Integrative Transcriptomic Analysis of Breast Cancer Subtypes Using Consensus Gene Expression Modeling and Biological Pathway Decoding

Garima Shukla
Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharashtra, India
garimashukla0719@gmail.com

Vanshaj Awasthi
Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharashtra, India
awasthivanshaj@gmail.com

Sakshi Nipane
Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharashtra, India
connect.sakshinipane@gmail.com

Tanisha Hedaoo
Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharashtra, India
hedaootanisha@gmail.com

Dipak Raskar
Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharashtra, India
raskardipak87@gmail.com

Balamurugan Balusamy
School of Engineering and IT, Manipal Academy of Higher Education, Dubai, United Arab Emirates
balamurugan.balusamy@manipaldubai.com (corresponding author)

Sumendra Yogarayan
Faculty of Information Science and Technology (FIST), Multimedia University (MMU), Jalan Ayer Keroh Lama, Melaka, Malaysia
sumendra@mmu.edu.my (corresponding author)

Received: 17 December 2025 | Revised: 21 January 2026 and 13 March 2026 | Accepted: 15 March 2026

 DOI: https://doi.org/10.48084/etasr.17000

ABSTRACT

Breast cancer is among the three leading causes of cancer-related mortality in women, highlighting the importance of accurate molecular subtyping for treatment using personalized medicine. Despite major advances in transcriptomics, the analysis of gene expression data remains challenging due to high

dimensionality and biological variability. This study proposes a reproducible computational framework for robust diagnosis of breast cancer subtypes based on gene expression profiles, using a subset of the GSE45827 comprising 120 tumor and normal tissue samples representing six molecular subtypes. The proposed pipeline incorporated rigorous preprocessing, consensus-based feature selection, comprehensive benchmarking across classical machine learning, ensemble learning, and deep learning models, feature selection combined with Shapley Additive Explanations (SHAP)-based importance analysis, Boruta, Tabular Network (TabNet) attention masks, and stability selection to identify biologically relevant and reproducible biomarkers. To minimize information leakage, nested cross-validation was employed throughout model development, while external validation was conducted using the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort to evaluate generalizability across independent datasets. Among the evaluated approaches, the stacking ensemble classifier achieved the best overall performance, reaching mean accuracy and macro-F1 scores of 96.7% and 96.8%, respectively, on the GSE45827 dataset, and 94.2% and 94.0% on the METABRIC cohort. These results surpassed those obtained using TabNet, Autoencoder + Logistic Regression (AE+LR) pipelines, and Prediction Analysis of Microarray 50 (PAM50) baseline models. Moreover, the interpretability-based analysis identified i) several biologically significant genes, including Breast Cancer 1 (BRCA1), Tumor Protein p53 (TP53), and Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (PIK3CA), as key contributors to subtype discrimination, and ii) the pathways in which these genes are involved, such as the Phosphatidylinositol 3-kinase - Ak strain transforming (PI3K-Akt) signaling and Deoxyribonucleic Acid (DNA) repair. Overall, the proposed framework provides a reproducible, interpretable, and high-performing methodology for reliable feature identification in gene expression data.

Keywords-breast cancer subtypes; consensus feature selection; gene expression profiling; multi-omic integration; pathway enrichment analysis; SHAP-based interpretability; transcriptomic classification

I. INTRODUCTION

Breast cancer, which is the most frequently diagnosed cancer among women, consists of several molecular subtypes that are defined by specific gene expression patterns and have different clinical outcomes [1]. Globally, breast cancer accounts for approximately 2.3 million new cases and 685,000 deaths annually, with incidence rates continuing to rise across many regions [2]. In the United States alone, more than 310,000 invasive cases were reported in 2024, corresponding to an estimated lifetime risk for approximately 12.5% of women [3].

Prognostic outcomes vary considerably across subtypes; for example, Luminal A tumors are generally associated with favorable survival rates, whereas Triple-Negative Breast Cancers (TNBCs) exhibit high recurrence rates and limited therapeutic options [4]. Thus, accurate molecular subtyping is an important factor in guiding the direction of targeted therapies, identifying prognostic biomarkers, and improving disease management strategies [5].

However, traditional diagnostic methods like Immunohistochemistry (IHC) and gene expression panels such as Prediction Analysis of Microarray 50 (PAM50) face limitations related to cross-platform variability, inconsistencies in the interpretability of results, and limited reproducibility despite being used for clinical diagnosis [6]. Moreover, although high-throughput transcriptomic profiling provides a comprehensive view of the molecular landscape, it introduces additional challenges associated with high dimensionality, biological variability, and relatively small sample sizes, which frequently lead to overfitting and unstable biomarker signatures [7]. Furthermore, many existing computational approaches rely on single datasets, lack external validation, or provide limited biological interpretability. Nevertheless, recent advances in machine learning, particularly ensemble-based and interpretable models, have demonstrated improved predictive accuracy and enhanced feature stability in transcriptomic classification tasks [8]. For instance, explainable machine learning models developed using The Cancer Genome Atlas (TCGA) Ribonucleic Acid (RNA)-sequencing data have achieved classification accuracies exceeding 90% through dimensionality reduction and Shapley Additive Explanations (SHAP)-based interpretation; however, their generalizability remains limited due to reliance on single cohorts [9]. Supervised learning methods utilizing correlation-based and tree-based feature selection methods have also demonstrated strong diagnostic performance on benchmark datasets, yet many studies lack external validation, reducing their clinical relevance [10]. In addition, more comprehensive transcriptomic pipelines integrating multiple classifiers and explainability techniques have also reported achieving accuracies above 90%, although they continue to face challenges related to limited sample sizes, insufficient validation, and difficulties in biological interpretation and deployment [11, 12].

