MACHINE LEARNING APPROACHES FOR PREDICTING ESTROGEN RECEPTOR STATUS FROM BREAST CANCER GENE EXPRESSION DATA

Predictive Analytics and Machine Learning 25WS

Daryna Pikulska

AHI RWTH 2025

Abstract

In this project breast cancer gene expression data from 327 patients, combined with clinical metadata, were analysed to predict target status. Different machine learning techniques were used including principal component analysis (PCA) (1) for dimensionality reduction of the high-dimensional gene expression profiles. Clinical metadata imputation of the non-target columns for overcoming issues with “Not a Number” (NaN) or missing values. Several machine learning models were run, tested and evaluated, including linear and ensemble methods, with explicit handling of class imbalance. Finally, the LLM-augmented analysis was performed and compared with manually created models.

Analysis

Dataset Overview

First, I looked into the gene expression profiles from 327 breast cancer patients across 6384 genes. The dataset contains no missing gene expression values. However, there are substantial missing values in clinical variables, particularly relapse status (33.3%) and lymph node involvement (24.5%). Age and tumour size are also missing in approximately 23% of cases, while estrogen receptor (ER) status is the most complete, with only 5.8% missing. For a more detailed overview, see Fig. 1.

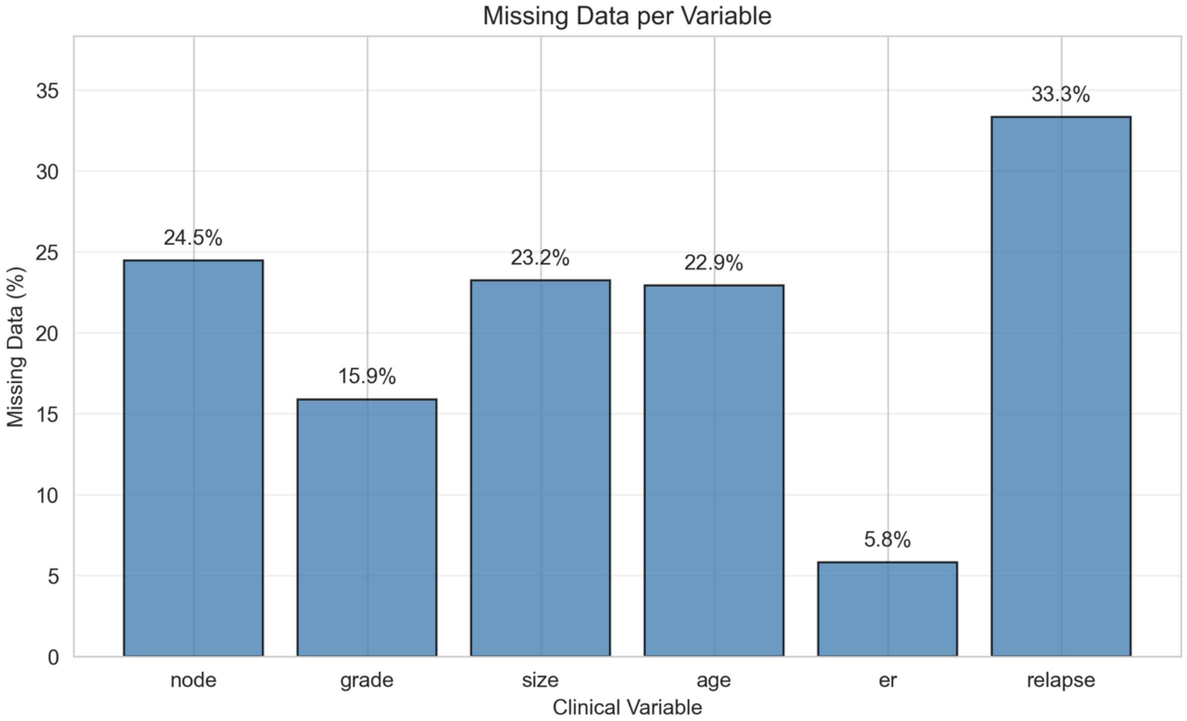

Figure 1: The barplot that represents the missing values in the clinical data, shown in percentages.

Key Data Characteristics

Class imbalance is evident across clinical outcomes: ER-positive tumours dominate (80% vs 14% ER-negative), lymph node involvement is present in only approximately 17% of cases, and relapse occurred in 27% of patients with available data. Tumour grade shows a bell-shaped distribution centred on grade 2 (44%), with fewer grade 1 (21%) and grade 3 (20%) tumours. Patients averaged 59 years old (SD=12) with tumours measuring 2.2 cm (SD=1.1). For more detailed results, see Fig. 2.

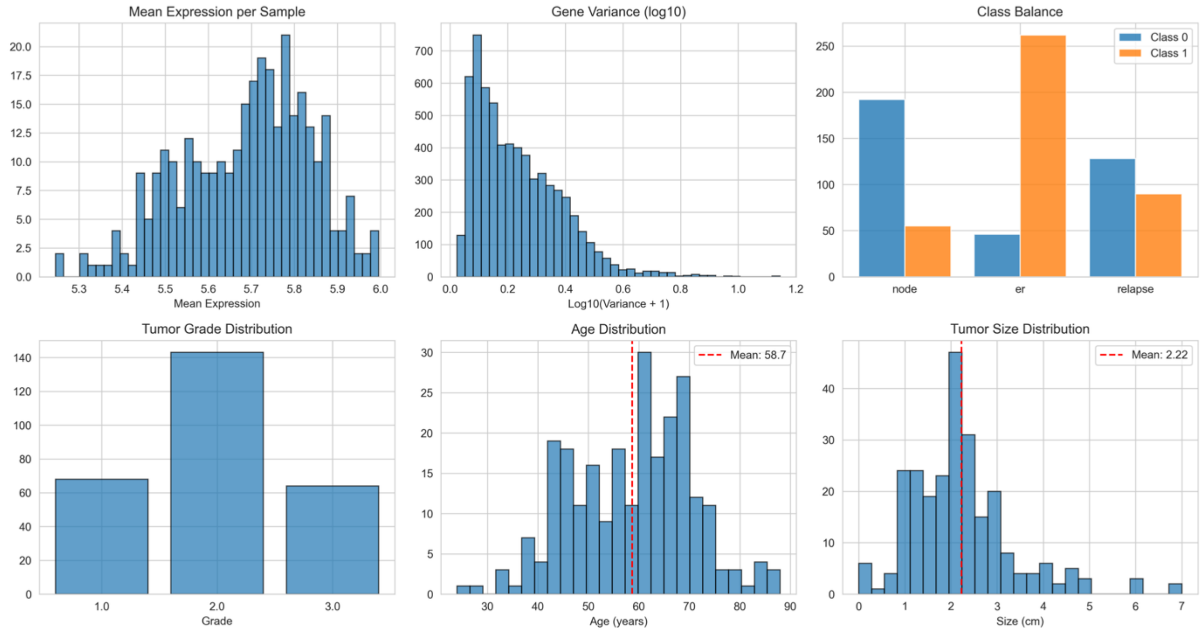

Figure 2: Visualisation of the mean gene expression, gene variance, class imbalance and clinical values distribution.

