Project Report “The War on Cancer“

by Dirk Westhölter, January 2026

Introduction:

In the Western world, cancer accounts for millions of deaths each year and represents the second leading cause of death. Machine learning approaches offer the potential to improve patient stratification by identifying individuals who are likely to benefit from specific therapies and by predicting patient trajectories more accurately than existing clinical scoring systems. Achieving this requires large, well-annotated multimodal datasets integrating molecular and clinical information. This project focuses on the development of predictive models for estrogen receptor (ER) status and relapse using gene expression data in combination with corresponding clinical metadata from a cohort of patients with breast (mammary) carcinoma.

Work plan/ methods:

An anonymized dataset comprising clinical variables (lymph node status, histological grade, tumor size, age, ER status, and relapse) and gene expression data for 6,385 genes from 327 patients was used in this study. All analyses were conducted in a Jupyter Notebook running Python 3.11.9 within Visual Studio Code (v1.107.1). Descriptive statistics were computed using the Python packages NumPy, SciPy (scipy.stats), Matplotlib, and Seaborn. Continuous variables were summarized as mean ± standard deviation as well as median and range, while categorical variables were reported as counts and percentages. Gene expression data were analyzed with respect to overall expression levels (mean ± standard deviation) and skewness; a skewness value with an absolute magnitude greater than 1 was considered highly skewed. For dimensionality reduction, principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) were applied to scaled gene expression data using the scikit-learn package. Clustering analyses were performed on PCA-transformed data using multiple approaches, including KMeans, Gaussian mixture models (GMMs), and agglomerative clustering, also implemented in scikit-learn. Estrogen receptor (ER) status and relapse were selected as primary outcome variables for prediction. Several supervised learning models were evaluated using packages from scikit-learn, XGBoost, and TensorFlow/Keras, including logistic regression, tree-based models, and deep learning models. Model performance was assessed using accuracy, F1-score, and ROC–AUC metrics. Finally, a commercial large language model (LLM; GPT-5.0) was prompted to identify a panel of genes considered most relevant for predicting ER status. Predictive models based on this LLM-derived gene set were trained and compared with the previously established models.

Results:

Part 0: Orient yourself

The 327 patients of the study had a mean age of 58.7 years and a mean tumor size of 2cm (Table 1, Figure 1). Histological grade, lymph node status, ER status and relapse status showed an imbalanced distribution across its categories, with ER status exhibiting the strongest shift towards ER positivity (80%) versus ER negativity in only 14%. Lymph node status and relapse status had a high percentage of missing data.

Table 1: Patient characteristics

Characteristic (total, n=327)	N (available)	Summary
Age (years)	252	58.7 ± 11.9 (median 60, range 24–88)
Tumor size	251	2.2 ± 1.1 (median 2.1, range 0–7)
Histological grade	275	
Grade 1	68 (20.8%)	
Grade 2	143 (43.7%)	
Grade 3	64 (19.6%)	
Missing	52 (15.9%)	
Lymph node status	247	
Node negative	192 (58.7%)	
Node positive	55 (16.8%)	
Missing	80 (24.5%)	
Estrogen receptor (ER) status	308	
ER positive	262 (80.1%)	
ER negative	46 (14.1%)	
Missing	19 (5.8%)	
Relapse status	218	
No relapse	128 (39.1%)	
Relapse	90 (27.5%)	
Missing	109 (33.3%)	

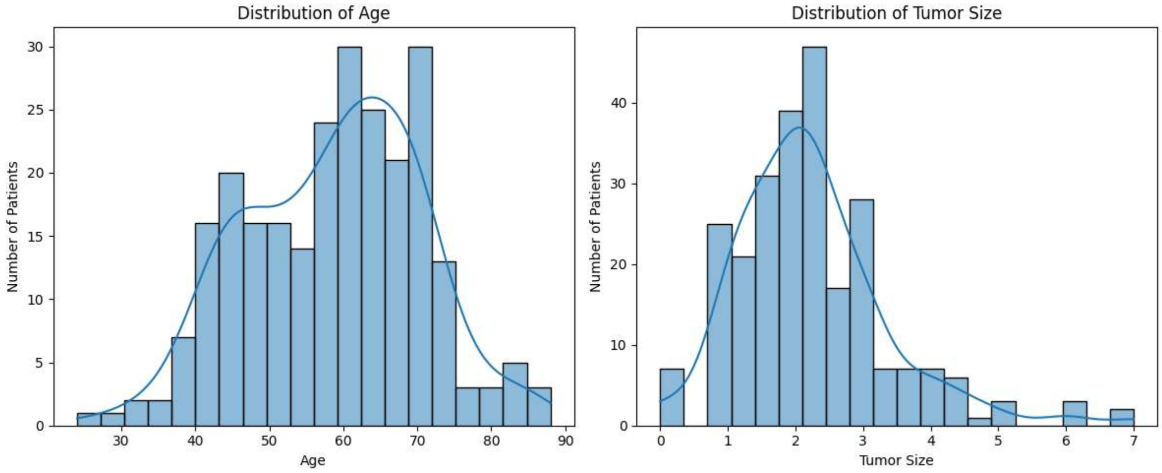

Figure 1: Distribution of age and tumor size show approximately a normal distribution, both right-skewed.

