Computer Methods and Programs in Biomedicine 279 (2026) 109288

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

A multimodal data-based model for breast cancer diagnosis

Huina Wang [a,d], Lan Wei [b], Jianqiang Li [a], Bo Liu [c], Juan Fang [a], Catherine Mooney [d,*]

[a] School of Computer Science, Beijing University of Technology, Beijing, 100124, China
[b] School of Electrical and Electronic Engineering, University College Dublin, Dublin, D04 V1W8, Ireland
[c] School of Mathematical and Computational Sciences, Massey University, Auckland, 0745, New Zealand
[d] FutureNeuro Research Ireland Centre, School of Computer Science, University College Dublin, Dublin, D04 V1W8, Ireland

ARTICLE INFO

Keywords:
Diagnostic systems
Contrastive learning
Cross-modal learning
Multimodal classification

ABSTRACT

Background and Objective: Developing multimodal data-driven diagnostic systems has become a key clinical strategy for improving breast cancer outcomes. However, effectively modeling multimodal features remains challenging due to substantial semantic heterogeneity, scale discrepancies, and the inherent difficulty of cross-modal alignment. Although existing studies have proposed various multimodal fusion methods, most rely on direct feature concatenation or shallow integration, which fail to capture fine-grained intra-modality semantics as well as the complex interactions between histopathological and genomic modalities.

Methods: In this study, we propose a multimodal diagnostic framework based on Feature Enhancement and Semantic Collaborative Alignment (FESCA). The method incorporates a semantic-guided modality feature enhancement mechanism that effectively extracts and strengthens diagnostic cues from both pathological images and genomic data. In addition, a contrastive-learning-based cross-modal alignment strategy is introduced to map heterogeneous modalities into a unified semantic space and achieve deep semantic collaboration through contrastive optimization. To ensure robust breast cancer classification under varying modality availability, a multimodal collaborative diagnostic strategy is employed to dynamically adapt the feature representations.

Results: We evaluate FESCA on the TCGA-BRCA dataset, and the experimental results demonstrate that it outperforms state-of-the-art methods in breast cancer classification while significantly improving both intra-modality representation quality and cross-modal semantic alignment.

Conclusion: To enhance accessibility and practical application, we developed a web-based breast cancer pathological staging diagnosis system to visualize and deploy the FESCA model, demonstrating a step toward clinical application and providing a benchmark for other research methods.

1. Introduction

Breast cancer (BC) has become the leading cause of death among women worldwide, and its global incidence continues to rise [1,2]. Therefore, it is crucial to develop a highly accurate and efficient diagnostic model to improve diagnostic efficiency and patient survival rates. With the continuous advancements in experimental techniques and sequencing technology, deep learning is widely used in the field of breast cancer disease diagnosis [3–5]. However, most previous studies have relied on single-modality data, which lack the richness and diversity necessary to comprehensively represent complex diseases, thereby limiting both the diagnostic performance and generalization ability of the resulting models [6]. Recently, multi-modal medical diagnostic models that combine histological and omics data have demonstrated significant promise [7–9].

With the rapid development of digital pathology and high-throughput sequencing technologies, integrating histopathological imaging and genomics has become an important direction in multimodal medical analysis [7–10]. Unlike conventional radiogenomic studies, this line of research focuses on whole-slide images (WSIs) as the primary imaging modality. By jointly modeling histological patterns and genomic features, these approaches enable cross-scale analyses that link tissue morphology to underlying molecular mechanisms. Recent multimodal diagnostic models combining imaging and genomic data have demonstrated strong potential in cancer subtyping, prognosis prediction, and molecular feature inference [7–9]. Such cross-modal modeling improves diagnostic robustness and supports biologically interpretable precision medicine. A key prerequisite for effective multimodal integration is the construction of stable and discriminative representations for both WSIs and genomic data. For WSIs, multiple

* Corresponding author.
E-mail addresses: huinawang@emails.bjut.edu.cn (H. Wang), lan.wei@ucd.ie (L. Wei), lijianqiang@bjut.edu.cn (J. Li), b.liu@massey.ac.nz (B. Liu), fangjuan@bjut.edu.cn (J. Fang), catherine.mooney@ucd.ie (C. Mooney).

https://doi.org/10.1016/j.cmpb.2026.109288
Received 10 August 2025; Received in revised form 22 January 2026; Accepted 17 February 2026
Available online 23 February 2026

instance learning (MIL) has become the dominant paradigm [11–13], typically involving patch extraction, patch-level feature encoding, and slide-level aggregation. For genomic data, representation learning methods such as autoencoders [14], neural network-based models [15], and Transformer-based encoders [16] have been widely adopted. These advances have laid the foundation for joint analysis of imaging and genomic features. Nevertheless, WSI and genomic data exhibit substantial differences in dimensionality, data distribution, and semantic structure, which makes it difficult to construct a unified diagnostic representation. Although various approaches have been proposed to address these challenges, most existing methods primarily emphasize architectural fusion or feature aggregation, with limited consideration of explicitly modeling diagnostically meaningful semantic priors within each modality and enforcing semantic consistency across modalities. As a result, fine-grained diagnostic cues and task-relevant semantic relationships are often underutilized in current multimodal diagnostic frameworks.

To address this gap, we propose a Multimodal Diagnostic Model Based on Feature Enhancement and Semantic Collaborative Alignment (FESCA), which explicitly injects diagnostically relevant semantic priors into intra-modality representation learning and enforces semantic collaboration across modalities at the representation level. Specifically, a semantic-guided feature enhancement mechanism is introduced, which incorporates semantic prior information such as annotated lesion regions and differentially expressed genes to enable efficient extraction and discriminative enhancement of fine-grained features within each modality. On this basis, a contrastive-learning-based semantic alignment strategy is developed to achieve deep semantic collaboration between histopathological images and genomic data through modality-specific feature mapping and contrastive optimization. This approach significantly enhances the semantic modeling capability and diagnostic performance of the multimodal diagnostic framework. Unlike prior multimodal diagnostic models that focus primarily on fusion architecture design, FESCA is built upon a semantic-driven modeling perspective that jointly enhances intra-modality representations and cross-modality semantic alignment. The main contributions of this research are as follows:

- We propose a multimodal diagnostic model (FESCA) that enables collaborative modeling between histopathological imaging and genomic data.
- A semantic-guided feature enhancement mechanism leverages prior knowledge, including lesion annotations and differentially expressed genes, to enhance critical modality-specific representations.
- A contrastive-learning-based cross-modal semantic alignment method leverages modality-specific feature mapping and contrastive optimization to achieve semantic collaboration between histopathological and genomic modalities.
- Based on the proposed FESCA model, an intelligent breast cancer diagnostic system is developed to validate the feasibility and potential of the proposed approach for clinical decision support and computer-aided diagnosis.

2. Related works

WSIs are ultra-high-resolution histopathological scans that provide detailed cellular and tissue-level information, but their gigapixel-scale size and high heterogeneity pose significant challenges for efficient computation and discriminative feature extraction. To address these issues, multiple instance learning (MIL) has become the dominant framework for WSI representation learning [11,12,17,18]. Existing MIL-based approaches can be broadly categorized into attention-based methods, which highlight cancer-relevant regions via adaptive weighting [19–21]; transformer-based methods, which model global contextual dependencies among patches [13,22]; and graph-based methods, which capture spatial and semantic tissue relationships through graph representations [23,24]. Although these approaches have achieved strong performance in cancer classification and prognosis tasks, image-only models remain constrained by weak supervision and the absence of molecular-level information, limiting their ability to fully capture tumor biological heterogeneity.

In parallel, RNA sequencing data provide complementary molecular insights into gene regulation and tumor heterogeneity, but present their own challenges due to high dimensionality, feature sparsity, and complex nonlinear interactions. To effectively encode such tabular genomic data, deep learning-based approaches have been developed, including data transformation methods [14], specialized architectures for tabular learning [25–27], and regularization-based models to improve robustness and generalization [28]. Despite these advances, each modality alone provides only a partial view of tumor biology. This complementary yet incomplete nature of histopathological and genomic data has motivated the development of multimodal fusion approaches that jointly leverage imaging and molecular information to achieve more comprehensive and accurate cancer diagnosis.

A variety of multimodal fusion strategies have been proposed to integrate histopathological and genomic information for cancer diagnosis, which can be broadly categorized into feature-level, decision-level, and joint fusion approaches. Feature-level fusion typically concatenates WSI and omics features, offering simplicity but failing to model high-order cross-modal interactions and being sensitive to missing modalities [29,30]. Decision-level fusion leverages shared latent spaces, attention mechanisms, or dual-tower architectures to align modalities [8,9,31], but often suffers from shallow cross-modal interactions. Joint fusion methods integrate modalities within unified architectures [32], yet challenges in modality alignment, data imbalance, and interpretability remain. To better capture complex cross-modal relationships, Transformer-based multimodal frameworks have gained increasing attention. Representative methods include MCAT [18], which applies co-attention for omics-guided WSI modeling; MOTCat [33], which employs optimal transport for improved alignment; SurvPath [34], which models interactions between tokenized genomic features and WSIs; and CMTA [35] as well as prototype-based approaches [36], which enhance semantic alignment and interpretability through parallel encoders and biological priors.

