Medical Image Analysis 97 (2024) 103251

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

MIST: Multi-instance selective transformer for histopathological subtype prediction

Rongchang Zhao [a], Zijun Xi [a], Huanchi Liu [a], Xiangkun Jian [a], Jian Zhang [a], Zijian Zhang [b], Shuo Li [a,c,*]

[a] School of Computer Science and Engineering, Central South University, Changsha, China
[b] National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, China
[c] Department of Computer and Data Science and Department of Biomedical Engineering, Case Western Reserve University, Cleveland, USA

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Multi-instance learning
Histopathological subtype prediction
Self-attention
Feature decoupling
Information bottleneck

ABSTRACT

Accurate histopathological subtype prediction is clinically significant for cancer diagnosis and tumor microenvironment analysis. However, achieving accurate histopathological subtype prediction is a challenging task due to (1) instance-level discrimination of histopathological images, (2) low inter-class and large intra-class variances among histopathological images in their shape and chromatin texture, and (3) heterogeneous feature distribution over different images. In this paper, we formulate subtype prediction as fine-grained representation learning and propose a novel multi-instance selective transformer (MIST) framework, effectively achieving accurate histopathological subtype prediction. The proposed MIST designs an effective selective self-attention mechanism with multi-instance learning (MIL) and vision transformer (ViT) to adaptive identify informative instances for fine-grained representation. Innovatively, the MIST entrusts each instance with different contributions to the bag representation based on its interactions with instances and bags. Specifically, a SiT module with selective multi-head self-attention (S-MSA) is well-designed to identify the representative instances by modeling the instance-to-instance interactions. On the contrary, a MIFD module with the information bottleneck is proposed to learn the discriminative fine-grained representation for histopathological images by modeling instance-to-bag interactions with the selected instances. Substantial experiments on five clinical benchmarks demonstrate that the MIST achieves accurate histopathological subtype prediction and obtains state-of-the-art performance with an accuracy of 0.936. The MIST shows great potential to handle fine-grained medical image analysis, such as histopathological subtype prediction in clinical applications.

1. Introduction

Histopathological subtype prediction is clinically significant for the diagnosis and treatment of cancer disease. Histopathological subtype prediction aims to identify different sub-categories associated with the pathological tissues in whole slide image (WSI) (Fig. 1(1)), *i.e.,* Normal mucosa, debris, pathological benign, lymphocytes, and invasive carcinoma (Han et al., 2022). Histopathological subtype prediction is always rested upon by clinical decision of optimal therapeutic schedule (Han et al., 2017). With the understanding of histopathological subtype, pathologists can control the metastasis of tumor cells early and make substantial therapeutic schedules according to special clinical performance and prognosis results of multiple cancers (Han et al., 2017), such as breast cancer, and colorectal cancer. In addition, the prediction of histopathological subtypes sheds light on the tumor microenvironment analysis and has a significant impact on clinical endpoints (Gurcan et al., 2009; Kather et al., 2019, 2018, 2017).

Although existing histopathological subtype prediction works have made promising progress in computer-aided diagnosis (CAD) (Hashimoto et al., 2020; Yang et al., 2019; Laleh et al., 2022; Srinidhi et al., 2022), achieving accurate histopathological subtype prediction is still challenging due to: **(1) Patch-level representation for cancer subtypes or pathological tissues**. Different from existing WSI works, the subtype prediction task is a fine-grained prediction problem that discriminates different categories of cancer subtypes or pathological tissues from each other. This task requires patch-level representations of histopathological images for precise subtype prediction. **(2) Low inter-class and large intra-class variation in their shape and chromatin texture**. The histopathology appearance is similar between

* Corresponding author at: Department of Computer and Data Science and Department of Biomedical Engineering, Case Western Reserve University, Cleveland, USA.
E-mail address: slishuo@gmail.com (S. Li).

https://doi.org/10.1016/j.media.2024.103251
Received 19 April 2023; Received in revised form 24 January 2024; Accepted 21 June 2024
Available online 26 June 2024

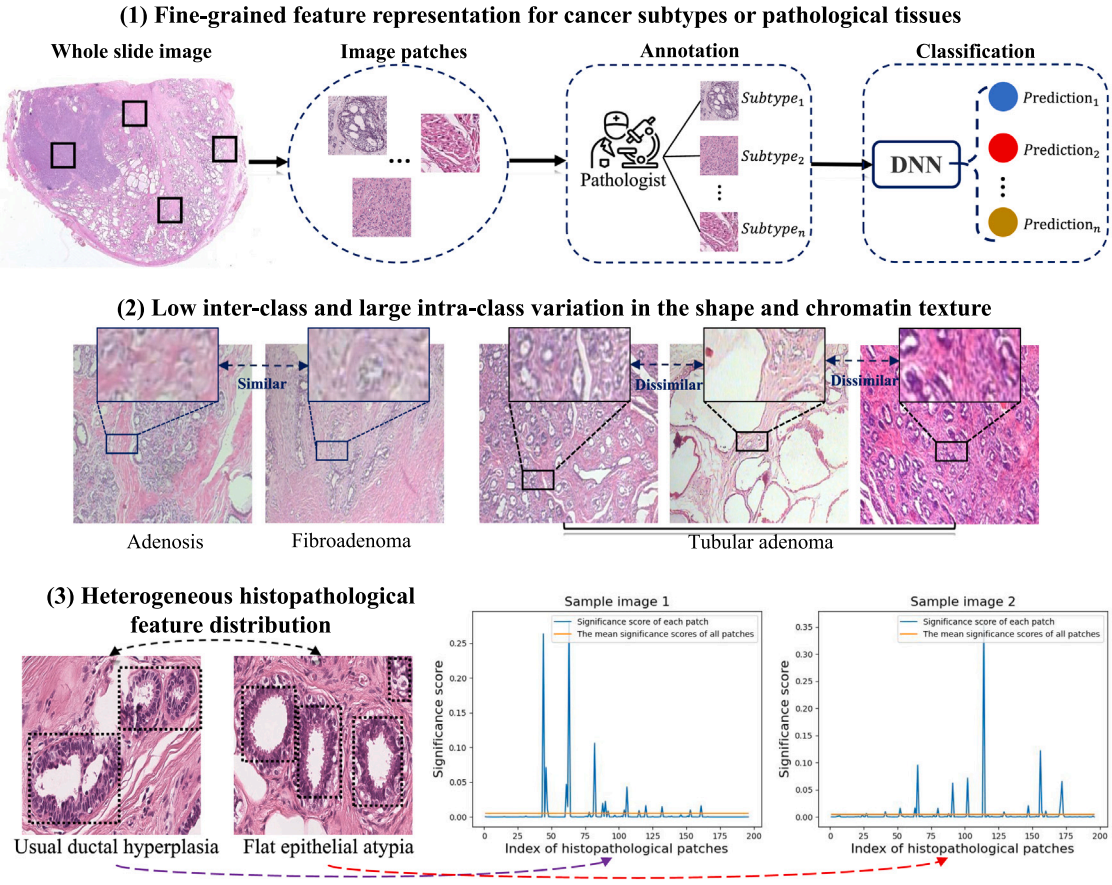


Fig. 1. Three challenges hinder the accurate histopathological subtype prediction. (1) Fine-grained feature representation for cancer subtypes or pathological tissues. (2) Low inter-class and large intra-class variances in their shape and chromatin texture, which are extremely hard to recognize. (3) Heterogeneous histopathological feature distribution over different images that uniform feature extraction is prone to lead to degraded prediction performance which needs to adaptively select highly significant pathological regions in different histopathology images.

different inter-class (Fig. 1(2)), such as the *staining*, *cell distribution*. Meanwhile, subtype characteristics of adenosis and fibroadenoma are quite similar, leading to extremely challenging subtype identification. Moreover, there are large differences between intra-class images of tubular adenomas, such as color, number of nuclei, texture, *etc*, bringing high ambiguity for accurate subtype prediction. **(3) Heterogeneous histopathological feature distribution over different images.** Based on the validation experimental results (Fig. 1(3)), it can be observed that the significant feature of the histopathological subtypes is widely distributed over the various locations for different images. Specifically, the significant histopathological regions (highlighted with dashed boxes) tend to clutter together and hold erratic morphology changes, hindering the accurate histopathological subtype prediction. The validation experiment justifies (curve of Fig. 1(3)) that the significance score of different histopathological patches varies with their location and different distribution of the number of significant histopathological patches in different images, where the horizontal coordinates represent the indices of the different patch tokens in vision transformer (ViT) and the vertical coordinates represent the attention scores of different patch tokens in ViT. Treating the mean of the significance scores of all patches as a measure, it can be seen that the number of highly significant patches is small and varies from image to image. Therefore, uniform feature extraction of all regions of the histopathology image using an encoder is prone to lead to degraded prediction performance.

Multi-instance learning (MIL) has gained popularity in histopathological image analysis, especially for WSI classification. MIL-based approach regards WSI as a bag of instances and each instance corresponds to one 2D patch of the image. One straightforward idea of MIL is to perform a pooling operation on instance features for bag prediction. To obtain the best performance, previous studies focus on the instance representation (Ilse et al., 2018; Hashimoto et al., 2020), feature pooling (Li et al., 2021a), and instance-relation modeling (Shao et al., 2021; Li et al., 2021b; Myronenko et al., 2021). Although those methods made progress in WSI prediction, it still suffers from some **defects** in histopathological subtype prediction (Fig. 2(a)): **(1)** Performance degeneration due to the missed simultaneous modeling of the local instance-to-instance and global instance-to-bag interactions. Because traditional MIL is often used for large-scale WSI prediction, it is difficult to learn the global bag representation in an end-to-end manner, leading to inconsistent features among the instances. **(2)** It predicts histopathological subtype based on instance representation **with equal contribution**, which lacks the precision identification of representative instances for the fine-grained bag representation. In histopathological subtype prediction, it is common for only a small number of instances within a histopathological image to be tightly associated with the disease of interest. However, existing approaches are hard to find the discriminative instances for subtype prediction. **(3)** Lack of effective instance selection and decoupling strategy for robust MIL. Specifically, the features of the histopathological image are often mixed with irrelevant noise, making the bag representation of the histopathological image a performance degeneration. The conventional MIL decision is inefficient due to the participated irrelevant instances, which obstruct the local instance-to-instance interactions for learning a fine-grained representation. Hence, continuously selecting and stripping local irrelevant instance representations from global information is necessary. In other words, decoupling local fine-grained features from global bag representation.

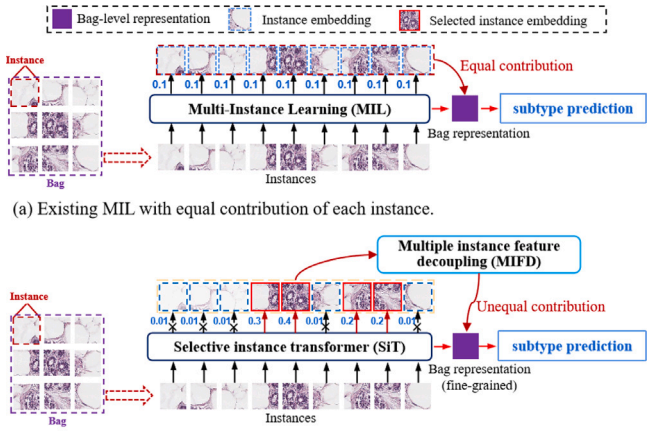


Fig. 2. Different from existing multi-instance learning (MIL), the proposed MIST innovatively formulates subtype prediction as fine-grained representation learning with MIL and ViT, and constructs novel selective self-attention mechanism to conduct histopathological subtype prediction, where each instance conducts unequal contributions to the representation of the histopathological image by the instance selection and feature decoupling.

Token-based MIL (Dosovitskiy et al., 2021; Shao et al., 2021; Li et al., 2021b; Myronenko et al., 2021) is a potential motivation for the prediction of histopathological subtypes by modeling the long-range interaction between different instances. Token-based MIL is conducted with the ViT framework, where the class token in ViT represents the bag and the patch token in ViT denotes the instance in the bag. However, existing token-based MIL methods still have some limitations in achieving fine-gained histopathological subtype prediction: **(1)** Learning of instance representation suffers from **negative impact of redundant patches**. The quality of the instance features greatly influences the performance of the subsequent bag representation and subtype prediction. Most existing methods extract instance features from deep neural networks with inductive biases such as locality and weight sharing, resulting in an inability to capture the global semantic representation of bags. **(2)** It is lack of an effective **instance scoring and selection for bag representation**. For most representation learning frameworks, the bag representation is affected by unessential instances, resulting in an inability to learn robust features to discriminate images with similar subtypes.

