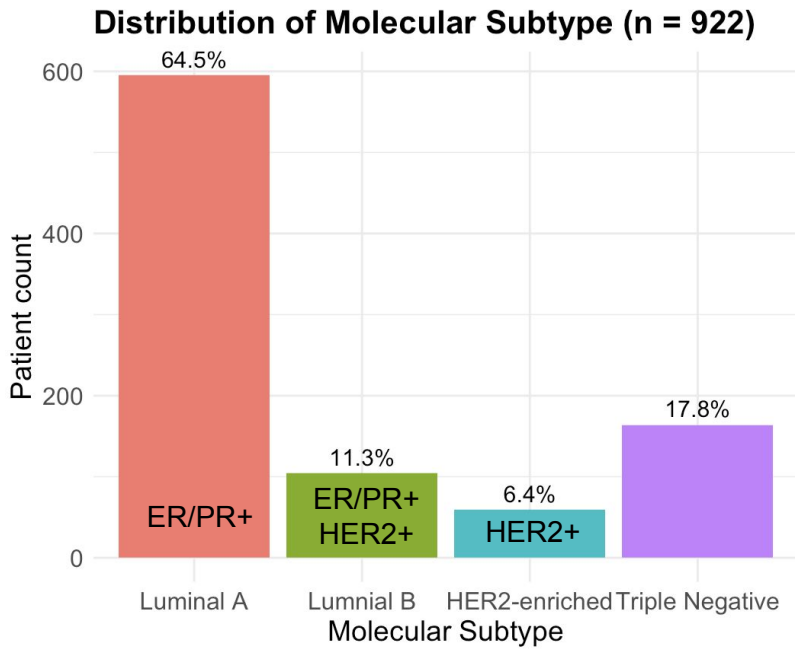


Biopsy



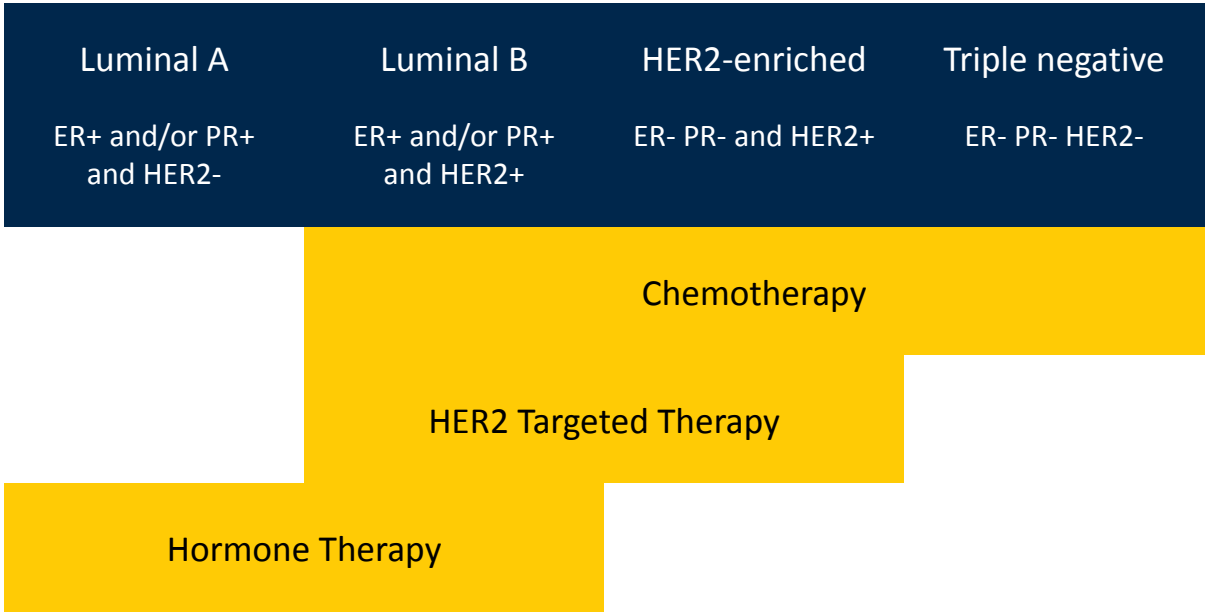
Background: Molecular Subtype

- Combination of Estrogen Receptor (ER), Progesterone Receptor (PR), and HER2 status
- Provides deeper insights into tumor pathology and progression than from histology
- Crucial for accurate diagnosis and precision treatment



