



## OPEN ACCESS

### EDITED BY

Ganesh Panzade,  
National Cancer Institute at Frederick  
(NIH), United States

### REVIEWED BY

Soudabeh Sabet,  
Shiraz University of Medical  
Sciences, Iran  
Polina Turova,  
BostonGene, United States

### \*CORRESPONDENCE

Chittibabu Guda,  
✉ [babu.guda@unmc.edu](mailto:babu.guda@unmc.edu)

RECEIVED 04 December 2025

REVISED 23 January 2026

ACCEPTED 16 February 2026

PUBLISHED 10 March 2026

### CITATION

Patel JC, Veerappa A and Guda C (2026)  
An explainable-AI framework reveals  
novel lncRNAs specific for breast cancer  
subtypes.  
*Front. Bioinform.* 6:1760987.  
doi: 10.3389/fbinf.2026.1760987

### COPYRIGHT

© 2026 Patel, Veerappa and Guda. This  
is an open-access article distributed  
under the terms of the [Creative  
Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).  
The use, distribution or reproduction in  
other forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which  
does not comply with these terms.

# An explainable-AI framework reveals novel lncRNAs specific for breast cancer subtypes

Jai Chand Patel<sup>1</sup>, Avinash Veerappa<sup>1</sup> and Chittibabu Guda<sup>1,2\*</sup>

<sup>1</sup>Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, United States, <sup>2</sup>Center for Biomedical Informatics, Research and Innovation, University of Nebraska Medical Center, Omaha, NE, United States

**Background:** Long non-coding RNAs (lncRNAs) have emerged as important regulators in cancer biology; yet their potential for cancer subtyping remains underexplored particularly in the context of large-scale, multi-class supervised classification frameworks, due to limited publicly available data or their use only as auxiliary features in classification tasks.

**Methods:** In this study, we utilized an expansive set of 7,177 lncRNAs obtained from 1,021 breast cancer (BRCA) transcriptomics datasets for subtyping using an explainable artificial intelligence (AI) framework. lncRNA, mRNA, and miRNA features were used to build machine learning (ML) models individually and in combination. Four ML classifiers: Naïve Bayes, Random Forest, Artificial Neural Network, and XGBoost were employed to evaluate subtype classification performance.

**Results:** Using lncRNAs alone, XGBoost demonstrated strong performance with an accuracy of 89.2% and AUROC of 0.99. Addition of miRNA or mRNA features to lncRNA marginally improved the accuracy to 90.8% and 92.2%, respectively, while using all the three features together provided no further gain. A sequential key feature identification pipeline (ANOVA, Boruta, SHAP) has identified interpretable subtype-specific biomarker panels, yielding 119, 66, 54, and 24 unique features for Luminal A, Luminal B, HER2+, and Basal subtypes, respectively. Further lncRNA characterization followed by survival analysis revealed significant subtype-specific novel lncRNAs, including CUFF.25255 (LumA), CUFF.20237 and CUFF.3888 (LumB), CUFF.22414 (HER2+), and CUFF.26607 and CUFF.1961 (Basal).

**Conclusion:** Our findings highlight the diagnostic and biomarker discovery potential of lncRNAs, and the explainable-AI framework implemented here provides a systematic large-scale evaluation of lncRNA-only and integrative models for multi-class BRCA subtyping for BRCA subtyping and can be adopted to other cancers using the existing cancer transcriptomics data in the public databases.

### KEYWORDS

breast cancer subtyping, explainable-AI, long non-coding RNA, multi-omics integration, novel lncRNA

## 1 Introduction

Long non-coding RNAs (lncRNAs) have emerged as key regulators of gene expression at epigenetic, transcriptional, post-transcriptional and translational levels (Grammatikakis et al., 2014). Mounting evidence shows that they are involved in a wide range of cellular processes including cell differentiation and development. Similarly, dysfunction or aberrant expression of lncRNAs has been associated with hundreds of human ailments

including several neurological diseases and cancers (Lin et al., 2024; Tuna et al., 2025). Due to their regulatory roles in various cancer-related processes like cell proliferation, apoptosis, and metastasis, lncRNAs are emerging as promising therapeutic targets for cancer treatment (Coan et al., 2024). Given the tissue-specific nature of transcriptional regulation, lncRNAs could serve as effective biomarkers for a specific cancer or different subtypes of a cancer (Mahato et al., 2024). Motivated by this, our study was designed to test whether lncRNA expression alone can stratify breast cancer subtypes and how the addition of mRNA and miRNA expression data affect the performance of subtype prediction using computational approaches.

Machine learning (ML) techniques have become extremely powerful for refining breast cancer subtype prediction beyond the conventional PAM50 gene panel framework (Ben Rabah et al., 2025; Cancer Genome Atlas, 2012; Cascianelli et al., 2020; Sarkar and Mali, 2022; Wu and Hicks, 2021). Several recent studies have explored deep learning (DL) and graph-based strategies to combine multiple modalities of molecular data. DeepMO, integrating mRNA, DNA methylation, and copy number variations (CNVs), achieved 78.2% accuracy for multi-class subtype classification (Lin et al., 2020). MOGONET, using mRNA, DNA methylation, and miRNA data features, has demonstrated 82.9% accuracy (Wang et al., 2021), while MoGCN, combining mRNA, CNV, and reverse phase protein array (RPPA) data obtained 89.8% accuracy (Li et al., 2022b). On the other hand, moBRCA-net by combining mRNA, DNA methylation, and miRNA data achieved 89.1% accuracy (Choi and Chae, 2023) and Moanna with mRNA, somatic mutations, and CNV data from METABRIC reported 85% accuracy (Lupat et al., 2023). Other hybrid models integrated multiple datatypes such as mRNA, CNV, and histopathology images resulting in 88% accuracy (Liu et al., 2022) and mRNA, CNA, and miRNA resulting in 87.5% accuracy (Cristovao et al., 2022). Recently, we developed GAIN-BRCA framework using mRNA, DNA methylation, and miRNA datatypes, which delivered the highest reported accuracy at 92% (Patel et al., 2025). Notably, none of these high-performing models incorporated lncRNAs, leaving their potential for supervised, multi-class BRCA subtype classification unexplored.

While lncRNAs are increasingly recognized for their regulatory roles in cancer, their potential use in predictive multi-omics models has been notably limited, with very few studies benchmarking their value as standalone transcriptomic features alongside mRNAs, miRNAs and fusion transcripts. Early work using hierarchical clustering revealed subtype-associated lncRNA signatures (Su et al., 2014), while later models incorporate Lasso-Cox, and nomogram frameworks to predict prognosis or identify biomarkers (Li et al., 2022a; Li et al., 2021). Other works proposed pathway- or phenotype-linked lncRNA panels, such as disulfidptosis-associated lncRNAs, where ML models like random forest (RF) and K-nearest neighbor (KNN) have reached AUCs around 0.84–0.87 (Xia et al., 2023); however, only 132 lncRNA features were used in this study. Other approaches, such as lncRIndiv, estimated patient-specific lncRNA expression to capture intra- and inter-subtype variability (Zhao et al., 2021). Additional efforts focused on specific clinical contexts such as diagnostic panels for TNBC (AUC ~0.80) (Shaath et al., 2021), immune-related lncRNA prognostic models (Liu et al., 2025), and radiomic frameworks linking MRI

features with lncRNA signatures (Yu et al., 2023). Moreover, large-scale surveys of subtype- and cell-type-specific lncRNA expression consistently highlight their discriminatory potential across the four subtypes of breast cancer (BRCA), Luminal A, Luminal B, HER2+, and Basal (Bjorklund et al., 2022). Despite this progress, systematic large scale benchmarking of lncRNA-only models for supervised, multi-class subtype classification remains limited, particularly in studies that directly compare lncRNAs against mRNA and miRNA features under a unified machine learning framework. The most prior efforts were predominantly focused on prognostic or binary subtype classification, relied on limited annotations, or incorporated lncRNAs merely as auxiliary features within multi-omics models. Consequently, the independent and integrative discriminatory potential of lncRNAs remains underexplored.

To address these gaps, our study introduces a comprehensive framework that systematically evaluates the predictive and biological value of lncRNAs in breast cancer subtype classification. Rather than claiming novelty based solely on the use of lncRNAs, our contribution lies in the scale and scope of evaluation. We utilized a high-quality in-house curated lncRNA dataset that captures a broader range of biologically relevant lncRNAs across breast cancer subtypes (Guda C, 2025). Using this dataset, we first trained models based solely on lncRNA expression to assess their standalone discriminatory power, followed by integrative modeling with mRNA and miRNA data to examine the additive and synergistic effects. To ensure robust and interpretable results, we implemented a multi-stage key feature identification strategy combining statistical filtering with ANOVA followed by a wrapper-based method, Boruta, and a model-agnostic interpretability tool, SHAP (Zhu et al., 2025; Ahmed et al., 2018; Zhang et al., 2018; Alkuhlani et al., 2017). Overall, our lncRNA-centric study provides a systematic assessment of the independent and combined discriminatory power of lncRNA in classification tasks and reveals several novel lncRNAs for the four BRCA subtypes.

## 2 Materials and methods

Patient-matched datasets across three distinct molecular data modalities including lncRNAs, mRNAs, and miRNAs were downloaded for 1,021 tumor and 104 normal samples obtained from the Cancer Genome Atlas (TCGA) BRCA collection. Patient IDs were matched across all data modalities, allowing direct performance comparison across different feature compositions. The pipeline was designed to individually preprocess and normalize each datatype before their use in developing independent and integrative models for ML-based classification. The overall workflow, encompassing data integration, model development, feature prioritization, and functional characterization is summarized in Figure 1.