In summary, multimodal diagnostic models that integrate histopathological imaging and genomic data have emerged as a promising direction in precision oncology, demonstrating strong potential in cancer subtyping, molecular phenotype prediction, and prognosis estimation. Nevertheless, existing approaches still exhibit notable limitations. First, unlike prior methods that implicitly assume modality-specific features are sufficiently discriminative once fused, FESCA explicitly strengthens intra-modality representations by incorporating domain semantic priors, including lesion-relevant regions in WSIs and differentially expressed genes in genomic data. This design enables fine-grained, task-relevant semantic enhancement within each modality before multimodal interaction. Second, while many transformer-based frameworks emphasize architectural fusion or attention-based aggregation to model cross-modal interactions, FESCA introduces a contrastive-learning-based cross-modal semantic alignment strategy that explicitly enforces semantic consistency between histopathological and genomic representations at the representation level, rather than relying solely on structural coupling. Finally, in contrast to prognosis-oriented or task-specific multimodal models that focus on prediction heads, FESCA emphasizes semantic collaboration and alignment as a general representation-learning principle, making it particularly suited for diagnostic classification tasks that require interpretable and complementary multimodal features. The related works are summarized in Table 1, which provides an overview of representative approaches across imaging, genomics, and multimodal integration.

Table 1
A summary of related work on diagnostic models based on WSIs and mRNA data.

Method	Year	Data modality	Method type	Model mechanism
Attention MIL [19]	2018	WSIs	Attention-based	Patch attention for instance weighting
TransMIL [13]	2021	WSIs	Transformer-based	Global patch relation modeling
CLAM [20]	2022	WSIs	Attention-based	Learn attention and instance clustering
DTFD-MIL [22]	2022	WSIs	Transformer-based	Dual transformer + feature decomposition fusion.
NAGCN [23]	2022	WSIs	Graph-based	Hierarchical global-to-local clustering strategy
Additive MIL [21]	2022	WSIs	Attention-based	Spatial credit assignment
MamMIL [24]	2024	WSIs	Graph-based	Selective structured state space(Mamba)+MIL
CatBoost [26]	2018	Tabular data	Hybrid models	Ordered Boosting+Target Encoding
RLNs [28]	2018	Tabular data	Regularization-based	Proposing a hyperparameter tuning strategy
NODE [25]	2019	Tabular data	Hybrid models	NODE extends traditional ensemble methods by building on oblivious decision trees
TabTransformer [37]	2020	mRNA	Transformer-based	Built upon self-attention based Transformers
TabNet [27]	2021	Tabular data	Transformer-based	Sequential attention to select salient features
SAINT [38]	2024	Tabular data	Transformer-based	Self-attention+intersample attention
AEmiGAP [14]	2024	micRNA	Transformation-based	Autoencoders+long short-term memory networks
MCAT [18]	2021	WSIs+mRNA	Attention-based	Generating omic-guided histology features for survival prediction
MOTCat [33]	2023	WSIs+mRNA	Co-attention Transformer	Optimal Transport to learn an optimal plan between histology and genes
CMTA [35]	2023	WSIs+mRNA	Transformer-based	Two parallel Transformer encoder–decoder modules
MMP [36]	2023	WSIs+mRNA	Transformer-based	Morphological+biological pathway
SurvPath [34]	2024	WSIs+mRNA	Transformer-based	Transcriptomics tokenizer+multimodal transformer

3. Methodology

As shown in Fig. 1, the FESCA consists of three modules. Specifically, the semantic-guided feature enhancement module incorporates semantic prior information from lesion annotations and differential genes to achieve efficient intra-modality feature extraction and discriminative enhancement, thereby generating high-quality modality embeddings with richer semantic representation and stronger discriminability. The cross-modal semantic alignment module maps pathology radiomics features and genomic features into a shared semantic space through modality-specific feature mappers, and leverages a contrastive learning mechanism to strengthen cross-modal feature alignment and complementary representation. The multimodal collaborative diagnosis module dynamically adjusts the feature representation strategy based on the availability of input modalities, enabling accurate cancer-type prediction.

3.1. Data preparation

The mRNA dataset used in this study was obtained from the Cancer Genome Atlas (TCGA, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). HTSeq-Counts data for breast cancer (TCGA-BRCA) were downloaded using R Studio, followed by ID transformation, $\log_2$ normalization, and the removal of samples with a high proportion of missing values or abnormal expression profiles.

The WSIs from the Breast Cancer Semantic Segmentation Challenge (BCSS) are available through the Grand Challenges website (https://bcsegmentation.grandchallenge.org/), which includes 151 histologically confirmed breast cancer WSIs scanned at 40× resolution and stained with haematoxylin and eosin (H&E). These slides are part of TCGA Program and have been manually annotated by pathologists, pathology residents, and medical students via a crowdsourcing process. In this study, we first perform color normalization using the Reinhard method [39], which aligns the statistical distribution (mean and standard deviation) of image colors in the LAB color space to a reference template to reduce staining variability. Subsequently, the original WSIs are processed using Otsu's binarization algorithm to separate tissue from background, followed by morphological post-processing with a 5×5 kernel closing operation to smooth tissue boundaries and fill small internal holes.

The details of the TCGA and BCSS datasets are shown in Table 2. From these datasets, 216 samples containing both WSIs and mRNA expression data were filtered, forming the paired dataset for this study (Table 3). As shown in Table 4, a two-stage data partitioning strategy was adopted. In the first stage, aimed at cross-modal semantic alignment, 80% of the WSIs (173 slides) were used for training and the remaining 20% (43 slides) were used for validation. In the second stage, which focused on multimodal collaborative diagnosis, the complete set of 216 WSIs was re-split to ensure a rigorous evaluation: 70% (151 slides) for training, 15% (33 slides) for validation, and 15% (32 slides) for testing. The model was trained and validated multiple times during development to ensure stable behavior. All results are reported under a fixed train–validation–test split. The final test set was completely held out from model development, enabling an unbiased evaluation of generalizability within the given dataset.

Table 2
Summary of datasets.

Data type	Original samples	Filtered samples	Feature dimension	Pairing status
mRNA	1187	216	19,938	WSI paired
WSIs	151	216	–	mRNA paired

Table 3
Summary of selected paired samples.

Data	Normal	Abnormal		
		Stage I	Stage II	Stage III
mRNA	98	181	624	269
WSIs	98	25	76	17

Table 4
Data partitioning.

Modules	WSIs	Percentage
Cross-modal semantic alignment module		
Training	173	80%
Validation	43	20%
Multimodal collaborative diagnosis module		
Training	151	70%
Validation	33	15%
Test	32	15%

3.2. Semantic-guided feature enhancement module

To achieve efficient extraction and discriminative enhancement of intra-modality features, we propose a semantic-guided modality feature enhancement mechanism, which introduces semantic prior information from the pathology and genomics domains to strengthen intra-modality feature representation. Specifically, this mechanism designs

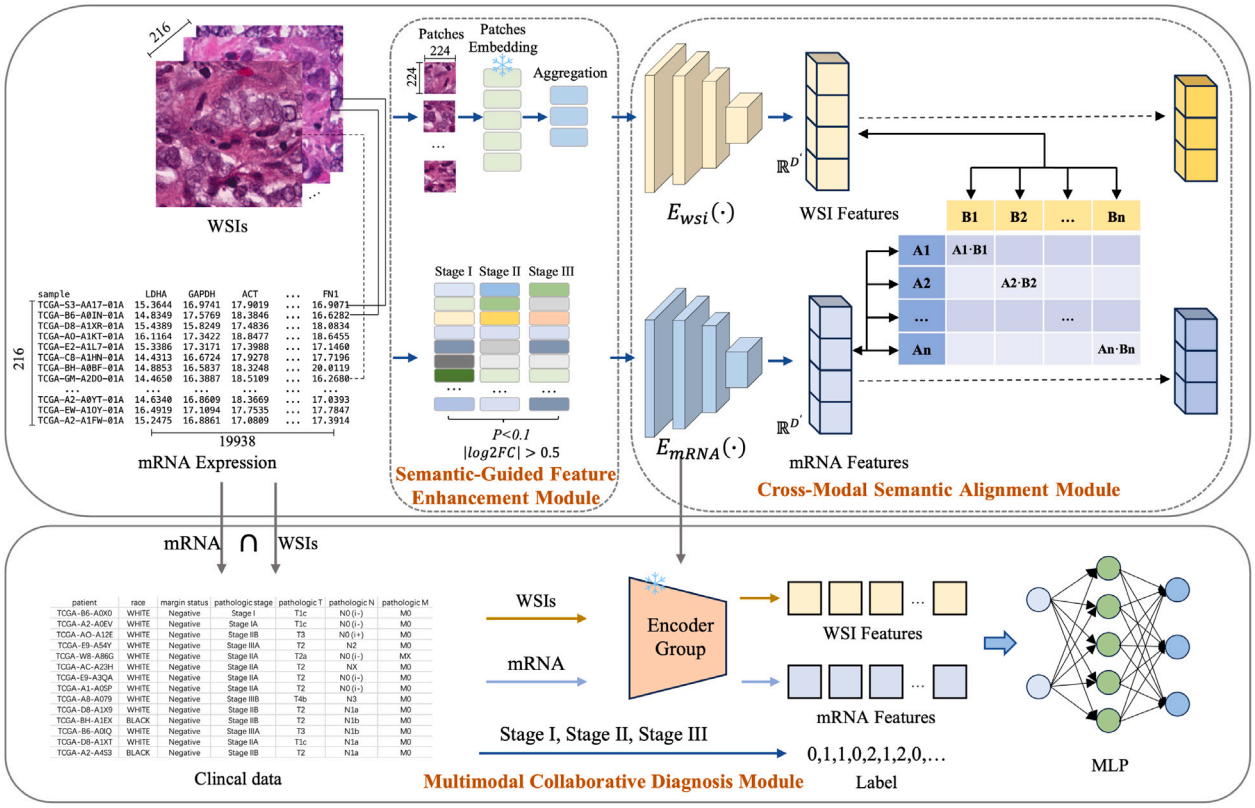

Fig. 1. Overview of the FESCA model and training strategy. The model includes three main modules: semantic-guided feature enhancement module, cross-modal semantic alignment module, and multimodal collaborative diagnosis module.

semantic-guided feature extraction modules for both the pathology imaging modality and the genomic modality, reinforcing semantic representations within each modality from the early stages of the model and providing high-quality, discriminative embeddings for subsequent cross-modal semantic alignment. The mechanism consists of two components: lesion-guided image feature embedding and differential-gene-driven mRNA feature extraction.

