UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Radiogenomic Prediction of Breast Cancer Subtypes Using the TCGA Dataset

Lucas Fayolle & Jose Valero

Biomedical Data Science
(ETSINF - UPV)

Course 2024/2025

Table of Contents

Table of Contents

Motivation

- Breast cancer is one of the leading causes of mortality in women [1].
- Identifying molecular subtypes:
  - Luminal A
  - Luminal B
  - HER2-enriched
  - Basal-like

  is key for personalized treatments [2].
- Current procedures (biopsies, genomic assays) are invasive, expensive, and time-consuming.
- **Radiomics**: a non-invasive method that extracts features from MRI images to capture phenotypic traits of the tumor.

Objectives

- **General Objective:** Develop a non-invasive approach based on MRI images to determine molecular subtypes.

- **Specific Objectives:**
  - Utilization of the provided radiomic features.
  - Incorporation of clinical data.
  - Integration of multigenic assay data.

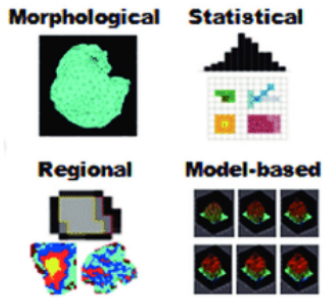
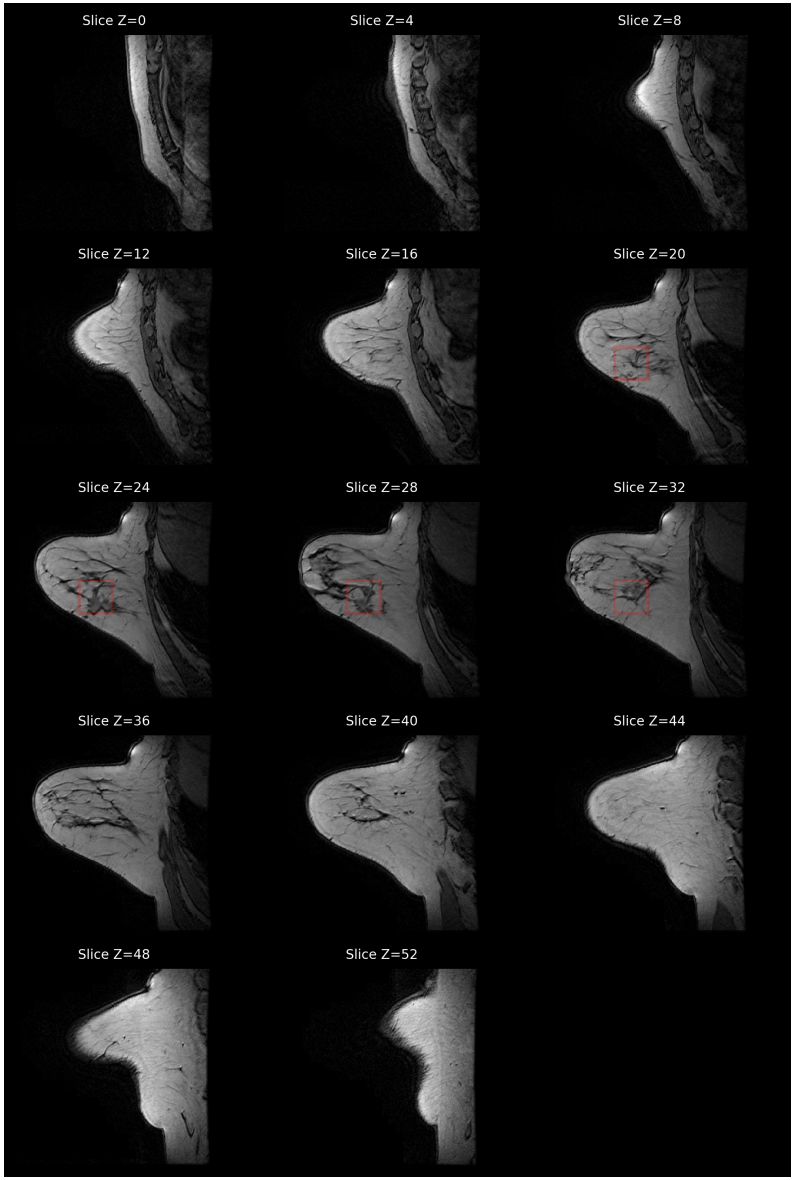


Table of Contents

Dataset description

Data obtained from TCGA Breast Radiogenomics [3]:

- **Radiomic Features:**
  - 36 quantitative features derived from MRI images.
  - Shape, texture, etc.
- **Multigenic Assay Results:** Genomic scores associated with breast cancer prognosis.
- **Clinical Data:** Patient demographics, tumor characteristics, and treatment.