Gene expression data were available from all patients and comprised data from 6,385 genes. Skewness was calculated as a measure of asymmetry of a distribution. As a rule of thumb, skewness values higher than 1 or lower than -1 were considered relevant. This was the case in 1,138 out of 6,384 genes (18%). Looking at the top- and bottom-expressed genes, the top-expressed genes had high mean expression with small standard deviation, while the bottom-expressed genes showed higher variability. This justifies using Python's StandardScaler to balance these disparities; otherwise, high-expression genes may dominate downstream models. Skewness can be handled with log transformation. However, not all models require this preprocessing step (as with XGBoost), and therefore log transformation was not applied in the methods used further downstream.

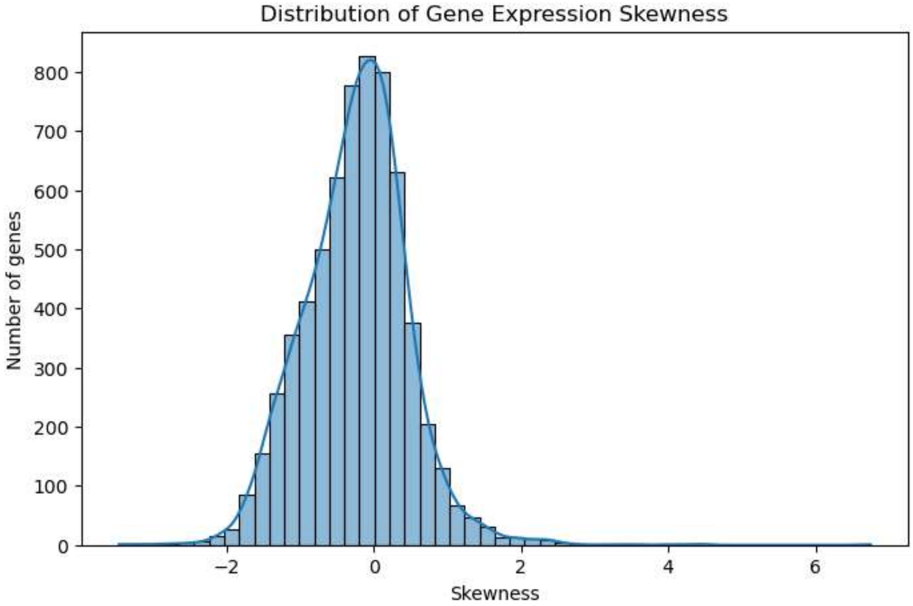

Figure 2: Skewness as calculated with scipy.stats package. Skewness can be handled by most downstream models and transformation/ filtering was not applied in this study.

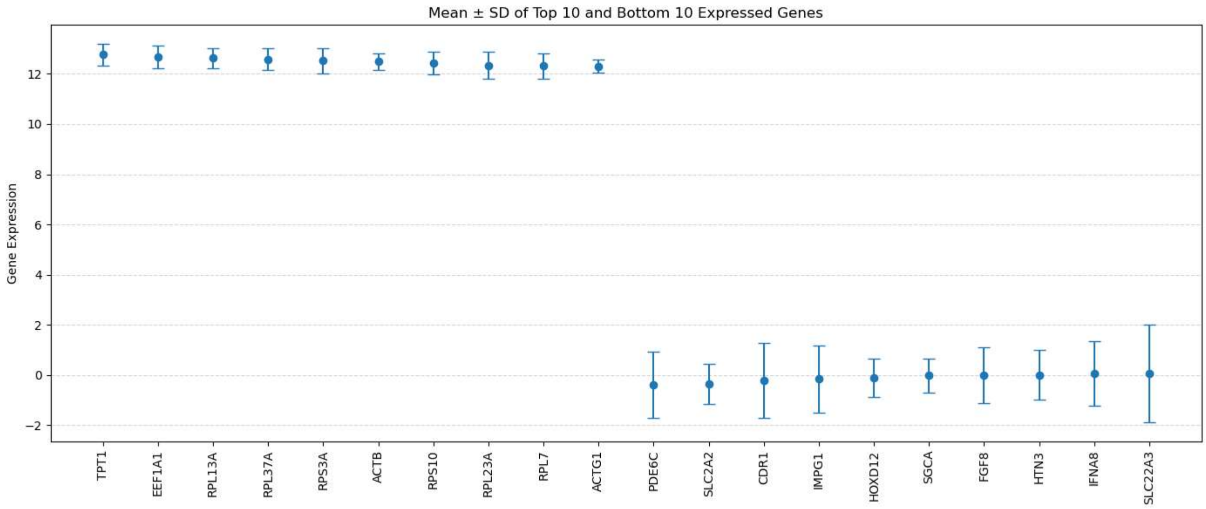

Figure 3: Top and bottom expressed genes. High expressed genes showed less standard deviation. Bottom genes showed higher variability.