### 2.1 lncRNA data: extraction and quantification

In this study, we utilized our in-house computational pipeline to extract an expansive high-quality lncRNA dataset (GUDA, 2025), addressing the limited coverage of lncRNAs in standard annotations. BAM files from the TCGA-BRCA dataset

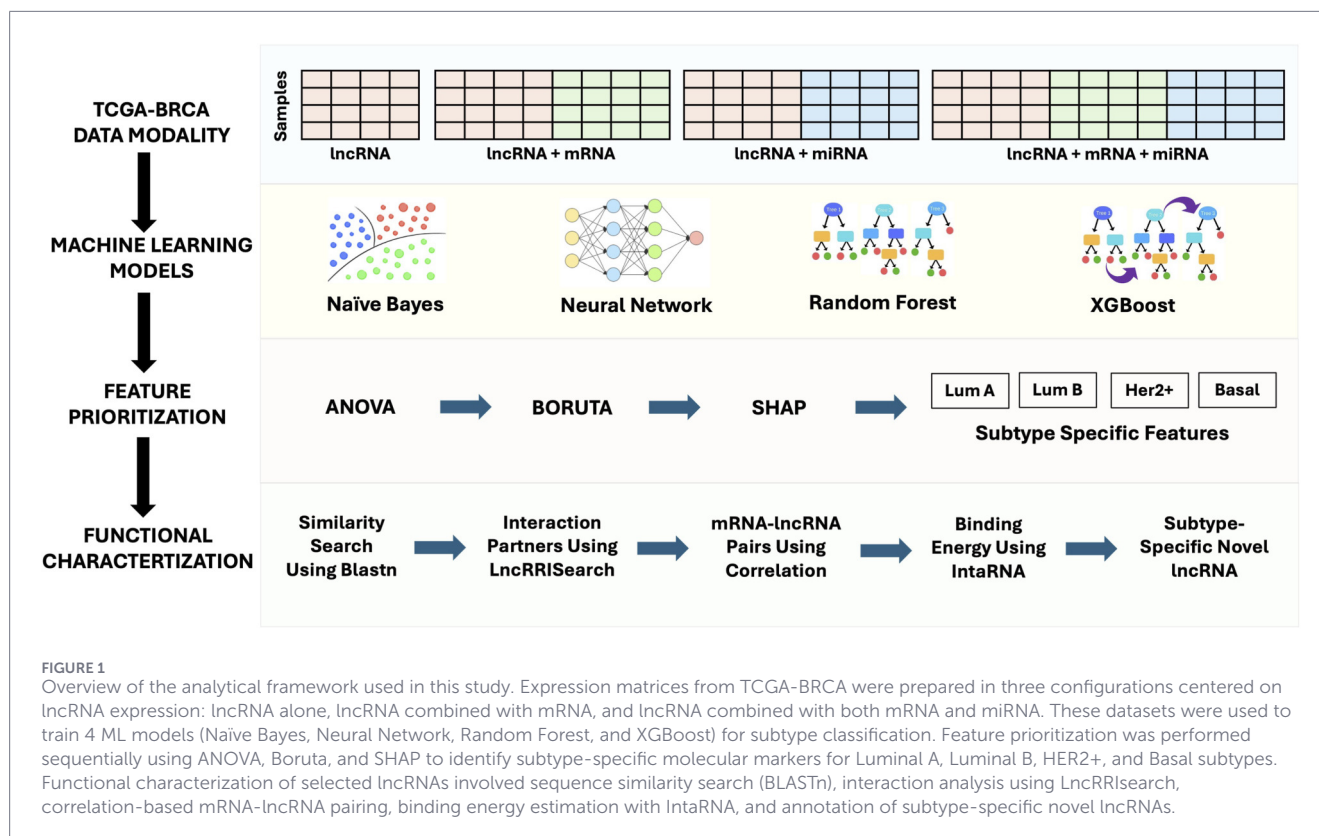


FIGURE 1

Overview of the analytical framework used in this study. Expression matrices from TCGA-BRCA were prepared in three configurations centered on lncRNA expression: lncRNA alone, lncRNA combined with mRNA, and lncRNA combined with both mRNA and miRNA. These datasets were used to train 4 ML models (Naïve Bayes, Neural Network, Random Forest, and XGBoost) for subtype classification. Feature prioritization was performed sequentially using ANOVA, Boruta, and SHAP to identify subtype-specific molecular markers for Luminal A, Luminal B, HER2+, and Basal subtypes. Functional characterization of selected lncRNAs involved sequence similarity search (BLASTn), interaction analysis using LncRRISearch, correlation-based mRNA-lncRNA pairing, binding energy estimation with IntaRNA, and annotation of subtype-specific novel lncRNAs.

were retrieved and converted into paired-end FASTQ files using *samtools* (Li et al., 2009). *STAR* aligner was employed to map reads to the *GRCh38.p14* genome reference using the *encode.v47.long\_noncoding\_RNAs.gtf* annotation file (Dobin et al., 2013). The aligned BAM files were processed through *Cufflinks* for transcript assembly (Trapnell et al., 2010). All resulting transcripts were merged into a unified transcriptome file (*merged.gtf*), followed by duplicate removal to obtain *merged\_without\_duplicates.gtf*. Specifically, transcript entries with redundant coordinates or identical transcript IDs were filtered out. Mitochondrial chromosome annotations were also excluded to improve relevance and reduce noise. Transcript quantification was performed using *featureCounts* from the *subread* package (Liao et al., 2013). Expression profiles were generated for tumor and normal samples separately, yielding count matrices. These raw counts were normalized using *DESeq2*'s variance-stabilizing transformation (VST), log<sub>2</sub>-transformed, and filtered to retain features with less than 80% missing or zero values (Love et al., 2014). To ensure non-coding purity of the retained transcripts, their coding potential was assessed using the Coding Potential Assessment Tool (CPAT) (Wang et al., 2013). Transcript sequences were extracted from the genome with *gffread*, and CPAT was applied using a species-specific cutoff for humans (0.364). Transcripts exceeding this threshold were removed, and a final high-confidence set of lncRNAs was stored in *filtered\_lncRNAs.fa*.

## 2.2 mRNA data processing

mRNA expression data for both tumor and solid normal tissues were downloaded using the *TCGAbiolinks* R package

(Colaprico et al., 2016). Raw count matrices were extracted from the GDC portal and merged into a unified expression matrix. A sample metadata table was constructed to assist in tracking and downstream labeling. The data were normalized using *DESeq2*'s variance-stabilizing transformation (VST), log<sub>2</sub>-transformed, and filtered to remove features with more than 80% missing or zero values (Love et al., 2014). The resulting cleaned mRNA matrix was used for multi-omics modeling.

## 2.3 miRNA data processing

miRNA expression data were processed in the same manner as the mRNA pipeline described above. Count matrices were obtained using *TCGAbiolinks*, merged, and normalized with *DESeq2* VST, followed by log<sub>2</sub> transformation and filtering of low-abundance features (>80% zeros or missing) (Love et al., 2014; Colaprico et al., 2016).

## 2.4 Data preparation for ML classification

Preprocessed datasets from the three omics datatypes were used to construct input matrices for ML classification. Each matrix was centered on lncRNA expression: (i) lncRNA alone, (ii) lncRNA combined with mRNA and (iii) lncRNA combined with both mRNA and miRNA. Expression matrices from each omics layer were aligned by patient identifiers and concatenated horizontally to create unified input matrices for downstream modeling, ensuring consistent sample representation across all configurations.

## 2.5 Methodology development

### 2.5.1 Selection of supervised learning models

We used the four subtypes of TCGA-BRCA patient samples, Luminal A, Luminal B, HER2+, and Basal for building supervised models. Four ML models were selected based on their diversity in learning paradigms, interpretability, and effectiveness in managing high-dimensional transcriptomic contexts. These methods include Naïve bayes (NB), random forest (RF), artificial neural networks (ANNs) and extreme gradient boosting (XGBoost). NB classifiers were implemented using the *GaussianNB* (Webb, 2011) module from scikit-learn (Fabian Pedregosa et al., 2011). Despite the assumption of conditional independence among features, NB often performs well on high-dimensional and sparse datasets, such as gene expression matrices. RF models were constructed using *RandomForestClassifier* with 100 estimators ( $n\_estimators = 100$ ) and a fixed seed ( $random\_state = 42$ ) (Breiman, 2001). As a tree-based ensemble method, RF can model complex non-linear interactions and offers robustness to overfitting. Another ensemble method, XGBoost classifier, was implemented with the following configuration:  $objective = 'multi:softprob'$ ,  $num\_class = 4$ ,  $eval\_metric = 'mlogloss'$ , and  $random\_state = 42$  (Guestrin, 2016). XGBoost offers efficient and scalable gradient-boosted tree learning and is particularly suited for large-scale structured datasets with redundant or correlated features. ANNs were built using the *TensorFlow/Keras* library (Jain, 1996). The architecture comprised an input layer equal to the number of features, followed by two dense hidden layers with 512 and 256 neurons, respectively, each using *ReLU* activation (Agarap, 2019). Dropout layers with a rate of 0.5 were applied to prevent overfitting. The final output layer consisted of four neurons with *softmax* activation for multi-class prediction (Jan Goodfello, 2016). The network was trained using the Adam optimizer and categorical cross-entropy loss for 50 epochs with a batch size of 32 (Diederik and Kingma, 2014; Anqi Mao and Yutao, 2023). All models were implemented in Python using scikit-learn (Fabian Pedregosa et al., 2011), TensorFlow/Keras, (Martín Abadi et al., 2016; Chollet, 2015), and XGBoost (Guestrin, 2016).

### 2.5.2 Evaluation metrics

All models were evaluated using stratified 10-fold cross-validation to maintain subtype size balance across folds. Performance metrics, including accuracy, precision, recall, F1-score, and AUROC were computed on each fold and averaged (Powers, 2020; Bradley, 1997; Shakyawar et al., 2022; Mishra et al., 2019; Sethi et al., 2024). ROC curves were generated for each subtype class, and probability estimates from each model were preserved for downstream comparison and interpretability analysis. All performance evaluations were conducted using the full pre-processed feature matrices. Feature selection using the ANOVA-Boruta-SHAP (ABS) pipeline (described below) was not incorporated into model training or cross-validation and was applied only after model benchmarking for feature interpretation and biomarker discovery. This modeling strategy allowed for a robust evaluation of predictive performance of four different ML models across four breast cancer subtypes.

## 2.6 Identification of important subtype-specific key features

Feature identification was performed on the combined lncRNA and mRNA dataset to identify key molecular markers relevant to BRCA subtype classification. Given the high dimensionality of this concatenated dataset (46,715 features), we implemented a sequential feature identification pipeline designed to progressively refine the feature space while reducing noise and redundancy. The ANOVA-Boruta-SHAP, referred to as the ABS pipeline, combined three complementary methods applied in sequence. ANOVA (Analysis of Variance) was used as a univariate statistical filter using the  $f\_classif$  function from the *scikit-learn* library (Lars Sthle, 1989) and features with a  $p < 0.05$  were retained. Next, Boruta, a RF-based wrapper method was applied to the ANOVA-filtered features in a one-vs-rest manner for each breast cancer subtype using the *BorutaPy* module in Python (Miron et al., 2010). Finally, SHAP (SHapley Additive exPlanations) was employed to quantify the contribution of each of the Boruta-selected features (Scott Lundberg, 2017). A multiclass XGBoost classifier implemented via the *SHAP* Python library was used to compute SHAP values. Notably, this model was trained solely for SHAP-based interpretation and is distinct from the final predictive classifier used in subtype predictions. This ABS pipeline enabled progressive feature refinement from broad statistical filtering to the identification of key model-based features.

## 2.7 Identification of subtype-enriched pathways

To interpret the functional significance of subtype-specific gene sets, we performed Ingenuity Pathway Analysis (IPA, QIAGEN Inc.). Pathway enrichment was performed using subtype informative mRNA features identified by the ABS pipeline, rather than transcriptomic wide or subtype wise differential expression gene sets. Subtype-specific gene sets identified above were uploaded separately into IPA using default core analysis settings, the Ingenuity Knowledge Base as the reference set; the analysis type set to direct and indirect relationships; and using experimentally observed and high confidence predicted interactions. Since IPA outputs uncorrected p-values for pathway enrichment and does not provide FDR-adjusted results by default, we applied a stringent significance threshold of  $p < 0.01$ , corresponding to  $-\log_{10}(p\text{-value}) > 2$  in IPA outputs, to obtain reliable enriched pathways while minimizing potential false positives.