Data preprocessing I

Preprocessing Steps:

- **Selection of complete instances:** Only instances with complete information across all datasets are included.
- **Removal of the “Normal” category:** Instances labeled as ”Normal” in the `Pam50.Call` variable are excluded, as they are not relevant for molecular subtype classification.

Data preprocessing II

Result: After preprocessing, the dataset is reduced from 84 to 76 instances.

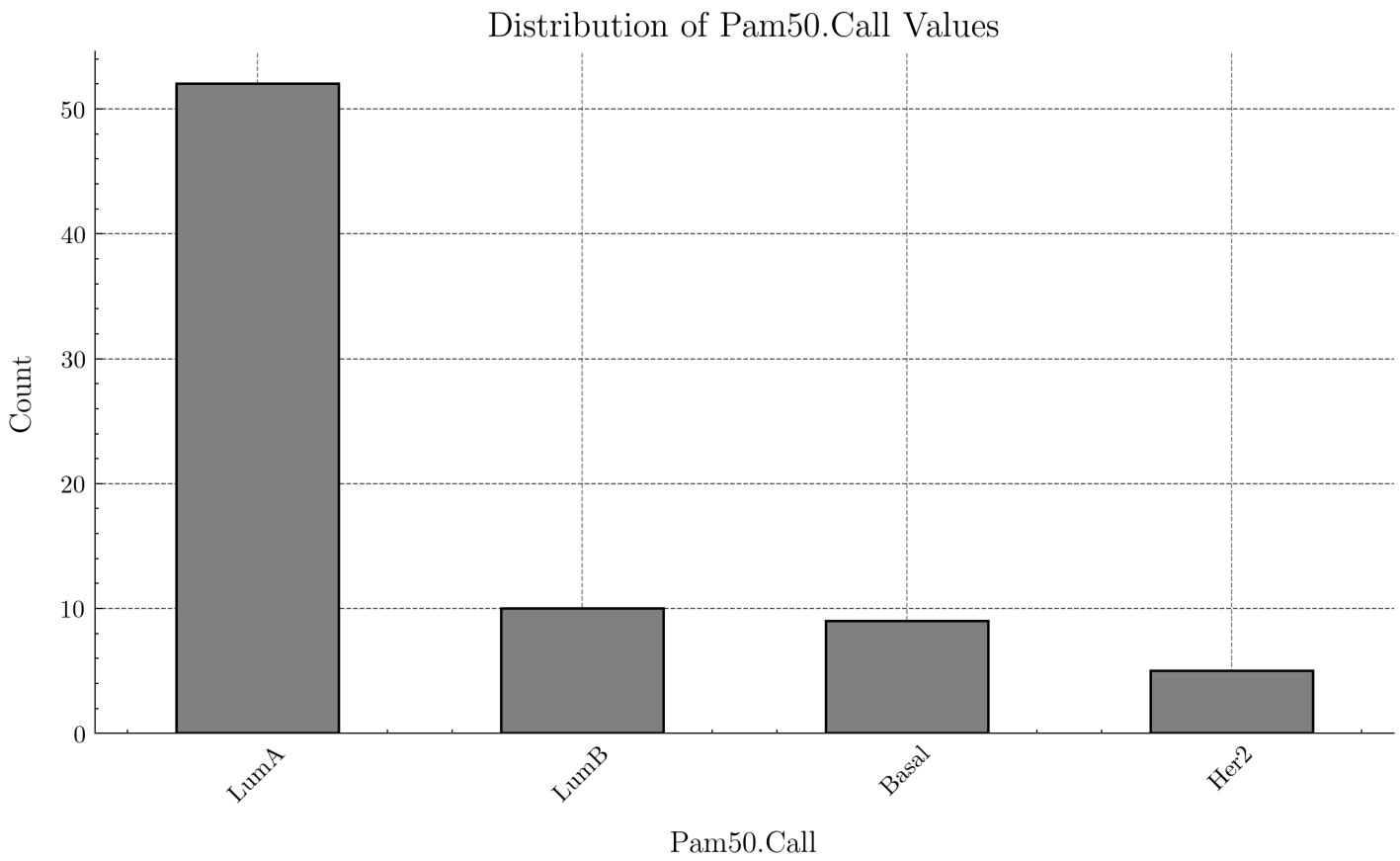


Figure: Distribution of the `Pam50.Call` variable.