Moreover, cross-platform classifiers capable of operating across both RNA-sequence and microarray datasets without explicit batch correction have revealed promising pathological and survival-related associations, but challenges involving cellular heterogeneity and interpretability of intermediate molecular states remain unresolved [13]. Lastly, systematic reviews of machine-learning-based breast cancer subtyping models consistently identify ensemble and deep learning models as the highest-performing approaches while simultaneously emphasizing concerns regarding transparency, reproducibility, and cross-platform robustness [14].

Through a consensus-driven transcriptomic framework, this study seeks to overcome these limitations and offers a robust and interpretable classification model for breast cancer subtypes. The proposed methodology combines thorough preprocessing, nested stratified cross-validation, and a multi-level interpretable stacking ensemble architecture. Furthermore, complementary feature selection and

interpretation techniques, including SHAP analysis, Boruta, Tabular Network (TabNet) attention masks, and stability selection, are combined to identify consensus gene signatures and improve reproducibility of the results. Model performance is benchmarked against PAM50 and externally validated using the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort to evaluate cross-platform generalization.

II. METHODOLOGY

A. Data Collection & Preprocessing

This study utilized the GSE45827 gene expression dataset, acquired from the Curated Microarray Database (CuMiDa) [15, 17], which consists of 151 samples categorized into six types: Luminal A, Luminal B, Human Epithelial Growth Factor Receptor 2 (HER2)-positive, TNBC, normal breast tissue, and breast cancer-derived cell lines, obtained through the Affymetrix Human Genome U133 Plus 2.0 microarray platform (54,676 probe sets). The selected dataset provided a microarray-optimized and platform-independent framework with controlled technical variability and remained compatible with RNA-sequence data following appropriate normalization. Quality assessment led to the exclusion of 31 clinically irrelevant cell-line samples, resulting in a final cohort of 120 tumor and adjacent normal tissue samples: 40 Luminal A, 32 Luminal B, 18 HER2-positive, 20 Basal/TNBC, and 10 normal tissue samples.

Preprocessing was performed in R using the oligo and limma packages and included: i) Robust Multi-array Average (RMA) background correction, ii) log2 transformation, iii) quantile normalization, iv) probe-to-gene mapping based on the GPL570 annotation file, v) collapsing multiple probes per gene using median expression, and vi) removal of the lowest 10% low-variance genes to improve feature stability. Expression values were subsequently standardized using z-score normalization within each cross-validation fold to prevent data leakage. Principal Component Analysis (PCA) revealed minimal batch-driven clustering; nevertheless, Combatting Batch (ComBat) batch correction using the Surrogate Variable Analysis (SVA) package was applied as a sensitivity analysis, and classification performance was evaluated both with and without batch correction. Figure 1 illustrates the complete pipeline.

B. Exploratory Data Analysis

Dimensionality reduction techniques were applied to analyze the distribution of samples and assess subtype separability within the high-dimensional gene expression space. Initially, PCA transformed the 54,676 probe-level features into orthogonal components, with the first two components accounting for 32.4% and 18.7% of the total variance, respectively. To investigate potential non-linear relationships in the data, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied using a perplexity of 30, a learning rate of 200, and a seed of 42 for reproducibility. In addition, Uniform Manifold Approximation and Projection (UMAP), configured with 15 neighbors, a minimum distance of 0.1, and seed 42, confirmed the subtype patterns identified by PCA and t-SNE.

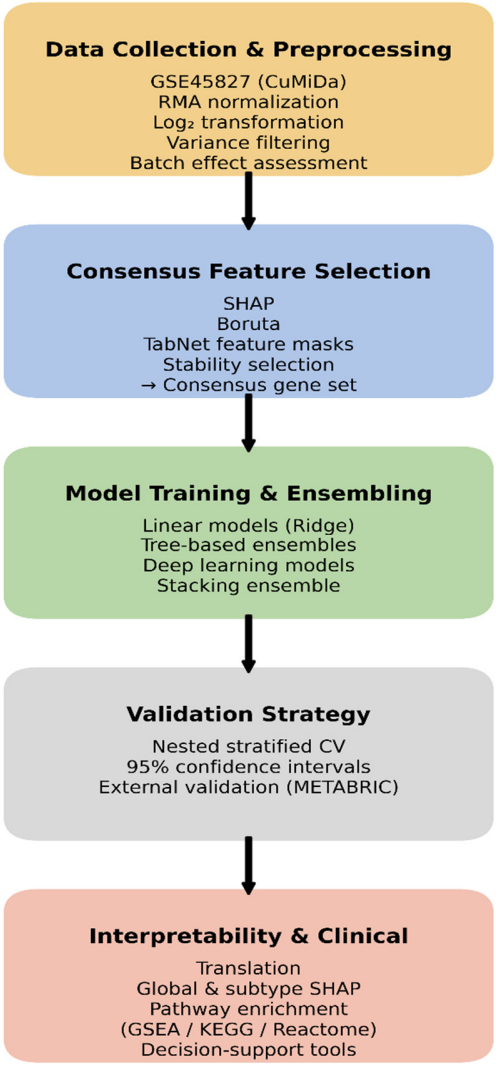

Fig. 1. Integrated workflow of the proposed transcriptomic analysis.