Part 1: Dimensionality Reduction

Dimensionality reduction is used to make the high dimensional data of this study interpretable and visible. In addition, the present data set contains many features compared with a rather small number of patients which may provoke the “curse of dimensionality”. Principal Component Analysis (PCA) is well understood method that handles correlated data properly to make data according to its maximum variance visible and ready for efficient processing. T-SNE is shown as an alternative method to reduce dimensionalities. It is able to handle nonlinear data and to reveal clusters, preserving local structures and supporting good visualization of data. Both methods may be used for further processing. PCA of gene expression data reveals an apparent separation in the projected space. Samples with negative or unavailable ER status predominantly project into one region of the first two principal components of PCA that spans a wider range along PC1 and PC2. ER-positive samples are found across both regions. This suggests that ER status is associated with major sources of variance in the data, but does not fully explain them. t-SNE shows more pronounced visual separation between samples due to its emphasis on preserving local neighborhood structure. Despite this, the qualitative interpretation is consistent with PCA (Figure 4, Figure 5). The PCA scree plot shows that the first principal component explains the largest proportion of variance, while PCs 2–5 provide additional but diminishing contributions. Subsequent components contribute little variance and likely capture noise rather than task-relevant structure (Figure 6). The PCA-transformed data (PC1–PC5) captures a substantial proportion of the total variance and is therefore suitable for downstream machine learning. This is further supported by a simple logistic regression model, which achieves above-chance performance when trained on these components (first model in Part 3). Other clinical targets (relapse, nodal status, tumor size, and grade) did not exhibit clear separation in the PCA projection. ER status appears to be the most suitable target variable for subsequent prediction tasks.

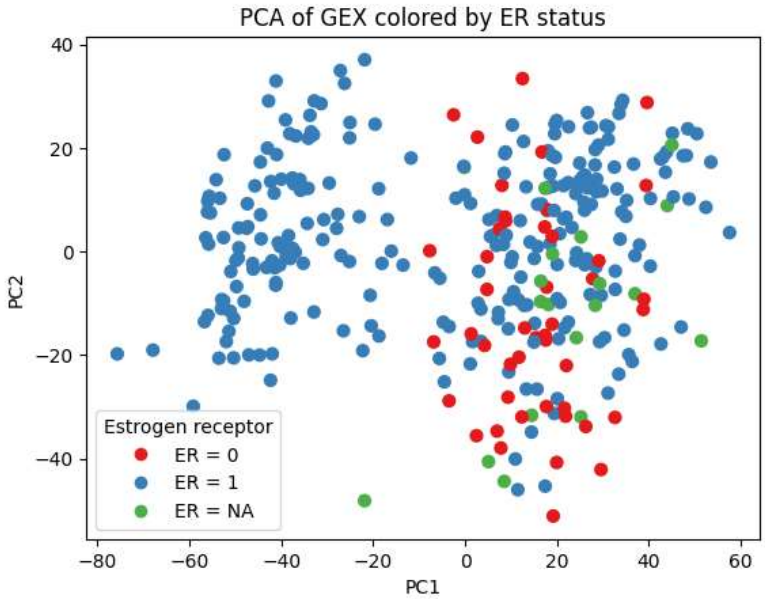

Figure 4: Principal component analysis of gene expression data colored by ER status. Positive ER status is found across both regions.

