Modeling results

1. Project overview

Our main goal is to train a classifier on the image features to predict the ER status from the MRI image features.

1.1 Input features.

The MRI-derived features can be categorized into several groups, each capturing different aspects of the breast tumor and surrounding tissue.

Feature Group	Number of Features	Description
Breast and FGT Volume Features	5	Quantifies breast and tissue volume and density.
Tumor Size and Morphology	10	Describes tumor size, shape, and structural characteristics.
FGT Enhancement	70	Measures enhancement patterns in FGT tissue.
Tumor Enhancement	40	Captures enhancement dynamics of the tumor.
FGT Enhancement Texture	120	Describes texture of FGT tissue based on enhancement patterns.
Tumor Enhancement Texture	60	Quantifies texture of the tumor based on enhancement patterns.
Combining Tumor and FGT Enhancement	18	Combines tumor and FGT enhancement features.
FGT Enhancement Variation	30	Measures variability in FGT enhancement patterns.
Tumor Enhancement Variation	40	Captures variability in tumor enhancement dynamics.
Tumor Enhancement Spatial Heterogeneity	4	Describes spatial heterogeneity of tumor enhancement.

- The FGT Enhancement Texture group has the highest number of features (120), indicating a strong focus on texture analysis in FGT tissue.
- The Tumor Enhancement and Tumor Enhancement Variation groups also have a significant number of features (40 each), highlighting the importance of dynamic and variable enhancement patterns in tumors.
- The Combining Tumor and FGT Enhancement group (18 features) suggests an effort to model interactions between tumor and surrounding tissue.

1.2 Target variable.

Estrogen receptor (ER) status is an important characteristic of breast tumors, as it affects prognosis and treatment decisions. There are 2 types of ER status:

1) ER-Positive Tumors:

- Definition: Tumors that have estrogen receptors on their cells, meaning they grow in response to estrogen.
- Prevalence: More common, accounting for about 70–80% of breast cancers.
- Growth Rate: Generally slower-growing compared to ER-negative tumors.
- Prognosis: Better overall prognosis because these tumors respond well to hormonal therapies.

2) ER-Negative Tumors

- Definition: Tumors that lack estrogen receptors and do not rely on estrogen for growth.
- Prevalence: Less common, about 20–30% of breast cancers.
- Growth Rate: Typically more aggressive and faster-growing.

- Prognosis: Worse than ER-positive tumors due to limited treatment options and higher recurrence rates.

Our target variable ER is imbalanced (25.6% ER-negative and 75.4% ER-positive) and the clinical importance of accurately predicting ER status for prognosis and treatment decisions, the choice of metric is critical.

2. Comprehensive Comparison of Model Results

2.1 Primary metrics to evaluate model's performance:

Given imbalanced dataset (25.6% ER-negative vs. 75.4% ER-positive) and the clinical significance of predicting Estrogen Receptor (ER) status for treatment decisions, it is suggested to use ***F1-Score, recall of ER-negative and AUC*** as our primary metrics, to ensure a balanced model, especially if the model needs high recall for ER-negative cases.

2.2 Comparison of the models

Below is a comparison of the models across the four approaches based on the provided metrics (accuracy, F1-score, and AUC):

Model Type	Definition	Accuracy (%)	F1-Score	AUC	Recall for ER-negative
Values_Lunit	**Supplied predictions in the file lunit_predictions.xlsx**	**76.0**	**0.861**	**0.618**	**0.08**
Approach 1 Logistic Regression	**1. Baseline (Approach 1)** Steps: ▪ This approach serves as an unprocessed benchmark, providing a baseline for comparison with more advanced methods. ▪ No data cleaning, feature selection, or outlier removal was performed. Datasets: ▪ Train: 617 observations, 529 features; ▪ Test: 305 observations, 529 features.	75.0	0.855	0.652	0.01
Approach 1 Random Forest		75.0	0.855	0.595	0.01
Approach 1 XGBoost		69.0	0.792	0.642	0.37
Approach 2 Logistic Regression	**2. Correlation-Based Selection & Outlier Removal (Approach 2)** Steps: ▪ Removed 359 highly correlated features (r > 0.8) to address multicollinearity and overfitting. ▪ Identified and removed 5 outliers from train set using the Isolation Forest algorithm. Datasets: ▪ Train: 612 observations, 170 features; ▪ Test: 305 observations, 170 features.	75.0	0.855	0.651	0.01
Approach 2 Random Forest		74.0	0.853	0.619	0.00
Approach 2 XGBoost		75.0	0.845	0.612	0.23
Approach 3 Logistic Regression	**3. Statistical Distribution-Based Selection (Approach 3)** Steps: ▪ Applied the two-sample Kolmogorov-Smirnov (KS) test to identify features with significant distributional differences between ER status groups. 380 insignificant features are removed. Datasets:	75.0	0.855	0.660	0.01
Approach 3 Random Forest		75.0	0.856	0.600	0.05
Approach 3 XGBoost		**69.0**	**0.790**	**0.649**	**0.44**