Background: Clinical Importance

- Decrease need for invasive biopsy procedures
- Decrease cost of care
- Faster treatment decisions



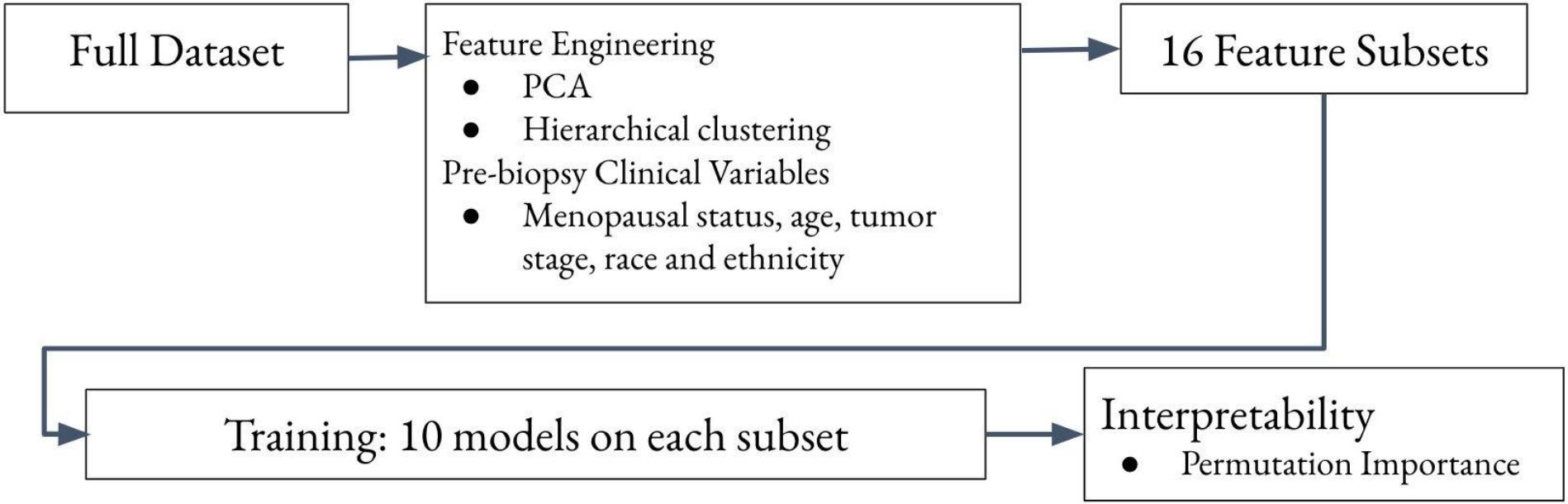
Background: Reference Paper + Dataset

- Radiomics = the field of study concerned with extracting large amounts of quantitative data from images
  - 529 radiomics features partitioned into 10 feature groups
- In a 2018 study[1] by Saha and colleagues, used random forest model to predict molecular subtype
- Can we do better?