As the data was preprocessed before to avoid removing key features or rare but potentially informative genes and interfering with the final results, the decision of keeping all genes was made.

Part I — Dimensionality Reduction

Gene expression data and corresponding clinical metadata were merged using a unique patient identifier. Samples with missing target labels were removed before analysis. Estrogen receptor (ER) status was selected as the target variable for this analysis. The remaining dataset was split into training (60%), validation (20%), and test (20%) subsets using stratified sampling to preserve class balance across splits. Gene expression features, consisting of several thousand genes, were treated separately from clinical metadata. Before dimensionality reduction, gene expression data were standardised using z-score normalisation (2):

$$z = \frac{X - \mu}{\sigma}$$

where X is a raw score, μ (mu) is the population mean, and σ (sigma) is the population standard deviation based on the training set only, and the same transformation was applied to the validation and test sets to avoid data leakage. PCA was then fitted on the scaled training gene expression data for dimensionality reduction. The number of principal components was selected based on scree and cumulative variance plots (see Fig.3). Although there is no clear “elbow” point, the decision of selecting eight components was made based on the slope flattening, which indicates decreasing significance of all following components. These eight principal components explained 37.4% of the total variance in the gene expression data, providing a balance between dimensionality reduction and information retention.

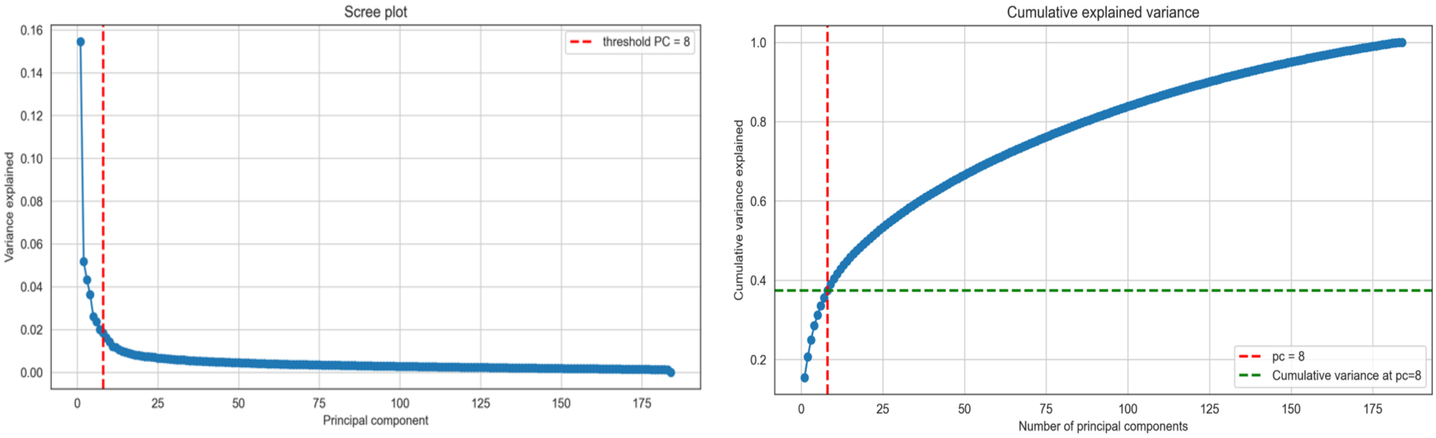

Figure 3. The scree and cumulative explained variance plots.

Clinical metadata (node status, tumour grade, tumour size, patient age, and relapse status) were scaled separately using standardisation fitted only on the training set. The PCA-transformed gene expression features and scaled metadata were concatenated to form the final feature representation used for downstream analysis.

Additionally, two nonlinear dimensionality reduction methods—Uniform Manifold Approximation and Projection (UMAP) (3) and t-distributed Stochastic Neighbour Embedding (t-SNE) (4) were applied to the scaled gene expression training data for exploratory analysis and visualisation. Both methods were configured to produce two-dimensional embeddings to facilitate visual inspection of class separability. Based on the 2D visualisation of the data, none of the methods separated low-dimensional data based on the ER parameter with the parameters selected for this particular example. Clearly, changing the number of neighbours or components in the UMAP function or perplexity in the t-SNE function would affect the final results. But for this purpose exercise the parameters were selected based on the trial-and-error method. The quality of separation between ER-positive and ER-negative samples in the reduced spaces was analysed using the silhouette score. Although all methods demonstrated extremely low scores, PCA, evaluated using the first eight components, demonstrated the highest silhouette score among the compared methods, indicating slightly more consistent separation (see Fig.4).

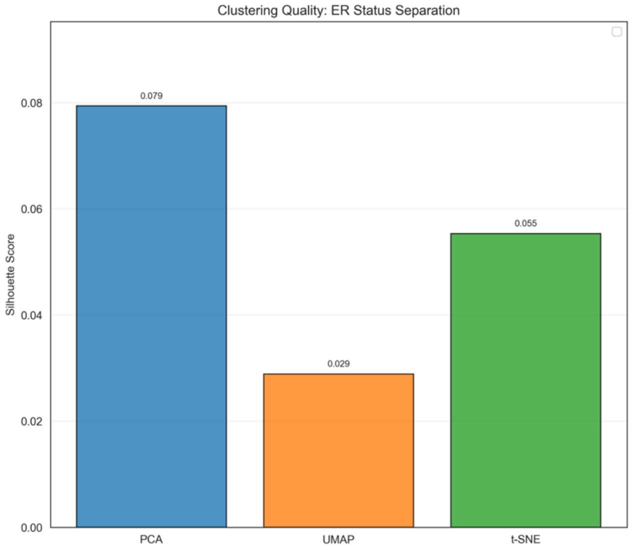

Figure 4. Bar plot visualisations of the results of the Silhouette Score evaluation of the PCA, UMAP and t-SNE dimensionality reduction.

Finally, the PCA-based feature representations combined with clinical metadata were saved separately for the training, validation, and test sets to support reproducible downstream modelling and clustering analyses.

