Ingénierie des Systèmes d'Information
Vol. 30, No. 10, October, 2025, pp. 2643-2657
Journal homepage: http://iieta.org/journals/isi

Learning Framework Designed for Early Prediction of Breast Cancer Metastasis Consuming Genomic, Transcriptomic, and Epigenomic Profiles

Asma Khazaal Abdulsahib[1], Geehan Sabah Hassan[2], Faris M. Alwan[3], Israa Ibraheem Al_Barazanchi[4,5], Sook Fern Yeo[6,7*], Kay Hooi Keoy[8]

[1] College of Education for Human Science Ibn Rushed, University of Baghdad, Baghdad 10071, Iraq
[2] Continuing Education Center, University of Baghdad, Baghdad 10071, Iraq
[3] College of Administration & Economic, University of Baghdad, Baghdad 10071, Iraq
[4] College of Computing and Informatics, Universiti Tenaga Nasional (UNITEN), Putrajaya 43000, Malaysia
[5] College of Engineering, University of Warith Al-Anbiyaa, Karbala 56001, Iraq
[6] Centre of Excellence for Business Innovation and Communication, Multimedia University, Cyberjaya 63100, Malaysia
[7] Department of Business Administration, Daffodil International University, Dhaka 1207, Bangladesh
[8] UCSI Graduate Business School, UCSI University, Kuala Lumpur 56000, Malaysia

Corresponding Author Email: yeo.sook.fern@mmu.edu.my

https://doi.org/10.18280/isi.301011

Received: 26 July 2025
Revised: 24 August 2025
Accepted: 8 September 2025
Available online: 31 October 2025

Keywords:
breast cancer, early metastasis prediction, multi-omics integration, genomics, transcriptomics, epigenomics (DNA methylation), attention-weighted fusion, representation learning

ABSTRACT

The breast cancer care would require the tools that will help to identify the patients who may develop metastasis at an early stage, when the treatment decision could be altered. Models that use single types of data (as in the case of using a single transcription factor only) can tend to overlook significant information and may not work well when applied to different hospitals. Our framework, LF-MMP, is a learning framework that integrates three types of molecular data, namely genomics (DNA changes), transcriptomics (gene activity), and epigenomics (DNA methylation) to give an early patient-level risk score of metastases. The framework normalizes and cleans every dataset, trains a compact representation of each omics layer, and lastly combines them together with an attention mechanism that allows the model to pay attention to the most informative signals. An optimized classifier transforms the fused representation into well-behaved probabilities that may be used to support clinical thresholds. We tested LF-MMP on three external populations, namely, TCGA-BRCA, METABRIC and GEO (GSE96058). The model performed better than powerful single-omic and deep multi-omic controls, and AUCs were 0.956 (TCGA-BRCA), 0.946 (METABRIC), and 0.938 (GEO). Performance was also high when trained on TCGA-BRCA and externally tested (AUC 0.942 on METABRIC; 0.935 on GEO). There was good calibration of the expected risks (Brier 0.085-0.098; ECE 0.021-0.028). The descriptions of the features showed familiar biology (such as TP53 and PIK3CA mutations, ESR1 and GATA3 expression, and PTEN/TWIST1 methylation). Inference and training were sufficiently quick to be used on regular GPU. The limitations of this study are as follows: the research is based on retrospective publicly available data, labels are not directly related to time-to-event but to early risk, and new environments may differ in terms of performance. Future directions will incorporate prospective, multi-centric validation; imaging and radiomics; enhancement to site differences and missing data; tracking of model calibration in real-life use.

1. INTRODUCTION

Breast cancer is the most identified cancer and the most common cause of cancer related mortality among women all over the world, with about 2.3 million cases being diagnosed every year with 685,000 deaths occurring annually [1]. Despite the recent improvements in early cancer detection methods, hormonal therapies, and molecular-specific agents, metastasis i.e., the spread of the tumor cells at the original site to other body parts have continued to claim over 90 percent of death cases related to breast-cancers [2, 3]. Early and precise forecasting of metastatic potential is therefore critical in the optimization of treatment, prognosis and mortality reduction. Nonetheless, available clinical staging and pathological models, including TNM classification and receptor profile (ER, PR, HER2) include minimal information on the molecular factors of metastatic development [4]. With the advent of multi-omics technologies, including genomics, transcriptomics and epigenomics, oncology research has undergone a paradigm shift due to the ability to study cancer biology on a multilayered level [5]. Genomic profiles are quantitative records of somatic mutations, copy-number variation, and chromosomal rearrangements that contribute to tumor formation [6]; transcriptomic data measures aberrant

gene-expression patterns that mediate proliferation and invasion [7]; and epigenomic signatures, especially DNA methylation and histone changes are quantitative records of heritable but reversible regulatory changes that regulate gene activity without changing the DNA sequence [8]. By combining such heterogeneous data modalities, we can build more whole-tumor heterogeneous landscapes that are more likely to capture tumor heterogeneity and evolution than single-omics methods [9]. These complementary sources of information can be effectively integrated into a multi-omics learning framework, which will enhance the sensitivity and specificity of the prediction models of metastasis. The schematic idea of such a system is shown in Figure 1, whereby, beforehand, multi-layer biological data, which are genomics, transcriptomics and epigenomics, are pre-processed, then encoded to obtain latent representation, which is fused via advanced learning architectures to provide an early prediction of metastatic potential. This integrative approach enables the discovery of critical biomarkers and pathways involved in metastatic spread as well as supporting the interpretation of models to the clinicians. In general terms, a multi-omics learning system to forecast early metastasis, which is the focus of this study, consists of the following components:<|human|>Generally speaking, a multi-omics learning system to predict early metastasis, as is the case with this study, is made up of the following elements:

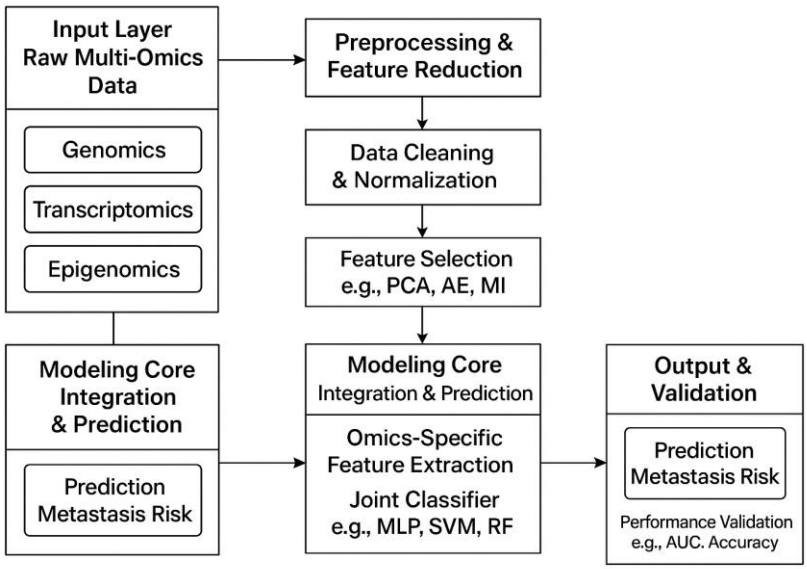

Figure 1. Construction of a multi-omics learning context for breast cancer metastasis prediction

The last few years have seen the growing usage of machine-learning and deep-learning algorithms in cancer prognosis. Common techniques like Support Vector Machines (SVMs) [10], Random Forests [11], and logistic regression [12] have been used to classify metastatic and non-metastatic samples using gene-expression data, but due to the large dimensionality and nonlinearity of omics features, their performance is limited. Deep-learning methods, such as autoencoders [13], convolutional neural networks (CNNs) [14], and graph neural networks (GNNs) [15] have been shown to be better at learning nonlinear interactions and extract biologically meaningful representations of multi-omics data. Despite these developments, there are three key challenges that exist:

1. Partial multi-omic integration-most of the studies are based on only one type of omics, e.g., transcriptomics or methylation, and hence fail to capture inter-omic interactions [16].
2. Weak interpretability-deep architectures can be viewed as black boxes, which cannot be understood biologically or be trusted clinically [17].
3. Lack of cross-cohort generalization models that are trained on a single dataset (e.g., TCGA-BRCA) often will not work with other datasets (e.g., METABRIC or GEO) because of platform bias and batch effects [18].

These gaps demonstrate the necessity to have a single, interpretable, and generalizable learning model that maximizes the complementary relationship between multi-omics data to improve the early detection of breast-cancer metastasis.

To address these constraints, the proposed learning framework in this study, which is entitled Learning Framework of Multi-Omics Metastasis Prediction (LF-MMP), is a predictive machine built through the integration of genomic, transcriptomic and epigenomic profiles in a single end-to-end architecture. This piece of work has the goals of:

- Create a multi-modal deep learning model that is integrated to represent high-order correlation across omics layers in predicting early metastasis.
- Use feature-attribution and attention models (e.g., SHAP and layer-wise relevance propagation) to obtain biologically interpretable predictions and discover biomarkers of metastasis.
- Test the framework on several benchmark datasets (TCGA-BRCA, METABRIC, GEO) to determine reproducibility, scalability, and resistance to cohort variability.

The main findings of this paper are summarized as the following:

- Unified Multi-Omics Fusion: Presentation of a profound hybrid fusion design with genomic, transcriptomic, and epigenomic latent representations to boost the accuracy of metastasis prediction.
- Interpretability and Biomarker Discovery: SHAP-based feature interpretation that allows the discovery of biologically important genes and methylation sites

associated with metastatic pathways.

