JOURNAL OF CLINICAL MEDICINE OF KAZAKHSTAN (E-ISSN 2313-1519)
Review Article DOI: https://doi.org/10.23950/jcmk/18568

Cross-Modal Self-Attention Fusion for Breast Cancer Subtype Classification Using Multi-Omics Data

Kurmash Zhumagozhayev[1], Tomiris Zhaksylyk[1], Beibit Abdikenov[1], Temirlan Karibekov[2], Liliya Skvortsova[3], Adil Faizullin[1]

[1]Science and Innovation Center "Artificial Intelligence", Astana IT University, Astana, Kazakhstan
[2]Science and Innovation Center "MedTech", Astana IT University, Astana, Kazakhstan
[3]Laboratory of Molecular Genetics, Institute of Genetics and Physiology, Committee of Science of the Ministry of Science and Higher Education, Almaty, Kazakhstan

Received: 2025-10-24.
Accepted: 2026-05-16.

This work is licensed under a Creative Commons Attribution 4.0 International License

J Clin Med Kaz 2026; 23(3): 40-51

Corresponding authors:

Tomiris Zhaksylyk
E-mail: zhaksylyk.tomiris@astanait.edu.kz.
ORCID: 0009-0002-8749-1967;

Beibit Abdikenov.
E-mail: beibit.abdikenov@astanait.edu.kz.
ORCID: 0000-0002-0284-0949.

ABSTRACT

Background: Accurate classification of breast cancer subtypes is essential for personalized therapy and prognosis. Traditional subtype classification basically relies on gene expression profiling, usually overlooking other genomic signals like copy-number alterations (CNA) and mutations. At the same time most of the multi-omics models often rely on early or late fusion strategies, which do not capture complex inter-modality interactions.

Methods: This study proposes a cross-modal transformer-based approach that integrates gene expression, copy number alterations, and mutation data for robust breast cancer subtype classification. Each omics modality is encoded as a separate sequence and projected into a shared embedding space. Gene expression is treated as the primary modality and enriched through cross-modal self-attention mechanisms with CNA and mutation features. The final enriched embeddings are flattened and passed through a residual-connected MLP classifier. We evaluate performance on the METABRIC dataset using ElasticNet-selected top-K features (K = 300, 500, 1000, 1500) and mostly focus on macro F1-score, weighted F1-score, and ROC AUC due to class imbalance.

Results: Integrating copy-number and mutation data with expression features improved subtype classification across most feature set sizes. The tri-omic model (EXP+CNA+MUT) achieved the best performance for smaller feature sets (K = 300–500), whereas for larger feature sets (K = 1000) the highest scores were obtained by the bi-omic model (EXP+CNA) with macro-F1 = 0.859, weighted F1 = 0.868, accuracy = 0.866 and ROC AUC = 0.969. Paired statistical tests across five folds showed that differences between modality configurations did not reach significance at any K (all $p > 0.09$), whereas feature-set size did.

Within the EXP+CNA configuration alone, macro-F1 increased significantly from K = 300 to K = 500 (paired t-test, $p = 0.012$) and from K = 300 to K = 1000 ($p = 0.036$); and in the higher-powered pooled analysis across all three modality configurations (n = 15 paired folds), K = 1000 also outperformed K = 300 ($p = 0.030$).

Conclusion: This pipeline demonstrates an application of cross-modal attention for omics integration in subtype classification task, offering a scalable and biologically grounded alternative to traditional fusion approaches.

Keywords: breast cancer; copy number alteration; cross-modal attention; gene expression; METABRIC; molecular subtype classification; multi-omics integration; mutation; transformer.

Introduction

Cancer remains a leading cause of mortality worldwide and in Kazakhstan, where socioeconomic development, urbanization, aging, and lifestyle transitions have driven a rising non-communicable disease burden. Nationwide data from 2014–2022 show stable overall cancer mortality but a temporary increase in the mortality-to-incidence ratio during the COVID-19 pandemic, as reported in recent national analyses [1, 2]. Breast cancer significantly contributes to this burden, with increasing prevalence (from 30.4 to 50.6 per 10,000 population between 2014 and 2019) and incidence rates (peaking at 7.3 per 10,000 in 2016), alongside high mortality in older age groups and associations with comorbidities like diabetes [3]. Against this backdrop, the present study develops a cross-modal Transformer that directly fuses gene-expression, copy-number and mutation signals for breast-cancer molecular-subtype classification, evaluated on the METABRIC cohort as a methodological testbed for future application to national cohorts.

Effective prognosis and diagnosis of breast cancer are crucial for improving outcomes, as early detection and risk stratification can address aggressive tumors, treatment related complications, and long-term effects [4-7]. Advanced imaging modalities, including mammography, ultrasound, MRI, digital breast tomosynthesis, and breast CT, enhance visualization and early detection, particularly in dense breasts (BI-RADS C/D, linked to larger tumors and poorer outcomes), while supplemental imaging raises overall sensitivity to 90-97% [4, 5].

Bridging imaging-based diagnosis with molecular insights, liquid biopsies provide non-invasive real-time monitoring of tumor evolution through biomarkers like circulating tumor DNA (ctDNA) and cells, enabling detection of resistance mutations and minimal residual disease to guide precision oncology [8]. This integration underscores the need for multi-omics approaches to capture breast cancer's genetic heterogeneity.

Beyond this phenotypic-surveillance layer, effective stratification ultimately requires access to the tumor’s underlying molecular program. Accurate molecular subtyping of breast cancer is essential for guiding personalized therapy and prognosis. Gene expression profiling studies have revealed that breast cancer is not a single disease but a collection of subtypes with distinct molecular signatures and clinical outcomes [9]. Pioneering work by Sorlie et al. identified intrinsic subtypes Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like based on unsupervised clustering of tumor mRNA expression profiles, showing their relevance to treatment outcomes and survival trajectories [10]. These findings laid the foundation for the development of clinical assay such as the PAM50 classifier, a 50-gene expression panel designed to assign breast tumors into the intrinsic subtypes in a supervised manner [11]. In addition, researchers proposed a sixth distinct group, the Claudin-low subtype, characterized by low expression of cell-cell adhesion genes (e.g., claudin 3/4/7, E-cadherin) and high expression of mesenchymal and immune response markers, marking a more stem-like, aggressive phenotype [12].

Despite the clinical utility of expression-based classifiers, they might not fully capture key genomic alterations that drive breast cancer heterogeneity. For instance, CNAs gains and losses of chromosomal segments contribute significantly to oncogenic pathways and are highly prevalent in breast tumors. Somatic mutations in genes like TP53, PIK3CA, and BRCA1/2 further influence tumor behavior and therapeutic sensitivity. The METABRIC dataset comprising matched gene expression, CNA, and clinical profiles from around 2,000 breast tumors has been pivotal in uncovering these relationships [13]. At present, no locally curated breast cancer cohort in Kazakhstan combines matched gene expression, CNA and somatic mutation profiles at the scale required for supervised deep learning. We therefore utilize METABRIC dataset as a methodological basis, with architecture deliberately designed to transfer to matched national cohorts as they mature.

Given this complexity, multi-omics integration has become a key strategy for enhancing molecular classification [14-16]. However, most current multi-omics models employ either early fusion (feature concatenation) or late fusion (decision-level merging), both of which might have certain limitations in capturing inter-modality interactions [17]. Early fusion may dilute important signals across modalities, while late fusion ignores dependencies between, for example, a gene expression level and its mutational or copy number status [18]. Emerging deep learning methods, especially those based on self-attention mechanisms such as Transformers, offer a way to dynamically model cross-modal relationships through learned attention patterns [9, 18, 19].

Dedicated cross-modal attention where gene expression tokens directly attend to copy number and mutation tokens, as opposed to attention applied separately within each modality remains mostly unexplored for breast cancer subtype classification. To address this, we propose a cross-modal Transformer-based approach for breast cancer subtype classification using gene expression (EXP), copy number alterations (CNA), and mutation (MUT) data from METABRIC. Each omics modality is independently represented as a sequence of features and encoded into a shared latent embedding space. Gene expression is designated as the primary modality and is subsequently enriched through cross-modal attention, leveraging CNA and mutation embeddings as contextual sources to incorporate complementary biological signals. After the enriched representation is flattened and passed through a residual-connected multilayer perceptron (MLP) classifier.