## 2.8 Survival analysis of subtype-specific mRNAs and lncRNAs

Subtype-specific survival analysis was conducted to evaluate the prognostic relevance of significant protein-coding genes and long non-coding RNAs (lncRNAs) identified from the pipeline described above. For each feature, Kaplan-Meier survival curves were generated (Kaplan and Meier, 1958), and statistical significance was assessed using the log-rank test (Kleinbaum and Klein, 2012; Shakyawar et al., 2024). Rather than employing a fixed median cutoff, we used a more refined data-driven strategy by scanning percentile-based cutoffs between the 20th and 80th percentiles of expression. The optimal threshold was selected based

on the minimum log-rank  $p$ -value, ensuring that both expression-defined groups (high and low) contained at least five patients to maintain statistical power. This approach offered a more precise assessment of the prognostic utility of each candidate biomarker while avoiding arbitrary or biased thresholding.

## 2.9 Novel lncRNA characterization

Subtype-specific lncRNA candidates were first identified using our ANOVA-Boruta-SHAP (ABS) pipeline (Figure 1). We then characterized the putative novel lncRNAs identified for each breast cancer subtype using several approaches described below. Nucleotide sequences of lncRNAs were retrieved in FASTA format from GENCODE v48 and sequence similarity searches were carried out using BLASTN against the human genome (GRCh38.p14.genome.fa) with stringent cutoffs (bit score > 40, alignment length  $\geq 200$  nt, identity = 100%,  $e$ -value <  $10^{-5}$ ) (Altschul et al., 1990). High-identity paralogous RNA sequences identified from this analysis were retained for interaction studies. Next, lncRNA-mRNA interaction analysis was performed using LncRRISearch (Fukunaga et al., 2019). Here, paralogous RNAs with 100% sequence identity were used as anchors to infer potential mRNA interaction partners (MIPs). The underlying assumption was that if two RNA sequences are identical, they are likely to base-pair with the same mRNA targets when expressed. Based on this, both ABS-derived novel lncRNAs and their paralogous RNAs were used to predict MIPs. Interactions with binding energies  $\leq -16$  kcal/mol were retained, and the top 100 most stable interactions were selected for each lncRNA. Finally, to assess clinical relevance, we evaluated whether the predicted MIPs of these novel lncRNAs included oncogenes or tumor suppressor genes (TSGs). This was accomplished by cross-referencing with the curated gene catalogue provided by OncoKB database (Chakravarty et al., 2017).

## 2.10 lncRNA-mRNA correlation and base pairing analysis

We computed expression-based correlations between novel lncRNAs and their predicted MIPs using both Pearson and Spearman correlation coefficients (Figure 1). Normalized raw count values for the lncRNAs as well as mRNAs were used to construct expression matrices with patients as rows and either lncRNA (CUFF IDs) or mRNA (gene symbols) as columns. Correlations were then calculated, and pairs with  $|r| \geq 0.40$  and  $FDR < 0.05$  were considered significant. For downstream analysis, we used the mean correlation value of Pearson and Spearman. To validate whether these correlated pairs could physically interact, we extracted the nucleotide sequences of the significant lncRNA-mRNA pairs (CUFF IDs and their targets) from the GENCODE v48 fasta file. Structural feasibility was evaluated using IntaRNA (Mann et al., 2017) which predicts RNA-RNA base-pairing by calculating hybridization events and corresponding binding energy scores ( $\Delta G$ ). Only interactions confirmed by both tools were retained. Stringent parameters such as energy threshold  $\leq -16$  kcal/mol, a maximum of five interactions per RNA pair, overlap restricted to the lncRNA sequence, no unstacked (lonely) base pairs, at least seven intermolecular base pairs in the seed region, and no GU base pairs or GU ends were applied. Additional folding parameters included: temperature

37 °C (RNAplfold -W), folding window size 150, maximum base-pair distance 100, and Turner-2004 energy model from the ViennaRNA package for base-pair probability estimation. The distribution of binding energies was then plotted to visualize the stability landscape across subtypes (Figure 1). LncRRISearch, uses IntaRNA output to generate RNA-RNA hybridization images (Fukunaga et al., 2019; Mann et al., 2017). From these distributions, we selected the most stable lncRNA-mRNA interactions for each subtype. Finally, for these top-ranked pairs, LncRRISearch was used to extract nucleotide-level hybridization images, illustrating specific base-pairing sites and interaction regions.

## 2.11 External cohort transferability analysis using CPTAC cohort

To provide an exploratory assessment of transferability beyond TCGA, we analyzed an independent breast cancer RNA-seq cohort from CPTAC breast cancer cohort of  $n = 356$  (Geffen et al., 2023). Transcripts per million (TPM) values were log-transformed as  $\log_2(\text{TPM}+1)$  for downstream analyses. Because intrinsic subtype labels were not available in CPTAC metadata, PAM50 subtypes were inferred directly from mRNA expression using a centroid-correlation strategy. Briefly, the gene-fu-based PAM50 centroid matrix was used as the reference. For each sample, Pearson correlation was computed between its  $\log_2(\text{TPM}+1)$  expression profile and each PAM50 centroid; the subtype corresponding to the maximum correlation was assigned as the inferred PAM50 label. Samples classified as Normal-like were excluded from subsequent analyses, yielding a tumor-only cohort comprising of Luminal A ( $n = 101$ ), Luminal B ( $n = 129$ ), HER2-enriched ( $n = 50$ ), and Basal-like tumors ( $n = 52$ ). Transferability was evaluated by testing subtype-associated expression patterns for a limited subset of gene-level counterparts linked to our prioritized CUFF-level candidates (Table 3): LINC01133 (CUFF.26607/L1RSM), NIFK-AS1 (CUFF.20237/NARUS), and TEC (CUFF.25255/TERCI). For each gene, expression differences across inferred PAM50 subtypes were assessed using the Kruskal-Wallis test on  $\log_2(\text{TPM}+1)$  values, and  $p$ -values were adjusted for multiple testing using the BH-FDR across the tested genes. In addition, pairwise subtype comparisons were performed using Wilcoxon rank-sum tests with BH-FDR correction applied within each gene to account for multiple pairwise comparisons.

## 3 Results

Expression matrices were constructed for ML classification using transcriptomic features from 1,021 BRCA patients. After preprocessing and filtering, the final dataset included differentially expressed transcripts from 1,021 BRCA patients, comprising 7,177 lncRNAs, 39,538 mRNAs, and 769 miRNAs. The resulting matrices contained 7,177 features for lncRNA alone, 46,715 features for lncRNA combined with mRNA, and 47,484 features for lncRNA combined with mRNA and miRNA. All datasets maintained identical sample counts and ordering, ensuring that performance differences among models reflected the underlying feature composition rather than sample variability.

**TABLE 1** Performance of four classifiers across lncRNA, lncRNA combined with miRNA, lncRNA combined with mRNA, and lncRNA combined with mRNA and miRNA feature sets on TCGA-BRCA.

Features used	ML models	Accuracy (%)	Precision	Recall	F1-Score	AUROC
lncRNA (7,177)	Naïve bayes	73.9	0.736	0.796	0.739	0.91
	ANN	87.7	0.875	0.869	0.867	0.98
	Random forest	86.2	0.895	0.752	0.783	0.98
	XGBoost	89.2	0.908	0.850	0.870	0.99
lncRNA + miRNA (7,177 + 769)	Naïve bayes	75.9	0.750	0.80	0.757	0.92
	ANN	88.6	0.867	0.856	0.856	0.98
	Random forest	87.5	0.904	0.774	0.807	0.98
	XGBoost	90.8	0.903	0.858	0.873	0.99
lncRNA + mRNA (7,177 + 39,538)	Naïve bayes	81.0	0.793	0.832	0.801	0.89
	ANN	89.4	0.874	0.858	0.860	0.95
	Random forest	89.3	0.917	0.821	0.849	0.99
	XGBoost	92.2	0.917	0.893	0.902	0.99
lncRNA + mRNA + miRNA (7,177 + 39,538 + 769)	Naïve bayes	81.1	0.800	0.835	0.804	0.89
	ANN	89.9	0.877	0.873	0.872	0.95
	Random forest	89.3	0.915	0.830	0.859	0.99
	XGBoost	92.1	0.920	0.894	0.903	0.99

Values are mean 10-fold cross-validation metrics (Accuracy, Precision, Recall, F1) and multi-class AUROC. lncRNA-only models already perform strongly (XGBoost: Accuracy = 89.2% and AUROC, 0.99). The lncRNA, combined with mRNA, configuration yielded the best overall results (XGBoost: Accuracy = 92.2%, AUROC, 0.99), followed by lncRNA, combined with miRNA (Accuracy = 90.8%, AUROC, 0.99), while including all three data types (lncRNA, mRNA, and miRNA) did not provide further improvement.

### 3.1 lncRNAs exhibit the same level of discriminatory power as mRNAs for subtyping

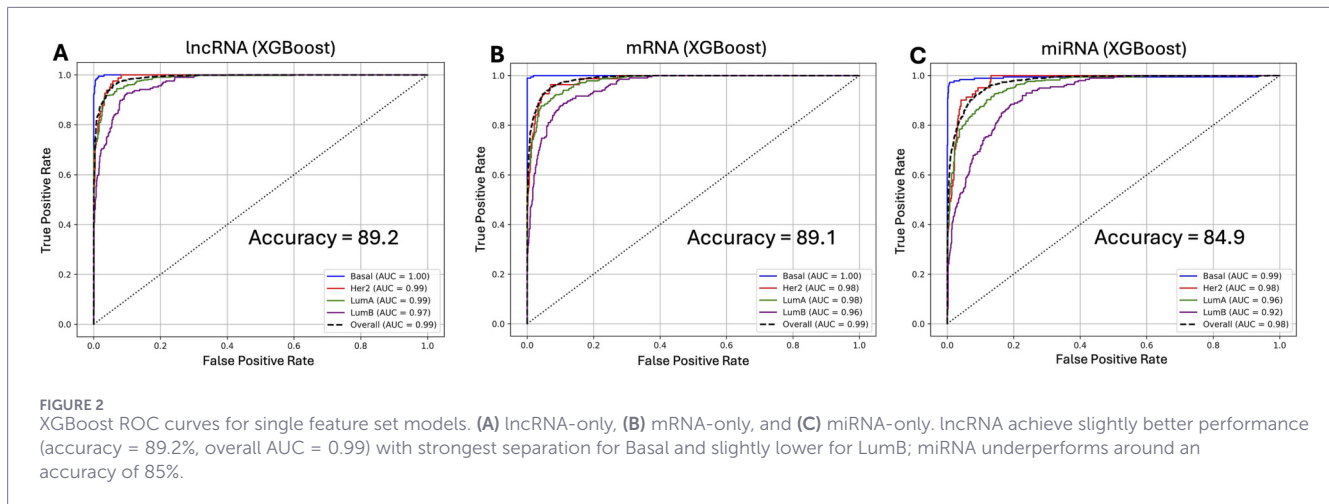
At first, four different ML models (NB, RF, XGBoost, and ANN) were implemented using lncRNA expression data containing 7,177 features to identify the best performing model for breast cancer subtype classification. Each of these models demonstrated varying degrees of classification performance. NB, a probabilistic model, showed a moderate discriminative ability, achieving an accuracy of 73.9% and an overall AUC of 0.91, yet performing exceptionally well on the Basal subtype (AUC = 0.98) (Table 1; Supplementary Figure S1A). In contrast, the ANN model markedly improved generalization, achieving an accuracy of 87.7% and AUC of 0.98 overall and maintaining consistent performance across the subtypes (AUCs of Basal: 0.99; HER2+: 0.97; Luminal A: 0.97; and Luminal B: 0.95) reflecting its strength in modeling non-linear relationships (Table 1; Supplementary Figure S1B). Random Forest, while also achieving an accuracy of 86.2%, with an overall AUC of 0.98 and a perfect Basal classification (AUC = 1.00) (Table 1; Supplementary Figure S1C). Among all, XGBoost emerged as the best performing classifier, delivering a near-perfect overall AUC of 0.99 and a strong cross-validation accuracy of 89.2% with F1 score of 0.87 (Figure 2A; Table 1). Notably, the Basal subtype was consistently the easiest to distinguish across all models, while Luminal A and B posed greater challenges possibly due to their overlapping transcriptional landscapes. These results highlight the predictive potential of lncRNA expression

alone, especially using interpretable and high-capacity classifiers like XGBoost.