Feature selection

Based on exploratory analysis and medical considerations, the following variables were selected from each dataset:

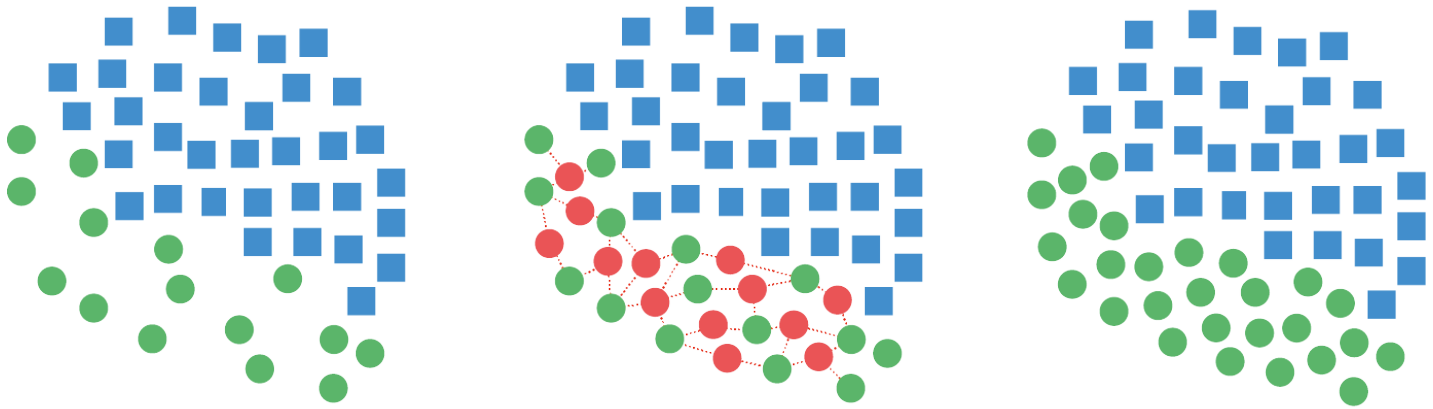
- **Clinical Data:**
  - **Age at diagnosis**
  - **Cancer stage and tumor size**
  - **Number of affected lymph nodes:** Reflects metastatic spread and indicates the severity of the disease [4].
  - **Hormone receptor status (estrogen and progesterone):** Essential for distinguishing Luminal subtypes [5], but also highly correlated with the target variable.
- **Multigenic Assay Data:**
  - **GHI RS Score:** Continuous score from the Oncotype DX assay measuring recurrence risk [6].
  - **Correlation with good outcome signature:** Indicates how strongly the sample correlates with a favorable prognosis gene profile from the MammaPrint assay [7].
  - **Proliferation-related gene expression:** Represents the average expression of genes linked to cell proliferation.