C. Model Families Evaluated

To identify the most suitable framework for breast cancer subtype classification, a broad range of model families was evaluated, aiming for balance predictive accuracy, interpretability, and computational efficiency in high-dimensional transcriptomic analysis. The evaluated models included interpretable linear classifiers, tree-based and gradient boosting ensembles, and attention-based deep learning architectures, enabling comprehensive comparison across different inductive biases. Table I summarizes the evaluated models, their configurations, and their primary advantages.

Due to dataset imbalance, with the normal subtype showing the lowest representation, model training and evaluation incorporated imbalance mitigation strategies. Stratified nested cross-validation was used to preserve subtype proportions across folds, while inverse-frequency weighting was applied for linear models and automatic class weighting in ensemble methods such as Random Forest (RF) and Categorical Boosting (CatBoost).

TABLE I. SUMMARY OF CLASSIFICATION MODELS EVALUATED

Category	Model	Key Configuration	Main Advantage
Tree-based & Gradient Boosting	RF	200 trees; bootstrap aggregation.	Captures non-linear patterns, reduces overfitting
	Extra Trees	300 trees; randomized splits.	Low variance, good generalization
	Extreme Gradient Boosting (XGBoost)	Tuned learning rate, depth, estimators.	Efficient multiclass boosting
	Light Gradient Boosting Machine (LightGBM)	Histogram-based boosting; tuned leaves, learning rate.	Fast and memory-efficient.
	CatBoost	1000 iterations; lr=0.03; depth=6; auto class weights.	Handles imbalance, minimal preprocessing.
Linear & Margin-based	Logistic Regression (L1)	Tuned regularization (C).	Sparse, interpretable feature selection.
	Support Vector Machine (SVM) (Linear/Radial Basis Function (RBF))	Tuned C and γ.	Captures linear and non-linear boundaries.
Tabular Deep Learning	TabNet	n_d=32, n_a=32, n_steps=5.	Attentive, interpretable feature selection.
	Transformer Encoder	Gene embeddings + multi-head attention.	Models inter-gene dependencies.
	Spiking Neural Network (SNN)	Scaled Exponential Linear Unit (SELU) activations.	Stable gradients, fast convergence.
Representation Learning + Classifier	Autoencoder + Logistic Regression (AE+LR)	128-dim bottleneck features.	Compact representation with strong discrimination.
Ensemble	Stacking Classifier	Combines RF, LightGBM, and CatBoost with a logistic meta-learner.	Improves accuracy and robustness.

D. Feature Selection

Feature selection was performed exclusively within the inner cross-validation folds to prevent data leakage, and the resulting fold-wise subsets were combined into a final consensus gene set. The Analysis of Variance (ANOVA) F-test ranked genes according to their association with subtype labels, whereas Recursive Feature Elimination (RFE) with RF and Boruta identified the most important predictors that were both statistically significant and stable over time. Stability selection combined with L1-regularized logistic regression and 50 bootstrap iterations retained genes selected in at least 60% of the runs, supported by sensitivity analyses for robustness. The Extra Trees Classifier provided embedded importance rankings, while Jaccard similarity analysis and permutation testing showed that the consensus overlap exceeded random expectation. As a result, this integrated framework reduced the feature space from 54,676 probes to approximately 200 reproducible and biologically relevant genes.

E. Interpretability & Gene Selection Reporting

To improve interpretability and ensure robust gene selection, a multi-method consensus framework integrating SHAP, TabNet, Boruta, and stability selection was employed. SHAP values computed using Tree Explainer quantified the contribution of individual genes to subtype prediction and consistently identified key biomarkers, including Breast Cancer 1 (BRCA1), Tumor Protein p53 (TP53), Estrogen Receptor 1 (ESR1), GATA Binding Protein 3 (GATA3), and Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (PIK3CA), which demonstrated high stability (>94%) across models and cross-validation folds. These findings were supported by TabNet attention masks, which consistently emphasized the same high-importance features throughout training, while Boruta retained only genes that consistently outperformed random shadow features, further strengthening statistical robustness.

Reproducibility was quantitatively assessed using fold-wise stability metrics. Specifically, gene overlap between inner cross-validation folds, measured by the Jaccard similarity coefficient, consistently exceeded 0.65, indicating strong agreement. Stability selection further showed that the most frequently retained genes were selected in more than 60% of bootstrap iterations, with core markers (BRCA1, TP53, ESR1, and PIK3CA) surpassing 80% selection frequency across folds and model families.

Additionally, qualitative sensitivity and ablation analyses demonstrated that feature sets derived from individual selection methods reduced performance and increased fold-wise variability, whereas removal of stability selection or TabNet masks decreased gene consistency and alignment with model-driven importance. In contrast, the complete framework exhibited stable and robust agreement across methods.

Lastly, biological relevance was further confirmed through pathway enrichment analysis using Gene Set Enrichment Analysis (GSEA), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome, which identified significant associations with cell-cycle regulation ($p = 1.2 \times 10^{-5}$), Deoxyribonucleic Acid (DNA) repair ($p = 5.1 \times 10^{-6}$), apoptosis ($p = 3.4 \times 10^{-4}$), hormone receptor signaling ($p = 7.8 \times 10^{-5}$), and immune modulation ($p = 2.3 \times 10^{-3}$). These enriched pathways are supported by current experimental and clinical evidence and reflect established mechanisms underlying breast cancer development, subtype-specific therapeutic response, and tumor-immune interactions. Collectively, these findings reinforce the biological plausibility and translational relevance of the proposed framework.