Given the superior performance of XGBoost, we selected this model to compare the performance of lncRNA model with those using only mRNA or miRNA features. Using only mRNA features (39,538), the XGBoost model delivered a near identical performance to that built from only-lncRNA features (7,177) with 89.1% accuracy and subtype AUC's at Basal = 1.00, HER2+/LumA = 0.98, LumB = 0.96) (Figure 2B). In contrast, the miRNA-only model showed weaker performance (accuracy = 84.9%, overall AUC = 0.98) compared to the lncRNA-only model (Figure 2C), which may be attributable to its very small feature size (769). These results demonstrate the remarkable discriminatory potential of lncRNAs as a sole source for breast cancer subtyping.

### 3.2 lncRNA models showed enhanced accuracy in combination with mRNAs or miRNA features

We also tested out pair-wise combinations of the three feature sets with four different ML models (NB, ANN, RF, XGBoost) to identify the best combination of features for breast cancer subtyping. When integrating lncRNA with miRNA and mRNA features separately, a substantial performance boost was observed across all ML models. NB, which previously performed moderately with the single-omics feature sets, has shown markedly improved metrics achieving an average 10-fold cross-validation accuracies of 75.9% and 81% when combined with miRNA and mRNA



feature sets, respectively, compared to 73.9% when only-lncRNA data was used (Table 1; Supplementary Figures S2A, S2D). ANN also demonstrated robust predictive capacity with significant improvements in all predictive metrics such as accuracies of 88.6% (lncRNA with miRNA) and 89.4% (lncRNA with mRNA), suggesting enhanced generalization and subtype sensitivity (Table 1; Supplementary Figures S2B, S2E). Similarly, RF benefited from the inclusion of additional features, achieving overall AUROC values of 0.98 (lncRNA with miRNA) and 0.99 (lncRNA with mRNA), outperforming the ANN model (AUROC = 0.95 with mRNA). It reached a perfect AUC for Basal subtype prediction and maintained strong AUC scores for HER2+ (0.99) in both the combinations. The mean accuracies of 87.5% and 89.3% with miRNA and mRNA integration, respectively, compared to 86.2% using lncRNA features alone, confirms its effectiveness for multi-omics learning (Table 1; Supplementary Figures S2C, S2F). Consistent with single omics results, XGBoost continued to outperform all other ML methods achieving an overall AUC of 0.99 and a perfect discrimination for the Basal subtype (AUC = 1.00). It also achieved the highest accuracy of 92.2% for the lncRNA in combination with mRNA features with an F1-score of 0.90. In comparison, the XGBoost model using lncRNA combined with miRNA features delivered only 90.8% accuracy with an F1-score of 0.87 (Figures 3A,B; Table 1). The most notable gain was observed in the Luminal B subtype classification, which had previously lagged in models based on only lncRNA feature set, with an AUC reached to 0.98 with the mRNA feature set. This indicates the complementary nature of mRNA features in improving resolution across challenging subtypes.

### 3.3 Predictive models using the combined features of lncRNA, mRNA, and miRNA

Given the consistently superior performance, XGBoost models were built using features from all three data modalities in a concatenated fashion resulting in a matrix of 1,021 patients and 47,484 features. Surprisingly, the addition of miRNA features to the lncRNA and mRNA framework did not yield any performance gains, suggesting a limited additive value from this third omics layer. All performance metrics as shown in Table 1 across the four models

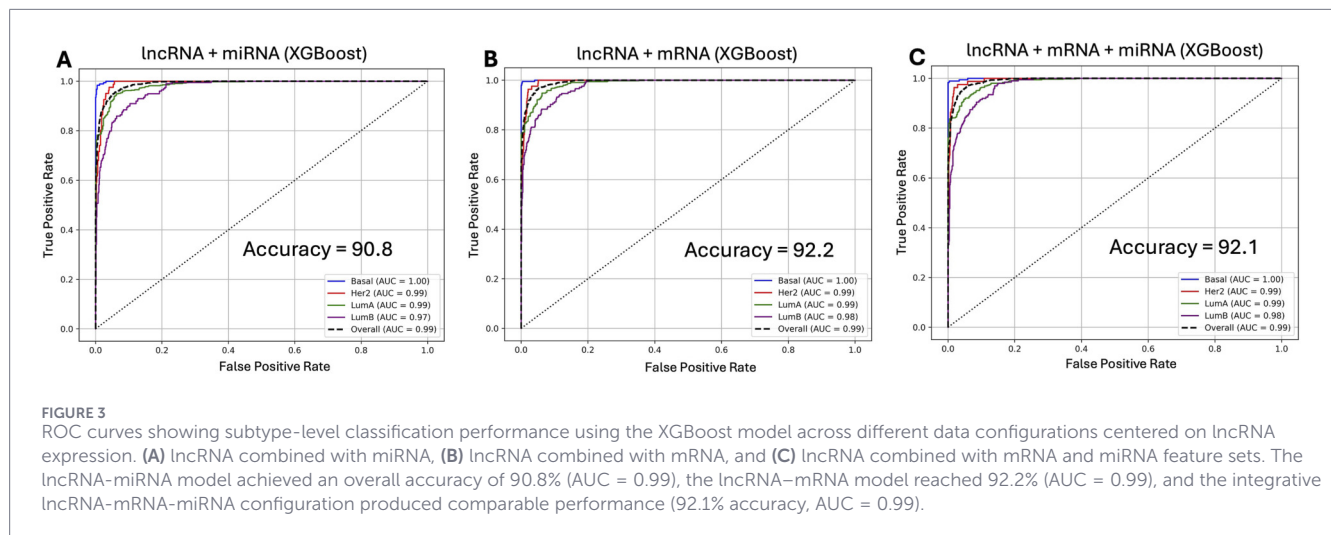
remained nearly identical between the lncRNA plus mRNA and lncRNA, mRNA plus miRNA featured models. For example, with the XGBoost model, the AUROC of 0.99, accuracy of 92.1%, and F1-score of 0.90, virtually mirrored corresponding performance with the lncRNA combined with mRNA model (Table 1; Figure 3C). Similar patterns were observed with the NB, ANN, and RF models (Table 1; Supplementary Figure S3). Collectively, these results underscore the sufficiency of lncRNA and mRNA data as the optimal feature sets for accurate breast cancer subtype classification.

### 3.4 Identification of important subtype-specific features

The best performing XGBoost model utilized a large number of combined features (46,715) that include 7,177 lncRNAs and 39,538 mRNAs from 1,021 patients. However, only a subset of these features is weighted heavily to discriminate the four subtypes. To identify such features, we used a three-step data filtering approach, referred to as ABS pipeline (Section 2.6) using ANOVA, Boruta, and SHAP tools in sequence. Initially, ANOVA identified 41,176 statistically significant features ( $p$ -value < 0.05), serving as the primary dimensionality reduction filter. These were further filtered first using Boruta and classified using SHAP. Boruta retained 1,336 features for Luminal A, 1,054 for Luminal B, 927 for HER2+, and 828 for Basal. Finally, SHAP analysis distilled them down to 139 for Luminal A, 85 for Luminal B, 60 for HER2+, and 26 for Basal. To improve subtype specificity, we removed features that appeared in more than one subtype class. This filtering step helped isolate features with high discriminatory value, reducing potential ambiguity caused by features commonly associated with multiple subtypes. The final refined sets of unique features are given in Table 2.

### 3.5 Functional pathway profiling of BRCA subtype-specific mRNA features

Because pathway enrichment was performed on ABS-selected subtype-informative mRNA features rather than on global differentially expressed gene sets, the resulting pathways



**TABLE 2** Summary of unique protein-coding genes and lncRNAs identified for each breast cancer subtype (LumA, LumB, HER2+, Basal) using the ABS pipeline.

Subtypes	Genes	lncRNAs	Total unique features
LumA	108	11	119
LumB	53	13	66
HER2+	49	5	54
Basal	19	5	24

The totals represent non-overlapping features specific to each subtype, highlighting the molecular distinctiveness of LumA (119 features) versus the compact signature observed in Basal (24 features).

should be interpreted as molecular programs contributing to subtype discrimination within a supervised learning framework. Consequently, enriched pathways do not necessarily reflect the dominant or canonical biological processes of each breast cancer subtype. Subtype-specific mRNA features presented in Table 2 were used for functional characterization of BRCA subtypes with Ingenuity Pathway Analysis. This analysis was performed to interpret the biological relevance of the mRNA features identified alongside of lncRNA features, providing functional context for the lncRNA-centered classification models. IPA revealed distinct yet overlapping biological themes across breast cancer subtypes, emphasizing pathways critical for cell proliferation, genomic stability, and metabolic regulation. In Luminal A, the most significantly enriched pathways centered around cell cycle regulation, including Cell Cycle Checkpoints ( $p$ -value =  $5.8e-07$ ), Mitotic Metaphase and Anaphase, and p53 signaling, which collectively highlight the involvement of genes such as CDC20, CHEK1, PCNA, and BCL2 in orchestrating DNA replication, damage checkpoints, and apoptotic resistance. Additional pathways included Regulation of TP53 Activity through Phosphorylation and Kinetochores Metaphase Signaling, suggesting coordinated control of chromosomal segregation and tumor suppressor activation within this subtype. These enrichments do not imply a globally high proliferative for Luminal A tumors rather reflect regulated cell cycle and checkpoint components retained within the compact,

classification-oriented feature panel that contribute to separating Luminal A from other subtypes. Similarly, for Luminal B, pathway enrichment was dominated by cell cycle and mitotic control pathways, with Mitotic Metaphase and Anaphase ( $p$ -value =  $3.1e-06$ ) and Kinetochores Metaphase Signaling Pathway at the top, implicating genes like BUB1, CCNB1, CENPK, and ESPL1. Furthermore, pathways such as Cell Cycle Checkpoints, Proteasomal PSMD10 Signaling, and Regulation of Apoptosis were significant, suggesting that Luminal B tumors exhibit dysregulation in protein degradation and apoptotic signaling alongside their proliferative drive.