Table of Contents

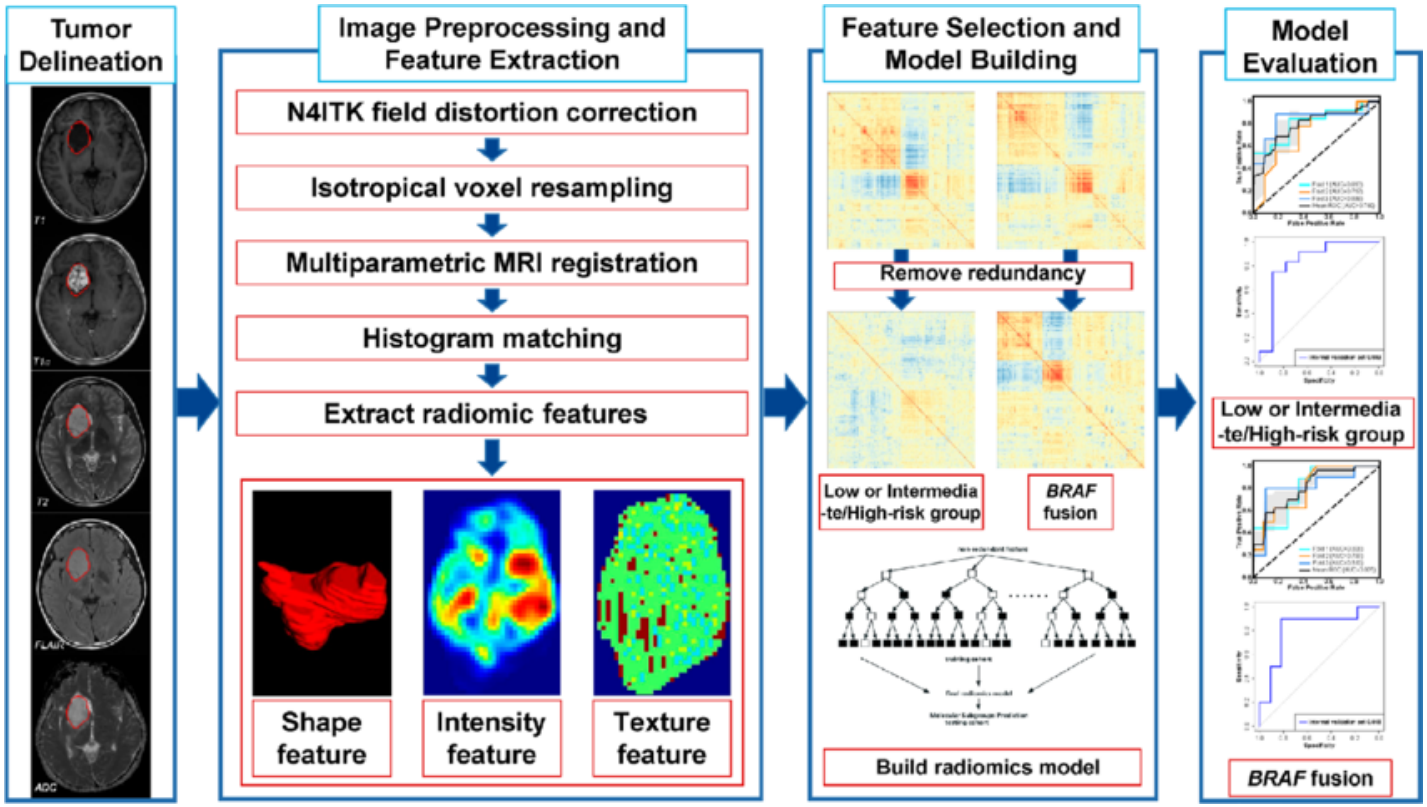
Data scaling and resampling

- **Data Scaling:**
  - The radiomic features were standardized using `StandardScaler` to ensure a mean of zero and a standard deviation of one.
  - This prevents features with larger magnitudes from disproportionately influencing the model's performance.

- **Data Resampling:**
  - **Undersampling:** The Luminal A class was reduced to 30 instances to decrease its dominance in the dataset.
  - **SMOTE (Synthetic Minority Oversampling Technique):** New synthetic samples were generated for minority classes to improve class balance.



Radiomic features-based model - Pipeline



In this project, we start with the given radiomic features, so feature selection and model training are yet to be performed.

