Radiogenomic Prediction of Breast Cancer Subtypes Using the TCGA Dataset

Lucas Fayolle
lfayoll@etsinf.upv.es

Jose Valero Sanchis
jvalsan@etsinf.upv.es

***Abstract*—Breast cancer subtype classification is essential for guiding personalized treatments. While molecular profiling provides valuable insights, these tests can be costly and invasive. Radiomics offers a non-invasive alternative by extracting quantitative features from MRI images that capture tumor characteristics.**

This study develops machine learning models to predict breast cancer molecular subtypes using radiomic features and evaluates the impact of incorporating clinical and genomic data. The model combining radiomic features + clinical data (including hormone receptor status) achieved the highest performance, with a macro F1-score of 0.78. Additionally, the integration of genomic data improved model performance, demonstrating the complementarity between radiomic and genomic features. These findings highlight the potential of combining imaging, clinical, and genomic data to improve subtype classification.

***Index Terms*—Breast cancer, radiomics, machine learning, molecular subtypes, genomic assays, clinical data, MRI.**

I. Introduction

A. Background and motivation

Breast cancer is one of the most common cancers worldwide and a leading cause of mortality among women [1]. Molecular profiling has identified subtypes such as Luminal A, Luminal B, HER2-enriched, and Basal-like, which are essential for guiding personalized treatments [2]. However, determining these subtypes often requires invasive procedures, such as biopsies and genomic assays, which can be costly and time-consuming.

Radiomics offers a non-invasive approach by extracting quantitative features from MRI images that can capture tumor phenotypic traits linked to molecular profiles. Integrating radiomic features with clinical and genomic data can enhance machine learning models for breast cancer subtype prediction. [3]

This project aims to develop a classifier that predicts molecular subtypes based on MRI-derived radiomic features and evaluates the impact of incorporating clinical and genomic data to improve performance.

B. Project objectives

The primary objective of this project is to develop a machine learning classifier capable of predicting the molecular subtypes of breast cancer (Luminal A, Luminal B, HER2-enriched, Basal-like) using radiomic features extracted from MRI images.

To achieve this, the following specific objectives are proposed:

- **Utilization of the provided radiomic features**. A preprocessing pipeline will be applied to the radiomic data provided, incorporating feature selection methods, such as Boruta, to refine the input features. Once processed, the machine learning models will be trained and evaluated using relevant performance metrics to assess their predictive accuracy.
- **Incorporation of clinical data**. Clinical variables, such as age and tumor stage, will be integrated into the dataset. Models trained with this comprehensive feature set will be compared to previous models to evaluate the impact of clinical data on predictive performance.
- **Integration of multigenic assay data**. Multigenic assay scores will be analyzed for correlations with radiomic features and added as predictors. Models will be trained with the extended dataset, and their performance will be assessed to determine if genomic data enhances prediction.

C. Report structure

This report is organized into five main sections. The Introduction presents the background, motivation, and objectives of the study. The Methodology describes the datasets used, the data preprocessing steps, and the exploratory data analysis, followed by the feature selection. The Model development section explains the different machine learning models created and the evaluation metrics used to assess their performance. The Results and discussion section compares the models' performance and analyzes the impact of incorporating clinical and genomic data. Finally, the Conclusions summarize the key findings, discuss limitations, and propose future work.

II. Methodology

A. Description of datasets

The data used in this project was obtained from the TCGA Breast Radiogenomics collection available in [4]. Among other files, such as the original MRI images and their corresponding tumor segmentations (as shown in Figure 1, illustrating the type of imaging data from which radiomic features are extracted for analysis), the dataset includes three types of data sources: radiomic features, multigenic assay results, and clinical data:

- **Quantitative Radiomic Features**. This dataset includes 41 quantitative features derived from MRI images, which describe different aspects of the tumor, such as shape, texture, and signal enhancement dynamics.
- **MammaPrint, Oncotype DX, and PAM50 multi-gene assays**. This dataset provides genomic scores associated with breast cancer prognosis and subtype classification.