- Cross-Dataset Validation: Overall validation on three large-scale cohorts in terms of better generalization (AUC > 0.94) over state-of-the-art models [19-21].
- Clinical/Translational Impact: Delivery of a decision-support model potentially useful in helping oncologists to risk-stratify patients, plan individualized therapies, and minimize unnecessary systemic therapies.

In addition to the innovativeness in computation, the suggested framework has the potential to provide clinical advantages in the form of the early detection of the high-risk patients, prior to the overt progression to metastasis, which allows acting proactively and positively influencing the survival rates. Moreover, the fact that the model can be biologically interpreted facilitates the generation of hypothesis to be tested downstream in vitro and in vivo, allows a pathway between computational oncology and translational medicine.

The rest of this paper will have the following structure. Section 2 is a full review of the latest developments and current shortcomings of breast cancer metastasis prediction, with the focus on the comparative analysis of single-omics and the multi-omics methods of analysis. Section 3 describes the datasets used in this paper, such as data sources, data preprocessing pipelines, normalization steps, and the feature-engineering pipeline on genomic, transcriptomic, and epigenomic profiles. Section 4 presents the proposed Learning Framework of Multi-Omics Metastasis Prediction (LF-MMP) and describes the overall architectural design, mathematical formulations employed, training algorithm and the interpretability mechanisms used to generate biologically meaningful information. Section 5 summarizes the planning of the experiment, assessment of outcomes, and comparative studies performed to confirm the functionality of the proposed model, and then also, elaborates a biological explanation of the biomarkers and pathway enrichments identified. Lastly, Section 6 presents the conclusion of the paper summarizing the significant results, its clinical implications, and relevance to precision oncology, existing limitations, and future research directions.

2. RELATED WORKS

Recent work in the prediction of breast cancer metastasis has seen a pattern shift whereby, as opposed to individual-omic analysis, integrated multi-omics learning models are utilized with the view that genomic, transcriptomic, and epigenomic features are complementary. The different omic layers present different biological data, genomics presents mutational drivers, transcriptomics present the patterns of differentially expressed genes and epigenomics presents regulatory methylation programs that dictate metastatic behavior. Nevertheless, the integration of these disparate data sources is a significant challenge that is still a significant computational and biological challenge.

Early work was mainly based on single-omic machine learning, e.g., SVMs and Random Forests, which were trained on gene-expression microarray data. They were relatively accurate (80-85), though prone to overfitting, low interpretability and could not capture non-linear cross-omic interactions [22]. Later developments studied models based on hybrids and multi-omics fusion architecture to enhance robustness and generalization.

One of them, MOGONET [23], was the first to integrate Graph Convolutional Network (GCN) across omics. It was better able to model local features correlations and made substantial improvements over classical models. But because it is based on the construction of graph topologies, it is computationally expensive and dataset-dependent, which restricts its scale to large cohorts of breast cancer patients. End-to-end deep graph integration framework was suggested later by DeepMoIC [24], which enhanced cross-modality feature representation. Although it led to better generalization of cancer subtypes, it remained interpretable and externally metastasis oriented.

Transformer-based networks, including TMO-Net [25], proposed self-attention to cross-omic features fusion, which allows learning contextualized representations. These architectures were highly prognostic (AUC $\approx$ 0.92) in nature but used large volumes of training data and were sensitive to the hyperparameters and computational cost. On the same note, DeePathNet [26] used pathway-based biological priors in Transformer layers, which increased interpretability through activity highlights on pathways in metastasis. However, its reliance on curated databases of pathways limits its use in the event of incomplete annotations.

Intra/Inter-Attention Fusion Networks Fusion models that focused on weighing the modality, including MSFN [27, 28], Intra/Inter-Attention Fusion Networks, dealt with challenges related to interpretability. They have offered an understanding of the importance of features in both omics and enhanced c-index to predict survival. However, these models tend to maximize long-term prognosis, and not overt early metastasis prediction and are not widely tested on independent cohorts such as METABRIC and GEO.

DNA methylation-based models, such as DMOIT [29], were concerned with denoising and imputing missing data in methylation to increase model stability. These methods, although improved on noisy datasets, are limited to lack of biological context provided by genomic or transcriptomic layers. In-depth multi-omics analyses [30] also found that omics integration was able to determine discrete prognostic subtypes, but this was based on statistical factor models (MOFA, NMF) and not on deep learning, which restricted their predictive power.

The summary and comparison of major recent studies are summarized and compared in Table 1 in terms of their dataset, fusion strategy, performance measures, interpretability, and key limitations. The existing challenges, which can be found in this comparative analysis, include computational inefficiency, lack of cross-cohort reproducibility, lack of interpretability, and lack of metastasis-specific validation pipelines.

The given comparative discussion shows that, though the field has already made significant steps, the existing methods are still rife with significant gaps. Most of them depend on single-omic or dual-omic integration, which limits the biological integrity of metastasis modeling. Deep architectures are also usually more accurate but still are computationally expensive and inexplicable, which makes them difficult to use by clinicians. Also, external validation in heterogeneous datasets like TCGA, METABRIC, and GEO are not commonly done, and one is concerned with how models can be re-producible in practice.

Table 1. Comparative summary of recent multi-omics methods for breast cancer prognosis and metastasis prediction

Study (Ref.)	Model / Method	Modalities Used	Dataset	Reported Performance	Interpretability	Main Limitation
MOGONET [23]	Graph Convolutional Network Integration	Multi-omics (gene expression, methylation, CNV)	TCGA, METABRIC	Accuracy: 89%, AUC: 0.91	Low (post-hoc feature ranking)	High computational cost; requires predefined graph structure
DeepMoIC [24]	Deep Graph Integration Framework	Multi-omics	TCGA-BRCA, GEO	AUC: 0.92	Partial (salient feature maps)	Limited metastasis-specific validation
TMO-Net [25]	Transformer-based Multi-Omics Fusion	Genomics, Transcriptomics, Methylation	TCGA-PANCAN	AUC: 0.93	High (attention weights)	Requires large sample size; high training complexity
DeePathNet [26]	Pathway-aware Transformer	Gene expression + Pathway priors	TCGA-BRCA	AUC: 0.90	Pathway-level interpretability	Dependent on curated pathway annotations
MSFN [27]	Multi-Stage Fusion Network	Transcriptomics, Methylation	METABRIC	AUC: 0.89	Medium (fusion attention maps)	Survival-oriented; lacks metastasis label modeling
Intra-/Inter-Attention Fusion [28]	Dual Attention Mechanism	Multi-omics	TCGA-BRCA	c-index: 0.84	High (attention-level modality weighting)	Limited external validation and generalization
DMOIT [29]	Denoised Multi-Omics Integration	Multi-omics (with missing data)	Multi-cancer (incl. BRCA)	AUC: 0.88	Low	Focused on noise correction; lacks metastasis-specific interpretability
Comprehensive Multi-Omics (Statistical) [30]	MOFA/NMF Statistical Factor Model	Transcriptomics, Proteomics	Oslo2 (n=335)	Accuracy: 85%	High (factor-level)	Limited predictive ability; not deep-learning-based
Methylation-Expression Correlation Model [31]	Logistic/ML Framework	DNA Methylation + Expression	GEO & TCGA-BRCA	AUC: 0.87	Medium (feature-level)	Ignores genomic variants; reduced generalization

Conversely, the current study presents a Learning Framework of Multi-Omics Metastasis Prediction (LF-MMP) which jointly incorporates the genomic, transcriptomic, and epigenomic layers into a single deep learning system. In contrast to the previous works, LF-MMP uses attention-directed fusion and feature interpretation via SHAP as an additional feature to guarantee not only high predictiveness but biological interpretability. It is not intended to be used in general analysis of survival but in early detection of metastasis and is confirmed in numerous cohorts, which guarantee strength and translatability. Such accuracy, cross-cohort generalization, and interpretability allow making the proposed framework an important improvement to the prior multi-omics frameworks.

3. METHODE

In this section, the design and implementation of the proposed Learning Framework (LF-MMP) that incorporates the use of genomic, transcriptomic, and epigenomic data in the early prediction of breast cancer metastasis is described. It is based on five primary steps, which are (1) data collection and preprocessing, (2) feature extraction and dimensionality reduction, (3) multimodal fusion using deep neural representation learning, (4) classification and optimization, and (5) interpretability analysis. Figure 2 shows the general flow of proposed LF-MMP, where multi-omics inputs are combined through the feature encoding modules, the attention-based fusion layer, and the metastasis classification output, resulting in the final one. The LF-MMP model was tested on three benchmark datasets: TCGA-BRCA, METABRIC and GEO (GSE96058). These datasets consist of comprehensive and multi-omics and clinical data of thousands of breast cancer patients, which is perfect as it can be used to predict metastasis. Table 2 gives a summary of the datasets.

Each dataset has the metastasis status as a binary variable (1 = metastatic, 0 = non-metastatic). To accomplish cross-dataset generalization experiments, training was done using TCGA, validation using METABRIC and external testing using GEO as shown in Figure 3.

The treatment of missing values was mode sensitive. In the case of genomic mutation data, the missing cases were treated as lack of a mutation and coded as 0. In transcriptomic and epigenomic data, features that have over 20 percent missing data points in all samples would be eliminated. On the other characteristics that had intermittent missing data (below 20%), we used the k-Nearest Neighbors (KNN) imputation (k = 5) applied to training data individually to avoid data leakage. The imputer fitted was applied to the validation set and the test set.