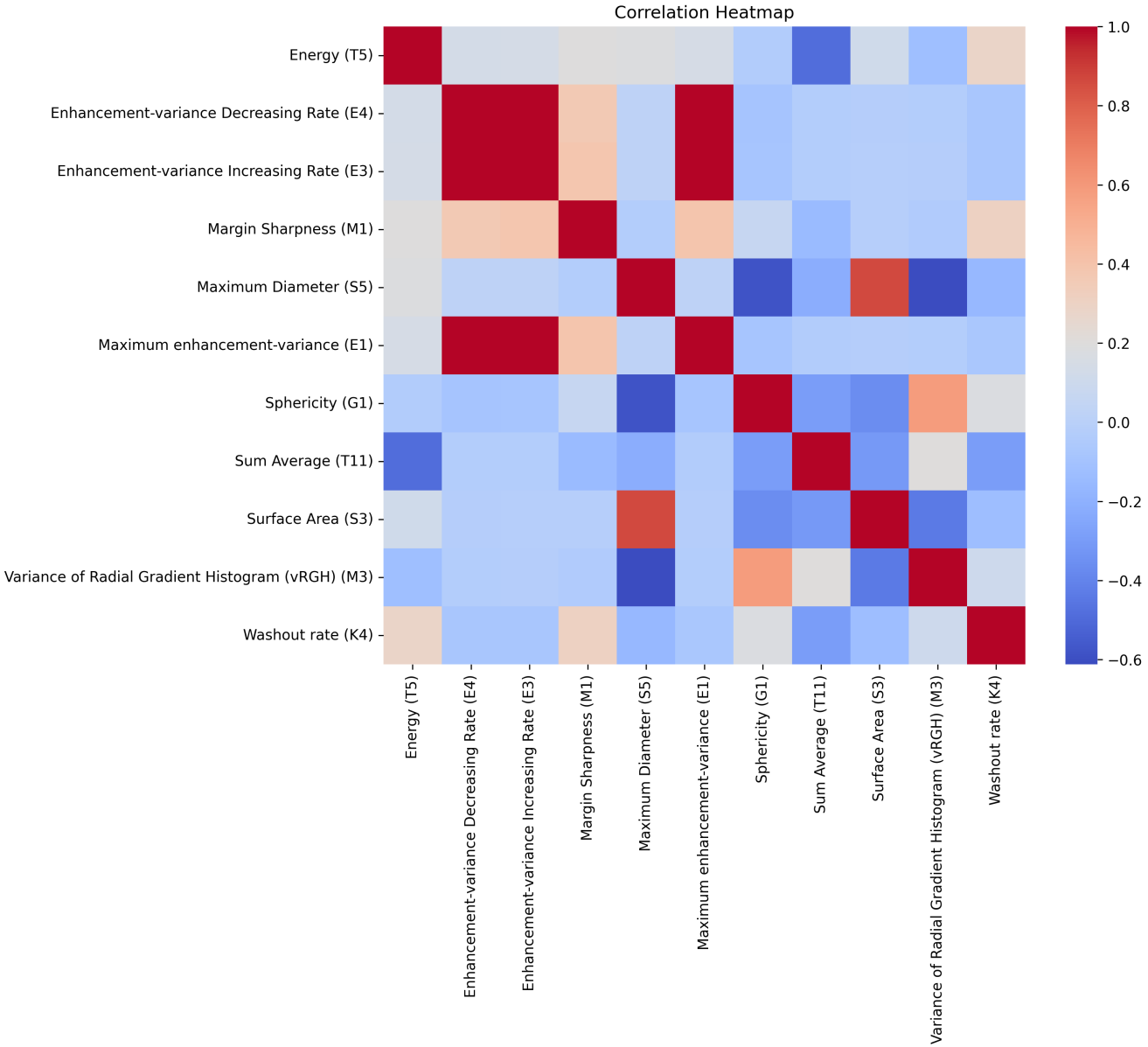
Radiomic features-based model - Feature Selection I

Feature selection steps:

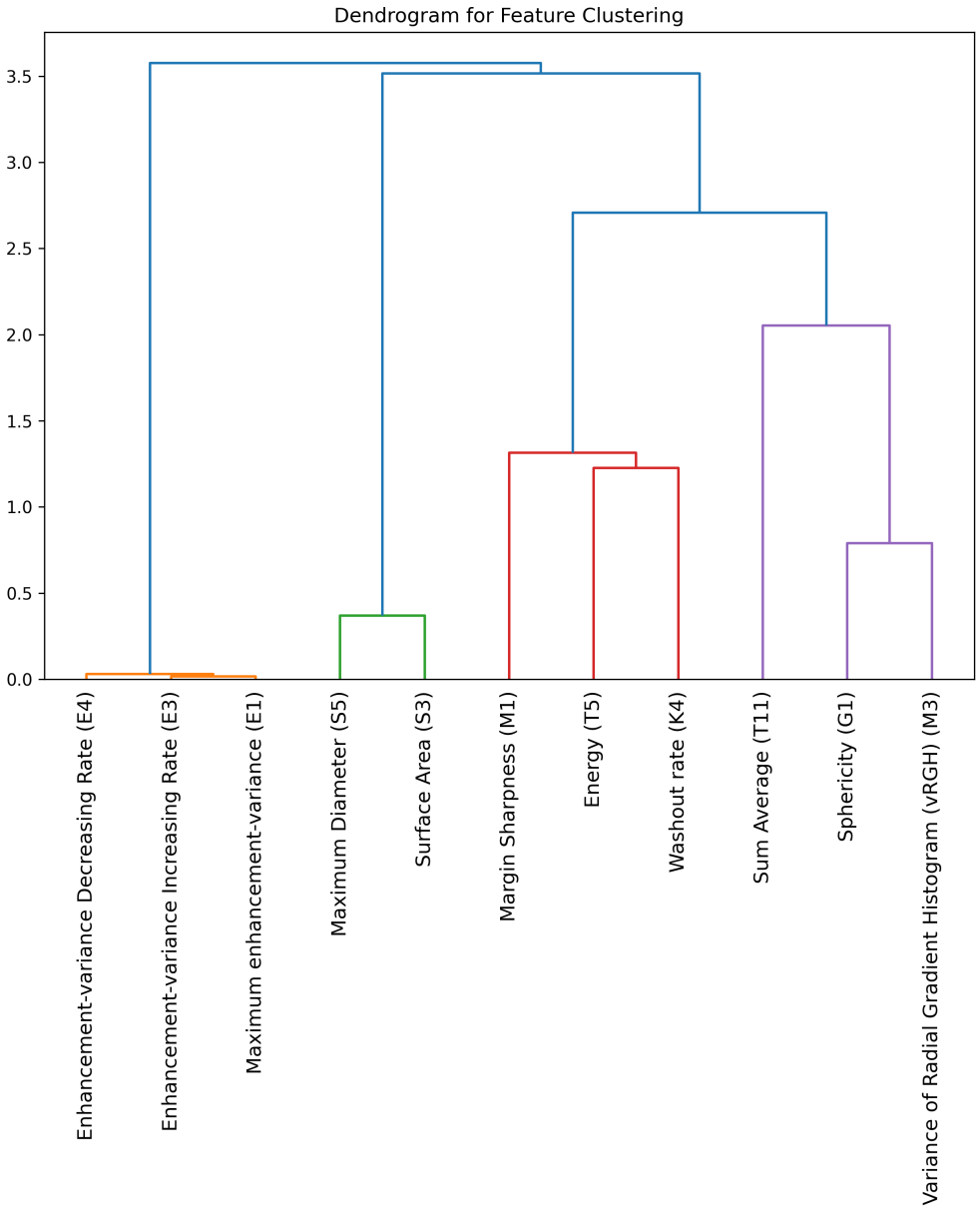
- **Boruta with Gradient Boosting:**
  - Applied to identify the most relevant radiomic features.
  - Reduced the feature set from 36 to 11 variables.