In this paper, we propose a novel multi-instance selective transformer (MIST, Fig. 3) framework to achieve accurate histopathological subtype prediction. The MIST (Fig. 2(b)) innovatively formulates subtype prediction as fine-grained representation learning and designs a novel selective self-attention mechanism with multi-instance learning (MIL) and vision transformer (ViT) to learn the instance-level fine-grained representation. It should be noted that the MIST entrusts each instance with different contributions for the bag prediction based on its interactions of instance-to-instance and instance-to-bag. Specifically, the selective instance transformer (SiT) is well-designed to adaptively identify the representative instances from the bag to participate in the bag representation learning by modeling the instance-to-instance interaction. Meanwhile, multiple instance feature decoupling (MIFD) is proposed to learn the instance-level fine-grained representation of histopathological images with the selected representative instances by modeling the instance-to-bag interactions. To learn the discriminative representation with those representative instances, an information bottleneck loss function is constructed for the novel MIST training. Therefore, the MIST is capable of achieving accurate histopathological subtype prediction by: **(1)** proposing the selective self-attention mechanism by coupling multi-instance learning with the vision transformer (ViT) to learn the fine-grained histopathology representation; **(2)** adaptively selecting the representative instances to be aggregated for the learning of discriminative bag representation with the selective self-attention mechanism; **(3)** leveraging the bag-level prior knowledge to the token-based MIL for the representative instance features learning with the multiple instance feature decoupling.

The main contributions are summarized as follows:

- The proposed MIST formulates histopathological subtype prediction as fine-grained representation learning and proposes a novel selective self-attention mechanism by advancing multi-instance learning (MIL) and vision transformer (ViT) in a unified framework.
- For the first time, the selective instance transformer (SiT) is designed by selecting the representative instances with a self-attention learning paradigm to learn the instance-level fine-grained representation in histopathological subtype prediction.
- The multiple instance feature decoupling (MIFD) is proposed to leverage information bottleneck into the fine-grained representation learning and conduct accurate histopathological subtype prediction.

2. Related work

2.1. Conventional MIL methods

MIL belongs to a weakly supervised learning problem, where instance labels are unknown but labels for instances of bags are given. MIL models consider customized aggregators, such as mean-pooling and max-pooling (Feng and Zhou, 2017; Pinheiro and Collobert, 2015). With the advent of CNN, the CNN-based aggregation operators make a significant performance (Feng and Zhou, 2017; Wang et al., 2018; Lee et al., 2017; Shi et al., 2020). Several methods (Ilse et al., 2018) are based on an attention mechanism that allocates the contribution of each instance to the embedding. In addition, the combination of MIL and multi-label learning (Feng and Zhou, 2017) is proposed to deal with the problem that each instance corresponds to multiple labels and obtain a promising result. The contextual information-based methods (Tu et al., 2019; Yan et al., 2018; Chikontwe et al., 2020; Lin et al., 2022) are proposed to model the relationships between the instances with capsule network, graph neural network, and causal inference. However, these methods do not simultaneously consider local instance-to-instance and global instance-to-bag interactions in MIL.

2.2. MIL for histopathological subtype prediction

MIL has been successfully applied to WSI prediction, and it consists of two parts: **(1)** CNN-based MIL methods (Ilse et al., 2018; Hashimoto et al., 2020; Li et al., 2021a; Macenko et al., 2009; Lin et al., 2022). Ilse et al. (2018) proposed an attention-based aggregation operator with CNN which includes the contribution of each instance to the bag embedding. The multi-scale is considered (Hashimoto et al., 2020; Li et al., 2021a) due to the WSIs being analyzed in a multi-scale environment. Hashimoto et al. (2020) proposed a multi-scale method with a domain adversarial mechanism, several instances from different scales are treated as a bag and the bag feature is extracted by CNN to firstly train single-scale domain adversarial-MIL, and then put each trained feature extractor together to accomplish multi-scale. Li et al. (2021a) adopted non-local attention in MIL to model the instance-to-instance and instance-to-bag relations between the highest-score instance and all the remaining instances. Macenko et al. (2009) assigned each instance a pseudo-label based on the predicted probability by a pre-trained model with a bag-level label and constructed an instance-level loss function, but it only works for binary prediction of WSI. Recently, Zhang et al. (2022) proposed the double-tier MIL framework which introduced a concept of pseudo-bags to alleviate the issue of the limited number

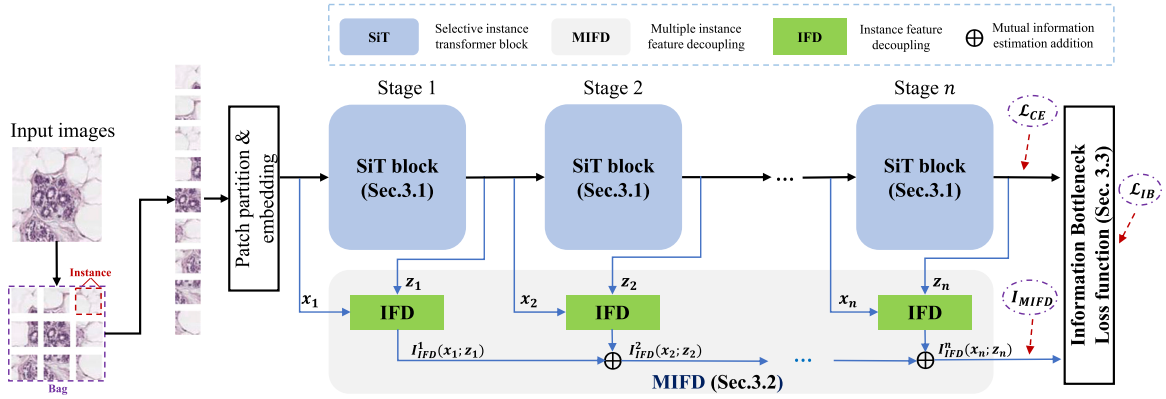


Fig. 3. The MIST formulates histopathological subtype prediction as a fine-grained representation learning problem and provides a novel learning paradigm of selective self-attention mechanism by advancing vision transformer architecture and multi-instance learning (MIL). The MIST achieves histopathological subtype prediction by progressively learning the fine-grained histopathological representation. The MIST consists of three key components: (1) selective instance transformer (SiT) with S-MSA to adaptively identify the representative instances for the representation of the histopathological image by modeling the instance-to-instance interactions with selective self-attention mechanism; (2) multiple instance feature decoupling (MIFD) to gradually learn instance-level fine-grained feature representation for histopathological subtype prediction by modeling the instance-to-bag interactions; (3) information bottleneck loss function to train the multiple-stage transformer with the idea of multi-instance learning by integrating the instance-to-instance and instance-to-bag interactions.

of WSIs and leveraged the idea of Grad-CAM to generate the instance probability. **(2)** Token-based MIL methods (Shao et al., 2021; Li et al., 2021b; Myronenko et al., 2021). Some instances of a bag are assigned with pseudo-label based on the attention weights (Myronenko et al., 2021) and a feature map with different scales extracted by CNN is connected after each transformer block to achieve the performance of multi-scale. Shao et al. (2021) A combination of CNN and ViT was adopted to obtain both local contextual and global spatial information, extracting multi-fine-grained features using group convolutions of different scales. Unfortunately, the existing works have not solved the special problem of token-based MIL in histopathological subtype prediction from tokens that are not fully exploited and lack noise alleviation and redundancy elimination.

2.3. Information bottleneck in deep learning

With the rapid development of deep learning, it is important to understand and interpret the essence of deep learning models. Recently, several works have applied information theory to the field of deep learning, especially computer vision. Among them, the information bottleneck (IB) has gained increasing popularity in analyzing deep learning model (Zhmoginov et al., 2020; Bang et al., 2021; Kim et al., 2020; Wang et al., 2022; Lai et al., 2021).

The information bottleneck (Tishby et al., 2001; Shamir et al., 2010; Alemi et al., 2016) is an information-theoretic learning principle for latent representation learning. It has also achieved considerable performance in medical image analysis (Bardera et al., 2009; Thirion and Faugeras, 2004; Bardera et al., 2007; Zuo et al., 2021; Song et al., 2022). However, existing approaches showed less interest in incorporating the information bottleneck into histopathological subtype prediction, not to mention histopathological subtype prediction. Furthermore, the information bottleneck is frequently used in traditional machine learning and CNNs, but rarely used in vision transformers due to its different network architecture. Therefore, it is promising to incorporate the information bottleneck into the vision transformer for sufficient discriminative representation learning.

3. Method

The proposed MIST (Fig. 3) formulates histopathological subtype prediction as fine-grained representation learning and achieves histopathological subtype prediction by constructing a multiple-stage vision transformer coupled with multi-instance learning. The newly-designed MIST learns the instance-level fine-grained features by adaptive entrusting each instance with different contributions for the histopathological image representation. Therefore, the MIST has three tightly connected components: **(1) selective instance transformer (SiT)** with selective self-attention mechanism (S-MSA) to adaptively identify the representative instances from the bag for discriminative representation of the histopathological image. The SiT conducts instance selection by modeling the instance-to-instance interactions with a selective self-attention mechanism; **(2) multiple instance feature decoupling (MIFD)** to gradually learn instance-level fine-grained feature representation for histopathological subtype prediction by modeling the instance-to-bag interactions; **(3) loss function with information bottleneck** to classify the histopathological subtypes with the bag representation learned with MIST by integrating the instance-to-instance and instance-to-bag interactions.

Mathematical preparation: The MIST follows the idea of multi-instance learning for bag-level prediction, where histopathological image patches are formulated as instances. Specifically, given a bag X with N instances $\{x_p^1, x_p^2, \dots, x_p^N\}$, where the instance is corresponding to the partitioned 2D patches of the histopathological image. Note that, the bag label Y is given, whereas instance-level labels $\{y_1, y_2, \dots, y_n\}$ are unavailable.

After the patch and position embedding of all image patches (Fig. 3), the input image (bag) $X_0 \in \mathbb{R}^{H\times W\times C}$ are encoded into a sequence of flattened 2D patches (instances) $x_p \in \mathbb{R}^{N\times(P^2\times C)}$ (where H, W, and C denote the height, width, and channel the number of images, respectively, each instance with the size of $P\times P$), and the input bag is divided into $N = HW/P^2$ instances. Then those instances are flattened and mapped to D dimensions with a trainable linear projection E to obtain the embedding space $\in \mathbb{R}^{N\times D}$. The bag-level representation $x_{\text{cls}} \in \mathbb{R}^{1\times D}$ is learned as the global semantic embedding to achieve the fine-grained prediction of the histopathological subtype, coupled with the instances. It can be formulated as:

$$X_0 = \left[\underbrace{x_{\text{cls}}}_{\text{bag}}; \underbrace{x_p^1 E}_{\text{instance 1}}; \underbrace{x_p^2 E}_{\text{instance 2}}; \dots; \underbrace{x_p^N E}_{\text{instance N}}\right] + E_{\text{pos}} \tag{1}$$

where $X_0, E_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$, E_{pos} denotes the position embedding which retains positional information.

3.1. Selective instance Transformer (SiT) block

The SiT (Fig. 4) is designed by integrating a selective self-attention mechanism into the vision transformer structure, adaptively identifying

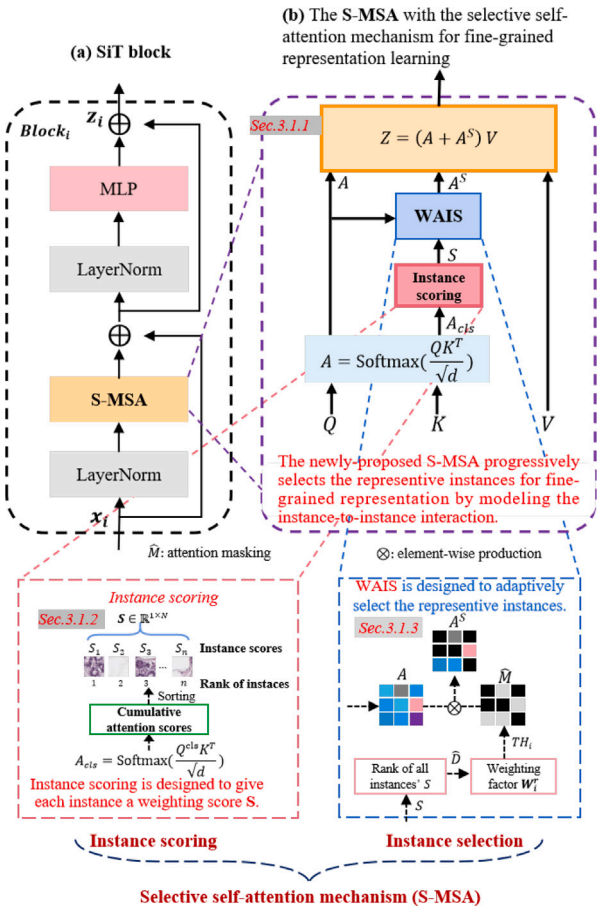


Fig. 4. (a) The SiT learns the fine-grained representation of the histopathological image. (b) The S-MSA with selective self-attention mechanism, which consists of instance scoring and WAIS, progressively selects the representative instances for fine-grained representation by modeling the instance-to-instance interaction. Instance scoring gives each instance a score of significance and WAIS adaptively selects the representative instances as fine-grained features.

and selecting the representative instances for the histopathological image bag representation. The selective self-attention mechanism contains instance identification (S-MSA) and instance selection (WAIS) through instance-to-instance self-attention interactions. In each SiT block, the selective multi-head self-attention (S-MSA) with instance scoring and weight adaptive instance selection (WAIS) is newly designed to empower the vanilla MSA for gradually extracting the local instance representation. Specifically, the instance scoring of S-MSA quantifies the contributions of different instances by the cumulative attention score. The WAIS of S-MSA adaptively selects the instances with high rank as the informative instances from more to less and re-weights them in the global information for accurate fine-grained histopathological subtype prediction.