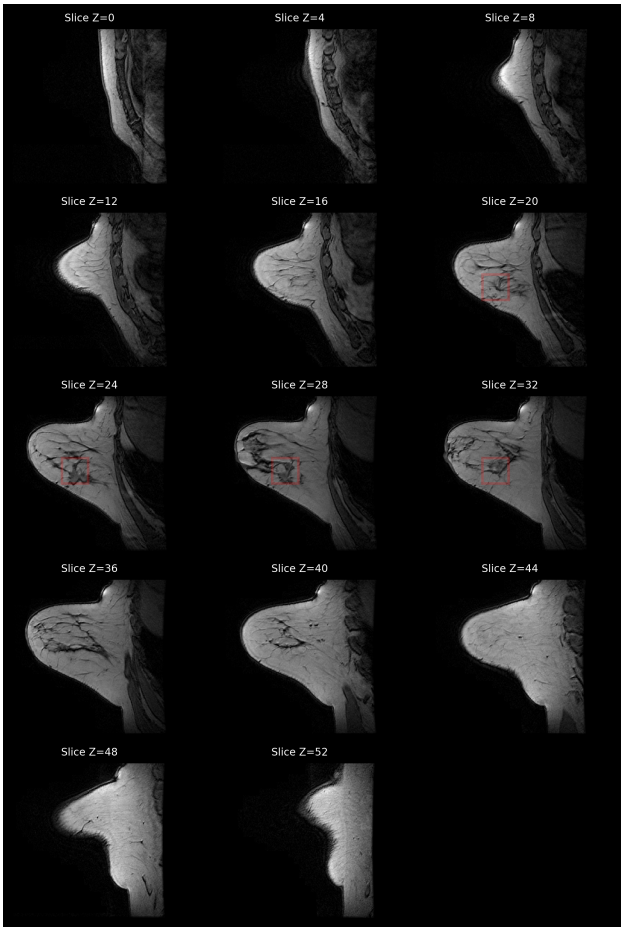

Fig. 1. Example MRI scan with corresponding tumor segmentation

The variable **Pam50.Call** serves as the target variable, indicating the molecular subtype for each sample.

- **Clinical data**. This dataset contains information on patient demographics, tumor characteristics, and treatment-related variables

B. Data preprocessing

The data preprocessing phase ensures that the dataset is consistent and focused on the relevant classification task by selecting complete instances and excluding irrelevant categories.

1) Identification of complete instances: In this step, we ensure that only instances with complete information across all datasets—radiomic features, multigenic assay results, and clinical data—are included. This guarantees that the training and evaluation of machine learning models are performed consistently with the same set of data points, regardless of the features being tested.

The approach involves checking for the presence of a common identifier (CLID) across the three datasets. Instances that do not have corresponding entries in the radiomic, multigenic, or clinical datasets are excluded.

As a result, a new dataset containing only the complete instances is created, ensuring uniformity in the data used across all experiments.

2) Removal of the "normal" category: In this step, instances labeled as "Normal" in the Pam50.Call variable are excluded from the dataset. The rationale behind this decision is that the focus of the project is to differentiate between molecular subtypes of breast cancer, not to distinguish between healthy and cancerous tissue. Including "Normal" as a class would introduce a category that lacks clinical relevance for the subtype prediction task, as "Normal" does not correspond to a specific cancer phenotype.

After this preprocessing, from the 84 cases present in the original dataset, 76 instances remain for analysis in this project. The distribution of the target variable in these instances is shown in Figure 2.