Lesion-guided image feature embedding: To address the computational challenges of utilizing all pixels in WSIs as input features, we adopt a Lesion-guided image feature embedding method. Specifically, the ultra-high-resolution WSI is first partitioned into patches and embedded into compact local representations, substantially reducing computational overhead while retaining critical pathological information. Subsequently, patches containing lesion regions are identified, and their deep feature representations are extracted to capture discriminative cues associated with tumor development. Finally, the lesion-related features are aggregated to derive a patient-level WSI representation, thereby establishing a coherent and discriminative foundation for subsequent cross-modal semantic alignment.

In the patch extraction and embedding stage, we employ a sliding window approach with a 112-pixel overlap to crop image patches of size $W_p \times H_p \times C_p$, where W_p, H_p, and C_p denote the width, height, and number of channels of each patch, respectively. Then, to effectively aggregate the numerous patch-level features into a compact slide-level representation for cosine similarity computation, we introduce a Transformer-based patch decoder. This decoder reduces the dimensionality of each patch feature from $W_p \times H_p \times C_p$ to $D \times 1$ to feature vector of size $D \times 1$, where D is the feature dimension. We utilize a ViT model pretrained on large-scale datasets such as ImageNet [40], which has demonstrated strong generalization capabilities when transferred to domain-specific tasks, including WSI analysis. The use of ViT substantially enhances the accuracy and robustness of the extracted patch features [41], making it a suitable backbone for our feature extraction pipeline. The specific implementation process is as follows:

Given an input image $I \in \mathbb{R}^{224\times224\times3}$, the pretrained ViT model constructs a sequence of 196 non-overlapping 16 × 16 patches. These patches are projected into token embeddings, and to capture their spatial information, positional embeddings $P \in \mathbb{R}^{196\times768}$ are incorporated. The resulting position-encoded features are formulated in Eq. (1).

$$X_{patch}^{pos} = X_{patch} + P \tag{1}$$

Each position-encoded patch feature x_i^{pos} undergoes a linear transformation to map it into an embedding space. Let the weight matrix for the linear transformation be $W \in \mathbb{R}^{768\times D}$, where D ($D = 768$) is the dimension of the embedding space. The transformed features are shown in Eq. (2).

$$E_{patch} = X_{patch}^{pos} W \tag{2}$$

Finally, each patch is flattened into a D-dimensional vector, so the patch feature matrix is shown in Eq. (3).

$$X_p = [x_{cls}; x_1, x_2, \ldots, x_{196}], x_i \in \mathbb{R}^D \tag{3}$$

where, x_{cls} is the embedding of the [CLS] token, which has a dimensionality of D. $x_1, x_2, \ldots, x_{196}$ represent the embeddings of the 196 patches, each with a dimensionality of D.

$$f_{ij} = \text{Transformer}(X_p^{ij}) \in \mathbb{R}^D \tag{4}$$

where X_p^{ij} denotes the jth patch of the ith WSI image, and f_{ij} represents the corresponding feature vector extracted from X_p^{ij}.

Based on the large number of patch-level features obtained from the patch extraction and embedding, we employed an attention-based multiple instance learning (MIL) model to identify patches containing lesion-related information. This model performs weakly supervised classification of patches as cancerous or non-cancerous, enabling the

selection of patches that are most relevant to breast cancer and reducing the impact of irrelevant regions on downstream classification tasks. To derive sample-level image representations, we aggregated the patch features using mean pooling, resulting in a compact and informative feature vector for each sample. The calculation formula is as follows:

$$F_p = \frac{1}{N_p}\sum_{j=1}^{N_p} f_{ij} \in \mathbb{R}^D \tag{5}$$

where F_p denotes the feature vector of the pth WSI, the N represents the number of patches extracted from the pth WSI by the MIL model.

Differential-gene-driven mRNA feature extraction: A common challenge in medical research is the Large-p, Small-n problem, where the number of input features (e.g., genes) is much greater than the number of available samples. This imbalance can lead to the so-called curse of dimensionality, which hinders the generalization ability of machine learning models. To address this issue and identify representative signature genes that are specifically associated with cancer, we perform differential expression analysis across stage I, stage II, and stage III samples.

There are several methods available for identifying significantly differentially expressed genes between groups. Considering the type of dataset that was used in this study, we selected the Differential Expression analysis based on the Negative Binomial distribution 2 (DESeq2) [42] as the differential analysis method in our study.

Assume the mRNA expression data is a $N \times M$ matrix, where N represents the number of samples and M represents the number of genes. Let $x_i (1 \le i \le N)$ represents ith sample in the dataset and $G = \{g_1, g_2, \ldots, g_M\}$ denotes the gene set. To correct for technical variability and ensure reliable downstream differential expression analysis, DESeq2 estimates the library size factors and gene-wise scaling factors. Specifically, the library size factor of sample x_i, computed using Eq. (6), accounts for differences in sequencing depth across samples. The gene-specific size factor for gene g_j, calculated using Eq. (7), captures expression variability across all samples. Finally, the standardized expression level of gene g_j in sample x_i, given by Eq. (8), is obtained by normalizing the raw read count $Count(x_i g_j)$ with both sample-wise and gene-wise size factors.

$$s_{x_i} = \frac{Count(x_i)}{\exp\left(\frac{1}{N}\sum_{i=1}^{N}\log\left(Count(x_i)\right)\right)} \tag{6}$$

$$s_{g_j} = \frac{\sum_{i=1}^{N} Count(x_i)}{\exp\left(\frac{1}{M}\sum_{j=1}^{M}\log\left(\sum_{i=1}^{N} Count(x_i)\right)\right)} \tag{7}$$

$$Stand_{Count_{ij}} = \frac{Count(x_i g_j)}{\frac{s_{x_i}}{\exp\left(\frac{1}{N}\sum_{i=1}^{N}\log(s_{x_i})\right)}} \tag{8}$$

where, s_{x_i} represents the library size factor of sample i, s_{g_i} represents the size factor of gene g_i, $Stand_{Count_{ij}}$ represents the standardized expression of each gene g_i, and $Count(x_i)$ represents the read count of sample x_i.

Following normalization, DESeq2 estimates the dispersion (variance) of gene counts across samples and applies an empirical Bayesian approach to shrink these estimates toward the overall average dispersion. The count data are modeled using a negative binomial distribution, and statistical testing is performed using the Wald test. For each gene, the Wald statistic is calculated based on the estimated log2 fold change, and a p-value is derived from the chi-square distribution (χ^2). Differentially expressed genes were selected based on the criteria of $|\log_2 \mathrm{FC}| > 1$ and an adjusted p-value < 0.1 (as shown in (9)), and their expression patterns across stage I, II, and III samples are visualized as a heatmap in Fig. 2.

$$G_{\text{signature}} = \{g_j : |\log_2(FC_j)| > 1 \text{ and } p_j < 0.1\} \tag{9}$$

where, $G_{signature}$ represents the obtained differential gene set, which serves as a key driving feature for subsequent cross-modal alignment and cancer diagnosis.

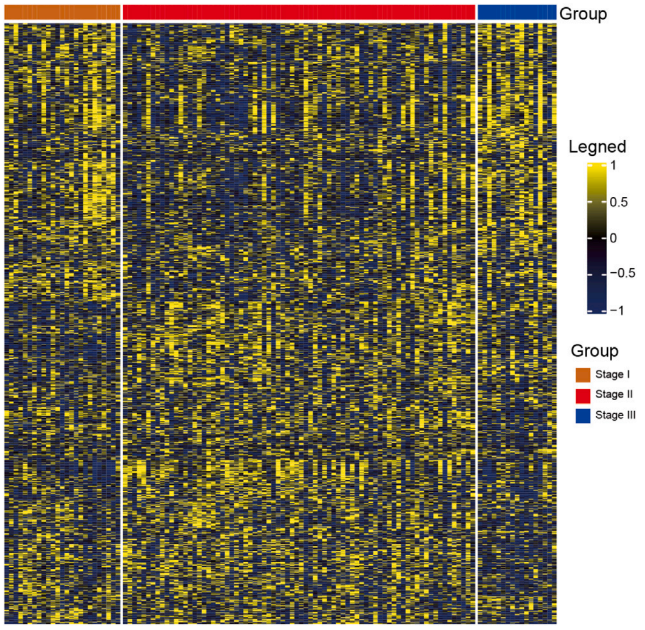

Fig. 2. Heatmap of signature genes. A heatmap illustrates the expression patterns of 889 signature genes, selected from a total of 18,031 genes, across samples from stage I, stage II, and stage III.