The SiT with S-MSA learns the instance-level fine-grained representation from instance-to-instance interactions among different histopathological instances, which conventional MSA does not address in ViT. According to the interaction between instance-to-instance, a corresponding proportion of attentive instances are selected as fine-grained features and continuously re-weighted in different stages of the global representation to progressively improve the fine-grained feature extraction capability of the model. In general, the procedure of SiT (according to Eq. (1)) can be summarized as:

$$X_l = \text{S-MSA}(\text{LN}(X_{l-1})) + X_{l-1}, \quad l \in 1, 2, \ldots, L \tag{2}$$

where $\text{LN}(\cdot)$ is the layer normalization operation. L is the number of selective instances transformer blocks.

3.1.1. Selective multi-head self-attention (S-MSA)

The S-MSA (Fig. 4(b)) models the instance-to-instance interaction to assign each instance a score of importance and progressively select the instances with significant scores as informative representations for fine-grained representation learning.

For selective multi-head self-attention, given the input bag $X \in \mathbb{R}^{H\times W\times C}$, each bag is encoded as an embedding $x \in \mathbb{R}^{(N+1)\times d}$ and each instance x_p is encoded as an embedding $x_i \in \mathbb{R}^{1\times d}$ (detailed in Section 3), where d denotes the head dimension. The queries $\mathbf{Q} \in \mathbb{R}^{(N+1)\times d}$, keys $\mathbf{K} \in \mathbb{R}^{(N+1)\times d}$ and values $\mathbf{V} \in \mathbb{R}^{(N+1)\times d}$ are computed from the input instance embedding x by a linear fully-connected neural network. The attention matrix $\mathbf{A} \in \mathbb{R}^{(N+1)^2}$ of the corresponding x_i in the self-attention can be calculated as:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\times\mathbf{K}^{\top}}{\sqrt{d}}\right) \tag{3}$$

where each row of $\mathbf{A}$ sums up to 1 due to $\text{Softmax}(\cdot)$ operation and each row of $\mathbf{A}$ contains the attention weights of an input instance. The values of the attention weights represent the contribution of different instances. $\mathbf{A}_{1,1}$ is the attention weight of the bag-level representation, it is closely related to the contribution of other instances. Hence, the attention weights $\mathbf{A}_{1,2}, \ldots, \mathbf{A}_{1,N+1}$ of all instances can represent the importance of the corresponding instance, respectively.

3.1.2. Attention-guided instance scoring

To entrust each instance with different contributions for the bag representation, adaptive weight is leveraged to identify the representative instances via the self-attention mechanism. It can be formulated as

$$\mathbf{A}_{cls} = \text{Softmax}\left(\frac{\mathbf{Q}^{cls}\times\mathbf{K}^{\top}}{\sqrt{d}}\right) \tag{4}$$

where $\mathbf{Q}^{cls} \in \mathbb{R}^{1\times d}$ denotes the first row of $\mathbf{Q}$ (*i.e.*, bag-level representation). $\mathbf{A}_{cls} = [\mathbf{A}_{1,1}, \mathbf{A}_{1,2}, \ldots, \mathbf{A}_{1,N+1}] \in \mathbb{R}^{1\times(N+1)}$ denotes the significance of the input instances for the output bag-level representation.

To jointly attend to information from different representation subspaces at different positions, multi-head self-attention (MSA) is defined by considering multiple attention heads (Zhu et al., 2022). Vanilla MSA performs equality processing for each query $\mathbf{Q}$ to calculate the global attention scores (Eq. (3)). In other words, each local position of the image interacted with all positions similarly. For identifying fine-grained histopathological representation, we expect to mine discriminative local information to facilitate learning subtle features. To this end, the S-MSA computes the cumulative attention scores $\mathbf{A}^* \in \mathbb{R}^{1\times(N+1)}$ from all different heads in the MSA $\hat{\mathbf{A}} \in \mathbb{R}^{M\times(N+1)^2}$ (according to Eq. (4)) for instance scoring. The significance score $\mathcal{S}_i$ of a instance i is calculated as:

$$\hat{\mathbf{A}}_{cls} = [\mathbf{A}^1_{cls}, \mathbf{A}^2_{cls}, \ldots, \mathbf{A}^M_{cls}] \tag{5}$$

$$\mathbf{A}^* = \sum_{j=1}^{M} \mathbf{A}^j_{cls} \tag{6}$$

$$\mathcal{S}_i = \frac{\mathbf{A}^*_i}{\sum_{j=2}^{N+1} \mathbf{A}^*_j} \tag{7}$$

where M is the number of self-attention heads in each SiT. $\hat{\mathbf{A}}_{cls} \in \mathbb{R}^{M\times1\times(N+1)}$ denotes $\mathbf{A}_{cls}$ set of MSA. The cumulative attention scores $\mathbf{A}^*$ denote the significance scores of all instances. $\mathbf{A}^*_i$ is one of the cumulative attention scores from all heads, representing the cumulative attention score of the corresponding ith instance. Note that $\mathcal{S} \in \mathbb{R}^{1\times N}$ does not contain the value of bag-level representation.

3.1.3. Weighted adaptive instance selection (WAIS)

The WAIS adaptively selects the representative instances with high $\mathcal{S}$ from more to less and re-weight them in the global information for accurate fine-grained histopathological subtype prediction. Many instances with similar keys, which occur in the early stages of the ViT, will lower their attention weights due to the *Softmax* operation (Fayyaz et al., 2022). Precisely, some instance features are less discriminative in early stages but may be useful in later stages where discriminability is improved.

To reasonably select the representative instances at different stages, the WAIS assigns different weighting factors $\mathcal{W}_i^r$ (Eq. (8)) to the corresponding S-MSA of each SiT $block_i$ from the early stages to the last stages. Specifically, different SiTs are sequentially and linearly assigned correspondingly large to small selection ratios of informative instances. Moreover, It can be described as:

$$\mathcal{W}_i^r = 1 - \frac{i-1}{L}, \quad 1 \le i \le L \tag{8}$$

where i denotes the ith stage of SiTs, L is the number of SiTs.

To adaptively extract the informative instances, a naive approach is to select the crucial instances whose significance scores $\mathcal{S}$ surpass the average value of all instances' $\mathcal{S}$, and mask the lower ones. However, experiments (Fig. 8, Table 11 and Fig. 9) show that this *average* method will mask a large number of instances with low attention scores in the early stage of ViT, resulting in the abandonment of those potential informative instances in the later stage and achieving adverse performance. To remedy this, the instance selection strategy of the WAIS is performed hierarchically, *i.e.*, it ranks all instances by their corresponding $\mathcal{S}$ (Eq. (7)) and constructs the threshold TH of instance selection (Eq. (10)), then progressively drops the uninformative instances via newly-proposed attention masking $\hat{\mathbf{M}} \in \mathbb{R}^{(N+1)^2}$ (Eq. (11), Algorithm 1) to indicate whether to drop or keep the current instances. It can be summarized as:

$$\hat{\mathbf{D}} = \text{argsort}(\text{argsort}(\mathcal{S})), \quad \hat{\mathbf{D}} \in \mathbb{R}^{1\times N} \tag{9}$$

$$TH_i = (1 - \mathcal{W}_i^r) \cdot N \tag{10}$$

$$\hat{\mathbf{M}}_{k,i,j} = \begin{cases} 1, j \in [2, N+1], & \text{if} \quad \hat{\mathbf{D}}_i \ge TH_k, \\ 1, i \in [2, N+1], & \text{if} \quad \hat{\mathbf{D}}_j \ge TH_k, \\ 0, & \text{else.} \end{cases} \tag{11}$$

where $TH_i \in \mathbb{N}^+$ denotes the threshold of informative instance selection at ith stage. N is the number of instances. $\hat{\mathbf{D}}$ denotes the rank of different instances' significance scores $\mathcal{S}$ among all instances.

To further reinforce the fine-grained local feature meanwhile ensuring the global perception, through the element-wise production between the attention masking $\hat{\mathbf{M}}$ and the attention matrix $\mathbf{A}$, obtaining the selected attention matrix $\mathbf{A}^S$ which is considered as the fine-grained attention weight matrix. Then the fusion of fine-grained local feature and global information is leveraged by the addition of $\mathbf{A}$ and $\mathbf{A}^S$ to generate the output instance embeddings $\mathbf{Z}$ of histopathological subtype feature for accurate prediction. Therefore, these can be summarized as

$$\mathbf{A}^S = \mathbf{A} \cdot \hat{\mathbf{M}} \tag{12}$$

$$\mathbf{Z} = (\mathbf{A} + \mathbf{A}^S)\mathbf{V} \tag{13}$$

where $\mathbf{A}^S \in \mathbb{R}^{M\times(N+1)^2}$ denotes the selected attention matrix, $\mathbf{Z} \in \mathbb{R}^{(1+N)\times D}$ denotes the output instance embeddings from S-MSA.

3.1.4. Algorithm of selective self-attention mechanism

Given the input instances x, multi-head self-attention scores, and the weighting factor from different weighted adaptive transformer blocks in Section 3.1. Algorithm 1 describes the implementation of the selective self-attention mechanism by the PyTorch-like Pseudocode.

Algorithm 1 Selective self-attention mechanism: PyTorch-like Pseudocode

```
# x: input instances => Tensor([B, N+1, D])
# attn_weights: MSA weights => Tensor([B, num_heads, N
    +1, N+1])
# weighting_factor: the weight factor according to Eq
    .(10)

def S_MSA(x, attn_weights, weighting_factor):
    B, _N, D = x.shape
    N = _N - 1 # reduce the class-instance
    A_S = attn_weights
    # obtain the attention scores for
    # bag representation to all the other instances
    S = attn_weights[..., 0, 1:]
    # sum across all heads with attention scores by
        column
    S = einsum("bhn->bn", S)
    # Normalize the attn scores of all instances to 1
    S = S / (S.sum(dim=-1) + eps)
    # sort the instances and get the indices of the
        instance
    # corresponding to the value of the attention score
    ids_restore = argsort(argsort(S, dim=1), dim=1)
    # keep the indices of informative instances and drop
        others
    TH = int((1 - weighting_factor)*N)
     _remain, _reduce = ones(1, N), zeros(1, N)
    mask = where(ids_restore >= TH, _remain, _reduce)
    # mask the negative instances (element-wise
        production)
    A_S[:,:,:,1:] = A_S[:,:,:,1:] * mask.reshape(B, 1, N
        , 1)
    A_S[:,:,1:,:] = A_S[:,:,1:,:] * mask.reshape(B, 1,
        1, N)

    return A_S # return the selected attention weights;
```

Notes: `einsum` denotes the Einstein summation convention.

Summarized Advantages: For the first time, SiT with a selective self-attention mechanism is proposed to identify and select representative instances for the representation of fine-grained histopathological images. The proposed SiT block conducts instance scoring and selection by adaptive entrusting each instance with a score of significance based on the selective self-attention mechanism.

3.2. Multiple instance feature decoupling (MIFD)

The MIFD (Fig. 5(a)) learns the fine-grained representation by modeling the instance-to-bag interactions with the selected instances and bag-level prior knowledge for histopathological subtype prediction. The MIFD progressively integrates the instance-level features into the fine-grained representation with the idea of feature decoupling, which reduces the correlation between the selected instance-level feature X and bag representation Z (obtained in Section 3.1) by the mutual information upper-bound estimation (*i.e.*, mutual information minimization). Specifically, the instance feature decoupling (IFD) measures the mutual information between individual instance-level feature X and bag representation Z to encode the fine-grained description into the bag representation. At the different stages of the SiT block, the IFD reduces the correlation between bag representation and instance-level features with the instance-to-bag interaction, preventing performance degradation of histopathological subtype prediction.

3.2.1. Instance feature decoupling (IFD)

The IFD (Fig. 5(b)) aims to decouple noisy or unessential instance-level features from fine-grained bag representation by an effective MI minimization. Usually, the conventional MIL decision is inefficient due

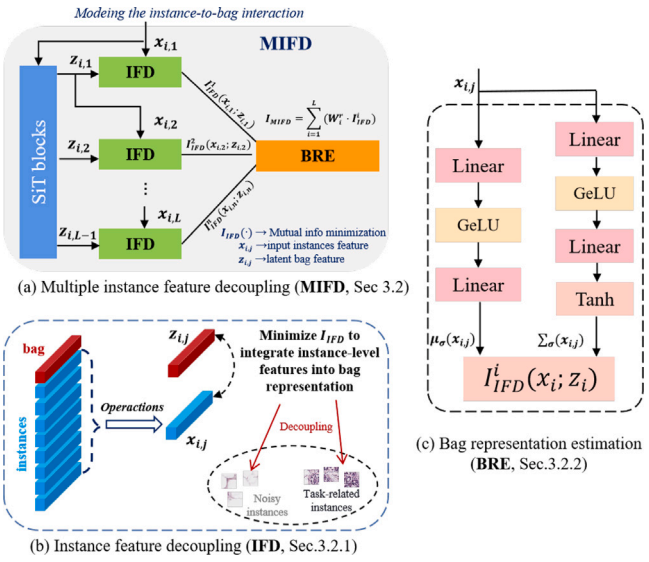


Fig. 5. The MIFD learns the fine-grained representation of histopathological images for subtype prediction by integrating the discriminative instance-level features into bag representation with the instance-to-bag interaction. It reduces the correlation between bag-level representation and task-independent information of input instance features in the instance-to-bag interaction, preventing performance degradation of histopathological subtype prediction.

to the existence of task-irrelevant instance features, which obstruct the effective instance-to-bag interactions for discriminative representation. Therefore, the IFD reduces the inherent relationships between the task-irrelevant input space and discriminatory information of the latent features by MI minimization. To better learn and fit the feature distribution between input space and latent space, the *operations*, including transpose, average pooling, and flatten, is leveraged to transform input instances feature into an instance embedding. Precisely, the input feature consists of the instances feature and bag feature, here the input instance-level feature $x_i \in \mathbb{R}^{N\times D}$ is stripped from the input feature, and transformed into $x_i \in \mathbb{R}^{1\times D}$ by a series of operations.