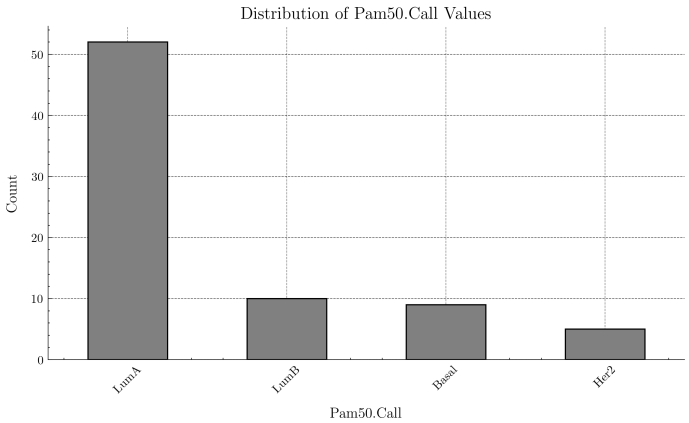

Fig. 2. Distribution of `Pam50.Call` values

As shown in the abovementioned figure, there is a significant class imbalance, with the Luminal A subtype being highly overrepresented compared to the other classes. This imbalance must be taken into account when selecting the evaluation metrics for the model. Additionally, as will be discussed in Section III-A2, undersampling and oversampling techniques will be applied to address this issue and improve the model's robustness.

C. Exploratory data analysis (EDA) and feature selection

The exploratory data analysis (EDA) provides insights into the clinical data and multigenic assay scores, helping to understand their distribution, correlations, and relevance to the target variable. In addition to statistical findings, medical considerations are taken into account to ensure that the selected variables are not only informative but also clinically meaningful for inclusion in the model.

1) Clinical data: The clinical variables included in the model are the patient's age at the time of diagnosis, the cancer's overall stage and tumor size, the number of lymph nodes affected, and the status of estrogen and progesterone receptors. Although an exploratory analysis was conducted to better understand the data distribution, due to space limitations, these details are not displayed. Therefore, the selection of variables is primarily based on clinical criteria:

- **Age at diagnosis**. A key prognostic factor, as younger patients often have more aggressive tumors, while older patients may present less aggressive subtypes. [5]
- **Cancer stage and tumor size**. These variables describe the extent of disease progression and are crucial for predicting molecular subtypes.
- **Number of affected lymph nodes**. Indicates metastatic spread and reflects the severity of the disease. [6]
- **Hormone receptor status (estrogen and progesterone)**. These statuses are essential for distinguishing Luminal subtypes [7]. However, due to their high correlation with the target variable, a model was also built without them to avoid redundancy.

The HER2 receptor status based on immunohistochemistry results was excluded due to a high proportion of missing values, despite its relevance to the HER2-enriched subtype.

2) Multigenic assay scores: The following multigenic assay variables were selected based on their relevance to tumor biology and their ability to complement the radiomic data:

- **GHI_RS Score**. This continuous score, derived from the Oncotype DX assay, measures the risk of recurrence based on gene expression related to tumor growth and proliferation. [8]
- **Correlation with good outcome signature**. This variable represents how strongly the sample correlates with a favorable prognosis gene profile from the MammaPrint assay. [9]
- **Proliferation-related gene expression**. This variable captures the average expression of a set of genes associated with cell proliferation, a key marker of tumor aggressiveness. While not directly predictive of molecular subtype, it is particularly informative for identifying more aggressive subtypes, such as Basal-like or HER2-enriched.

III. Model Development

A. Data scaling and resampling

Before selecting the radiomic features and training the models, the data underwent preprocessing steps to standardize feature values and address class imbalance.

1) Data scaling: The radiomic features were standardized using StandardScaler to ensure that all features have a mean of zero and a standard deviation of one. Standardizing the features helps prevent features with larger magnitudes from disproportionately influencing the model's performance.

2) Resampling for class balance: The dataset exhibited a significant class imbalance, with the Luminal A subtype being overrepresented. To mitigate this, a two-step resampling strategy was applied:

- **Undersampling**. The Luminal A class was undersampled to 30 instances to reduce its dominance and balance the dataset.
- **SMOTE (Synthetic Minority Oversampling Technique)**. This technique was then applied to synthetically generate new samples for the minority classes until they were more balanced with the other classes.

This combination of undersampling and oversampling created a more balanced training set, improving the model's ability to learn meaningful patterns from all classes rather than being biased toward the majority class.