3.3. Cross-modal semantic alignment module

To achieve deep semantic collaboration between pathology imaging and genomic features, we propose a contrastive-learning-based cross-modal semantic alignment framework. The proposed framework first employs modality-specific projection heads to unify the dimensionality of features across modalities. These features are then projected into a shared semantic space via a contrastive learning scheme, which maximizes cross-modal consistency for the same patient while suppressing semantic discrepancies across different patients. This procedure effectively strengthens the semantic coherence and discriminative power of the multimodal fusion framework.

Specifically, based on the WSI and gene embeddings obtained through image and mRNA feature extraction, respectively, we employ two modality-specific deep projection heads based on Deep Neural Networks (DNNs) [43] to perform cross-modal alignment learning. These projection heads are defined as $E_{\text{wsi}}(\cdot)$ for WSIs and $E_{\text{mRNA}}(\cdot)$ for mRNA expression data, respectively, as follows:

$$F' = E_{\text{wsi}}(F) = \begin{cases} h_1 = \phi(W_1 x + b_1), W_1 \in \mathbb{R}^{512\times D},\ b_1 \in \mathbb{R}^{512} \\ h_2 = \phi(W_2 h_1 + b_2), W_2 \in \mathbb{R}^{256\times 512},\ b_2 \in \mathbb{R}^{256} \\ h_3 = \phi(W_3 h_2 + b_3), W_3 \in \mathbb{R}^{D'\times 256},\ b_3 \in \mathbb{R}^{D'} \end{cases} \tag{10}$$

$$G' = E_{mRNA}(G_{\text{signature}}) = \begin{cases} h_1 = \phi(W_1 x + b_1), W_1 \in \mathbb{R}^{1024\times D},\ b_1 \in \mathbb{R}^{1024} \\ h_2 = \phi(W_2 h_1 + b_2), W_2 \in \mathbb{R}^{512\times 1024},\ b_2 \in \mathbb{R}^{512} \\ h_3 = \phi(W_3 h_2 + b_3), W_3 \in \mathbb{R}^{256\times 512},\ b_3 \in \mathbb{R}^{256} \\ h_4 = \phi(W_4 h_3 + b_4), W_4 \in \mathbb{R}^{128\times 256},\ b_4 \in \mathbb{R}^{128} \\ z = W_5 h_4 + b_5, W_5 \in \mathbb{R}^{D'\times 128},\ b_5 \in \mathbb{R}^{D'} \end{cases} \tag{11}$$

where $\phi(\cdot)$ denotes the ReLU activation function. F denotes the feature vectors extracted from WSIs, $G_{signature}$ denotes the signature gene vectors derived from mRNA expression data, and D' represents the output dimensionality of both $E_{wsi}(\cdot)$ and $E_{mRNA}(\cdot)$.

A contrastive loss function is applied to measure the similarity between the outputs of $E_{wsi}(\cdot)$ and $E_{mRNA}(\cdot)$. Eqs. (12), (13), and (14) represent the loss function used during the model training. In training we maximize $\cos(S_i^{F'}, S_i^{G'})$, while minimizing $\cos(S_i^{F'}, S_j^{G'})$, where $i \neq j$.

$$\mathcal{L}_{pos}^{WSI}(i) = -\log \frac{\exp\left(cos(S_i^{F'}, S_i^{G'})/\tau\right)}{\sum_{j=1}^{N} \exp\left(cos(S_i^{F'}, S_j^{G'})/\tau\right)} \tag{12}$$

$$\mathcal{L}_{pos}^{mRNA}(i) = -\log \frac{\exp\left(cos(S_i^{G'}, S_i^{F'})/\tau\right)}{\sum_{j=1}^{N} \exp\left(cos(S_i^{G'}, S_j^{F'})/\tau\right)} \tag{13}$$

$$\mathcal{L}_{contrast} = \frac{1}{2N} \sum_{i=1}^{N} \left(\mathcal{L}_{pos}^{WSI}(i) + \mathcal{L}_{pos}^{mRNA}(i)\right) \tag{14}$$

where $cos(\cdot)$ denotes the cosine similarity, $S_i^{F'}$ represents the embedding features of the ith sample obtained from $E_{wsi}(\cdot)$, $S_i^{G'}$ represents the embedding feature of ith sample from $E_{mRNA}(\cdot)$, and τ denotes the temperature parameter to controlling the smoothness of softmax function.

3.4. Multimodal collaborative diagnosis module

To enable precise diagnostic modeling for cancer patients, we introduce a modality-adaptive fusion strategy that dynamically adjusts the feature integration process based on the availability of input modalities, thereby fully leveraging the complementary strengths of multimodal representations. Specifically, the strategy adapts the feature encoding pathway according to modality accessibility: when only a single modality is available, the model directly utilizes the embedding generated by the corresponding encoder, preserving its discriminative power and stability in unimodal diagnostic tasks. When multiple modalities are available, the model employs an attention-based fusion mechanism to adaptively integrate WSI and mRNA features, enabling modality-aware interaction and hierarchical feature aggregation, and ultimately capturing richer cross-modal semantic relationships. This adaptive fusion strategy not only mitigates performance degradation caused by missing modalities but also enhances the model's flexibility and robustness across diverse data conditions.

When only a single modality is available, let X_i^{wsi} and X_i^{rna} denote the WSI and mRNA inputs of the ith sample, respectively. The image or gene embedding representations are obtained as shown in Eqs. (15) and (16).

$$Z_i^{wsi} = E_{wsi}(X_i^{wsi}) \tag{15}$$

$$Z_i^{rna} = E_{mRNA}(X_i^{rna}) \tag{16}$$

When both WSI and mRNA data are available, we employ a fixed-weight fusion strategy in which the modality-specific encoders first produce the embeddings Z_i^{wsi} and Z_i^{rna}, and the two representations are subsequently integrated using equal-weight averaging. This design avoids introducing additional learnable fusion parameters and provides a simple yet stable mechanism for integrating heterogeneous modalities.

$$\alpha_i^{rna} = 1 - \alpha_i^{wsi}, \alpha_i^{wsi} = 0.5 \tag{17}$$

$$Z_i = \alpha_i^{wsi} Z_i^{wsi} + \alpha_i^{rna} Z_i^{rna} \tag{18}$$

where, α_i^{wsi} and α_i^{rna} represent the weights of WSI and mRNA features, respectively.

Subsequently, the fused feature Z_i is fed into an MLP classifier to predict the cancer category, as defined in Eq. (19). The model is trained by minimizing the loss function (20).

$$\hat{y}_i = f_{mlp}(Z_i) \tag{19}$$

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c} \tag{20}$$

where $\hat{y}_i \in \mathbb{R}^C$ denotes the predicted class–probability vector for the ith sample, $f_{\text{mlp}}(\cdot)$ is the MLP classifier mapping the embedding Z_i to the probability space, $\mathcal{L}_{cls}$ is the classification loss, $y_{i,c}$ is the one-hot ground-truth label, and $\hat{y}_{i,c}$ is the predicted probability for class c.

To further assess the clinical applicability and diagnostic performance of the FESCA method, we developed a unified multimodal embedding-based classification system. The system is designed to flexibly support inputs from two modalities, WSI or mRNA expression data, by dynamically selecting the corresponding encoder from an encoder set to project them into a shared embedding space, followed by classification using an MLP classifier. The system ultimately provides an automated prediction of breast cancer pathological staging, offering a scalable and effective tool for clinical decision support.

All modules described above are trained under a unified optimization protocol to ensure stability and reproducibility. All experiments were conducted in R and Python environments. Models were optimized using the Adam optimizer (learning rate 1×10^{-4}), with a batch size of 16 and early stopping within 100 epochs. Regularization strategies were applied, and all hyperparameters were selected based on validation performance and kept consistent across experiments.

3.5. Evaluation criterion

Given the class imbalance in our dataset, we selected evaluation metrics that are robust to this characteristic. We use the Area Under the Precision–Recall Curve (AUC-PR), which offers a more reliable performance measure for the minority class than ROC-AUC [44]. Additionally, we report weighted precision, weighted recall, and weighted F1-score. These weighted averages account for class imbalance by weighting each class's metric by its sample size, preventing the majority class from skewing the overall performance evaluation.

$$P_i = \frac{TP_i}{TP_i + FP_i}, \quad P_{weighted} = \frac{\sum_{i=1}^{C}(P_i \cdot n_i)}{\sum_{i=1}^{C} n_i} \tag{21}$$

$$R_i = \frac{TP_i}{TP_i + FN_i}, \quad R_{weighted} = \frac{\sum_{i=1}^{C}(R_i \cdot n_i)}{\sum_{i=1}^{C} n_i} \tag{22}$$

$$F1_i = \frac{2 \cdot (P_i \cdot R_i)}{P_i + R_i}, \quad F1_{weighted} = \frac{\sum_{i=1}^{C}(F1_i \cdot n_i)}{\sum_{i=1}^{C} n_i} \tag{23}$$

where, TP_i represents True Positives, FP_i represents False Positives, and FN_i represents False Negatives, corresponding to the elements of the confusion matrix for class i. C represents the total number of classes, and n_i represents the number of samples in class i.