Inspired by the MI upper-bound estimation method (Cheng et al., 2020), the IFD is developed to integrate the discriminative instance-level feature into fine-grained bag representation, while eliminating the noisy features based on the instance-to-bag interaction. Considering the x_i as the input instance-level feature before the SiT $block_i, j \in [1, L]$, and the $z_i \in \mathbb{R}^{1\times D}$ indicates the latent bag representation (in Section 3.1.3 **Z**). Meanwhile, the z_i is treated as the input instance-level feature x_{i+1} for the next SiT block. As discussed earlier, it is needed to keep more representative instances in the early stage and fewer in the later stage. Therefore, the weighted mechanism (Eq. (8)) is also applied to the MI upper-bound estimation of different SiTs.

3.2.2. Bag representation estimation (BRE)

The bag representation of the histopathological image is formulated by a Gaussian distribution parameterized with a fully connected neural network. Given the sample pairs $\{(\boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^{N}$, $\boldsymbol{x}_i$ represents the input instance-level feature and $\boldsymbol{z}_i$ represents the latent representation of the histopathological image (bag) learned by the ith SiT block. For the training dataset, given input instance-level feature x_i, the conditional probability distribution $p(\boldsymbol{z}|\boldsymbol{x})$ is approximated by the variational distribution $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ with parameter θ. The variational approximation $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ is treated as Gaussian distribution parameterized for bag representation by neural networks. Formally, $q_\theta(z|x) = \mathcal{N}\left(\boldsymbol{z}|\mu_\sigma(\boldsymbol{x}), \Sigma_\sigma(\boldsymbol{x})\right)$, where μ_σ and Σ_σ are two linear combinations of fully-connected neural networks (Fig. 5(c)).

The BRE transforms the input instance-level feature $\boldsymbol{x}$ into two different distributions by the two fully connected neural networks, then constructs the positive and negative samples in the instance-to-bag interaction. Note that the mutual information is symmetric, *i.e.*, $I(x; z) = I(z; x)$. The estimation of IFD $I_{\text{IFD}}(\boldsymbol{x}; \boldsymbol{z})$ can be formulated as:

$$I^i_{\text{IFD}}(\boldsymbol{x}_i; \boldsymbol{z}_i) = \frac{1}{N}\sum_{j=1}^{N}\left[\log q_\theta(\boldsymbol{z}_{i,j}|\boldsymbol{x}_{i,j}) - \log q_\theta(\boldsymbol{z}_{k'_{i,j}}|\boldsymbol{x}_{i,j})\right] \tag{14}$$

where $\boldsymbol{x}$ and $\boldsymbol{z}$ are random vectors. N denotes the batch size. L denotes the number of SiT blocks. $q_\theta(\boldsymbol{z}_{i,j}|\boldsymbol{x}_{i,j})$ is the probabilities of all negative pairs. $q_\theta(\boldsymbol{z}_{k'_{i,j}}|\boldsymbol{x}_{i,j})$ is an unbiased estimation with a randomly sampled negative pair, $z_{k'_{i,j}}$ is uniformly selected from a set of positive integers $\{1, 2, \ldots, N\}$. Therefore, the overall mutual information $I_{\text{MIFD}}(\boldsymbol{x}; \boldsymbol{z})$ is estimated as follows to obtain the fine-grained bag representation.

$$I_{\text{MIFD}} = \sum_{i=1}^{L}(\mathcal{W}^r_i \cdot I^i_{\text{IFD}}) \tag{15}$$

Summarized Advantages: The MIFD learns the fine-grained representation of the histopathological image by progressively integrating the informative and discriminative instance-level feature into the bag representation based on the instance-to-bag interactions. The MIFD innovatively formulates the bag representation in MIL as instance-to-bag interaction with mutual information upper-bound estimation for fine-grained representation learning.

3.3. Loss function with information bottleneck

The loss function with information bottleneck (IB) models the effective instance-to-instance and instance-to-bag interactions to guide the model to learn a minimum adequate discriminative fine-grained histopathology representation.

The IB (Tishby et al., 2001; Shamir et al., 2010; Alemi et al., 2016) is an information-theoretic learning principle for latent representation learning, which restricts the information flow, forcing the model to localize the discriminatory information (Wang et al., 2023). Given an input space X and a corresponding ground-truth space Y, the IB aims to maximize the mutual information between the latent feature Z of the histopathology prediction task and ground-truth Y and minimize the mutual information between the input feature X and Z. In a nutshell, the IB seeks to find the sufficient representation of x concerning y, with minimum information used from x. The objective of the IB is summarized as:

$$min\{\underbrace{-I(Y;Z)}_{\text{instance-to-instance}} + \beta \underbrace{I(X;Z)}_{\text{instance-to-bag}}\} \tag{16}$$

where β is a Lagrange multiplier that controls the trade-off between the performance of Z on Y and the complexity of Z.

To make a precise prediction of histopathological subtype, the $\max I(Y; Z)$ is equivalent to minimize cross-entropy (CE) loss $\mathcal{L}_{CE}$ (Alemi et al., 2016), which can be formulated as:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j}\log p_{i,j} \tag{17}$$

where N denotes the sample number, and C denotes the number of the histopathology categories. $y_{i,j}$ represents the ground truth label, and $p_{i,j}$ represents the prediction of bag representation.

Since the global features contain task-irrelevant information, to improve the robustness of the informative instance features and strengthen the effectiveness of instance-to-instance relationship modeling, the I_{MIFD} (Eq. (14)–(15)) is the MI upper-bound estimation for $I(Z; X)$, which is inserted into the information bottleneck loss function. Hence the final loss function is summarized as:

$$\begin{aligned}\mathcal{L}_{IB} &= I(Y;Z) - \beta I(X;Z)\\ &= \mathcal{L}_{CE} + \beta \cdot I_{\text{MIFD}}\end{aligned} \tag{18}$$

Summarized Advantages: The loss function with information bottleneck is well-designed to conduct the instance-to-bag and instance-to-instance interactions to guide the model to learn a minimal and

necessary discriminative representation for accurate histopathological subtype prediction.

3.4. Algorithm of MIST

Given a training set $D = \{(x_i, y_i)\}_{i=1}^{N}$, learning a model with the MIST method (described in Section 3) leads to the minimizing of the information bottleneck loss function and then iterative updating of SiT and the MIFD. The training procedure requires estimating the parameters of the SiTs W and updating the parameters of the MIFD W^{σ}. The learning procedure of the MIST is iterative with four stages: SiT, MIFD, information bottleneck loss function for sufficient informative representation, and optimizing and updating the parameters of SiT and MIFD network. Algorithm 2 summarizes the detailed procedure of training for the MIST.

Algorithm 2 : Training procedure of the proposed MIST

Input : Training set D and its corresponding ground truth; maximum epoch T; Lagrange multiplier β;
Output: Parameters of SiT: W; Parameters of MIFD: W^{σ}
Initialization: $l = 1$; randomly initialize W, W^{σ}
while $l < T$ **do**
  Shuffle the training set D, and fetch mini-batch D_n `/* Forward */`
  for $i \leftarrow 1$ **to** N **do**
    Obtain the fine-grained feature z_i of the sample x_i by *SiT* module in Sec.3.1
    `/* MIFD estimation */`
    Estimate the $I^i_{\mathrm{IFD}}(x_i; z_i)$ of global-local fine-grained z_i and sample x_i by Eq. (14)
    Compute the I_{MIFD} of cumulative sum of I^i_{IFD} from all SiTs by Eq.(15)
    Compute the cross-entropy loss $\mathcal{L}_{CE}$ between the fine-grained z_i and ground truth y_i
    `/* information bottleneck loss */`
    Calculate the information bottleneck loss by:
    Eq.(18): $\mathcal{L}^i_{IB} = \mathcal{L}^i_{CE} + \beta \cdot I^i_{\mathrm{MIFD}}$
  end
  Calculate the summarized loss $\mathcal{L}_{IB} = \sum_{i=1}^{N} \mathcal{L}^i_{IB}$ by Eq. (18)
  `/* Backward */`
  Compute the gradients and optimize the parameters W of SiT and parameters W^{σ} of MIFD, respectively
  `/* optimizing SiT network with Adam */`
  `/* Update */`
  Update W with back propagation from $\mathcal{L}_{IB}$ (Eq. (18))
  Update W^{σ} with back propagation from I_{MIFD} (Eq. (15));
end
return W,W^{σ}

4. Experiments

4.1. Datasets

In this section, five widely used and competitive clinical datasets, NCT-CRC-HE (Kather et al., 2019), BreakHis (Brancati et al., 2022), BRACS (Brancati et al., 2022), Camelyon16 (Bejnordi et al., 2017) and private data from Fujian Medical University Union Hospital, are employed to evaluate the effectiveness of MIST on histopathological subtype prediction. These histopathology datasets contain multiple categories with complex and ambiguous information especially the images acquired from a large number of patients encompassing large variability, which is difficult to distinguish their categories in clinical practice.

Table 1
Experiment results illustrate MIST achieves advanced histopathological subtype prediction performance on five challenging datasets.

Dataset	ACC	AUC	F1	REC	PRE
NCT-CRC-HE	0.935	0.996	0.910	0.916	0.909
BreaKHis	0.922	0.997	0.907	0.905	0.913
BRACS	0.546	0.876	0.532	0.542	0.539
Camelyon16	0.893	0.956	0.892	0.893	0.893
Private	0.818	0.951	0.711	0.691	0.741

NCT-CRC-HE is H&E-stained colorectal cancer tissue slides from four patient cohorts (NCT biobank), containing 107,180 WSI patches, divided into 93,000 training images, 7000 validating images, and 7180 test images. All images are 224 × 224 pixels at 0.5 μm per pixel (MPP). The tissue classes are adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM).

BreakHis is composed of 7909 microscopic images of breast tumor subtypes collected from 82 patients using different magnifying factors 40× (1995), 100× (2081), 200× (2013) and 400× (1820). All images are 700 × 460 pixels and divided into 6327 images for training, 791 images for validating, and 791 images for testing. The subtype classes are adenoids (ADE), ductal carcinoma (DC), fibroadenoma (FIB), phyllodes tumor (PT), tubular adenoma (TA), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC).

BRACS consists of 547 WSIs extracted on 189 different patients. It also includes 4539 RoIs extracted from 387 WSI collected on 151 patients. All images were scanned at 0.25 μ/pixel using a magnification factor of 40 × and randomly separated into 3657 training images, 312 validating images, and 570 test images. The BRACS is challenging data for histopathological subtyping with seven subtypes, *i.e.*, normal tissue (N), pathological benign (PB), usual ductal hyperplasia (UDH), flat epithelial atypia (FEA), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), invasive carcinoma (IC). Due to the atypical pathological subtypes and their small data volume and large intra-class differences, existing methods have achieved quite low performance.

Camelyon16 is a benchmark dataset for breast cancer analysis, including 270 training sets and 130 test sets. After pre-processing, a total of about 15,403 annotated bag-level patches with 1280 × 1280 resolution were obtained.

Private data from Fujian Medical University Union Hospital contains approximately 10,000 manually processed and annotated 512 × 512-sized histopathological subtype patches of colorectal cancer collected from Fujian Medical University Union Hospital, with eight subtype categories.

4.2. Evaluation criteria

To fair quantitative comparison with other methods, the experiments are evaluated with the widely-used evaluation metrics via macro-average on multi-class: Accuracy (ACC) = $\frac{TP+TN}{TP+TN+FP+FN}$, Recall (REC) = $\frac{TP}{TP+FN}$, Precision (PRE) = $\frac{TP}{TP+FP}$, F1-score (F1) = $\frac{2\times REC\times PRE}{REC+PRE}$, and Area Under Curve (AUC). Here, *TP*, *TN*, *FP*, and *FN* are the numbers of true positive, true negative, false positive, and false negative, respectively. Note that, REC measures the performance at detecting the positives, which is significant to evaluate how excellent a model is at classifying atypical cases, especially rare cases. However, the indicators of PRE and REC sometimes appear contradictory, so they need to be considered comprehensively, which is frequently used by the F1-score.

In addition, the statistical significance of the proposed method versus ground truth is examined by the paired t-test with a significance level of 0.1%. The *p*-value for each pair of measurements is computed to demonstrate the significant improvement of the MIST. A lower *p*-value than 0.001 indicates that the method achieves the quantification of the histopathological subtype prediction with no significant differences.

Table 2
Performance of the MIST under different configurations for histopathological subtype prediction with five evaluation criteria on three datasets.