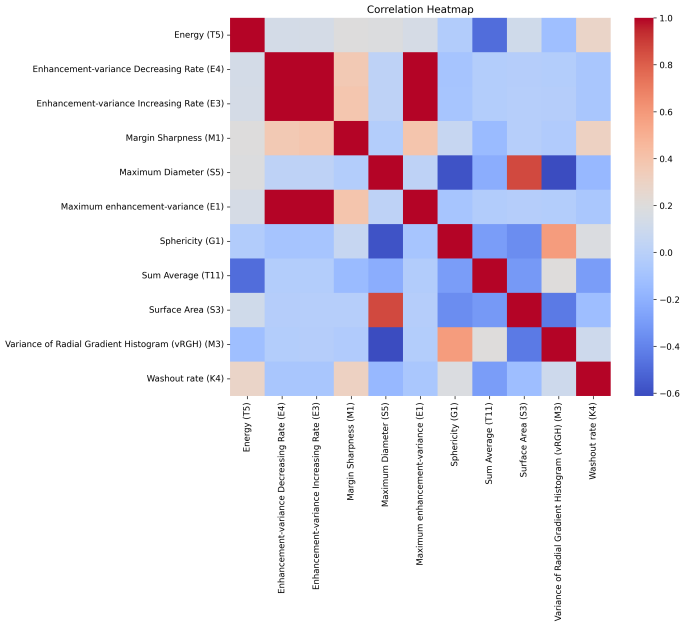

Fig. 3. Correlation heatmap displaying the pairwise correlations between the selected features.

B. Model creation

1) Radiomic features-based model: Once the radiomic features were scaled and a more balanced dataset was obtained, several feature selection techniques and model training steps were performed.

For feature selection, a multi-step approach was used. The Boruta method with Gradient Boosting was first applied to identify the most relevant features from the full set of radiomic variables. Out of the initial 36 features, 11 were retained. These selected features were then analyzed using a correlation heatmap to understand their interrelationships, followed by the creation of a dendrogram to visualize feature clusters.

The correlation heatmap (Figure 3) displays the pairwise correlations between the selected features, highlighting several highly correlated feature pairs, indicating redundancy. To address this, hierarchical clustering was applied, as shown in the dendrogram (Figure 4), where each branch represents a cluster of related features based on correlation distances.

From these clusters, representative features were selected to reduce redundancy while retaining the most informative variables. After testing different numbers of clusters, it was determined that using three clusters provided a good balance without significantly affecting the model's performance. The final representative features selected were:

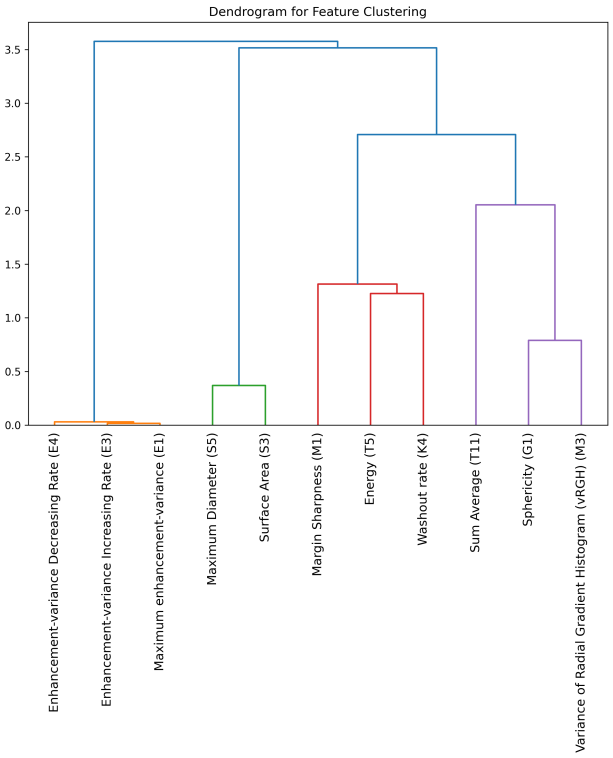

Fig. 4. Dendrogram resulting from hierarchical clustering

- **Margin Sharpness (M1)**. Describes the abruptness of intensity changes at the tumor's boundary, indicating how clearly the tumor is demarcated from surrounding tissues. The sharpness is quantified by averaging the magnitude of the intensity gradient across all boundary voxels.
- **Maximum Enhancement-Variance (E1)**. Measures the variance in the enhancement signal across the most enhancing regions of the tumor, reflecting heterogeneity in vascularization.
- **Surface Area (S3)**. Represents the surface area of the tumor boundary, providing information about the tumor's size and shape complexity.