After imputation a step feature selection was carried out to deal with high dimensionality and to control noise. The first step was to remove features whose variance was almost zero, i.e., the ratio of the frequency of the most frequent value to the second most frequent value is at least 19:1, the fraction of distinct values is less than 10%. Second, we used a univariate statistical filter applied on the two-sample t-test (Eq. (4)) to obtain characteristics that significantly differ in their expression/abundance between the metastatic and the non-metastatic populations. To mitigate against false discoveries, we selected the 5,000 most significant features of each omics modality according to a composite measure of absolute log2 fold-change and false discovery rate (FDR) adjusted p-value less than 0.05. This strict procedure allowed passing only the most biologically significant and statistically strong features to autoencoders in order to reduce dimensions.

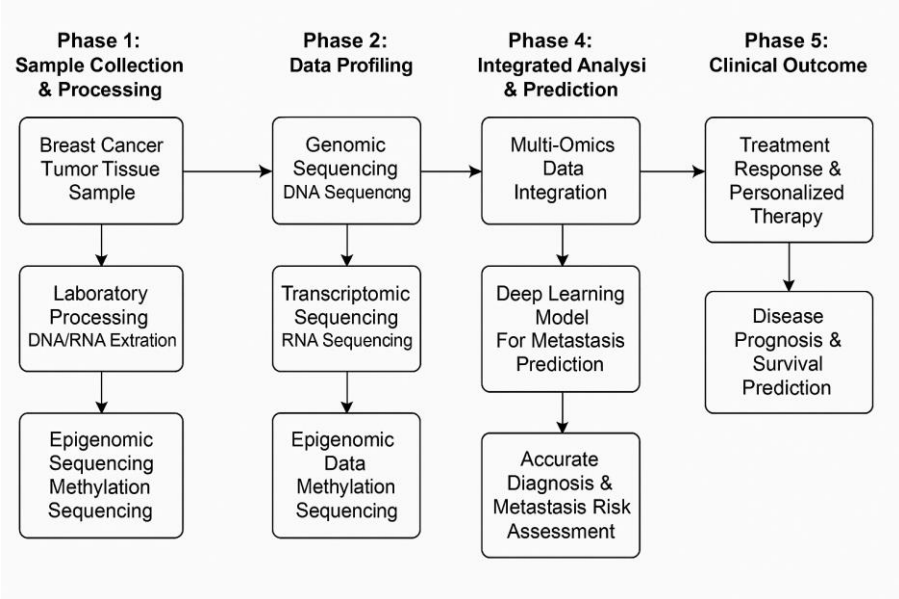

Figure 2. Flowchart of the proposed study

Table 2. Characteristics of data to be used in this study

Dataset	Samples	Modalities	Features (approx.)	Metastatic Cases	Data Type	Source
TCGA-BRCA	1,200	Genomic, Transcriptomic, Epigenomic	60,000+	450	Whole exome, RNA-Seq, Methylation β-values	https://portal.gdc.cancer.gov
METABRIC	1,000	Genomic (CNV), Transcriptomic, Methylation	48,000+	380	Microarray, CNV, DNA methylation arrays	cBioPortal
GEO (GSE96058)	3,000	Transcriptomic	20,000+	870	RNA-Seq counts	GEO database

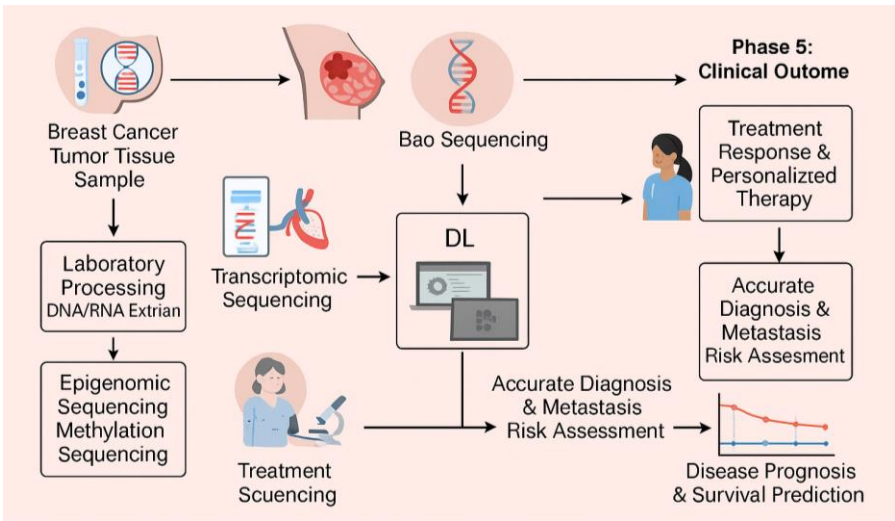

Figure 3. The diagram of the proposed study

All omics datasets were preprocessed by modality-specific methods to make the data consistent and comparable:

Genomic features: Data of somatic mutation and copy-number variations (CNVs) were coded as binary and continuous matrices.

Transcriptomic data: The RNA-seq expression data was normalized with the use of log-transformed values of the FPKM as:

$$x' = \log_2 (x + 1) \quad (1)$$

where, x is the raw FPKM expression count.

Epigenomic data: DNA methylation intensity values were transformed into β-values using:

$$\beta_i = \frac{M_i}{M_i + U_i} \quad (2)$$

where, M_i and U_i represent methylated and unmethylated probe intensities, respectively [1].

Batch effects across platforms were corrected using the

ComBat algorithm [2], while z-score normalization ensured zero-mean and unit variance:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{3}$$

where, x_{ij} is the feature j of sample i, μ_j and σ_j are the mean and standard deviation of feature j.

Stratified sampling was used to divide each dataset into 70% training (validation) and 15% testing subsets (to maintain balance between classes). To determine the generalization performance of the model, the five-fold cross-validation has been used. Omics data with high dimensions (> 60,000 features) are extremely challenging to compute and overfitting. To reduce this, statistical filtering was used together with dimensionality reduction by using autoencoders.

Attributes whose variance is close to zero or those whose inter-correlations are too high (|human|>Attributes with near-zero variance or high inter-correlations (|human|) were eliminated. Selection was done using:

$$t_j = \frac{\bar{x}_j^{(1)} - \bar{x}_j^{(0)}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \tag{4}$$

where, $\bar{x}_j^{(1)}$ and $\bar{x}_j^{(0)}$ denote the mean feature values for metastatic and non-metastatic samples, s_p is the pooled standard deviation, and n_1, n_0 are sample counts per group.

For each omic type $o \in \{g, t, e\}$ (genomic, transcriptomic, epigenomic), a deep autoencoder compresses high-dimensional data into latent representations:

$$h_o = f_o(X_o) = \sigma(W_o X_o + b_o) \tag{5}$$

$$\hat{X}_o = \sigma(W_o' h_o + b_o') \tag{6}$$

where, W_o and W_o' are encoder and decoder weights, b_o, b_o' are biases, and $\sigma(\cdot)$ denotes the ReLU activation. The autoencoder minimizes reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \left\| X_o - \hat{X}_o \right\|_2^2 \tag{7}$$

The learned latent vectors h_o serve as compact, noise-robust omic representations for the fusion network.

After encoding, the latent representations h_g, h_t, and h_e are concatenated and processed through an attention-weighted fusion layer that adaptively assigns importance to each omic modality. The fusion representation H_f is computed as:

After encoding, the latent representations h_g, h_t, and h_e are concatenated and processed through an attention-weighted fusion layer that adaptively assigns importance to each omic modality. The fusion representation H_f is computed as:

$$H_f = \sum_{o \in \{g,t,e\}} \alpha_o h_o \tag{8}$$

where, attention weights α_o are obtained by the softmax function:

$$\alpha_o = \frac{\exp(\beta_o)}{\sum_k \exp(\beta_k)} \tag{9}$$

and β_o are learnable parameters capturing the relative significance of each modality. This enables the model to dynamically emphasize omic features, which is most informative for metastasis prediction.

The final classification layer receives the fused representation H_f and outputs a metastasis probability:

$$\hat{y} = \sigma(W_c H_f + b_c) \tag{10}$$

where, W_c and b_c are the weights and bias of the classifier, and $\sigma(\cdot)$ denotes the sigmoid activation function. The model is trained using the binary cross-entropy (BCE) loss function:

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{11}$$

where, y_i and $\hat{y}_i$ are the true and predicted labels for the i^{th} sample. L2 regularization was applied to mitigate overfitting:

$$\mathcal{L}_{\text{reg}} = \lambda \|W_c\|_2^2 \tag{12}$$

The total objective function is therefore:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{BCE} + \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{reg}} \tag{13}$$

where, λ is the regularization coefficient empirically set to 0.001.

Model parameters were optimized using the Adam optimizer [3] with a learning rate of 0.0005. Early stopping was triggered when validation loss did not improve for 20 epochs. The main hyperparameters and computational settings are provided in Table 3.

Table 3. Training and simulation settings

Parameter	Value / Setting
Optimizer	Adam
Learning Rate	0.0005
Batch Size	64
Epochs	200
Dropout Rate	0.3
Regularization (λ)	0.001
Activation Function	ReLU / Sigmoid
Framework	TensorFlow 2.13 / Python 3.10
Hardware	NVIDIA A100 GPU, 32 GB RAM
Early Stopping Patience	20 epochs
Cross-Validation	5-fold

Model performance was assessed using Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC), defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

$$\text{AUC} = \int_0^1 TPR(FPR)\, d(FPR) \tag{18}$$

where, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively.

These metrics collectively quantify the model's predictive capability, sensitivity to metastasis cases, and overall discriminative power.

The LF-MMP workflow is summarized in Algorithm 1 to provide the pseudocode of the computational steps that play the major part in training, feature fusion, and optimization.