4. Results

4.1. Comparisons with state-of-the-art

To comprehensively evaluate the performance of FESCA, we reproduced several representative unimodal and multimodal models as baseline methods for comparison.

WSI-Based Unimodal Models:

- ABMIL [19]: It is a representative attention-based MIL model that is widely used in pathological image analysis.
- TransMIL [13]: It is a Transformer-based MIL model widely applied in WSI-based pathology analysis.
- Low-rank MIL [17]: This model enhances feature compactness and generalization through low-rank constraints.
- AlexNet-BC [45]:This model extracts morphological and semantic tissue patterns to enable effective WSI classification.
- Xception [46]: This model performs efficient WSI feature extraction with reduced computational cost.

- InceptionV3 [47]: This model performs efficient multi-scale feature extraction for WSI representation learning.
- SEDenseNet [48]: This model improves WSI feature representation and classification performance through channel-wise feature recalibration.

mRNA-Based Unimodal Models:

- RF [49]: It is a classical method widely used for gene expression data classification.
- XGBoost [50]: It is an efficient ensemble learning method commonly applied to genomic data analysis.
- LightGBM [51]: It is a highly efficient gradient boosting framework that performs well on large-scale and high-dimensional genomic datasets.
- gcForest [52]: It is a cascade forest model with strong robustness and adaptability for small-sample, high-dimensional data.
- TabNet [27]: It is a model that dynamically selects the most informative features at each decision step for efficient feature utilization.
- TabTransformer [37]: This model automatically learns feature interactions without manual feature engineering, improving tabular data classification.
- SAINT [53]: This model combines self-attention and inter-sample attention to capture both feature dependencies and sample relationships in tabular data.

WSI–mRNA Multimodal Models:

- MCAT [18]: A representative multimodal diagnostic model known for its effective feature fusion and semantic interaction mechanisms, and thus selected as a baseline.
- MOTCat [33]: A model that introduces a cross-modal Transformer architecture to further enhance semantic interaction and feature fusion across modalities, making it a suitable Transformer-based baseline for comparison.

Based on the aforementioned baseline models, we systematically evaluate FESCA across different modalities under consistent experimental settings.

Comparison with WSI-based Unimodal Models: To evaluate the semantic extraction capability and discriminative feature representation of FESCA within the pathology image modality, we conducted a systematic comparison with WSI-based unimodal models. The experimental results are summarized in Table 5, where $\text{FESCA}_{\text{WSI}}$ denotes the FESCA model using only the WSI modality.

Specifically, $\text{FESCA}_{\text{WSI}}$ achieves the highest AUC-PR (0.796), exceeding the best-performing baseline, Xception (0.782). This improvement demonstrates that FESCAWSI can better identify subtle lesion-related patterns, particularly in minority-class samples. Furthermore, the balanced precision (0.722) and recall (0.778), together with an F1-score of 0.720, suggest that the semantic-guided feature enhancement mechanism enables the model to extract more reliable and meaningful pathological representations. In comparison, CNN-based baselines exhibit clear trade-offs between precision and recall, and Transformer-based methods such as TransMIL, although achieving high recall, still fail to match the overall discriminative performance of $\text{FESCA}_{\text{WSI}}$. Traditional MIL and shallow CNN methods show even larger performance gaps, underscoring their limitations in capturing high-level semantic patterns from gigapixel WSIs.

These findings collectively confirm that $\text{FESCA}_{\text{WSI}}$ not only yields superior classification performance but also demonstrates enhanced intra-modality semantic extraction and feature representation capabilities compared with existing WSI-based unimodal approaches.

Comparison with mRNA-based Unimodal Models: To evaluate the semantic modeling capability and discriminative feature extraction of FESCA within the genomics modality, we conducted a systematic comparison with several classical gene expression-based classification models. The experimental results are presented in Table 6, where $\text{FESCA}_{\text{mRNA}}$ denotes the FESCA variant that uses only mRNA data.

Table 5
Performance comparison between the FESCA model and seven baseline algorithms on the test data of WSIs.

Model	AUC-PR	Precision	Recall	F1-score
ABMIL	0.430	0.277	0.526	0.363
SEDenseNet	0.560	0.399	0.632	0.489
Xception	0.782	0.493	0.611	0.495
InceptionV3	0.682	0.468	0.684	0.556
AlexNet-BC	0.675	0.592	0.632	0.611
Low-rank MIL	0.638	0.573	0.737	0.640
TransMIL	0.697	0.709	**0.842**	**0.770**
$\text{FESCA}_{\text{WSI}}$	**0.796**	**0.722**	0.778	0.720

Table 6
Performance comparison of Baselines and FESCA models on mRNA data.

Model	AUC-PR	Precision	Recall	F1-score
XGBoost	0.460	0.389	0.389	0.389
SAINT	0.626	0.556	0.389	0.456
TabNet	0.552	0.458	0.611	0.524
gcForest	0.643	0.444	0.667	0.533
TabTransformer	0.574	0.444	0.667	0.533
RF	0.473	0.471	0.668	0.552
LightGBM	0.511	0.471	0.668	0.552
SVM	0.587	0.693	0.722	0.641
$\text{FESCA}_{\text{mRNA}}$	**0.816**	**0.850**	**0.833**	**0.839**

Specifically, $\text{FESCA}_{\text{mRNA}}$ achieves an AUC-PR of 0.816, yielding a notable 17.3% improvement over the strongest baseline, gcForest (0.643). This demonstrates its superior ability to capture subtle, disease-related signals embedded in high-dimensional gene expression profiles. The model also attains the highest precision (0.850) and recall (0.833), indicating that its learned representations not only better distinguish positive samples but also effectively suppress irrelevant variation, as further evidenced by its F1-score of 0.839. In contrast, ensemble models such as XGBoost, RF, and LightGBM struggle with both precision and recall, implying that tree-based decision boundaries are insufficient for modeling complex nonlinear gene interactions. Deep tabular models (TabNet, TabTransformer, SAINT) show moderate gains yet still fall well short of $\text{FESCA}_{\text{mRNA}}$, suggesting that feature-level attention or local feature selection alone cannot fully capture gene–gene dependencies or latent biological semantics.

In conclusion, these results demonstrate that FESCA's semantic-guided feature enhancement mechanism effectively identifies disease-associated gene groups and constructs a more discriminative and biologically meaningful embedding space, rather than relying on shallow statistical correlations.

Comparison with WSI–mRNA multimodal Models: To evaluate the multimodal fusion performance of the proposed FESCA model, we compare $\text{FESCA}_{\text{WSI}}$, $\text{FESCA}_{\text{mRNA}}$, and the full FESCA model with two WSI–mRNA multimodal baselines (as shown in Table 7). In the full FESCA model, the input is the fused feature vector derived from the outputs of the Image Encoder and the mRNA Encoder.

The results show that the proposed FESCA model consistently outperforms both multimodal models across all evaluation metrics. FESCA achieves the highest AUC-PR (0.829), Precision (0.856), and F1-score (0.822), demonstrating clear advantages in predictive accuracy and overall classification performance. Although $\text{FESCA}_{\text{mRNA}}$ reaches the same Recall (0.833), the full multimodal model achieves stronger overall results by effectively leveraging complementary cross-modal information. In contrast, MCAT and MOTCat exhibit pronounced imbalances between precision and recall, suggesting that their fusion strategies fail to fully exploit cross-modal complementary cues, thereby limiting their semantic interaction capabilities. These findings validate the effectiveness of FESCA's cross-modal alignment and fusion strategy, highlighting

Table 7
Performance comparison of classification performance among multimodal baseline models and the proposed FESCA model on test data.

Model	AUC-PR	Precision	Recall	F1-score
MCAT	0.746	0.719	0.278	0.187
MOTCat	0.746	0.759	0.333	0.362
$FESCA_{WSI}$	0.796	0.722	0.778	0.720
$FESCA_{mRNA}$	0.816	0.850	**0.833**	**0.839**
FESCA	**0.829**	**0.856**	**0.833**	0.822

Table 8
Comparison of ablation experiments for the FESCA on test data.

Model	AUC-PR	Precision	Recall	F1-score
$Single_{WSI}$	0.705	0.500	0.333	0.370
$FESCA_{WSI}$	0.796	0.722	0.778	0.720
$Single_{mRNA}$	0.676	0.432	0.500	0.463
$FESCA_{mRNA}$	0.816	0.850	**0.833**	**0.839**
FESCA	**0.829**	**0.856**	**0.833**	0.822

its capability to more fully capture and utilize complementary information from WSI and mRNA data, and demonstrating its strong potential for complex multimodal medical analysis.

4.2. Ablation experiment

To assess the contributions of the semantic-guided feature enhancement module and cross-modal collaboration module in FESCA, we conducted ablation studies from both unimodal and multimodal perspectives. Specifically, we constructed two unimodal variants using only the pathology image modality ($Single_{WSI}$) and only the genomic modality ($Single_{mRNA}$), respectively. These models were systematically compared with their corresponding enhanced versions ($FESCA_{WSI}$ and $FESCA_{mRNA}$) as well as with the full FESCA model. The experimental results are summarized in Table 8.