- **Redundancy Analysis:**
  - **Correlation Heatmap (15):** Displayed pairwise correlations between the selected features, highlighting highly correlated pairs, indicating redundancy.
  - **Hierarchical Clustering (16):** Visualized using a dendrogram to group correlated features into clusters based on correlation distances.

Redundancy Analysis - Correlation Heatmap



Redundancy Analysis - Dendrogram



Radiomic features-based model - Feature Sel. II

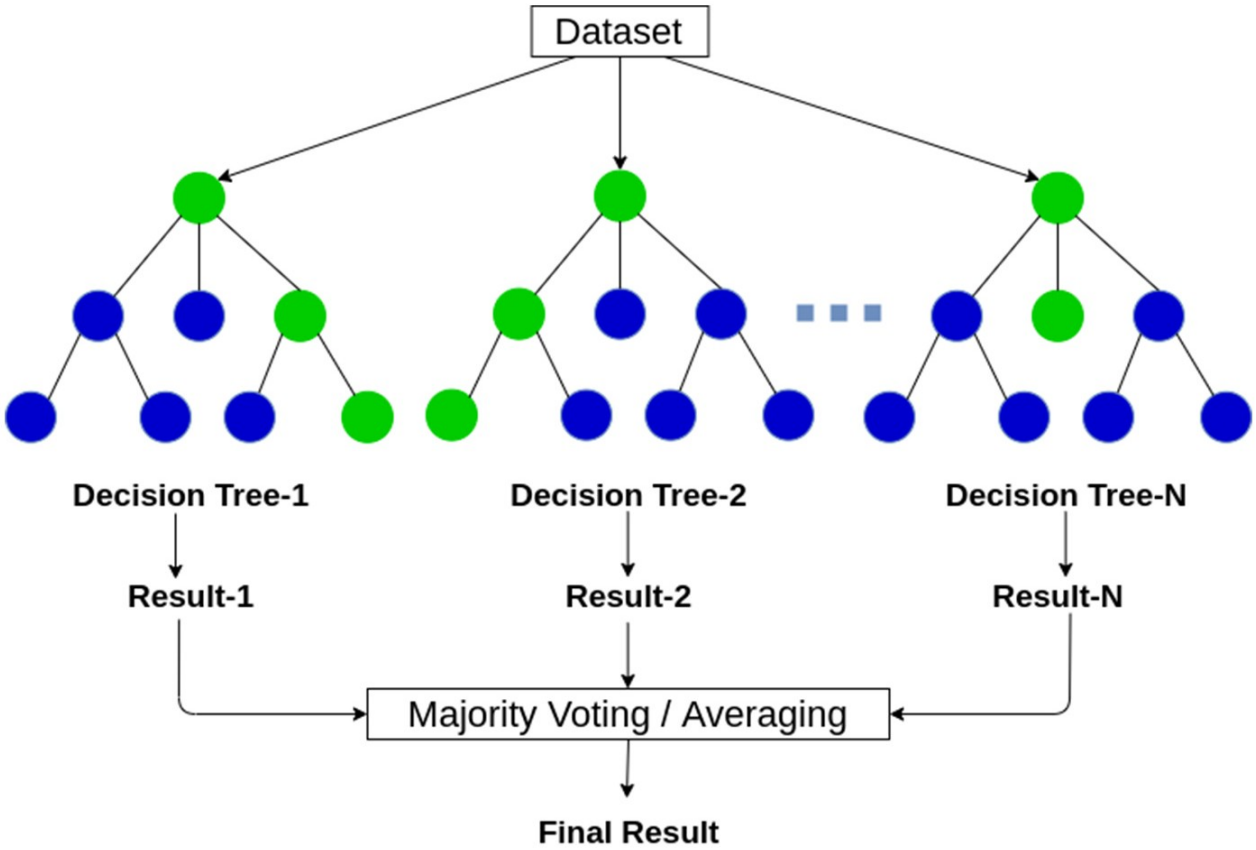
After testing different numbers of clusters, selecting 3 clusters provided a good balance between reducing redundancy and maintaining model performance.