Related work

learning has emerged as a powerful tool in bioinformatics, but it faces challenges with high-dimensional omics data and limited samples [20]. High feature counts (tens of thousands of genes) and class imbalance complicate model training and interpretation. A review by Nasser and Yusof highlights that while deep learning has achieved remarkable results in breast cancer imaging diagnostics and some genomic tasks, issues like interpretability and data heterogeneity remain obstacles [14]. Similarly, Abdikenov et al. survey the landscape of machine learning in breast cancer diagnosis and emphasize emerging trends in multimodal data integration combining different data sources (imaging, genomic, clinical) is seen as a key to improving predictive performance [21]. Our work aligns with these trends by focusing on multi-omics fusion via a novel transformer-based approach to better capture the complexity of breast tumors.

Feature Selection and Dimensionality Reduction

Because of the "large p, small n" problem in genomics (more features than samples), feature selection is often a crucial step for subtype classification tasks. One classical approach is to identify a small gene panel that discriminates against subtypes, as done with PAM50 [11]. Beyond expert-curated genes, data-driven methods are widely used. Differential expression analysis can filter genes; for instance, Choi and Chae (moBRCA-net) selected the top 1000 differentially expressed genes and associated features from other omics in their multi-omics model

[22]. Other works have used mutual information or entropy-based filtering. Liu et al. applied a mutual information feature selection to identify genes distinguishing triple-negative breast cancer subgroups [23]. An emerging strategy is to incorporate feature selection into the model training itself. Similarly, Guo et al. introduced a neural-network-based gene selection using knockoff filters to identify predictive genes with statistical control of false discoveries [24]. Considering high number of predictors versus number of observations, the ElasticNet regularization, which linearly combines L1 and L2 penalties, is another choice to enforce sparse gene selection while training a classifier [25]. Evolutionary and heuristic algorithms have also been explored for gene selection: Molaei et al. used particle swarm optimization to pick informative microRNAs before classification, and Andelic and Segota evolved symbolic expressions (combinations of a small number of genes) that achieved high accuracy in subtype classification (reporting ~99% accuracy on a small microarray dataset of 6 subtypes after balancing) [26, 27]. Unsupervised dimensionality reduction can likewise alleviate the curse of dimensionality. Bruno and Calimeri demonstrated that applying dimensionality reduction techniques such as PCA on gene expression data, alongside the integration of clinical features, could improve the visualization and classification of breast tumors [28]. In addition, autoencoder-based compression has been used to denoise and reduce genomic data: Arafa et al. developed a reduced-noise autoencoder that mitigated class imbalance effects and improved cancer genomic classification [29]. These methods show that carefully reducing feature space either through statistical selection or learned lower-dimensional representations can significantly benefit subtype classifiers by focusing on models on the most relevant biological signals.

Single-Omics Deep Learning Models

Early applications of machine learning for subtype classification often used single-omics data, primarily gene expression. Traditional classifiers (SVMs, Random Forests, etc.) achieved moderate success, but deep learning models have started to outperform them by capturing nonlinear patterns. Mostavi et al. proposed CNN-based models operating on gene expression profiles to classify tumor types across cancers; for breast cancer subtypes, such CNNs can automatically learn groups of co-expressed genes relevant to each subtype [30]. Mohamed et al. more recently designed a "bio-inspired" deep CNN for breast cancer detection using gene expression data, demonstrating that network architectures tailored to genomic data (e.g., using layers that mimic gene-gene interaction patterns) can improve accuracy [31]. Beykikhoshk et al. introduced an attention mechanism in a model called DeepTRIAGE to compute personalized biomarker scores from gene expression; they specifically classified Luminal A vs Luminal B subtypes and used attention to highlight genes contributing to the distinction for each patient (improving interpretability). These single-omics deep models typically report high performance on their focused tasks (e.g., DeepTRIAGE achieved over 84% accuracy distinguishing Luminal A vs B), but they inherently ignore other molecular information [9]. Our approach builds on the successes of deep learning in capturing transcriptomic patterns and extends it by integrating additional omics modalities to capture a more holistic view of the tumor.

Multi-Omics Integration Strategies. Integrating multiple data types (gene expression, CNAs, mutations, methylation, etc.) is challenging due to differing data scales and the potential lack of one-to-one feature correspondence. Nonetheless, many works have shown that multi-omics models can outperform single-omic ones for subtyping [16, 22]. Fusion strategies are generally categorized as early fusion (concatenate raw features or learned representations from each modality and then classify), late fusion (build separate models on each modality and then combine their predictions), or hybrid approaches in between. Early fusion is exemplified by Lin et al.'s DeepMO model, which trained parallel neural network subnetworks for mRNA, DNA methylation, and CNA data and then concatenated the learned features for final subtype prediction. DeepMO showed that a simple fully connected integration of multi-omic features already improved accuracy over single-omic models on TCGA data [16]. However, early fusion can struggle when one omics type dominates the signal or when there are many irrelevant features. Late fusion approaches, such as ensemble methods, build an expert model per modality and then aggregate decisions. For instance, Arya and Saha proposed a stacked ensemble where separate deep models for different modalities (expression CNA and clinical variables) were combined to predict breast cancer prognosis while their task was prognosis rather than subtype, the principle is similar – each data source is first mined independently for predictive insight [32]. Late fusion can be robust if one data type fails, but it may miss cross-modality feature interactions. More recent methods therefore explore intermediate or joint fusion methods that preserve modality-specific modeling while enabling inter-modality interaction. Graph-based models interpret multi-omics data as a network: Tanvir et al. introduced MOGAT, which builds a graph where nodes represent samples (with multi-omic feature vectors) and uses graph attention networks to learn sample embeddings that consider multi-omic similarity – yielding improved cancer subtype predictions by leveraging patient–patient relationships across omics [33]. Another graph-based approach, MoGCN by Li et al. compresses each omics layer with an autoencoder and fuses the resulting sample representations into a patient similarity network via Similarity Network Fusion; a graph convolutional network is the applied over this patient graph to classify cancer subtypes [18]. These graph-based models explicitly leverage cross-sample relationships induced by multi-omic similarity and have demonstrated strong performance in pan-cancer subtype classification.

Cascianelli et al. leveraged a pan-cancer multi-omic dataset, first training deep (especially semi-supervised) models across various cancer types including breast cancer and then tuning specifically for breast cancer subtype classification [15]. Their results indicate that both data quantity and heterogeneity, even beyond breast-specific samples, can improve model generalization and subtype discrimination. Attention mechanisms have also been applied at the feature level for multi-omics integration. One notable example is the proposed moBRCA-net framework by Choi and Chae [22]. In moBRCA-net, each omics modality gene expression, DNA methylation, and microRNA were processed independently through a self-attention module that assigns an importance weight to each feature within its respective modality. The resulting weighted representations are then concatenated and used for subtype classification. This design allows the model to highlight which genes, CpG sites, or miRNAs are most relevant, thus addressing the interpretability challenge commonly associated with deep models.

MoBRCA-net achieved an average accuracy of 89,1% and a F1-score of 0.887 on breast cancer subtypes from the TCGA dataset, outperforming ML-based models without attention mechanisms. However, it is important to note that moBRCA-net applies attention separately within each modality and does

not explicitly model cross-modal interactions. In contrast, our proposed method takes a step further by employing cross-modal attention, enabling features in one modality to directly attend to and integrate information from another. In this Transformer mechanism, cross-attention enables elements in one sequence (e.g., decoder tokens) to "attend" to elements in another sequence (e.g., encoder outputs), allowing rich contextualization. Similarly, in our model, a gene expression level can "look at" CNA or mutation features either from the same gene or from others when forming its representation [19]. By doing so, we allow direct cross-modal interactions, enhancing integration and interpretation of multi-omics data. To our knowledge, few works in cancer genomics have used transformer-style cross-modal attention for data fusion.

In summary, the state-of-the-art in breast cancer molecular subtype classification includes a spectrum of approaches: from conventional biomarkers and simple model ensembles to advanced deep learning architectures with attention, graph networks, and multi-omics integration.