Part II — Clustering

The PCA-transformed datasets generated in the previous step were used as input for unsupervised clustering analyses. The training dataset containing eight principal components and scaled clinical metadata was loaded, and the ER status was excluded from the feature matrix to ensure a fully unsupervised setting. The ER label was retained only as ground truth for the evaluation and visualisation. The first two principal components were additionally extracted for two-dimensional visualisations. Missing values in the clinical metadata were imputed before clustering. Mean imputation was applied using statistics computed from the training set and the same values were used to impute the validation and test sets. This approach ensured consistency across splits while avoiding data leakage. The imputed datasets were saved for reproducibility and downstream analyses.

Several clustering algorithms were tested including k-means clustering, Gaussian Mixture Models (with full, spherical, diagonal, and tied covariance structures), agglomerative hierarchical clustering and DBSCAN. For methods requiring a predefined number of clusters, two clusters were initially specified to represent the binary metadata of the ER status. All algorithms were configured with fixed random seeds to ensure reproducibility. The resulting clusters were visually compared to the ground truth ER labels using scatter plots in the space of the first two principal components (Fig.5).

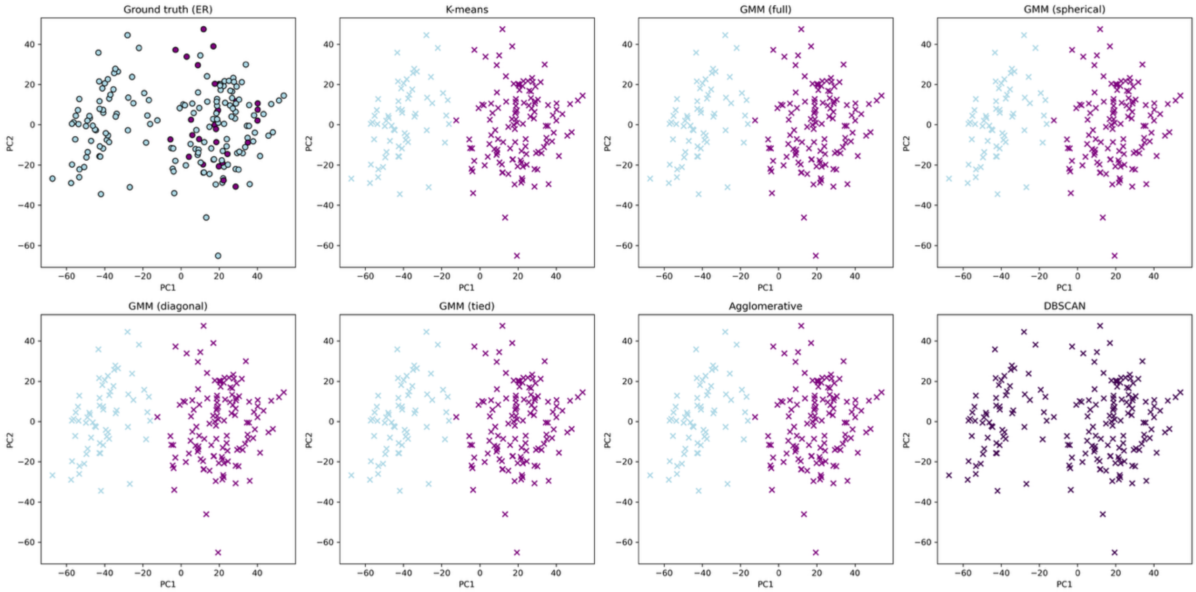

Figure 5. Unsupervised clustering comparison with the ground truth (ER status) using a predetermined number of clusters.

As one can see, none of the methods worked for the ER separation. Therefore, to see if clustering captured any other known metadata features, additional visualisations were generated to examine how clinical metadata variables were distributed within the PCA space (Fig.6). Unfortunately, that was not the case.

Quantitative evaluation was performed using the Jaccard similarity score (5) to measure agreement between clustering results and the ER ground truth. All clustering methods had a Jaccard score approximately equal to 0.5 (saved in 02_clustering.ipynb), except DBSCAN, which had 0.8. It is a clear example of how one can make a mistake assuming that DBSCAN worked best in this case, however, it is clear after visualisation that there is just one cluster based on the DBSCAN and therefore the highest overlap without any biological meaning.

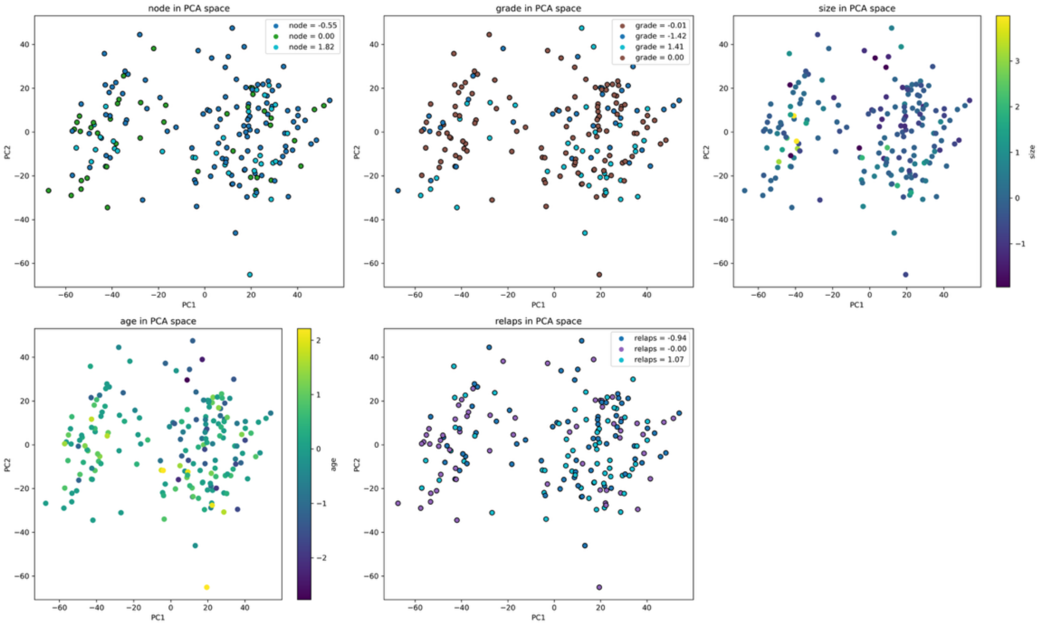

Figure 6. Unsupervised clustering comparison of the clustering results with clinical metadata.