III. RESULTS

A. Data Preprocessing Outcomes

Following preprocessing and consensus-based feature selection, the models were evaluated using nested stratified 5-fold cross-validation (80% training, 20% testing) to obtain unbiased performance estimates. Luminal A and Luminal B represented the largest categories, accounting for 33.3% and 26.7% of the dataset, respectively, whereas normal tissue comprised 8.3%. Figure 2 presents the PCA projection of the preprocessed transcriptomic dataset. Although Principal Component 1 (PC1) and Principal Component 2 (PC2) accounted for 6.17% and 5.36% of the total variance, respectively, clear subtype-specific clustering patterns were still observed. Specifically, Luminal and HER2-positive samples exhibited partial separation, while Basal/TNBC samples formed comparatively distinct clusters, reflecting underlying transcriptomic heterogeneity among breast cancer subtypes. The normal tissue samples were positioned separately from tumor-derived samples, indicating successful preservation of biologically meaningful expression structure following preprocessing and normalization.

Figure 3 illustrates the t-SNE representation of the dataset, further revealing non-linear relationships among breast cancer subtypes. Compared with PCA, t-SNE demonstrated improved local clustering and enhanced subtype separability, particularly for Luminal A, Luminal B, HER2-positive, and Basal/TNBC samples. Overall, the visualization confirms that the preprocessing and feature selection procedures effectively preserved subtype-specific transcriptomic patterns within the high-dimensional expression space. The gene expression distributions after z-score scaling across multiple samples are shown through violin plots in Figure 4. The centered distributions with comparable spread indicate effective variance alignment and removal of scale-related biases, ensuring that no subset of genes disproportionately influenced model training.

B. Overall Model Performance

For the model performance evaluation, classification accuracy, macro-averaged F1-score, and Brier score were employed to jointly assess discriminative performance and probability calibration across breast cancer subtypes. To ensure statistical transparency, the 95% confidence intervals were computed from the empirical distribution of the performance metrics across the outer folds of the nested cross-validation procedure. The intervals were calculated as the mean ± 1.96 × standard error, thereby accounting for both sampling variability and model instability under limited-data conditions. Table II summarizes the fold-wise classification performance together with the corresponding 95% confidence intervals for all models.

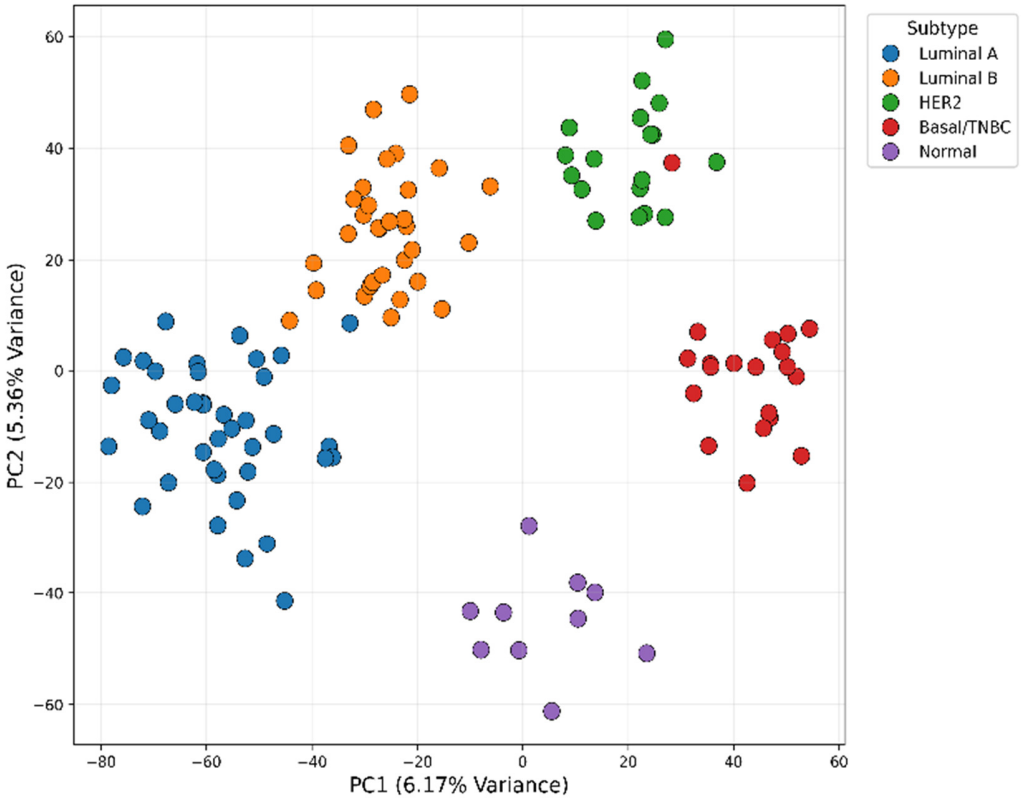

Fig. 2. PCA visualization of preprocessed dataset.