Our approach is a cross-modal transformer that enriches gene expression features with CNA and mutation context, and which captures complex inter-modality relationships that early fusion or per-modality attention methods might miss. We hypothesize that this leads to more discriminative and robust representations for subtype classification, especially in scenarios of class imbalance or when the signal in any single modality is weak.

Methods

Data and Preprocessing

We evaluated our approach on the METABRIC dataset, a large breast cancer cohort with clinical, genomic, and transcriptomic data, that was acquired via cBioPortal [13, 34]. METABRIC provides gene expression profiles (originally from Illumina HT-12 v3 microarrays), somatic copy-number alterations (CNA), and somatic mutation data for each tumor, along with detailed clinical annotations.

We used the molecular subtype labels provided by METABRIC, which originally had six subtypes (5 PAM50 intrinsic subtypes: Luminal A, Luminal B, HER2-enriched, Basal-like, Normal-like) and the Claudin-low subtype (the latter was assigned based on gene expression clustering) [12]. Due to the low number of samples of Normal-like and Claudin-low we decided to omit them. This resulted in a 4-class classification task.

The dataset comprises 1523 breast tumor samples with complete data across all three omics modalities. Samples with missing data in any modality were excluded. Gene expression data (20603 genes) were used as log2-transformed and z-score normalized values relative to the expression distribution across all samples. CNA values (22544 genes) were represented as discrete copy-number calls: "- 2" homozygous deletion, "-1" hemizygous deletion, "0" neutral, "1" gain, and "2" high-level amplification, following contamination correction and thresholding.

Mutation data for 173 genes were encoded as multi-hot vectors indicating the presence or absence of specific Gene + Variant Classification combinations, resulting in 885 binary features. In the training set, 13 distinct Variant Classification categories were observed, including Missense_Mutation, Nonsense_Mutation, Frame_Shift_Del, Frame_Shift_Ins, In_Frame_Del, In_Frame_Ins, Nonstop_Mutation, Translation_Start_Site, Splice_Site, Splice_Region, as well as synonymous and non-coding categories such as Silent, Intron, and 5'UTR. We used this encoding to allow the model to capture mutation-specific patterns relevant to subtype classification, while mitigating sparsity from rare individual mutations. For example, if a tumor sample had a missense mutation in the gene PIK3CA, the corresponding binary feature "PIK3CA_Missense_Mutation" was set to 1, while all other PIK3CA-related mutation features remained 0. Similarly, if a frameshift deletion was present in TP53, the feature "TP53_Frame_Shift_Del" was set to 1. If multiple mutations of different types occur in the same gene, multiple Gene + Variant Classification indicators can be simultaneously active for that sample.

We stratified the dataset into training and test subsets using an 80/20 split, preserving the subtype class distribution (which is imbalanced: Luminal A and Luminal B were the most common, while HER2-enriched and Basal-like were least represented) (see Table 1). The protocol governing use of this hold-out test set is described further under Evaluation protocol.

Table 1 Sample distributions by class

Subtype	Count
Luminal A	662
Luminal B	449
HER2-enriched	214
Basal-like	198
Total	**1523**

Feature selection

To reduce dimensionality and focus the model on the most informative genomic features, we applied ElasticNet logistic regression exclusively on the training set to avoid information leakage [25]. ElasticNet, combining L1 and L2 penalties, was fitted separately for gene expression (EXP) and copy-number alteration (CNA) data to predict the four molecular subtypes. This procedure produced two independent ranked lists of features based on the absolute values of their coefficients, yielding 1691 non-zero features for EXP and approximately 4000 for CNA.

Rather than fixing a single number of features, we explored four feature-set sizes (K = 300, 500, 1000, 1500), ranging from more stringent to more inclusive selections (limited by the EXP modality). For each value of K (K = 300, 500, 1000, 1500), we selected the top-K features from the EXP ranking and, in parallel, the top-K features from the CNA ranking, using the same K for both modalities. Thus, each experiment used K expression features and K CNA features, while the mutation (MUT) modality was kept fixed.

This design ensured consistent and controlled dimensionality across EXP and CNA inputs while preserving modality-specific feature selection. Moreover, it avoids any forced feature alignment between modalities, allowing the cross-modal attention mechanism to learn interactions without assuming shared feature indices.

Model training was performed on the training subset using stratified 5-fold cross-validation. It was used for model optimization and early stopping. In each fold, a portion of the training data was held out as a validation subset to monitor convergence and select the best model checkpoint. The overall steps of the proposed work are presented in Figure 1.

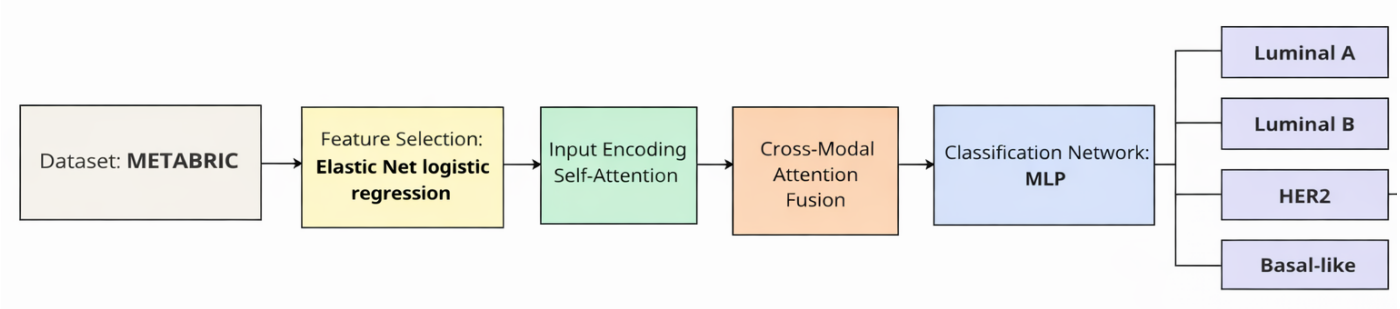

Figure 1 – Overall pipeline of the proposed method

Evaluation protocol

To prevent any information leakage, the external 20 % hold-out test set was fixed once at the start and was never used during feature ranking, pretraining, supervised training, hyperparameter tuning, or model-checkpoint selection. Within the 80 % training set we applied stratified 5-fold cross-validation: in each fold, four fifths of the training set were used for parameter updates and the remaining fifth served as an internal validation subset for early stopping and checkpoint selection. Five independent models (one per fold) were trained end-to-end, and each was then evaluated on the same, disjoint hold-out test set. The numbers reported in Table 4 are the mean (and, for macro-F1, the standard deviation) across these five independent test-set evaluations. No union of predictions, no re-fitting on the full training set, and no post-hoc thresholding on the test set were performed. Because the 5-fold split is defined strictly on the training 80 %, no internal validation fold can coincide with the external test set by construction. The best model checkpoint within each fold was selected based on minimum validation loss; no test-set information ever fed back into checkpoint selection, and hyperparameters were fixed a priori and kept identical across all Top-K experiments.

Our architecture consists of three key stages: (1) modality-specific self-attention encoding, (2) cross-modal transformer fusion that enriches EXP features using CNA and mutation signals, and (3) a residual MLP classifier that predicts the breast cancer subtype. The model supports ablation by selectively excluding modalities from the fusion step.

Input Encoding and Self-Attention

Each omics modality is treated as a sequence of scalar values one per gene or feature. Given a batch of N samples, the model processes the following inputs:

– Gene expression: $EXPraw \in R^{(N \times G_{exp} \times 1)}$

– Copy number alteration: $CNAraw \in R^{(N \times G_{cna} \times 1)}$

– Mutation features: $MUTraw \in R^{(N \times 885 \times 1)}$

Each tensor is first projected to a shared latent space using a linear layer ($1 \rightarrow 128$), then passed through a modality-specific Transformer encoder with self-attention:

– $EXPembed = SelfAtt(Linear_{1 \rightarrow 128}(EXPraw)) \in R^{(N \times G_{exp} \times 128)}$

– $CNAembed = SelfAtt(Linear_{1 \rightarrow 128}(CNAraw)) \in R^{(N \times G_{na} \times 128)}$

– $MUTembed = SelfAtt(Linear_{1 \rightarrow 128}(MUTraw)) \in R^{(N \times 885 \times 128)}$

The self-attended embeddings capture within-modality dependencies and are forwarded to the cross-modal attention blocks.