	▪ Train: 617 observations, 149 features; ▪ Test: 305 observations, 149 features.				
Approach 4 Logistic Regression	**4. Non-Linear Dimensionality Reduction + Robust Scaling (Approach 4)** Steps: ▪ Applied Robust Scaling to handle outliers and non-normal distributions. ▪ Used UMAP for non-linear dimensionality reduction, achieving a trustworthiness score of 0.865. 100 UMAP components are obtained. Datasets: ▪ Train: 617 observations, 100 UMAP components; ▪ Test: 305 observations, 100 UMAP components.	26.0	0.000	0.508	1.00
Approach 4 Random Forest		74.0	0.853	0.482	0.00
Approach 4 XGBoost		74.0	0.853	0.517	0.00

The most balanced model across key metrics is ***XGBoost in Approach 3***, as its two metrics surpass those of the Supplied predictions by Lunit. It attains the highest recall for ER-negative at 0.44 and a solid AUC of 0.649. However, its F1-score is comparatively lower at 0.79. ***However, these results may vary depending on the input parameters of the model, such as threshold selection, observation weighting, and other tuning factors.***

2.3 Classification report for the test set across ER status for Values_Lunit and XGBoost in Approach 3

1) Supplied predictions by lunit

category	precision	recall	f1-score	count
0	0.860	0.080	0.140	78
1	0.760	1.000	0.860	227

1. Class "0" (ER-negative)

- Precision: 0.86 → When the model predicts ER-negative (0), it is correct 86% of the time.
- ***Recall: 0.08 → The model captures only 8% of all true ER-negative cases, meaning it misses most of them.***

2. Class "1" (ER-positive)

- Precision: 0.76 → When the model predicts ER-positive (1), it is correct 76% of the time.
- Recall: 1.0 → The model correctly identifies all ER-positive cases.

Conclusions

1. ***Severe Class Imbalance Issue:*** The model is heavily biased towards ER-positive cases, performing exceptionally well at identifying them but very poorly at detecting ER-negative cases.
2. ***High Recall for ER-positive but Poor for ER-negative:*** The model predicts almost all samples as ER-positive (class "1"), leading to perfect recall (1.0) for this class but almost no ability to detect ER-negative cases.
3. ***Misleading Precision for ER-negative:*** Although the precision of 0.86 looks good, the model rarely predicts ER-negative, leading to high false negatives.

2) XGBoost in Approach 3

category	precision	recall	f1-score	count
0	0.400	0.440	0.420	78
1	0.800	0.780	0.790	227

1. Class "0" (ER-negative)

- **Precision:** 0.40 → When the model predicts ER-negative (0), it is correct **40%** of the time.

- ***Recall: 0.44 → The model captures 44% of all true ER-negative cases, meaning it misses 56% of them.***
- **F1-score:** 0.42 → A low balance between precision and recall, indicating poor overall performance for this class.

2. Class "1" (ER-positive)

- **Precision:** 0.80 → When the model predicts ER-positive (1), it is correct **80%** of the time.
- **Recall:** 0.78 → The model correctly identifies **78%** of all ER-positive cases.
- **F1-score:** 0.79 → A strong balance between precision and recall, meaning the model performs well for this class.

2.3 Summary of comparison

1) Best Performing Models:

- The most balanced model across key metrics is ***XGBoost in Approach 3***, as its two metrics surpass those of the Supplied predictions by Lunit. It attains the highest recall for ER-negative at 0.44 and a solid AUC of 0.649. However, its F1-score is comparatively lower at 0.79. ***However, these results may vary depending on the input parameters of the model, such as threshold selection, observation weighting, and other tuning factors.***

2) Approach Comparison:

• ***Approach 3 (Statistical Distribution-Based Selection)*** consistently performed well across all models, with the highest AUC for logistic regression and the highest recall for ER-negative for XGBoost.

• Approach 1 (Baseline) performed surprisingly well, indicating that the raw data contains strong predictive signals.

3) Model Type Comparison:

• XGBoost generally outperformed logistic regression and random forest in terms of recall for ER-negative.

• Logistic Regression performed well in terms of AUC, especially in Approach 3, indicating better class separation.

• Approach 4 (Non-Linear Dimensionality Reduction + Robust Scaling) performed poorly, especially for logistic regression, suggesting that UMAP components may not have captured the relevant information for this model.

3. XGBoost using Statistical Distribution-Based Selection.

The most balanced model was ***XGBoost in Approach 3*** using Statistical Distribution-Based Selection. The possible factors that improved the model's performance:

1)Effectiveness of the Kolmogorov-Smirnov (KS) Test for Feature Selection.

2) Multicollinearity and Overfitting Avoidance.

Approach 3 removes 380 features that do not have statistically significant differences in distribution between ER-positive and ER-negative groups. This means the retained features are more informative and better separated between the two classes.

3) Using of built-in scale_pos_weight (the ratio of ER-negative to ER-positive) parameter, making it more effective for imbalanced classification tasks.

4) Takes into account non-linear relationships and interactions through boosted decision trees, making it more effective for structured data.

4. Conclusion

- **The results revealed moderate associations between MRI imaging features and ER status.**
- **MRI imaging features by themselves lack the statistical strength needed to reliably predict ER status.**