Target	Reference AUC
ER	0.649
PR	0.622
HER2	0.500

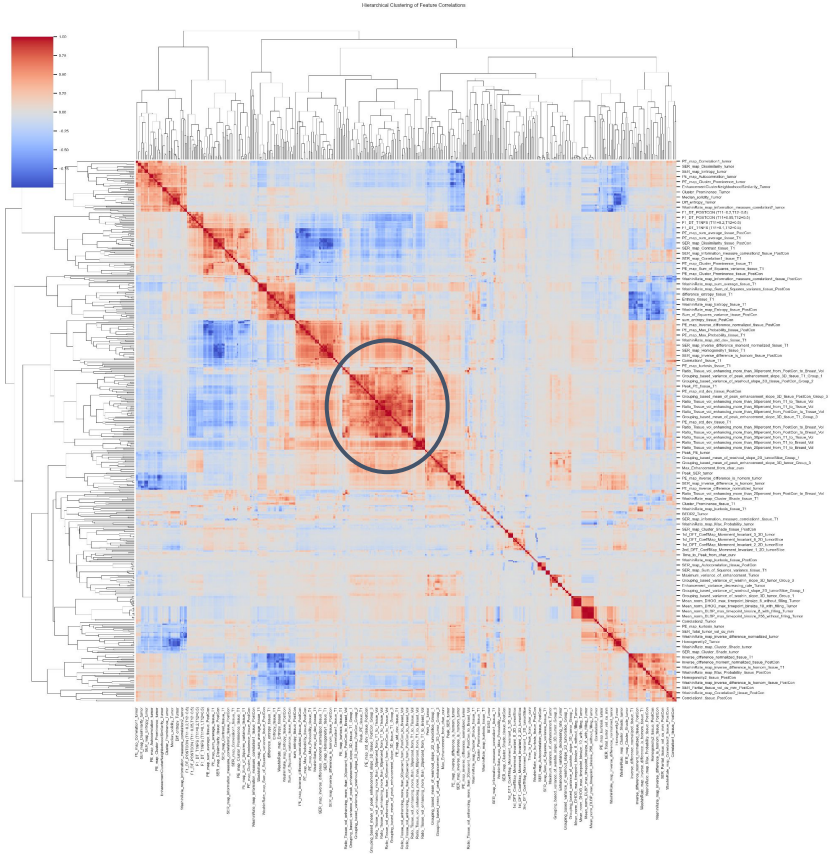
[1]A Machine Learning Approach to Radiogenomics of Breast Cancer: a study of 922 subjects and 529 DCE-MRI features

Our Process


Full Dataset
Feature Engineering
PCA
Hierarchical clustering
Pre-biopsy Clinical Variables
Menopausal status, age, tumor stage, race and ethnicity
16 Feature Subsets
Training: 10 models on each subset
Interpretability
Permutation Importance

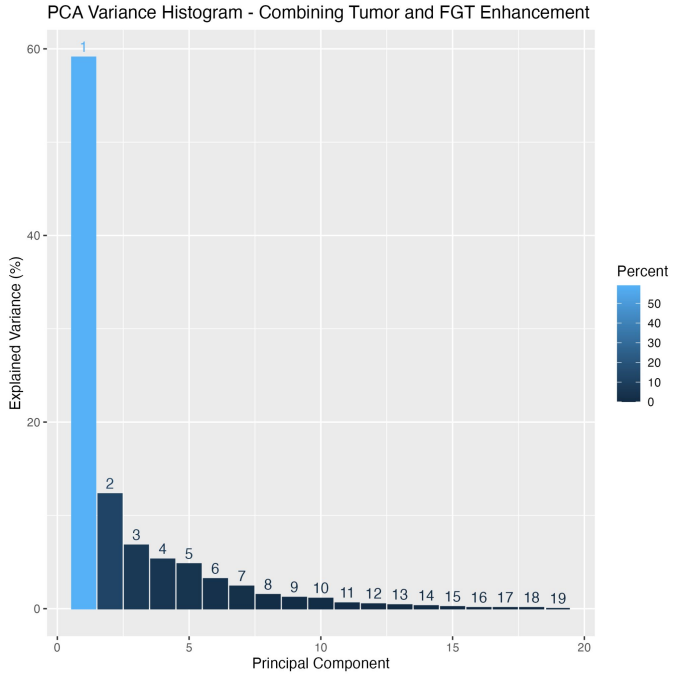
Feature Engineering: Hierarchical Clustering

- Correlation-based hierarchical clustering
- “Slice” the dendrogram at a level corresponding to correlation > .90
- Select one feature from each of these groups
- 529 → 251 covariates



Feature Engineering:
Principal Component Analysis (PCA)

- Transforms high dimensional data into a new coordinate system
- First principal component explains the most variation in the data
- Successive components explain less variation
- Each covariate has one loading factor per component
    - Represents how much that covariate contributes to the component



Training: Models

Generalized Linear Models:

- Logistic Regression
- LASSO
- Elastic Net

Clustering:

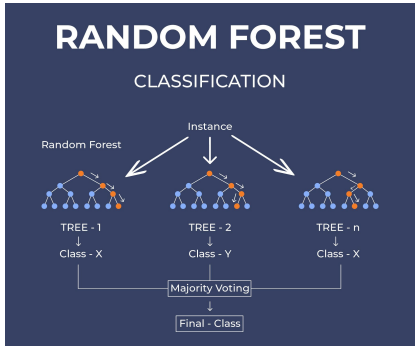
- K-Means
- Gaussian Mixture Model

Machine Learning Models:

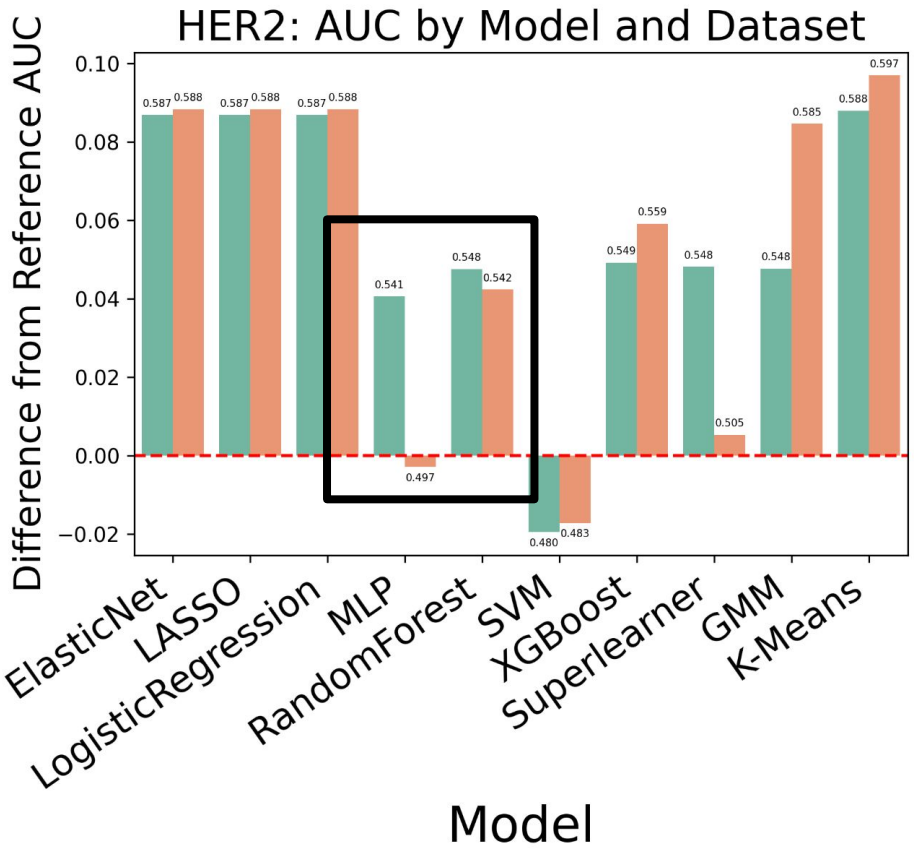
- Support Vector Machine (SVM)
- Multilayer Perceptron (MLP)

Ensemble Methods:

- Random Forest
- XGBoost
- Superlearner
  - Metalearner: Gradient Boosting Classifier
  - Base Learners: Random Forest, Decision Tree, Logistic Regression, Gradient Boosting Classifier



Source : https://diagramskiniatinguiccnf.z21.web.core.windows.net/random-forest-model-diagram.html



