This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

Multimodal Co-attention Fusion Network with Online Data Augmentation for Cancer Subtype Classification

Saisai Ding, Juncheng Li, Jun Wang, Member, IEEE, Shihui Ying, Member, IEEE, Jun Shi, Member, IEEE

***Abstract*—It is an essential task to accurately diagnose cancer subtypes in computational pathology for personalized cancer treatment. Recent studies have indicated that the combination of multimodal data, such as whole slide images (WSIs) and multi-omics data, could achieve more accurate diagnosis. However, robust cancer diagnosis remains challenging due to the heterogeneity among multimodal data, as well as the performance degradation caused by insufficient multimodal patient data. In this work, we propose a novel multimodal co-attention fusion network (MCFN) with online data augmentation (ODA) for cancer subtype classification. Specifically, a multimodal mutual-guided co-attention (MMC) module is proposed to effectively perform dense multimodal interactions. It enables multimodal data to mutually guide and calibrate each other during the integration process to alleviate inter- and intra-modal heterogeneities. Subsequently, a self-normalizing network (SNN)-Mixer is developed to allow information communication among different omics data and alleviate the high-dimensional small-sample size problem in multi-omics data. Most importantly, to compensate for insufficient multimodal samples for model training, we propose an ODA module in MCFN. The ODA module leverages the multimodal knowledge to guide the data augmentations of WSIs and maximize the data diversity during model training. Extensive experiments are conducted on the public TCGA dataset. The experimental results demonstrate that the proposed MCFN outperforms all the compared algorithms, suggesting its effectiveness.**

***Index Terms*—Cancer subtype, Multimodal learning, Co-attention, Whole slide images, Multi-omics data.**

I. INTRODUCTION

HISTOPATHOLOGICAL images are considered as the gold standard for cancer diagnosis [1], [2]. With the development of deep learning in computational pathology, the computer-aided diagnosis (CAD) for cancers with whole slide images (WSIs) has gained its reputation in recent years [3], [4]. Due to the huge size of WSIs, conducting pixel-level annotation for WSIs analysis becomes a challenging and time-consuming task. To address this challenge, the weakly-supervised learning frameworks, such as multiple instance learning (MIL), specifically tailored for CAD based on WSIs [5]–[7]. In MIL, a WSI is cropped into numerous patches as instances, and each WSI is considered as a bag. Then, the patch embeddings (instances) are extracted and aggregated to produce a slide-level prediction for different tasks, such as cancer grading, subtyping, and survival prediction[8]–[10].

In recent years, with the development of high-throughput sequencing technology, many studies have utilized WSIs and multi-omics data for more comprehensive cancer diagnosis[11]–[13]. It is known that WSIs can provide phenotypic information about cell types and tissues, while multi-omics data can also offer complementary information for identifying cancers [14]. However, a significant heterogeneity gap exists between WSIs and multi-omics data. For example, the WSIs are gigapixel images, while the multi-omics data are composed of thousands dimensional sequences. This disparity necessitates the development of multimodal approaches to effectively address the issue. Existing multimodal works generally employ the feature-based fusion strategies in deep learning, such as vector concatenation, element-wise summation, bilinear pooling (Kronecker Product), and co-attention techniques, to fuse different modality-level representations [15]–[21].

However, existing methods have not tapped the full potential of multimodal data to produce superior representations due to following limitations. 1) Modeling dense multimodal interactions imposes significant computational and memory requirements; 2) Multi-omics data generally come from multiple platforms with different representations and biological attributes [22], [23], but existing works usually simply concatenate standardized vectors from different omics data, and do not consider the distinctions and correlations among various omics data; 3) These multi-omics data typically have high-dimensional small-sample sizes (HDLSS) problem, which also poses a great challenge to the robustness of CAD models. Therefore, it is worth developing an effective approach to process multimodal data, and fully exploit the correlations between multimodal data for more efficient multimodal learning.

On the other hand, although multimodal data can provide a more comprehensive reflection of the health status of

This work is supported by This work is supported by the National Key R&D Program of China (2021YFA1003004), National Natural Science Foundation of China (62271298) and 111 Project (D20031). (Corresponding authors: Jun Shi)

S. Ding, J. Li, J. Wang and J. Shi are with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, China. (Email: junshi@shu.edu.cn)

S. Ying is with the Department of Mathematics, School of Science, Shanghai University, China.

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

patients, the costs of acquiring these different modalities are significant. Thus, the performance of deep learning models may be affected by insufficient multimodal patient data. Data augmentation is a simple yet effective technology in deep learning for improving data diversity[24]. It generally performs a series of random transformations on the original image to generate new samples, thereby improving the generalization ability and robustness of the model. Unfortunately, the data augmentation methods in the WSIs have not been fully explored, because it is extremely inefficient to use image processing operations, such as cropping, flipping, or shifting, for all patches in WSIs [25]. Although there are some data augmentation methods specifically designed for the MIL-based WSI classification, they generally only utilize knowledge from the WSI modality to obtain instance attention scores for importance ranking and instance alignment [26]–[31]. However, the attention scores may not always accurately reflect instance importance due to potential biases in WSI modality [32], [33]. Therefore, it is significant to incorporate useful knowledge from another modality to guide the data augmentations in WSI classification.

In this paper, we propose a novel co-attention multimodal fusion network (MCFN) with online data augmentation (ODA) for cancer diagnosis. Specifically, a novel multimodal mutual-guided co-attention (MMC) module is proposed to efficiently perform dense multimodal interactions and alleviate inter- and intra-modal heterogeneity. Subsequently, to better integrate multi-omics data, we develop a self-normalizing network (SNN)-Mixer to enable intra-modal interaction among multi-omics data. This allows for the exchange of information and the extraction of meaningful representations from the multi-omics data. Furthermore, to facilitate the model inference under insufficient training data, we incorporate a new ODA module into MCFN. The ODA module can utilize the useful knowledge of multimodal data to guide the data augmentation in WSI modality, thus improving the data diversity for model training. Extensive experiments are conducted on The Cancer Genome Atlas (TCGA) project, and the proposed MCFN outperforms state-of-the-art (SOTA) algorithms on the cancer subtype classification tasks.

The main contributions of this work are four-fold as follows:

1) We propose an effective MCFN for cancer subtype classification based on WSIs and multi-omics data. The MCFN can capture the correlations between different modalities and enhance feature representation in each modality.
2) We develop a novel MMC module that enables multimodal data to guide and calibrate each other to generate superior representations. MMC leverages the symmetry of the attention score matrix to simplify the calculation process, thereby reducing the computational cost of co-attention mechanism.
3) We propose an ODA module, a simple yet effective method for instance-level data augmentation for MIL-based WSI classification. The ODA module utilizes the multimodal information to divide instances into attentive and inattentive groups. It then fuses inattentive instances and matches similar attentive ones using cosine similarity to maximize data diversity.
4) We develop a new SNN-Mixer to learn the correlations among different omics features and to alleviate the HDLSS problem in multi-omics data. The SNN-Mixer employs two types of SNN layers to allow information communication across different dimensions of data, thus enhancing the interaction of features.

II. RELATED WORK

A. Multimodal Fusion of Histology and Genomics

Multimodal fusion via deep learning is the current clinical practice for many cancer types that seeks to correlate and combine disparate heterogeneous data modalities [34]. With the development of medical imaging technology and advanced genomic methods, many works have focused on integrating histology images and genomic data for more comprehensive cancer diagnosis [15]–[20]. For example, Vale-Silva et al. [16] used a concatenation operation to fuse histology and genomic features for survival prediction; Chen et al. [14] adopted a Kronecker product to fuse morphological and molecular information for cancer diagnosis and prognosis. Although these approaches can successfully fuse multimodal data, they usually utilized late fusion strategies and provide limited multimodal interactions.

Recently, several works have successfully adopted early fusion strategies to exploit the complementarity within different modalities. For example, Li et al. [35] proposed a Multi-modal Multi-instance (MMMI) model to generate a cross-modal representation for re-calibrating the features in each modality. Since MMMI used a non-linear layer to combine features from different modalities, the correlation of features lacked preference and attention mechanism, which was not conducive to information reduction and selection. Subsequently, Chen et al. [19] proposed a Multimodal Co-Attention Transformer (MCAT) for survival prediction, which used a genomic-guided co-attention layer to model correlations between histology and genomic features. However, the co-attention in MCAT is one-sided and only models the histology-to-genomic interaction. In addition, MCAT only simply concatenated standardized vectors from different omics data, and did not consider the distinctions and correlations among various omics data. Recently, Liu et al. [20] proposed a Mutual-Guided Cross-Modality Transformer (MGCT) that could combine histology features and genomic features to model the genotype-to-phenotype interactions. Hou et al. [36] also developed a Hybrid Graph Convolutional Network (HGCN) for multimodal survival prediction, which utilized the multimodal information of patients to realize intra- and inter-modal interactions between multimodal graphs. However, modeling dense multimodal interactions imposes significant computational and memory requirements, which brings great challenges to the efficient multimodal learning in the CAD model.

