MFMF: Multiple Foundation Model Fusion Networks for Whole Slide Image Classification*

Thao M. Dang
thaomai.dang@uta.edu
University of Texas at Arlington
Arlington, Texas, USA

Yuzhi Guo
yuzhi.guo@mavs.uta.edu
University of Texas at Arlington
Arlington, Texas, USA

Hehuan Ma
hehuan.ma@mavs.uta.edu
University of Texas at Arlington
Arlington, Texas, USA

Qifeng Zhou
qifeng.zhou@uta.edu
University of Texas at Arlington
Arlington, Texas, USA

Saiyang Na
saiyang.na@uta.edu
University of Texas at Arlington
Arlington, Texas, USA

Jean Gao
gao@uta.edu
University of Texas at Arlington
Arlington, Texas, USA

Junzhou Huang†
jzhuang@uta.edu
University of Texas at Arlington
Arlington, Texas, USA

ABSTRACT

Tumor detection and subtyping remain a significant challenge in histopathology image analysis. As digital pathology progresses, the applications of deep learning become essential. Whole Slide Image (WSI) classification has emerged as a crucial task in digital pathology, vital for accurate cancer diagnosis and treatment. In this paper, we introduce an innovative abnormal-guided **M**ultiple **F**oundation **M**odel **F**usion (MFMF) framework, aimed at enhancing WSI classification by integrating multi-level information from pathology images with Multiple Instance Learning (MIL). Traditional methods often focus on patch-level features while neglecting the rich contextual and morphological details at the cell and text levels, thus failing to fully exploit the multidimensional nature of WSIs. Our method enhances traditional models by efficiently integrating patch-level, cell-level, and text-level features using three foundation models. These are then fused through a novel three-step cross-attention module that effectively leverages cell and text information with patch-level features. Furthermore, unlike most studies that use attention scores to select instances based on the assumption that high scores indicate the presence of a tumor, we design an abnormality-aware module to naturally identify and detect abnormal features (i.e., tumors) as the criteria for selecting important instances, thereby reducing computational costs and boosting overall performance. We validate our approach against leading benchmarks on the CAMELYON16 and TCGA-Lung datasets, achieving superior classification performance. Our study not only tackles the challenges of sparsity and noise in multi-level features but also enhances the efficiency and accuracy of WSI classification by exploiting abnormal features.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Applied computing** → **Health informatics**; **Bioinformatics**.

KEYWORDS

WSI analysis; Multimodal fusion; Abnormal detection; Foundation model

ACM Reference Format:
Thao M. Dang, Yuzhi Guo, Hehuan Ma, Qifeng Zhou, Saiyang Na, Jean Gao, and Junzhou Huang. 2024. MFMF: Multiple Foundation Model Fusion Networks for Whole Slide Image Classification. In *15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '24), November 22–25, 2024, Shenzhen, China.* ACM, Shenzhen, China, 8 pages. https://doi.org/10.1145/3698587.3701372

1 INTRODUCTION

Cancer significantly threatens human health, making tumor detection and subtyping crucial for effective diagnosis and treatment [5]. Tissue examination by a pathologist remains the gold standard [25, 27]. With the adoption of digital slide scanning, advances in deep learning, and increased access to large datasets, computational pathology has transformed remarkably in recent years [21], especially in training models on whole slide images (WSIs) from Hematoxylin and Eosin (H&E)-stained specimens. However, WSIs can be gigapixel in size, making data collection and annotation labor-intensive [14]. A popular solution is weakly supervised learning based on Multiple Instance Learning (MIL) [11], where the WSI is tokenized into many patch embeddings using a pretrained vision encoder. These embeddings are then fed into pooling networks, such as attention networks, for downstream tasks [7, 20].

*This work was partially supported by US National Science Foundation IIS-2412195, CCF-2400785 and the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363).

†Corresponding author.

BCB '24, November 22–25, 2024, Shenzhen, China

ACM ISBN 979-8-4007-1302-6/24/11
https://doi.org/10.1145/3698587.3701372

Recently, pretrained models have achieved significant success in bioinformatics [9] and medical image analysis. However, most existing WSI-based methods focus on extracting patch-level features, neglecting the multi-level information intrinsic to pathology images. Some methods try to utilize cellular microenvironment information by using smaller patches, such as 16×16 pixel images [2, 3, 23], but this does not capture true cell-level information, missing critical features like cell contours and shapes. Moreover, reducing patch size results in losing the ability to learn about cellular context and background, limiting their effectiveness in providing comprehensive medical and pathological insights. Recent state-of-the-art (SOTA) multimodal methods, such as image-text models, offer additional textual knowledge to enhance analysis. However, obtaining accurate textual descriptions for pathology images usually requires manual annotations from experts, which is expensive and time-consuming.

Foundation models offer potential solutions to these issues. For example, foundation models for pathology images generate patch-level features and convert medical captions into text features, thereby enhancing image-text and text-image retrieval tasks [15]. For cell-level analysis, some fine-tuning technologies focus on cell segmentation [18]. Additionally, recent medical image-based large language models (LLMs) [12, 19] generate textual descriptions, reducing the need for manual annotations. Therefore, combining patch-level, cell-level, and text-level features can create a comprehensive representation, leveraging each foundation model's strengths for a more detailed understanding of pathology images. However, integrating these features introduces challenges, such as information sparsity in cell-level features and reliability issues in text-level descriptions, which may contain noise. Patch-level features are considered highly reliable, while cell-level and text-level features, derived from patches, can lack necessary information or include irrelevant details. Addressing these discrepancies is crucial to maximize the effectiveness of combined feature models in medical imaging.

A simple strategy is to concatenate embeddings from different foundation models, but this often leads to suboptimal solutions due to sparsity in cell-level information and noise in text-level features. This approach fails to effectively integrate the necessary multi-level information, limiting model performance and diagnostic accuracy. Additionally, some cross-attention multimodal fusion methods have limited capabilities because they are designed for only two modalities [26], making them difficult to reuse when the number of modality types exceeds two. Moreover, to our knowledge, there is little research on scenarios where the confidence level of one modality significantly surpasses others.