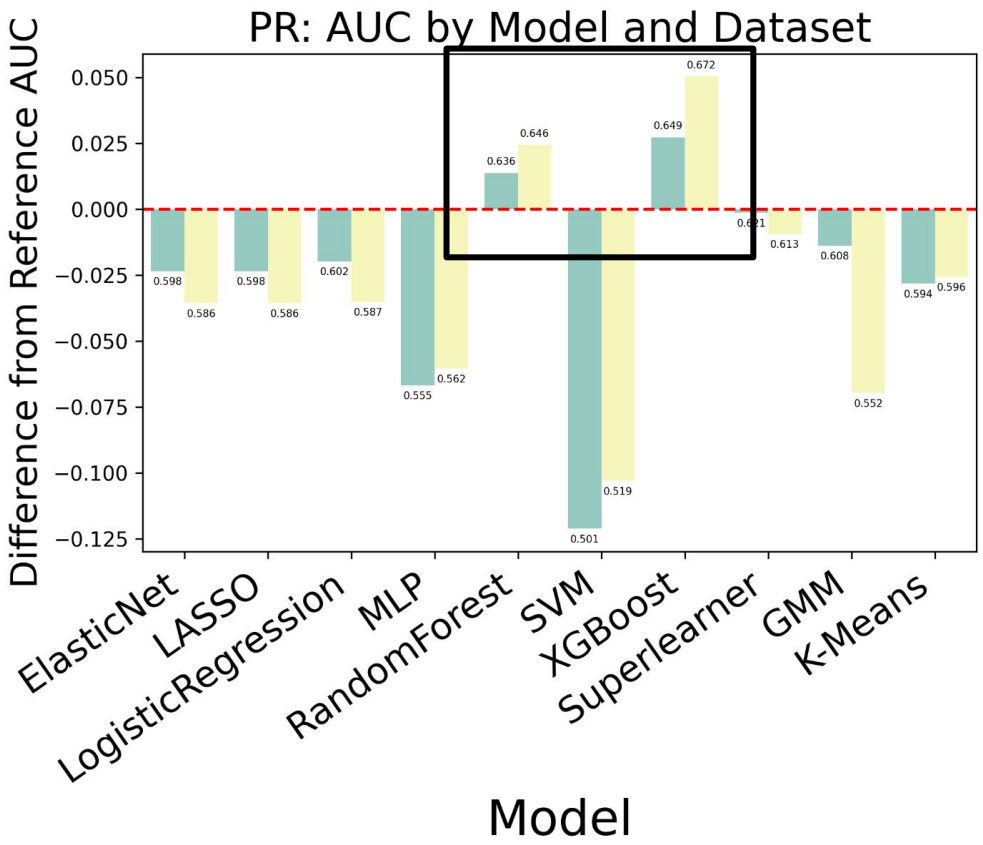
Feature Subsets

Top Three Principal Components Per Feature Group

Raw Data Values from Top Three Covariates in PC1

PCA scores yield better results than **raw data values.**

Improvement in performance by capturing more information in the same number of predictors.

