Journal of Pathology Informatics 15 (2024) 100359

Contents lists available at ScienceDirect

Journal of Pathology Informatics

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

Multimodal Gated Mixture of Experts Using Whole Slide Image and Flow Cytometry for Multiple Instance Learning Classification of Lymphoma

Noriaki Hashimoto [a,*], Hiroyuki Hanada [a], Hiroaki Miyoshi [b], Miharu Nagaishi [b], Kensaku Sato [b], Hidekata Hontani [c], Koichi Ohshima [b], Ichiro Takeuchi [a,d,*]

[a] RIKEN Center for Advanced Intelligence Project, Furo-cho, Chikusa-ku, Nagoya, 4648603, Japan
[b] Department of Pathology, Kurume University School of Medicine, 67 Asahi-machi, Kurume, 8300011, Japan
[c] Department of Computer Science, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 4668555, Japan
[d] Department of Mechanical Systems Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 4648603, Japan

ARTICLE INFO

Keywords:
Digital pathology
Mixture of experts
Multiple instance learning
Whole slide image
Flow cytometry

ABSTRACT

In this study, we present a deep-learning-based multimodal classification method for lymphoma diagnosis in digital pathology, which utilizes a whole slide image (WSI) as the primary image data and flow cytometry (FCM) data as auxiliary information. In pathological diagnosis of malignant lymphoma, FCM serves as valuable auxiliary information during the diagnosis process, offering useful insights into predicting the major class (superclass) of subtypes. By incorporating both images and FCM data into the classification process, we can develop a method that mimics the diagnostic process of pathologists, enhancing the explainability. In order to incorporate the hierarchical structure between superclasses and their subclasses, the proposed method utilizes a network structure that effectively combines the mixture of experts (MoE) and multiple instance learning (MIL) techniques, where MIL is widely recognized for its effectiveness in handling WSIs in digital pathology. The MoE network in the proposed method consists of a gating network for superclass classification and multiple expert networks for (sub)class classification, specialized for each superclass. To evaluate the effectiveness of our method, we conducted experiments involving a six-class classification task using 600 lymphoma cases. The proposed method achieved a classification accuracy of 72.3%, surpassing the 69.5% obtained through the straightforward combination of FCM and images, as well as the 70.2% achieved by the method using only images. Moreover, the combination of multiple weights in the MoE and MIL allows for the visualization of specific cellular and tumor regions, resulting in a highly explanatory model that cannot be attained with conventional methods. It is anticipated that by targeting a larger number of classes and increasing the number of expert networks, the proposed method could be effectively applied to the real problem of lymphoma diagnosis.

1. Introduction

The development of machine learning algorithms and the availability of whole slide images (WSIs) have greatly accelerated studies in digital pathology.[1–16] These advancements have particularly focused on class (subtype) prediction from hematoxylin-and-eosin (H&E)-stained tissue specimens, which are commonly used in pathological diagnosis. This approach provides a quantitative second opinion during the diagnostic process, aiming to reduce costs for pathologists and ensure more consistent diagnostic results. While current machine learning methods for digital pathology primarily utilize WSIs, the observation of H&E-stained tissue specimens alone is often insufficient for definitive diagnosis.In practical diagnosis, pathologists consider various factors, including basic clinical information, interview results such as performance status, and other test outcomes, in addition to the examination of tissue specimens. Furthermore, most cases require additional information such as immunohistochemically stained tissue specimens and genetic tests for making final decisions. To enhance the accuracy of diagnostic support tasks and provide explanations that closely mimic pathologists' decision making processes, a combination of digital images and auxiliary data is desired. In this paper, we propose a multimodal classification method that incorporates both WSIs and flow cytometry (FCM) for lymphoma pathology.

Conventional multimodal analysis methods typically treat auxiliary data as input features. For instance, previous studies such as[17–19] simply combine image feature and auxiliary data, while[20,21] encode the relationship between images and auxiliary data using transformer architecture.[22] However, the contribution of auxiliary data to the classification greatly depends on its characteristics. Some auxiliary data may be useful for classifying major categories, while others may only be useful for specific types of diseases. Therefore it is crucial to consider which information among the

* Corresponding authors.
E-mail addresses: noriaki.hashimoto.jv@riken.jp (N. Hashimoto), ichiro.takeuchi@mae.nagoyau.ac.jp (I. Takeuchi).

http://dx.doi.org/10.1016/j.jpi.2023.100359
Received 2 August 2023; Received in revised form 7 December 2023; Accepted 23 December 2023
Available online 29 December 2023

images and auxiliary data was considered and how it influenced the decision. From an explainability standpoint, it is desirable to be able to mimic the diagnostic process of pathologists and consider how images and auxiliary information contribute to the decision-making process of the diagnostic task. In practical lymphoma diagnosis, FCM is a commonly used auxiliary data that provides valuable information for classifying groups of subtypes (*superclasses*) such as B-cell or T-cell lymphoma. Hence, to replicate a pathologist's lymphoma diagnosis, a model is required that distinctly delineates the role of inputs, mimicking the actual diagnostic process. This may involve utilizing FCM solely for superclass classification, with final classification performed based on image data. We hypothesize that optimal integration of FCM and images would empower the model to yield highly explanatory classification results that are acceptable to the pathologist. Therefore, we develop a novel classification model that effectively combines multimodal inputs.