Self-attention block

Each modality uses a single standard Transformer encoder block with pre-norm residual structure. The block has four attention heads sharing a model width of d_model = 128 (per-head dimension 32), followed by a position-wise feed-forward network with hidden size 4•d_model = 512 and a GLU-style gated activation. Dropout of 0.1 is applied to attention weights, to the feed-forward output, and within the gating, together with a small stochastic-depth probability (DropPath, 0.05) on each residual branch. Residual branches are scaled by learnable LayerScale parameters initialised to 10-4. No positional encoding is applied. The feature order is fixed once per modality by the ElasticNet importance ranking (Feature selection) and is identical across all samples, so a gene's identity is carried by its position in the input tensor; adding sinusoidal or learned positional encodings is therefore neither necessary nor desirable and would introduce a spurious ordinal signal. The cross-modal attention block shares the same block structure, replacing self-attention with multi-head cross-attention in which the expression sequence is the query, and CNA or mutation sequences serve as key and value.

Cross-Modal Attention Fusion

To enrich the expression modality with auxiliary signals, we apply cross-attention using multi-head attention blocks. Gene expression embeddings serve as the query, and either CNA or mutation features act as the key and value:

$EXP_{(}CNA\ enriched_{)} = CrossAttn(query = EXPembed, context = CNAembed)$

$EXP_{(}MUT\ enriched_{)} = CrossAttn(query = EXPembed, context = MUTembed)$

We then compute the enriched expression tensor via residual addition:

$EXPenriched\_final = EXPembed + \alpha\ EXP_{(}CNA\ enriched_{)} + \beta\ EXP_{(}MUT\ enriched_{)}$

Where α and β are learnable scalars obtained via sigmoid(alpha) and sigmoid(beta) respectively. This mechanism allows the model to modulate the contribution of each auxiliary modality and refine the expression embeddings using cross-modal cues without overwriting original signals.

These gates are not driven to zero by the Stage 1 reconstruction term, for two complementary reasons. First, the reconstruction objective is optimized jointly with the supervised classification loss; trivially setting α = β = 0 would yield zero reconstruction but leave the classifier no better than an EXP-only baseline, and the joint objective therefore produces non-zero α and β whenever CNA or MUT information improves discrimination. Second, per-sample modality dropout applied during training randomly zeroes the CNA or MUT context with positive probability, so a solution that relies solely on EXP cannot attain a low joint loss across the training distribution. The

reconstruction term therefore acts as a stabilizing anchor against major drift of the enriched representation, not as an instruction to ignore auxiliary modalities.

Classifier Network

For classification, the enriched expression tensor of shape (N, G, 128) is flattened to a fixed-size vector of length G•128 (e.g., 128 000 features for K = 1000) and passed through a compact residual MLP. The MLP has a single hidden width H = 128, uses batch normalization after each linear layer, ReLU activations, and a skip connection that adds the output of the first block to the output of a bottleneck-and-expand sub-block:

– Linear: G 128 → 128 → BatchNorm → ReLU → Dropout (0.2) – produces the residual branch

– Linear: 128 → 64 → BatchNorm → ReLU → Dropout (0.1)

– Linear: 64 → 128→ Add residual branch → ReLU

– Output layer: 128 → 4 (corresponding to four molecular subtypes)

Ablation Variants

To investigate the contribution of each modality, we performed ablation studies using the following variants:

1. EXP – self-attention, no fusion.
2. EXP enriched by CNA only.
3. EXP enriched by both CNA and MUT.

The diagrams in Figures 2-4 illustrate the core model variants in our study.

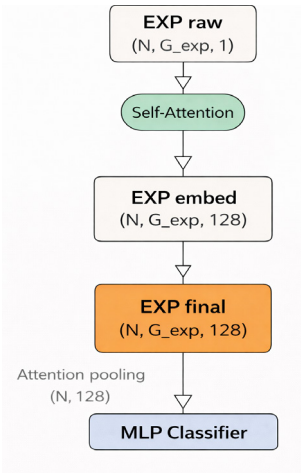

Figure 2 – EXP pipeline

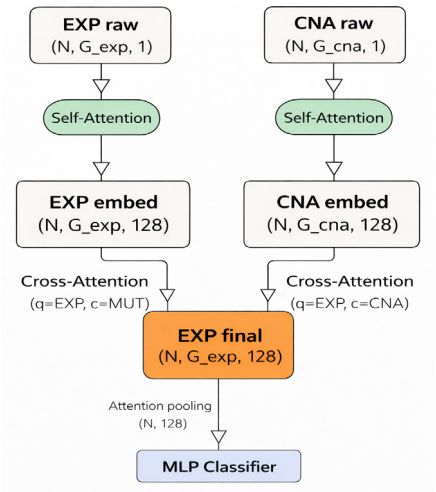

Figure 3 – Cross-modal attention from CNA

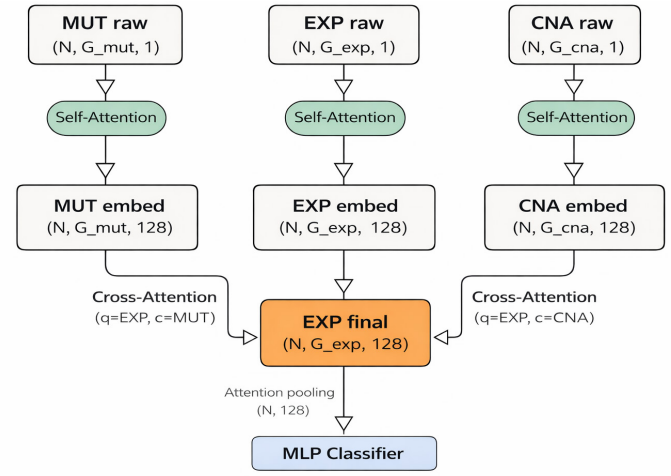

Figure 4 – Cross-modal attention from CNA and MUT

Training Procedure

Our training pipeline consisted of three sequential steps designed to learn within- and cross-modality representations before supervised classification.

Stage 1: Self and Cross-Modal Attention

Before classification, each input modality (EXP, CNA, MUT) was independently encoded using a self-attention Transformer. The resulting self-attended embeddings were then used to enrich the expression representation through cross-modal attention: the expression embedding served as the query, while CNA and mutation embeddings acted as context.

Let E denote the self-attended expression embedding, and AC, AM be the outputs of the CNA→EXP and MUT→EXP cross-attention blocks, respectively. The enriched expression tensor was computed as:

$$E' = E + \alpha \bullet AC + \beta \bullet AM$$

where α and β are learnable gating scalars obtained via sigmoid activation.

To guide this fusion, the model minimized a reconstruction loss between E′ and E:

$$L_{recon} = \| E' - E \|^2$$

Stage 1 uses a masked-feature reconstruction objective combined with an auxiliary classification head, the two losses balanced by a schedule that gradually shifts emphasis from reconstruction toward classification. This forces the encoder to recover masked gene positions from within-modality context while producing embeddings that are already discriminative for the four subtypes. Full numerical hyperparameters are given in Supplementary Table S1.

Stage 2: Cross-modal enrichment training

The cross-modal block is trained on the outputs of Stage 1 under the same masked-reconstruction plus auxiliary-classification recipe, with λ_rec fixed at 0.10 and the classification weight warmed up over the first eight epochs. Two scheduled regularisers are central to the behavior of the learned gates σ(α), σ(β): per-sample modality dropout zeroes the CNA or MUT context with positive probability during training, so a solution that relies solely on EXP cannot attain a low joint loss and the gates cannot collapse to zero; and a scheduled mask ratio on the expression query (5% → 12%) keeps the reconstruction anchor active throughout training. Full numerical hyperparameters are given in Supplementary Table S1.

Stage 3: Supervised Classification

Using the enriched embeddings E′, the MLP classifier was trained to predict molecular subtypes.