To address these challenges, we propose an abnormal guided **M**ultiple **F**oundation **M**odel **F**usion (MFMF) network with MIL for WSI classification. Our method generates multi-level information from WSIs using three foundation models: the CONCH [15] model for *patch-level* embeddings, the Segment Any Cell (SAC) [18] model for *cell-level* embeddings, and the Quilt-LLaVA [19] model for *text-level* embeddings. We introduce a three-step cross-attention module to integrate cell-level and text-level information with primary patch-level features. This module first fuses cell-level and text-level embeddings, then combines the merged cell-text features with patch-level embeddings. To further enhance predictive performance, we design an abnormal detection module to generate abnormality-aware features based on patch-level embeddings and fuse these with the patch-cell-text features. Extensive experiments on cancer classification and subtyping demonstrate the effectiveness of our framework, showing promising performance improvements. Overall, our contributions can be summarized as follows:

- We propose an innovative MFMF framework across pathology image, cell, and text-based foundation models, achieving superior classification performance on both cancer classification and subtyping datasets. Comparative analysis with multiple datasets further demonstrates that our method surpasses SOTA techniques.
- To the best of our knowledge, we are the first to introduce an abnormal detection module based on Variational Autoencoder (VAE) to naturally select top-k instances, thereby reducing computational costs and enhancing the learning of patch-level features for abnormal detection. This component not only boosts performance but also accelerates progress in tumor classification tasks.
- The proposed method integrates diverse feature types derived exclusively from WSIs without requiring any additional manually curated data, such as micro environment annotations or expert-provided ground truth descriptions. Instead, annotations are effectively replaced by automatically extracted cell features, and expert descriptions are substituted with text features.

2 METHODOLOGY

2.1 Multiple Instance Learning

We adopt the MIL approach for WSI classification, as MIL effectively handles large data with only slide-level labels, given that obtaining instance-level annotations in medical imaging is costly and time-consuming. Particularly, each WSI is treated as a bag of multiple instances. A bag is labeled as *positive* if it contains at least one positive instance (i.e., tumor cropped patch) and *negative* otherwise.

$$\mathrm{Y}_i = \begin{cases} 0, & \text{iff } \sum y_{i,j} \in \{0,1\}, \text{with } j = 1 \dots m, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

Given a bag of instances $\mathbf{X}_i = \{\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,m}\}$, where the instance-level labels $\{y_{i,1}, \dots, y_{i,m}\}$ are unknown, our goal is to predict $\hat{\mathbf{Y}}_i \in \{0,1\}$ such that the prediction $\hat{\mathbf{Y}}_i$ matches the target value $\mathbf{Y}_i \in \{0,1\}$, for $i = 1, \dots, b$. Here, b represents the total number of bags, and m is the number of instances in the i-th bag. Notably, the value of m can vary across different bags.

2.2 Autoencoder-based Abnormal Detection

The vanilla Variational Autoencoder (VAE) consists of two primary components: an encoder, Enc_{VAE}, and a decoder, Dec_{VAE}. The encoder compresses the input data x into a lower-dimensional latent space z, and the decoder attempts to reconstruct the original feature from this latent representation. The VAE is trained by minimizing the following loss function:

$$\mathcal{L}_{VAE} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - KL(q(z|x)||p(z)). \tag{2}$$