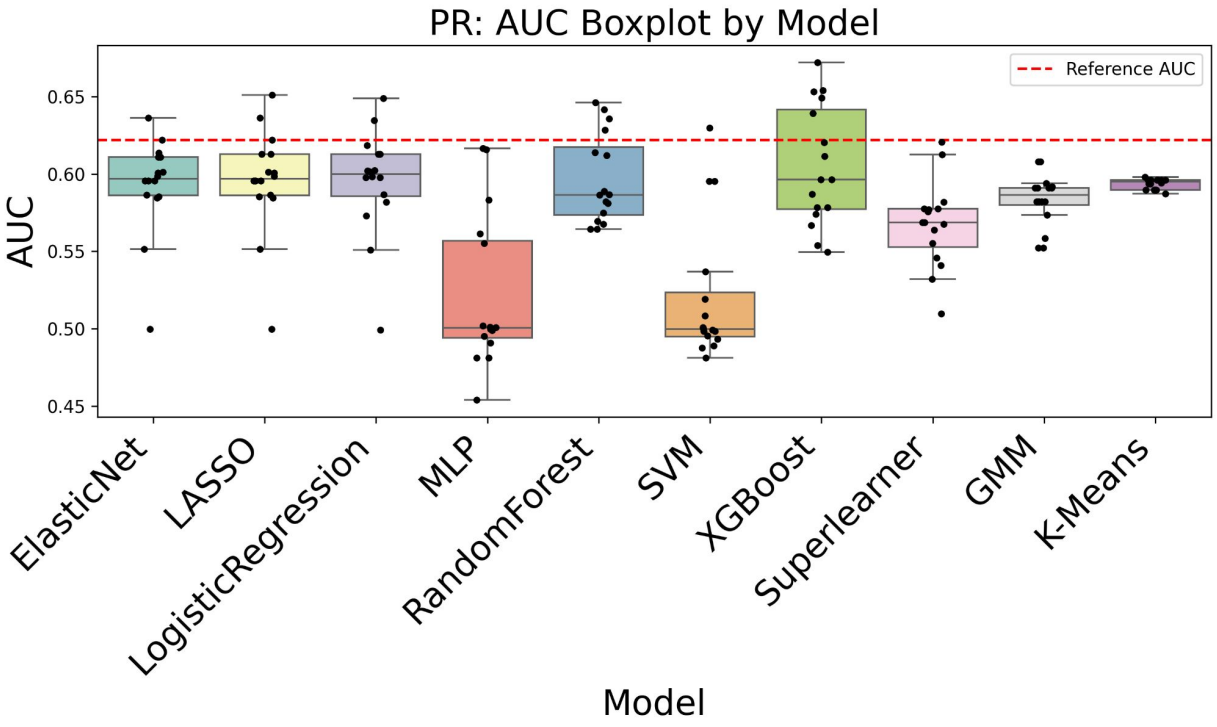


Feature Subsets

Unengineered Data

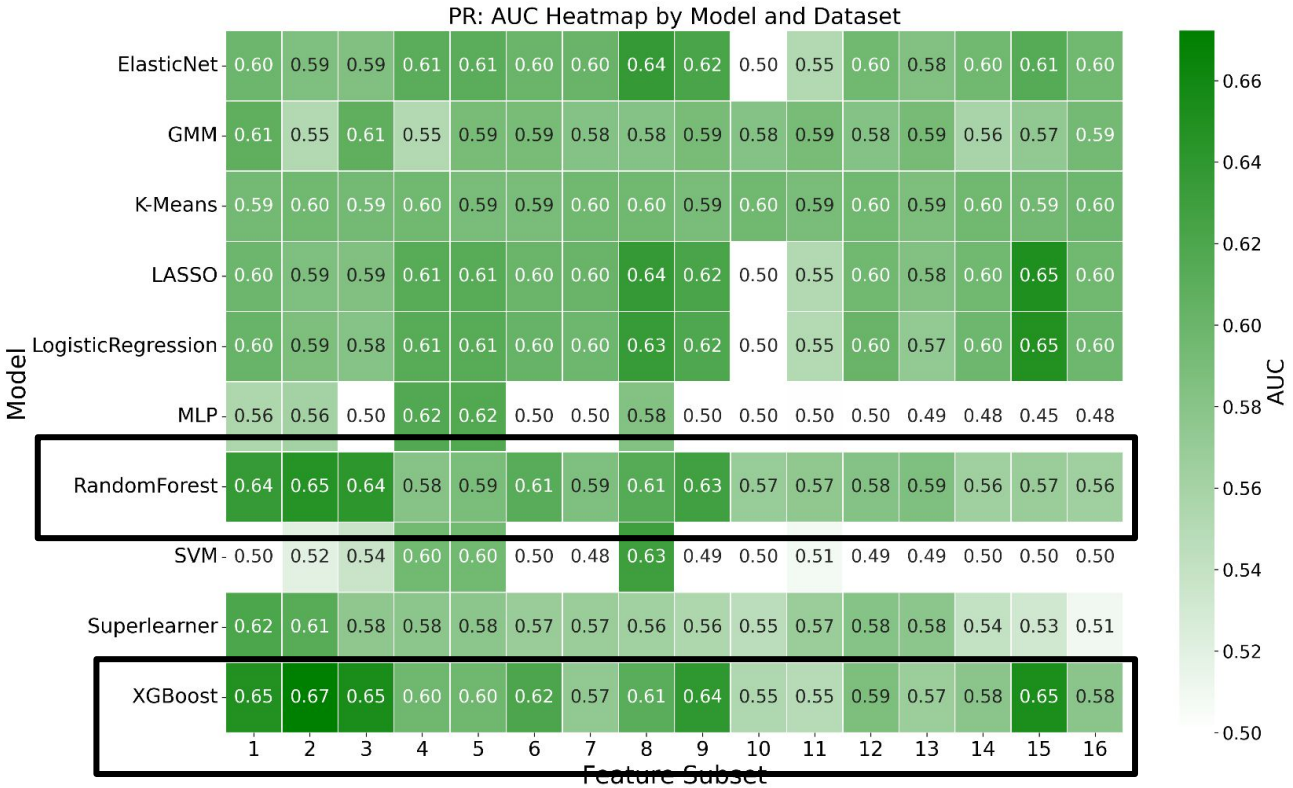
Unengineered Data + Pre-Biopsy Clinical

Adding **pre-biopsy clinical features** improves performance as compared to **the original imaging data** in some cases.