In the Basal subtype, while fewer pathways surpassed the stringent significance threshold, notable enrichment was observed for ESR-mediated signaling ( $p$ -value =  $5.1e-05$ ), involving FOXA1, TFF1, and CXXC5, which is intriguing given Basal tumors' typical ER-negativity but consistent with reports of partial estrogen pathway activation in subsets. Additionally, UFMylation signaling and ceramide biosynthesis pathways were enriched, driven by genes such as XBP1 and DEGS2, indicating potential alterations in protein homeostasis and lipid metabolism that may underlie aggressive Basal phenotypes. For HER2+ tumors, enriched pathways included Epithelial Membrane Protein Signaling ( $p$ -value =  $2.2e-03$ ), highlighting canonical HER2 (ERBB2) pathway activation alongside IGF1R and ITGA10, as well as metabolic pathways such as the Pentose Phosphate Pathway (Oxidative Branch) and Catecholamine Biosynthesis, driven by G6PD and PNMT. These collectively suggest enhanced proliferative signaling coupled with metabolic reprogramming characteristic of HER2-enriched cancers. Overall, cell cycle and mitotic pathways emerged as the predominant functional themes in Luminal subtypes. In contrast, Basal and HER2+ subtypes demonstrated enrichment in stress adaptation, membrane signaling, and metabolic pathways, underpinned by subtype-specific gene signatures identified in this study. These results provide a mechanistic context for the subtype classification models, supporting their biological validity and offering potential mRNA linked pathways that may interact with or to be regulated by lncRNA signatures.

### 3.6 Prognostic utility of lncRNAs and mRNAs across BRCA subtypes

Survival analysis revealed multiple subtype-specific features with significant prognostic associations, reinforcing their clinical significance. In the Luminal A subtype, while several genes such as *CHEK1*, *PCNA*, and *CDC20* showed a trend toward prognostic relevance, none reached statistical significance under the log-rank test. Among lncRNAs, *CUFF.14265* ( $p = 0.0014$ ) and *CUFF.29662* ( $p = 0.0451$ ) emerged as significantly associated with patient survival, indicating that non-coding elements may contribute more prominently to prognosis in this subtype. In contrast, the Luminal B subtype yielded a richer set of prognostic markers. Genes including *BUB1* ( $p = 0.0228$ ), *ESPL1* ( $p = 0.0442$ ), and *E2F7* ( $p = 0.0081$ ) showed significant survival differences, consistent with their known roles in mitotic regulation. Similarly, lncRNAs such as *CUFF.10077* ( $p = 0.0002$ ), *CUFF.10442* ( $p = 0.0085$ ), *CUFF.3888* ( $p = 0.0032$ ), and *CUFF.538* ( $p = 0.0008$ ) were significantly associated with outcomes, highlighting a multifaceted regulatory landscape influencing prognosis in Luminal B tumors. In the Basal subtype, *FOXAI* ( $p = 0.0148$ ) emerged as a significant gene-level marker despite the typically reduced number of predictive features in this aggressive subtype. Among lncRNAs, *CUFF.23894* showed a modest but significant association with survival ( $p = 0.0193$ ), suggesting limited but notable non-coding contributions. For the HER2+ group, mRNAs *IGF1R* ( $p = 0.0338$ ) and *G6PD* ( $p = 0.0395$ ) showed significant prognostic value, consistent with HER2-driven signaling and metabolic reprogramming. Prognostic lncRNAs in HER2+ included *CUFF.10593* ( $p = 0.0135$ ), *CUFF.1961* ( $p = 0.0247$ ), and *CUFF.25223* ( $p = 0.0473$ ), adding additional layers of subtype-specific risk stratification (Figures 4, 5).

### 3.7 Identification of mRNA targets for novel subtype-specific lncRNAs

From the ABS pipeline, we identified subtype-specific candidate lncRNAs: 124 in LumA, 193 in LumB, 61 in HER2+, and 460 in Basal. BLASTN similarity searches against the human reference genome revealed that a large fraction of these candidates had no prior matches, supporting their classification as putative novel lncRNAs. Paralogous RNA sequences were defined based on the similarity criteria described in the Methods section. Using these criteria, we identified 45 paralogous RNA sequences in LumA, 56 in LumB, 26 in HER2+, and 17 in Basal, which were advanced for interaction analysis. Predicted lncRNA-mRNA interactions showed that LumA (~2,800 pairs) and LumB (~3,000 pairs) had the largest networks, while HER2+ (~1,900 pairs) and Basal (~1,000 pairs) had fewer. Mapping to gene symbols yielded hundreds of unique targets per subtype. Integration with OncoKB identified multiple regulators, including 51 clinically verified genes in LumA, 69 in LumB, 60 in HER2+, and 33 in Basal. Collectively, these findings indicate that novel subtype-specific lncRNAs may regulate key oncogenes and tumor suppressors, providing potential mechanistic insights into breast cancer subtypes.

### 3.8 Refinement of target mRNAs by correlation analysis and structural validation

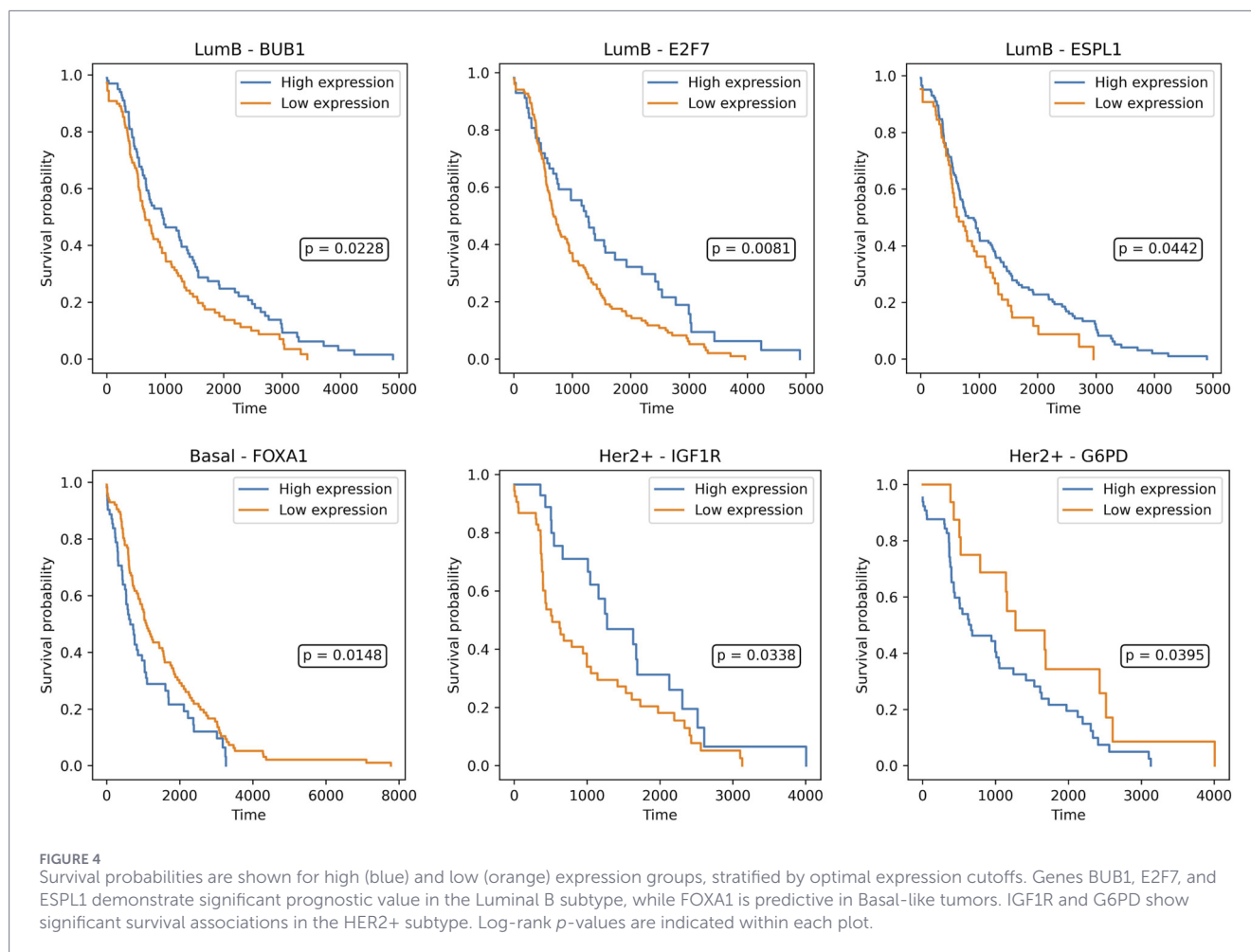
Correlation analysis revealed numerous subtype-specific lncRNA-mRNA pairs with strong associations. After applying the correlation thresholds ( $|r| \geq 0.40$ , FDR < 0.05), we retained nine highly correlated pairs in LumA, five in LumB, and three each in HER2+ and Basal (Figure 6A). To further refine these candidates, we used IntaRNA to generate binding energy distributions and rank interactions by stability. Pairs from LumA and LumB showed steeper declines in their energy plots, indicating a greater proportion of very stable interactions compared to those from HER2+ and Basal. From these distributions, we selected the top-scoring pairs with the most favorable  $\Delta G$  values (Figure 6B) for each subtype. For these high-confidence pairs, lncRRIsearch was used to generate hybridization maps, which highlighted discrete and stable base-paired regions. These visualizations confirmed that the shortlisted interactions are structurally capable of strong and specific RNA-RNA binding. Notably, eight novel subtype-specific lncRNA-mRNA pairs emerged as the most stable: *CUFF.25255-CIITA* and *CUFF.25255-IKZF3* (LumA); *CUFF.20237-USP8* and *CUFF.3888-TYRO3* (LumB); *CUFF.22414-CDKN2A* predicted via BLASTN to be similar to *lnc-ZNF624-1* and *LINC02875* (HER2+); *CUFF.26607-SMAD2* (Basal) and *CUFF.1961-POU2F2* (Basal) consistently appeared among the most stable interactions (Figures 6C–F). These interactions involved mRNA targets associated with oncogenes and tumor suppressors, underscoring their potential functional relevance in subtype-specific cancer regulation (Table 3).

### 3.9 Proposed nomenclature of novel lncRNAs

Based on similarity and predicted regulatory interactions, we propose new names for novel prognostic lncRNAs identified in our study (Table 1). These names reflect both their reference-like origin and target gene association, providing functional interpretability.

### 3.10 External cohort transferability analysis in CPTAC using inferred PAM50 subtypes

We assessed whether gene-level counterparts linked to prioritized CUFF-level candidates (Table 3) exhibit reproducible subtype-associated expression patterns. Three representative counterparts-LINC01133 (*CUFF.26607/L1RSM*), *NIFK-AS1* (*CUFF.20237/NARUS*), and *TEC* (*CUFF.25255/TERCI*) - showed significant expression differences across the subtypes (Kruskal-Wallis test on  $\log_2$  (TPM+1), BH-FDR-adjusted p-values: *LINC01133* =  $3.53 \times 10^{-7}$ , *NIFK-AS1* =  $4.00 \times 10^{-5}$ , *TEC* =  $6.53 \times 10^{-6}$ ; Supplementary Figure S4; Supplementary Table S1). Pairwise Wilcoxon comparisons indicated that the strongest subtype contrasts typically involved Basal-like tumors versus other subtypes, consistent with the marked transcriptional separation of Basal-like disease (Supplementary Table S2). Collectively, these results provide independent-cohort support that a subset of lncRNA-associated signals prioritized in TCGA display reproducible subtype-associated behavior in an external RNA-seq cohort.