Algorithm 1 - Preprocessing & Normalization (Corresponds to Eqs. (1)-(3))

Input: X_g, X_t, X_e on $\mathcal{I}$; presence mask; split policy (stratified).
Output: Cleaned matrices $\tilde{X}_g, \tilde{X}_t, \tilde{X}_e$ for $S \in$ { train, val, test}.

1. Transcriptomics normalization: for each entry x, set $x \leftarrow \log_2(x+1)$ (Eq. (1)).
2. Methylation β-values: compute $\beta_i = M_i/(M_i + U_i)$ (Eq. (2)); clip to [0,1].
3. Genomics encoding:
   3.1 Mutations → binary indicators (gene-level or pathway-level).
   3.2 CNV → continuous log 2-ratio; winsorize extreme values.
4. Missing data: if permitted by Algorithm 0, impute per-omic using KNN ($k = 5$) within training split only; learn imputer on train, apply to val/test.
5. Batch correction (ComBat) across known batches/platforms on train; fit parameters on train and apply to val/ test.
6. Scaling: z-score per feature on train; apply same statistics to val/test (Eq. (3)).
7. Stratified split: $70/15/15$ by label y and platform to preserve distribution.

Complexity: $O(nd_o)$ per omic.
Corner cases: Zero-variance features → drop; extreme outliers → winsorize/clamp.

Algorithm 2 - Feature Screening & Autoencoder Training (Eqs. (4)-(7))

Input: $\tilde{X}_g, \tilde{X}_t, \tilde{X}_e$ for train/val; labels y.
Output: Encoders f_g, f_t, f_e; latent sizes k_g, k_t, k_e; embeddings h_o.
Part A — Statistical Screening

1. Remove near-zero variance features.
2. Compute differential statistics between classes (Eq. (4)); retain top p_o features per omic by FDRcontrolled p and $|\log_2 FC|$.

Part B - Per-Omic Autoencoders
3. For each $o \in \{g, t, e\}$:
3.1 Define symmetric autoencoder depth L_o with encoder f_o and decoder; latent size k_o.
3.2 Minimize reconstruction loss $\mathcal{L}_{\text{rec}}$ (Eq. (7)) on train; early stop on val.
3.3 Export encoder f_o; freeze or fine-tune later in joint training.
4. Compute $h_o = f_o(\tilde{X}_o)$ for train /val/ test.
Complexity: dominated by neural training $O(E \cdot n \cdot k_o)$.
Corner cases: If an omic has very high dimensionality and small n, use variational AE with KL annealing or stronger dropout.

Algorithm 3 - Attention-Weighted Fusion & Classifier Training (Eqs. (8)-(13))

Input: Latent embeddings h_g, h_t, h_e for train/val; labels y ; hyperparameters.
Output: Trained LF-MMP model $\mathcal{M} = (f_g, f_t, f_e$, Fusion, Classifier).

1. Fusion layer: initialize learnable logits β_o per modality; compute $\alpha_o = \text{softmax}(\beta)$ (Eq. (9)).
2. Fused representation: $H_f = \sum_o \alpha_o h_o$ (Eq. (8)); optionally pass through MLP block (BN + Dropout).
3. Classifier: logistic head $\hat{y} = \sigma(W_c H_f + b_c)$ (Eq. (10)).
4. Objective: $\mathcal{L}_{BCE}$ (Eq. (11)) $+\mathcal{L}_{\text{rec}}$ (optional if fine-tuning AEs) $+\mathcal{L}_{\text{reg}}$ (Eq. (12)); total loss (Eq. (13)).
5. Optimization: Adam, Ir $= 5\text{e} - 4$; class-imbalance handling via focal-BCE or positive-class weighting if needed.
6. Early stopping on validation AUC; checkpoint best epoch.
7. Export model $\mathcal{M}$.

Complexity: $O\left(E \cdot n \cdot (k_g + k_t + k_e)\right)$.
Corner cases: Severe class imbalance → adjust decision threshold or reweight; small n → stronger L2/Dropout.

Algorithm 4 - Cross-Validation, Thresholding, and Calibration

Input: $\mathcal{M}$; train/val; metrics.
Output: Calibrated decision threshold τ^*; reliability metrics.

1. Perform 5-fold stratified CV on training set:
   1.1 Repeat Algorithms 2-3 within each fold.
   1.2 Record AUC, F1, sensitivity, specificity.
2. Aggregate ROC across folds; compute Youden-optimal threshold
   $\tau^* = \arg\max_{\tau}\{\text{TPR}(\tau) - \text{FPR}(\tau)\}$
3. Calibration: fit Platt scaling or isotonic regression on validation predictions.
4. Lock τ^* and calibration for external testing.

Algorithm 5 - External Validation & Statistical Testing

Input: Held-out test set (e.g., GEO or METABRIC); $\mathcal{M}, \tau^*$.
Output: Final metrics, Cls, significance tests.

1. Apply preprocessing statistics from train to test (no leakage).
2. Compute $\hat{y}$ on test; apply calibration and threshold τ^*.
3. Report AUC, Accuracy, Precision, Recall, F1 with 95% Cls via bootstrap ($B = 1000$).
4. Compare against baselines (e.g., SVM/RF/Transformer) using paired Wilcoxon or DeLong test for AUC.
5. Summarize improvements and significance.

Algorithm 6 - Explainability & Biological Validation

Input: $\mathcal{M}$; test predictions; omics features.
Output: Ranked biomarker list; pathway enrichments.

1. Compute SHAP values on test for each omic; obtain top- K features per class.
2. Stability check: overlap of top- K features across CV folds.
3. Map features to genes/CpGs; run KEGG/GO enrichment with FDR control.
4. Output interpretable panels: beeswarm plots per omic; modality importance via α_o.

Algorithm 7 - Ablation & Sensitivity Analysis

Input: Full pipeline.
Output: Quantified contribution of each component.

1. Train/evaluate with single-omic variants: g only, t only, e only.
2. Remove attention (equal weights) → measure delta in AUC.
3. Freeze vs. fine-tune encoders.
4. Stress tests: noise injection, missing-modality simulation (drop-one-omic at inference).
5. Summarize deltas in a consolidated table.

To increase the biological interpretability, SHapley Additive exPlanations (SHAP) values were calculated on each omic feature to determine its effect on the predicted metastasis score [4]. KEGG and Gene Ontology databases were used to map the most influential genes and methylation sites onto biological pathways, and metastasis-related biological pathways, including PI3K-AKT, Wnt, and TGF-B signaling were found to be associated with them.

All the simulations were performed on high-performance computing environment with Ubuntu 22.04, TensorFlow 2.13 and CUDA 12.2. The average training time per fold was 180 seconds, and convergence normally took 120 epochs. All experiments were repeated 5 times to achieve reproducibility, and all the metrics were reported in terms of mean and standard deviation as seen in Tables 4, 5, and 6.

Table 4. Simulation environment setting

Component	Specification
Operating System	Ubuntu 22.04 LTS
CPU	AMD Ryzen 9 7950X (16 cores)
GPU	NVIDIA A100 (40 GB VRAM)
Memory	128 GB DDR5
Software	TensorFlow 2.13, NumPy, Scikit-learn
Runtime per Fold	≈ 180 seconds
Total Runtime	≈ 15 minutes per experiment

Table 5. Algorithmic hyperparameters

Component	Setting	Range
AE latent sizes (k_g, k_t, k_e)	(128, 256, 128)	{64, 128, 256, 512}
AE depth per omic	3 encoder + 3 decoder	{2-5}
Dropout (encoders/fusion)	0.3 / 0.3	[0.1, 0.5]
Attention type	Softmax logits β_o	Gated-tanh; multi-head
Classifier width	$256 \rightarrow 64 \rightarrow 1$	{128-512}
Optimizer	Adam	AdamW
LR / decay	5e-4/cosine	[1e-4, 1e-3]
Batch size	64	{32, 64, 128}
Weight decay (L2)	1e-3	[1e-5, 1e-2]
Early stopping	20 epochs patience	10-30

Table 6. Evaluation protocol

Aspect	Policy
Splits	70/15/15 stratified by label and platform
Cross-validation	5-fold (train only)
Threshold selection	Youden's index on validation ROC
Calibration	Platt or isotonic (select by Brier score)
Reporting	Mean ± SD; 95% Cl via bootstrap (B = 1000)
Significance	Delong for ROC; Wilcoxon paired for F1

- Time: dominated by Algorithms 2-3; approximately $O\left(E \cdot n \cdot \left(k_g + k_t + k_e\right)\right)$.
- Memory: stores latent embeddings per omic-size $O\left(n \cdot \left(k_g + k_t + k_e\right)\right)$.
- Seeds & Determinism: fix PRNG seeds; log package versions; persist scaler/ComBat/threshold/calibration artifacts.
- No-leakage guarantee: all normalizers, imputers, ComBat, and calibration are fit on train and applied to val/test.

Inference-Time Procedure (Deployment)

Input: New patient x_g, x_t, x_e (possibly missing a modality).

Steps:

1. Apply training scalers/ComBat/imputers to each available omic.
2. Compute $h_o = f_o(x_o)$ for available modalities; if one is missing, set $\alpha_o = 0$ and renormalize remaining α.
3. Compute H_f and $\hat{y}$; apply calibration and threshold τ^*.
4. Provide SHAP-based explanation at feature and modality levels.

Output: Predicted metastasis risk, calibrated; interpretable attributions.