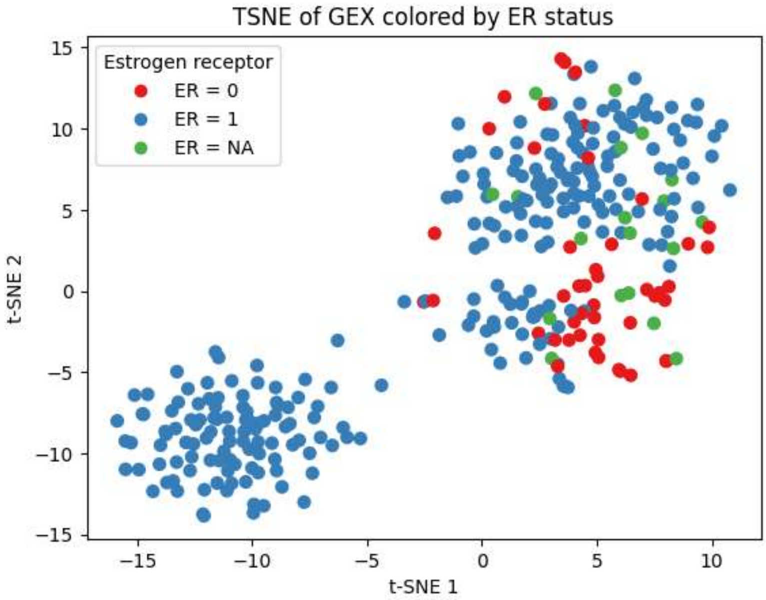

Figure 5: T-SNE analysis showing more pronounced separation of regions, but similar distribution of ER status.

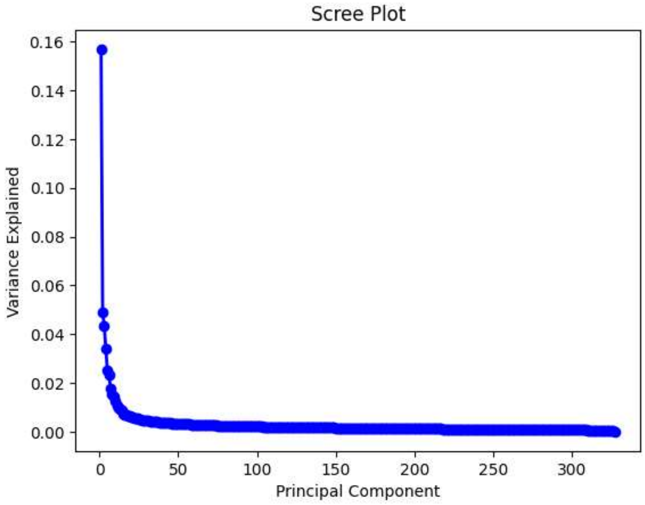

Figure 6: Screeplot. PC1 explains the largest proportion of the variance.

Part 2: Clustering

PCA-transformed data or scaled original data were used to explore clusters in the gene expression dataset. The silhouette score, which measures how well each data point fits within its assigned cluster compared to other clusters, was used to estimate the optimal number of clusters. In this analysis, the silhouette score was highest for two clusters, indicating that two clusters best capture the structure of the data (Figure 7). Overall, the optimal silhouette score was moderate, potentially reflecting the effects of the curse of dimensionality due to the high number of features. Different clustering methods were applied, and the resulting clusters were largely consistent across methods (Figure 8). In Cluster 0, patients were older, had larger tumors, and were all ER-positive. In Cluster 1, patients showed a mix of ER status, with ER positivity being dominant (Table 2). Although gene expression levels were generally higher in Cluster 1, this did not correspond to greater cancer aggressiveness, as assessed by lymph node status, histological grade, tumor size, or relapse (missing data in Cluster 0 may affect interpretation). Differences in expression likely reflect molecular subtype differences rather than severity or prognosis (Figure 9). Finally, cluster stability was assessed using the Adjusted Rand Index (ARI) for KMeans. A mean ARI close to 1 indicated that the clustering results were highly stable and nearly identical regardless of initialization.

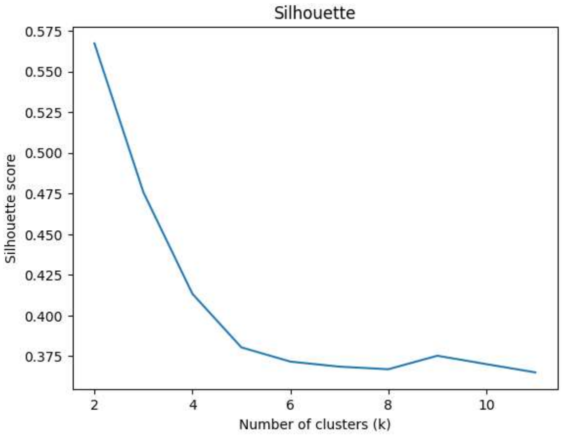

Figure 7: The silhouette score was highest for two clusters. Overall, silhouette scores were moderate when calculated on the PCA-transformed data (data shown) and low for the scaled original data, suggesting that clustering structure is clearer in the reduced PCA space than in the full feature space.