To further assess k-means clustering, internal validation metrics were computed for different numbers of clusters ranging from two to ten. Silhouette scores and Davies–Bouldin indices were calculated to evaluate cluster cohesion and separation (Fig.7). Based on these metrics, an alternative k-means solution with nine clusters was explored and visually compared to the ground truth. While this configuration altered the clustering structure, it did not show a clear biological correspondence to the clinical features.

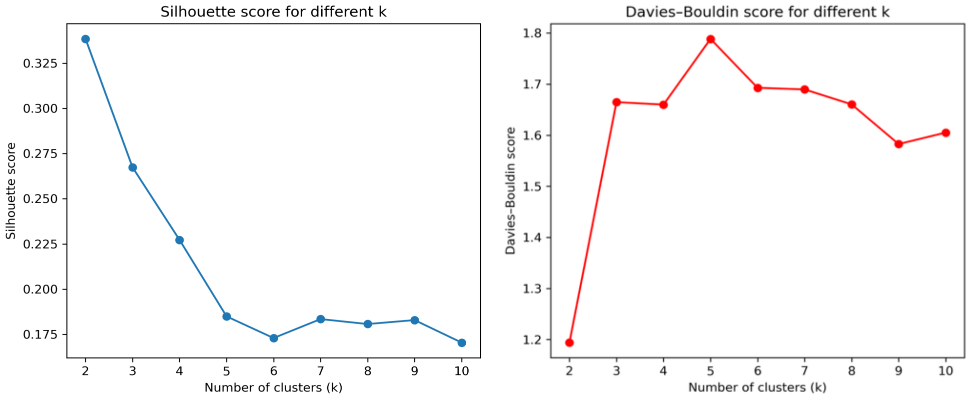

Figure 7. Visualisation of the Silhouette and Davies–Bouldin scores for different numbers of clusters.

Finally, clustering methods were also evaluated without specifying the number of clusters in advance, allowing algorithms such as DBSCAN and Gaussian Mixture Models to infer structure directly from the data. That revealed that most methods identified only one cluster, unable to recover the expected two-group structure corresponding to ER-positive and ER-negative samples (Fig.8). As a result, the high Jaccard score of 0.7 was considered misleading again, reflecting the method's tendency to collapse samples into one dominant cluster rather than meaningful biological separation.

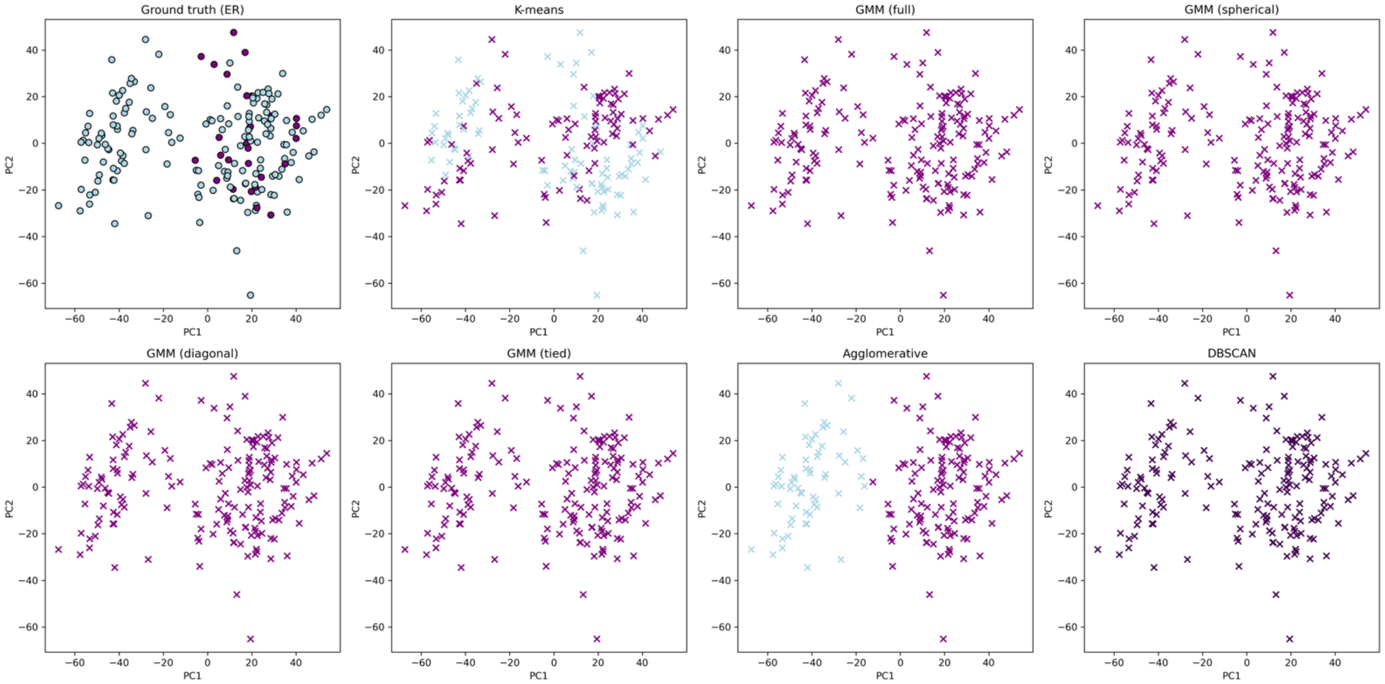

Figure 8. Unsupervised clustering comparison with the ground truth (ER status) without using a predetermined number of clusters.

The final conclusion is that none of the methods consistently recovered the known binary structure of the reported ER status in an unsupervised manner. That might be due to the complexity of the underlying data and the limitations of clustering approaches when applied without supervision in high-dimensional biological settings.

Part III — Prediction

In this part of the analysis, supervised machine learning models were tested to predict ER status using the previously constructed feature representation. The eight principal components from gene expression data and scaled clinical metadata were used as an input. Missing metadata values were imputed using training-set means as described above. Separate training and test datasets were used to ensure an unbiased evaluation of model performance. To capture different modelling assumptions, three contrasting models were evaluated. Logistic regression (6) was used as a linear baseline model, providing a simple and interpretable reference. A random forest classifier (7) was employed to model non-linear relationships through a group of decision trees. In addition, an XGBoost (8) classifier was included as a gradient-boosted tree-based method capable of learning complex interactions with built-in regularisation.