- **SVM** and **MLP** consistently perform poorly
- Linear models (**ElasticNet**, **LASSO**, **LogisticRegression**) comparable performance
- **XGBoost**, **RandomForest** consistently best

Feature Subsets



1: All Imaging Data

2: All Imaging
+ Pre-Biopsy Clinical

3: Hierarchical Clustering

4: Hierarchical Clustering
+ Clinical

5: Top PC\ Per Feature
Group

6: Top Three
PCs Per
Feature Group

7: Top Three PCs
s Per Feature Group
+ Clinical

8: Top PC Per
Feature Group
+ Clinical

9: Enough PCs
to Explain 90%
Variance per
Feature Group

10: Enough PCs to
Explain 90% of
Variance per
Feature Group
+ Clinical

11: Raw Data Values
from Top
Covaraiates in PC1

12: Raw Data Values
from Top
Covaraiate in
PC1 + Clinical

13: Raw Data Values
from Top Three Covariates
in PC1

14: Raw Data Values
from Top Three Covariates
in PC1
+ Clinical

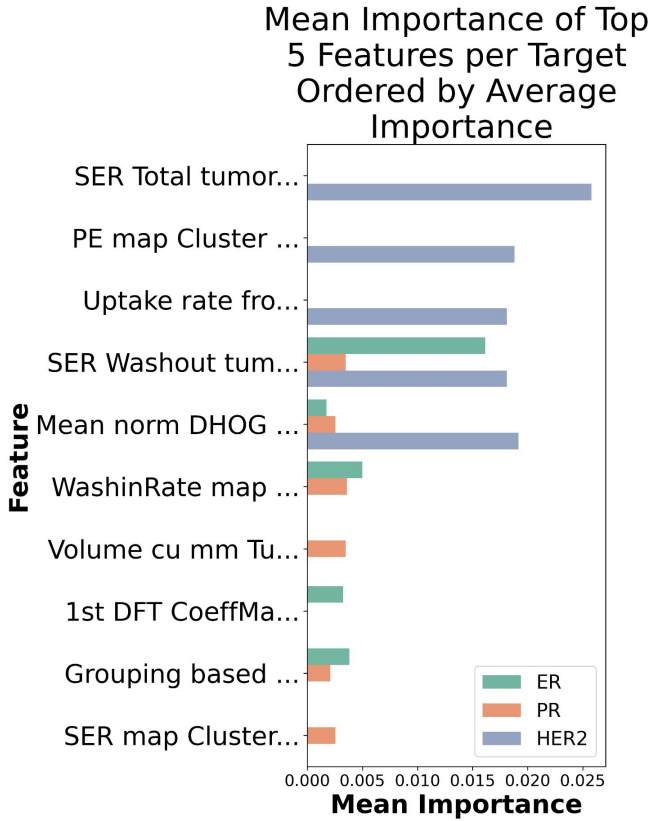
15: Raw Data Values
from Enough PCs to
Explain 90% of
Variance per
Feature Group

16: Raw Data Values
from PCs to
Explain 90% of
Variance per
Feature Group
+ Clinical

Results

Target	Best Model	Best Dataset	AUC*	Reference AUC
ER	XGBoost	Hierarchical Clustering + Clinical	0.661	0.649
PR	XGBoost	Unengineered + 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

Final model and dataset combination with highest *mean AUC across 5-fold cross-validation.



Feature Importance

- Permutation importance measures change in AUC when values of each covariate are permuted
- Insight into what covariates are most important in determining outcome

Important Features

Mean norm DHOG… (3)

- Quantification of texture

SER Washout tumor… (3) + WashinRate map… (2)

- Speed of contrast agent movement

Conclusion

Summary:

- Performing dimensionality reduction was able to increase prediction performance in some settings
- Adding pre-biopsy clinical predictors increased performance
- Training multiple models helped us find the best model for each target

Limitations:

- Imbalanced molecular subtype outcome
- Dataset is restricted to one hospital

Future:

- Using raw MRI data for SVM and MLP

Source: https://www.svchs.com/blog/southwest-virginia-community-health-systems-recognizes-the-month-of-october-as-breast-cancer-awareness-month/

Thank You!