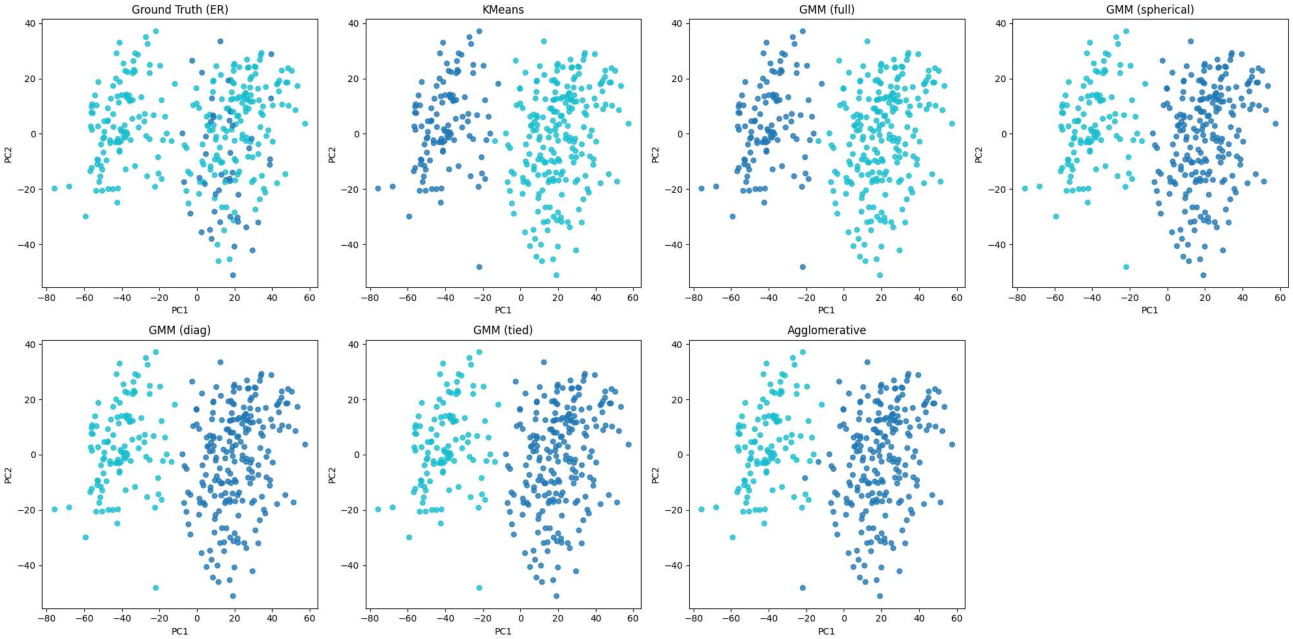

Figure 8: Various clustering methods—including KMeans, Gaussian Mixture Models (GMM) with different covariance types (full, spherical, diagonal, tied), and agglomerative clustering—consistently identified two distinct clusters in PCA-transformed data set.

Table 2: KMeans- based clusters. In cluster 0, patients were older and exhibited 100% ER positivity.

Characteristic	Cluster 0 (n=111)	Cluster 1 (n=216)
Age (years)	62.9 ± 9.2 (median 64, range 43–80, n=71)	57.0 ± 12.4 (median 57, range 24–88, n=181)
Tumor size (cm)	2.6 ± 1.4 (median 2.3, range 0–7, n=72)	2.1 ± 1.0 (median 2.0, range 0–6, n=179)
Histological grade	2: 52 (46.8%)1: 22 (19.8%)3: 18 (16.2%)Missing: 19 (17.1%)	2: 91 (42.1%)3: 46 (21.3%)1: 46 (21.3%)Missing: 33 (15.3%)
Lymph node status	0: 45 (40.5%)1: 22 (19.8%)Missing: 44 (39.6%)	0: 147 (68.1%)1: 33 (15.3%)Missing: 36 (16.7%)
ER status	1: 110 (99.1%)Missing: 1 (0.9%)	1: 152 (70.4%)0: 46 (21.3%)Missing: 18 (8.3%)
Relapse status	0: 34 (30.6%)1: 21 (18.9%)Missing: 56 (50.5%)	0: 94 (43.5%)1: 69 (31.9%)Missing: 53 (24.5%)