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

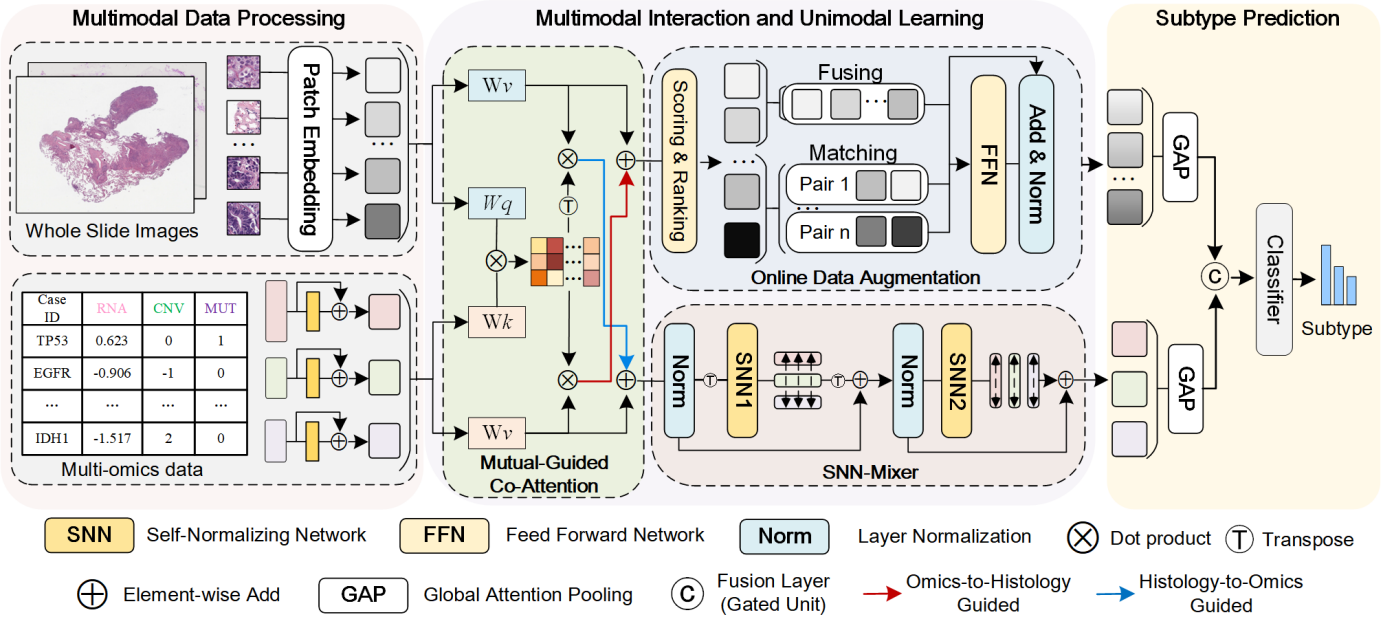


Fig. 1. Overview of MCFN for WSI classification. (a) Multimodal data processing for WSI and multi-omics bags construction. (b) Multimodal interaction and unimodal learning for online data augmentation and multi-omics feature learning. (c) A multimodal fusion layer that combines the histology and genomic features for subtype prediction.

B. Data Augmentation in MIL-based WSI Analysis

Data augmentation always improves the generalization ability and robustness of the deep models. In MIL, several works have adopted traditional image-level data augmentation functions, such as cropping, flipping, or shifting to generate diverse instance-level features for the same patch image [24], [25]. However, the traditional data augmentation methods are computationally expensive for the gigapixel images because a WSI generally typically consists of tens of thousands of patches.

In MIL-based WSI analysis, data augmentation could be roughly divided into three categories [32]: 1) instance-level augmentation, 2) bag-level augmentation, and 3) bag combination augmentation. The first category focuses on using bag prototypes [27], [30], generative adversarial networks [26], or diffusion models [29]. The second category mainly augments the entire bag by generating new subsets through hierarchical [30]or random sampling [31], rather than augmenting individual instances. The last category creates new bags by combining instances from different bags, typically selecting them randomly to introduce data diversity for improving the generalization of model [27], [33].

However, these methods only utilized knowledge within the WSIs to guide MIL-based data augmentation and overlooked potentially valuable complementary information from other modalities. While previous studies have employed multimodal information for guiding instance-level aggregation in MIL [19], [35], which involves leveraging the knowledge from another modality to distribute instance-level attention weights in the WSI modality optimally. To the best of our knowledge, our MCFN is the first work that incorporates useful multimodal knowledge to guide the instance-level data augmentations in MIL.

III. METHODOLOGY

In this work, a novel MCFN is proposed for cancer diagnosis using both WSIs and multi-omics data. Fig. 1 shows the overall pipeline of MCFN, which includes three main steps: (1) Multimodal data pre-processing for WSI and genomic bag construction; (2) Multimodal interaction and unimodal learning; and (3) Multimodal features fusion for cancer subtyping. In the following sections, we will introduce three steps in details.

A. Multimodal Data Processing

1) Preliminaries

MIL is a typical weakly supervised learning method in which the training data consists of a set of bags, and each bag contains multiple instances. Given a bag $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots \boldsymbol{x}_M\}$ with label $\boldsymbol{Y}$, the goal of MIL is to predict the bag-level label without the instance-level annotations. For our task, let $\boldsymbol{S} = \{\boldsymbol{X}_i, \boldsymbol{G}_i, \boldsymbol{Y}_i\}_{i=1}^{i=N}$ represents patient dataset, where $\boldsymbol{X}_i$ denotes WSI data of i-th patient, $\boldsymbol{G}_i$ is a matrix of multi-omics data attributes matched with $\boldsymbol{X}_i$, $\boldsymbol{Y}_i$ is the cancer subtype label of i-th patient, and N is the number of patients. Our goal is to develop a multimodal fusion network that integrates $\boldsymbol{X}_i$ and $\boldsymbol{G}_i$ to predict cancer subtyping.

2) WSI and Genomic Bag Construction

Before MIL, we should preprocess each WSI and multi-omics data into bags for training and testing. To represent WSI data as a bag data structure, we follow conventional MIL approach to crop non-overlapping 299 × 299 patches from WSIs, and a threshold is set to filter out background ones. After patching, we use a pre-trained TransPath [37] model to extract instance-level feature representations from these patches and convert each 299 × 299 patch into a feature vector. Finally, for M histology patches within $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots \boldsymbol{x}_M\}$, we stack the

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

extracted patch embeddings $\{\boldsymbol{h}_m \in \mathbb{R}^{d_k \times 1}\}_{m=1}^{M}$ into a bag $\boldsymbol{H}_{bag} \in \mathbb{R}^{M \times d_k}$.

To construct the genomic bag, we use multi-omics data that includes mutation status, copy-number variation, and RNA-seq expression. We first perform three pre-processing steps: outlier deletion, biological attributes alignment, and normalization. For RNA-seq expression, we select the top 2000 genes with the largest median absolute deviation to limit the number of features from RNA-seq. After that, we adopt the SNN [38] on biological attributes of each omics to obtain the genomics-based instance-level feature representations. Finally, for N omics data within $\boldsymbol{G} = \{\boldsymbol{g}_1, \cdots \boldsymbol{g}_N\}$, we stack the extracted genomic embeddings $\boldsymbol{g}_n \in \mathbb{R}^{1 \times d_k}$ into a bag $\boldsymbol{G}_{bag} \in \mathbb{R}^{N \times d_k}$. In our implementation, we use three different omics features (N=3) for multi-omics representation learning.

B. Multimodal Interaction and Unimodal Learning

Our MCFN is different from previous multimodal methods by employing a unique early fusion strategy. This approach enables the capture of local relevance between different modalities and enhances representation learning for each modality. As shown in Fig. 1, the MCFN comprises three main components: the MMC, ODA, and the SNN-Mixer modules. The MMC module initiates multimodal interactions that enables multimodal data to mutually guide and calibrate each other during the integration process. Subsequently, the ODA module comes into play in decoupling and matching instances based on attention scores obtained from MMC for data augmentation. Finally, the SNN-Mixer adopts two types of SNN layers to facilitate communication among different omics features, resulting in a more effective representation of multi-omics data. In the following subsections, we will provide a detailed introduction to each of the three main components.

1) Multimodal Mutual-guided Co-attention

After constructing bags that represent $\boldsymbol{H}_{bag}$ and $\boldsymbol{G}_{bag}$ for WSIs and multi-omics data, we aim to model dense pairwise interactions between patch embeddings and genomic embeddings. As shown in Fig. 1, there are significant heterogeneities not only between WSIs and multi-omics data, but also among different omics data due to their distinct representations and biological attributes. Therefore, the key idea of MMC is to add the co-attention mechanism during feature interaction, which facilitates information reduction and selection.

The co-attention mechanism is similar to the self-attention principle in that it maps query and key-value pairs to outputs. However, different from the self-attention principle that considers only one modality, the co-attention mechanism simultaneously computes attention scores for both modalities by generating query and key-value pairs from two different modalities. Specifically, our MMC models two interactions, one from histology to omics and the other from omics to histology interaction. Through multimodal interaction, the multimodal data can guide and calibrate each other, thereby alleviating inter- and intra-modal heterogeneity. More importantly, we can utilize the symmetry of the attention score matrix to simplify the calculation process, thus reducing the computational cost of the co-attention mechanism.

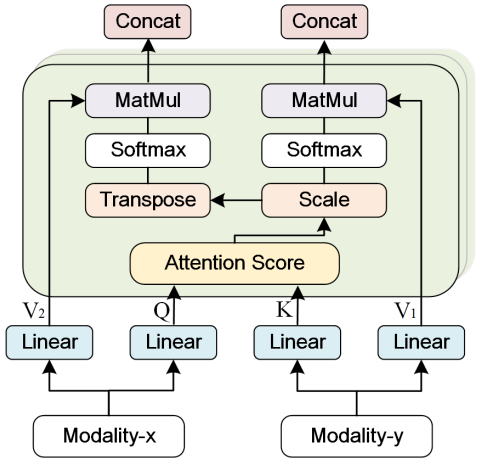


Fig. 2. The multi-head co-attention layer for two modalities.

As shown in Fig. 2, given a pair of bags for two modalities $\boldsymbol{H}_{bag} \in \mathbb{R}^{M \times d_k}$ and $\boldsymbol{G}_{bag} \in \mathbb{R}^{N \times d_k}$, the matrices of query $\boldsymbol{Q}$, key $\boldsymbol{K}$ and values $\boldsymbol{V}_1, \boldsymbol{V}_2$ are first calculated through four different linear projections. Assume the $\boldsymbol{G}_{bag}$ is the query and the $\boldsymbol{H}_{bag}$ is the key, the calculation process is as follows:

$$\begin{aligned} \boldsymbol{Q} &= Liner(\boldsymbol{G}_{bag}) = \boldsymbol{G}_{bag}\boldsymbol{W}_Q, \\ \boldsymbol{K} &= Liner(\boldsymbol{H}_{bag}) = \boldsymbol{H}_{bag}\boldsymbol{W}_K, \\ \boldsymbol{V}_1 &= Liner(\boldsymbol{H}_{bag}) = \boldsymbol{H}_{bag}\boldsymbol{W}_{V_1}, \\ \boldsymbol{V}_2 &= Liner(\boldsymbol{G}_{bag}) = \boldsymbol{G}_{bag}\boldsymbol{W}_{V_2}. \end{aligned} \tag{1}$$

where $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_{V_1}, and\ \boldsymbol{W}_{V_2} \in \mathbb{R}^{d_k \times d_m}$ are the corresponding weight matrices of linear projections.