The flattened enriched tensor is standardized per fold using mean and standard deviation computed on the training indices of that fold only, so that no validation or test statistic ever leaks into the fit. The classifier input is regularized with input dropout and MixUp, and optimization uses AdamW with a ReduceLROnPlateau schedule and early stopping on validation loss. Class imbalance is addressed with a focal loss weighted by inverse class frequency:

$$L_{focal} = -(1/N) \Sigma \alpha y_i (1 - py_i)^{\gamma} \log(py_i),$$

where py_i is the predicted softmax probability for the true class y_i, αy_i is the inverse-frequency class weight, and γ is a focusing parameter whose value is given with all other Stage 3 hyperparameters in Supplementary Table S1. The combination of focal loss, class weighting, and MixUp penalises confident misclassifications while preserving gradient signal on minority subtypes.

All experiments were implemented in PyTorch using stratified 5-fold cross-validation on the training subset. Within each fold, 20% of the training data served as a validation subset for early stopping and model selection. The full procedure was repeated across different expression-feature sizes (K = 300, 500, 1000, 1500) and modality configurations (EXP-only, EXP + CNA, EXP + CNA + MUT).

The best model checkpoint was selected based on minimum validation loss and then evaluated on the fixed external hold-out test set.

Evaluation Metrics

We evaluated model performance using metrics that address class imbalance, which is critical in subtype classification. The macro-F1 score served as the primary criterion, emphasizing balanced performance across all subtypes, including minority classes. To complement this, weighted F1 reflected the overall performance adjusted for class frequencies, while accuracy was reported only as a secondary reference due to its bias toward dominant subtypes. We also computed ROC AUC as a threshold-independent measure of separability, and tracked precision, recall, and confusion matrices to analyze subtype-specific behavior. Macro-F1 and ROC AUC are highlighted as key indicators.

Statistical analysis

To quantify the uncertainty of the reported metrics and to evaluate whether observed differences between configurations are meaningful, we applied paired two-sided statistical tests across the five fold-specific models. All configurations — different modality setups at a given K, or different K values within a modality — share the same training and test samples fold-for-fold, so paired tests on the five per-fold macro-F1 values are the appropriate design. Two complementary tests were used for each comparison: the paired Student's t-test, and the non-parametric Wilcoxon signed-rank test (with n = 5 paired observations the Wilcoxon test has a minimum attainable two-sided p-value of 0.0625, which we note explicitly at borderline results). For the cross-K analysis we additionally ran a Friedman omnibus test across the four K-levels within each modality, and a pooled analysis across modalities (n = 15 paired observations per comparison, obtained by stacking the five folds of the three modality configurations). The Holm–Bonferroni correction was applied to the family of three pooled comparisons of K = 1000 against the other K-levels. Raw two-sided p-values are reported, with Holm-adjusted p-values where applicable; $p < 0.05$ is considered statistically significant and $0.05 \leq p < 0.10$ as suggestive. Effect sizes are reported as mean paired differences Δ in macro-F1 together with per-fold standard deviations. All tests were implemented with SciPy in Python 3.

Results

Tables 2 and 3 summarize the top-20 expression (EXP) and top-20 copy-number (CNA) features ranked by the absolute

Table 2 Top-20 ElasticNet-ranked features for gene expression (EXP) modality. Features are ordered by the absolute value of their coefficients

Rank	Gene	Coefficient
1	KRT17	0,04600026
2	KRT14	0,039189994
3	SLC39A6	0,038983226
4	COL17A1	0,03479651
5	SFRP1	0,03312954
6	KRT6B	0,03296845
7	FGFR4	0,03000498
8	ELP2	0,02835113
9	TYMSOS	0,025144786
10	CLDN11	0,023658585
11	KRT5	0,022679463
12	PBK	0,02123858
13	LDLRAD4	0,020831855
14	ESR1	0,020572215
15	NKX2-1	0,01995619
16	CALML3	0,019574217
17	FADS2	0,019254878
18	TUBB1	0,019099653
19	AXIN2	0,019011276
20	ERBB2	0,018967364

Table 3 Top-20 ElasticNet-ranked features for copy number alterations (CNA) modality. Features are ordered by the absolute value of their coefficients

Rank	Gene	Coefficient
1	PPP1R1B	0,012399
2	TCAP	0,012062
3	PNMT	0,012062
4	GRB7	0,011504
5	STARD3	0,011289
6	PGAP3	0,011123
7	ERBB2	0,011042
8	MIEN1	0,011042
9	MIR4728	0,011042
10	CTNNA2	0,010845
11	NEUROD2	0,00982
12	SH3RF3	0,009491
13	SHANK2-AS1	0,009424
14	MED1	0,009093
15	CDK12	0,008316
16	DEFB108B	0,008201
17	FBXL20	0,007518
18	ZNF92	0,007361
19	MUC16	0,007319
20	STRN3	0,007299

Table 4 Classification performance on METABRIC breast-cancer subtypes for different feature-set sizes and modality combinations. Bold values indicate the best results within each Top-K group.

Features (Top-K)	Modality	Acc	Pre	Rec	F1 (weighted)	Macro Pre	Macro Rec	Macro F1	ROC AUC
300	EXP	0.8472	0.8582	0.8472	0.8501	0.8336	0.8534	0.8402	0.9663
300	EXP+CNA	0.8400	0.8501	0.8400	0.8428	0.8233	0.8439	0.8306	0.9642
300	EXP+CNA+MUT	0.8623	0.8705	0.8623	0.8645	0.8488	0.8669	0.8551	0.9684
500	EXP	0.8518	0.8596	0.8518	0.8538	0.8360	0.8595	0.8451	0.9633
500	EXP+CNA	0.8518	0.8599	0.8518	0.8537	0.8370	0.8622	0.8469	0.9664
500	EXP+CNA+MUT	0.8557	0.8637	0.8557	0.8578	0.8461	0.8628	0.8519	0.9674
1000	EXP	0.8682	0.8727	0.8682	0.8694	0.8535	0.8654	0.8580	0.9680
1000	EXP+CNA	0.8656	0.8727	0.8656	0.8677	0.8548	0.8671	0.8589	0.9691
1000	EXP+CNA+MUT	0.8577	0.8619	0.8577	0.8590	0.8477	0.8551	0.8502	0.9669
1500	EXP	0.8511	0.8550	0.8511	0.8524	0.8443	0.8485	0.8453	0.9649
1500	EXP+CNA	0.8584	0.8631	0.8584	0.8598	0.8516	0.8610	0.8549	0.9682
1500	EXP+CNA+MUT	0.8544	0.8586	0.8544	0.8558	0.8495	0.8533	0.8503	0.9665

value of their ElasticNet coefficients, together with their ranks and coefficient values. These genes correspond to the highest-ranked entries in the modality-specific ElasticNet rankings used to construct all Top-K feature sets.

Further, we examined how the number of selected features (K) and the integration of omics modalities affected classification performance. For each configuration, the selected model checkpoint was applied to the independent hold-out test set (20% of the data), and all metrics reported in Table 4 were computed on this external test set. Experiments are grouped by feature-set size (Top-K) and evaluated under three modality setups:

- Expression (EXP) only;
- Expression + Copy Number Alterations (CNA);
- Expression + CNA + Mutation (MUT).

The experimental results indicate that both feature-set size (K) and the integration of multiple omics modalities significantly influence classification performance. As shown in Table 4, performance improved as the number of selected genes increased up to K = 1000, with macro-F1 scores rising from 0.84 to 0.86. Beyond this point (K = 1500), gains plateaued or slightly declined, with metrics stabilizing at around 0.85. Multi-omics integration consistently outperformed the single-modality (EXP-only) approach. Specifically, the tri-modal model (EXP+CNA+MUT) achieved the highest macro-F1 for smaller feature sets (K = 300–500). However, for larger feature sets ($K \geq 1000$), the bi-modal (EXP+CNA) configuration demonstrated the best overall performance. Across all configurations, ROC AUC remained high (~0.96–0.97), indicating robust class separability.

We tested whether the numerical differences observed in Table 4 are statistically meaningful. Across the three modality configurations at any K, paired differences in per-fold macro-F1 were consistent in direction but did not reach the $\alpha = 0.05$ threshold (smallest $p = 0.091$ for EXP+CNA vs. EXP+CNA+MUT at K = 1000).