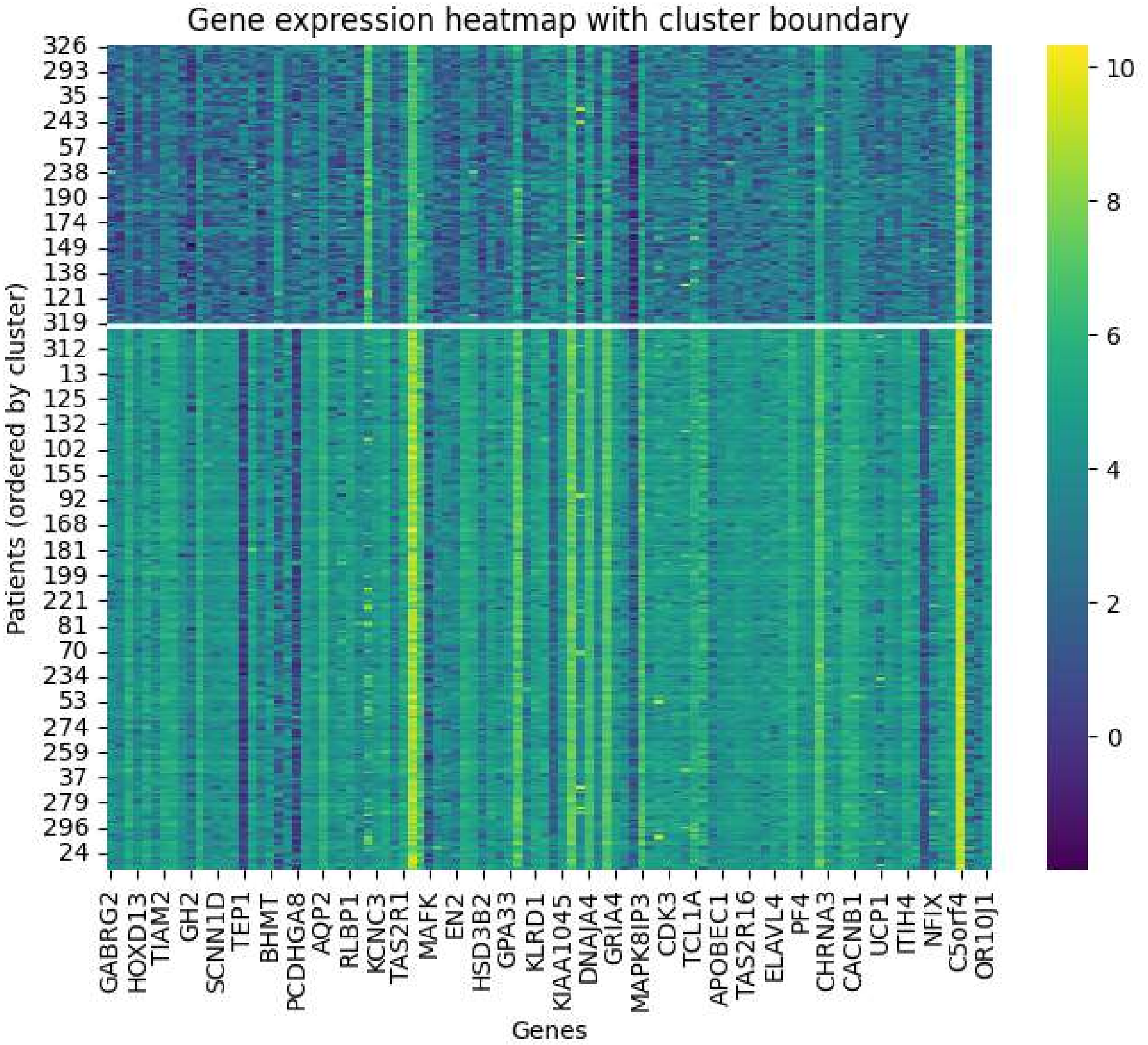

Figure 9: Top differently expressed genes shown as heatmap. Cluster 0 in the upper part, cluster 1 in the lower part. With only a few higher expressed genes in cluster 0, overall expression levels are higher in cluster 1 reflecting potentially some molecular subtype.

Part 3: Prediction

The models were trained and evaluated on scaled and PCA-transformed gene expression data after splitting the original dataset into training and test sets, with the test set kept completely untouched during model development. For each target variable (ER status and relapse), models were trained on complete cases only, without imputation of missing values. ER status was selected as a target variable because it exhibited the greatest structure and separability following PCA and clustering analyses. Relapse was chosen as a second target because it represents the only clinically relevant outcome measure available in the dataset. Class imbalance, particularly for ER status and relapse, was taken into account during performance evaluation. Using a majority-class baseline, an accuracy of 0.86 was obtained for ER status prediction. A logistic regression model was first fitted to predict ER status, using the class_weight="balanced" option in scikit-learn to address class imbalance by weighting classes inversely proportional to their frequencies. This model achieved an accuracy of 0.87, indicating slight improvement over the majority-class baseline, although the ROC–AUC was 0.93, suggesting acceptable discrimination despite limited gains in accuracy. Next, an XGBoost classifier was trained, with class weights manually calculated from the training data. This model demonstrated improved performance on the test set (accuracy 0.93, F1-score 0.96, ROC–AUC 0.95). As a third approach, a deep learning model (TensorFlow/Keras) was trained with adjusted class weights showing moderate performance (accuracy 0.89, F1-score 0.93, ROC–AUC 0.74). In addition, logistic regression and XGBoost models were trained to predict relapse. However, both models exhibited weak predictive performance, indicating limited potential for clinical usefulness in this setting. XGBoost outperformed logistic regression and deep learning, potentially due to its capability to handle nonlinear relationships and interactions between the features. Relapse status proved difficult to predict using gene expression data alone. This suggests that relapse probability is likely influenced by additional factors beyond gene expression, including clinical, pathological, and treatment-related parameters not captured in the dataset.