Final representative features:

- **Margin Sharpness (M1):** Describes the abruptness of intensity changes at the tumor's boundary, indicating how clearly the tumor is demarcated from surrounding tissue.
- **Maximum Enhancement-Variance (E1):** Measures the variance in the enhancement signal across the most enhancing regions, reflecting vascular heterogeneity.
- **Surface Area (S3):** Represents the surface area of the tumor boundary, indicating tumor size and shape complexity.

Radiomic features-based model - Model training

Once the features are selected, we train a Random Forest (RF) model.



Radiomic model with aditional data

- **Radiomic Model with Clinical Data:** Adds to the previous radiomic features the following variables:
  - **Full Clinical Data Model:**
    - Age at diagnosis
    - Cancer stage and tumor size
    - Number of affected lymph nodes
    - Hormone receptor status
  - **Reduced Clinical Data Model:** This model excludes the hormone receptor status variables to avoid an overly optimistic evaluation.

- **Radiomic Model with Multigenic Assays:** Adds to the previous radiomic features the following variables:
  - GHI RS Score
  - Correlation with good outcome signature
  - Proliferation-related gene expression

Table of Contents

Comparison of model performance

Model	Accuracy	F1-score
Only Radiomic	0.62	0.45
Radiomic + Reduced clinical data	0.56	0.42
Radiomic + Full clinical data	0.75	**0.78**
Radiomic + Multigenic	0.81	0.74
All (Radiomic + Full clinical + Multigenic)	0.75	0.60
Only full clinical data	0.67	**0.40**
Only Multigenic	0.75	**0.40**

Table: Performance results for the different models based on accuracy and macro F1-score.

Detailed performance of the best model

Due to the strong performance of the Radiomic + Full Clinical Data model, we consider it relevant to explore its performance using the confusion matrix from the test set.