The results in Table 8 indicate that incorporating the proposed semantic-guided and cross-modal collaboration mechanisms yields substantial performance improvements across both modalities. For the WSI modality, FESCAWSI surpasses SingleWSI on all evaluation metrics, with AUC-PR increasing from 0.705 to 0.796 and F1-score from 0.370 to 0.720. This improvement reflects the effectiveness of lesion-prior-driven semantic enhancement and contrastive alignment in enabling more reliable extraction of discriminative pathological patterns. For the mRNA modality, FESCAmRNA likewise demonstrates notable gains, achieving an approximately 20.7% increase in AUC-PR and improving the F1-score from 0.463 to 0.839 compared with SinglemRNA. These results highlight the benefit of incorporating differential-gene priors and cross-modal alignment to obtain more discriminative and noise-resilient genomic representations. Finally, the complete FESCA model further improves multimodal performance, achieving an AUC-PR of 0.829 and a Precision of 0.856, outperforming both unimodal variants. This enhancement is primarily attributable to the attention-based fusion mechanism, which enables more effective integration of complementary information from WSIs and mRNA profiles.

In conclusion, the ablation studies provide strong empirical evidence for the effectiveness of both the semantic-guided feature enhancement module and the cross-modal collaboration module within FESCA. These two components contribute fundamentally to the framework: the former substantially improves intra-modality semantic representation learning, while the latter markedly strengthens multimodal fusion and cross-modal information integration.

Furthermore, to validate the effectiveness of using ViT as the feature extractor in the lesion-guided image feature embedding module, we conducted an additional ablation study. Under the same dataset and parameter settings, the ViT encoder in the original model was replaced with two commonly used convolutional neural network backbones (ResNet-50 and EfficientNet-B0) and the extracted features were used to train the attention-based MIL model. Subsequently, MIL prediction results on different image patches were visualized for comparison (as shown in Fig. 3). The experimental results show that the ViT-based feature extractor learns more discriminative and semantically enriched representations from WSI patches, achieving significantly higher accuracy and robustness in lesion-region identification compared with ResNet-50 and EfficientNet-B0. These findings indicate that ViT is more effective at capturing diagnostically meaningful patch-level features from WSIs and provides higher-quality inputs for subsequent semantic representation learning and classification tasks. This further confirms both the value and necessity of employing ViT within the FESCA framework.

4.3. Multimodal correlation analysis of radiology images and mRNA expression profiles

To evaluate the performance of the multimodal alignment module between WSIs and mRNA expression data, the cosine similarity between the encoded vectors of image and mRNA data obtained through the FESCA model was visualized. Fig. 4A shows the distributions of self-similarity and cross-similarity on both the training and test sets. Specifically, self-similarity quantifies the similarity between paired image and mRNA representations, while cross-similarity captures the similarity between randomly paired samples across modalities. The results reveal that, in both the training and test sets, the similarity between paired data (self-similarity) is significantly higher than that of unpaired data (cross-similarity), indicating an effective alignment of multimodal representations. Furthermore, in the independent test set, 87.5% of the samples exhibit a strong correlation between their WSI and mRNA feature representations derived from the multimodal alignment module, with a mean similarity of 0.6737 and a median similarity of 0.7004, suggesting that the learned multimodal representations maintain stable alignment behavior on unseen samples (as shown in Fig. 4B). The high similarity indicates that FESCA effectively captures semantic consistency between histopathological and mRNA feature representations.

To provide an intuitive representation-level perspective, Fig. 5 visualizes the feature distributions before and after multimodal alignment using t-SNE. As shown in Fig. 5A, unaligned WSI and mRNA representations exhibit disparate geometric structures, reflecting limited correspondence between modalities. In contrast, after alignment by FESCA (Fig. 5B), the two modalities display more consistent structural patterns in the embedding space. Importantly, the mean paired distance between corresponding WSI–mRNA representations is reduced from 38.25 (unaligned) to 19.18 after alignment, quantitatively supporting the enhanced cross-modal consistency observed in the visualization.

In summary, the experimental results provide evidence that FESCA not only enables effective mapping of WSI-derived image features and mRNA expression features into a unified semantic representation space, but also markedly strengthens the semantic coherence and complementarity between the two modalities. These aligned multimodal representations provide a favorable basis for downstream classification tasks.

4.4. System application

The FESCA framework has been further extended for breast cancer pathological staging tasks and is referred to as the Multimodal Data-Based Model for Breast Cancer Diagnosis (MMBCD). The trained MMBCD model has been deployed on an online platform http://mmbcd.ucd.ie/MMBCD/, forming an intelligent prototype system for breast cancer pathological staging diagnosis.

The system integrates histopathological data from WSIs and genomic mRNA expression data through modality-specific deep projection heads. These heads map both data types into a shared semantic

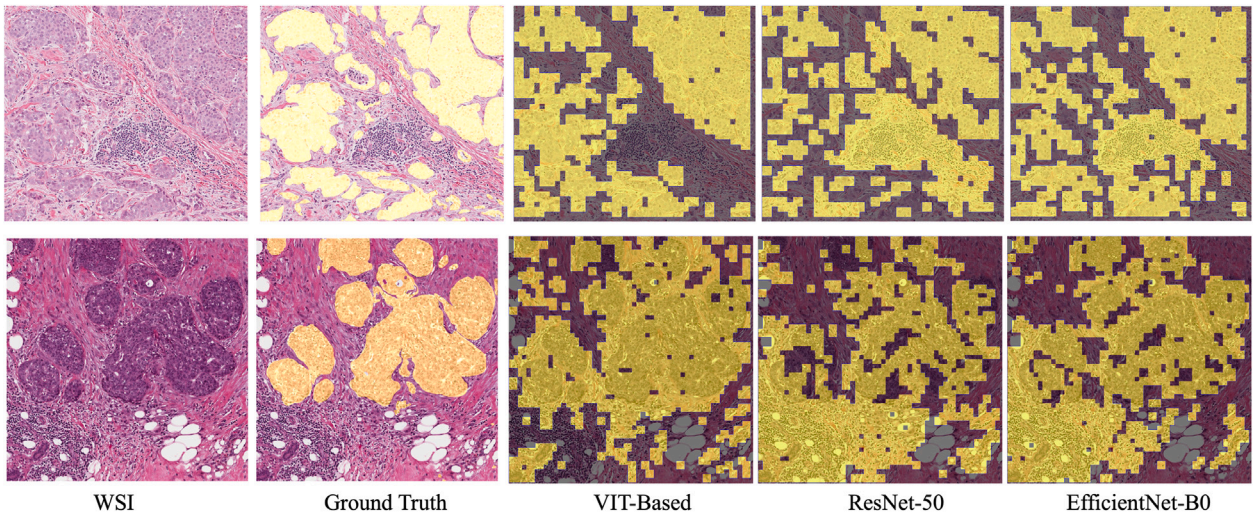

Fig. 3. Qualitative visualization of lesion-guided attention maps generated by the Attentional Multi-Instance Learning model using features extracted from different pretrained encoders, including an ImageNet-pretrained ViT, ResNet-50, and EfficientNet-B0. Regions with higher cancer-related attention are highlighted in yellow. This visualization is provided for interpretability purposes to illustrate how different feature extractors attend to lesion-related regions in high-resolution WSIs.

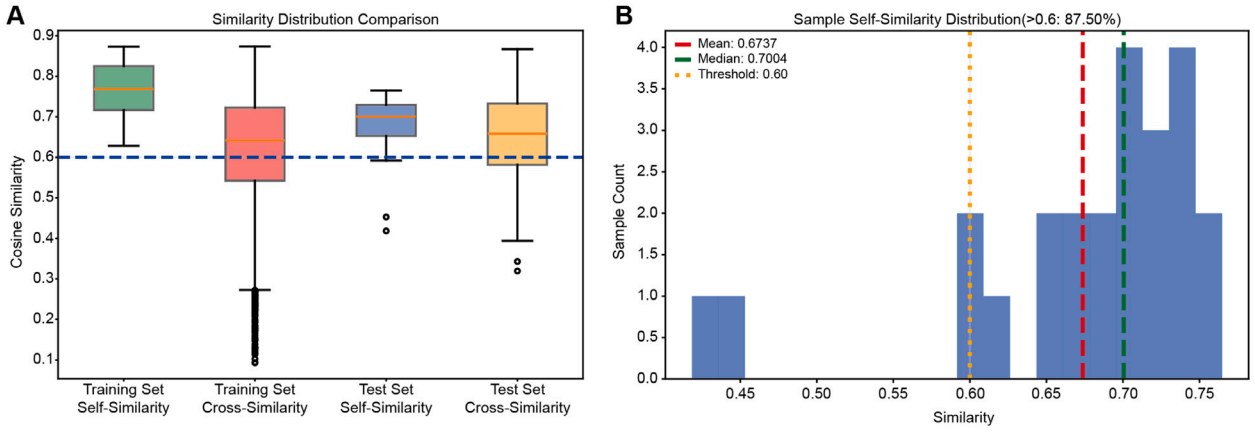

Fig. 4. Cosine similarity statistics between the encoded vectors of image and mRNA data obtained using the FESCA model. A. t-SNE visualization of unaligned WSI and mRNA representations, showing limited shared structure between the two modalities. B. t-SNE visualization of representations aligned by FESCA, where WSI and mRNA exhibit more coherent structural patterns, indicating improved cross-modal semantic alignment.