Following that, we incorporate a multi-head structure to enhance the co-attention mechanism. Illustrated in Fig. 2, this structure allows us to project inputs into different subspaces, enabling the learning of various features through attention mechanisms, thereby improving the performance of the model. Specifically, the input features are evenly split into h parts, and the attention score matrix $\boldsymbol{S}_i \in \mathbb{R}^{N \times M}$ can be calculated as follows:

$$\boldsymbol{S}_i = Score(\boldsymbol{Q}_i, \boldsymbol{K}_i) = \frac{\boldsymbol{Q}_i \boldsymbol{K}_i^T}{\sqrt{d_m/h}}. \tag{2}$$

where $\boldsymbol{Q}_i \in \mathbb{R}^{N \times \frac{d_m}{h}}, \boldsymbol{K}_i \in \mathbb{R}^{M \times \frac{d_m}{h}}, and\ 1/\sqrt{d_m/h}$ is a scaling factor. To weight the features of the two modalities, we simply transpose the matrix $\boldsymbol{S}_i$ because of its symmetry:

$$\begin{aligned} \boldsymbol{V}_1' &= softmax(\boldsymbol{S}_i)\boldsymbol{V}_{1i} \\ \boldsymbol{V}_2' &= softmax(\boldsymbol{S}_i^T)\boldsymbol{V}_{2i}. \end{aligned} \tag{3}$$

where $\boldsymbol{V}_{2i} \in \mathbb{R}^{N \times \frac{d_m}{h}}, \boldsymbol{V}_{1i} \in \mathbb{R}^{M \times \frac{d_m}{h}}$, and the softmax function is used to normalize each row vector of the $\boldsymbol{S}_i$.

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

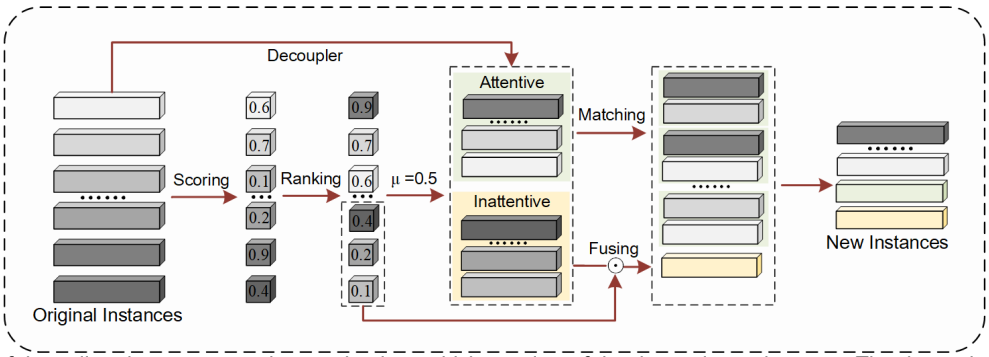


Fig. 3. Illustration of the online data augmentation mechanism, which consists of the decoupler and merger. The decoupler divides the original instances into attentive and inattentive groups based on the attention scores of MMC. The merger then fuses the inattentive instances and matches similar attentive ones to maximize the data diversity.

Finally, the outputs of the multi-head structure are concatenated together and subsequently feed into linear projections to obtain the complete output as follows:

$$\begin{aligned} \boldsymbol{H}'_{bag} &= [Concat(\boldsymbol{V}'_{2i}, \dots \boldsymbol{V}'_{2h} + \boldsymbol{V}_1)]W_{o1} \\ \boldsymbol{G}'_{bag} &= [Concat(\boldsymbol{V}'_{1i}, \dots \boldsymbol{V}'_{1h}) + \boldsymbol{V}_2]W_{o2}. \end{aligned} \quad (4)$$

where $\boldsymbol{W}_{o1}$ *and* $\boldsymbol{W}_{o2} \in \mathbb{R}^{d_m \times d_k}$ are the weight matrices of linear projections.

The multimodal bags $\boldsymbol{H}_{bag}$ and $\boldsymbol{G}_{bag}$ are used in the MMC layer to model dense pair-wise interactions between histology and genomic features. Then, the outputs $\boldsymbol{H}'_{bag}$ and $\boldsymbol{G}'_{bag}$ are respectively inputted to the following branch for unimodal learning.

2) Online Data Augmentation

To enhance the data diversity during model training, we propose an ODA module, which increases the diversity of bags by changing the distribution of instances. This strategy is inspired by Mixup [32], [33], a data augmentation technique that combines the feature vectors of two different groups in some proportion to create a new training sample. The ODA module allows the MIL to be exposed to a more diverse range of bags during training, so as to improve the model generalization. Since the unimportant instances may degrade the performance of ODA module, we do not use all instances for data augmentation, but divide them into two groups based on the attention scores. As shown in Fig. 3, the ODA block comprises two components, namely the instances decoupler and merger. The decoupler divides the patch embeddings into attentive and inattentive sections based on the attention score matrix $\boldsymbol{S}_i$ of MMC. Then the merger fuses inattentive instances and matches similar attentive ones to maximize the data diversity.

In order to preserve the most important instances, we first score and rank the instances. As mentioned above, the MMC layer scores pairwise similarity between genomic embeddings $\boldsymbol{g}_k \in \boldsymbol{G}_{bag}$ and patch embeddings $\boldsymbol{h}_m \in \boldsymbol{H}_{bag}$, which can be written as a row vector $[\boldsymbol{a}_{n1,\dots,}\boldsymbol{a}_{nm}] \in \boldsymbol{S}_i$, thus, the matrix $\boldsymbol{S}_i$ reflects the scores of the image patches to gene expression. Therefore, the attention scores can be used to determine the importance of each patch embedding. As shown in Fig. 3, we the split original patch embeddings into two groups according to the attention score. We first calculate the average of the attention scores on all genomic embeddings to obtain the genomic-guided attention vector $\boldsymbol{a} \in \mathbb{R}^{1\times M}$, then the average attention scores $\bar{\boldsymbol{a}}$ of all heads is computed by:

$$\bar{\boldsymbol{a}} = \sum_{h=1}^{H} \boldsymbol{a}_h / H \quad (5)$$

For M patch embeddings in total, we preserve the top-K instances as attentive ones according to $\bar{\boldsymbol{a}}$ and the remained $M-K$ instances are identified as inattentive ones. The keep rate is defined as $\mu = K/M$.

Although inattentive instances contain less information, they may still contribute to the classification results. Instead of discarding them directly, we use the attention scores in $\bar{\boldsymbol{a}}$ to weight these instances to generate a new instance, which can be written as:

$$\boldsymbol{h}_{inatt} = \sum_{i=1}^{i=M-K} \boldsymbol{h}_i \, \bar{\boldsymbol{a}}_i \quad (6)$$

where $\boldsymbol{h}_i$ represents the i-th instance in $\boldsymbol{H}'_{bag}$, and $\bar{\boldsymbol{a}}_i$ represents the i-th attention score in $\bar{\boldsymbol{a}}$.

For attentive instances, we consider instance diversity while maintaining their importance. Specifically, we adopt the cosine similarity metric to compute the similarity of different patch embeddings and obtain their cosine similarity scores $\boldsymbol{R}_{i,j} \in [-1,1]$, which represents the relation value between i-th and j-th instances:

$$\boldsymbol{R}_{i,j} = \frac{\sum_{k=1}^{d_k} \boldsymbol{h}_{i,k} \times \boldsymbol{h}_{j,k}}{\sqrt{\sum_{k=1}^{d_k} (\boldsymbol{h}_{i,k})^2} \times \sqrt{\sum_{k=1}^{d_k} (\boldsymbol{h}_{j,k})^2}} \quad (7)$$

where d_k represents the dimension of patch embeddings.

To avoid noisy edges and over-smoothing problems, we use a binary strategy to filter and re-weight values:

$$\boldsymbol{R}'_{i,j} = \begin{cases} 0, & if\ \boldsymbol{R}_{i,j} < \tau \\ 1, & if\ \boldsymbol{R}_{i,j} \geq \tau \end{cases} \quad (8)$$

where τ is a threshold.

Then, according to the values of $\boldsymbol{R}'_{i,j}$, we combine each pair of similar instances into a new one:

$$\boldsymbol{h}^i_{att} = p * \boldsymbol{h}^i_{att} + (1-p) * \boldsymbol{h}^j_{att},\ if\ \boldsymbol{R}'_{i,j} = 1 \quad (9)$$

where p is a strength hyper-parameter for data augmentation. We don't discard these attentive instances after generating new ones because the attentive instances contain the most discriminative information for the final prediction.