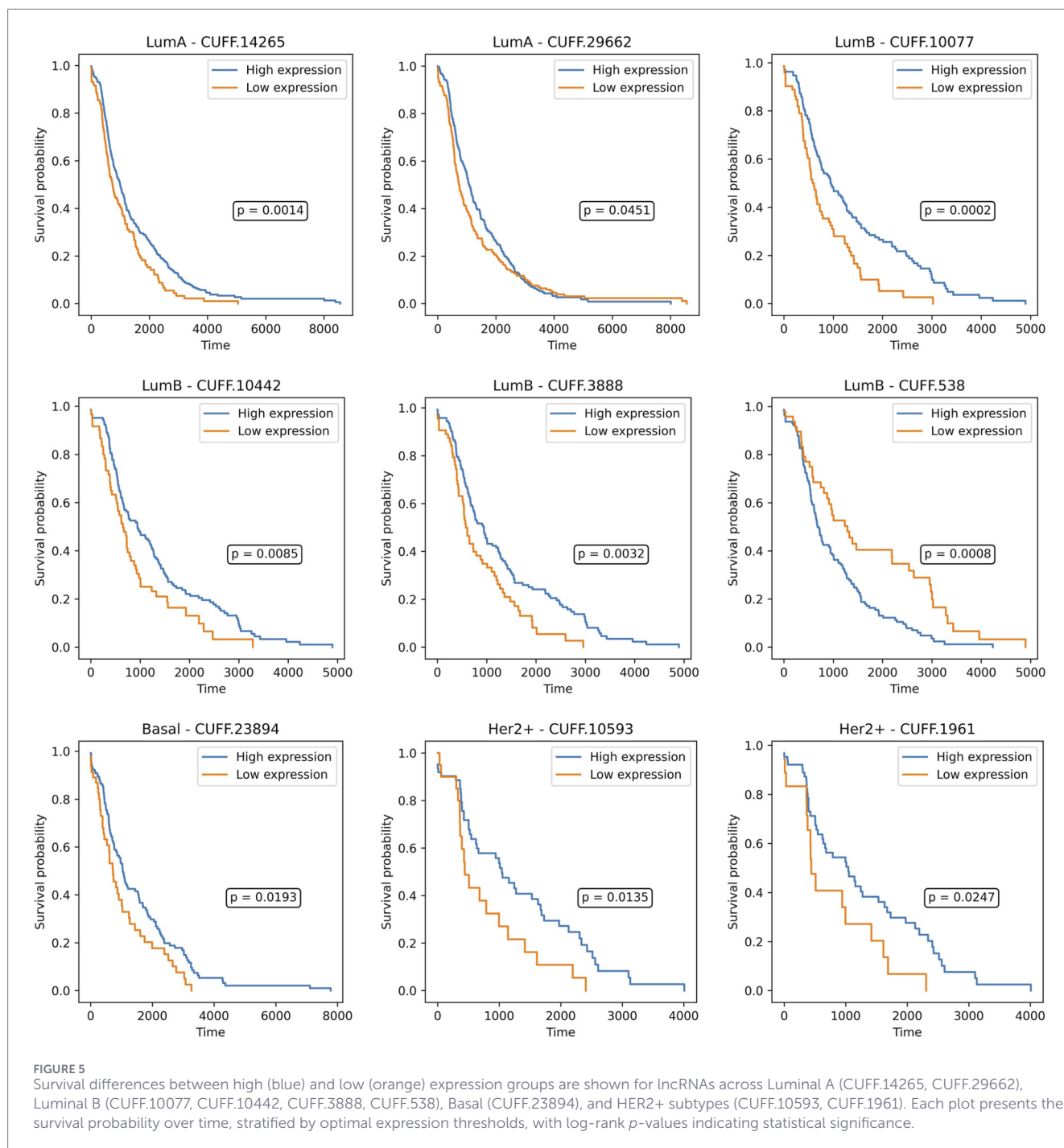


### 3.11 Liquid biopsy detectability analysis of candidate lncRNA counterparts

To evaluate whether prioritized lncRNA counterparts are detectable in blood-associated transcriptomic contexts, we queried LINC01133 (CUFF.26607/L1RSM) and TEC (CUFF.25255/TERCI) across publicly available blood-derived expression studies. This meta-level screening identified statistically significant signals for both candidates in peripheral blood and/or bone marrow-associated datasets (Supplementary Table S3; Supplementary Figure S5). Specifically, LINC01133 (CUFF.26607/L1RSM) showed multiple significant blood/bone marrow comparisons ( $Z \approx +1.98$ ;  $p \sim 6.5 \times 10^{-4}$ – $3.9 \times 10^{-3}$ ), while TEC (CUFF.25255/TERCI) exhibited consistent peripheral blood-associated signals with stronger statistical support ( $Z \approx +1.09$  to  $+1.21$ ;  $p \sim 3 \times 10^{-6}$ – $7 \times 10^{-6}$ ). Collectively, these findings indicate that at least a subset of the candidate lncRNA counterparts prioritized by our framework are detectable in blood-associated expression datasets, supporting feasibility for future evaluation in true liquid biopsy analytes and breast cancer-specific clinical settings.

## 4 Discussion

While prior studies have explored lncRNA associated subtype or prognostic signatures in breast cancer, those efforts have largely focused on binary classification tasks, limited lncRNA panels, or outcome prediction, rather than systematic multi-class subtype benchmarking at transcriptome scale. Breast cancer remains a clinically and molecularly heterogeneous disease, with subtype-specific biology influencing prognosis, treatment response, and therapeutic resistance. While traditional gene expression-based classifiers have primarily focused on protein-coding mRNAs, recent discoveries have established long non-coding RNAs (lncRNAs) as central regulators of cancer biology, including proliferation, metastasis, immune evasion, and hormone response (Zhang et al., 2013; Eptaminotaki et al., 2021; Campos-Parra et al., 2018; Mondal and Meeran, 2020; Qi and Du, 2013; Yin et al., 2023; Denaro et al., 2019; Yu et al., 2018; Sikder et al., 2024). Our study supports this growing evidence that lncRNA expression profiles alone can reliably distinguish breast cancer subtypes and serve as effective biomarkers. Notably, lncRNA-based models demonstrated a nearly identical performance to mRNA-only



models despite using one-sixth as many features, reinforcing that the non-coding transcriptome contains rich, subtype-specific information that rivals the coding transcriptome. When mRNA features were integrated with lncRNAs, classification performance improved further, the most notable for Luminal B, which is typically difficult to separate due to its mixed expression profile with other subtypes and clinical ambiguity (Ades et al., 2014; Creighton, 2012). The mRNA appears to add complementary resolution to the transcriptomic landscape, helping to refine borderline cases and enhance subtype fidelity. These observations align with previous reports indicating that integrative transcriptomic modeling yields

superior predictive and diagnostic accuracy in heterogeneous cancers (Yuan et al., 2019; Wang et al., 2023).

Furthermore, the consistent superiority of XGBoost across all omics configurations highlights the power of gradient boosting in modeling complex transcriptomic relationships. The robustness of these models across 10-fold cross-validation underscores their reliability and generalizability for practical use in clinical settings. Our subtype-specific biomarker discovery framework offers biological interpretability beyond classification. For example, several lncRNAs identified through SHAP-based filtering exhibited strong subtype-enriched expression patterns,



TABLE 3 Suggested names for novel prognostic lncRNAs.

Subtype	CUFF IDs	Predicted similar gene	Target gene	Suggested new name	Symbol
Luminal A	CUFF.25255	TEC	CIITA	TEC-like -reg-CIITA	TERCI
Luminal A	CUFF.25255	PCAT18	IKZF3	PCAT18-like-reg-IKZF3	PARIK
Luminal B	CUFF.20237	NIFK-AS1	USP8	NIFK-AS1-like-reg-USP8	NARUS
Luminal B	CUFF.3888	SLC30A4-AS1	TYRO3	SLC30A4-AS1-like-reg-TYRO3	SARTR
HER2+	CUFF.22414	Lnc-ZNF624-1	CDKN2A	ZNF624-like-reg-CDKN2A	ZRCDA
HER2+	CUFF.22414	LINC02875	CDKN2A	LINC02875-like-reg-CDKN2A	L2RCA
Basal	CUFF.26607	LINC01133	SMAD2	LINC01133-like-reg-SMAD2	L1RSM
Basal	CUFF.1961	MVP-DT	POU2F2	MVP-DT-like-reg-POU2F2	MVRPO

It is important to emphasize that pathway enrichment results in this study arise from supervised, classification-driven feature selection rather than unbiased transcriptome-wide differential expression analyses. Following ABS pipeline, we performed two complementary downstream analyses: (i) functional pathway profiling of mRNA features and (ii) functional characterization of lncRNAs to assess their prognostic and regulatory relevance. IPA analysis on the mRNA components showed clear differences in pathway activity across breast cancer subtypes. Luminal B tumors had strong signals related to cell cycle control and p53 pathways, which fit with their higher growth rates and more frequent p53 changes (Marvalim et al., 2023). Although Luminal A associated feature sets also showed enrichment for cell cycle related pathways, the signals reflect regulated checkpoint and replication associated genes that contribute to subtype discrimination. This is consistent with the known lower proliferation rates and relatively intact p53 function in Luminal A tumors (Marvalim et al., 2023; Shan et al., 2013). Basal and HER2+ subtypes showed enrichment in stress response, metabolic, and hypoxia-associated signaling pathways, consistent with microenvironmental adaptation (Gatza et al., 2011; Jarman et al., 2019; Kjölle et al., 2023). These subtype-specific enrichments help contextualize the mRNA features within known oncogenic pathways and support biological validity of selected panels.

In parallel, survival analysis and lncRNA characterization were performed on the lncRNA feature sets to evaluate the prognostic and functional significance. Survival analysis used a log-rank based approach that systematically evaluated all possible expression cutoffs, allowing for more sensitive identification of prognostic genes than median-based splits. Both mRNA and lncRNA features yielded subtype-specific markers associated with overall survival. Notably, several lncRNAs from our selected feature sets were found to stratify patients within their respective subtypes, highlighting their translational potential as noncoding prognostic biomarkers. For instance, CUFF.14265 and CUFF.29662 showed prognostic separation in Luminal A, while CUFF.23894 and CUFF.10593 were informative in HER2+ and Basal subtypes, respectively. In Luminal B, both coding genes, such as BUB1, ESPL1, and E2F7, and lncRNAs including CUFF.10077, CUFF.10442, and CUFF.538 were significantly associated with survival, reinforcing the multifactorial nature of prognosis in this aggressive subtype. Similarly, FOXA1 showed prognostic relevance in Basal tumors,

and IGF1R and G6PD were informative in HER2+, alongside additional lncRNAs CUFF.1961 and CUFF.25223. For lncRNA characterization, expression correlation with structural interaction prediction, we were able to filter out false positives and ensure that the identified lncRNA-mRNA pairs were both statistically supported and biologically plausible. The integration of OncoKB annotations further strengthened our findings by directly linking several novel lncRNAs to clinically validated oncogenes and tumor suppressors, underscoring their potential translational value. Our proposed naming framework, which connects each lncRNA to both its similarity gene and predicted target, provides a systematic way to interpret and track these biomarkers across studies. Importantly, the averaged scoring strategy based on the mean of Pearson and Spearman correlation coefficients and binding energy values revealed subtype-specific prognostic lncRNAs, with Luminal A and Luminal B subtypes showing particularly stable and confident interactions. Together, these analyses provide a comprehensive regulatory and clinical characterization of lncRNAs derived from ABS pipeline, which could be clinically meaningful lncRNA-mRNA signatures for breast cancer subtypes.