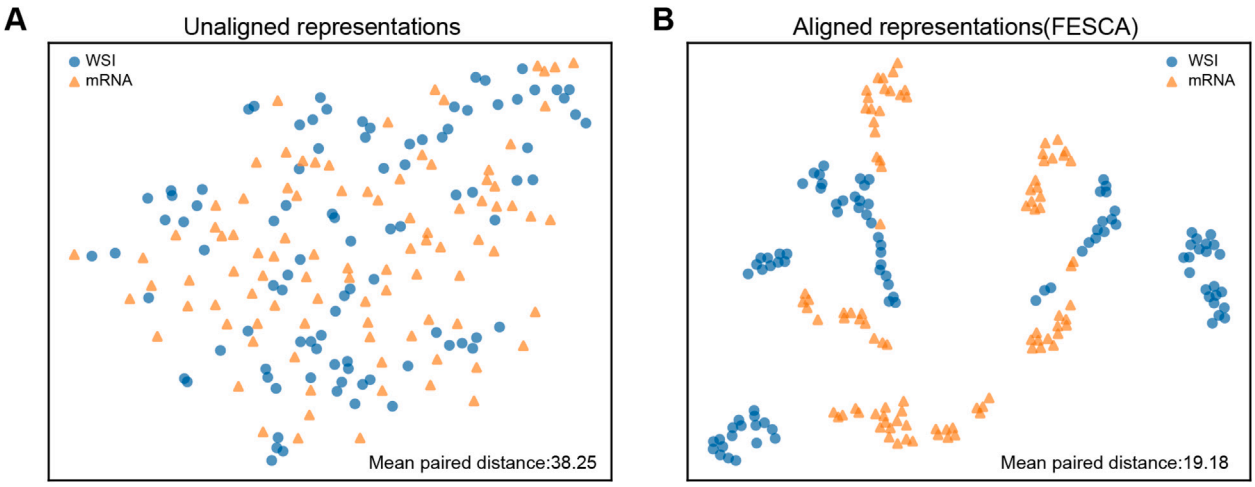

Fig. 5. Visualization of cross-modal representation alignment before and after semantic alignment. A. Box plot illustrating the cosine similarity distributions between image and mRNA feature representations in the training and test sets. B. Bar chart showing the similarity distribution in the test set.

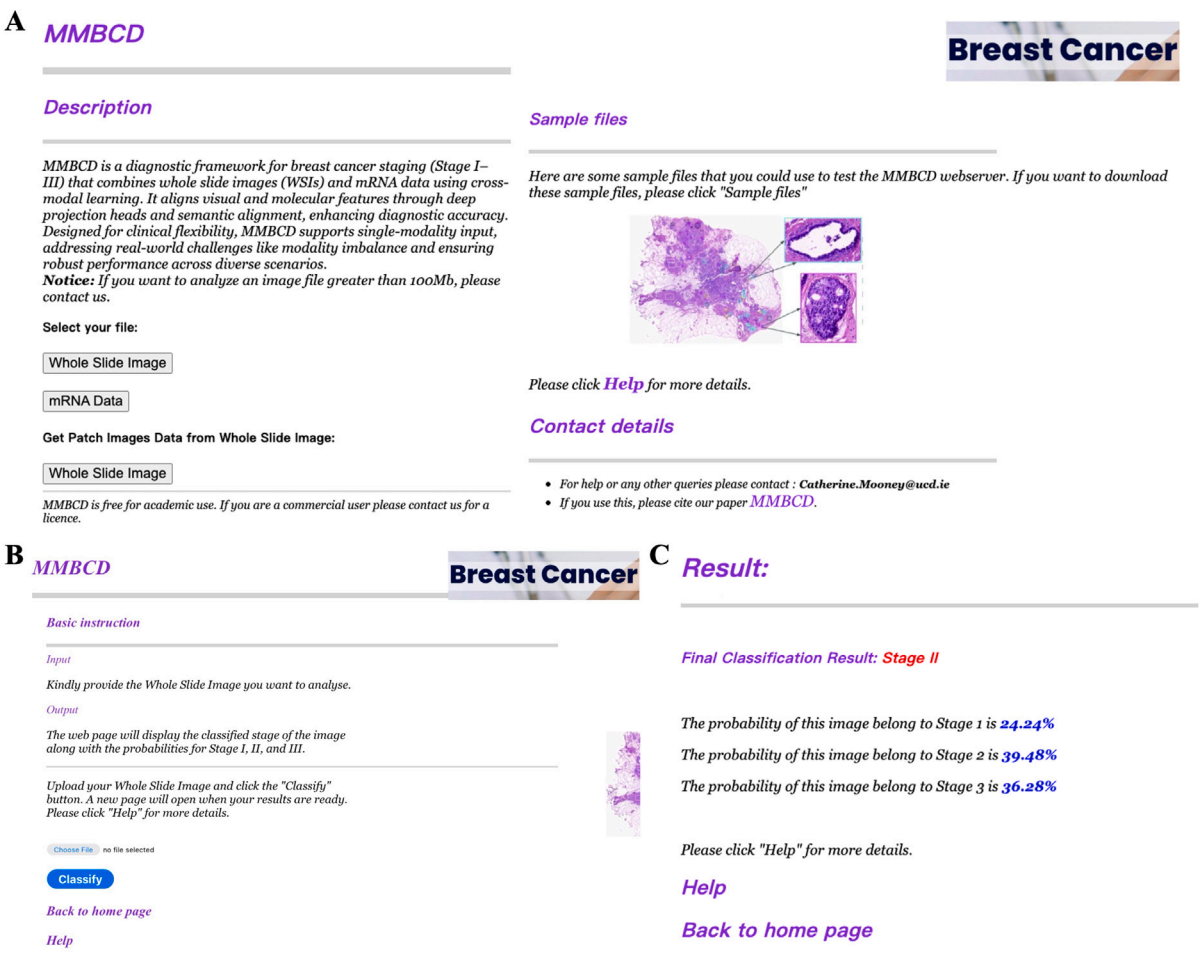

Fig. 6. User interface of the MMBCD system. A. Main interface of the system, which supports input of either WSIs or mRNA expression data. B. Interface for image input, allowing users to upload WSI images from local storage. C. Output interface displaying the predicted pathological stage of the patient.

space, facilitating alignment. Advanced cross-modal semantic alignment techniques resolve the semantic inconsistencies between visual (WSI) and molecular (mRNA) modalities, thereby enhancing diagnostic accuracy and reliability.

Users can select between two input options: a WSI file (e.g., .png, or .jpg) can be uploaded on the image classification page; for mRNA input, an expression data files (e.g., .csv, or .txt) can be uploaded on the mRNA classification page. After uploading, users click "Classify", and the system processes the data to return cancer stage probabilities (as shown in Fig. 6). This ensures that the system can adapt to evolving research needs, marking a step toward clinical application. However, its deployment in clinical settings would require extensive validation through large-scale clinical trials.

5. Discussion

This study proposes FESCA, a semantic-guided multimodal framework that effectively integrates WSI features and mRNA expression profiles for breast cancer stage classification. The experimental results consistently demonstrate the superiority of FESCA over existing unimodal and multimodal approaches. These improvements highlight the importance of semantic-guided feature enhancement and cross-modal alignment in addressing the inherent heterogeneity and complexity of medical multimodal data.

Compared with CNN-based WSI models, the lesion-guided image feature embedding method enables FESCA to capture more discriminative and fine-grained pathological patterns. Similarly, the incorporation of differential-gene-guided semantic priors allows the model to extract biologically meaningful genomic representations, leading to clear advantages over traditional ensemble methods and tabular deep learning approaches. Furthermore, the cross-modal alignment module reinforces semantic coherence between modalities, enabling FESCA to exploit complementary diagnostic information more effectively than existing multimodal fusion frameworks such as MCAT and MOTCat. From a clinical perspective, these results suggest that FESCA holds strong potential for supporting more reliable and comprehensive diagnostic assessment by jointly leveraging histopathological morphology and molecular-level signatures.

Despite its strong performance, this study has several limitations. First, the number of paired multimodal samples remains relatively limited, and further validation on large-scale, multi-center cohorts is required to ensure the robustness and generalizability of the proposed model. Second, FESCA does not explicitly address the modality-missing scenario, which is common in real-world clinical practice. Third, FESCA adopts an ImageNet-pretrained ViT for WSI feature extraction rather than pathology-specific pretrained models, which may further enhance histopathological representation learning. In future work, we plan to collect larger and more diverse datasets to further validate the generalization ability of the proposed framework. Moreover, we will explore generative modeling strategies to improve the robustness and clinical applicability of FESCA under modality-incomplete conditions. In addition, incorporating pathology-specific pretrained visual backbones into the proposed multimodal framework will be investigated to further enhance performance. As multimodal diagnostic and prognostic tasks share common representation learning foundations, exploring unified modeling frameworks for different clinical task objectives represents another direction for future work.

6. Conclusions

In this work, we presented FESCA, a multimodal diagnostic framework designed to address the limited capability of existing fusion methods in modeling fine-grained intra-modality semantics and cross-modal semantic interactions. FESCA incorporates a semantic-guided feature enhancement module and a contrastive-learning-based alignment strategy to improve representation quality for both WSI and mRNA modalities. Experimental results on unimodal and multimodal tasks demonstrate that FESCA consistently outperforms state-of-the-art methods, achieving more discriminative and robust feature representations. These findings indicate that FESCA provides an effective and scalable solution for integrative pathological–genomic analysis and holds strong potential for future applications in computer-aided diagnosis.