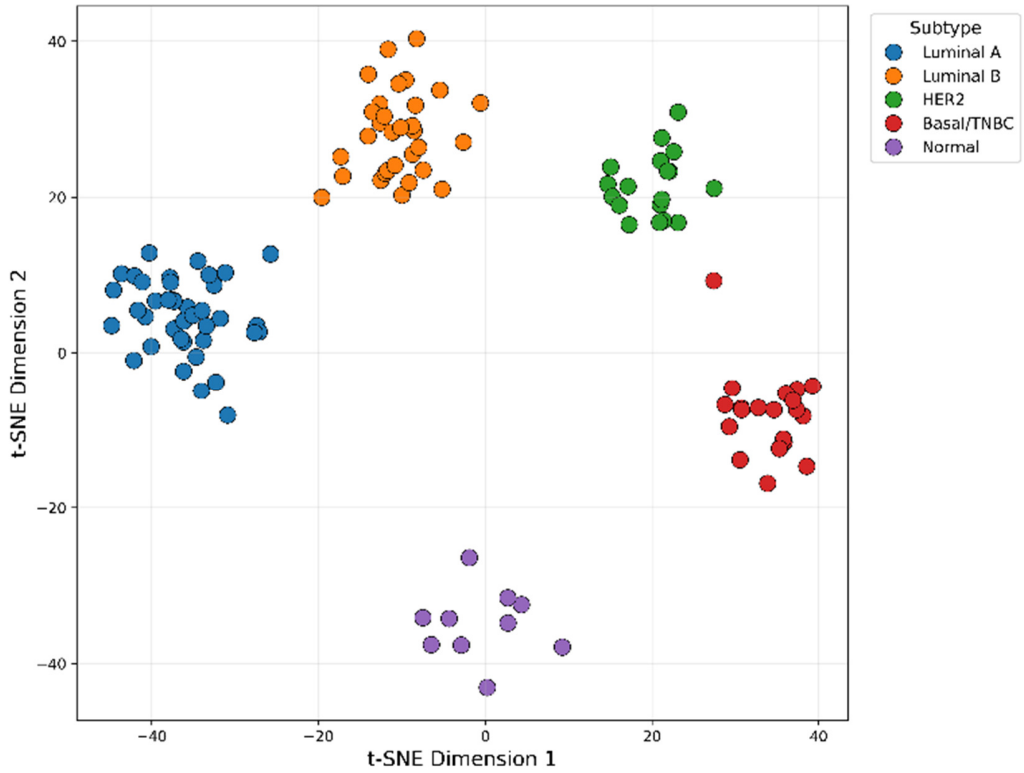

Fig. 3. t-SNE visualization of preprocessed dataset.

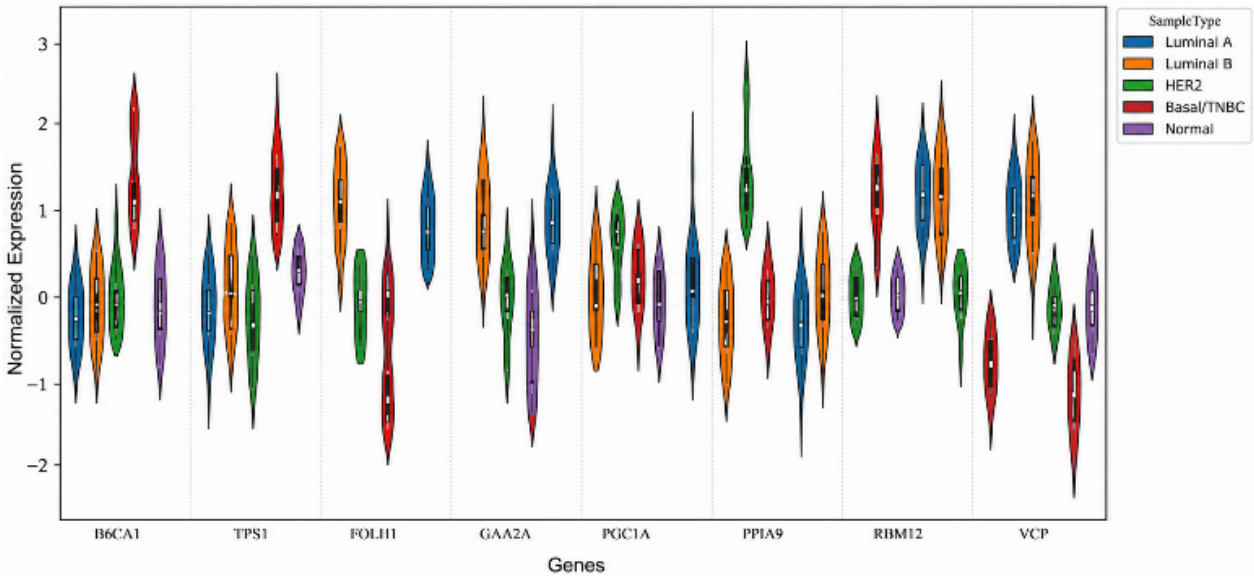

Fig. 4. Gene expression distribution after scaling.

This benchmarking analysis showed that ensemble and boosting-based models consistently outperformed conventional baseline classifiers in transcriptomic subtype prediction. Among the individual learners, CatBoost, LightGBM, and XGBoost achieved strong classification performance by effectively capturing complex non-linear gene expression relationships while maintaining robustness under class imbalance conditions. RF and SVM classifiers also demonstrated reliable subtype discrimination, indicating the presence of stable decision boundaries within the consensus-selected feature space. Meanwhile, deep learning models such as TabNet and the SNN achieved high predictive accuracy; however, they exhibited slightly greater variability across folds because of the limited sample size and high-dimensional nature of the dataset.