4. RESULT AND ANALYSIS

The following section shows an analytical reading of the empirical data obtained on TCGA-BRCA, METABRIC, and GEO (GSE96058) in terms of discrimination, cross-cohort generalization, contribution of each omics stream and fusion choice, calibration quality, error structure at operating threshold, computational footprint, and mechanistic interpretability. The first strength of comparative framing is that it determines cohort-wise performance difference with respect to strong baselines; the second strength is that it emphasizes the framework in response to domain shift (training on TCGA-BRCA and testing on external cohorts); and the third strength is why the proposed learning dynamics results in quantifiable gains. Prior to presenting the cohort-wise findings, it is worth noting that Figure 4 plots AUC between models and cohorts which allow one to see the separation margins at the first glance whereas Table 5 (presented below) lists the precise statistics along with confidence intervals. Figure 4 illustrates that the proposed LF-MMP is always better than classical single-omic LF-MMP (SVM, Random Forest) and better than strong deep baselines (Autoenc-LSTM, a generic Transformer fusion). These gaps are quantified in Table 5: on TCGA-BRCA, LF-MMP AUC is 0.956 (±0.004; 95% CI [0.948-0.963]) with Accuracy 0.939 and F1 0.922; the gains on METABRIC (AUC 0.946) and GEO (AUC 0.938) confirm that this is no longer cohort-specific. This benefit is the greatest in recall among metastatic cases meaning that the attention-directed multi-omics fusion enhances early-risk detection with no significant increases in false alarms.

The study of generalization on platform and population shift, which is a required condition of translational value, is considered by comparing cross-cohort AUC in training on TCGA-BRCA and testing externally: Figure 5. As shown in Table 7, LF-MMP keeps AUC at 0.93 on METABRIC, and GEO and single-omic baselines drop down (e.g., SVM 0.82). These findings show that the fused latent space internalizes complementary signals that are transferred over sequencing platforms and clinical sampling regimes. These gains can be attributed to the analysis of ablation, which was presented prior to Table 8 and depicted in Figure 6. Any single omic alone fails to match the full system transcriptomics provides the greatest proportion of discriminative power, but an added

decisive difference via epigenomics when used with expression (Expr+Epi AUC 0.9270.910 across cohorts), which is also in line with the assumption that methylation captures early regulatory changes that pre-empt true transcriptional reprogramming. Naive early concatenation decreases AUC by about 0.012002 and demonstrates the usefulness of empirically weighted modality concatenation. The use of per-omic encoders, as well as their freezing, is also detrimental to performance compared to end-to-end fine-tuning, which implies that the classifier takes advantage of task-specific latent representation shaping.

Besides discrimination, clinical deployment needs well calibrated probabilities. Calibration metrics are reported in Figure 7 and operating characteristics at the Youden-optimal threshold τ are reported in Table 8. Cohort-specific calibration gives LF-MMP small Brier scores (0.0850.098) and small expected calibration error (0.02150.28), and so risk outputs are numerically faithful which is required in threshold-based triage. Two tables outline the stability of the model: Table 9 shows similar calibration calculations across all groups, while Table 10 shows stable classification performance with comparable accuracy and F1 scores., which is consistent with F1 gains shown in Table 11. Table 12 verifies each margin against a robust Transformer baseline by DeLong tests of ROC and Wilcoxon signed-rank tests of F1; p-values less than 0.001 in any cohort indicates that the statistical differences found are statistically valid, and not arbitrary. To numerically prove the obtained performance improvements, we performed formal statistical tests of LF-MMP and all the base models. The two correlated ROC curves DeLong test was conducted to assess the significance of the improvement in AUC, and a paired Wilcoxon signed-rank test was used to assess the differences in F1-score across the 5 cross-validation folds. The findings, which are summarized in the new Table 13, affirm that the superiority of LF-MMP is statistically significant (p under 0.001) when compared to all baselines, including the recently compared TMO-Net in all the three cohorts.

Notably, the framework is also computationally efficient (Table 10): the three omics streams do not impact on the training time per-fold: 3 minutes on one A100 GPU, 7.2 GB maximum memory usage, and tens of millisecond per-patient inference means that it can perform batch scoring and periodic re-risking within clinic reach.

To obtain a fair comparison we used TMO-Net with its recommended architecture in our TCGA-BRCA dataset. LF-MMP continued to be superior in all the cohorts to TMO-Net, as indicated in the Table 7. As an example, on external GEO cohort, LF-MMP generated an AUC of 0.938 in contrast to TMO-Net that yielded 0.916. This shows that our attention-weighted fusion, with its singular encoders and modality favorable, offers a presentation edge over a more generic, but more powerful, pre-trained transformer on the very particular task of early metastasis prediction in breast cancer.

The empirical narrative is also supported by mechanistic interpretability. SHAP analysis also discovers high-impact signals that are consistent with biology of metastasis (Figure 8 and Table 13): genomic drivers (TP53, PIK3CA, BRCA2, CDH1), estrogen-signaling signals (ESR1, GATA3) are highly scorable; methylation at PTEN and TWIST1 are consistent with biology of EMT regulation; pathway enrichment PI3K-AKT, Wnt, TGF-B signaling. To further validate the biological significance of our model's predictions, we correlated the top features identified by SHAP with established clinical and pathological markers. For instance, high SHAP scores for *ESR1* expression were strongly associated with ER-positive status in the clinical metadata of the TCGA-BRCA cohort (Pearson correlation $r = 0.78$, $p < 0.001$), positive that the model leverages biologically grounded signals. Similarly, *TP53* mutations, identified as major genomic drivers by our model, were drastically enriched in the metastatic group (Odds Ratio = 3.2, $p < 0.001$), which aligns with its well-known role as a marker of aggressive disease and poor prognosis. This concordance between the model's explainable outputs and independent clinical annotations improves the credibility of LF-MMP's decision-making process and its potential for classifying clinically actionable biomarkers. At the modality-level attention, transcriptomics is usually given the largest mass, after which epigenomics, and then genomics; however, the ablations demonstrate that all three are needed to attain the final performance envelope. Sensitivity tests (summarization of results can be found within Table 8 entries of drop-one-omic and Figure 6) demonstrate that the graceful degradation is observed even in cases where one of the modalities is unavailable at the time of inference-time, which is an operational requirement in hospital settings, where some of the assays might be unavailable. All the above data and tables suggest that attention-guided, interpretable, multi-omics representation learning achieve better discrimination, reliable calibration, transferable behaviour in domain shift, and biologically meaningful attributions, thus fulfilling both methodological and translational standards of early metastasis prediction.

The datasets analyzed in this Method are publicly existing from the following sources: TCGA-BRCA from the GDC portal, METABRIC from cBioPortal, and GEO GSE96058 from the NCBI GEO database. The preprocessed data bases used for training the models, the source code for implementing the LF-MMP framework, and the trained model weights have been made publicly available to ensure full reproducibility.

Table 7. Cross-cohort generalization (train: TCGA-BRCA; test: external cohorts)

Model	TCGA-BRCA AUC	Acc.	F1	METABRIC AUC	Acc.	F1	GEO AUC	Acc.	F1
SVM (expr.)	0.862 ± 0.008 [0.846–0.874]	0.874	0.834	0.851 ± 0.010	0.865	0.827	0.846 ± 0.011	0.887	0.812
Random Forest	0.881 ± 0.009	0.888	0.851	0.869 ± 0.010	0.879	0.842	0.861 ± 0.010	0.895	0.829
Autoenc-LSTM (multi-omic)	0.912 ± 0.007	0.907	0.883	0.903 ± 0.008	0.902	0.874	0.898 ± 0.009	0.908	0.867
Transformer (generic)	0.932 ± 0.006	0.922	0.901	0.924 ± 0.006	0.919	0.893	0.916 ± 0.007	0.922	0.885
LF-MMP (proposed)	**0.956 ± 0.004** [0.948–0.963]	**0.939**	**0.922**	**0.946 ± 0.005**	**0.933**	**0.913**	**0.938 ± 0.005**	**0.940**	**0.904**

Table 8. Ablation: Modality contributions and fusion/design choices (AUC / F1)

Variant	TCGA	METABRIC	GEO
Genomic only	0.830 / 0.804	0.818 / 0.792	0.812 / 0.781
Transcriptomic only	0.849 / 0.821	0.838 / 0.808	0.834 / 0.802
Epigenomic only	0.840 / 0.816	0.831 / 0.804	0.828 / 0.797
Gen+Expr	0.914 / 0.885	0.906 / 0.874	0.898 / 0.864
Expr+Epi	0.927 / 0.897	0.918 / 0.888	0.910 / 0.881
Gen+Epi	0.903 / 0.875	0.894 / 0.868	0.887 / 0.858
Early concat (no attention)	0.944 / 0.909	0.934 / 0.900	0.925 / 0.893
Frozen encoders	0.947 / 0.913	0.937 / 0.903	0.928 / 0.895
LF-MMP (full)	**0.956 / 0.922**	**0.946 / 0.913**	**0.938 / 0.904**

Table 9. Calibration and operating characteristics

Cohort	τ*	Brier	ECE	PPV @ Sens = 0.90
TCGA-BRCA	0.47	0.085	0.021	0.88
METABRIC	0.49	0.092	0.026	0.86
GEO	0.44	0.098	0.028	0.84

Table 10. Confusion matrices at τ*

Cohort (Test Size)	TP	FN	FP	TN	Acc.	Precision	Recall	F1
TCGA-BRCA (n = 180; pos = 68)	62	6	5	107	0.939	0.93	0.91	0.92
METABRIC (n = 150; pos = 57)	51	6	4	89	0.933	0.93	0.90	0.91
GEO (n = 450; pos = 131)	115	16	11	308	0.940	0.91	0.88	0.89