Finally, a Random Forest classifier was trained using the selected features obtained from the clustering process.

2) Radiomic model with clinical data: To evaluate the impact of clinical variables on the predictive performance of the model, two versions of the radiomic model were created by adding clinical data to the selected radiomic features:

- **Full clinical data model**: This model incorporates the following clinical variables:
  - **Age at diagnosis**
  - **Cancer stage and tumor size**
  - **Number of affected lymph nodes**
  - **Hormone receptor status (estrogen and progesterone)**
- **Reduced clinical data model**: This model excludes the hormone receptor status variables to avoid an overly optimistic evaluation. Clinically, these hormone receptor statuses are highly correlated with the molecular subtype, particularly in distinguishing Luminal A and Luminal B. Removing them allows the model to focus on more indirect clinical features and ensures that the predictions are not dominated by highly correlated variables.

In both cases, the selected radiomic features obtained from the clustering process—Margin Sharpness (M1), Maximum enhancement-variance (E1), and Surface Area (S3)—were combined with the clinical variables, and a Random Forest classifier was trained.

3) Radiomic model with multigenic assays: To evaluate the impact of genomic information on the model's performance, a radiomic model was created by incorporating multigenic assay variables. The selected radiomic features were combined with the following genomic variables:

- **GHI_RS Score**
- **Correlation with good outcome signature**
- **Proliferation-related gene expression**

A Random Forest classifier was trained on this combined feature set.

C. Evaluation metrics used

Given the significant class imbalance in the dataset, the macro F1-score was chosen as the primary evaluation metric. The macro F1-score calculates the F1-score for each class independently and then averages them, giving equal importance to each class, regardless of its frequency in the dataset. This is particularly useful in scenarios where some classes (such as Luminal A) are overrepresented, ensuring that the model's performance is not biased toward the majority class.

The F1-score for a single class is defined as the harmonic mean of precision and recall:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

where:

- **Precision**: The proportion of true positive predictions out of all predicted positives for a class.
- **Recall**: The proportion of true positive predictions out of all actual positives for a class.

The macro F1-score is then computed as:

$$\text{F1}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i \tag{2}$$

where C is the number of classes, and F1_i is the F1-score for class i.

D. Model performance results

All the models described in Section III-B, including an additional model trained with all features (radiomic, clinical, and multigenic data), as well as models trained using only clinical data and only multigenic assays (without radiomic features), were evaluated using the metric explained in Section III-C. The results of this evaluation are presented in Table I.

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	0.60

TABLE I
PERFORMANCE RESULTS FOR THE DIFFERENT MODELS BASED ON ACCURACY AND MACRO F1-SCORE.

IV. RESULTS AND DISCUSSION

A. Comparison of model performance

The performance comparison across the different models, as shown in Table I, highlights significant differences based on the type of features included. The model using only radiomic features achieved a macro F1-score of 0.45, indicating limited predictive power. The highest performance (macro F1-score of 0.78) was observed for the radiomic features combined with full clinical data, showcasing the importance of including relevant clinical variables. Interestingly, the "All" model, which includes radiomic, clinical, and multigenic data, did not surpass the radiomic + multigenic model, suggesting potential redundancy or noise when combining all feature sets.

B. Impact of clinical data inclusion

The inclusion of clinical data, particularly hormone receptor status and tumor staging information, significantly enhanced the performance of the radiomic model. The Radiomic + Full clinical data model, which included hormone receptor status variables, achieved an F1-score of 0.78. In contrast, the Radiomic + Reduced clinical data model, excluding these variables to avoid optimistic bias, had a lower F1-score of 0.42. This outcome confirms the predictive value of hormone receptor status in distinguishing molecular subtypes.