Dataset		w/*S-MSA*	w/*IB*	ACC	AUC	F1	REC	PRE
NCT-CRC-HE	Baseline (ViT-B/16)	✗	✗	0.913	0.987	0.889	0.893	0.896
	Baseline (ViT-B/16)	✓	✗	0.918	0.992	0.895	0.899	0.903
	Baseline (ViT-B/16)	✗	✓	0.924	0.992	0.900	0.900	0.903
	Baseline (ViT-B/16)	✓	✓	**0.928**	**0.993**	**0.908**	**0.914**	**0.906**
BreaKHis	Baseline (ViT-B/16)	✗	✗	0.898	0.994	0.878	0.876	0.883
	Baseline (ViT-B/16)	✓	✗	0.909	0.994	0.895	0.889	0.903
	Baseline (ViT-B/16)	✗	✓	0.902	0.993	0.893	0.874	0.893
	Baseline (ViT-B/16)	✓	✓	**0.912**	**0.995**	**0.900**	**0.895**	**0.905**
BRACS	Baseline (ViT-B/16)	✗	✗	0.463	0.811	0.419	0.460	0.428
	Baseline (ViT-B/16)	✓	✗	0.465	0.801	0.470	0.464	0.481
	Baseline (ViT-B/16)	✗	✓	0.491	0.816	0.467	0.488	0.473
	Baseline (ViT-B/16)	✓	✓	**0.519**	**0.826**	**0.492**	**0.518**	**0.495**

We indicate the backbone prediction network without selective multi-head self-attention, multiple instance feature decoupling, and information bottleneck loss function as experimental *baseline*, selective multi-head self-attention as *S-MSA* and information bottleneck loss function with MIFD as *IB*.

4.3. Implementation details

The MIST framework was implemented with Pytorch and all experiments were conducted on a cluster with four NVIDIA Tesla T4 (16 GB) and four NVIDIA GeForce RTX 3090 GPUs. In the experiments, the ViT-B/16 is indicated as *baseline*. The initial learning rate is set to 1e−4 and the weight decay of 1e−5. The decay strategy of the learning rate is cosine annealing. Except for the scalability experiment in the ablation study, the rest of the bag-level histopathological images are resized to 224 × 224, and the instance size $P \times P = 16 \times 16$. Data augmentation, including random cropping, rotation, horizontal flipping, and color jitting, is used during training. The information bottleneck loss function (Eq. (18)) is minimized with Adam optimizer and 0.9 momentum. Furthermore, for those baselines that were originally proposed for processing whole slide images, we have adapted them in our experiments by sizing the input bags to the appropriate dimensions, *e.g.,* $224 \times 224, 256 \times 256$, as needed, then dividing the instances into equally sized or randomly selected $n \times n$ sized patches carrying the labels of the corresponding bags via CNN or ViT networks.

4.4. Comparison methods

In the experiments, the MIST is compared with three groups of methods:

- **Baseline methods**. Leveraging the basic prediction network (ViT-B/16) and cross-entropy loss as the ***baselines***.
- **Instance selection strategies**. To validate the effectiveness of the MIST, especially S-MSA, the experiments also compare it with the different fixed selection ratio methods and naive adaptive instance selection methods.
- **State-of-the-art methods**. Compared with state-of-the-art (SOTA) MIL methods on the histopathological subtype prediction (Ilse et al., 2018; Campanella et al., 2019; Lu et al., 2021; Li et al., 2021a; Chikontwe et al., 2022; Zhang et al., 2022; Li et al., 2021b; Shao et al., 2021; Yu et al., 2021), which achieve superior prediction accuracy on five aforementioned datasets.

5. Results and analysis

Extensive experiment results with high prediction accuracy, F1-score, AUC, Recall, and Precision demonstrate that the MIST gains superior histopathological subtype prediction performance. The effectiveness of the MIST framework is validated in three folds: (1) The best quantification results are examined on five challenging datasets: NCT-CRC-HE, BreaKHis, BRACS, private data, and Camelyon16. (2) The effectiveness of each component in the MIST is probed to demonstrate its capacity in histopathological subtype prediction. (3) The advantages of MIST over existing methods on histopathological subtype prediction are revealed compared with the state-of-the-art methods.

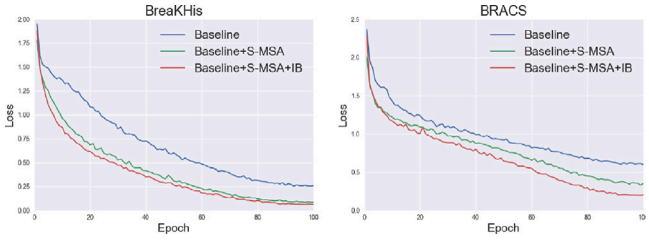


Fig. 6. The curve of train loss demonstrates significant improvements in the training convergence of the MIST with different modules on different datasets.

5.1. Prediction performance on five clinical datasets

Experimental results on NCT-CRC-HE As shown in Table 1, MIST delivers accurate histopathological subtype prediction on the NCT-CRC-HE dataset with the top performance on all the evaluation metrics with 0.935 of ACC, 0.996 of AUC, 0.910 of F1-score, 0.916 of Recall and 0.909 of Precision. The results indicate that the MIST well handles the aforementioned challenges and obtains accurate histopathological subtype prediction with multi-instance selection learning.

Experimental results on BreaKHis Extensive experiments are conducted on BreaKHis dataset with different magnifying factors 40×, 100×, 200×, and 400× (Table 1). The experimental results demonstrate that the MIST consistently achieves the best performance on histopathological subtype prediction. Especially for a dataset with the four magnifying factors, the MIST gets 0.922 for ACC, 0.997 for AUC, 0.907 for F1-score, 0.905 for Recall, and 0.913 for Precision.

Experimental results on BRACS An extensive experiment is conducted on the small dataset BRACS, which contains only 4539 images. Table 1 shows the results which demonstrate that the MIST can obtain the best performance on histopathological subtype prediction with a high 0.546 of ACC, 0.876 of AUC, 0.532 of F1-score, 0.542 of Recall, and 0.539 of Precision.

Experimental results on Camelyon16 To fairly contextualize the scalability of the MIST with large bag sizes, we conducted experiments on the Camelyon16 dataset. Table 1 depicts that our MIST achieves superior performance with the best result on all the evaluation metrics with 0.893 of ACC, 0.956 of AUC, 0.892 of F1-score, 0.893 of Recall and 0.893 of Precision.

Experimental results on private data Table 1 depicts that MIST achieves the best performance on all evaluation metrics with 0.818 of ACC, 0.951 of AUC, 0.711 of F1-score, 0.691 of Recall, and 0.741 of Precision.

Table 3
Experimental results show that the MIST obtains the competitive performance compared with the SOTA methods on the NCT-CRC-HE dataset. Notes: the *bold* indicates the SOTA result, while the *underline* states the second-best result.

Method	Backbone	NCT-CRC-HE				
		ACC	AUC	F1	REC	PRE
CNN-based:						
AB-MIL (*ICML'18*) (Ilse et al., 2018)	LeNet5	0.886	0.992	0.853	0.861	0.863
Gated-AB-MIL (*ICML'18*) (Ilse et al., 2018)	LeNet5	0.891	0.993	0.864	0.873	0.877
RNN-MIL (*Nat.Med'19*) (Campanella et al., 2019)	ResNet34	0.924	0.991	0.900	0.908	0.904
CLAM (*Nat.Bio.Eng'21*) (Lu et al., 2021)	ResNet50	0.923	0.993	0.899	0.904	0.906
DS-MIL (*CVPR'21*) (Li et al., 2021a)	ResNet18	0.920	0.991	0.899	0.900	0.898
DTFD-MIL (*CVPR'22*) (Zhang et al., 2022)	ResNet50	0.925	0.993	0.898	0.903	0.901
FRMIL (*MICCAI'22*) (Chikontwe et al., 2022)	ResNet18	0.915	0.982	0.893	0.902	0.901
Token-based without CNN:						
Baseline (ViT) (Dosovitskiy et al., 2021)	ViT-B/16	0.913	0.987	0.889	0.893	0.896
MIL-VT (*MICCAI'21*) (Yu et al., 2021)	ViT-B/16	0.914	0.989	0.890	0.897	0.899
MIST (Ours)	ViT-B/16	0.928	0.993	0.908	0.914	0.906
Token-based with pre-trained CNN:						
Baseline (ViT) (Dosovitskiy et al., 2021)	ResNet50 + ViT-B	0.922	0.993	0.896	0.903	0.900
TransMIL (*NeurIPS'21*) (Shao et al., 2021)	ResNet50 + ViT-B	0.926	0.984	0.899	0.908	0.904
MIST (Ours)	ResNet50 + ViT-B	**0.935**	**0.996**	**0.910**	**0.916**	**0.909**

Table 4
Experimental results show that the MIST obtains the competitive performance compared with the SOTA methods on the BreaKHis dataset. Notes: the *bold* indicates the SOTA result, while the *underline* states the second-best result.

Method	Backbone	BreaKHis				
		ACC	AUC	F1	REC	PRE
CNN-based:						
AB-MIL (*ICML'18*) (Ilse et al., 2018)	LeNet5	0.733	0.948	0.637	0.616	0.687
Gated-AB-MIL (*ICML'18*) (Ilse et al., 2018)	LeNet5	0.728	0.953	0.644	0.628	0.679
RNN-MIL (*Nat.Med'19*) (Campanella et al., 2019)	ResNet34	0.859	0.986	0.834	0.828	0.846
CLAM (*Nat.Bio.Eng'21*) (Lu et al., 2021)	ResNet50	0.896	0.994	0.876	0.873	0.880
DS-MIL (*CVPR'21*) (Li et al., 2021a)	ResNet18	0.891	0.994	0.873	0.872	0.861
DTFD-MIL (*CVPR'22*) (Zhang et al., 2022)	ResNet50	0.894	0.994	0.875	0.878	0.877
FRMIL (*MICCAI'22*) (Chikontwe et al., 2022)	ResNet18	0.889	0.992	0.869	0.868	0.865
Token-based without CNN:						
Baseline (ViT) (Dosovitskiy et al., 2021)	ViT-B/16	0.898	0.994	0.878	0.876	0.883
MIL-VT (*MICCAI'21*) (Yu et al., 2021)	ViT-B/16	0.900	0.992	0.884	0.885	0.885
MIST (Ours)	ViT-B/16	0.912	0.995	0.900	0.895	0.905
Token-based with pre-trained CNN:						
Baseline (ViT) (Dosovitskiy et al., 2021)	ResNet50 + ViT-B	0.917	0.996	0.900	0.904	0.897
TransMIL (*NeurIPS'21*) (Shao et al., 2021)	ResNet50 + ViT-B	0.882	0.991	0.859	0.861	0.861
MIST (Ours)	ResNet50 + ViT-B	**0.922**	**0.997**	**0.907**	**0.905**	**0.913**

Table 5
Experimental results show that the MIST obtains the competitive performance compared with the SOTA methods on the BRACS dataset. Notes: the *bold* indicates the SOTA result, while the *underline* states the second-best result.

Method	Backbone	BRACS				
		ACC	AUC	F1	REC	PRE
CNN-based:						
AB-MIL (*ICML'18*) (Ilse et al., 2018)	LeNet5	0.446	0.806	0.398	0.443	0.408
Gated-AB-MIL (*ICML'18*) (Ilse et al., 2018)	LeNet5	0.488	0.813	0.448	0.484	0.453
RNN-MIL (*Nat.Med'19*) (Campanella et al., 2019)	ResNet34	0.511	0.834	0.505	0.510	0.505
CLAM (*Nat.Bio.Eng'21*) (Lu et al., 2021)	ResNet50	0.528	0.845	0.513	0.526	0.509
DS-MIL (*CVPR'21*) (Li et al., 2021a)	ResNet18	0.517	0.823	0.494	0.516	0.511
DTFD-MIL (*CVPR'22*) (Zhang et al., 2022)	ResNet50	0.515	0.819	0.493	0.522	0.510
FRMIL (*MICCAI'22*) (Chikontwe et al., 2022)	ResNet18	0.502	0.813	0.489	0.503	0.499
Token-based without CNN:						
Baseline (ViT) (Dosovitskiy et al., 2021)	ViT-B/16	0.463	0.811	0.419	0.460	0.428
MIL-VT (*MICCAI'21*) (Yu et al., 2021)	ViT-B/16	0.461	0.800	0.462	0.460	0.474
MIST (Ours)	ViT-B/16	0.519	0.826	0.492	0.518	0.495
Token-based with pre-trained CNN:						
Baseline (ViT) (Dosovitskiy et al., 2021)	ResNet50 + ViT-B	0.523	0.849	0.508	0.522	0.521
TransMIL (*NeurIPS'21*) (Shao et al., 2021)	ResNet50 + ViT-B	0.533	0.866	0.515	0.531	0.512
MIST (Ours)	ResNet50 + ViT-B	**0.546**	**0.876**	**0.532**	**0.542**	**0.539**

5.2. Ablation study

Several ablation experiments are conducted by comparing against the baselines to analyze the effectiveness of each module in MIST. From the results shown in Table 2, we have the following observations:

(1) Effectiveness of S-MSA. Two contrast experiments are conducted to demonstrate the effectiveness of the newly-designed selective multi-head self-attention (*S-MSA*) from two aspects:

Effectiveness for fine-grained feature mining. The ablation studies (Table 2) on three datasets illustrate that when the *S-MSA* is utilized,

Table 6
Experimental results show that the MIST obtains a competitive performance compared with the SOTA methods on the histopathological subtype dataset of private data from Fujian Medical University Union Hospital. Notes: the *bold* indicates the SOTA result, while the *underline* states the second-best result.