In contrast, the effect of feature-set size was statistically demonstrable. Within EXP+CNA, macro-F1 increased significantly from K = 300 to K = 500 (paired t-test $p = 0.012$) and from K = 300 to K = 1000 ($p = 0.036$); the corresponding Wilcoxon signed-rank p-values (0.063 in both cases) hit the minimum attainable two-sided value for $n = 5$ paired observations and should therefore be read as suggestive rather than as failing to confirm the t-test result.

The Friedman omnibus test over the four K-levels within EXP+CNA returned $\chi^2 = 7.08$, $p = 0.069$, and the EXP+CNA+MUT configuration was flat across K ($\chi^2 = 0.12$, $p = 0.99$). Going from K = 500 to K = 1000 produced a further numerical gain that did not reach significance in any configuration (pooled $\Delta = +0.008$, $p = 0.20$), and increasing K from 1000 to 1500 yielded no significant gain either, with all three configurations showing a flat or slightly negative trend (pooled $\Delta = -0.006$, $p = 0.25$); this absence of further gain is consistent with the plateau evident in Figure 5 and supports K = 1000 as the operating point at which the significant gains end and the plateau begins.

Pooling across modalities ($n = 15$ paired observations per comparison) confirmed the same picture at the cohort level: K = 1000 significantly outperformed K = 300 (paired t-test $p = 0.030$; Wilcoxon $p = 0.031$), with the Friedman test across the four K-levels giving $\chi^2 = 6.84$, $p = 0.077$. After Holm–Bonferroni correction for the three headline pooled comparisons of K = 1000 against K = 300, 500 and 1500, the adjusted p-values are 0.089, 0.41 and 0.25. The most salient pairwise comparisons are reported in Table 5. Figure 5 shows macro-F1 as a function of K for the three modality configurations. The tri-modal configuration (EXP + CNA + MUT) saturates at K = 300 and remains statistically flat across K, while the bi-modal EXP + CNA configuration benefits from additional features and peaks at K = 1000.

A detailed analysis of the subtype-level performance for the top-performing model (K=1000, EXP+CNA) is presented in Figure 6 and Table 6. The use of Macro Average metrics ensures that the model's performance is evaluated equally across all subtypes, regardless of their prevalence in the dataset. The model demonstrated the highest discriminative power for Basal-like tumors, achieving an F1-score of 0.93. Similarly, Luminal A samples were classified with high precision (0.93) and recall (0.87). In contrast, the HER2-enriched subtype proved to be the most challenging, with a precision of 0.70. As shown in the confusion matrix, a significant portion of HER2-enriched misclassifications involved Luminal B and Luminal A, with 22 and 10 cases respectively. Conversely, Luminal B was most frequently confused with Luminal A (36 cases) and HER2-enriched (32 cases), highlighting the molecular overlap between these categories.

Table 5 Key paired statistical tests on per-fold macro-F1 (two-sided). Δ is the mean paired difference. n = 5 folds per comparison within a single modality; n = 15 for the pooled rows

Family	Comparison	Δ	p (t-test)	p (Wilcoxon)
Modality (K=300)	EXP+CNA+MUT vs EXP+CNA	+0.0246	0.096	0.125
Modality (K=1000)	EXP+CNA+MUT vs EXP+CNA	−0.0087	0.091	0.313
Cross-K (EXP+CNA)	K=500 vs K=300	+0.0163	0.012	0.063
Cross-K (EXP+CNA)	K=1000 vs K=300	+0.0284	0.036	0.063
Cross-K (EXP+CNA+MUT)	K=1000 vs K=300	−0.0050	0.473	0.813
Pooled (n=15)	K=1000 vs K=300	+0.0137	0.030	0.031
Pooled (n=15)	K=1000 vs K=500	+0.0077	0.204	0.244

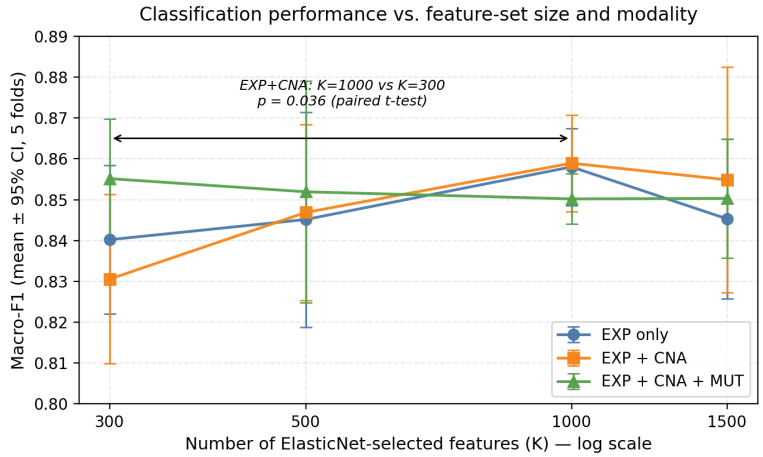

Figure 5 – Macro-F1 vs the number of features (K) for the three modality configurations

Each point is the mean across five fold-specific models; error bars show the 95% confidence interval of the mean. The bracket highlights the only within-modality comparison that reached significance at α = 0.05 (EXP+CNA, K = 300 vs K = 1000; paired t-test, p = 0.036).

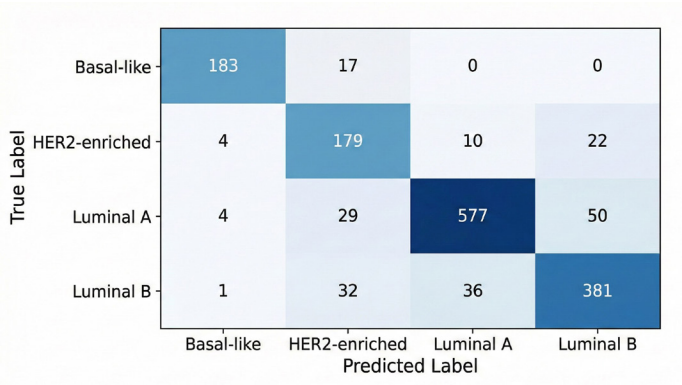

Figure 6 – Cross-modal attention from CNA

Table 6 Cumulative subtype-specific classification metrics (N=1525) Combined results for the K=1000 EXP_CNA model across five checkpoints (5 x 305 test predictions)

Subtype	Pre	Rec	F1	Support
Basal-like	0.9536	0.9150	0.9338	200
HER2-enriched	0.6978	0.8326	0.7587	215
Luminal A	0.9262	0.8742	0.8994	660
Luminal B	0.8417	0.8467	0.8438	450
Macro-Average	**0.8548**	**0.8671**	**0.8589**	**1525**

Discussion

ElasticNet-selected features demonstrate strong biological coherence with established molecular programs. The prioritization of KRT17, KRT14, and KRT5 underscores the unique epithelial differentiation of the Basal-like subtype, reflecting aggressive phenotypes [12]. Similarly, the ranking of ERBB2, GRB7, and STARD3 captures the focal 17q12 amplification characteristic of HER2-enriched tumors [12, 35]. The inclusion of ESR1 and FOXA1 further validates the model's ability to identify the regulatory backbone of luminal differentiation [35]. This alignment confirms that the cross-modal transformer is not merely identifying statistical noise but is operating on a biologically grounded hierarchy of breast cancer drivers.

Our findings highlight how feature-set size and multi-omics integration jointly influence the classification of molecular subtypes. Performance improved as the number of selected genes increased to approximately K = 1000, beyond which gains plateaued or slightly declined. This suggests that the top 1000 genes capture most of the biologically informative variance, while larger sets introduce redundancy or low-rank noise that reduces model efficiency.

The statistical analysis refines this qualitative picture. Pairwise differences between modality configurations at any K did not reach significance on five folds, so claims of strict superiority of one modality configuration over another should be interpreted as directional rather than established at this cohort size. In contrast, the effect of feature-set size is statistically supported: within EXP+CNA, K = 500 and K = 1000 both outperform K = 300 significantly (p = 0.012 and

0.036) and pooled across modalities K = 1000 outperform K = 300 (p = 0.030). The biologically meaningful lever at this scale is therefore the amount of information available to the cross-modal module, not which auxiliary modality is added.