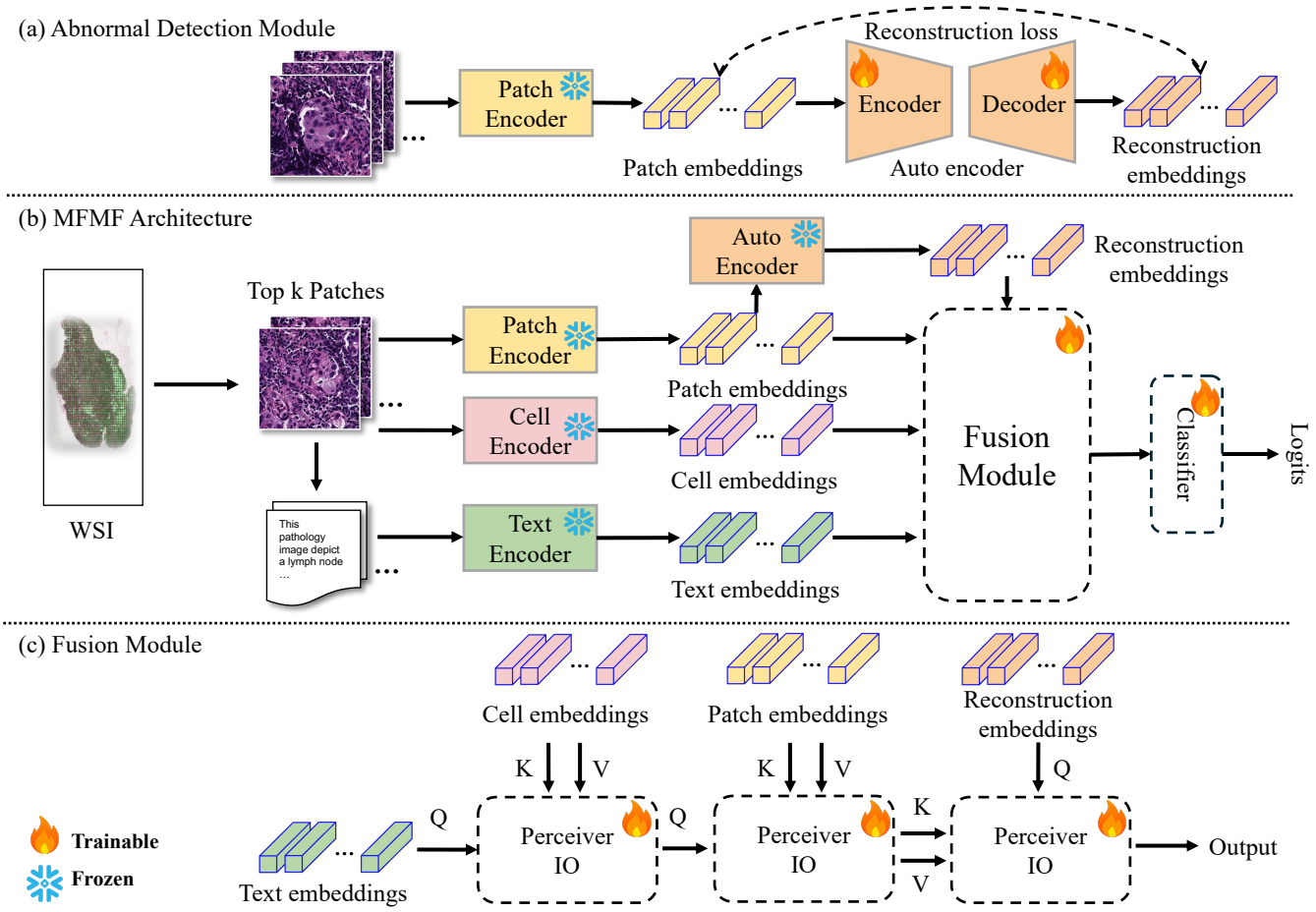


Figure 1: Overview of MFMF. (a) Abnormal Detection Module: The process begins by dividing input WSIs into patches and extracting patch-level embeddings using an image encoder. These embeddings are used to train an abnormal detection module, generating reconstruction features and calculating reconstruction errors to select the top-k potential patches. (b) MFMF: The selected patches are further processed to extract cell-level and text-level embeddings. (c) Fusion Module: These embeddings are integrated using cross-attention mechanisms. Finally, the integrated embedding is passed through a fully connected layer to predict the slide-level label.

The intrinsic capability of VAE to detect abnormalities aligns well with the MIL classification described in Eq. 1. Since instance-level labels are not available, it is only certain that all instances in a normal bag are normal, whereas a tumor bag contains both normal and tumor embeddings. To implement the concept of VAE in tumor classification, the abnormal detection (AD) module is trained exclusively on WSIs labeled as *normal*. For cancer subtype classification, any subtype can be designated as a *normal* case.

We utilize AD model to generate embeddings that are sensitive to abnormalities from tiles extracted from WSIs. Intuitively, this approach focuses on identifying and encoding deviations from typical histopathological patterns, facilitating a more refined analysis in subsequent processing stages. Further than that, to completely leverage the power of AD, we introduce the reconstruction error-based selection module that will be applied before extracting cell and text-level embeddings (see Figure 1). This module, equipped with AD, filters out patches that do not exhibit significant abnormal features, thus focusing computational resources on top-k promising candidates for detailed analysis. This strategy helps to reduce both the data preparation stage and the volume of processed data in training, streamlining the overall workflow and enhancing the efficiency of the system. The specifics of the error-based instance selection approaches will be thoroughly detailed in the next subsection.

2.3 Multiple Foundation Model Fusion

Feature encoding with foundation models. Given m cropped patches from the bag $\mathbf{X}_i$, cell features $\mathbf{F}_c \in \mathbb{R}^{m \times d_1}$ and patch features $\mathbf{F}_p \in \mathbb{R}^{m \times d_3}$ are derived from the corresponding encoders, where SAC [18] and CONCH [15] are employed, respectively. For text features, we apply Quilt-LLaVA [19] as a caption generation function $Gen_{cap}(\mathbf{X}_i)$ to produce descriptions of the patches. The prompts in MFMF are designed to elicit responses using three

specific types of queries proposed in the LLaVA structure [13]: short conversation, detailed description, and complex reasoning. We use the prompts introduced by Quilt-LLaVA directly and supplement them with relevant medical terms depending on the tumor or subtype classification tasks. The length of the response is limited to a maximum of 1024 tokens. These text responses of arbitrary length are then encoded to obtain fixed-size text-based embeddings, $\mathbf{F}_t \in \mathbb{R}^{m \times d_2}$. The diverse feature extraction techniques employ in this phase are summarized in the Appendix section.

Guided abnormal features with VAE. Given a set of training samples $B = \{(\mathbf{X}_i, \mathbf{Y}_i)\}^b$, where b is the number of bags, the abnormal detection component is trained using patch features from bags $\mathbf{X}_i$ if the corresponding labels $\mathbf{Y}_i$ are denoted as *normal*. For instance, in tumor classification, the selected bags have $\mathbf{Y}_i = 0$. We design the loss function of the Abnormal Detection (AD) module based on the VAE architecture. Thus, the Eq. 2 is updated to combine the Mean Squared Error (MSE) for the reconstruction loss and the Kullback-Leibler Divergence (KLD) for the regularization loss, defined as follows:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{f}_p^{(i)} - Dec_{VAE}(Enc_{VAE}(\mathbf{f}_p^{(i)}))||^2, \tag{3}$$

$$\mathcal{L}_{KLD} = -\frac{1}{2}\sum_{j}^{d}\left(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right), \tag{4}$$

$$\mathcal{L}_{VAE} = \mathcal{L}_{MSE} + \mathcal{L}_{KLD}, \tag{5}$$

with N is the total number of patches that belong to training bags labeled as *normal*, d is the dimensionality of the latent space, and μ and σ are the mean and standard deviation of the latent variable, respectively. The combined loss function, $\mathcal{L}_{VAE}$, ensures that the model not only accurately reconstructs the input features but also maintains a well-behaved latent space by regularizing it to follow a standard normal distribution. During both the training and testing phases of MFMF, the trained autoencoder is kept frozen. Each patch's image feature, $\mathbf{f}_p$, is processed through the frozen autoencoder to obtain the reconstructed feature, $\mathbf{f}_r \leftarrow Dec_{VAE}(Enc_{VAE}(\mathbf{f}_p))$.

Integrating multimodal features. Let the quadruplet ($\mathbf{F}_p, \mathbf{F}_c, \mathbf{F}_t, \mathbf{F}_r$) represent the matrices that contain all the corresponding features for each modality, which we denote as $\mathbf{F}$ for simplicity. Based on the defined MIL classification problem in Section 2.1, to predict the bag-level label, the MIL models need to aggregate all instances and then produce a conclusion in the form of: $\hat{\mathbf{Y}}_i = g(l(\mathbf{F}))$, where $l(\cdot)$ is the aggregation function and $g(\cdot)$ is the bag-level classifier.