By fusing inattentive instances and matching attentive ones, we can increase the diversity of bags and still maintain importance instances. Therefore, the combination of bag

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

$\boldsymbol{H}'_{bag} = [\boldsymbol{h}_{inatt}, \boldsymbol{h}^i_{att}]$ is no longer static and fixed, but rather diverse and dynamic in each epoch. Subsequently, the output of ODA is fed into the global attention pooling [6], which can adaptively aggregate all instances within $\boldsymbol{H}'_{bag} \in \mathbb{R}^{K\times d_k}$ to obtain the bag-level representations $\boldsymbol{z}_h \in \mathbb{R}^{1\times d_k}$ by:

$$\boldsymbol{z} = \sum_{i=1}^{N} \boldsymbol{a}_i \boldsymbol{h}_i \tag{10}$$

with

$$\boldsymbol{a}_i = \frac{exp\{\boldsymbol{W}_a^T(\tanh(\boldsymbol{W}_v h_i) \odot sigm(\boldsymbol{W}_u h_i))\}}{\sum_{j=1}^{N} exp\{\boldsymbol{W}_a^T(\tanh(\boldsymbol{W}_v h_j) \odot sigm(\boldsymbol{W}_u h_j))\}} \tag{11}$$

where $\boldsymbol{h}_i$ represents the i-th instance in $\boldsymbol{H}'_{bag}$, $\boldsymbol{a}_i$ is the attention score of i-th instance, $\boldsymbol{W}_a$, $\boldsymbol{W}_v$ and $\boldsymbol{W}_u \in \mathbb{R}^{d_k\times d_k}$ are the weight matrices of FC layers, and $\odot$ is an element-wise multiplication.

3) SNN-Mixer

After using histology features to calibrate different omics features in the MMC module, two main challenges still remain: the HDLLS problem in multi-omics data and the correlation learning among different omics features. To address these issues, we develop an SNN-Mixer based on MLP-Mixer [39] to integrate different omics features.

MLP-Mixer is a recently proposed simple architecture that relies solely on MLPs. It introduces two types of layers to allow information communication across different dimensions of data, thereby enhancing the interaction of features. However, the multi-omics data generally have hundreds to thousands of features with relatively few training samples, and thus the traditional MLP is prone to overfitting, as well as training instabilities from current deep learning regularization technique, s such as activation function and Dropout. Therefore, we replace the GELU activation and Dropout in MLP-Mixer with the ELU activation and Alpha Dropout from the SNN [3]. The ELU activation has non-zero gradients when the input is negative, which can help alleviate the problem of vanishing gradients. Alpha Dropout is a variant of the Dropout regularization technique, where instead of randomly dropping out neurons during training, a random value is drawn from an alpha distribution and multiplied with each neuron's output to enhance the stabilities of the model during training.

As shown in Fig. 1, SNN-Mixer contains one token-mixing SNN and one channel-mixing SNN, each consisting of two fully-connected layers, two alpha Dropout layers, and an ELU activation. The token-mixing SNN is a cross-location operation that acts on columns of the input to mix all omics, while the channel-mixing SNN is a pre-location operation that acts on rows of the input to mix features of each omics. For the input multi-omics feature $\boldsymbol{X} = \boldsymbol{G}'_{bag} \in \mathbb{R}^{N\times d_k}$, SNN-Mixer obtains the corresponding output representation $\boldsymbol{Y} \in \mathbb{R}^{N\times d_k}$ as follows:

$$\begin{aligned} \boldsymbol{U} &= \boldsymbol{X}^T + \boldsymbol{W_2}\sigma(\boldsymbol{W_1}\mathrm{LN}(\boldsymbol{X}^T)) \\ \boldsymbol{Y} &= \boldsymbol{U}^T + \boldsymbol{W_4}\sigma(\boldsymbol{W_3}\mathrm{LN}(\boldsymbol{U}^T)) \end{aligned} \tag{12}$$

where LN denotes the layer normalization, σ denotes the activation function implemented by ELU, $\boldsymbol{W}_1 \in \mathbb{R}^{K\times C}$, $\boldsymbol{W}_2 \in \mathbb{R}^{C\times K}$, $\boldsymbol{W}_3 \in \mathbb{R}^{d_k\times d_s}$ *and* $\boldsymbol{W}_4 \in \mathbb{R}^{d_s\times d_k}$ are the weight matrices of fully-connected layers. C and d_k are tunable hidden widths in the token-mixing and channel-mixing SNNs, respectively. Note that the $K = 3$ and $d_k = 256$ in this paper. To further alleviate the HDLSS problem in G_{bag}, We increase $K \to \mathrm{C}: 3 \to 16$, and decrease $d_k \to d_s: 256 \to 64$ to reduce the quantitative gap in the two dimensions. Subsequently, the outputs of SNN-Mixer are also fed into the GAP to obtain the bag-level representations $\boldsymbol{z}_g \in \mathbb{R}^{1\times d_k}$.

C. Subtype Prediction

To reduce the influence of noisy unimodal features and increase the expressiveness of important modality, we use a gated bimodal unit [40] to weight the multimodal bag-level features to generate final representation $\boldsymbol{h}_{final}$ for cancer subtyping, which can be defined by:

$$\begin{aligned} \boldsymbol{h}_{\mathrm{a}} &= tanh(\mathbf{W}_{\mathrm{a}}\boldsymbol{z}_h) \\ \boldsymbol{h}_{\mathrm{b}} &= tanh\big(\boldsymbol{W}_{\mathrm{b}}\boldsymbol{z}_g\big) \\ \boldsymbol{z} &= \mathrm{Sigmoid}(\boldsymbol{W}_z[\boldsymbol{h}_{\mathrm{a}}, \boldsymbol{h}_{\mathrm{b}}]) \\ \boldsymbol{h}_{final} &= z\boldsymbol{h}_{\mathrm{a}} + (1-\boldsymbol{z})\boldsymbol{h}_{\mathrm{b}} \end{aligned} \tag{13}$$

where $\boldsymbol{W}_a, \boldsymbol{W}_b,$ and $\boldsymbol{W}_z$ are the weight matrices of FC layers, and $[\cdot,\cdot]$ represents the concatenation operation. The weights of the two modalities are $\boldsymbol{z}$ and $(1-\boldsymbol{z})$ based on the output of the Sigmoid activation function.

Finally, the $\boldsymbol{h}_{final}$ is fed into the classifier and softmax function to get the bag class predictions $\widehat{\boldsymbol{Y}}$. The class loss function $\mathcal{L}_{bag}$ is defined by the cross entropy between the bag class label $\boldsymbol{Y}$ and bag class prediction $\widehat{\boldsymbol{Y}}$, which can be expressed as:

$$\mathcal{L}_{bag} = \sum_{p=1}^{P} \mathcal{L}_{bag}\big(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}\big) \tag{14}$$

where P is the number of patients.

IV. EXPERIMENTS AND RESULTS

A. Datasets

To validate the effectiveness of our proposed method, we used three TCGA datasets in The Cancer Genome Atlas (https://tcga-data.nci.nih.gov/tcga/): 1) Invasive Ductal Carcinoma (IDC) versus Invasive Lobular Carcinoma (ILC) for invasive breast carcinoma (BRCA) subtyping, 2) Lung Adenocarcinoma (LUAD) versus Lung Squamous Cell Carcinoma (LUSC) for non-small cell lung carcinoma (NSCLC) subtyping, and 3) Kidney Chromophobe (KICH), Kidney Renal Clear (KIRC), and Kidney Renal Papillary (KIRP) for Renal Cell Carcinoma (RCC) subtyping.

For the BRCA, a total of 955 WSI slides were collected from 895 patients, including 787 IDC slides from 737 patients, and 168 ILC slides from 158 patients. For the NSCLC, a total of 999 slides were collected from 894 cases, including 515 LUAD slides from 453 patients, and 484 LUSC slides from 450 patients. For the RCC, a total of 936 slides were collected from 903 cases, including 518 KIRC slides from 512 patients, 297

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

KIRP slides from 273 patients, and 121 KICH slides from 109 patients. After WSI pre-processing, the total number of patches extracted at 10×magnification on the BRCA, NSCLC, and RCC datasets were 1.54 million, 3.31 million and 2.07 million, respectively. To further perform multimodal learning, we also collected the multi-omics data containing mutation status, copy number variation, and RNA-Seq abundance from the cBioPortal (https://www.cbioportal.org/). The data statistics of those three datasets are concluded in Table I.

B. Experiment Setup and Evaluation Metrics

We adopted the Slideflow [41], a deep-learning library for digital pathology, to preprocess WSIs. We first applied the Otsu threshold algorithm to filter out the background, and then each WSI was cropped into 299×299 non-overlapping patches. After patching, a pre-trained TransPath [37] model was used to extract a feature vector with a dimensional of 768 from each patch. We then applied a 5 repeated 5-fold cross-validation to the proposed method and evaluated the model performance in accuracy and area under the curve (AUC). Note that during the 5-fold cross-validation, 25% of the training dataset was also randomly selected as the validation sets for choosing the checkpoints. For each dataset, the 5-fold evaluation procedure was run 5 times.

C. Implementation Details

For training, we utilized the Adam optimizer to update our models with a weight decay of 1e-5. The size of mini-batch was set as 1 (bag). The models were trained for 100 epochs with a cross-entropy loss function, and they would early stop if the loss did not decrease in 20 epochs. The learning rate was tuned from {5e-4, 1e-4, 5e-5}, and the dropout ratio of linear layers was tuned from {0, 0.1, 0.2, 0.3, 0.4}. All models were implemented by Python 3.8 with PyTorch toolkit 1.13 on a platform with an NVIDIA GeForce RTX 3090 GPU.

D. Comparison with State-of-the-art Methods

We compared the proposed MCFN with several SOTA algorithms. For a fair comparison, the same 5-fold cross-validation splits were used for evaluating all methods. All reference approaches were conducted using the original code implementation and the parameters were consistent with the original experimental settings.

1) **MLP**: It is a common feedforward network for processing genomic features, which includes two fully-connected layers and a ReLU nonlinearity.
2) **SNN** [38]: It is a current SOTA unimodal baseline on genomic data, where ELU nonlinearity and Alpha Dropout replace ReLU nonlinearity and Dropout in MLP.
3) **Deep Sets** [9]: It is one of the first neural network architectures for MIL that proposes a sum pooling to aggregate instance-level features.
4) **ABMIL** [6]: This MIL model replaces the sum pooling in Deep Sets with global attention pooling, which can use attention scores to weigh the instance-level features.
5) **TransMIL** [8]: It is a SOTA MIL model for WSI classification that approximates self-attention with Nyström method in Transformer.