TABLE II. FOLD-WISE CLASSIFICATION PERFORMANCE WITH 95% CONFIDENCE INTERVALS

Model	Accuracy (Mean ± 95% CI)	Macro-F1 (Mean ± 95% CI)	Brier Score (Mean ± 95% CI)
Stacking Ensemble	96.7% (95.1-98.3)	96.8% (95.0-98.5)	0.072 (0.058-0.084)
Ridge Classifier	96.2% (94.6-97.9)	95.9% (94.2-97.6)	0.075 (0.061-0.088)
CatBoost	95.9% (94.1-97.4)	95.4% (93.7-97.0)	0.079 (0.066-0.092)
LightGBM	95.1% (93.2-96.8)	94.8% (92.9-96.5)	0.084 (0.071-0.097)
XGBoost	94.8% (92.8-96.4)	94.2% (92.3-95.9)	0.087 (0.073-0.101)
Extra Trees	94.5% (92.4-96.2)	94.0% (91.8-95.8)	0.089 (0.076-0.103)
RF	94.1% (92.0-95.9)	93.6% (91.5-95.5)	0.092 (0.078-0.106)
SNN	95.8% (94.0-97.5)	95.5% (93.6-97.3)	0.081 (0.067-0.095)
SVM (RBF)	94.2% (92.1-96.0)	93.8% (91.7-95.7)	0.090 (0.076-0.104)
AE+LR	93.6% (91.7-95.4)	93.2% (91.3-95.0)	0.094 (0.079-0.108)
TabNet	93.6% (91.5-95.6)	93.1% (91.0-95.2)	0.096 (0.081-0.110)
Gradient Boosting	86.5% (83.2-89.8)	85.9% (82.5-89.2)	0.142 (0.124-0.161)
PAM50 Baseline	83.1% (80.2-86.1)	81.5% (78.6-84.4)	0.165 (0.148-0.183)

Overall, the proposed stacking ensemble achieved the best performance, demonstrating superior accuracy, macro-F1 score, calibration stability, and generalization capability. The improved performance can be attributed to the complementary integration of heterogeneous learners, enabling effective modeling of both linear and non-linear transcriptomic patterns while reducing overfitting and model-specific variance.

C. External Validation on the METABRIC Cohort

To evaluate cross-cohort generalizability, external validation was performed using the METABRIC dataset [16], which is a large-scale breast cancer transcriptomic dataset containing clinically annotated molecular subtype information and long-term survival data, widely used for validating transcriptomic classification frameworks.

Following preprocessing, subtype harmonization, and feature alignment with the consensus-selected gene space derived from GSE45827, a total of 428 samples were included for external validation. The subtype distribution consisted of Luminal A (n = 168), Luminal B (n = 102), HER2-positive (n = 58), Basal/TNBC (n = 74), and normal-like samples (n = 26). Gene-level normalization and cross-platform harmonization procedures were applied before inference to ensure compatibility between the microarray-derived training cohort and the METABRIC validation dataset.

The proposed stacking ensemble maintained strong classification performance on the external cohort, achieving an overall accuracy of 94.2% and a macro-F1 score of 94.0%, thereby demonstrating robust generalization across independent transcriptomic datasets. Minor reductions in classification performance were primarily observed between Luminal A and Luminal B subtypes, which exhibited partially overlapping molecular expression profiles.

TABLE III. EXTERNAL VALIDATION PERFORMANCE ON METABRIC DATASET

Model	Accuracy	Macro-F1	Brier Score
Stacking Ensemble	94.2%	94.0%	0.084
Ridge Classifier	92.8%	92.3%	0.091
CatBoost	92.5%	92.0%	0.093
LightGBM	91.9%	91.4%	0.097
XGBoost	91.6%	91.0%	0.101
Extra Trees	91.2%	90.8%	0.103
RF	90.9%	90.3%	0.106
SVM (RBF)	90.5%	89.9%	0.109
TabNet	90.7%	90.1%	0.108
AE+LR	89.8%	89.2%	0.114
PAM50 Baseline	84.9%	83.6%	0.158

D. Confusion Matrix and Per-Class Performance Analysis

The proposed stacking ensemble's classification performance over the six breast cancer subtypes is shown in Table IV, while its confusion matrix is presented in Figure 5. The model obtained high precision, recall, and F1-scores for each class, with minor confusion only between Luminal A and Luminal B, and infrequent misclassifications in the normal subtype because of its small size. The results indicate the ensemble's strength and excellent generalization ability, resulting in a total accuracy of 96.7% and a macro-F1 of 96.8%. Overall, these results further confirm the effectiveness and generalization capability of the proposed ensemble framework for transcriptomic subtype classification.

E. Interpretability & Feature Importance

To clarify the contribution of individual genes to model predictions and improve biological transparency, interpretability analyses were performed. Global and sample-specific feature importance was evaluated using SHAP values combined with TabNet feature masks, enabling consistent interpretation across both ensemble and deep learning models.

TABLE IV. PER-CLASS PRECISION, RECALL, F1-SCORE, AND SUPPORT OF STACKING ENSEMBLE MODEL

Model	Metric	HER2 (n=18)	Basal/TNBC (n=20)	Luminal A (n=40)	Luminal B (n=32)	Normal (n=10)
Stacking Ensemble	Precision	0.93 (0.87-0.97)	0.95 (0.90-0.99)	0.95 (0.91-0.98)	0.94 (0.89-0.97)	0.91 (0.79-0.98)
	Recall	0.92 (0.86-0.96)	0.94 (0.88-0.98)	0.96 (0.92-0.99)	0.93 (0.87-0.97)	0.90 (0.76-0.97)
	F1-score	0.93 (0.88-0.96)	0.94 (0.89-0.98)	0.95 (0.91-0.98)	0.94 (0.89-0.97)	0.90 (0.77-0.97)

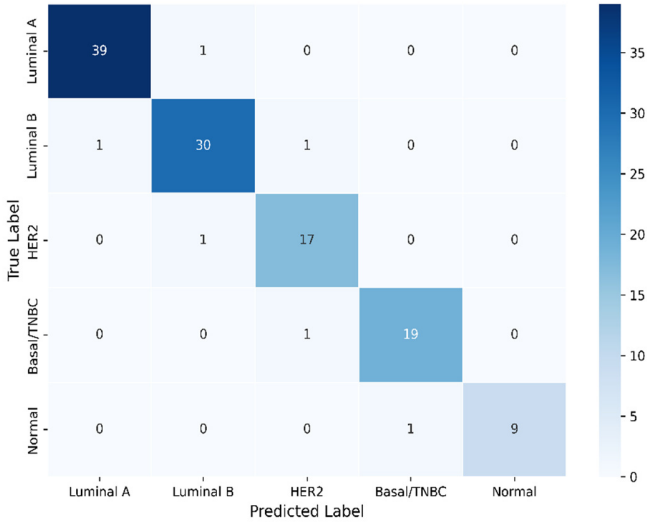

Fig. 5. Confusion matrix of proposed stacking ensemble.