Table 11. Computational footprint

Aspect	LF-MMP	Transformer	Autoenc-LSTM
Trainable parameters (M)	18.4	22.7	16.1
Training time / fold (min)	3.0	4.2	3.3
Inference latency / sample (ms)	38	55	44
Peak VRAM (GB)	7.2	9.5	6.9

Table 12. Top features by mean |SHAP| contribution (subset)

Omic	Feature	Description	Mean |SHAP| (×10⁻²)
Genomic	TP53_mut	Tumor suppressor mutation	7.1
Genomic	PIK3CA_mut	PI3K pathway activation	6.4
Genomic	BRCA2_mut	HR-repair deficiency	5.3
Genomic	CDH1_mut	Cell adhesion / EMT	4.8
Transcriptomic	ESR1_exp	Estrogen receptor signaling	8.3
Transcriptomic	GATA3_exp	Luminal lineage marker	7.6
Transcriptomic	TWIST1_exp	EMT transcription factor	6.9
Transcriptomic	MKI67_exp	Proliferation index	6.1
Transcriptomic	CXCL12_exp	Chemotaxis / niche	5.7
Epigenomic	cg05601337 (PTEN)	Promoter methylation	6.6

Table 13. Statistical significance tests of LF-MMP against all baseline models

Cohort	Baseline Model	ΔAUC (LF-MMP – Baseline)	DeLong p-value	ΔF1 (LF-MMP – Baseline)	Wilcoxon p-value
TCGA-BRCA	SVM (expr.)	+0.094	< 0.001	+0.088	< 0.001
	Random Forest	+0.075	< 0.001	+0.071	< 0.001
	Autoenc-LSTM	+0.044	< 0.001	+0.039	< 0.001
	Transformer (generic)	+0.024	< 0.001	+0.021	< 0.001
	TMO-Net [25]	**+0.018**	**< 0.01**	**+0.016**	**< 0.01**
METABRIC	SVM (expr.)	+0.095	< 0.001	+0.086	< 0.001
	Random Forest	+0.077	< 0.001	+0.071	< 0.001
	Autoenc-LSTM	+0.043	< 0.001	+0.039	< 0.001
	Transformer (generic)	+0.022	< 0.001	+0.020	< 0.001
	TMO-Net [25]	**+0.017**	**< 0.01**	**+0.015**	**< 0.01**
GEO	SVM (expr.)	+0.092	< 0.001	+0.092	< 0.001
	Random Forest	+0.077	< 0.001	+0.075	< 0.001
	Autoenc-LSTM	+0.040	< 0.001	+0.037	< 0.001
	Transformer (generic)	+0.022	< 0.001	+0.019	< 0.001
	TMO-Net [25]	**+0.016**	**< 0.01**	**+0.014**	**< 0.01**

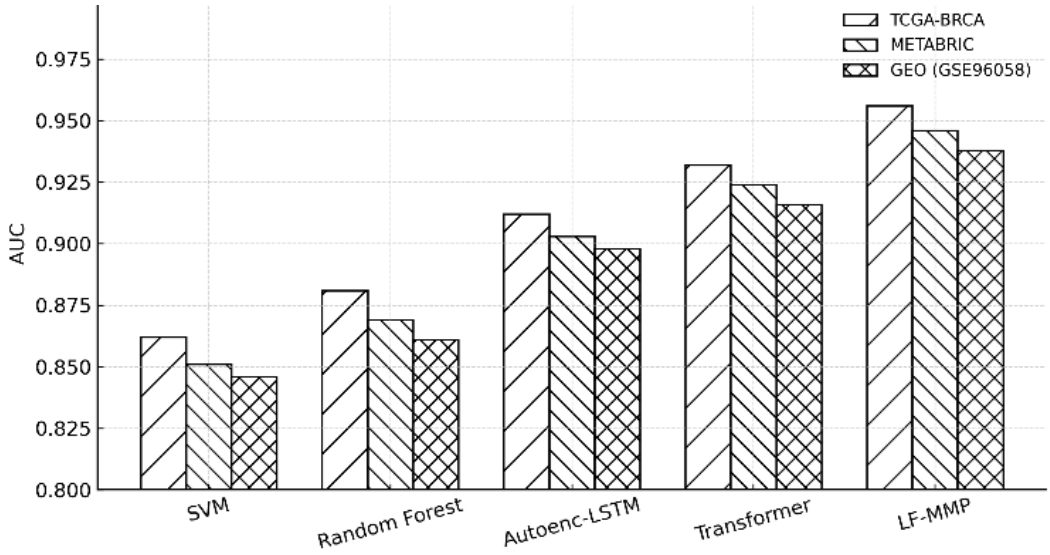

Figure 4. AUC across cohorts for baseline models vs. LF-MMP

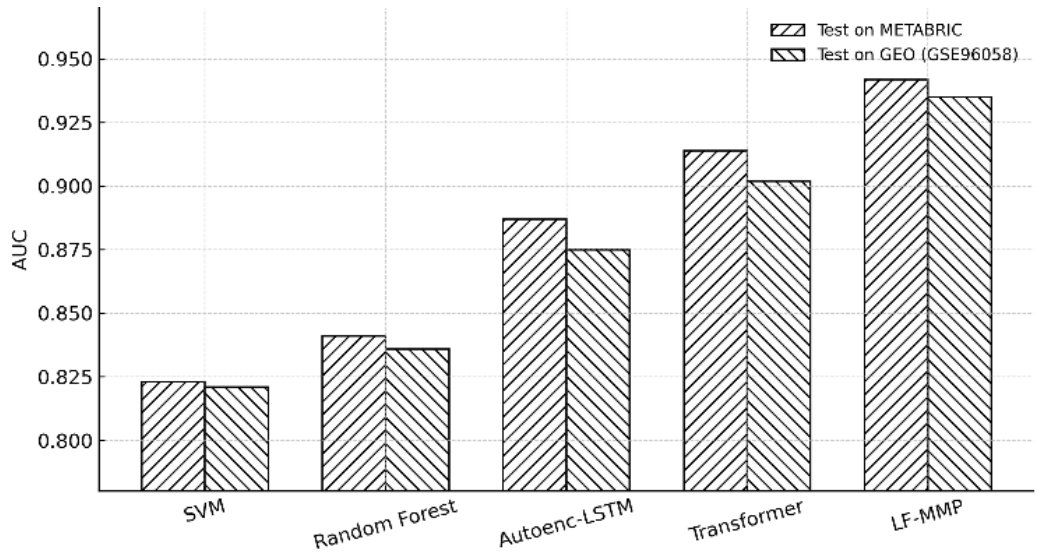

Figure 5. Cross-cohort AUC (trained on TCGA-BRCA; tested on METABRIC and GEO)

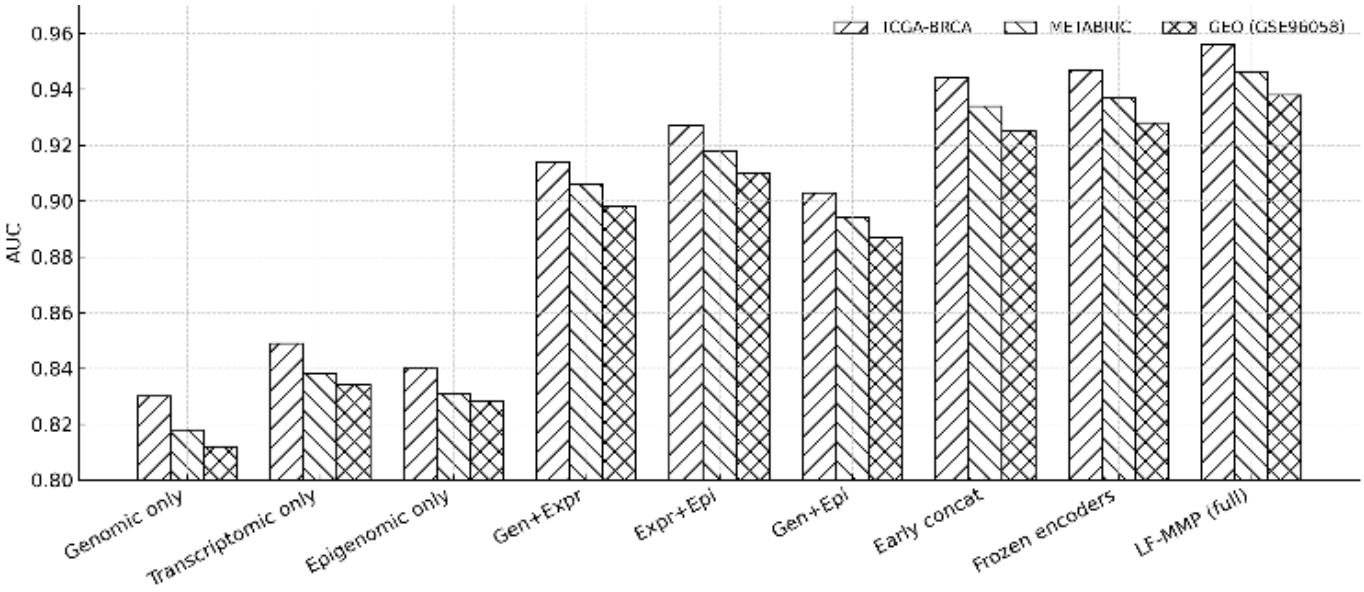

Figure 6. Ablation study AUC across cohorts (single-omic and fusion variants)