TABLE I
DATA STATISTICS OF BRCA, NSCLC AND RCC DATASETS

Label Description	Dataset		
	BRCA	NSCLC	RCC
Total Patients	IDC: 737 ILC: 158	LUAD: 453 LUSC: 450	KIRC: 512 KIRP: 273 KICH: 109
Total Slides	IDC: 787 ILC: 168	LUAD: 515 LUSC: 484	KIRC: 518 KIRP: 297 KICH: 121
Total patches	1.54 million	3.31 million	2.07 million
Average Patches	1622	3601	2208
Total Genes	3637	2936	2755
Mutation Status	746	864	262
Copy Number Variation	1333	517	937
RNA-Seq Abundance	1558	1555	1556

6) **GSCNN** [16]: It is a common late multimodal fusion algorithm, which uses concatenation operation to combine histology and genomic features.
7) **Bilinear Pooling (BP)** [14]**:** It is a common late multimodal fusion algorithm, which uses bilinear pooling operation to combine histology and genomic features.
8) **MCAT** [19]: It is a common early multimodal fusion algorithm using both WSI and genomic data, which uses the co-attention mechanism to capture the relationships between the histology and genomic features.
9) **PORPOISE** [21]: It is a SOTA late multimodal fusion algorithm using both WSI and multi-omics data. The Kronecker Product operation is employed to combine histology and genomic features.
10) **HGCN** [36]: It is a SOTA early multimodal algorithm that utilizes the multimodal information of patients to realize intra- and inter-modal interaction between multimodal graphs.

TABLE II shows the classification results of different algorithms on the TCGA-BRCA, TCGA-NSCLC, and TCGA-RCC datasets. It can be observed that the proposed MCFN outperforms all the compared algorithms with statistical significance on all datasets. Specifically, on the TCGA-BRCA dataset, MCFN achieves the best mean accuracy of 92.82%, and AUC of 97.33%. Compared to other algorithms, it improves at least 1.19%, and 1.06% on the corresponding indices, respectively. MCFN also outperforms all the compared algorithms with the best accuracy of 94.53%, and AUC of 99.03% on the TCGA-NSCLC dataset, improving by 1.12%, and 1.09% on corresponding indices, respectively. Similarly, on the TCGA-RCC dataset, the MCFN also achieves the best mean accuracy of 93.89% and AUC of 94.73%. Compared to other algorithms, it improves at least 1.31%, and 1.24% on the corresponding indices, respectively. We attribute the high performance of MCFN to (1) an effective utilization of both modalities, (2) an efficient online data augmentation, and (3) a meaningful multi-omics data learning scheme.

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

TABLE II
CLASSIFICATION RESULTS OF DIFFERENT ALGORITHMS ON TCGA-BRCA, TCGA-NSCLC AND TCGA-RCC DATASETS (UNIT: %). THE SUBSCRIPTS ARE THE CORRESPONDING 95% CONFIDENCE INTERVALS. THE BEST ONES ARE IN BOLD.

Algorithm	Study	TCGA-BRCA		TCGA-NSCLC		TCGA-RCC		Complexity	
		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	FLOPs	Params
MLP	Omics	87.92(87.41, 88.43)	89.85(89.19, 90.51)	84.77(89.26, 85.28)	91.96(91.38, 92.54)	82.38(81.76, 83.00)	84.67(84.18, 85.16)	10.1M	878K
SNN [38]	Omics	88.75(88.07, 89.43)	90.01(89.22, 90.80)	85.00(84.13, 85.87)	90.92(90.07, 91.77)	83.94(83.17, 84.71)	85.89(85.27, 86.51)	10.1M	878K
Deep Sets [9]	WSI	87.62(87.07, 88.17)	89.70(89.11, 90.29)	87.79(87.27, 88.31)	93.86(93.58, 94.14)	89.55(88.70, 90.40)	90.94(90.35, 91.53)	260M	346K
ABMIL [6]	WSI	88.57(87.78, 89.36)	89.73(88.87, 90.59)	90.23(89.42, 91.04)	94.90(94.72, 95.08)	90.74(90.30, 91.18)	91.84(91.40, 92.28)	263M	329K
TransMIL [8]	WSI	89.07(88.69, 89.45)	90.48(89.67, 91.29)	90.35(89.35, 91.35)	94.74(94.54, 94.94)	89.77(89.21, 90.33)	91.33(91.01, 91.65)	844M	746K
GSCNN [16]	Omics & WSI	90.43(89.62, 91.24)	92.79(92.21, 93.37)	91.95(91.26, 92.64)	95.78(95.42, 96.14)	91.70(91.23, 92.17)	92.72(92.19, 93.25)	264M	1.41M
BP [14]	Omics & WSI	89.99(89.51, 90.47)	92.30(91.72, 92.88)	91.91(91.36, 92.46)	95.63(95.54, 95.72)	91.89(91.41, 92.37)	92.63(92.10, 93.16)	284M	5.65M
PORPOISE [21]	Omics & WSI	90.47(89.77, 91.17)	93.08(92.50, 93.66)	92.12(91.79, 92.45)	95.91(95.53, 96.29)	92.01(91.63, 92.39)	93.17(92.48, 93.86)	309M	2.58M
MCAT [19]	Omics & WSI	90.78(90.13, 91.43)	93.17(92.73, 93.61)	92.65(92.12, 93.18)	96.05(95.59, 96.51)	92.17(91.69, 92.65)	92.77(92.17, 93.37)	336M	3.78M
HGCN [36]	Omics & WSI	91.63(91.12, 92.14)	93.66(93.08, 94.24)	93.41(92.93, 93.89)	96.24(95.98, 96.50)	92.58(91.96, 93.20)	93.49(92.76, 94.22)	446M	3.27M
MCFN (Ours)	Omics & WSI	**92.82(92.34, 93.30)***	**94.72(94.39, 95.05)***	**94.53(94.16, 94.90)***	**97.33(96.96, 97.70)***	**93.89(93.43, 94.35)***	**94.73(94.51, 94.95)***	393M	1.94M

* DENOTES MCFN GETS STATISTICALLY SIGNIFICANT IMPROVEMENT ON THIS RESULT ($P < 0.05$, TWO-SAMPLE T-TEST) COMPARED TO OTHER COMPARED ALGORITHMS.

Unimodal vs. Multimodal Methods: The multimodal algorithms achieve better results than the unimodal ones, indicating the effectiveness of integrating multimodal information in cancer diagnosis. Compared to the unimodal methods, the multimodal methods require more model parameters, as they need to simultaneously process data from two modalities. However, their FLOPs do not increase significantly, because genomic data is very small and hardly incurs computational costs. As a SOTA multimodal algorithm, the HGCN achieves the second-best results due to the multimodal graphs in it, which can effectively perform intra- and inter-modal interaction. Nevertheless, our MCFN still outperforms HGCN with fewer model parameters and FLOPs.

Early vs. Late Fusion: In our experiments, the early fusion algorithms (MCAT, HGCN, and MCFN) mostly outperform the late fusion multimodal ones (GSCNN, Bilinear Pooling, and PORPOISE) on all datasets. We attribute this observation to the utilization of the multimodal interactions, which effectively capture the relationships between the histology and genomic features. Although these multimodal interactions will increase FLOPs, they can significantly improve the model's performance. This observation strongly supports our design choice of incorporating the co-attention mechanism for joint learning in the multimodal feature space.

E. Ablation Study

We conducted an ablation study to describe the contributions of three major components in the proposed MCFN: MMC, ODA, and SNN-Mixer modules. The proposed MCFN was compared with four variants:

1) **MCFN-m**: This variant replaced the MMC, ODA, and SNN-Mixer modules in MCFN with linear layers.
2) **MCFN-MMC**: This variant only maintained the MMC for multimodal interactions, while the ODA and SNN-Mixer modules were replaced by the linear layers.
3) **MCFN-ODA**: This variant maintained the MMC and ODA for data augmentation, as ODA relies on the attention scores of MMC to perform token decoupling, while the SNN-Mixer was replaced by the linear layer.
4) **MCFN-SNN**: This variant maintained the MMC and SNN-Mixer for multi-omics features learning, as the SNN-Mixer relies on the MMC to calibrate different omics features, while the ODA was replaced by the linear layer.

TABLE III
CLASSIFICATION RESULTS OF ABLATION EXPERIMENT ON TCGA-BRCA, TCGA-NSCLC AND TCGA-RCC DATASETS (UNIT: %). THE RESULTS ARE PRESENTED IN THE FORMAT OF MEAN ± SD (STANDARD DEVIATION). THE BEST ONES ARE IN BOLD.

	Study/Strategy	TCGA-BRCA		TCGA-NSCLC		TCGA-RCC		Complexity	
		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	FLOPs	Param.
Ablation Study	MCFN-m	90.39±1.22	92.27±1.51	91.94±0.87	95.19±0.87	91.60±1.34	92.52±1.59	193M	1.38 M
Ablation Study	MCFN-MMC	91.56±1.33	92.76±1.80	92.96±0.74	96.16±0.37	92.72±1.09	93.59±1.34	378M	1.76 M
Ablation Study	MCFN-ODA	92.32±0.89	93.87±1.58	93.63±1.15	96.34±0.73	93.19±1.34	93.92±1.46	391M	1.86 M
Ablation Study	MCFN-SNN	92.17±1.14	93.52±1.12	93.42±1.04	96.47±0.70	93.24±1.18	93.87±1.29	380M	1.84 M
Ablation Study	MCFN	**92.82±1.23**	**94.72±0.85**	**94.53±0.94**	**97.33±0.95**	**93.89±1.17**	**94.73±0.55**	393M	1.94M
Interaction Strategy	$\mathbf{A}_{h\to g}$	90.54±1.34	91.87±1.27	91.67±1.27	95.29±1.19	92.34±1.38	93.09±1.37	315M	1.63M
Interaction Strategy	$\mathbf{A}_{g\to h}$	90.98±1.24	92.17±1.09	92.54±1.21	95.71±1.16	92.07±1.23	92.99±1.29	316M	1.63M
Interaction Strategy	$\mathbf{A}_{h\to g} + \mathbf{A}_{g\to h}$	91.31±1.57	92.28±1.70	92.54±1.21	95.71±1.16	92.51±1.18	93.25±1.13	507M	1.90M
Interaction Strategy	**MMC**	**91.56±1.33**	**92.76±1.80**	**92.96±0.74**	**96.16±0.37**	**92.72±1.09**	**93.59±1.34**	378M	1.76 M

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

Table III shows the classification results of the four variants on TCGA-BRCA, TCGA-NSCLC, and TCGA-RCC datasets, respectively. It can be found that MCFN-MMC, MCFN-ODA, and MCFN-SNN all achieve better results than MCFN-m, demonstrating the effectiveness of the three components in the proposed MCFN. Moreover, the MCFN-ODA achieves superior performance compared to MCFN-MMC, indicating that ODA can utilize complementary information from genomic information to guide the data augmentation, and thus its performance improves. The MCFN outperforms all variants on all datasets, indicating that the combination of three components can capture correlations between different modalities as well as enhance representation learning for each modality, thus improving the overall performance of model.