C. Impact of multigenic assays on prediction

The addition of genomic data from multigenic assays further enhanced the model's performance. The Radiomic + Multigenic model achieved an F1-score of 0.74, outperforming the radiomic-only model. Notably, the Only Multigenic model achieved a comparable F1-score of 0.60, indicating that genomic features alone can provide substantial predictive power. However, when combined with radiomic features, the predictive accuracy improved, suggesting that radiomic data provide complementary information about tumor phenotype that is not captured by genomic assays alone. This finding supports the integration of imaging and genomic data for more robust subtype classification.

D. 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. As shown in Figure 5, the model accurately predicts the Basal and HER2-enriched subtypes, demonstrating reliable classification for these categories. However, it tends to confuse Luminal A and Luminal B subtypes, as evidenced by the misclassification of several Luminal A instances as Luminal B.

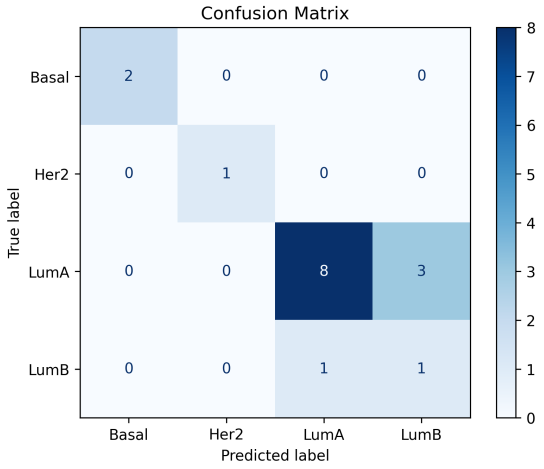

Fig. 5. Confusion matrix for the **Radiomic + Full clinical data** model.

V. CONCLUSIONS

A. Key findings

This study demonstrated that 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 also improved the model's performance, with the Radiomic + Multigenic model achieving an F1-score of 0.74, showing the complementarity between genomic data and radiomic features. However, the "All" model did not outperform the others, suggesting potential redundancy and the need for careful feature selection.

B. Study limitations

Several limitations should be acknowledged:

- **Small dataset size**: The number of instances available for training and evaluation was limited, which may have impacted the model's generalizability and robustness.
- **Limited computational resources**: The complexity of feature selection and model training was constrained by computational resources, potentially limiting the exploration of more advanced model architectures or hyperparameter optimization.

C. Suggestions for future work

Future work could address these limitations through the following actions:

- **Increased data collection**: Incorporating more patient data, especially from diverse cohorts, would improve model generalization and allow for a more accurate evaluation.

- **Better data organization and integration**: A more structured approach to organizing radiomic, clinical, and genomic data would facilitate preprocessing, feature extraction, and model training, potentially improving performance and interpretability.
- **Extraction of radiomic features from raw images**: Performing feature extraction directly from the original MRI scans would allow for a more customized approach and could uncover more relevant imaging biomarkers.
- **Hyperparameter tuning**: Future studies should focus on optimizing hyperparameters such as learning rate, number of estimators, and regularization parameters to improve model performance and reduce overfitting.

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] A. Saha, M. R. Harowicz, L. J. Grimm, C. E. Kim, S. V. Ghate, R. Walsh, and M. A. Mazurowski, "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features," *Br. J. Cancer*, vol. 119, no. 4, pp. 508–516, Aug. 2018, doi: 10.1038/s41416-018-0185-8.

[4] 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.

[5] J. Brandt, J. P. Garne, I. Tengrup, et al., "Age at diagnosis in relation to survival following breast cancer: a cohort study," *World J. Surg. Oncol.*, vol. 13, no. 33, 2015. doi: 10.1186/s12957-014-0429-x.

[6] 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].

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

[8] 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.

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