The subtype-specific lncRNA panels identified in this study have potential translational relevance. Compact lncRNA signatures could be incorporated into diagnostic or molecular subtyping assays, either alone or in combination with protein-coding genes, to refine intrinsic subtype classification, particularly in ambiguous cases such as Luminal B tumors. In addition, the association of selected lncRNAs with patient survival within specific subtypes suggests potential utility for risk stratification and prognostic modeling. Finally, the inferred lncRNA-mRNA regulatory relationships provide a framework for generating therapeutic hypotheses, whereby subtype-enriched lncRNAs may modulate key oncogenic pathways and warrant further functional investigation (Zhang et al., 2019).

A limitation of this study is the absence of large-scale external validation of CUFF-level lncRNAs across independent breast cancer cohorts. Widely used resources such as METABRIC are primarily microarray-based and therefore do not provide RNA-seq data that can be re-processed using the same transcriptome assembly and quantification strategy applied in this study. As a result, many CUFF-level lncRNAs identified here are not represented on array platforms, limiting the feasibility of direct external validation using such datasets.

## Conclusion

This study underscores the emerging significance of long non-coding RNAs (lncRNAs) as robust transcriptomic markers for breast cancer subtyping. By leveraging a robust multi-omics dataset and evaluating multiple ML models, we demonstrate that lncRNA expression alone achieves high predictive performance, particularly when modeled using XGBoost, which consistently outperformed other classifiers. Integrating lncRNA with mRNA features further enhanced classification and interpretability, most notably for the clinically challenging Luminal B subtype. The ABS framework enabled the identification of potential lncRNA-centered subtype-specific biomarker panels that retained high discriminative power while offering biological interpretability. Overall, these findings position the lncRNAs as functional regulators and promising biomarkers that merit further experimental validation for their potential role in precision oncology.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://portal.gdc.cancer.gov/>; <https://zenodo.org/records/17820359>.

## Author contributions

JP: Writing – original draft, Conceptualization, Data curation, Visualization, Validation, Methodology. AV: Data curation, Methodology, Validation, Writing – review and editing. CG: Writing – review and editing, Conceptualization, Funding acquisition, Investigation, Resources, Supervision, Visualization.

## Funding

The author(s) declared that financial support was received for this work and/or its publication. The authors would like to thank the

## References

- Ades, F., Zardavas, D., Bozovic-Spasojevic, I., Pugliano, L., Fumagalli, D., DE Azambuja, E., et al. (2014). Luminal B breast cancer: molecular characterization, clinical management, and future perspectives. *J. Clin. Oncol.* 32, 2794–2803. doi:10.1200/JCO.2013.54.1870
- Agarap, A. F. (2019). Deep learning using rectified linear units (ReLU). *Comput. Sci.* doi:10.48550/arXiv.1803.08375
- Ahmed, S., Kabir, M., Ali, Z., Arif, M., Ali, F., and Yu, D. J. (2018). An integrated feature selection algorithm for cancer classification using gene expression data. *Comb. Chem. High. Throughput Screen* 21, 631–645. doi:10.2174/1386207322666181220124756
- Alkuhlani, A., Nassef, M., and Farag, I. (2017). Multistage feature selection approach for high-dimensional cancer data. *Soft Comput.* 21, 6895–6906. doi:10.1007/s00500-016-2439-9
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2
- Anqi Mao, M. M., and Yutao, ZHONG (2023). Cross-entropy loss functions: theoretical analysis and applications. *Comput. Sci.* doi:10.48550/arXiv.2304.07288
- Ben Rabah, C., Sattar, A., Ibrahim, A., and Serag, A. (2025). A multimodal deep learning model for the classification of breast cancer subtypes. *Diagn. (Basel)* 15, 995. doi:10.3390/diagnostics15080995
- Bjorklund, S. S., Aure, M. R., Hakkinen, J., Vallon-Christersson, J., Kumar, S., Evensen, K. B., et al. (2022). Subtype and cell type specific expression of lncRNAs provide insight into breast cancer. *Commun. Biol.* 5, 834. doi:10.1038/s42003-022-03559-7
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 30, 1145–1159. doi:10.1016/s0031-3203(96)00142-2
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324
- Campos-Parra, A. D., Lopez-Urrutia, E., Orozco Moreno, L. T., Lopez-Camarillo, C., Meza-Menchaca, T., Figueroa Gonzalez, G., et al. (2018). Long non-coding RNAs as new master regulators of resistance to systemic treatments in breast cancer. *Int. J. Mol. Sci.* 19, 1–20. doi:10.3390/ijms19092711
- Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70.

Bioinformatics and Systems Biology Core (BSBC) facility at UNMC for providing the computational infrastructure and support. BSBC is partly supported by NIH awards (5P30CA036727, 2P20GM103427) to CG. JP was partly supported by an NIH award, 2P01AG029531.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fbinf.2026.1760987/full#supplementary-material>

- Cascianelli, S., Molineris, I., Isella, C., Masseroli, M., and Medico, E. (2020). Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer. *Sci. Rep.* 10, 14071. doi:10.1038/s41598-020-70832-2
- Chakravarty, D., Gao, J., Phillips, S. M., Kundra, R., Zhang, H., Wang, J., et al. (2017). OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* 2017, 1–16. doi:10.1200/PO.17.00011
- Choi, J. M., and Chae, H. (2023). moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. *BMC Bioinforma.* 24, 169. doi:10.1186/s12859-023-05273-5
- Chollet, F. (2015). Keras. *GitHub*.
- Coan, M., Haefliger, S., Ounzain, S., and Johnson, R. (2024). Targeting and engineering long non-coding RNAs for cancer therapy. *Nat. Rev. Genet.* 25, 578–595. doi:10.1038/s41576-024-00693-2
- Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44, e71. doi:10.1093/nar/gkv1507
- Creighton, C. J. (2012). The molecular profile of luminal B breast cancer. *Biologics* 6, 289–297. doi:10.2147/BTT.S29923
- Cristovao, F., Cascianelli, S., Canakoglu, A., Carman, M., Nanni, L., Pinoli, P., et al. (2022). Investigating deep learning based breast cancer subtyping using pan-cancer and multi-omic data. *IEEE/ACM Trans. Comput. Biol. Bioinform* 19, 121–134. doi:10.1109/TCBB.2020.3042309
- Denaro, N., Merlano, M. C., and Lo Nigro, C. (2019). Long noncoding RNAs as regulators of cancer immunity. *Mol. Oncol.* 13, 61–73. doi:10.1002/1878-0261.12413
- Diederik, P., and Kingma, J. B. (2014). Adam: a method for stochastic optimization. *Comput. Sci.* doi:10.48550/arXiv.1412.6980
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635
- Eptaminotaki, G. C., Wolff, N., Stellas, D., Sifakis, K., and Baritaki, S. (2021). Long non-coding RNAs (lncRNAs) in response and resistance to cancer immunosurveillance and immunotherapy. *Cells* 10, 3313. doi:10.3390/cells10123313
- Fabian Pedregosa, G. V., Gramfort, ALEXANDRE, Vincent, MICHEL, Bertrand, THIRION, Grisel, OLIVIER, Blondel, MATHIEU, et al. (2011). AND ÉDOUARD DUCHESNAY 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 2825–2830. doi:10.48550/arXiv.1201.0490
- Fukunaga, T., Iwakiri, J., Ono, Y., and Hamada, M. (2019). LncRRsearch: a web server for lncRNA-RNA interaction prediction integrated with tissue-specific expression and subcellular localization data. *Front. Genet.* 10, 462. doi:10.3389/fgene.2019.00462
- Gatza, M. L., Kung, H. N., Blackwell, K. L., Dewhurst, M. W., Marks, J. R., and Chi, J. T. (2011). Analysis of tumor environmental response and oncogenic pathway activation identifies distinct basal and luminal features in HER2-related breast tumor subtypes. *Breast Cancer Res.* 13, R62. doi:10.1186/bcr2899
- Geffen, Y., Anand, S., Akiyama, Y., Yaron, T. M., Song, Y., Johnson, J. L., et al. (2023). Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation. *Cell* 186, 3945–3967 e26. doi:10.1016/j.cell.2023.07.013
- Grammatikakis, I., Panda, A. C., Abdelmohsen, K., and Gorospe, M. (2014). Long noncoding RNAs (lncRNAs) and the molecular hallmarks of aging. *Aging (Albany NY)* 6, 992–1009. doi:10.18632/aging.100710
- Guda, C. (2025). VEERAPPA A. *Cancer Epigenetics*. London: Springer Nature
- Guestin, T. C. A. C. (2016). “XGBoost: a scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. San Francisco, California, USA: Association for Computing Machinery, 785–794.
- Ian Goodfello, Y. B. A. A. C. (2016). *Deep learning*. MIT Press.
- Jain, A. K., Mao, J., and Mohiuddin, K. M. (1996). Artificial neural networks: a tutorial. *Computer* 29, 31–44. doi:10.1109/2.485891
- Jarman, E. J., Ward, C., Turnbull, A. K., Martinez-Perez, C., Meehan, J., Xintaropoulou, C., et al. (2019). HER2 regulates HIF-2alpha and drives an increased hypoxic response in breast cancer. *Breast Cancer Res.* 21, 10. doi:10.1186/s13058-019-1097-0
- Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* 53, 457–481. doi:10.1080/01621459.1958.10501452
- Kjolle, S., Finne, K., Birkeland, E., Ardawatia, V., Winge, L., Aziz, S., et al. (2023). Hypoxia induced responses are reflected in the stromal proteome of breast cancer. *Nat. Commun.* 14, 3724. doi:10.1038/s41467-023-39287-7
- Kleinbaum, D. G., and Klein, M. (2012). “Kaplan-meier survival curves and the log-rank Test. Survival analysis,” in *Statistics for biology and health*. New York, NY: Springer.
- Lars Sthle, S. W. (1989). Analysis of variance (ANOVA). *Chemom. Intelligent Laboratory Syst.* 6, 259–272.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352
- Li, Z., Zheng, J., Feng, Y., Li, Y., Liang, Y., Liu, Y., et al. (2021). Integrated analysis identifies a novel lncRNA prognostic signature associated with aerobic glycolysis and hub pathways in breast cancer. *Cancer Med.* 10, 7877–7892. doi:10.1002/cam4.4291
- Li, C., Wang, X., Chen, T., Li, W., and Yang, Q. (2022a). A novel lncRNA panel for risk stratification and immune landscape in breast cancer patients. *Int. J. Gen. Med.* 15, 5253–5272. doi:10.2147/IJGM.S366335
- Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F., et al. (2022b). MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Front. Genet.* 13, 806842. doi:10.3389/fgene.2022.806842
- Liao, Y., Smyth, G. K., and Shi, W. (2013). The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 41, e108. doi:10.1093/nar/gkt214
- Lin, Y., Zhang, W., Cao, H., Li, G., and Du, W. (2020). Classifying breast cancer subtypes using deep neural networks based on multi-omics data. *Genes (Basel)* 11, 1–18. doi:10.3390/genes11080888
- Lin, X., Lu, Y., Zhang, C., Cui, Q., Tang, Y. D., Ji, X., et al. (2024). LncRNADisease v3.0: an updated database of long non-coding RNA-Associated diseases. *Nucleic Acids Res.* 52, D1365–D1369. doi:10.1093/nar/gkad828
- Liu, T., H., J., Liao, T., Pu, R., Liu, S., and Peng, Y. (2022). A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data. *IRBM* 43, 62–74. doi:10.1016/j.irbm.2020.12.002
- Liu, Y., Chen, J., Yang, D., Liu, C., Tang, C., Cai, S., et al. (2025). Machine learning combined with multi-omics to identify immune-related lncRNA signature as biomarkers for predicting breast cancer prognosis. *Sci. Rep.* 15, 23863. doi:10.1038/s41598-025-10186-9
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8
- Lupat, R., And, S. L. O. I., and Li, J. (2023). Moanna: multi-omics autoencoder-based neural network algorithm for predicting breast cancer subtypes. *IEEE Access* 11, 10912–10924. doi:10.1109/access.2023.3240515
- Mahato, R. K., Bhattacharya, S., Khullar, N., Sidhu, I. S., Reddy, P. H., Bhatti, G. K., et al. (2024). Targeting long non-coding RNAs in cancer therapy using CRISPR-Cas9 technology: a novel paradigm for precision oncology. *J. Biotechnol.* 379, 98–119. doi:10.1016/j.jbiotec.2023.12.003
- Mann, M., Wright, P. R., and Backofen, R. (2017). IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res.* 45, W435–W439. doi:10.1093/nar/gkx279
- Martin Abadi, P. B., Chen, JIANMIN, Chen, ZHIFENG, Davis, ANDY, Dean, JEFFREY, Devin, MATHIEU, et al. (2016). “TensorFlow: a system for large-scale machine learning,” in *Proceedings of the 12th USENIX conference on operating systems design and implementation (OSDI'16)* (USA: USENIX Association), 265–283.
- Marvalim, C., Datta, A., and Lee, S. C. (2023). Role of p53 in breast cancer progression: an insight into p53 targeted therapy. *Theranostics* 13, 1421–1442. doi:10.7150/tno.81847
- Mathias, C., Muzzi, J. C. D., Antunes, B. B., Gradia, D. F., Castro, M. A. A., and Carvalho DE Oliveira, J. (2021). Unraveling immune-related lncRNAs in breast cancer molecular subtypes. *Front. Oncol.* 11, 692170. doi:10.3389/fgene.2021.692170
- Miano, V., Ferrero, G., Reineri, S., Caizzi, L., Annaratone, L., Ricci, L., et al. (2016). Luminal long non-coding RNAs regulated by estrogen receptor alpha in a ligand-independent manner show functional roles in breast cancer. *Oncotarget* 7, 3201–3216. doi:10.18632/oncotarget.6420
- Miron, B., Kurs, A. J., and Rudnicki, WITOLD R. (2010). Boruta - a system for feature selection. *Fundam. Inf.* 101, 271–285. doi:10.3233/FI-2010-288
- Mishra, N. K., Southekal, S., and Guda, C. (2019). Survival analysis of multi-omics data identifies potential prognostic markers of pancreatic ductal adenocarcinoma. *Front. Genet.* 10, 624. doi:10.3389/fgene.2019.00624
- Mondal, P., and Meeran, S. M. (2020). Long non-coding RNAs in breast cancer metastasis. *Noncoding RNA Res.* 5, 208–218. doi:10.1016/j.ncrna.2020.11.004
- Patel, J. C., Shakyawar, S. K., Sethi, S., and Guda, C. (2025). GAIN-BRCA: a graph-based AI-net framework for breast cancer subtype classification using multiomics data. *Bioinform Adv.* 5, vbaf116. doi:10.1093/bioadv/vbaf116
- Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Comput. Sci.* doi:10.48550/arXiv.2010.16061
- Qi, P., and Du, X. (2013). The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine. *Mod. Pathol.* 26, 155–165. doi:10.1038/modpathol.2012.160
- Sarkar, S., and Mali, K. (2022). Breast cancer subtypes classification with hybrid machine learning model. *Methods Inf. Med.* 61, 68–83. doi:10.1055/s-0042-1751043
- Scott Lundberg, S.-I. L. (2017). A unified approach to interpreting model predictions. *Comput. Sci.* doi:10.48550/arXiv.1705.07874
- Sethi, S., Shakyawar, S., Reddy, A. S., Patel, J. C., and Guda, C. (2024). A machine learning model for the prediction of COVID-19 severity using RNA-Seq, clinical, and Co-Morbidity data. *Diagn. (Basel)* 14, 1284. doi:10.3390/diagnostics14121284