Computational Complexity: Table III presents the computational complexity of the three components in MCFN. It can be observed that the ODA and SNN-Mixer modules significantly enhance the model's performance without substantially increasing parameters and Flops, demonstrating the efficiency and effectiveness of both modules. As a precursor module for the ODA and SNN-Mixer modules, although there is an increase in parameters and Flops of MMC, it improves the overall performance of the model. In summary, our MCFN achieves a good balance between computational complexity and model performance.

We further conducted an ablation study to investigate how the model performed under different interaction strategies: $\mathbf{A}_{h\to g}$(it models only histology-to-omics interaction, this design resembles MCAT [19]), $\mathbf{A}_{g\to h}$ (it models only omics-to-histology interaction), $\mathbf{A}_{h\to g}+\mathbf{A}_{g\to h}$ (it models both $\mathbf{A}_{h\to g}$ and $\mathbf{A}_{g\to h}$ interactions, this design resembles MGCT [20]), and our MMC module. As shown in Table III, the $\mathbf{A}_{g\to h}$ performs better than $\mathbf{A}_{h\to g}$ on the TCGA-BRCA and NSCLC datasets, but the $\mathbf{A}_{h\to g}$ demonstrates a stronger capability in TCGA-RCC than the $\mathbf{A}_{g\to h}$. That is, both $\mathbf{A}_{h\to g}$ and $\mathbf{A}_{g\to h}$ show their advantages on the datasets of different cancers. It can further be found that both $\mathbf{A}_{h\to g}+\mathbf{A}_{g\to h}$ and MMC strategies consistently outperform either $\mathbf{A}_{h\to g}$ or $\mathbf{A}_{g\to h}$ on all three datasets. Moreover, the MMC also outperforms $\mathbf{A}_{h\to g}+\mathbf{A}_{g\to h}$ with fewer model parameters and FLOPs. It indicates that the proposed MMC can efficiently exploit the complementary information in the histology and genomics to provide more accurate predictions.

1) Study on Different Instance Merger Strategy

We also studied the performance of MCFN-SNN with different instance merging strategies. For the inattentive instances, we compared two strategies to reduce the influence of noisy or irrelevant instances on GAP: discarding them or fusing them into one new instance. All attentive instances were preserved without any processing to prevent interference. The result in Table IV shows the fusing strategy achieves superior performance over the discarding strategy. It indicates that the fusing strategy can retain potentially valuable information, thereby facilitating the model's generalization. Moreover, for the attentive instances, we compared two strategies to increase the diversity of bags: random matching and cosine similarity-based similarity matching. As shown in Table IV, the similarity matching far outperforms the random matching on all three datasets, mainly because the random matching maximizes diversity, but may introduce noise by combining unrelated instances. In contrast, the cosine similarity-based matching ensures that instances with similar features are combined, striking a balance between diversity and relevance. Overall, these results prove the effectiveness of our ODA module in enriching the diversity of bags.

2) Study on Data Augmentation in MIL

To verify the effectiveness of the ODA module, we further compared the MCFN-ODA with several other available data augmentation techniques, including:

(1) **ReMix** [27]: It is a prototype-based method that mixes the prototypes of two bags to generate a new bag.

(2) **RankMix** [33]: It is an improved interpolation-based Mixup method, in which the instances in each bag are ranked based on their attention scores, and then two bags are aligned by dropping instances with lower scores from the bag with larger instances.

(3) **PseMix** [32]: It is a pseudo-bag method that divides each bag into n pseudo-bags based on prototypes for size alignment, and then mixes the pseudo-bag for semantic alignment.

Fig. 4 shows the results of different data augmentation algorithms. It can be found that our MCFN-ODA consistently outperforms other methods on all three datasets, which ensures this method has clinical value. First, both ReMix and PseMix, need to use the K-Means clustering to generate the prototypes. Since different initial values may lead to different cluster results, the final results may not be the global optimal solution, thus affecting the performance of MIL. Second, RankMix needs to use the attention scores to drop instances from the bag with more instances, and then align two irregular bags for interpolation. However, the attention scores may not always accurately reflect instance importance due to potential biases in

TABLE IV
CLASSIFICATION RESULTS OF DIFFERENT INSTANCE MERGER STRATEGY ON TCGA-BRCA, TCGA-NSCLC AND TCGA-RCC DATASETS (UNIT: %). THE RESULTS ARE PRESENTED IN THE FORMAT OF MEAN ± SD (STANDARD DEVIATION). THE BEST ONES ARE IN BOLD.

Instance	Strategy	TCGA-BRCA		TCGA-NSCLC		TCGA-RCC	
		ACC	AUC	ACC	AUC	ACC	AUC
Inattentive	Discarding	89.77±1.57	91.04±1.67	90.95±1.38	93.21±1.91	91.49±1.76	92.31±1.56
	Fusing	92.07±1.34	93.52±1.12	93.51±1.46	96.60±1.11	93.17±1.59	93.94±1.49
Attentive	Random matching	91.28±1.64	92.67±1.57	92.38±1.52	95.89±1.43	92.76±1.67	93.69±1.34
	Similarity matching	**92.82±1.23**	**94.72±0.85**	**94.53±0.94**	**97.33±0.95**	**93.89±1.17**	**94.73±0.55**

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

WSI data, and then ignore the most important instances that contain critical diagnostic information. In contrast, the proposed ODA utilizes the useful complementary knowledge from the genomics modality to guide the instance-level augmentation of WSI data in an end-to-end manner, which can improve data diversity for model training.

3) Study on Multi-omics Feature Learning

We also investigate the effect of different multi-omics feature learning schemes on the performance of the proposed MCFN-SNN. The proposed MCFN-SNN was compared with the following three variants:

(1) **MFN (Multimodal Fusion Network)-SNN:** This variant only maintained the SNN-Mixer for multi-omics interactions, while the MMC module was replaced by the linear layer.

(2) **MCFN-MLP:** This variant had the same network structure as MCFN-SNN, while the SNN-Mixer module was replaced by the MLP-Mixer.

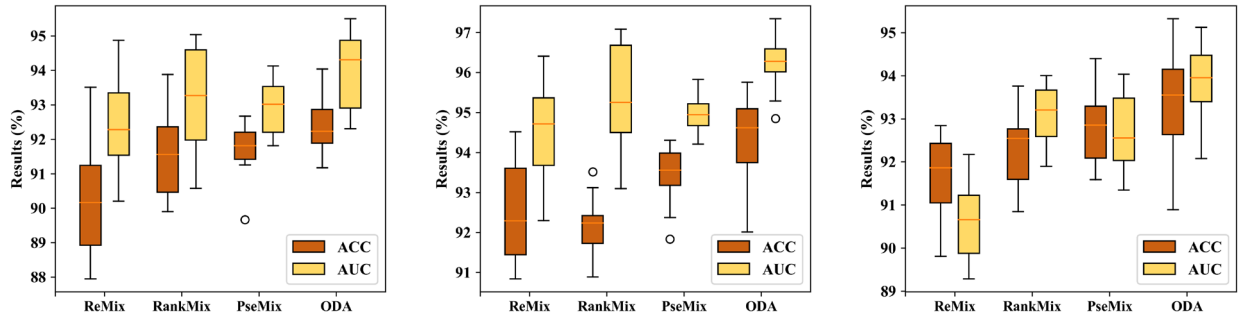
(3) **MCFN-Trans:** This variant had the same network structure as MCFN-SNN, while the SNN-Mixer module was replaced by the Transformer block.

Fig. 5 shows the classification results on the three datasets. It can be found that MCFN-SNN achieves better results than MFN-SNN, MCFN-MLP, and MCFN-Trans, indicating that the MCFN-SNN can effectively alleviate the HDLSS problem in multi-omics features and fully learn the correlation between different omics. The MCFN-MLP and MCFN-Trans also outperform MFN-SNN on all the three datasets. It indicates that the MMC module can effectively calibrate different omics features by using histology features, thereby reducing the impact of noise on the interaction of multi-omics features. Moreover, MCFN-MLP obtains superior performance to MCFN-Trans, this is because Transformer usually requires abundant training samples to get a robust model, while multi-omics data belongs to small samples, thus leading to performance degradation.