Class imbalance in ER status was handled without advanced transformations of the data distribution, such as SMOTE (9), to avoid potential amplification of the noise or overfitting. Therefore, for logistic regression and random forest models, class weighting was applied to penalise misclassification of the minority class. In the XGBoost model, imbalance was addressed using the scale_pos_weight parameter, which adjusts the loss function according to the ratio of negative to positive samples. Therefore, the minority-class or negative ER predictions were not ignored during training. To balance performance and generalisation model hyperparameters were selected. L2 regularisation (10) with a standard regularisation strength and increased iteration limits to guarantee higher convergence was used for logistic regression. The random forest model employed a relatively large number of trees to stabilise predictions and the bootstrap aggregation was used for variance reduction. For XGBoost, hyperparameters were optimised using randomised search with five-fold cross-validation (11), using ROC-AUC as the optimisation criterion.

Model performance was evaluated on a held-out test set using accuracy, precision, and ROC-AUC. Confusion matrices were generated for each model to visualise classification errors and class-specific performance. The random forest and XGBoost models achieved the highest accuracy (0.90), with the random forest obtaining the highest ROC-AUC (0.95), indicating better overall class separation. It was also the only model that predicted 0 false negative results, but more false positive results compared to the two other models. XGBoost achieved the highest precision (0.94), showing strong confidence in positive ER predictions. Logistic regression demonstrated slightly lower overall performance but remained comparably accurate, demonstrating that linear decision boundaries still captured relevant structure in the data. The confusion matrices are presented in Figure 9.

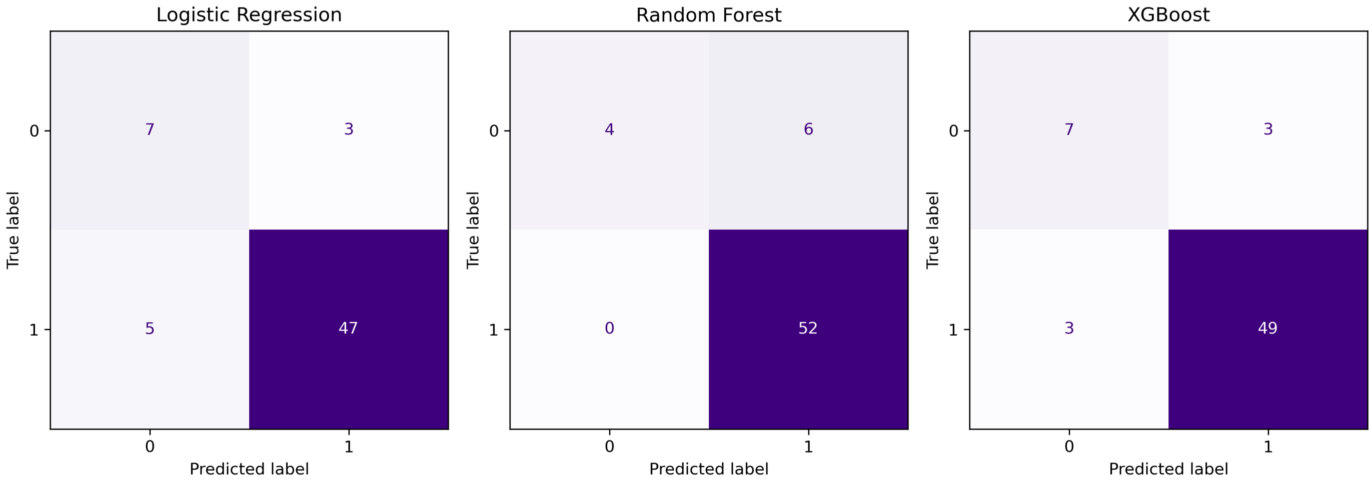

Figure 9. Confusion matrices of the Logistic Regression, Random Forest and XGBoost prediction results.

Overall, ensemble-based methods were found to slightly outperform the linear baseline using selected parameters, with the random forest model providing the best balance between discrimination ability and robustness. The results suggest that non-linear interactions between gene expression components and clinical variables play an important role in ER status prediction, while careful handling of class imbalance and regularisation is essential for reliable performance.

Part IV — Evaluation

To evaluate the performance and robustness of the selected Random Forest model for ER prediction, the model selection was executed on the training data, while evaluation depended on cross-validation, held-out data, and resampling procedures to prevent information leakage. For context, a majority-class baseline and a regularised logistic regression model were included. Cross-validated results showed that the majority baseline achieved high apparent accuracy due to class imbalance but exhibited no discriminative ability (ROC–AUC = 0.50). Logistic regression substantially improved performance, achieving a mean accuracy of 0.911 and ROC–AUC of 0.945. The Random Forest achieved the best overall performance, with a mean accuracy of 0.923, F1-score of 0.957, and ROC–AUC of 0.939, and was therefore selected for further evaluation. Performance stability was evaluated using bootstrap resampling of the ROC–AUC. The bootstrap mean ROC–AUC was 0.987 with a small standard deviation (0.008), and the 95% confidence interval ranged from 0.966 to 0.999 which indicates low variability and robust discrimination (Fig.10).

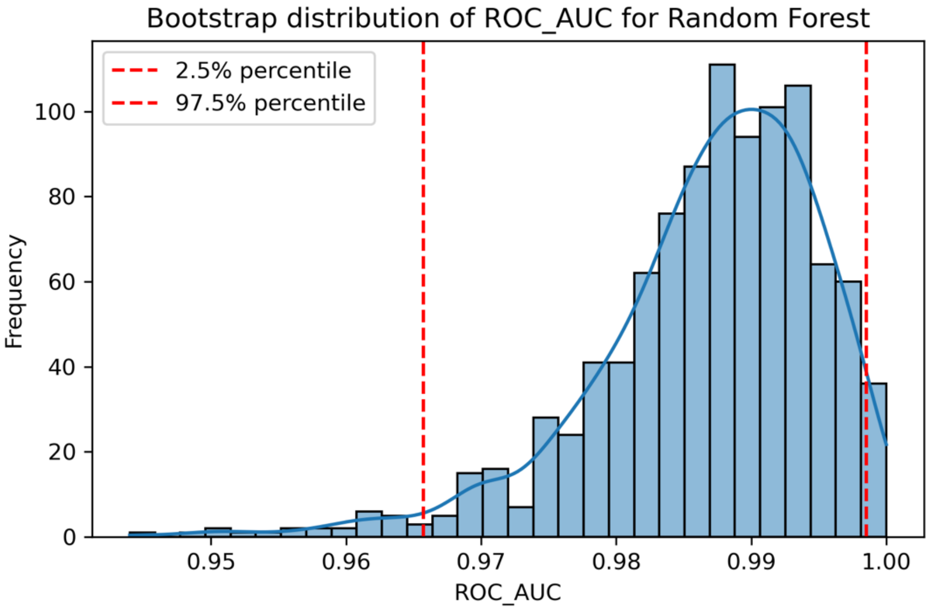

Figure 10. The histogram plot represents the Bootstrap distribution of ROC_AUC for the Random Forest model.