Method	Backbone	Private data				
		ACC	AUC	F1	REC	PRE
CNN-based:						
AB-MIL (*ICML'18*) (Ilse et al., 2018)	LeNet5	0.759	0.927	0.629	0.608	0.684
Gated-AB-MIL (*ICML'18*) (Ilse et al., 2018)	LeNet5	0.779	0.928	0.658	0.631	0.707
RNN-MIL (*Nat.Med'19*) (Campanella et al., 2019)	ResNet34	0.781	0.933	0.668	0.641	0.713
CLAM (*Nat.Bio.Eng'21*) (Lu et al., 2021)	ResNet50	0.792	0.951	0.677	0.674	0.709
DS-MIL (*CVPR'21*) (Li et al., 2021a)	ResNet18	0.801	0.954	0.679	0.676	0.719
DTFD-MIL (*CVPR'22*) (Zhang et al., 2022)	ResNet50	0.805	0.951	0.682	0.678	0.723
FRMIL (*MICCAI'22*) (Chikontwe et al., 2022)	ResNet18	0.799	0.929	0.678	0.673	0.711
Token-based without CNN:						
Baseline (ViT) (Dosovitskiy et al., 2021)	ViT-B/16	0.793	0.944	0.671	0.645	0.719
MIL-VT (*MICCAI'21*) (Yu et al., 2021)	ViT-B/16	0.798	0.949	0.692	0.669	0.732
MIST (Ours)	ViT-B/16	0.809	0.949	0.690	0.677	0.733
Token-based with pre-trained CNN:						
Baseline (ViT) (Dosovitskiy et al., 2021)	ResNet50 + ViT-B	0.795	0.941	0.680	0.663	0.718
TransMIL (*NeurIPS'21*) (Shao et al., 2021)	ResNet50 + ViT-B	0.811	0.951	0.694	0.672	0.728
MIST (Ours)	ResNet50 + ViT-B	**0.818**	**0.951**	**0.711**	**0.691**	**0.741**

Table 7
Experimental results show that the MIST obtains a competitive performance compared with the SOTA methods on the Camelyon16 dataset. Notes: the *bold* indicates the SOTA result, while the *underline* states the second-best result.

Method	Backbone	Camelyon16				
		ACC	AUC	F1	REC	PRE
CNN-based:						
AB-MIL (*ICML'18*) (Ilse et al., 2018)	LeNet5	0.813	0.889	0.813	0.822	0.824
Gated-AB-MIL (*ICML'18*) (Ilse et al., 2018)	LeNet5	0.806	0.890	0.806	0.816	0.820
RNN-MIL (*Nat.Med'19*) (Campanella et al., 2019)	ResNet34	0.855	0.941	0.854	0.858	0.855
CLAM (*Nat.Bio.Eng'21*) (Lu et al., 2021)	ResNet50	0.880	0.951	0.879	0.881	0.878
DS-MIL (*CVPR'21*) (Li et al., 2021a)	ResNet18	0.877	0.950	0.876	0.878	0.876
DTFD-MIL (*CVPR'22*) (Zhang et al., 2022)	ResNet50	0.879	0.952	0.879	0.878	0.880
FRMIL (*MICCAI'22*) (Chikontwe et al., 2022)	ResNet18	0.859	0.939	0.859	0.865	0.863
Token-based without CNN:						
Baseline (ViT) (Dosovitskiy et al., 2021)	ViT-B/16	0.817	0.900	0.817	0.821	0.819
MIL-VT (*MICCAI'21*) (Yu et al., 2021)	ViT-B/16	0.853	0.929	0.853	0.856	0.853
MIST (Ours)	ViT-B/16	0.882	0.950	0.881	0.881	0.882
Token-based with pre-trained CNN:						
Baseline (ViT) (Dosovitskiy et al., 2021)	ResNet50 + ViT-B	0.863	0.933	0.862	0.862	0.862
TransMIL (*NeurIPS'21*) (Shao et al., 2021)	ResNet50 + ViT-B	0.868	0.941	0.867	0.867	0.867
MIST (Ours)	ResNet50 + ViT-B	**0.893**	**0.956**	**0.892**	**0.893**	**0.893**

Table 8
Experiments of S-MSA Sensitivity with different stages on histopathological subtype dataset of private data from Fujian Medical University Union Hospital.

Stages	ACC	AUC	F1	REC	PRE
4	0.798	0.945	0.696	0.672	0.734
8	0.798	0.945	0.696	0.672	0.734
12	**0.803**	**0.947**	**0.700**	**0.677**	**0.736**
16	0.798	0.945	0.696	0.672	0.734
20	0.798	0.945	0.696	0.672	0.734

Table 9
Experiments of S-MSA Sensitivity with different heads on histopathological subtype dataset of private data from Fujian Medical University Union Hospital.

Heads	ACC	AUC	F1	REC	PRE
4	0.803	0.947	0.700	0.677	0.736
8	0.803	0.947	0.700	0.677	0.736
12	**0.803**	**0.947**	**0.700**	**0.677**	**0.736**
16	0.798	0.945	0.696	0.672	0.734
20	0.798	0.945	0.696	0.672	0.734

the experiment results exceed the results of *Baseline* on all evaluation metrics, which demonstrates that the *S-MSA* effectively captures the local fine-grained information to participate in the histopathological subtype prediction under the framework of MIST.

Convergence speed. Fig. 6 depicts that *Baseline* with *S-MSA* during training achieves the optimal convergence speed and optimal convergence point due to the *S-MSA* designing. Compared with *Baseline*, the downward trend of loss and convergence speed is faster. The results show that the *S-MSA* when leveraged in the MIST framework leads to remarkably superior results on all the evaluation metrics and convergence.

(2) S-MSA Sensitivity. One extensive experiment is conducted to demonstrate the S-MSA sensitivity in the proposed MIST. Tables 8 and 9 illustrate the experimental results of S-MSA Sensitivity on the histopathological subtypes of private data from Fujian Medical University Union Hospital. We conducted ablation experiments on different numbers of stages/heads. Specifically, Table 8 shows that the head number is fixed to 12, the S-MSA of MIST achieves the best results when the number of stages is 12. Moreover, the prediction results degrade as the number of stages increases or decreases. Similarly, Table 9 demonstrates that the stage number is fixed to 12, and the S-MSA of MIST also achieves the best performance when the number of heads is 12. The prediction results also degrade as the number of heads increases or decreases. Hence, the number of stages and heads is set to 12 in our experiments.

(3) Effectiveness of the information bottleneck with MIFD. This experiment illustrates the efficacy of the proposed information bottleneck loss function with MIFD (*IB*). The experimental results (Table 2) of

Table 10
Experiments of Scalability of different size of bags and instances on Camelyon16 dataset.

Size of bags	Size of instances	No. of instances	ACC	AUC	F1	REC	PRE
224 × 224	32 × 32	49	0.851	0.918	0.851	0.855	0.852
	16 × 16	196	**0.882**	0.950	**0.881**	**0.881**	0.882
384 × 384	32 × 32	144	0.873	0.943	0.878	0.834	**0.908**
	16 × 16	576	0.879	0.949	0.875	0.847	0.907
512 × 512	32 × 32	256	0.877	**0.951**	0.879	0.855	0.907
	16 × 16	1024	0.874	0.942	0.873	0.876	0.873

Baseline with *S-MSA+IB* exceed the results of *Baseline* on all evaluation metrics, which indicate the *IB* module is significant for eliminating the negative impact of task-irrelevant in instance-to-instance correlation based on S-MSA and learning a minimum adequate discriminative representation. The compared experimental results between *Baseline* and *Baseline* w/ *S-MSA+IB*, it can be observed that the *IB* achieves an average improvement of ACC with 1.4%, AUC with 0.1%, F1 with 2.2%, REC with 1.9%, and PRE with 2.2% on BreaKHis.

(4) Scalability of different bag sizes and instance sizes. To fairly contextualize the scalability of our approach, we conduct ablations on the Camelyon16 dataset. In our experiments, we provide three different bag image sizes (*i.e.*, 224 × 224, 384 × 384, 512 × 512), and their corresponding two different instance sizes (*i.e.*, 16 × 16, 32 × 32). Table 10 demonstrates that when the bag size is 224 × 224 and the instance size is 16 × 16, the MIST achieves the best results in three of five metrics with 0.882 of *ACC*, 0.881 of *F1* and 0.881 of *REC*. Specifically, MIST has the worst prediction results when there are 49 instances in a bag, and as the number of instances goes from 49 to 196, the experimental results increased by 3. 1%, 3. 2%, 3. 0%, 2. 6%, and 3. 0%, respectively, in all evaluation metrics *ACC*, *AUC*, *F1*, *REC* and *PRE*. However, as the number of instances in a bag increases from 196 to 256, the experimental results dropped by 0.5%, 0.2%, and 2.6% in the three evaluation metrics *ACC*, *F1*, and *REC* respectively. The other two evaluation metrics AUC and PRE increased by 0.1% and 2.5% respectively. Furthermore, when the number of instances goes from 256 to 576 and 1024, the prediction results are as follows: Firstly, the experimental results increased by 0.2% in *ACC*, and decreased by 0.2%, 0.4%, and 0.8% in *AUC*, *F1* and *REC* respectively. Secondly, the experimental results decreased by 0.5%, 0.7%, 0.2%, and 3.4% respectively in *ACC*, *AUC*, *F1*, and *PRE*, and increased by 2.9% in *REC*. In a nutshell, when our MIST handles larger bag sizes and a larger number of instances, we found that the Recall values are high but the Precision values are low, indicating that our MIST has high recognition of normal tissue samples and misclassifies some cancer-negative samples into positive normal samples. To further explore the reasons, a bag usually contains smaller cancer subtype lesion areas if the bag size and the number of instances are large, therefore the receptive fields of most instances are normal tissue areas, causing the cancer subtype area to be relatively small and easily misclassified.

5.3. Qualitative analysis on the instance selection of S-MSA

To explain the effectiveness of the S-MSA, Fig. 7 depicts the progressively discarded unessential instances in the WAIS of the S-MSA module, where the masked regions represent the unessential instances that are discarded. It shows the input image and the instance selection results after the 12 stages of ViT. The results demonstrate that the proposed MIST can gradually focus on the most discriminative regions in the image, where the dashed box represents the significant area of histopathology. Specifically, instances associated with discriminative histopathological regions are adaptively preserved, continuously eliminating the unessential information (*e.g.*, backgrounds located on the right side of the image in the second row). The WAIS in S-MSA obtains

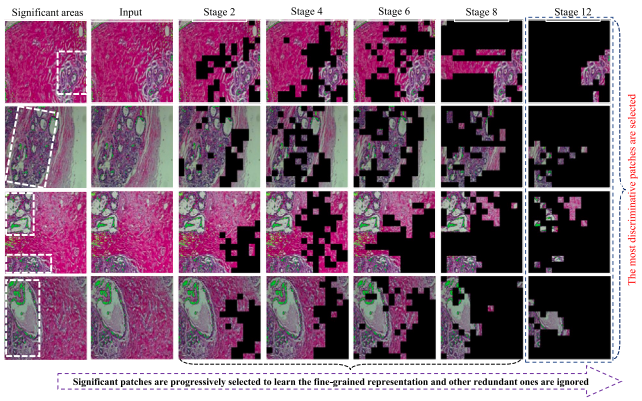


Fig. 7. Visualization results demonstrate that the MIST with WAIS of S-MSA progressively focuses on the most discriminative instances, where the dashed box represents the significant area of histopathology and the masked regions represent the unessential instances that are discarded after the 12 stages. This phenomenon indicates that the S-MSA has adequate interpretability.

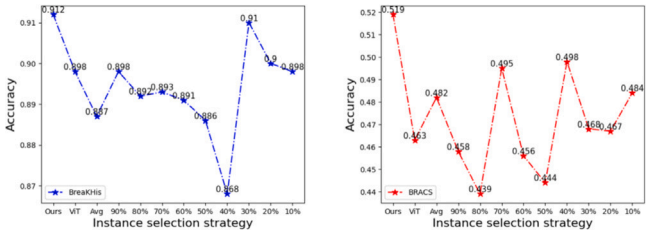


Fig. 8. The S-MSA of MIST yields the highest accuracy in histopathological subtype prediction by capturing the discriminative subtle features, compared with different instance selection strategies *i.e.*, fixed and average instance selection methods on BreaKHis and BRACS datasets.

the weighted fine-grained instance-level features from different stages (according to Eq. (8)) and then the S-MSA increases the weights of the corresponding instance score in the global histopathology feature for accurate histopathological subtype prediction (Eqs. (11)–(13)). This phenomenon indicates that the S-MSA has adequate interpretability.