We designed the function $l(\cdot)$ by cascading three cross-attention blocks to integrate the information from the quadruplet $\mathbf{F}$. The cross-attention [22] module can be mathematically expressed as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})V, \tag{6}$$

where $\sqrt{d_k}$ is the dimension of the key vectors, which scales the dot-product of $\mathbf{Q}$ and $\mathbf{K}^T$.

In the initial attention block, $\mathbf{H}_1 = \text{Attention}(\mathbf{F}_t, \mathbf{F}_c, \mathbf{F}_c)$, cell features are used as keys ($\mathbf{K}$) and values ($\mathbf{V}$), while text features serve as queries ($\mathbf{Q}$). This layer integrates information from text descriptions with cell-based embeddings to enhance the representation of cell features. In the subsequent block, patch features serve as $\mathbf{K}$ and $\mathbf{V}$, and the output from the first block is then used as $\mathbf{Q}$, represented as $\mathbf{H}_2 = \text{Attention}(\mathbf{H}_1, \mathbf{F}_p, \mathbf{F}_p)$. Reconstructed patch features, which capture abnormal information, are used as queries in the third cross-attention block, while $\mathbf{K}$ and $\mathbf{V}$ are derived from $\mathbf{H}_2$. This integration, denoted as $\mathbf{H}_3$, ensures that the model leverages both original and reconstructed features for robust decision-making. We then average the resultant $\mathbf{H}_3$ to obtain a bag representation. One linear layer, $Linear(\cdot)$, is applied to the aggregated features, yielding the logits. The decision classifier $g(\cdot)$ is trained using the Binary Cross-Entropy loss:

$$\mathcal{L}_{BCE} = -\frac{1}{b}\sum_{i=1}^{b}\left[\mathbf{Y}_i \log(\hat{\mathbf{Y}}_i) + (1 - \mathbf{Y}_i)\log(1 - \hat{\mathbf{Y}}_i)\right]. \tag{7}$$

Scaling MFMF with Perceiver IO. The Transformer architecture suffers from quadratic complexity, leading to inefficient scaling in both computational and memory resources. To address this issue, we utilized Perceiver IO [8], which performs the attention mechanism in the latent space. Thus, the three lightweight Perceiver IO blocks are applied as the cross-attention mechanism in the function $l(\cdot)$ to process data F efficiently.

Scaling MFMF with reconstruction error. Furthermore, we design a reconstruction error-based approach to scale the data and decrease the computational costs, as mentioned in Section 2.2. This approach involves selecting instances in a bag based on the Euclidean distances between the original features $\mathbf{F}_p$ and the reconstructed features $\mathbf{F}_r$, using two strategies:

- Maximum selection: We select the top-k instances with the highest reconstruction error scores to be forwarded to our MFMF. This strategy focuses on the instances most likely to contain tumor information or potential false positives.
- MinMax selection: This approach selects the top-k instances based on both the highest and lowest reconstruction error scores. By including instances with minimal reconstruction error, we balance the model's learning space and prevent it from becoming overconfident, reducing the likelihood of false negatives.

Given a bag $\mathbf{X}_i$, the number of instances remaining in the bag after applying the select function $Select(\cdot)$ is $\lceil k * m \rceil$, where $k \in (0, 1]$ represents the percentage of instances in the bag to be processed. By doing so, we reduce the number of instances to be processed by $(1 - k) \times 100\%$, making the model more efficient.

3 EXPERIMENTS

3.1 Implementation Details

3.1.1 Datasets: We evaluate the proposed method on three different histopathological datasets: CAMELYON16 [4], TCGA-LUAD [1], and TCGA-LUSC [10]. TCGA-LUAD and TCGA-LUSC are combined into a single large dataset, TCGA-Lung, for the cancer subtype classification task.

In the pre-processing stage, each WSI is cropped into 1024×1024 patches without overlap to form a bag, with magnifications of $40\times$ for CAMELYON16 and $20\times$ for TCGA-Lung. We apply Macenko

color normalization [17] and discard patches with more than 30% background.

CAMELYON16 consists of 398 WSIs, producing 569,533 patches, while TCGA-Lung includes 1,042 WSIs, yielding 729,193 patches. Both datasets are split for 5-fold cross-validation, with CAMELYON16 tested on its official test set and TCGA-Lung on the DSMIL GitHub test set [11]. The training-to-testing ratios are 269:129 for CAMELYON16 and 828:214 for TCGA-Lung..

3.1.2 Evaluation: All experiments are conducted on a single Nvidia GTX 4090 GPU with 32GB of RAM and an Intel Core i7 processor. We report the mean and standard deviation of the 5-fold cross-validation results for the area under the curve (AUC), accuracy, and recall scores of eight baselines and our proposed MFMF in the task of WSI classification on both datasets. We evaluate our method's robustness against eight baselines in both unimodal and multimodal settings, ensuring fairness by using the same input features for all methods.

In the unimodal setting, the input to baselines is a set of patch feature vectors $\mathbf{F}_p$ extracted using the foundation CONCH model. MFMF utilizes these $\mathbf{F}_p$ as key and value, while using its reconstruction features $\mathbf{F}_r$ as query, and combines the two features with a single Perceiver IO block. In the multimodal setting, the inputs are tuples of image, cell, and text features, denoted as $\mathbf{F}_p$+$\mathbf{F}_c$+$\mathbf{F}_t$. To implement the multimodal mode on eight baseline methods, the image, cell, and text features are concatenated as described in [6]. The normal class in CAMELYON16 and the LUSC class in TCGA-Lung are used to train the AD module.

3.2 Results

MFMF demonstrates superior performance in both unimodal and multimodal settings, as shown in the comparisons in Tables 1-2.