- Shaath, H., Elango, R., and Alajez, N. M. (2021). Molecular classification of breast cancer utilizing long non-coding RNA (lncRNA) transcriptomes identifies novel diagnostic lncRNA panel for triple-negative breast cancer. *Cancers (Basel)* 13, 5350. doi:10.3390/cancers13215350
- Shakyawar, S., Southekal, S., and Guda, C. (2022). minRULS: prediction of miRNA-mRNA target site interactions using regularized least square method. *Genes (Basel)* 13, 1528. doi:10.3390/genes13091528
- Shakyawar, S. K., Sajja, B. R., Patel, J. C., and Guda, C. (2024). iCluF: an unsupervised iterative cluster-fusion method for patient stratification using multiomics data. *Bioinform Adv.* 4, vbae015. doi:10.1093/bioadv/vbae015
- Shan, M., Zhang, X., Liu, X., Qin, Y., Liu, T., Liu, Y., et al. (2013). P16 and p53 play distinct roles in different subtypes of breast cancer. *PLoS One* 8, e76408. doi:10.1371/journal.pone.0076408
- Sikder, S., Bhattacharya, A., Agrawal, A., Sethi, G., and Kundu, T. K. (2024). Micro-RNAs in breast cancer progression and metastasis: a chromatin and metabolic perspective. *Heliyon* 10, e38193. doi:10.1016/j.heliyon.2024.e38193
- Su, X., Malouf, G. G., Chen, Y., Zhang, J., Yao, H., Valero, V., et al. (2014). Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. *Oncotarget* 5, 9864–9876. doi:10.18632/oncotarget.2454
- Su, Y., Bai, Q., Zhang, W., Xu, B., and Hu, T. (2025). The role of long non-coding RNAs in modulating the immune microenvironment of triple-negative breast cancer: mechanistic insights and therapeutic potential. *Biomolecules* 15, 454. doi:10.3390/biom15030454
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi:10.1038/nbt.1621
- Tuna, M., Mills, G. B., and Amos, C. I. (2025). The role of long non-coding RNAs in lung cancer metastasis: molecular mechanisms, pathogenesis and clinical implications. *Clin. Transl. Med.* 15, e70429. doi:10.1002/ctm2.70429
- Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., and Li, W. (2013). CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41, e74. doi:10.1093/nar/gkt006
- Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., et al. (2021). MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* 12, 3445. doi:10.1038/s41467-021-23774-w
- Wang, H., Liu, B., Long, J., Yu, J., Ji, X., Li, J., et al. (2023). Integrative analysis identifies two molecular and clinical subsets in luminal B breast cancer. *iScience* 26, 107466. doi:10.1016/j.isci.2023.107466
- Wasson, M. D., Venkatesh, J., Cahill, H. F., Mclean, M. E., Dean, C. A., and Marcato, P. (2024). LncRNAs exhibit subtype-specific expression, survival associations, and cancer-promoting effects in breast cancer. *Gene* 901, 148165. doi:10.1016/j.gene.2024.148165
- Webb, G. I. (2011). *Naïve bayes. Encyclopedia of machine learning*. Boston, MA: Springer.
- Wu, J., and Hicks, C. (2021). Breast cancer type classification using machine learning. *J. Pers. Med.* 11, 61. doi:10.3390/jpm11020061
- Xia, Q., Yan, Q., Wang, Z., Huang, Q., Zheng, X., Shen, J., et al. (2023). Disulfidptosis-associated lncRNAs predict breast cancer subtypes. *Sci. Rep.* 13, 16268. doi:10.1038/s41598-023-43414-1
- Yin, Q., Ma, H., Bamunarachchi, G., Zheng, X., and Ma, Y. (2023). Long non-coding RNAs, cell cycle, and human breast cancer. *Hum. Gene Ther.* 34, 481–494. doi:10.1089/hum.2023.074
- Yu, W. D., Wang, H., He, Q. F., Xu, Y., and Wang, X. C. (2018). Long noncoding RNAs in cancer-immunity cycle. *J. Cell Physiol.* 233, 6518–6523. doi:10.1002/jcp.26568
- Yu, Y., Ren, W., He, Z., Chen, Y., Tan, Y., Mao, L., et al. (2023). Machine learning radiomics of magnetic resonance imaging predicts recurrence-free survival after surgery and correlation of lncRNAs in patients with breast cancer: a multicenter cohort study. *Breast Cancer Res.* 25, 132. doi:10.1186/s13058-023-01688-3
- Yuan, C. L., Jiang, X. M., Yi, Y., E, J. F., Zhang, N. D., Luo, X., et al. (2019). Identification of differentially expressed lncRNAs and mRNAs in luminal-B breast cancer by RNA-Sequencing. *BMC Cancer* 19, 1171. doi:10.1186/s12885-019-6395-5
- Zhang, H., Chen, Z., Wang, X., Huang, Z., He, Z., and Chen, Y. (2013). Long non-coding RNA: a new player in cancer. *J. Hematol. Oncol.* 6, 37. doi:10.1186/1756-8722-6-37
- Zhang, D., Zhou, X., and He, F. (2018). Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. *IEEE Access* 6, 28936–28944. doi:10.1109/access.2018.2837654
- Zhang, J., Le, T. D., Liu, L., and Li, J. (2019). Inferring and analyzing module-specific lncRNA-mRNA causal regulatory networks in human cancer. *Brief. Bioinform* 20, 1403–1419. doi:10.1093/bib/bby008
- Zhao, Z., Guo, Y., Liu, Y., Sun, L., Chen, B., Wang, C., et al. (2021). Individualized lncRNA differential expression profile reveals heterogeneity of breast cancer. *Oncogene* 40, 4604–4614. doi:10.1038/s41388-021-01883-6
- Zhu, J., Zhao, Z., Yin, B., Wu, C., Yin, C., Chen, R., et al. (2025). An integrated approach of feature selection and machine learning for early detection of breast cancer. *Sci. Rep.* 15, 13015. doi:10.1038/s41598-025-97685-x