F. Interpretation of Results

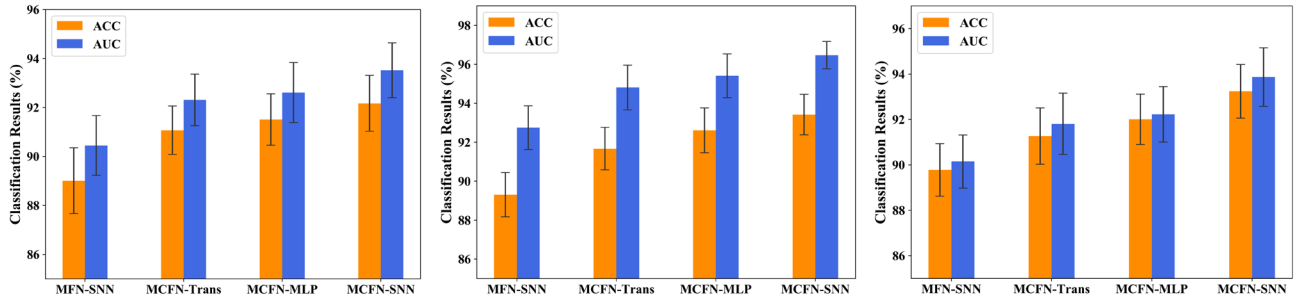
To further validate the interpretability of our model, we applied the attention scores of MMC to visualize the resulting attention map, and Integrated Gradient (IG) analysis [42] on multi-omics data. We first normalized the co-attention scores to a range of 0 to 1 (from blue to red), then generated the attention maps by overlaying computed co-attention scores for each histology patch onto WSIs. Fig. 6(a) shows several subtype cases in the TCGA-BRCA datasets. We observe that these attention maps can localize the tumor regions. For example, the high-attention patches of the IDC cases are mainly focused on the high-grade tumor morphology such as dense tumor cellularity, while the high-attention patches of the ILC cases are mainly focused on the invasive and tumor-infiltrated stroma. A mosaic map is generated by overlaying remained patches in a grid-wise fashion. It proves that our framework can effectively preserve the most discriminative patches via the MMC, which is beneficial for the ODA module.

Fig. 6(b) shows the distribution of genomic features corresponding to the patients with the IDC (blue points) and ILC (red points) subtypes. The x-axis represents the IG attribution scores, and the y-axis indicates the selected genomic features. For each genomic feature of a patient, the attribution score is computed through IG analysis to quantify the feature's impact on the predictive outcomes [42]. Thus, for the cancer



(a) TCGA-BRCA (b) TCGA-NSCLC (c) TCGA-RCC

Fig. 4. Classification results of different data augmentation techniques on different datasets.

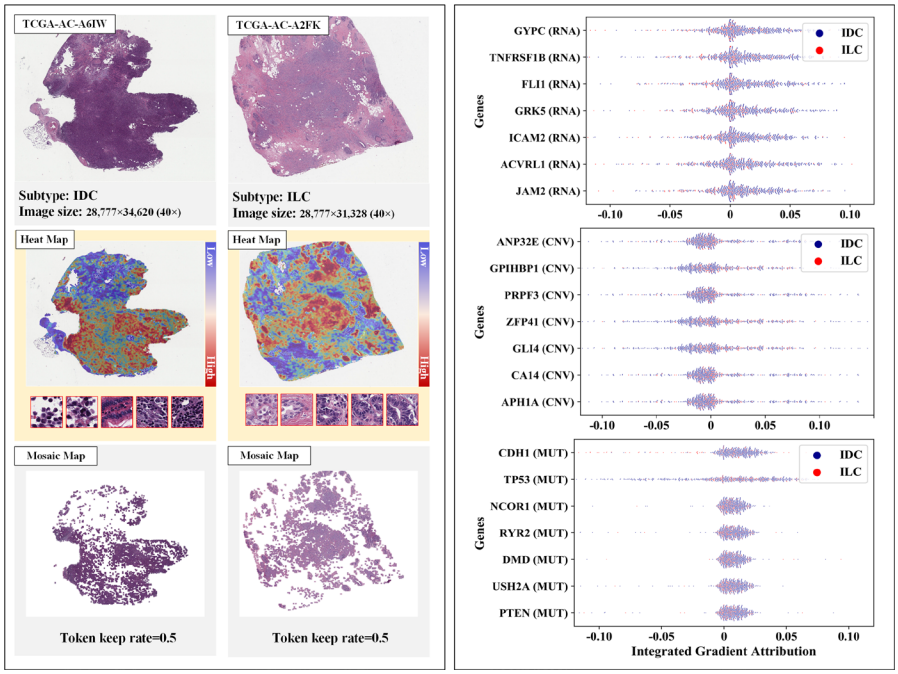


(a) TCGA-BRCA (b) TCGA-NSCLC (c) TCGA-RCC

Fig. 5. Classification results of different multi-omics features learning schemes on different datasets.

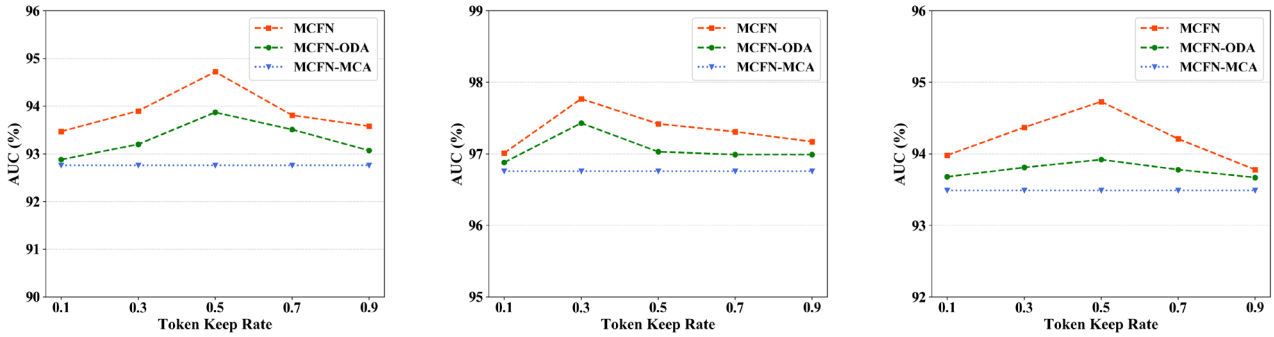
Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535



(a) (b)

Fig. 6. Interpretation of the proposed framework on BRCA dataset. (a) WSIs with the corresponding heat maps and mosaic maps. (b) Integrated gradient analysis of multi-omics data, where the x-axis represents the attribution score to indicate the contribution of a feature to the prediction results, and the y-axis represents the selected genomic features.



(a) TCGA-BRCA (b) TCGA-NSCLC (c) TCGA-RCC

Fig. 7. Classification results of ODA block with different keep-rate on different datasets.

subtyping task in this work, the features with positive attribution scores (above zero) tend to increase the probability of a sample being classified into a particular subtype, whereas those with negative scores (below zero) decrease this probability. As shown in Fig. 6(b), the top 7 most influential genomic features are selected from each omics to further investigate these subtypes based on the prediction results. It can be observed that different features contribute differently to the classification of IDC and ILC. For example, the CDH1 mutation is more related to IDC patients than ILC patients within the MUT features, as most IDC cases show positive attribution values for this mutation. It suggests that the presence of a CDH1 mutation is a significant indicator for predicting IDC. On the contrary, the TP53 mutation tends to support classification as ILC, since most IDC patients exhibit negative attribution values for this mutation. By analyzing these selected omics features, we can gain insights into the molecular characteristics that discriminates IDC from ILC subtypes. It helps to identify specific genetic variations that may play a crucial role in determining the subtypes of breast cancers. In summary, these visualization results provide valuable information for understanding the underlying mechanisms in different cancer subtypes, which has the potential to improve personalized treatment strategies for patients.

G. Hyperparameter Sensitivity Analysis

A hyperparameter sensitivity analysis was also conducted for the proposed MCFN. Two hyperparameters in MCFN will affect the classification performance, i.e., the patch keeping rate μ and the strength hyperparameter p for data augmentation.

The value of patch keeping rate μ is an important hyperparameter to affect the performance of ODA module. We investigated the performance of proposed MCFN and MCFN-w-ODA with different keep rate$\mu \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and the results are shown in Fig. 7. It is observed that for BRAC and RCC subtyping tasks, the best performance of MCFN is

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

TABLE V
THE AUC PERFORMANCE WITH DIFFERENT HYPERPARAMETERS ON TCGA-BRCA, TCGA-NSCLC AND TCGA-RCC DATASETS (UNIT: %).

Dataset	Augmentation Strength p					
	0	0.1	0.3	0.5	0.7	0.9
BRCA	92.17	92.71	92.77	92.82	92.70	92.77
NSCLC	93.42	94.45	94.50	94.53	94.45	94.44
RCC	93.24	93.73	93.82	93.89	93.95	93.81
Average	92.94	93.63	93.70	93.75	93.69	93.67

achieved by setting μ to 0.5, while for NSCLC subtyping task, the best performance of MCFN is achieved by setting μ to 0.3. Compared to the BRCA and RCC datasets (with a total of 1.54 and 2.02 million patches, respectively), the NSCLC dataset (with a total of 3.31 million patches) has a relatively larger proportion of tumor regions, leading to the positive bags containing a large portion of the positive patches. Therefore, a lower keep rate in the NSCLC dataset can effectively reduce redundant instances and prevent noise from adversely affecting the ODA module.

After determining the optimal token keep rate μ_i for each dataset, we further conducted experiments on all the datasets starting from $p = 0$, where the MCFN degenerated into the MCFN-SNN without any data augmentation. We then gradually increased $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. As shown in Table V, apart from $p = 0$, there is no significant difference in the performance of MCFN under different strength values of p on all datasets. The result indicates the robustness of our MCFN to the choice of augmentation strength.

H. Survival Prediction Results

We also performed the survival prediction task on the BRCA, LUSC, LUAD, and KIRC cancer types, which have relatively large sample sizes in this study. All datasets were evaluated using a cross-validated concordance index (c-Index). Table VI presents the c-Index of different algorithms on these four datasets. It can be observed that the MCFN achieves the best c-Index performance across all four cancer types, with significant improvements demonstrated in BRCA (0.672) and LUAD (0.657) compared to other methods. Interestingly, the multimodal algorithms are not always superior to the unimodal ones. For instance, the TransMIL outperforms most multimodal algorithms in LUSC. We attribute this observation to the fact that survival prediction is a more challenging regression task compared to the cancer subtype classification, and therefore it requires higher model robustness. Compared with previous multimodal methods, our MCFN mainly addresses the heterogeneities and insufficient among multimodal data, thus improving model robustness and achieving better performance.