For the unimodal setting on CAMELYON16, MFMF achieves an AUC of 0.9402, an accuracy of 0.9302, and a recall of 0.9090. These results demonstrate that MFMF is highly competitive, with excellent performance across all metrics. Significantly, MFMF achieved the best recall, underscoring its robustness in detecting relevant features from image data. In the multimodal setting, MFMF outperforms other methods, achieving an AUC of 0.9746, an accuracy of 0.9566, and a recall of 0.9429. Specifically, MFMF with instance selection achieves the highest performance. This substantial improvement with multimodal inputs highlights the effectiveness of integrating multiple feature types for pathological image analysis.

Our method again demonstrates superior performance for the unimodal setting on TCGA-Lung, achieving the best results for accuracy and the second-best recall. This validates its robustness across different datasets. In the multimodal setting, MFMF maintains its superior performance, achieving the best results for AUC, accuracy, and recall metrics. This further demonstrates the advantage of using multimodal inputs.

Notably, our recall scores consistently surpass those of other methods, indicating that the abnormal detection module successfully guides MFMF to identify positive features. This enhanced recall is crucial, as false negatives are particularly dangerous in fields such as diagnosis and medical imaging, where missing information about metastases or other critical abnormalities can have serious consequences.

4 CONCLUSIONS

In this study, we introduced MFMF, an innovative framework that integrates multiple foundation models to enhance whole slide image classification performance. This framework distinctively incorporates an abnormal-guided Variational Autoencoder, which significantly boosts classification accuracy by effectively integrating patch-level, cell-level, and text-level features through a cross-attention mechanism. Extensive experiments are conducted on the CAMELYON16 and TCGA-Lung datasets to demonstrate the superior performance of our model over existing SOTA methods, particularly in multimodal settings. Looking ahead, we plan to establish a benchmark for multi-class classification to further validate and refine our framework's capabilities. MFMF represents a significant and promising step forward in computational pathology.

REFERENCES

[1] Watson M. Holback C. Jarosz R. Kirk S. Lee Y. Rieger-Christ K. Lemmerman J. Albertina, B. 2016. The Cancer Genome Atlas Lung Adenocarcinoma Collection (TCGA-LUAD). The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.JGNIHEP5

[2] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. 2022. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16144–16155.

[3] Richard J Chen and Rahul G Krishnan. 2021. Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology. *Learning Meaningful Representations of Life, NeurIPS 2021* (2021).

[4] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. 2017. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* 318, 22 (12 2017), 2199–2210. https://doi.org/10.1001/jama.2017.14585

[5] Olivier Elemento, Christina Leslie, Johan Lundin, and Georgia Tourassi. 2021. Artificial intelligence in cancer research, diagnosis and therapy. *Nature Reviews Cancer* 21, 12 (2021), 747–752.

[6] Simon Holdenried-Krafft, Peter Somers, Ivonne Montes-Mojarro, Diana Silimon, Cristina Tarín, Falko Fend, and Hendrik P. A. Lensch. 2023. Dual-Query Multiple Instance Learning for Dynamic Meta-Embedding based Tumor Classification. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA.

[7] Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*. PMLR, 2127–2136.

[8] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2022. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *International Conference on Learning Representations*.

[9] Feng Jiang, Yuzhi Guo, Hehuan Ma, Saiyang Na, Wenliang Zhong, Yi Han, Tao Wang, and Junzhou Huang. 2024. GTE: a graph learning framework for prediction of T-cell receptors and epitopes binding specificity. *Briefings in Bioinformatics* 25, 4 (07 2024), bbae343. https://doi.org/10.1093/bib/bbae343

[10] Lee Y. Kumar P. Filippini J. Albertina B. Watson M. Rieger-Christ K. Lemmerman J. Kirk, S. 2016. The Cancer Genome Atlas Lung Squamous Cell Carcinoma Collection (TCGA-LUSC). https://doi.org/10.7937/K9/TCIA.2016.TYGKKFMQ

[11] Bin Li, Yin Li, and Kevin W Eliceiri. 2021. 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*. 14318–14328.

[12] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890* (2023).

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning.

[14] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. 2024. A visual-language foundation model for computational pathology. *Nature Medicine* 30, 3 (2024), 863–874.

Table 1: Classification performance comparison on CAMELYON16. The best result is shown in bold, the second-best result is underlined, and the third-best result is in *italics*. "MFMF ∗" represents our methods.

Feature	$\mathbf{F}_p$			$\mathbf{F}_p$+$\mathbf{F}_c$+$\mathbf{F}_t$		
Method	AUC	Accuracy	Recall	AUC	Accuracy	Recall
Mean Pooling	0.6371 ± 0.0273	0.7023 ± 0.0093	0.6200 ± 0.0107	0.6194 ± 0.0279	0.5953 ± 0.0257	0.5346 ± 0.0310
Max Pooling	0.8018 ± 0.0124	0.7721 ± 0.0062	0.7087 ± 0.0077	0.6831 ± 0.0437	0.5581 ± 0.1372	0.5607 ± 0.0744
ABMIL [7]	0.9302 ± 0.0026	0.9108 ± 0.0079	0.7837 ± 0.0208	0.7711 ± 0.0114	0.8031 ± 0.0125	0.5020 ± 0.0099
CLAM-SB [16]	0.9233 ± 0.0013	0.9077 ± 0.0084	0.7633 ± 0.0163	0.7446 ± 0.0240	0.7769 ± 0.0341	0.4449 ± 0.0473
CLAM-MB [16]	0.9092 ± 0.0119	0.8954 ± 0.0104	0.7265 ± 0.0208	0.7417 ± 0.0182	0.7538 ± 0.0188	0.4000 ± 0.0306
DSMIL [11]	0.9334 ± 0.0028	0.9339 ± 0.0062	0.8367 ± 1.1e-16	0.8090 ± 0.0269	0.8154 ± 0.0262	0.6000 ± 0.0420
TransMIL [20]	0.9373 ± 0.0030	0.9132 ± 0.0130	0.8928 ± 0.0124	0.8283 ± 0.0274	0.8248 ± 0.0227	0.7923 ± 0.0135
ILRA-MIL [24]	0.9402 ± 0.0062	0.9400 ± 0.0031	0.8489 ± 0.0099	0.7742 ± 0.0499	0.7723 ± 0.0593	0.5469 ± 0.0757
MFMF w/o Top-k	0.9392 ± 0.0197	0.9116 ± 0.0159	0.8901 ± 0.0126	*0.9627 ± 0.0091*	*0.9426 ± 0.0079*	*0.9277 ± 0.0094*
MFMF (MinMax)	0.9391 ± 0.0309	0.9240 ± 0.0173	0.9079 ± 0.0179	0.9702 ± 0.0110	0.9457 ± 0.0245	0.9327 ± 0.0345
MFMF (Max)	0.9402 ± 0.0132	0.9302 ± 0.0110	0.9090 ± 0.0151	**0.9746 ± 0.0098**	**0.9566 ± 0.0079**	**0.9429 ± 0.0104**