The behavior of auxiliary modalities across different values of K provides deeper insight into the genomic architecture of breast cancer. At small feature sets (K = 300-500), mutation data provided the most significant performance boost, acting as a critical genetic anchor. When transcriptional coverage is limited, discrete somatic events in driver genes such as the high prevalence of TP53 mutations in Basal-like tumors or PIK3CA mutations in Luminal subtypes provide stable diagnostic cues that compensate for a sparse transcriptomic landscape.

Conversely, the dominance of the bi-modal EXP+CNA setup at K≥1000 highlights breast cancer as a primarily copy-number driven disease. In larger feature sets, the model effectively captures the downstream transcriptional effects of large-scale genomic events, such as the 17q12 (ERBB2) amplification. Because copy-number alterations affect entire gene clusters rather than single points, they provide a more robust and stable reinforcement of subtype identity compared to sparse mutational data, which may include non-functional passenger variants that introduce variability at higher dimensions. These results support a biologically coherent interpretation where gene expression captures the primary phenotypic state, while CNA provides a stable genomic backbone that reinforces subtype identity.

The granular analysis of the confusion matrix reveals specific areas where molecular biology challenges discrete classification. The lower precision for the HER2-enriched subtype (~0.70) highlights a significant genetic gray zone involving Luminal B tumors. Clinically, these triple-positive cases often harbor the HER2 amplification but are simultaneously driven by a dominant Estrogen Receptor program. The fact that many HER2-enriched samples were predicted as Luminal B suggests that the luminal transcriptional program can occasionally mask the HER2-driven signal [12]. Furthermore, the confusion between Luminal A and Luminal B reflects a biological spectrum rather than a binary divide. The transition between these subtypes is largely defined by a gradient of proliferation markers, and our results indicate that while cross-modal attention improves separation, tumors with intermediate proliferation remain difficult to categorize using standard thresholds.

Overall, our cross-modal transformer model achieved robust and competitive performance on the METABRIC dataset for PAM50-like breast cancer subtype classification, surpassing several previously reported multi-omics approaches. The comparative results are given in Table 7.

Limitations

Despite the promising performance of the cross-modal Transformer, several limitations constrain the interpretation of this work. Primarily, our evaluation was restricted to the METABRIC cohort; while this dataset provides a robust benchmark, external validation on independent cohorts such as TCGA-BRCA is essential to confirm the framework's generalizability across diverse populations. Furthermore, while the study is adequately powered (K=1000 vs. K=300) to detect significant architectural improvements, it remains underpowered for resolving sub-percent differences between specific modality configurations. Future iterations should utilize repeated cross-validation or bootstrap resampling to clarify these borderline comparisons. Additionally, the exclusion of Normal-like and Claudin-low subtypes due to class imbalance—reducing the task to a four-class problem—limits the current clinical scope. Technically, the reliance on Illumina HT-12 v3 microarrays necessitates a domain-adaptation step before the model can be applied to modern RNA-seq data. Finally, while our modality-agnostic design allows for the future integration of methylation and proteomic layers, the current lack of a formalized per-sample biomarker scoring system limits the immediate clinical interpretability of the produced attention maps.

Conclusion

This study demonstrates that integrating multi-omics data through a cross-modal transformer framework enables an accurate and biologically interpretable classification of breast cancer subtypes. By employing ElasticNet-based gene selection, we identified expression features aligned with canonical Basal, Luminal, and HER2 molecular programs, ensuring that the model's inputs are grounded in established oncological drivers. Our analysis revealed that a feature set of approximately 1000 genes maximizes predictive performance, beyond which the inclusion of lower-ranked features introduces redundancy that plateaus model accuracy.

A key finding of this work is the complementary role of different omics modalities. We propose that mutation data serves as a critical "genetic anchor" for classification when transcriptomic signals are limited, while copy-number alterations provide a stable genomic backbone that reinforces subtype identity in higher-dimensional spaces. The cross-modal attention mechanism effectively mimics the biological flow of information, weighting genomic alterations in the context of their transcriptional consequences.

The model's performance, characterized by a macro F1-score of 0.8589, proves robust across imbalanced classes. However, the persistent "gray zones" observed in the

Table 7 Comparison of Selected Methods for Breast Cancer Subtype Classification. Bold values indicate our results (Top-1000 EXP+CNA model)

Method	Modalities	Dataset/classes	Macro F1	Weighted F1/ Acc.
DeepMO [16]	EXP+CNA+Meth	TCGA (5 cls)	N/A	78.2% acc.
Islam et al. [17]	EXP+CNA	M-BRIC (5 cls)	N/A	79.2% acc.
moBRCA-net [22]	EXP+Meth+miRNA	TCGA (5 cls)	N/A	0.887 / 89.1%
MOGAT [33]	Multi-omics (8)	TCGA, M-BRIC (5 cls)	0.804	N/A
MoGCN [18]	EXP+CNV+RPPA+Clin	TCGA (4 cls)	~ 0.90	~0.90 / ~89%
Proposed	EXP+CNA	M-BRIC (4 cls)	0.8589	0.8677 / 0.8656

N/A = not reported.

misclassification of HER2-enriched and Luminal B tumors underscore the biological continuum of breast cancer, where luminal-driven transcriptional programs can partially mask HER2-driven signals. Overall, this pipeline offers a scalable and biologically grounded alternative to traditional fusion approaches, providing a foundation for more precise, data-driven diagnostics in precision oncology.

Supplementary materials

The Supplementary information includes tables:

- Supplementary Table 1. _____;
- Supplementary Table 2. _____.

This supplemental materials have been provided by the authors to give readers additional information about their work.

The file can be accessed using: https://www.editorialpark.com/download/article-supp/_____/Supplementary-data.docx.

Author Contributions: Conceptualization, K. Zh., T. Zh.; methodology / planning and organization, K. Zh.; validation, K. Zh., T. Zh., A. F.; formal analysis, T. Zh., A. F.; investigation, K. Zh., T. Zh.; clinical interpretation of the results, medical relevance of the study design, and validation the findings from a translational oncology perspective, T. K.; guidance on biological interpretation of omics features, validation of the relevance of selected molecular markers to breast cancer subtypes, L. S.; resources, B. A.; software, K. Zh.; data curation, K. Zh.; writing – original draft preparation, K. Zh.; writing – review and editing, K. Zh., T. Zh.; visualization, K. Zh.; supervision, B. A.; project administration, A. F., B. A.; funding acquisition, A. F., B. A. All authors have read and agreed to the published version of the manuscript.

Disclosures: The authors have no conflicts of interest.

Acknowledgments: None.

Funding: This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number BR24993145.

Data availability statement: The corresponding author can provide the data supporting the study's conclusions upon request. Due to ethical and privacy constraints, the data are not publicly accessible.

Artificial Intelligence (AI) Disclosure Statement: The authors declare no AI Tools used for preparation of this work.