5.4. Performance with different instance selection strategies

Quantitative comparison under different selection strategies. Table 11 and Fig. 8 demonstrate the performance of different instance selection strategies, which consist of fixed instance selection, average instance selection, and the S-MSA. Furthermore, the fixed instance selection is composed of various top-*k* selection ratios. Specifically, the top-*k* selection selects the *k* instances with the largest attention scores, and the rest are masked. With the gradual decrease of the top-*k* selection ratio, the performance of the top-*k* selection method varies according to the selection ratios. It means that the number of different instances of sampling varies from image to image. To alleviate the reliance on the fixed selection ratio, a naive adaptive method (*i.e., Average* method) is further proposed, which calculates the attention score of all instances to get the average value, and keeps the instances that exceed the average value, is that the average instance selection. Due to the attention scores of instances being less discriminatory in the early stages of ViT, the average instance selection can easily discard most instances that are informative for later stages. It should be noted that the proposed MIST w/ S-MSA module captures the discriminative subtle representation and yields high efficacy in histopathological subtype prediction (Fig. 8 and Table 11). The MIST w/ S-MSA has stronger robustness than the second-best results of other instance selection strategies, the MIST w/ S-MSA brings

Table 11
Experimental results under different instances selection strategies on BreaKHis and BRACS datasets.

Dataset	Backbone	Instance selection strategy	ACC	AUC	F1	REC	PRE
BreaKHis	ViT-B/16	Top-100% (*baseline*)	0.898	0.994	0.878	0.876	0.883
	ViT-B/16	Top-90%	0.898	0.995	0.881	0.868	0.897
	ViT-B/16	Top-80%	0.892	0.992	0.872	0.863	0.883
	ViT-B/16	Top-70%	0.893	0.994	0.876	0.868	0.884
	ViT-B/16	Top-60%	0.891	0.992	0.866	0.862	0.871
	ViT-B/16	Top-50%	0.886	0.993	0.864	0.860	0.869
	ViT-B/16	Top-40%	0.868	0.992	0.839	0.837	0.846
	ViT-B/16	Top-30%	0.910	**0.996**	0.894	0.892	0.897
	ViT-B/16	Top-20%	0.900	0.993	0.879	0.868	0.892
	ViT-B/16	Top-10%	0.898	0.993	0.874	0.865	0.884
	ViT-B/16	Average	0.887	0.993	0.867	0.862	0.875
	ViT-B/16	**MIST w/ S-MSA** (Ours)	**0.912**	0.995	**0.900**	**0.895**	**0.905**
BRACS	ViT-B/16	Top-100% (*baseline*)	0.463	0.811	0.419	0.460	0.428
	ViT-B/16	Top-90%	0.458	0.818	0.440	0.457	0.456
	ViT-B/16	Top-80%	0.439	0.795	0.442	0.438	0.452
	ViT-B/16	Top-70%	0.495	0.824	0.481	0.493	0.503
	ViT-B/16	Top-60%	0.456	0.820	0.441	0.456	0.452
	ViT-B/16	Top-50%	0.444	0.808	0.428	0.443	0.452
	ViT-B/16	Top-40%	0.498	0.825	0.491	0.497	0.505
	ViT-B/16	Top-30%	0.468	0.820	0.471	0.468	0.488
	ViT-B/16	Top-20%	0.467	0.822	0.463	0.465	0.469
	ViT-B/16	Top-10%	0.484	0.823	0.464	0.483	0.464
	ViT-B/16	Average	0.482	0.815	0.479	0.481	0.482
	ViT-B/16	**MIST w/ S-MSA** (Ours)	**0.519**	**0.826**	**0.492**	**0.518**	**0.495**

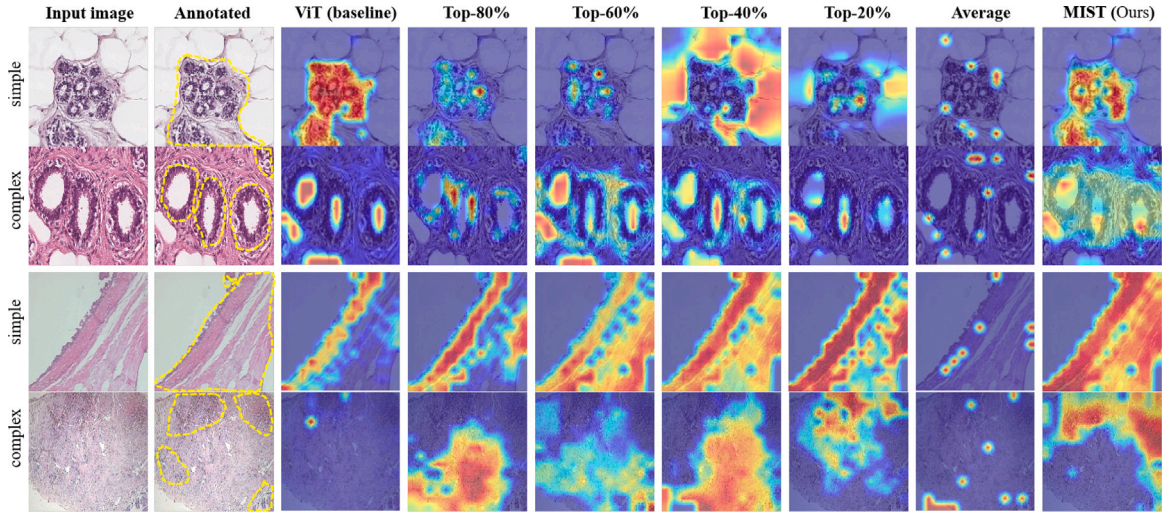


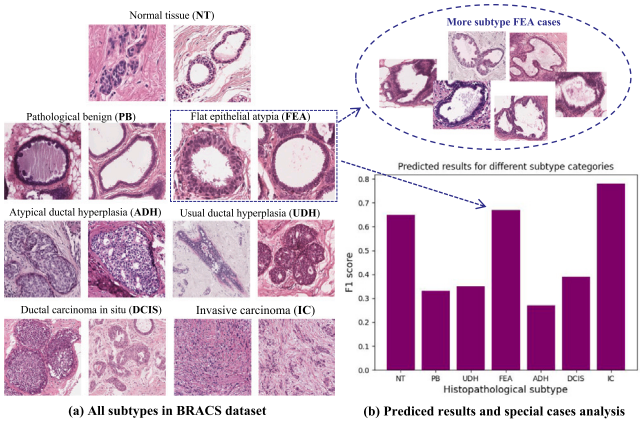
Fig. 9. Visualization results illustrate that owed to the proposed selective self-attention mechanism, the MIST can effectively identify the most informative regions in images with either simple informative regions (in the first and third rows) or complex informative regions (in the second and last rows), comparing with different instance selection methods. Note that the second-column annotated images with dashed polygon represent the significant histopathological regions.

obvious improvement of histopathological subtype prediction performance on BRACS dataset with the improvement of 2.1%, 0.1%, 0.1%, 2.1%, and 0.7% in terms of ACC, AUC, F1, Recall, and Precision, respectively.

Visualization results of comparisons with different instance selection strategies. To interpret the prediction performance of different instance selection strategies, including the fixed selection ratio (*i.e.,* Top-K) and average selection method, visualization results via Grad-CAM (Selvaraju et al., 2017) on simple and complex images (Fig. 9) illustrates that owe to the selective self-attention mechanism, the MIST can effectively identify the most informative regions in images with either simple informative regions (in the first and third rows) or complex informative regions (in the second and last rows), where the second-column annotated images with dashed polygon represent the significant histopathological regions. But other methods more or less partially ignore or misclassify informative regions in the image, especially the image of the last row containing plenty of cancer subtypes with a similar texture.

5.5. Analysis of special histopathological subtype cases

To further explore whether the proposed MIST can perform well on images with the background included in the center of the tissue, we conduct experiments on the BRACS dataset which is consistent with this characteristic. Fig. 10(a) illustrates several images of 7 different histopathological subtypes from the BRACS dataset, FEA subtype images usually have large background areas located in their centers. Moreover, Fig. 10(b) shows that the bar chart prediction results of our MIST for the seven different subtypes of the BRACS test set. Our MIST achieves the second-best result for special cases like the FEA subtype only to that of the simple case IC subtype which has small intra-class gaps between different image features, whereas other subtypes that do not contain a large background in the center of the image have bad prediction results. Some cases of the PB subtype also have the background region in its center, but it has a large variation in intra-class pathological features leading to a poorer prediction result. Therefore, Our MIST with an adaptive S-MSA mechanism does not suffer from



(a) All subtypes in BRACS dataset (b) Prediced results and special cases analysis

Fig. 10. Visualization of special histopathological subtype cases. (a) Several images of 7 different histopathological subtypes from the BRACS dataset, FEA subtype images usually have large background areas located in their centers. (b) Prediction results for the seven different subtypes of the BRACS test set with bar chart presentation, our MIST achieves the second-best result for special subtype cases whose background is included in the center.

Table 12
Experiments of Lagrange multipliers β with different assignments on BreaKHis dataset.

Parameter β	ACC	AUC	F1	REC	PRE
$\beta = 0.0001$	0.891	0.993	0.871	0.871	0.873
$\beta = 0.0003$	0.879	0.993	0.856	0.862	0.853
$\beta = 0.0005$	0.873	0.993	0.846	0.844	0.851
$\beta = 0.001$	**0.912**	**0.995**	**0.900**	**0.895**	**0.905**
$\beta = 0.003$	0.899	0.993	0.873	0.870	0.876
$\beta = 0.005$	0.888	0.993	0.861	0.864	0.860
$\beta = 0.01$	0.830	0.982	0.787	0.782	0.795
$\beta = 0.03$	0.727	0.953	0.645	0.636	0.677
$\beta = 0.05$	0.703	0.934	0.588	0.569	0.639

performance degradation due to the background included in the center of the tissue image.

5.6. Performance with different Lagrange multiplier β

To validate the performance of the MIST with different Lagrange multiplier β in information bottleneck loss function (Eq. (18)), extensive experimental results about β (Table 12) illustrate that decreasing β leads to better results, but the performance degenerate when β decreases from 0.001 to 0.0001. Hence the parameter β is assigned as 0.001 in all experiments to alleviate the task-irrelevant instances for accurate histopathological subtype prediction.

5.7. Performance comparison

Comparisons reveal the great superiority of the MIST for histopathological subtype prediction over existing MIL approaches. MIST compares with all the other state-of-the-art methods, including AB-MIL (Ilse et al., 2018), Gated-AB-MIL (Ilse et al., 2018), RNN-MIL (Campanella et al., 2019), CLAM (Lu et al., 2021), DS-MIL (Li et al., 2021a), FR-MIL (Chikontwe et al., 2022), DTFD-MIL (Zhang et al., 2022), TransMIL (Shao et al., 2021), ViT (Dosovitskiy et al., 2021) and MIL-VT (Yu et al., 2021) that have been tested on the NCT-CRC-HE dataset, BRACS dataset, BreaKHis dataset, private dataset, and Camelyon16 dataset, respectively. It can be observed from the compared results as follows.

(1) MIST with *Baseline* (ViT-B/16 w/ CNN) outperforms the state-of-the-art histopathological subtype prediction methods significantly on five challenging datasets. As shown in Table 3, the comparison results with the state-of-the-art method on the NCT-CRC-HE dataset show that MIST achieves the average improvement of 0.9%, 1.2%, 1.1%, 0.8%, 0.5% in terms of *ACC*, *AUC* and *F1*, *Recall* and *Precision*, respectively. Table 4 illustrates that with state-of-the-art tested on BreaKHis dataset, MIST also achieves the improvement of 0.5%, 0.1%, 0.7%, 0.1%, 1.6% in terms of *ACC*, *AUC*, *F1*, *Recall* and *Precision*, respectively. Comparing the results with others of Table 5, it clearly shows MIST obtains more accurate histopathological subtype prediction than other SOTA methods and achieves the improvement of 1.3%, 1.0%, 1.7%, 1.1%, 2.7% in terms of *ACC*, *AUC*, *F1*, *Recall* and *Precision*, respectively. Table 6 depicts that with state-of-the-art tested on the private dataset, MIST also achieves the improvement of 0.7%, 1.7%, 1.9%, 1.3% in terms of *ACC*, *F1*, *Recall* and *Precision*, respectively. Table 7 show that with state-of-the-art TransMIL method tested on the Camelyon16 dataset, MIST also achieves the improvement of 2.5%, 0.5%, 1.3%, 1.2%, and 1.3% in terms of *ACC*, *AUC*, *F1*, *Recall* and *Precision*, respectively. In a nutshell, MIST outperforms the other state-of-the-art methods.

(2) MIST with *Baseline* (ViT-B/16 w/o CNN) outperforms the best of existing token-based without CNN methods on histopathological subtype prediction. The comparison results with SOTA methods on NCT-CRC-HE dataset of Table 3 shows that the proposed MIST achieves the improvement of 1.2%, 0.3%, 1.6%, 1.0% and 2.0%in terms of *ACC*, *AUC*, *F1*, *Recall* and *Precision*, respectively. Table 4 illustrates that with SOTAs tested on the BreaKHis dataset, MIST also achieves the improvement of 1.4%, 0.4%, 1.8%, 1.7%, 0.7% in terms of *ACC*, *AUC*, *F1*, *Recall* and *Precision*, respectively. Moreover, comparing the results with others of Table 5, it clearly shows that MIST obtains more accurate histopathological subtype prediction than other SOTA methods and achieves the improvement of 5.6%, 1.5%, 7.3%, 5.8%, 6.7% in terms of *ACC*, *AUC*, *F1*, *Recall* and *Precision*, respectively. Table 6 depicts that with SOTAs tested on the private dataset, MIST also achieves the improvement of 0.4%, 0.1% in terms of *ACC*, *PRE*. The comparison results with SOTA methods on the Camelyon16 dataset of Table 7, MIST also achieves the improvement of 0.2%, 0.2%, and 0.2% in terms of *ACC*, *F1*, and *PRE*. In a nutshell, the MIST leads to remarkably superior results on all five challenging datasets.