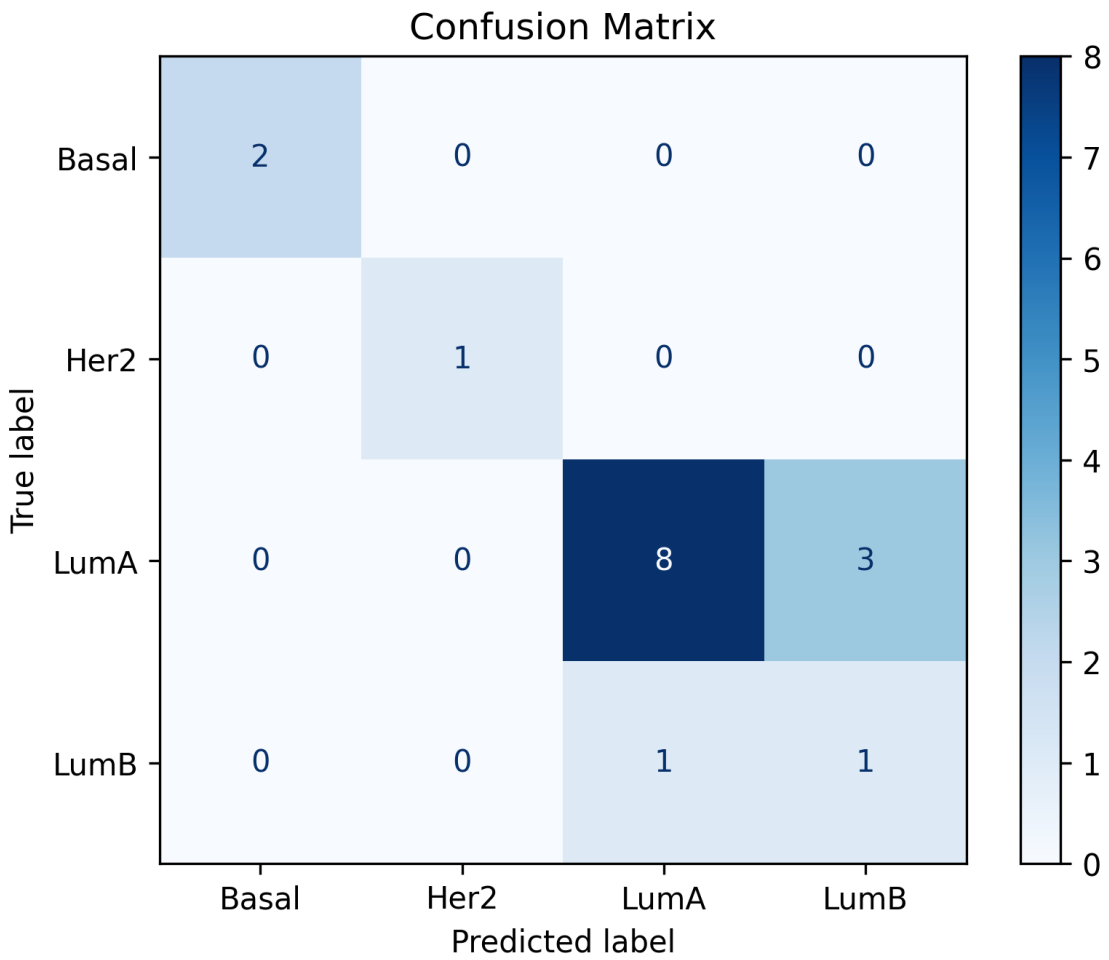


Table of Contents

Key findings

- Combining radiomic, clinical, and genomic data enhances the prediction of molecular subtypes of breast cancer.
- The **Radiomic + Full Clinical Data** model achieved the highest performance, with an F1-score of 0.78, highlighting the importance of hormone receptor status and other clinical variables in subtype classification.
- The inclusion of multigenic assays improved the model's performance, with the **Radiomic + Multigenic model** achieving an F1-score of 0.74, demonstrating the complementarity between genomic data and radiomic features.

Study limitations and suggestions for future work

Study Limitations

- Small dataset size.
- Limited computational resources.

Suggestions for Future Work

- Increased data collection.
- Better data organization and integration.
- Extraction of radiomic features from raw images.
- Hyperparameter tuning.

References

[1] World Health Organization. ”Cancer: Fact Sheets.” Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed: Jan. 5, 2025].

[2] S. Komen Foundation. ”Molecular Subtypes of Breast Cancer.” Available: https://www.komen.org/breast-cancer/diagnosis/molecular-subtypes. [Accessed: Jan. 4, 2025].

[3] E. Morris, E. Burnside, G. Whitman, M. Zuley, E. Bonaccio, M. Ganott, E. Sutton, J. Net, K. Brandt, H. Li, K. Drukker, C. Perou, and M. L. Giger, ”Using Computer-extracted Image Phenotypes from Tumors on Breast MRI to Predict Stage [Data set],” *The Cancer Imaging Archive*, 2014. Available: https://doi.org/10.7937/K9/TCIA.2014.8SIPIY6G.

[4] S. Komen Foundation. ”Factors that Affect Prognosis: Lymph Node Status.” Available: https://www.komen.org/breast-cancer/diagnosis/factors-that-affect-prognosis/lymph-node-status. [Accessed: Jan. 5, 2025].

[5] Wikipedia. ”Breast cancer classification.” Available: https://en.wikipedia.org/wiki/Breast_cancer_classification. [Accessed: Jan. 5, 2025].

[6] Y. Y. Syed, ”Oncotype DX Breast Recurrence Score®: A Review of its Use in Early-Stage Breast Cancer,” *Mol. Diagn. Ther.*, vol. 24, pp. 621–632, 2020. doi: 10.1007/s40291-020-00482-7.

[7] Wikipedia. ”MammaPrint.” Available: https://en.wikipedia.org/wiki/MammaPrint. [Accessed: Jan. 5, 2025].