Part 4: Evaluation

Model evaluation was performed for ER prediction models. Prediction models for relapse did not achieve satisfactory performance and were therefore not considered for further in-depth evaluation. To assess and control potential overfitting, early stopping was applied to the XGBoost model developed for ER prediction which is preferred model from Part 3. For this purpose, the training dataset was further split into an internal training and validation set. The logloss metric was selected as the optimization metric for the early stopping procedure. The optimal model was obtained after 70 boosting iterations, yielding a best validation log loss of 0.41, indicating that model performance plateaued at this point. The optimized model was subsequently evaluated on an independent hold-out test set, where it demonstrated performance comparable to the original model described in Part 3 (accuracy = 0.94, AUC = 0.96, F1-score = 0.97). These results suggest that optimization of the XGBoost model via early stopping produced an equally performing model, indicating that the previously selected hyperparameters did not lead to substantial overfitting. Five-fold cross-validation (CV) of the adjusted ER XGBoost model with 70 boosting iterations yielded a mean accuracy of 0.86 (0.82, 0.84, 0.89, 0.89, 0.86) and a mean AUC of 0.90. These results provide an estimate of the model's expected performance on independent data, which is not superior compared with the accuracy achieved by the majority-class (naïve) model. However, the model's performance is sensitive to data partitioning, and limited sample size per fold may contribute to variability in cross-validation estimates. Mean AUC was 0.90 across validation sets which is below the adjusted original model, suggesting some limitations in model robustness. GridSearchCV was used to improve models performance by adjusting the XGBoost hyperparameters. The improved model demonstrated comparable performance in cross-validation (CV accuracy mean 0.87 and mean AUC 0.89). The final XGBoost model (70 boosting iterations, GridSearchCV recommendations) achieved ACC 0.95 and AUC 0.95 on the hold out test set. For the purpose of analyzing sensitivity to key choices, the final XGBoost model was also applied on the original data (no PCA) and showed satisfying performance (ACC 0.93, AUC 0.96) as well as when applied across different seeds of the model (ACC: mean = 0.94 , std = 0.01, AUC: mean = 0.96, std = 0.003, final model on PCA-transformed data). Learning curves were calculated to estimate the effect of larger sample sizes of ER patient data. Both curves rise sharply initially, showing that the model learns quickly from small amounts of data and the plateau indicates no significant impact of larger samples (Figure 10).

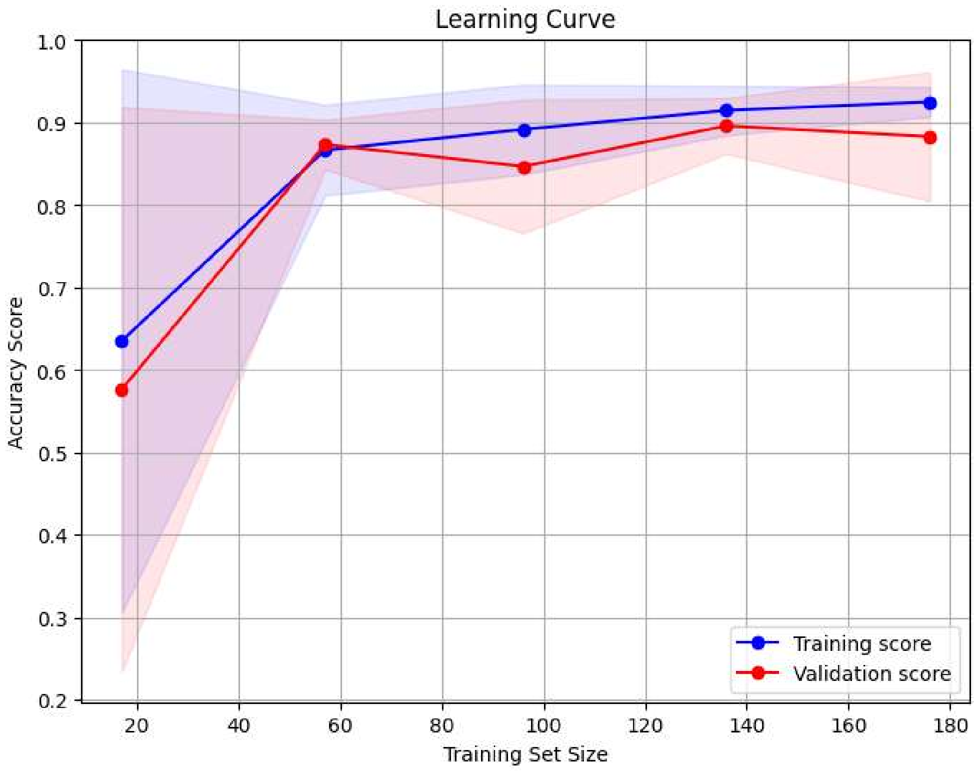

Figure 10: Learning curves for training and validation data. Accuracy improves only slightly with larger sample sizes.