The confusion matrix on the held-out evaluation set showed no misclassifications, confirming strong class separation under the chosen representation. Learning curve analysis indicated that training performance remained stable across all sample sizes, while validation ROC–AUC increased steadily as more data were added (Fig.11). This suggested that generalisation performance benefited from additional samples and had not yet fully plateaued. Overall, the evaluation supported the reliability of the Random Forest model while highlighting that further data could improve robustness.

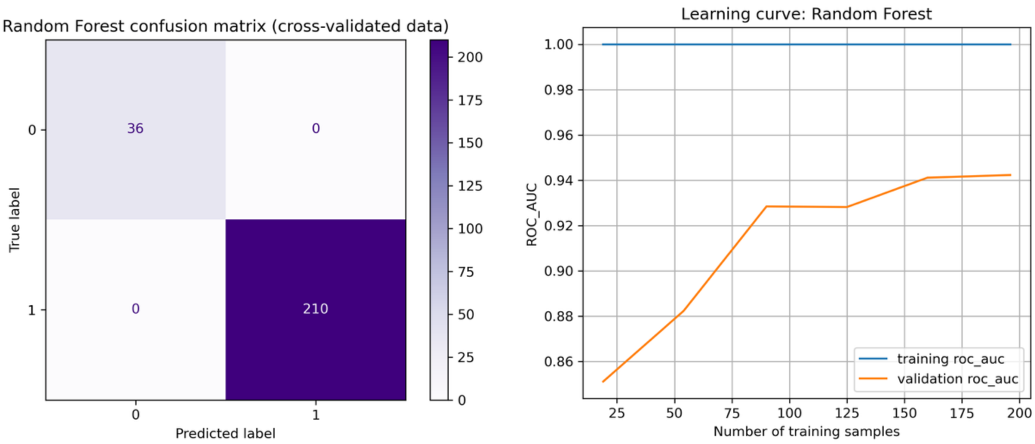

Figure 11: The confusion matrix on the held-out dataset and learning curve plots for the Random Forest model.

Part V – LLMs-augmented analysis

In this part Large Language Model Claude Sonnet 4.5 (12) was utilised as a research assistant to identify a biologically relevant gene panel for predicting an ER status. The full input prompt, output response and model version are saved in the appendix. Out of 18 suggested by LLM genes 2 were not present in the original dataset. The remaining 16 genes were reviewed and known to be associated with breast cancer metabolism and estrogen receptor regulation. Next, the performance of classifiers trained on three different feature representations was compared: the full feature set, a low-dimensional PCA representation, and a biologically motivated gene panel proposed by a large language model (LLM). The objective was to assess whether a small, knowledge-driven gene panel could achieve predictive performance comparable to data-driven feature sets when predicting ER status.

Two model families were evaluated for each feature representation: logistic regression and random forest. Performance was evaluated using accuracy, F1 score, and ROC–AUC on a held-out evaluation set, ensuring a fair and consistent comparison across all different settings.

Overall, the training results were more similar than was expected. Interestingly, the highest overall accuracy (0.919) and F1 score (0.954) were achieved with logistic regression using all available features, while the random forest reached a comparable ROC–AUC (0.940). When restricted to the LLM-proposed gene panel, performance decreased modestly but remained competitive. In particular, the random forest trained on LLM-selected genes achieved an accuracy of 0.907 and an F1 score of 0.946, indicating that much of the predictive signal was retained despite the substantial reduction in feature dimensionality. Logistic regression with LLM genes showed slightly lower accuracy and F1 scores, but a comparable ROC-AUC to that obtained with all features.

Models trained on PCA features exhibited similar behaviour: logistic regression performed comparably to the LLM gene model, while the random forest achieved a high F1 score (0.952) but a lower ROC–AUC than when trained on the full feature set. This suggests that dimensionality reduction preserved discriminative power but may have removed information relevant for ranking predictions across thresholds. The final results of the evaluation scores are presented in Table 1.

Table 1. Summary of classification performance (accuracy, F1 score, and ROC–AUC) for the Logistic Regression and Random Forest models trained using all features, PCA-derived features, and the LLM-proposed gene panel (sorted based on ROC_AUC).

Feature set	Model	Accuracy	F1	ROC_AUC
All features	Random forest	0.866	0.927	0.940
All features	Logistic regression	0.919	0.954	0.940
LLM genes	Logistic regression	0.870	0.920	0.939
LLM genes	Random forest	0.907	0.946	0.922
PCA features	Logistic regression	0.870	0.923	0.920
PCA features	Random forest	0.914	0.952	0.913

Overall, the results indicate that the LLM-proposed gene panel provides a strong and biologically relevant compromise between interpretability and performance for the ER status identification. While the full feature set yielded the best absolute results, the LLM-based approach achieved comparable predictive quality with far fewer features, supporting its potential usability for clinically gene-panel-based modelling.

Conclusion

In this project, a range of dimensionality reduction, clustering, and supervised learning approaches were applied to breast cancer gene expression data to assess how well key clinical targets are encoded in the molecular signal. Across all experiments, estrogen receptor (ER) status emerged as the most predictably encoded target in gene expression. Consistently high ROC–AUC values were achieved using multiple modelling strategies, including logistic regression, random forests, and gradient-boosted trees, indicating a strong and biologically plausible association between transcriptional profiles and ER status. In contrast, unsupervised clustering showed limited alignment with ER labels, highlighting that separability becomes apparent primarily when supervised objectives are introduced.

Dimensionality reduction by PCA preserved sufficient signal for both prediction and exploratory analysis, despite retaining only a small fraction of total variance. Furthermore, restricting models to a small, biologically relevant gene panel proposed by a large language model led to only a moderate reduction in performance compared with using all features. That suggests that a small knowledge-based set of genes captures a lot of the clinically relevant information for ER prediction. Lower quality performance, particularly for clustering or reduced feature representations, can possibly be caused by label imbalance, limited sample size, and the presence of biological heterogeneity that is not well captured by global expression patterns alone.

The next step can be including more samples or integrating additional data modalities, such as single-cell RNA-seq or spatial transcriptomic data, to improve clinical usefulness. Additionally, using more patient samples and validating models on an external cohort can further improve generalisability and better represent real-world clinical data.

Limitations

First and most obvious limitation is the dataset size. It is relatively small, which decreases the model's complexity that can be reliably trained on it. Although cross-validation, bootstrapping, and held-out validation sets were used to mitigate overfitting, performance estimates are probably too optimistic and sensitive to individual samples. This is particularly relevant for high-capacity models such as random forests and gradient-boosted trees, which can easily overfit in the case of a low number of samples.