TABLE VI
THE C-INDEX OF DIFFERENT ALGORITHMS ON BRCA, LUSC, LUAD AND KICH DATASETS. THE RESULTS ARE PRESENTED IN THE FORMAT OF MEAN ± SD (STANDARD DEVIATION). THE BEST ONES ARE IN BOLD.

Dataset	BRCA	LUSC	LUAD	KIRC
MLP	0.561±0.058	0.531±0.051	0.539±0.069	0.649±0.071
SNN [38]	0.586±0.072	0.522±0.022	0.554±0.074	0.633±0.063
Deep Sets [9]	0.521±0.022	0.527±0.057	0.496±0.008	0.555±0.051
ABMIL [6]	0.560±0.066	0.561±0.062	0.548±0.042	0.567±0.068
TransMIL [8]	0.530±0.057	0.584±0.110	0.557±0.071	0.589±0.067
GSCNN [16]	0.574±0.041	0.560±0.013	0.617±0.014	0.658±0.111
BP [14]	0.583±0.048	0.509±0.034	0.600±0.046	0.661±0.078
PORPOISE [21]	0.628±0.053	0.538±0.033	0.626±0.018	0.659±0.075
MCAT [19]	0.652±0.087	0.564±0.012	0.620±0.032	0.672±0.040
HGCN [36]	0.657±0.069	0.598±0.012	0.633±0.071	0.686±0.054
MCFN (Ours)	**0.672±0.042**	**0.611±0.080**	**0.657±0.048**	**0.693±0.039**

V. DISCUSSION

With the advancement of deep learning in computational pathology, there is an increasing utilization of multimodal medical data (such as WSIs and genomics) for achieving precise cancer diagnosis and personalized treatment. However, the data heterogeneity brings challenges for designing automated analysis methods. Previous studies have primarily focused on late fusion strategy in multimodal data learning. However, such approaches offer limited opportunities for effective multimodal interactions. Our MCFN can capture the local relevance between different modalities intuitively, making it convenient for subsequent processing and interpretable analysis.

Based on our proposed MMC module, we further present a novel ODA block for data augmentation in MIL. Specifically, this block utilizes the attention scores generated by MMC to separate histological instances into attention and inattention ones. By combining the inattentive instances with matching attentive ones, we effectively enhance the diversity of instances while preserving their importance. This augmentation technique is particularly significance for MIL as it has traditionally lacked data augmentation techniques. Moreover, the MMC and ODA modules can also be used in other approaches without significant modifications, which indicates the versatility of our module for practical applications.

This manuscript also investigated multi-omics data learning schemes. To handle the HDLSS problem in multi-omics data, we proposed an SNN-Mixer and integrated it into our MCFN model. Experimental results have demonstrated that our SNN-Mixer can effectively capture the correlations among different omics features and alleviate the overfitting problem caused by HDLSS. Our research provides new insights, particularly on learning HDLSS data.

Our method does have some limitations. Firstly, the proposed MCFN is only applicable to complete data in this study, meaning that all modalities of all samples in the dataset must be available. However, missing data in certain modalities is a common problem in real clinical scenarios. Future work would focus on addressing the missing data issue for multimodal learning. Secondly, in MMC, all modalities are considered equally important for multimodal interactions, and therefore, using unimportant modality to guide the learning of important modality may introduce noise. In the future, weighting mechanism will be investigated in multimodal interaction.

VI. CONCLUSION

In summary, we propose a novel MCFN for cancer subtype classification, leveraging WSI and multi-omics data. The

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2024.3405535

MCFN effectively learns multimodal information through the MMC module to improve performance of CAD. Additionally, we introduce a dedicated unimodal feature learning block that incorporates the ODA and SNN-Mixer layer to enhance representation learning for each modality. Experimental results on three public cancer datasets validate the effectiveness of the proposed MCFN, highlighting its potential for clinical applications in WSI-based CAD.

REFERENCES

[1] J. Stingl and C. Caldas, "Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis," *Nat Rev Cancer*, vol. 7, no. 10, pp. 791–799, Oct. 2007.

[2] K. Bera et al., "Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology," *Nat Rev Clin Oncol*, vol. 16, no. 11, pp. 703–715, Nov. 2019.

[3] K.-H. Yu et al., "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nat Commun*, vol. 7, no. 1, p. 12474, Aug. 2016.

[4] Z. Gao et al., "A Convolutional Neural Network and Graph Convolutional Network Based Framework for Classification of Breast Histopathological Images," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 7, pp. 3163–3173, Jul. 2022.

[5] G. Campanella et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nat Med*, vol. 25, no. 8, pp. 1301–1309, Aug. 2019.

[6] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," *arXiv:1802.04712 [cs, stat]*, Jun. 2018.

[7] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14313–14323.

[8] Z. Shao et al., "TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification," in *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 2136–2147.

[9] M. Zaheer et al., "Deep Sets," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[10] S. Ding et al., "Multi-Scale Efficient Graph-Transformer for Whole Slide Image Classification," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 12, pp. 5926–5936, Dec. 2023.

[11] T. Zhou et al., "M2Net: Multi-modal Multi-channel Network for Overall Survival Time Prediction of Brain Tumor Patients," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 221-231.

[12] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, Jul. 2019.

[13] Z. Li et al., "Survival Prediction via Hierarchical Multimodal Co-Attention Transformer: A Computational Histology-Radiology Solution," *IEEE Trans. Med. Imaging*, vol. 42, no. 9, pp. 2678–2689, Sep. 2023.

[14] R. J. Chen et al., "Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis," *IEEE Trans. Med. Imaging*, vol. 41, no. 4, pp. 757–770, Apr. 2022.

[15] J. Gamper and N. Rajpoot, "Multiple Instance Captioning: Learning Representations from Histopathology Textbooks and Articles," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16544–16554, Jun. 2021.

[16] P. Mobadersany et al., "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, no. 13, Mar. 2018.

[17] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear Attention Networks," in *Advances in neural information processing systems*, 2018, no. 31.

[18] J. H. Kim et al., "Hadamard Product for Low-rank Bilinear Pooling," *arXiv preprint arXiv*: 1610.04325, 2016.

[19] R. J. Chen et al., "Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3995–4005, Oct. 2021.

[20] M. Liu et al., "MGCT: Mutual-Guided Cross-Modality Transformer for Survival Outcome Prediction using Integrative Histopathology-Genomic Features," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1306-1312, 2023.

[21] R. J. Chen et al., "Pan-cancer integrative histology-genomic analysis via multimodal deep learning," *Cancer Cell*, vol. 40, no. 8, pp. 865-878.e6, Aug. 2022.

[22] L. Cantini et al., "MicroRNA–mRNA interactions underlying colorectal cancer molecular subtypes," *Nat Commun*, vol. 6, no. 1, p. 8878, 2015.

[23] J. Xu et al., "A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data," *BMC Bioinformatics*, vol. 20, no. 1, p. 527, Dec. 2019.

[24] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J Big Data*, vol. 6, no. 1, p. 60, 2019.

[25] D. Tellez et al., "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical Image Analysis*, vol. 58, p. 101544, 2019.

[26] I. Zaffar et al., "Embedding Space Augmentation for Weakly Supervised Learning in Whole-Slide Images." In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1-4, 2023.

[27] J. Yang et al., "ReMix: A General and Efficient Framework for Multiple Instance Learning based Whole Slide Image Classification." in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 34-45, 2022.

[28] M. Kang et al., "Benchmarking Self-Supervised Learning on Diverse Pathology Datasets." in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3344-3354, 2023.

[29] Z. Shao et al., "AugDiff: Diffusion based Feature Augmentation for Multiple Instance Learning in Whole Slide Image." *arXiv preprint arXiv*: 2303.06371, 2023.

[30] R. Yang, P. Liu, and L. Ji, "ProtoDiv: Prototype-guided Division of Consistent Pseudo-bags for Whole-slide Image Classification," *arXiv preprint arXiv*: 2304.06652, 2023.

[31] H. Zhang et al., "DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18802-18812, 2022.

[32] P. Liu et al., "Pseudo-Bag Mixup Augmentation for Multiple Instance Learning-Based Whole Slide Image Classification," *IEEE Trans. Med. Imaging*, 2024.

[33] Y.-C. Chen and C.-S. Lu, "RankMix: Data Augmentation for Weakly Supervised Learning of Classifying Whole Slide Images with Diverse Sizes and Imbalanced Categories," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23936–23945, Jun. 2023.

[34] C. Akkus et al., "Multimodal Deep Learning," in *Proceedings of the 28th international conference on machine learning*, pp. 689-696, 2023.

[35] H. Li et al., "Multi-modal Multi-instance Learning Using Weakly Correlated Histopathological Images and Tabular Clinical Information," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, vol. 12908, 2021, pp. 529–539.

[36] W. Hou et al., "Hybrid Graph Convolutional Network with Online Masked Autoencoder for Robust Multimodal Cancer Survival Prediction," *IEEE Trans. Med. Imaging*, pp. 1–12, 2023.

[37] X. Wang et al., "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical Image Analysis*, vol. 81, p. 102559, Oct. 2022.

[38] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-Normalizing Neural Networks," *Advances in neural information processing systems,* vol. 30. 2017.

[39] I. Tolstikhin et al., "MLP-Mixer: An all-MLP Architecture for Vision," *Advances in neural information processing systems*, vol. 34, pp. 24261-24272, 2021.

[40] J. Arevalo et al., "Gated Multimodal Units for Information Fusion," *arXiv preprint arXiv*: 1702.01992, 2017.

[41] J. M. Dolezal et al., "Slideflow: Deep Learning for Digital Histopathology with Real-Time Whole-Slide Visualization," *arXiv preprint arXiv*: 2304.04142, 2023.

[42] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 3319–3328.

Authorized licensed use limited to: Chinese University of Hong Kong. Downloaded on May 29,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.