Table 2: Classification performance comparison on TCGA-Lung. The best result is shown in bold, the second-best result is underlined, and the third-best result is in *italics*. "MFMF ∗" represents our methods.

Feature	$\mathbf{F}_p$			$\mathbf{F}_p$+$\mathbf{F}_c$+$\mathbf{F}_t$		
Method	AUC	Accuracy	Recall	AUC	Accuracy	Recall
Mean Pooling	0.9695 ± 0.0029	0.9075 ± 0.6669	0.9075 ± 0.0069	0.9120 ± 0.0050	0.8458 ± 0.0102	0.8452 ± 0.0103
Max Pooling	0.9711 ± 0.0019	0.9071 ± 0.0048	0.9069 ± 0.0048	0.8586 ± 0.0050	0.7116 ± 0.0102	0.7109 ± 0.0103
ABMIL [7]	0.9756 ± 0.0034	0.9131 ± 0.0048	0.8972 ± 0.0135	0.9656 ± 0.0055	0.9112 ± 0.0089	0.9101 ± 0.0158
CLAM-SB [16]	0.9729 ± 0.0037	0.9084 ± 0.0096	0.9083 ± 0.0082	0.9662 ± 0.0044	0.9140 ± 0.0113	0.9046 ± 0.0149
CLAM-MB [16]	0.9738 ± 0.0047	0.9234 ± 0.0076	0.9083 ± 0.0164	0.9698 ± 0.0059	0.9168 ± 0.0035	0.9064 ± 0.0158
DSMIL [11]	0.9685 ± 0.0055	0.9112 ± 0.0051	0.9266 ± 0.0154	0.9506 ± 0.0057	0.8757 ± 0.0124	0.9028 ± 0.0179
TransMIL [20]	0.9706 ± 0.0048	0.9121 ± 0.0176	0.9126 ± 0.0180	0.9405 ± 0.0157	0.8738 ± 0.0145	0.8739 ± 0.0143
ILRA-MIL [24]	0.9742 ± 0.0055	0.9206 ± 0.0084	0.9276 ± 0.0249	0.9531 ± 0.0049	0.8869 ± 0.0116	0.8844 ± 0.0179
MFMF w/o Top-k	0.9737 ± 0.0006	0.9149 ± 0.0069	0.9151 ± 0.0074	*0.9791 ± 0.0030*	*0.9346 ± 0.0029*	*0.9348 ± 0.0028*
MFMF (MinMax)	0.9724 ± 0.0007	0.9196 ± 0.0171	0.9191 ± 0.0175	0.9804 ± 0.0035	**0.9374 ± 0.0069**	**0.9377 ± 0.0067**
MFMF (Max)	0.9737 ± 0.0030	0.9271 ± 0.0029	0.9269 ± 0.0028	**0.9815 ± 0.0029**	0.9355 ± 0.0062	0.9358 ± 0.0058

[15] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. 2024. A visual-language foundation model for computational pathology. *Nature Medicine* 30 (2024), 863–874.

[16] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* 5, 6 (2021), 555–570.

[17] Marc Macenko, Marc Niethammer, J. Marron, David Borland, John Woosley, Xiaojun Guan, Charles Schmitt, and Nancy Thomas. 2009. A Method for Normalizing Histology Slides for Quantitative Analysis. *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009* 9, 1107–1110. https://doi.org/10.1109/ISBI.2009.5193250

[18] Saiyang Na, Yuzhi Guo, Feng Jiang, Hehuan Ma, and Junzhou Huang. 2024. Segment Any Cell: A SAM-based Auto-prompting Fine-tuning Framework for Nuclei Segmentation. *ArXiv* abs/2401.13220 (2024).

[19] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. 2023. Quilt-LLaVA: Visual Instruction Tuning by Extracting Localized Narratives from Open-Source Histopathology Videos. *arXiv e-prints* (2023), arXiv–2312.

[20] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems* 34 (2021), 2136–2147.

[21] Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. 2023. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering* 1, 12 (2023), 930–949.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[23] Hangchen Xiang, Junyi Shen, Qingguo Yan, Meilian Xu, Xiaoshuang Shi, and Xiaofeng Zhu. 2023. Multi-scale representation attention based deep multiple instance learning for gigapixel whole slide image analysis. *Medical Image Analysis* 89 (2023), 102890. https://doi.org/10.1016/j.media.2023.102890

[24] Jinxi Xiang and Jun Zhang. 2023. Exploring Low-Rank Property in Multiple Instance Learning for Whole Slide Image Classification. In *The Eleventh International Conference on Learning Representations*.

[25] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis* 65 (2020), 101789.

[26] Fengtao Zhou and Hao Chen. 2023. Cross-Modal Translation and Alignment for Survival Analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 21485–21494.

[27] Qifeng Zhou, Wenliang Zhong, Yuzhi Guo, Michael Xiao, Hehuan Ma, and Junzhou Huang. 2024. PathM3: A Multimodal Multi-Task Multiple Instance Learning Framework for Whole Slide Image Classification and Captioning. *arXiv preprint arXiv:2403.08967* (2024).