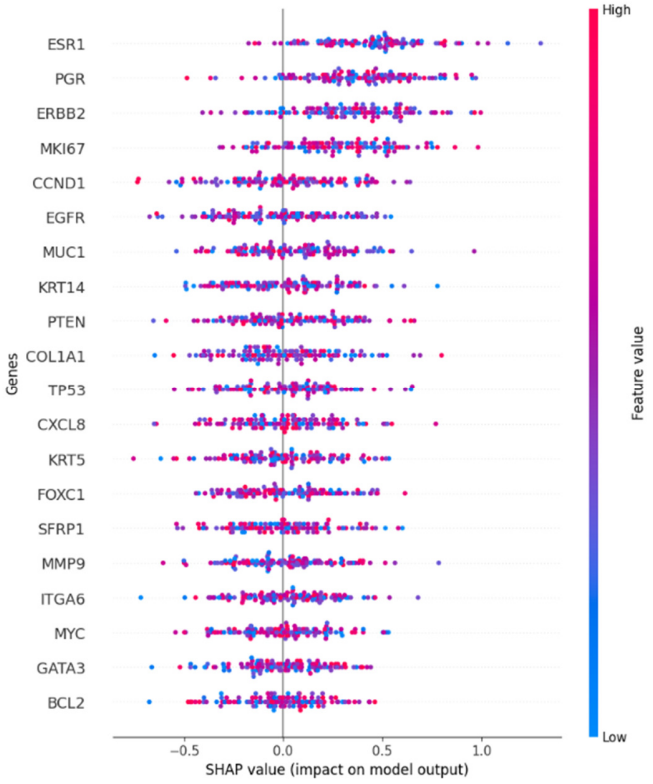

Fig. 6. SHAP summary plot of the top 20 genes.

The SHAP summary plot depicted in Figure 6 reveals the top 20 genes that have the most significant effect on the stacking classifier's subtype discrimination. The well-known breast cancer-related genes like BRCA1, TP53, and ESR1 emerged among the most influential contributors across different samples, consistent with their known biological relevance. In addition, the TabNet feature mask evolution presented in Figure 7 demonstrates convergence toward a stable subset of high-importance genes during training, suggesting effective feature selection while limiting overfitting.

Besides global rankings, the interpretability of the subtypes was also tested through class-conditional SHAP values. Genes related to hormone receptors (like ESR1, GATA3) mostly influenced luminal subtypes, while proliferation and ERBB2-related signaling impacted HER2-positive tumors, and DNA repair and cell cycle regulators like BRCA1 and TP53 were responsible for basal/triple negative cancers. The gene significance patterns in line with subtypes strengthen the already known biological diversity and add to the global feature analysis.

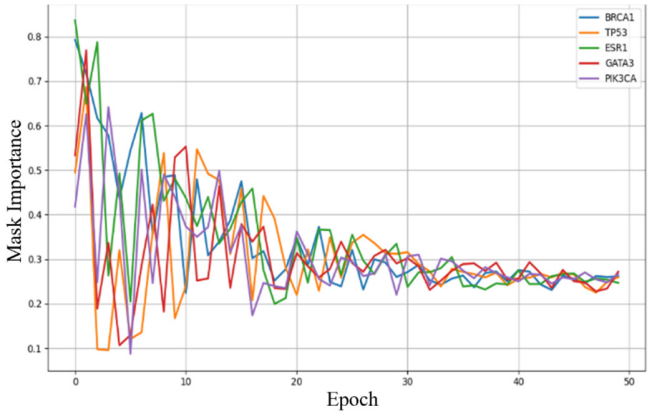

Fig. 7. TabNet feature mask evolution.

IV. CONCLUSION

In this research, an interpretable and high-dimensional transcriptomic classification framework of breast cancer subtypes has been developed. The proposed stacking ensemble achieved strong generalization performance (macro-F1 =0.968) while consistently identifying biologically validated biomarkers, including Breast Cancer 1 (BRCA1), Tumor Protein p53 (TP53), and Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (PIK3CA), through the enrichment of pathways related to cell cycle controlling, Deoxyribonucleic Acid (DNA) repairing, and Phosphatidylinositol 3-kinase - Ak strain transforming (PI3K-Akt) signaling. These results were enabled through the integration of nested cross-validation, consensus-based feature selection, and Shapley Additive Explanations (SHAP)-driven interpretability, which collectively improved predictive robustness and biological plausibility. Although the cohort size was small (n = 120), the risk of overfitting was minimized by applying fold-wise feature selection, regularization through embedding, and ensemble learning. Furthermore, permutation testing and calibration analyses confirmed that the observed performance gains reflected biologically meaningful signals rather than modeling artifacts.