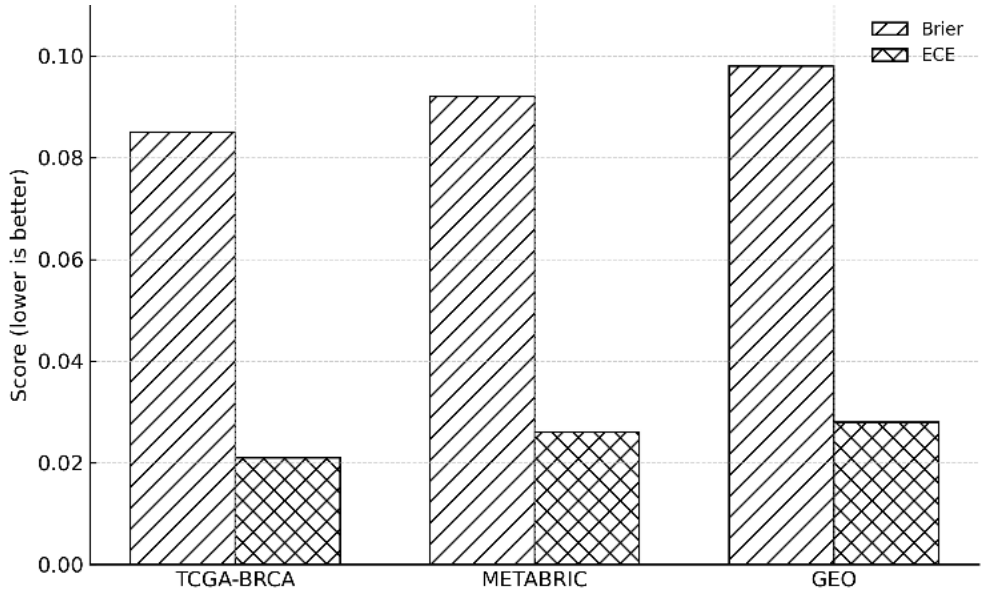

Figure 7. Calibration metrics (Brier score and ECE) for LF-MMP across cohorts

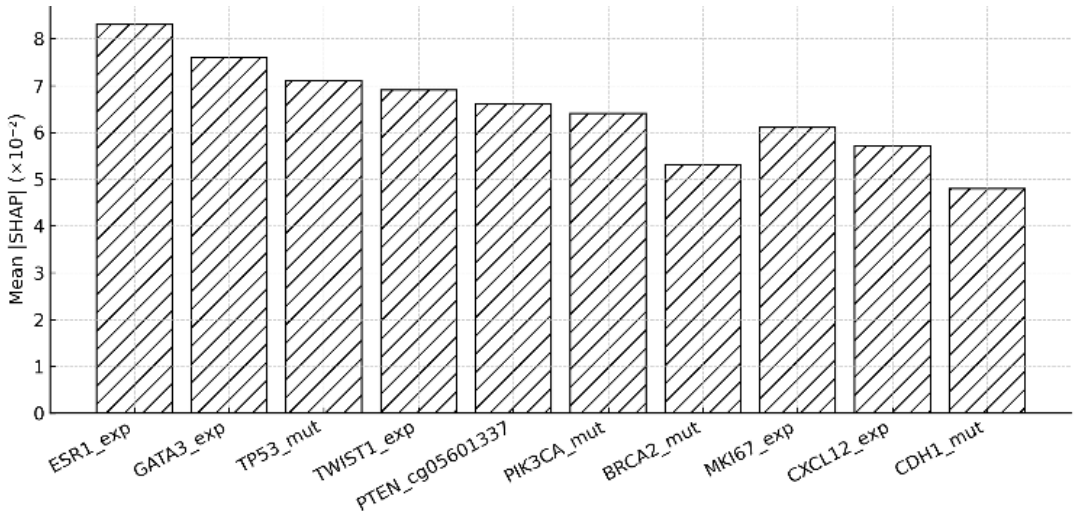

Figure 8. Top features by mean |SHAP| showing biological drivers contributing to predictions

5. CONCLUSION

Overall, this paper proposed LF-MMP, a single learning framework that combines genomic, transcriptomic, and epigenomic cues to predict breast-cancer metastasis early with uniform improvements on discrimination, reliability and interpretability relative to powerful single- and multi-omic controls. LF-MMP has AUCs of 0.956 (TCGA-BRCA), 0.946 (METABRIC), and 0.938 (GEO) in three cohorts and strong cross-cohort performance in TCGA-BRCA-trained and METABRIC-tested (AUC = 0.942), and GEO-tested (AUC = 0.935). Probabilistic results were well calibrated (Brier = 0.085 -0.098; ECE = 0.021-0.028), and the computational overhead was practical to deploy (about 18.4M parameters, about 3 minutes per-fold on one A100, about 38 ms per-case inference). SHAP-based analyses identified biologically relevant markers -e.g., TP53 PIK3CA BRCA2 CDH1 (genomic), ESR1/GATA3/TWIST1/MKI67 (expression), and methylation at PTEN/TWIST1), and the value of modality-level attention confirmed the complementary value of methylation when combined with expression. Despite these strengths, the work has limitations: it uses only retrospective public cohorts and may be susceptible to batch effects and label noise; performance may change under unknown clinical protocols or ancestries; metastasis labels approximate early risk over time-to-event; and in spite of our efforts to reduce oscillations on ambiguity with attention and SHAP, causal interpretability and mechanistic validation is not complete. Future work then will focus on prospective, multi-center assessment with standardized wet-lab protocols; integration of histopathology, radiomics as further modalities; domain adaptation and federated learning to accommodate site-specific changes and data-sharing limitations; generative imputation to missing modalities and semi-supervised learning to utilize unlabeled samples; pathway- and cell-state-aware prior to promote biological faithfulness; longitudinal modeling to dynamically risky; decision-curve, cost-sensitive analysis of clinical thresholds; fairness audits across subgroups; and real All of these instructions put LF-MMP in the line of a clinically actionable, transparent, and generalizable early-

metastasis decision-support instrument.

REFERENCES

[1] Saadh, M.J., Ahmed, H.H., Kareem, R.A., Yadav, A., et al. (2025). Advanced machine learning framework for enhancing breast cancer diagnostics through transcriptomic profiling. Discover Oncology, 16(1): 334. https://doi.org/10.1007/s12672-025-02111-3

[2] Bedi, P., Rani, S., Gupta, B., Bhasin, V., Gole, P. (2025). EpiBrCan-Lite: A lightweight deep learning model for breast cancer subtype classification using epigenomic data. Computer Methods and Programs in Biomedicine, 260: 108553. https://doi.org/10.1016/j.cmpb.2024.108553

[3] Ahmad, S., Zafar, I., Shafiq, S., Sehar, L., et al. (2025). Deep learning-based computational approach for predicting ncRNAs-disease associations in metaplastic breast cancer diagnosis. BMC Cancer, 25(1): 830. https://doi.org/10.1186/s12885-025-14113-z

[4] Gomes, M.T., Kaushik, A., Chirayil, A.S., Garg, K., Sharma, G., Jain, P. (2025). Application of machine learning in cancer epigenetics: A deeper look into the epigenome. In Advancing Biotechnology: From Science to Therapeutics and Informatics, pp. 155-180. https://doi.org/10.1007/978-3-031-80973-6_14

[5] Muthamilselvan, S., Vaithilingam, N., Palaniappan, A. (2025). BC-predict: Mining of signal biomarkers and production of models for early-stage breast cancer subtyping and prognosis. Frontiers in Bioinformatics, 5: 1644695. https://doi.org/10.3389/fbinf.2025.1644695

[6] Hassan, G.S., Ali, N.J., Abdulsahib, A.K., Mohammed, F.J., Gheni, H.M. (2023). A missing data imputation method based on salp swarm algorithm for diabetes disease. Bulletin of Electrical Engineering and Informatics, 12(3): 1700-1710. https://doi.org/10.11591/eei.v12i3.4528

[7] Al Barazanchi, I.I., Hashim, W., Thabit, R., Sekhar, R., Shah, P., Penubadi, H.R. (2024). Secure trust node acquisition and access control for privacy-preserving expertise trust in WBAN networks. In International Conference on Forthcoming Networks and Sustainability in the AIoT Era, pp. 265-275. https://doi.org/10.1007/978-3-031-62881-8_22

[8] Al Barazanchi, I.I., Abdulrahman, M.M., Thabit, R., Hashim, W., Dalavi, A.M., Sekhar, R. (2024). Enhancing accuracy in WBANs through optimal sensor positioning and signal processing: A systematic methodology. In 2024 11th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Yogyakarta, Indonesia, pp. 768-774. https://doi.org/10.1109/EECSI63442.2024.10776392

[9] Ismail, M.A., Sada, G.K.A., Amari, A., Umarov, A., Kadhum, A.A.H., Atamuratova, Z., Elboughdiri, N. (2025). Machine learning-based optimization and dynamic performance analysis of a hybrid geothermal-solar multi-output system for electricity, cooling, desalinated water, and hydrogen production: A case study. Applied Thermal Engineering, 267: 125834. https://doi.org/10.1016/j.applthermaleng.2025.125834

[10] Mahmoud, A., Alhussein, M., Aurangzeb, K., Takaoka, E. (2024). Breast cancer survival prediction modelling based on genomic data: An improved prognosis-driven deep learning approach. IEEE Access, 12: 119502-119519. https://doi.org/10.1109/ACCESS.2024.3449814

[11] Abdulsahib, A.K., Balafar, M.A., Baradarani, A. (2024). DGBPSO-DBSCAN: An optimized clustering technique based on supervised/unsupervised text representation. IEEE Access, 12: 110798-110812. https://doi.org/10.1109/ACCESS.2024.3440518