A APPENDIX

A.1 Experiment Setup

The process of MFMF begins with the extraction of patch-level embeddings from the input instances $\mathbf{X}_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,m}\}$ using an image encoder, resulting in feature representations $\mathbf{F}_p$, with $\mathbf{F}_p$ is a set that contains m feature vectors $\mathbf{f}_p$. These embeddings are then utilized to train an abnormal detection module (VAE) which produces reconstruction features $\mathbf{F}_r$. By calculating the reconstruction errors $||\mathbf{f}_p^{(i)} - \mathbf{f}_r^{(i)}||^2$, potential abnormal instances are identified, and a subset of these instances is selected based on a threshold k, where $||\cdot||^2$ denotes the squared Euclidean norm. The selected instances undergo further feature extraction, where cell-level and text-level embeddings are generated using respective encoders. These embeddings $\mathbf{f}_c$ and $\mathbf{f}_t$, along with the refined patch-level and reconstruction features, are integrated using attention mechanisms. Finally, the integrated features are passed through a classifier to predict the WSI-level label $\hat{\mathbf{Y}}_i$. The diverse feature extraction techniques employed in this research are summarized in Table 3.

Table 3: Feature extraction methods.

Embedding	Process	Dim.
$\mathbf{f}_p$	Patches of size 1024×1024 pixels were fed into the image encoder of the CONCH foundation model, which uses ViT-Base-16 as a backbone.	$\mathbb{R}^{512}$
$\mathbf{f}_c$	SAC fine-tuned the image encoder, a ViT from SAM, with a cell segmentation task using LoRA. We then extracted the features using SAC's image encoder, processing them with max pooling to generate cell features.	$\mathbb{R}^{1280}$
$\mathbf{f}_t$	The Quilt-LLaVA foundation model was used to generate a description for each patch. These descriptions were then input into the text encoder of the CONCH model.	$\mathbb{R}^{512}$
$\mathbf{f}_r$	An image-based embedding extracted by CONCH was then encoded and decoded by the Abnormal Detection model (see Section 2.2) to obtain the 512-dimensional reconstructed embedding.	$\mathbb{R}^{512}$

In terms of the prompts used for the CAMELYON16 and TCGA-Lung datasets to obtain patch descriptions generated by Quilt-LLaVA, we employed the following prompts, respectively:

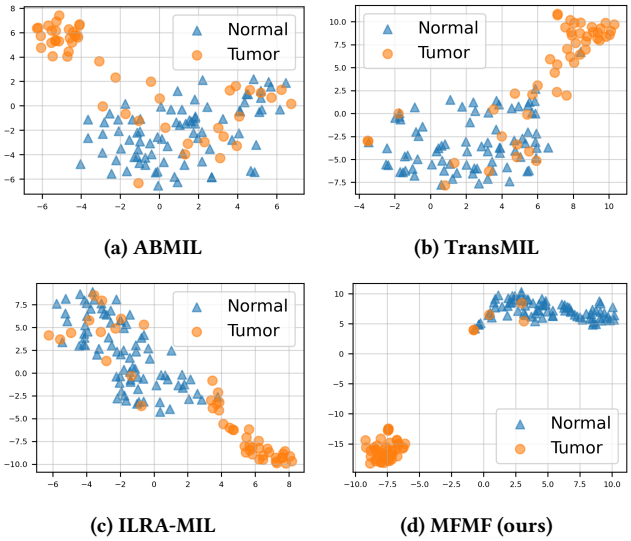
'Can you describe the main features visible in this histopathology image? In a few words, what does the histopathology image depict? Is there a tumor in this pathology image? Are there abnormal, neoplastic, atypical, or metastatic cells in this pathology image? Make a diagnosis based on this single patch of histopathology image.'

'Can you describe the main features visible in this histopathology image? In a few words, what does the histopathology image depict? Is it lung adenocarcinoma or lung squamous cell carcinoma?'

A.2 Classification Results

To emphasize the robustness of our method, we plot bag embeddings produced by four different methods on the CAMELYON16 test set under multimodal conditions.

In Figure 2, the clear distinction between 'Normal' and 'Tumor' categories in the MFMF model's scatter plot underscores its effectiveness in differentiating between tissue types. In contrast, the other methods show more category mixing, suggesting less effective feature integration and classification capabilities.



(a) ABMIL (b) TransMIL

(c) ILRA-MIL (d) MFMF (ours)

Figure 2: t-SNE visualizations of the CAMELYON16 test set in multimodal mode.

Similarly, in Figure 3, the proposed method continues to outperform other methods in distinguishing between two subtypes, 'LUAD' and 'LUSC', with clear clustering of each category. This consistent performance across different datasets underscores the effectiveness of the MFMF model in feature integration and classification tasks, highlighting its potential for broader applications in histopathological image analysis.

A.3 Ablation Results

A.3.1 Top-k settings: We conduct a grid search experiment to gain an in-depth understanding of the Maximum and MinMax selection strategies.

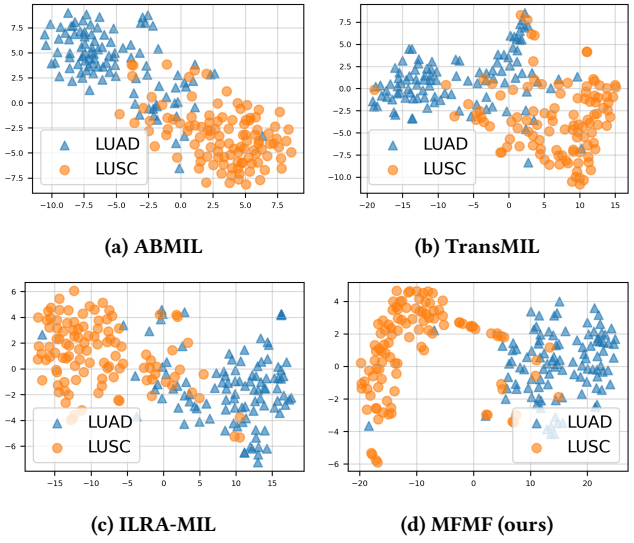
From Figure 4, in CAMELYON16, the best performance for Maximum selection is at $k = 0.3$, while in TCGA-Lung, the best performance for both strategies is around $k = 0.6$. The MinMax selection strategy demonstrates more stable performance across different k values compared to Maximum selection. This stability suggests that including instances with minimal reconstruction error helps in maintaining a balanced learning space, reducing overconfidence in the model. Maximum selection tends to peak at specific k values but shows more variability as k changes. However, although the MinMax selection strategy generally demonstrates stable performance across different k values, there are cases, such as at $k = 0.2$ in the TCGA-Lung dataset, where its performance is more variable. In contrast, the Maximum selection method performs better at specific k values, particularly on the TCGA-Lung dataset, where $k \in \{0.2, 0.4, 0.8\}$. This suggests that while MinMax might be a more robust choice when the optimal k value is uncertain, the Maximum

Table 4: Classification performance for different instance selection strategies on CAMELYON16.