Table 13
Statistical significance of MIST versus baseline model is examined by the *t*-test with a significance level of 0.1. A lower *p*-value than 0.001 indicates that the MIST achieves a significantly different performance than the baseline model.

Method	NCT-CRC-HE	BreaKHis	BRACS
MIST/TransMIL	3.12×10^{-9}	1.98×10^{-6}	4.02×10^{-3}
MIST/Baseline	1.64×10^{-6}	7.41×10^{-3}	7.71×10^{-4}
MIST/Baseline + *S-MSA*	9.11×10^{-4}	8.57×10^{-5}	2.38×10^{-7}
MIST/Baseline + *IB*	5.76×10^{-8}	6.26×10^{-3}	1.54×10^{-5}

5.8. Significant difference analysis

Statistical significance of MIST versus baseline model and second-best method TransMIL are examined by the t-test with a significance level of 0.1%. The p-values are computed to demonstrate the significant improvement of the MIST. A lower *p*-value than 0.001 indicates that the method achieves a significantly different performance than the *baseline* model and second-best method TransMIL. Table 13 indicates that MIST significantly outperforms *baseline* model and the previous TransMIL method. Moreover, the statistical significance versus previous TransMIL with a value smaller than 0.001 indicates a significant improvement of the MIST.

5.9. Complexity analysis

We now discuss the computational complexity of our proposed method. Suppose that the input bag $X \in \mathbb{R}^{h \times w \times c}$, multiplying the input bag X and the three coefficient matrices $M \in \mathbb{R}^{c \times c}$ to get the three vectors of $QKV \in \mathbb{R}^{hw \times c}$. Hence the computational complexity of the current stage is $\mathcal{O}(3hwc^2)$. Moreover, Q and K are multiplied

to obtain a matrix $A \in \mathbb{R}^{hw \times hw}$, so the complexity of current stage is $\mathcal{O}((hw)^2c)$. With the proposed S-MSA module, the selected attention matrix $\mathbf{A}^S \in \mathbb{R}^{hw \times hw}$ is obtained by multiplying the matrix A with the row attention masking $\in \mathbb{R}^{c \times hw}$ and column attention masking $\in \mathbb{R}^{hw \times c}$, respectively. The computational complexity is $\mathcal{O}(2(hw)^2c)$. By adding the matrices $\mathbf{A}$ and $\mathbf{A}^S$ and multiplying them with V, the computational complexity is $\mathcal{O}((hw)^2c)$. After the MLP layer processing, the computational complexity is $\mathcal{O}(hwc^2)$. Therefore, by adding the complexity computed at each stage above, the final computational complexity of each SiT block is $\mathcal{O}(4hwc^2 + 4(hw)^2c)$. Assume that given N pairs of samples, the computational complexity of multi-instance feature decoupling (MIFD) is $\mathcal{O}(N)$.

6. Conclusion and limitations

In this paper, the multi-instance selection transformer (MIST) framework is proposed to achieve histopathological subtype prediction for cancer prognosis. MIST designs a novel structure to simultaneously model the instance-to-bag and instance-to-instance interactions, creatively proposing a selective instance transformer (SiT) with selective multi-head self-attention (S-MSA) to progressively extract the significant instance-level feature in the instance-to-bag interactions for fine-grained representation, developing a multiple instance feature decoupling (MIFD) progressively learns the task-related representation by leveraging bag-level prior knowledge for redundancy alleviation that prevents prediction performance degradation by modeling the instance-to-bag interactions, constructing an information bottleneck loss function for guiding the network to learn a minimum adequate discriminative representation. Experimental results show that MIST is capable of achieving impressive performance for fine-grained prediction of the histopathological image. The proposed method has great potential in clinical cancer diagnoses.

The motivation of the proposed MIST is to achieve histopathological subtype prediction by adaptively identifying the representative instances from the bag for discriminative representation. It is a general transformer structure that can be used in any other visual scene not only histopathological subtype prediction. Inevitably, the MIST is unable to be directly used for WSI classification because of the large image resolution. A whole slide image with 150 000 × 150 000 pixels will be partitioned into 87 million 16 × 16 patches or 0.34 million 256 × 256 patches. Regardless of the partition method, the other challenge about the balance between complexity and effectiveness is arising. Therefore, the combination of the MIST and hierarchical structure is the future topic for the WSI classification problem.

CRediT authorship contribution statement

Rongchang Zhao: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Zijun Xi:** Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing – original draft. **Huanchi Liu:** Software. **Xiangkun Jian:** Software. **Jian Zhang:** Conceptualization, Methodology, Writing – review & editing. **Zijian Zhang:** Conceptualization, Writing – review & editing. **Shuo Li:** Writing – review & editing, Supervision.

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.

Data availability

All the data we used in this paper are publicly available except for the private data from Fujian Medical University Union Hospital.

Acknowledgments

We sincerely thank Jun Lu, and Binbin Xu with the Department of Gastric Surgery, Department of General Surgery, and Fujian Medical University Union Hospital for the testing data. This work is supported by the National Natural Science Foundation of China (62372474, 61702558), the National Key R&D Program of China (2020YFC2008500), the 111 Project (B18059), Hunan Science and Technology Project (2020SK2055), Natural Science Foundation of Hunan Province of China (2021JJ30879), and Postgraduate Scientific Research Innovation Project of Hunan Province (CX20220292). We are grateful for resources from the High Performance Computing Center of Central South University.

References

Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K., 2016. Deep variational information bottleneck. In: International Conference on Learning Representations.

Bang, S., Xie, P., Lee, H., Wu, W., Xing, E., 2021. Explaining a black-box by using a deep variational information bottleneck approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35, pp. 11396–11404.

Bardera, A., Feixas, M., Boada, I., Rigau, J., Sbert, M., 2007. Registration-based segmentation using the information bottleneck method. In: Iberian Conference on Pattern Recognition and Image Analysis. Springer, pp. 130–137.

Bardera, A., Rigau, J., Boada, I., Feixas, M., Sbert, M., 2009. Image segmentation using information bottleneck method. IEEE Trans. Image Process. 18 (7), 1601–1612.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 318 (22), 2199–2210.

Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al., 2022. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database 2022, baac093.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., Carin, L., 2020. Club: A contrastive log-ratio upper bound of mutual information. In: International Conference on Machine Learning. PMLR, pp. 1779–1788.

Chikontwe, P., Kim, M., Nam, S.J., Go, H., Park, S.H., 2020. Multiple instance learning with center embeddings for histopathology classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 519–528.

Chikontwe, P., Nam, S.J., Go, H., Kim, M., Sung, H.J., Park, S.H., 2022. Feature recalibration based multiple instance learning for whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 420–430.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.

Fayyaz, M., Koohpayegani, S.A., Jafari, F.R., Sengupta, S., Joze, H.R.V., Sommerlade, E., Pirsiavash, H., Gall, J., 2022. Adaptive token sampling for efficient vision transformers. In: European Conference on Computer Vision. Springer, pp. 396–414.

Feng, J., Zhou, Z.-H., 2017. Deep MIML network. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 31.

Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B., 2009. Histopathological image analysis: A review. IEEE Rev. Biomed. Eng. 2, 147–171.

Han, C., Lin, J., Mai, J., Wang, Y., Zhang, Q., Zhao, B., Chen, X., Pan, X., Shi, Z., Xu, Z., et al., 2022. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. Med. Image Anal. 80, 102487.

Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., Li, S., 2017. Breast cancer multi-classification from histopathological images with structured deep learning model. Sci. Rep. 7 (1), 1–10.

Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I., 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3852–3861.

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

Kather, J.N., Halama, N., Jaeger, D., 2018. Genomics and emerging biomarkers for immunotherapy of colorectal cancer. In: Seminars in Cancer Biology, vol. 52, Elsevier, pp. 189–197.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al., 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med. 16 (1), e1002730.

Kather, J.N., Poleszczuk, J., Suarez-Carmona, M., Krisam, J., Charoentong, P., Valous, N.A., Weis, C.-A., Tavernar, L., Leiss, F., Herpel, E., et al., 2017. In silico modeling of immunotherapy and stroma-targeting therapies in human colorectal cancer. Cancer Res. 77 (22), 6442–6452.

Kim, J., Kim, M., Woo, D., Kim, G., 2020. Drop-Bottleneck: Learning discrete compressed representation for noise-robust exploration. In: International Conference on Learning Representations.

Lai, Q., Li, Y., Zeng, A., Liu, M., Sun, H., Xu, Q., 2021. Information bottleneck approach to spatial attention learning. In: International Joint Conference on Artificial Intelligence.

Laleh, N.G., Muti, H.S., Loeffler, C.M.L., Echle, A., Saldanha, O.L., Mahmood, F., Lu, M.Y., Trautwein, C., Langer, R., Dislich, B., et al., 2022. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med. Image Anal. 79, 102474.

Lee, S.J., Yun, J.P., Choi, H., Kwon, W., Koo, G., Kim, S.W., 2017. Weakly supervised learning with convolutional neural networks for power line localization. In: 2017 IEEE Symposium Series on Computational Intelligence. SSCI, IEEE, pp. 1–8.

Li, B., Li, Y., Eliceiri, K.W., 2021a. 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. pp. 14318–14328.

Li, H., Yang, F., Zhao, Y., Xing, X., Zhang, J., Gao, M., Huang, J., Wang, L., Yao, J., 2021b. DT-MIL: Deformable transformer for Multi-instance learning on histopathological image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 206–216.

Lin, T., Xu, H., Yang, C., Xu, Y., 2022. Interventional multi-instance learning with deconfounded instance-level prediction. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, the Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. AAAI Press, pp. 1601–1609.

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

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano To Macro. IEEE, pp. 1107–1110.

Myronenko, A., Xu, Z., Yang, D., Roth, H.R., Xu, D., 2021. Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 329–338.

Pinheiro, P.O., Collobert, R., 2015. From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1713–1721.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV.

Shamir, O., Sabato, S., Tishby, N., 2010. Learning and generalization with the information bottleneck. Theoret. Comput. Sci. 411 (29–30), 2696–2711.

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

Shi, X., Xing, F., Xie, Y., Zhang, Z., Cui, L., Yang, L., 2020. Loss-based attention for deep multiple instance learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34, pp. 5742–5749.

Song, J., Zheng, Y., Wang, J., Ullah, M.Z., Li, X., Zou, Z., Ding, G., 2022. Multi-feature deep information bottleneck network for breast cancer classification in contrast enhanced spectral mammography. Pattern Recognit. 131, 108858.

Srinidhi, C.L., Kim, S.W., Chen, F.-D., Martel, A.L., 2022. Self-supervised driven consistency training for annotation efficient histopathology image analysis. Med. Image Anal. 75, 102256.

Thirion, B., Faugeras, O., 2004. Feature characterization in fMRI data: the Information Bottleneck approach. Med. Image Anal. 8 (4), 403–419.

Tishby, N., Pereira, C., Bialek, W., 2001. The information bottleneck method. In: Proceedings of the 37th Allerton Conference on Communication, Control and Computation. Vol. 49.

Tu, M., Huang, J., He, X., Zhou, B., 2019. Multiple instance learning with graph neural networks. arXiv preprint arXiv:1906.04881.

Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W., 2018. Revisiting multiple instance neural networks. Pattern Recognit. 74, 15–24.

Wang, Y., Ye, S., Yu, S., You, X., 2022. R2-Trans: Fine-grained visual categorization with redundancy reduction. arXiv preprint arXiv:2204.10095.

Wang, J., Zheng, Y., Ma, J., Li, X., Wang, C., Gee, J., Wang, H., Huang, W., 2023. Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation. Med. Image Anal. 83, 102687.

Yan, Y., Wang, X., Guo, X., Fang, J., Liu, W., Huang, J., 2018. Deep multi-instance learning with dynamic pooling. In: Asian Conference on Machine Learning. PMLR, pp. 662–677.

Yang, H., Kim, J.-Y., Kim, H., Adhikari, S.P., 2019. Guided soft attention network for classification of breast cancer histopathology images. IEEE Trans. Med. Imaging 39 (5), 1306–1315.

Yu, S., Ma, K., Bi, Q., Bian, C., Ning, M., He, N., Li, Y., Liu, H., Zheng, Y., 2021. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 45–54.

Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y., 2022. 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. CVPR, pp. 18802–18812.

Zhmoginov, A., Fischer, I., Sandler, M., 2020. Information-bottleneck approach to salient region discovery. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 531–546.

Zhu, H., Ke, W., Li, D., Liu, J., Tian, L., Shan, Y., 2022. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4692–4702.

Zuo, L., Dewey, B.E., Liu, Y., He, Y., Newsome, S.D., Mowry, E.M., Resnick, S.M., Prince, J.L., Carass, A., 2021. Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. NeuroImage 243, 118569.