Second, class imbalance, especially for certain clinical targets, including ER status. While this imbalance was addressed through class weighting and appropriate evaluation metrics (e.g. F1 score and ROC–AUC), it may still bias model learning towards the majority class. In addition, some evaluation metrics, such as accuracy, are less informative under imbalance and should be interpreted with caution. Future work would benefit from larger and more balanced cohorts, or from external validation datasets reflecting real-world prevalence.

Third, missing values in the patients' metadata constituted an additional source of uncertainty. Mean imputation was applied using statistics computed from the training set only, which avoided data leakage but implicitly assumed that missingness was random. However, this approach could change feature distributions and decrease the true associations if missing values were informative (for example, related to disease severity or clinical practice). More sophisticated imputation strategies, such as model-based or multiple imputation, could be explored in future studies.

Fourth, dimensionality reduction and feature selection steps introduced further constraints. PCA, while effective at compressing gene expression data, produces components that are difficult to interpret biologically. Conversely, the LLM-proposed gene panel relied on prior biological knowledge rather than dataset-specific optimisation. Although this improved interpretability and yielded competitive performance, it may not capture cohort-specific signals or novel biomarkers present in the data.

Finally, all analyses were performed within a single dataset, and no independent external data was used for validation. As a result, the model's generalisability to other populations, platforms, or clinical settings is untested. External validation is necessary before any relevant conclusions.

Future improvements could include the integration of larger multi-cohort datasets, the use of cross-validation for more robust model selection, deeper analysis and better selection of model hyperparameters and the exploration of hybrid approaches combining biologically informed gene panels with data-driven feature refinement. In addition, uncertainty-aware modelling and calibration analysis could further improve the reliability of predictions in clinically relevant settings.

Code availability

All code used for analysis is available at https://github.com/Dana162001/ML_project_AHI_2025. Note parts of the code were written with the help of ChatGPT version GPT-5.2 and GPT-5 mini.

Bibliography

1. *LIII. On lines and planes of closest fit to systems of points in space.* **F.R.S., Karl Pearson.** 11, s.l. : The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901, Vol. 2. DOI: 10.1080/14786440109462720.

2. **scikit-learn.** scikit-learn.org. *StandardScaler.* [Online] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

3. *UMAP: Uniform Manifold Approximation and Projection.* **McInnes et al.** s.l. : Journal of Open Source Software, 2018. https://doi.org/10.21105/joss.00861.

4. *Visualizing High-Dimensional Data Using t-SNE.* **van der Maaten, and Hinton, .** s.l. : Journal of Machine Learning Research, 2008.

5. *Étude comparative de la distribution florale dans une portion des Alpes et des Jura.* **Jaccard, .** s.l. : Bulletin de la Société vaudoise des sciences naturelles, 1901.

6. *The regression analysis of binary sequences.* **Cox, David R.** 2, s.l. : Wiley Online Library Journal of the Royal Statistical Society: Series B (Methodological), 1958, Vol. 20.

7. **Ho, Tin Kam.** Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition.* s.l. : IEEE, Volume 1, p. 278--282, 1995.

8. **Chen, Tianqi and Guestrin, Carlos.** XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Francisco, California, USA : ACM, 2016.

9. **Kim, .** medium.com. *smote-practical-consideration-limitations-f0d926b661a8.* [Online] 16 Dec 2023. https://medium.com/@minjukim023/smote-practical-consideration-limitations-f0d926b661a8.

10. **Tewari, .** Regularization — Understanding L1 and L2 regularization for Deep Learning. *Medium.* [Online] 9 Nov 2021. https://medium.com/analytics-vidhya/regularization-understanding-l1-and-l2-regularization-for-deep-learning-a7b9e4a409bf.

11. **Banerjee, .** Kaggle. [Online] 2020. https://www.kaggle.com/code/prashant111/xgboost-k-fold-cv-feature-importance.

12. **Anthropic, .** *Claude Sonnet 4.5.* 2025. https://platform.claude.com/docs/en/about-claude/models/overview.

Appendix

LLM Prompt

You are acting as a biomedical research assistant with expertise in breast cancer biology, hormone receptor signalling, and translational genomics. I am working with a breast cancer dataset containing 6384 gene expression features and clinical metadata from 327 patients. I want to test whether a small, biologically plausible gene panel proposed by an expert LLM can be used to predict clinically relevant outcomes using machine learning. Your task is to propose a compact gene panel of 10–20 genes that could plausibly predict one of the following targets (choose the most appropriate and state which one you chose):

- node: Involvement of lymph nodes (0=no, 1=yes)
- grade: Histologic grade (Elston-Ellis) category
- size: Tumor size (in centimeters)
- age: Age of the patient (in years)
- er: Estrogen receptor (ER) status (0=negative, 1=positive)
- relapse: Relapse after surgery (0=no, 1=yes)

Constraints

1. The gene panel must be biologically motivated, based on known mechanisms (e.g. ER signalling, proliferation, invasion, cell cycle, tumour microenvironment).
2. Genes should be well-studied and interpretable, not unknown identifiers.
3. The panel must be small (10–20 genes) to reflect a realistic clinical assay.
4. Do not lie and imagine facts

Required output format

Please return your answer in the following structured format only:

Chosen prediction target:

Proposed gene panel (10–20 genes):

* GENE1

* GENE2

* …

Biological justification (2–4 sentences per gene. If possible, add references):

* GENE1: explanation

* GENE2: explanation

* …

Expected modelling implications (short paragraph): Briefly describe why this panel should work well in linear and tree-based classifiers (e.g. signal strength, redundancy, non-linearity).

Reproducibility and logging

At the very end of your response, explicitly report the following:

LLM metadata (mandatory):

* Model name and version

* Date of generation

* Temperature (if known; otherwise state “unknown”)

* Any other relevant generation settings if available

LLM Output

Chosen prediction target:

er: Estrogen receptor (ER) status (0=negative, 1=positive)

Proposed gene panel (10–20 genes):

ESR1

PGR

GATA3

FOXA1

XBP1

TFF1

MKI67

AURKA

CCNB1

ERBB2

EGFR

VIM

CDH1

FOXC1

KRT5

KRT14

MAPT

NAT1

Biological justification:

ESR1: Encodes the estrogen receptor alpha protein itself. Direct measurement of ESR1 mRNA expression is the most direct molecular correlate of ER status and is a defining feature of luminal breast cancers. ESR1 high expression is virtually synonymous with ER-positive status (Perou et al., Nature 2000; Sørlie et al., PNAS 2001).