	Maximum selection				MinMax selection			
Top-k	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision
$k = 0.4$	0.9648 ± 0.0124	0.9442 ± 0.0158	0.9265 ± 0.0208	0.9570 ± 0.0108	0.9708 ± 0.0109	0.9426 ± 0.0267	0.9249 ± 0.0351	0.9583 ± 0.0173
$k = 0.3$	0.9746 ± 0.0098	0.9566 ± 0.0079	0.9429 ± 0.0104	0.9674 ± 0.0056	0.9702 ± 0.0109	0.9457 ± 0.0245	0.9327 ± 0.0345	0.9604 ± 0.0161
$k = 0.2$	0.9626 ± 0.0165	0.9504 ± 0.0152	0.9363 ± 0.0204	0.9608 ± 0.0103	0.9695 ± 0.0102	0.9442 ± 0.0237	0.9306 ± 0.0338	0.9593 ± 0.0155
$k = 0.1$	0.9613 ± 0.0198	0.9379 ± 0.0129	0.9215 ± 0.0154	0.9491 ± 0.0131	0.9633 ± 0.0108	0.9504 ± 0.0126	0.9355 ± 0.0168	0.9619 ± 0.0087

Table 5: Classification performance for different instance selection strategies on TCGA-Lung.

	Maximum selection				MinMax selection			
Top-k	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision
$k = 0.8$	0.9769 ± 0.0038	0.9299 ± 0.0029	0.9303 ± 0.0028	0.9305 ± 0.0027	0.9744 ± 0.0046	0.9168 ± 0.0075	0.9168 ± 0.0075	0.9177 ± 0.0075
$k = 0.6$	0.9804 ± 0.0035	0.9374 ± 0.0069	0.9377 ± 0.0067	0.9386 ± 0.0056	0.9815 ± 0.0029	0.9355 ± 0.0062	0.9358 ± 0.0058	0.9365 ± 0.0048
$k = 0.4$	0.9769 ± 0.0038	0.9308 ± 0.0019	0.9312 ± 0.0018	0.9314 ± 0.0016	0.9736 ± 0.0018	0.9206 ± 0.0066	0.9209 ± 0.0069	0.9221 ± 0.0069
$k = 0.2$	0.9762 ± 0.0018	0.9280 ± 0.0023	0.9286 ± 0.0022	0.9294 ± 0.0020	0.9639 ± 0.0097	0.9001 ± 0.0179	0.8999 ± 0.0183	0.9008 ± 0.0173



(a) ABMIL (b) TransMIL

(c) ILRA-MIL (d) MFMF (ours)

Figure 3: t-SNE visualizations of the TCGA-Lung test set in multimodal mode.

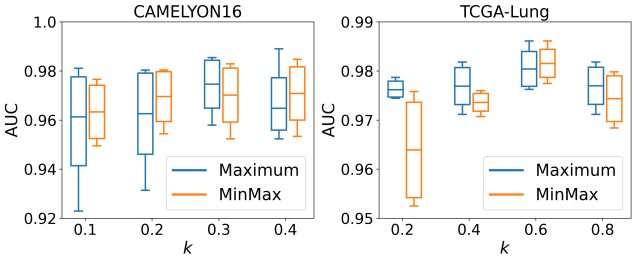


Figure 4: AUC performance for different instance selection strategies on CAMELYON16 and TCGA-Lung. Results for other metrics can be found in the Appendix section.

method could be more effective in certain cases, depending on the dataset and selected k value. Overall, the worst performance for both strategies is still comparable to SOTA methods, indicating the reliability of our approach. The results also suggest that the effectiveness of instance selection strategies can be dataset-dependent. Please refer to Tables 4-5 for the full report.

A.3.2 Abnormal guided component: To highlight the importance of reconstruction features and the proposed instance selection strategies, we present the classification results of MFMF when these features are excluded. In this scenario, the input features consist of the triplet $(\mathbf{F}_p, \mathbf{F}_c, \mathbf{F}_t)$ only.

From the results of Table 6, it is evident that the inclusion of the AD component improves the overall performance of the MFMF model across both datasets. For the CAMELYON16 dataset, the AUC increases from 0.9478 to 0.9746, indicating stronger class distinction. Additionally, accuracy improves from 0.9395 to 0.9566, and recall rises from 0.9236 to 0.9429, showing enhanced identification of positive samples. In the TCGA-Lung dataset, the AUC increases slightly from 0.9806 to 0.9815, while accuracy improves from 0.9271 to 0.9374, and recall from 0.9271 to 0.9377, reflecting better true positive identification. These results underscore the AD component's role in enhancing the MFMF model's performance. Consistent improvements in AUC, accuracy, and recall across both datasets highlight the effectiveness of the AD component in refining classification, which is crucial for precise abnormality detection in medical imaging.

Table 6: Classification performance of MFMF without the abnormal detection component (no instance selection and no reconstruction features $\mathbf{F}_r$) in multimodal mode.

Dataset	AUC	Accuracy	Recall
CAMELYON	0.9478 ± 0.016	0.9395 ± 0.008	0.9236 ± 0.011
TCGA-Lung	0.9806 ± 0.003	0.9271 ± 0.009	0.9271 ± 0.009