References

1. Akhmedullin R, Aimyshev T, Zhakhina G, Yerdessov S, Beyembetova A, Ablayeva A, Biniyazova A, Seyil T, Abdukhakimova D, Segizbayeva A, Semenova Y, Gaipov A. In-depth analysis and trends of cancer mortality in Kazakhstan: a joinpoint analysis of nationwide healthcare data 2014–2022. BMC Cancer. 2024;24:1340. https://doi.org/10.1186/s12885-024-13128-2
2. Beyembetova A, Ablayeva A, Akhmedullin R, Abdukhakimova D, Biniyazova A, Gaipov A. National Electronic Oncology Registry in Kazakhstan: Patient's Journey. Epidemiol Health Data Insights. 2025;1(1):ehdi004. https://doi.org/10.63946/ehdi/16385
3. Midlenko A, Mussina K, Zhakhina G, Sakko Y, Rashidova G, Saktashev B, Adilbay D, Shatkovskaya O, Gaipov A. Prevalence, incidence, and mortality rates of breast cancer in Kazakhstan: data from the Unified National Electronic Health System, 2014–2019. Front Public Health. 2023;11:1132742. https://doi.org/10.3389/fpubh.2023.1132742
4. Chuvakova E, Zaripova L, Segizbayeva A, Baigenzhin A, Yegembay A, Idrissova D. Visualization of Breast Cancer and Safety: Review. J Clin Med Kaz. 2025;22(2):4–11. https://doi.org/10.23950/jcmk/16273
5. Iztleuov Y, Mutigulina G, Almagambetova A, Iztleuova G. Prognostic Role of Breast Architecture in Imaging, Histopathology, and Breast Cancer Outcome. J Clin Med Kaz. 2025; 22(5):73–79. https://doi.org/10.23950/jcmk/16879
6. Tombak Y, Umay EK, Unkazan FN, Karaahmet OZ, Sezer MK, Akyuz EU, Gurcay E. The Effect of Breast Cancer History on Bone Mineral Density in the Treatment of Postmenopausal Osteoporosis: One-Year Follow-Up Results. J Clin Med Kaz. 2024;21(6):85–90. https://doi.org/10.23950/jcmk/15703
7. Tlegenova Z, Balmagambetova S, Zholdin B, Kurmanalina G, Talipova I, Koyshybaev A, Nurmanova D, Sultanbekova G, Baspayeva M, Madinova S, Kubenova K, Urazova A. Stratifying breast cancer patients by baseline risk of cardiotoxic complications linked to chemotherapy. J Clin Med Kaz. 2023;20(3):75-81. https://doi.org/10.23950/jcmk/13325
8. Oladosu TA, Okafor CP, Nwosu PC, Ibukunoluwa AE, Monica UI, Aderanti TA. The Role of Liquid Biopsies in Tracking Tumor Evolution and Overcoming Therapeutic Resistance in Cancer. Oncol Nucl Med Transplantol. 2025;1(1):onmt006. https://doi.org/10.63946/onmt/17244
9. Beykikhoshk A, Quinn TP, Lee SC, Tran T, Venkatesh S. DeepTRIAGE: interpretable and individualised biomarker scores using attention mechanism for the classification of breast cancer sub-types. BMC Med Genomics. 2020;13(Suppl 3):20. https://doi.org/10.1186/s12920-020-0658-5
10. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lønning PE, Brown PO, Borresen-Dale A-L, Botstein D. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci. 2003;100(14):8418-8423. https://doi.org/10.1073/pnas.0932692100
11. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo JP, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160-1167. https://doi.org/10.1200/JCO.2008.18.1370
12. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, He X, Perou CM. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. 2010;12(5):R68. https://doi.org/10.1186/bcr2635
13. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, METABRIC Group, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale AL, Brenton JD, Tavaré S, Caldas C, Aparicio S. The genomic and

transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346-352. https://doi.org/10.1038/nature10983
14. Nasser M, Yusof UK. Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. Diagnostics. 2023;13(1):161. https://doi.org/10.3390/diagnostics13010161
15. Cristovao F, Cascianelli S, Canakoglu A, Carman M, Nanni L, Pinoli P. Investigating deep learning-based breast cancer subtyping using pan-cancer and multi-omic data. IEEE/ACM Trans Comput Biol Bioinform. 2022;19:121-134. https://doi.org/10.1109/TCBB.2020.3042309
16. Lin Y, Zhang W, Cao H, Li G, Du W. Classifying Breast Cancer Subtypes Using Deep Neural Networks Based on Multi-Omics Data. Genes. 2020;11(8):888. https://doi.org/10.3390/genes11080888
17. Islam MM, Huang S, Ajwad R, Chi C, Wang Y, Hu P. An integrative deep learning framework for classifying molecular subtypes of breast cancer. Comput Struct Biotechnol J. 2020;18:2185-2199. https://doi.org/10.1016/j.csbj.2020.08.005
18. Li X, Ma J, Leng L, Han M, Li M, He F, Zhu Y. MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. Front Genet. 2022;13:806842. https://doi.org/10.3389/fgene.2022.806842
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is All You Need. Adv Neural Inf Process Syst. 2017;30. Available at: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
20. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, Xie W, Rosen GL, Lengerich BJ, Israeli J, Lanchantin J, Woloszynek S, Carpenter AE, Shrikumar A, Xu J, Cofer EM, Lavender CA, Turaga SC, Alexandari AM, Lu Z, Harris DJ, DeCaprio D, Qi Y, Kundaje A, Peng Y, Wiley LK, Segler MHS, Boca SM, Swamidass SJ, Huang A, Gitter A, Greene CS. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141):20170387. https://doi.org/10.1098/rsif.2017.0387
21. Abdikenov B, Zhaksylyk T, Shortanbaiuly O, Orazayev Y, Makhanov N, Karibekov T, Suvorov V, Imasheva A, Zhumagozhayev K, Seitova A. Future of Breast Cancer Diagnosis: A Review of DL and ML Applications and Emerging Trends for Multimodal Data. IEEE Access. 2025;13:136101–136143. https://doi.org/10.1109/ACCESS.2025.3585377
22. Choi JM, Chae H. moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. BMC Bioinformatics. 2023;24(1):169. https://doi.org/10.1186/s12859-023-05273-5
23. Liu J, Su R, Zhang J, Wei L. Classification and gene selection of triple-negative cancer subtype using ensemble learning and mutual information-based selection. Brief Bioinform. 2021;22(5):1-12. https://doi.org/10.1093/bib/bbaa395
24. Guo J, Jin M, Chen Y, Liu J. An embedded gene selection method using knockoffs optimizing neural network. BMC Bioinformatics. 2020;21:414. https://doi.org/10.1186/s12859-020-03717-w
25. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005;67(2):301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
26. Molaei S, Cirillo S, Solimando G. Cancer Detection Using a New Hybrid Method Based on Pattern Recognition in MicroRNAs Combining Particle Swarm Optimization Algorithm and Artificial Neural Network. Big Data Cogn Comput. 2024;8(3):33. https://doi.org/10.3390/bdcc8030033
27. Anđelić N, Šegota SB. Development of symbolic expressions ensemble for breast cancer type classification using genetic programming symbolic classifier and decision tree classifier. Cancers. 2023;15(1):3411. https://doi.org/10.3390/cancers15133411
28. Bruno P, Calimeri F, Kitanidis AS, Momi E. Data reduction and data visualization for automatic diagnosis using gene expression and clinical data. Artif Intell Med. 2020;107:101884. https://doi.org/10.1016/j.artmed.2020.101884
29. Arafa A, El-Fishawy N, Badawy M, Radad M. RN-Autoencoder: Reduced Noise Autoencoder for classifying imbalanced cancer genomic data. J Biol Eng. 2023;17(1):7. https://doi.org/10.1186/s13036-022-00319-3
30. Mostavi M, Chiu Y-C, Huang Y, Chen Y. Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics. 2020;13(Suppl 5):44. https://doi.org/10.1186/s12920-020-0677-2
31. Mohamed T, Ezugwu A, Fonou-Dombeu JV, Ikotun AM, Mohammed M. A bio-inspired convolution neural network architecture for automatic breast cancer detection and classification using RNA-Seq gene expression data. Sci Rep. 2023;13:14644. https://doi.org/10.1038/s41598-023-41731-z
32. Arya N, Saha S. Multi-modal classification for human breast cancer prognosis prediction: Proposal of deep-learning based stacked ensemble model. IEEE/ACM Trans Comput Biol Bioinform. 2022;19:1032-1041. https://doi.org/10.1109/TCBB.2020.3018467
33. Tanvir R, Islam M, Sobhan M, Luo D, Mondal AM. MOGAT: A Multi-Omics Integration Framework Using Graph Attention Networks for Cancer Subtype Prediction. Int J Mol Sci. 2024;25:2788. https://doi.org/10.3390/ijms25052788
34. cBioPortal for Cancer Genomics. METABRIC Breast Cancer Study – cBioPortal. 2024. Available at: https://www.cbioportal.org/study?id=brca_metabric
35. Zhang MH, Man HT, Zhao XD, Dong N, Ma SL. Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials. Biomed Rep. 2014;2(1):41-52. https://doi.org/10.3892/br.2013.187