PGR: Encodes the progesterone receptor, which is a direct transcriptional target of ESR1. PGR expression is a hallmark of functional ER signaling and is highly correlated with ER-positive tumors. Loss of PGR in ER+ tumors can indicate endocrine resistance (Horwitz et al., Cancer Res 1975; Early Breast Cancer Trialists' Collaborative Group, Lancet 2011).

GATA3: A master transcription factor essential for luminal epithelial differentiation and maintenance of the ER+ phenotype. GATA3 is required for ESR1 expression and luminal lineage specification. Mutations and high expression are characteristic of ER+ breast cancers (Asselin-Labat et al., Nat Cell Biol 2007; Usary et al., Breast Cancer Res 2004).

FOXA1: A pioneer transcription factor that facilitates estrogen receptor binding to chromatin. FOXA1 is essential for ER-mediated transcription and is highly expressed in luminal/ER+ tumors. It helps define the ER cistrome and luminal identity (Carroll et al., Cell 2005; Hurtado et al., EMBO J 2011).

XBP1: An ER stress response transcription factor that is also an ER target gene. XBP1 is highly expressed in luminal breast cancers and contributes to ER+ tumor cell survival and hormone responsiveness (Ding et al., Cancer Res 2003; Davies et al., Breast Cancer Res 2008).

TFF1: Trefoil factor 1 (also known as pS2) is a classic ER-regulated gene and a marker of functional ER signaling. TFF1 expression is highly specific to ER+ breast cancers and is used clinically as a surrogate for ER pathway activity (Brown et al., PNAS 1984; Henry et al., Br J Cancer 1991).

MKI67: Encodes the Ki-67 protein, a marker of cellular proliferation. ER-negative tumors, particularly basal-like subtype, tend to have higher proliferation rates than ER+ luminal A tumors. Including MKI67 helps distinguish low-proliferating ER+ from high-proliferating ER- tumors (Urruticoechea et al., J Clin Oncol 2005; Cheang et al., J Natl Cancer Inst 2009).

AURKA: Aurora kinase A, involved in mitotic progression and frequently overexpressed in basal-like and ER-negative breast cancers. AURKA is associated with high-grade, aggressive tumors and is typically lower in ER+ disease (Bischoff et al., EMBO J 1998; Nadler et al., Clin Cancer Res 2008).

CCNB1: Cyclin B1, a key regulator of G2/M transition. Like MKI67 and AURKA, CCNB1 is a proliferation marker enriched in ER-negative and basal-like tumors compared to ER+ luminal tumors (Winters et al., Mol Cancer Ther 2010; Sotiriou et al., J Natl Cancer Inst 2006).

ERBB2: Encodes HER2/neu receptor tyrosine kinase. While HER2+ tumors can be ER+ or ER-, HER2 amplification/overexpression is more common in ER-negative contexts and defines a distinct subtype. HER2 expression helps distinguish ER-/HER2+ from other ER- tumors (Slamon et al., Science 1987; Perou et al., Nature 2000).

EGFR: Epidermal growth factor receptor, enriched in basal-like (triple-negative) breast cancers which are predominantly ER-negative. EGFR expression is inversely correlated with ER status and is a marker of basal phenotype (Nielsen et al., Clin Cancer Res 2004; Rakha et al., Am J Surg Pathol 2009).

VIM: Vimentin, an intermediate filament protein and marker of epithelial-to-mesenchymal transition (EMT). Vimentin expression is higher in basal-like/ER-negative tumors and lower in luminal/ER+ tumors, reflecting differences in differentiation state (Laakso et al., Clin Cancer Res 2006; Karihtala et al., BMC Cancer 2013).

CDH1: E-cadherin, an epithelial adhesion molecule. Loss or reduction of CDH1 is associated with lobular histology and can occur in both ER+ and ER- contexts, but high expression is typical of well-differentiated luminal tumors. CDH1 loss is more common in some ER- tumors (Berx et al., Hum Mol Genet 1998; Rakha et al., Mod Pathol 2010).

FOXC1: A transcription factor associated with basal-like breast cancer and invasive/metastatic phenotypes. FOXC1 is enriched in ER-negative tumors and contributes to basal stem cell characteristics. It is inversely correlated with ER expression (Ray et al., Oncogene 2010; Jensen et al., Breast Cancer Res 2015).

KRT5: Cytokeratin 5, a basal epithelial marker. KRT5 expression defines basal-like breast cancers, which are predominantly ER-negative and triple-negative. KRT5 is essentially absent in luminal/ER+ tumors (Perou et al., Nature 2000; Nielsen et al., Clin Cancer Res 2004).

KRT14: Cytokeratin 14, another basal cytokeratin marker highly specific for basal-like/ER-negative breast cancers. Like KRT5, KRT14 expression is mutually exclusive with ER+ luminal phenotype (Abd El-Rehim et al., J Pathol 2004; Livasy et al., Mod Pathol 2006).

MAPT: Microtubule-associated protein tau, recently identified as highly specific for luminal A/ER+ breast cancers. MAPT expression is strongly associated with ER positivity and good prognosis, and is virtually absent in basal-like tumors (Bonneau et al., Breast Cancer Res 2015; Li et al., Oncotarget 2016).

NAT1: N-acetyltransferase 1, identified in intrinsic subtype studies as a luminal/ER+ marker. NAT1 is consistently upregulated in ER+ tumors and has been used in gene expression signatures to classify luminal breast cancers (Sørlie et al., PNAS 2001; Parker et al., J Clin Oncol 2009).

Expected modelling implications:

This 18-gene panel is designed to provide strong discriminatory power for ER status prediction across multiple machine learning approaches. The panel includes direct ER pathway genes (ESR1, PGR, GATA3, FOXA1, XBP1, TFF1) that should provide highly correlated, strong positive signals for ER+ tumors, as well as basal markers (KRT5, KRT14, FOXC1, EGFR, VIM) that provide negative signals (inverse correlation). Proliferation markers (MKI67, AURKA, CCNB1) capture the typically lower proliferation of ER+ luminal A tumors. This combination creates both linear separability (sum of luminal markers minus basal markers) and opportunities for non-linear interactions that tree-based models can exploit (e.g., ESR1-high AND KRT5-low definitively indicates ER+). The biological redundancy within marker groups (multiple basal keratins, multiple ER targets) provides robustness against measurement noise while maintaining interpretability. The panel avoids purely data-driven but biologically obscure features, ensuring that model predictions can be mechanistically explained.

LLM metadata:

Model name and version: Claude Sonnet 4.5 (claude-sonnet-4-5-20250929)

Date of generation: December 27, 2025

Temperature: unknown (system default)

Other settings: Standard conversation mode, no specialized parameters applied