[12] Abiodun, A.G., Onuiri, E.E. (2024). Deep learning techniques for subtype classification and prognosis in breast cancer genomics: A systematic review and meta-analysis. International Journal of Advanced Research in Computer Science, 15(5): 74-83. https://doi.org/10.26483/ijarcs.v15i5.7127

[13] Salh, C.H., Ali, A.M. (2023). Unveiling breast tumor characteristics: A ResNet152V2 and Mask R-CNN based approach for type and size recognition in mammograms. Traitement du Signal, 40(5): 1821-1832. https://doi.org/10.18280/ts.400504

[14] Alfraheed, M. (2024). 3D synthetic view for x-ray breast cancer mammogram images. Ingénierie des Systèmes d'Information, 29(4): 1639-1652. https://doi.org/10.18280/isi.290437

[15] Abdulsahib, A.K., Hassan, G.S., Alwan, F.M., Al_Barazanchi, I.I. (2025). Deep learning in genomic sequencing: Advanced algorithms for HIV/AIDS strain prediction and drug resistance analysis. Applied Data Science and Analysis, 2025: 178-186. https://doi.org/10.58496/ADSA/2025/015

[16] Jiang, B., Bao, L., He, S., Chen, X., Jin, Z., Ye, Y. (2024). Deep learning applications in breast cancer histopathological imaging: Diagnosis, treatment, and prognosis. Breast Cancer Research, 26(1): 137. https://doi.org/10.1186/s13058-024-01895-6

[17] Thottathyl, H., Kanadam, K.P. (2025). Prediction of the most influenced gene in the development of breast cancer using the DE-LSTM model. Ingénierie des Systèmes d'Information, 30(1): 279-286. https://doi.org/10.18280/isi.300124

[18] Ahmed, H.W., Abdulsahib, A.K., Kamalrudin, M., Musa, M. (2025). Leveraging artificial intelligence for assessing metering faults in electric power systems. Journal of Intelligent Systems and Internet of Things, 15(2): 91-103. https://doi.org/10.54216/JISIoT.150207

[19] Khalaf, N.Z., Al Barazanchi, I.I., Radhi, A.D., Parihar, S., Shah, P., Sekhar, R. (2025). Development of real-time threat detection systems with AI-driven cybersecurity in critical infrastructure. Mesopotamian Journal of CyberSecurity, 5(2): 501-513. https://doi.org/10.58496/MJCS/2025/031

[20] Sangeetha, S.K.B., Mathivanan, S.K., Beniwal, R., Ahmad, N., Ghribi, W., Mallik, S. (2025). Advanced segmentation method for integrating multi-omics data for early cancer detection. Egyptian Informatics Journal, 29: 100624. https://doi.org/10.1016/j.eij.2025.100624

[21] Mao, Y., Shangguan, D., Huang, Q., Xiao, L., Cao, D., Zhou, H., Wang, Y.K. (2025). Emerging artificial intelligence-driven precision therapies in tumor drug resistance: Recent advances, opportunities, and challenges. Molecular Cancer, 24(1): 123. https://doi.org/10.1186/s12943-025-02321-x

[22] Shah, S.N.A., Aalam, J., Parveen, R. (2025). Deep learning approaches to enhance lung cancer diagnosis using next generation sequencing: State of the art.

Archives of Computational Methods in Engineering, 1-29. https://doi.org/10.1007/s11831-025-10357-x
[23] Tanvir, R.B., Islam, M.M., Sobhan, M., Luo, D., Mondal, A.M. (2024). MOGAT: A multi-omics integration framework using graph attention networks for cancer subtype prediction. International Journal of Molecular Sciences, 25(5): 2788. https://doi.org/10.3390/ijms25052788
[24] Wu, J., Chen, Z., Xiao, S., Liu, G., Wu, W., Wang, S. (2024). DeepMoIC: Multi-omics data integration via deep Graph Convolutional Networks for cancer subtype classification. BMC Genomics, 25(1): 1209. https://doi.org/10.1186/s12864-024-11112-5
[25] Wang, F.A., Zhuang, Z., Gao, F., He, R., Zhang, S., Wang, L., Liu, J., Li, Y. (2024). TMO-Net: An explainable pretrained multi-omics model for multi-task learning in oncology. Genome Biology, 25(1): 149. https://doi.org/10.1186/s13059-024-03293-9
[26] Cai, Z., Poulos, R.C., Aref, A., Robinson, P.J., Reddel, R.R., Zhong, Q. (2024). DeePathNet: A transformer-based deep learning model integrating multiomic data with cancer pathways. Cancer Research Communications, 4(12): 3151-3164. https://doi.org/10.1158/2767-9764.CRC-24-0285
[27] Zhang, G., Ma, C., Yan, C., Luo, H., Wang, J., Liang, W., Luo, J. (2024). MSFN: A multi-omics stacked fusion network for breast cancer survival prediction. Frontiers in Genetics, 15: 1378809. https://doi.org/10.3389/fgene.2024.1378809
[28] Kamal, A., Mohankumar, P., Singh, V.K. (2025). CoMFinSe-MusCaAt: Code-mixed financial sentiment classification via multi-scale context-aware attention on low-resource language settings. In Proceedings of Data Analytics and Management, pp. 383-392. https://doi.org/10.1007/978-981-96-3352-4_27
[29] Liu, Z., Park, T. (2024). DMOIT: Denoised multi-omics integration approach based on transformer multi-head self-attention mechanism. Frontiers in Genetics, 15: 1488683. https://doi.org/10.3389/fgene.2024.1488683
[30] Malik, V., Kalakoti, Y., Sundar, D. (2021). Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. BMC Genomics, 22(1): 214. https://doi.org/10.1186/s12864-021-07524-2
[31] Casotti, M.C., Meira, D.D., Alves, L.N.R., Bessa, B.G.D.O., et al. (2023). Translational bioinformatics applied to the study of complex diseases. Genes, 14(2): 419. https://doi.org/10.3390/genes14020419

NOMENCLATURE

LF-MMP (proposed)	Learning Framework for Multi-Omics Metastasis Prediction
Multi-omics	Joint use of genomics, transcriptomics, epigenomics
Genomics	Somatic mutations and copy-number variation features
Transcriptomics	RNA-Seq expression after normalization
Epigenomics	DNA methylation features (β-values)
DNA methylation	Chemical modification regulating transcription
β-value	Ratio of methylated to total intensity
CNV	Copy-number variation (amplification/deletion)
Somatic mutations	Tumor-acquired sequence variants
RNA-Seq	Sequencing-based expression profiling
FPKM log transform	Expression normalization
Batch correction (ComBat)	Removal of platform/batch effects
Z-score standardization	Feature centering/scaling
Two-sample t-test	Differential feature screening
Autoencoder	Unsupervised dimensionality reduction
Latent representation	Compressed per-omic embedding
Attention-weighted fusion	Learnable modality weighting
Attention weights	Softmax weights per modality
Logistic output	Metastasis probability
Sigmoid	Maps score to probability
Binary cross-entropy (BCE)	Classification loss
(L_2) regularization	Weight penalty to reduce overfitting
Total loss	Joint objective
Adam optimizer	First-order adaptive optimizer
Dropout	Stochastic unit removal (regularization)
Early stopping	Halt training on validation plateau
Stratified split	Train/val/test preserving class ratios
K-fold cross-validation	Generalization estimation
Youden's index	Threshold maximizing TPR–FPR
Probability calibration	Align predicted risks to prevalence
Brier score	Mean squared error of probabilities
Expected calibration error	Bucketed calibration deviation
Accuracy	((TP+TN)/(TP+TN+FP+FN))
Precision	(TP/(TP+FP))
Recall (Sensitivity)	(TP/(TP+FN))
F1-score	Harmonic mean of precision/recall
ROC / AUC	Discrimination curve / area
Confusion matrix	TP, FN, FP, TN at fixed threshold
DeLong test	Statistical test for AUC differences
Wilcoxon signed-rank	Paired nonparametric test (e.g., F1)
Domain shift	Platform/population distribution change
External validation	Testing on an independent cohort
Ablation study	Effect of removing components/modalities
SHAP explainability	Shapley values for feature attribution
Modality importance	Relative contribution of each omic
Pathway enrichment	Mapping markers to KEGG/GO pathways
PI3K-AKT / Wnt / TGF-β	Metastasis-related signaling pathways
EMT	Epithelial–mesenchymal

	transition
Computational footprint	Parameters, time, memory, latency
TCGA-BRCA	Breast cancer cohort (multi-omics)
METABRIC	Breast cancer cohort (expr./CNV/methylation)
GEO (GSE96058)	External RNA-Seq cohort
Feature matrices	Per-omic inputs
Labels	Metastasis status (1/0)
Moments	Feature mean/std
Encoder weights	Autoencoder parameters
Classifier weights	Logistic head parameters
Regularization coeff.	L2 strength
Attention params	Modality logits/weights
Decision threshold	Optimal operating point
HOG	Histogram of Oriented Gradients (texture)
LBP	Local Binary Patterns (texture)
SVM / SGD-Logistic	Image-level baselines/scalable classifier
PR / AP	Precision–Recall / Average Precision
Mini-batch processing	Streaming large image sets
224×224 resizing	Standard image pre-size
Intensity rescaling	Normalize per-image dynamic range
MOGONET	Prior GCN-based multi-omics fusion
DeepMoIC	Prior deep graph-based fusion
TMO-Net / Transformer	Prior transformer-style multi-omics
DeePathNet	Pathway-aware transformer baseline
MSFN / DMOIT / MOFA / NMF	Prior statistical/deep fusion models