Part 5: LLM-augmented analysis

Gene expression and clinical data were provided to GPT-5 to identify genes most relevant for predicting ER status (prompt in Appendix A1). For machine learning purposes, the large language model recommended a panel of 15 genes, which were subsequently extracted from the dataset. In addition, the LLM suggested the use of logistic regression as the primary modeling approach. However, logistic regression showed limited performance on both PCA-transformed data and the original data restricted to the 15 selected genes (PCA-transformed: accuracy 0.79, ROC–AUC 0.88; original: accuracy 0.79, ROC–AUC 0.87). As XGBoost had previously demonstrated superior performance, it was also applied to the original, non-transformed data using the 15-gene panel. Although this approach yielded improved performance compared with logistic regression, it did not outperform the models trained on the full feature set described in Parts 3 and 4 (Table 3). Cross-validation of the 15-gene XGBoost model resulted in a mean accuracy of 0.84 and a mean ROC–AUC of 0.92, consistent with observations from the other models. Overall, LLM-based gene preselection did not improve predictive performance in this setting, suggesting that LLM (expert)-guided feature selection without direct optimization on the dataset may be insufficient for high-dimensional gene expression modeling.

Table 3: Overview of final XGBoost models.

Model	Dataset	Target	PCA	Result: ACC	Result: AUC
XGBoost_final	All genes	ER	YES	0.95	0.95
XGBoost_final	All genes	ER	NO	0.93	0.96
XGBoost	LLM 15 genes	ER	YES	0.89	0.84
XGBoost	LLM 15 genes	ER	NO	0.87	0.93

Limitations:

This study has several limitations. Information regarding the origin, year of collection, and validity of the dataset was unavailable, as were details on the gene expression measurement technique and the criteria used to select the gene panel. In addition, the clinical data were incomplete and limited in scope; other outcome measures may have been more appropriate targets for prediction. Limiting the analysis to complete data may introduce selection bias. Moreover, predicting ER status from gene expression data has limited clinical relevance, as ER status is routinely determined using established methods which are faster and less expensive. Furthermore, no external validation dataset was provided. Thus, the generalizability and performance of the models on independent datasets from other institutions remains uncertain.

The models themselves were trained on high-dimensional data with a limited sample size, increasing the risk of overfitting and limiting their robustness. ER status exhibited class imbalance (as did the other targets). Despite the use of techniques to handle class imbalance, this imbalance could have an impact on model training and reduce or inflate performance metrics. Some models were thresholddependent, but thresholds were not further explored. Only a limited number of models (regression, XGBoost, deep learning) were tested. Other ML tools, for example newer techniques developed to handle tabular data, may outperform the models of this study. The final and best model (XGBoost) performed well on the test data. However, cross validation suggests reduced transferability to other datasets. Gene selection by the LLM was not optimized on the dataset and lacked biological or statistical validation within the cohort, as this was not supported by the free version of GPT-5.

Conclusion:

PCA and clustering analyses of gene expression data from patients with breast (mammary) carcinoma revealed the presence of two distinct molecular subtypes. However, using gene expression data alone, it was not possible to develop high-quality predictive models for the available staging parameters (histological grade, lymph node status, and tumor size) or for the clinical outcome relapse. Although the best-performing models outperformed the majority-class baseline, their robustness and transferability to external datasets remain uncertain. Integrating gene expression data with multimodal information, such as additional clinical variables or imaging features, may enable the development of more accurate and clinically useful prediction models. Furthermore, the provided gene expression data and cluster-specific differences provide a starting point for future investigations into the underlying molecular mechanisms, which may support the identification of therapeutic targets and contribute to more personalized treatment strategies.

Concrete recommendations include information on the presence of distant metastases into baseline data, providing proliferation markers such as Ki-67, and integrating imaging features. In addition, overall survival should be included as the primary outcome measure, as it is the most clinically relevant endpoint in oncology. Combining these data in a multimodal model may enable more accurate survival prediction and improved patient stratification.

Appendix

A1 GPT 5.0 prompt

“You are a data scientist working with raw gene expression data (mammacarcinoma_gex.csv) and corresponding clinical metadata (mammacarcinona_pat.csv). Your task is to identify the most informative genes for predicting estrogen receptor (ER) status. Select and report the top 15 genes most predictive of ER status. Specify the machine learning model used for ER prediction, including the final hyperparameters. Report the model’s performance using appropriate evaluation metrics (e.g., accuracy, AUC, F1-score) on an independent test set.”