CRediT authorship contribution statement

Huina Wang: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Lan Wei:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Investigation, Conceptualization. **Jianqiang Li:** Supervision, Resources, Investigation. **Bo Liu:** Supervision, Resources, Investigation. **Juan Fang:** Supervision, Investigation. **Catherine Mooney:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Investigation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT in order to improve the clarity and language expression. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by National Natural Science Foundation of China (Grant Number: 62076015), the 2023 International Cooperation Training Program for Innovative Talents (“Double First-class” Construction Special Program - “Artificial Intelligence + Internet of Things”) of the China Scholarship Council (CSC). This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland, under Grant number 21/RC/10294_P2 at FutureNeuro Research Ireland Centre for Translational Brain Science.

Data availability

All data used in this research were obtained from publicly available sources. We confirm that no private, confidential, or personally identifiable data were used in the analysis. All datasets utilized were either freely accessible or provided under open access licenses, ensuring compliance with relevant ethical guidelines. No new data collection was conducted during this research.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R.L. Siegel, I. Soerjomataram, A. Jemal, Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J. Clin. 74 (3) (2024) 229–263.

[2] R.L. Siegel, A.N. Giaquinto, A. Jemal, Cancer statistics, 2024, CA: Cancer J. Clin. 74 (1) (2024) 12–49.

[3] A.S. Elkorany, Z.F. Elsharkawy, Efficient breast cancer mammograms diagnosis using three deep neural networks and term variance, Sci. Rep. 13 (1) (2023) 2663.

[4] N. Aidossov, V. Zarikas, Y. Zhao, A. Mashekova, E.Y.K. Ng, O. Mukhmetov, Y. Mirasbekov, A. Omirbayev, An integrated intelligent system for breast cancer detection at early stages using IR images and machine learning methods with explainability, SN Comput. Sci. 4 (2) (2023) 184.

[5] A. Thakur, M. Gupta, D.K. Sinha, K.K. Mishra, V.K. Venkatesan, S. Guluwadi, Transformative breast cancer diagnosis using CNNs with optimized ReduceLROnPlateau and early stopping enhancements, Int. J. Comput. Intell. Syst. 17 (1) (2024) 14.

[6] J. Venugopalan, L. Tong, H.R. Hassanzadeh, M.D. Wang, Multimodal deep learning models for early detection of Alzheimer’s disease stage, Sci. Rep. 11 (1) (2021) 3254.

[7] J.N. Acosta, G.J. Falcone, P. Rajpurkar, E.J. Topol, Multimodal biomedical AI, Nature Med. 28 (9) (2022) 1773–1784.

[8] T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, Y. Peng, A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data, Irbm 43 (1) (2022) 62–74.

[9] R.J. Chen, M.Y. Lu, D.F. Williamson, T.Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo, et al., Pan-cancer integrative histology-genomic analysis via multimodal deep learning, Cancer Cell 40 (8) (2022) 865–878.

[10] R. Li, X. Wu, A. Li, M. Wang, HFBSurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction, Bioinformatics 38 (9) (2022) 2587–2594.

[11] B. Li, Y. Li, K.W. Eliceiri, Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14318–14328.

[12] W. Hou, L. Yu, C. Lin, H. Huang, R. Yu, J. Qin, L. Wang, Hˆ 2-MIL: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 933–941.

[13] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al., Transmil: Transformer based correlated multiple instance learning for whole slide image classification, Adv. Neural Inf. Process. Syst. 34 (2021) 2136–2147.

[14] S. Yoon, H. Yoon, J. Cho, K. Lee, AEmiGAP: AutoEncoder-based miRNA–Gene association prediction using deep learning method, Int. J. Mol. Sci. 25 (23) (2024) 13075.

[15] A. Licciardi, A. Fiannaca, M. La Rosa, M.A. Urso, L. La Paglia, A deep learning multi-omics framework to combine microbiome and metabolome profiles for disease classification, in: International Conference on Artificial Neural Networks, Springer, 2024, pp. 3–14.

[16] J. Wang, N. Liao, X. Du, Q. Chen, B. Wei, A semi-supervised approach for the integration of multi-omics data based on transformer multi-head self-attention mechanism and graph convolutional networks, BMC Genomics 25 (1) (2024) 86.

[17] J. Xiang, J. Zhang, Exploring low-rank property in multiple instance learning for whole slide image classification, in: The Eleventh International Conference on Learning Representations, 2023.

[18] R.J. Chen, M.Y. Lu, W.-H. Weng, T.Y. Chen, D.F. Williamson, T. Manz, M. Shady, F. Mahmood, Multimodal co-attention transformer for survival prediction in gigapixel whole slide images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4015–4025.

[19] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 2127–2136.

[20] M.Y. Lu, D.F. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (6) (2021) 555–570.

[21] S.A. Javed, D. Juyal, H. Padigela, A. Taylor-Weiner, L. Yu, A. Prakash, Additive mil: Intrinsically interpretable multiple instance learning for pathology, Adv. Neural Inf. Process. Syst. 35 (2022) 20689–20702.

[22] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S.E. Coupland, Y. Zheng, Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18802–18812.

[23] Y. Guan, J. Zhang, K. Tian, S. Yang, P. Dong, J. Xiang, W. Yang, J. Huang, Y. Zhang, X. Han, Node-aligned graph convolutional network for whole-slide image representation and classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18813–18823.

[24] Z. Fang, Y. Wang, Y. Zhang, Z. Wang, J. Zhang, X. Ji, Y. Zhang, Mammil: Multiple instance learning for whole slide images with state space models, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2024, pp. 3200–3205.
[25] S. Popov, S. Morozov, A. Babenko, Neural oblivious decision ensembles for deep learning on tabular data, 2019, arXiv preprint arXiv:1909.06312.
[26] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, Adv. Neural Inf. Process. Syst. 31 (2018).
[27] S.Ö. Arik, T. Pfister, Tabnet: Attentive interpretable tabular learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 6679–6687.
[28] I. Shavitt, E. Segal, Regularization learning networks: deep learning for tabular datasets, Adv. Neural Inf. Process. Syst. 31 (2018).
[29] K. Ranipa, W.-P. Zhu, M. Swamy, A novel feature-level fusion scheme with multimodal attention CNN for heart sound classification, Comput. Methods Programs Biomed. 248 (2024) 108122.
[30] K. Atrey, B.K. Singh, N.K. Bodhey, Multimodal classification of breast cancer using feature level fusion of mammogram and ultrasound images in machine learning paradigm, Multimedia Tools Appl. 83 (7) (2024) 21347–21368.
[31] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, NPJ Digit. Med. 3 (1) (2020) 136.
[32] J. Dhar, N. Zaidi, M. Haghighat, S. Roy, P. Goyal, A. Alavi, V. Kumar, Multimodal fusion learning with dual attention for medical imaging, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, IEEE, 2025, pp. 4362–4371.
[33] Y. Xu, H. Chen, Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21241–21251.
[34] G. Jaume, A. Vaidya, R.J. Chen, D.F. Williamson, P.P. Liang, F. Mahmood, Modeling dense multimodal interactions between biological pathways and histology for survival prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 11579–11590.
[35] F. Zhou, H. Chen, Cross-modal translation and alignment for survival analysis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21485–21494.
[36] A.H. Song, R.J. Chen, G. Jaume, A.J. Vaidya, A.S. Baras, F. Mahmood, Multimodal prototyping for cancer survival prediction, 2024, arXiv preprint arXiv:2407.00224.
[37] X. Huang, A. Khetan, M. Cvitkovic, Z. Karnin, Tabtransformer: Tabular data modeling using contextual embeddings, 2020, arXiv preprint arXiv:2012.06678.
[38] J. Gutheil, Self-attention and intersample attention transformer for tabular healthcare data/author Julian Gutheil, 2024.
[39] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput. Graph. Appl. 21 (5) (2002) 34–41.
[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[41] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, D. Shen, Transformers in medical image analysis, Intell. Med. 3 (1) (2023) 59–78.
[42] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (2014) 1–21.
[43] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[44] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PloS One 10 (3) (2015) e0118432.
[45] M. Liu, L. Hu, Y. Tang, C. Wang, Y. He, C. Zeng, K. Lin, Z. He, W. Huo, A deep learning method for breast cancer classification in the pathology images, IEEE J. Biomed. Health Inform. 26 (10) (2022) 5025–5032.
[46] S. Sharma, S. Kumar, The xception model: A potential feature extractor in breast cancer histology images classification, ICT Express 8 (1) (2022) 101–108.
[47] M. Xiao, Y. Li, X. Yan, M. Gao, W. Wang, Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example, in: Proceedings of the 2024 7th International Conference on Machine Vision and Applications, 2024, pp. 145–149.
[48] W. Wang, Y. Li, X. Yan, M. Xiao, M. Gao, Breast cancer image classification method based on deep transfer learning, in: Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition, 2024, pp. 190–197.
[49] A. Cutler, D.R. Cutler, J.R. Stevens, Random forests, Ensemble Mach. Learn.: Methods Appl. (2012) 157–175.
[50] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[51] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017).
[52] Z.-H. Zhou, J. Feng, Deep forest: towards an alternative to deep neural networks, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI '17, AAAI Press, 2017, pp. 3553–3559.
[53] G. Somepalli, M. Goldblum, A. Schwarzschild, C.B. Bruss, T. Goldstein, SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training, 2021, arXiv preprint arXiv:2106.01342.