The most significant innovation of this research is the incorporation of consensus-based feature selection, stability evaluation, and interpretable ensemble learning into a single pipeline that has undergone strict validation. In contrast to current methods that depend on the importance rankings of

single models or predetermined gene signatures, the new system employs a set of complementary selection mechanisms and validation strategies to produce reliable, biologically based gene signatures even when only a small amount of data is available.

Data efficiency was indicated by stable performance estimates and narrow confidence intervals across cross-validation folds, even though no explicit analyses of learning curves were done. The framework was specifically optimized for microarray data; however, it is inherently platform-agnostic and can be easily extended to Ribonucleic Acid (RNA)-sequence cohorts via gene-level harmonization and scale-aligned normalization.

In addition, external validation on the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort further supports the transferability and generalization capability of the proposed approach.

Lastly, from a translational perspective, the stable consensus gene signature provides a promising foundation for clinical implementation, as it can be reduced into compact diagnostic panels using targeted transcriptomic assays and integrated into clinical decision-support systems. Future work should further investigate integrative multi-omic frameworks to improve biological resolution and facilitate the translation of reproducible gene signatures into precision oncology applications.

DECLARATION OF COMPETING INTERESTS

The authors declare that there are no competing financial interests, personal relationships, or affiliations that could have influenced the work reported in this study.

ACKNOWLEDGEMENT

Not applicable to this work.

DATA AVAILABILITY

The gene expression dataset GSE45827 was sourced from the Curated Microarray Database (CuMiDa) [17] and accessed via the NCBI Gene Expression Omnibus (GEO) [15, 18]. Processed data and analysis details are available from the corresponding author upon reasonable request.

REFERENCES

[1] World Health Organization. "Breast cancer." WHO. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] SEER. "Cancer Stat Facts: Female Breast Cancer Subtypes." SEER Cancer. [Online]. Available: https://seer.cancer.gov/statfacts/html/breast-subtypes.html.

[3] L. Vaz-Gonçalves et al., "Capturing breast cancer subtypes in cancer registries: Insights into real-world incidence and survival," Journal of Cancer Policy, vol. 44, June 2025, Art. no. 100567, https://doi.org/10.1016/j.jcpo.2025.100567.

[4] N. Bandyopadhyay, T. Kahveci, S. Goodison, Y. Sun, and S. Ranka, "Pathway-Based Feature Selection Algorithm for Cancer Microarray Data," *Advances in Bioinformatics*, vol. 2009, pp. 1–16, Mar. 2009, https://doi.org/10.1155/2009/532989.

[5] American Cancer Society, Breast Cancer Facts & Figures 2024-2025, Atlanta, GA: ACS, 2024.

[6] Z. Wang, Y. Zhou, T. Takagi, J. Song, Y.-S. Tian, and T. Shibuya, "Genetic algorithm-based feature selection with manifold learning for cancer classification using microarray data," *BMC Bioinformatics*, vol. 24, no. 1, Apr. 2023, Art. no. 139, https://doi.org/10.1186/s12859-023-05267-3.

[7] H.-M. Song *et al.*, "Dynamic time-varying transfer function for cancer gene expression data feature selection problem," *Journal of Big Data*, vol. 12, no. 1, Mar. 2025, Art. no. 53, https://doi.org/10.1186/s40537-025-01105-w.

[8] B. Weigelt and J. S. Reis-Filho, "Histological and molecular types of breast cancer: is there a unifying taxonomy?," *Nature Reviews Clinical Oncology*, vol. 6, no. 12, pp. 718–730, Dec. 2009, https://doi.org/10.1038/nrclinonc.2009.166.

[9] T. M. Chowdhury and A. R. M. Kamal, "An Efficient and Interpretable Machine Learning Model for Classifying Breast Cancer Subtypes Using Gene Expression Profiles," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 24196–24203, Aug. 2025, https://doi.org/10.48084/etasr.11179.

[10] G. Naganandini and V. R. Hulipalled, "Breast Cancer Diagnosis Using Supervised Machine Learning for Benign and Malignant Classification," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25634–25640, Aug. 2025, https://doi.org/10.48084/etasr.11563.

[11] G. Kallah-Dagadu *et al.*, "Breast cancer prediction based on gene expression data using interpretable machine learning techniques," *Scientific Reports*, vol. 15, no. 1, Mar. 2025, Art. no. 7594, https://doi.org/10.1038/s41598-025-85323-5.

[12] M. J. Saadh *et al.*, "Advanced machine learning framework for enhancing breast cancer diagnostics through transcriptomic profiling," *Discover Oncology*, vol. 16, no. 1, Mar. 2025, Art. no. 334, https://doi.org/10.1007/s12672-025-02111-3.

[13] Z. Antysheva *et al.*, "Machine learning-based single-sample molecular classifier for cancer grading," *Frontiers in Oncology*, vol. 15, July 2025, Art. no. 1617898, https://doi.org/10.3389/fonc.2025.1617898.

[14] S. Rezaei *et al.*, "Role of machine learning in molecular pathology for breast cancer: A review on gene expression profiling and RNA sequencing application," *Critical Reviews in Oncology/Hematology*, vol. 213, Sept. 2025, Art. no. 104780, https://doi.org/10.1016/j.critrevonc.2025.104780.

[15] *Expression data from Breast cancer subtypes*. (2016), NCBI, T. Gruosso, Y. Kieffer, T. Dubois, F. Mechta-Grigoriou. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45827.

[16] METABRIC Group *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, June 2012, https://doi.org/10.1038/nature10983.

[17] B. C. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, "CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research." Apr. 2019, https://doi.org/10.1089/cmb.2018.0238.

[18] "Home - GEO - NCBI." [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/.