The mixture-of-experts (MoE) framework,[23] which employs hierarchies as a network structure, is particularly effective for classification problems with hierarchical superclasses and classes. In the MoE framework, the final output of the model is obtained by applying weights from a gating network to the outputs of multiple weak classifiers, known as experts. For instance, in lymphoma classification, each expert corresponds to a different superclass (B-cell lymphoma, T-cell lymphoma, or Others). The expert for B-cell lymphoma well classifies diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL), while the expert for T-cell lymphoma well classifies angioimmunoblastic T-cell lymphoma (AITL) and adult T-cell leukemia/lymphoma (ATLL). Moreover, the expert for Others effectively classifies classical Hodgkin lymphoma (CHL) and reactive lymphoid hyperplasia (RL). Cases of CHL exhibit low detectability in FCM, and RL is indicative of suspected lymphoma but not cancer. Both have been categorized as Others, constituting a superclass in which FCM does not identify abnormal cells. By utilizing the FCM data, which is informative for superclass classification, as input for the gating process, the model's final output can focus on the output of the expert corresponding to the superclass indicated by the FCM data. The MoE framework is effective not only for lymphoma classification but also for any problem with hierarchical class structures. In this paper, our objective is to achieve high classification accuracy by pretraining experts that have knowledge of a hierarchical class structure, encompassing both superclasses and classes. Fig. 1 provides an overview of the MoE model, which incorporates multimodal input for a classification problem with a hierarchical class structure.

Since WSIs are large digital images (e.g., 100,000 × 100,000 pixels), they cannot be directly fed into the classification model. Typically, image patches extracted from small regions of the tissue slide are used as inputs instead. Additionally, in most pathological image datasets, labels are assigned to WSIs rather than individual image patches due to the high cost of pathologists' annotations. This problem setting of WSI analysis can be addressed using multiple instance learning (MIL),[14–16,24,25] a weakly supervised learning method. It is widely acknowledged that MIL is valuable in digital pathology and MIL techniques are applied not only to classification task but also to segmentation[26] and survival analysis tasks.[27] In our proposed method, we incorporate the MIL mechanism into the MoE architecture to create a multimodal classification model. To validate the effectiveness of our proposed method, we conducted experiments using 600 lymphoma WSIs where we compared the accuracy of our approach against several baseline methods. Additionally, we visualized the computed weights in the model to confirm that each network worked well as expected. In this paper, as an initial step, the number of target classes was set to six, comprising three superclasses, each with two classes as illustrated in Fig. 1, while the total number of lymphoma subtypes is approximately 100 and there should exist the more suitable number of superclasses. To ensure successful learning and achieve the expected visualization with weights, the number of cases in each class was equalized. As the number of target classes in the model increases, aligning with the complexity of the actual lymphoma diagnosis problem, the proposed method aids the pathologist's decision-making process as the clinical application. It provides class predictions and the rationale for decisions for a given case through the combination of FCM and WSI. Table 1 shows a list of abbreviations used in this paper.

Our contributions in this paper are summarized as follows:

- We developed a multimodal classification method that incorporates WSI and FCM data, closely resembling pathologists' diagnostic process.
- We introduced a feature aggregation mechanism that combines MIL and MoE architectures.
- We provided explanations for the decision-making process by visualizing gating and attention weights.

2. Related works

2.1. Digital pathology

In the field of digital pathology, numerous machine learning algorithms have been developed for various tasks using digital pathological images. These tasks include classification,[1,28,29] segmentation,[2,3,30,31] detection,[32–34] survival analysis,[4,5,35] similar image retrieval,[6–8,36] and pathomics.[9,10] A WSI is a digital image that represents a pathological tissue

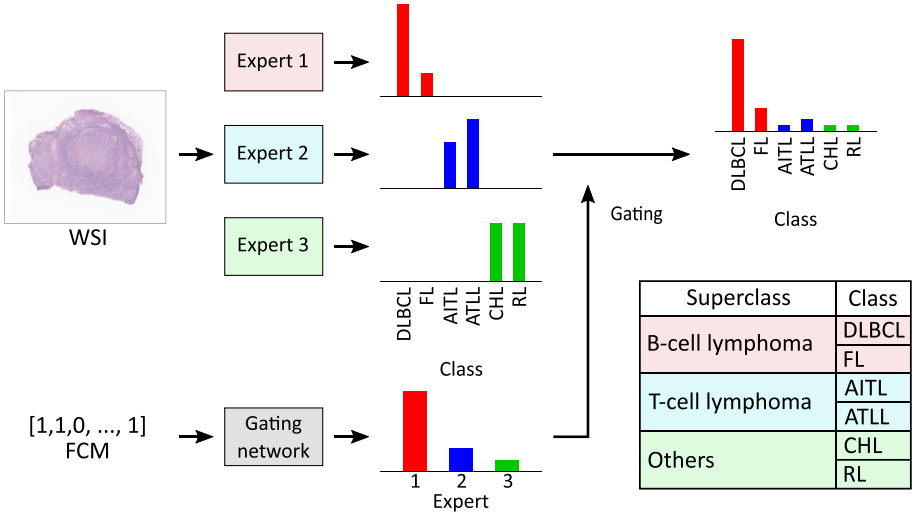

Fig. 1. An overview of the multimodal classification performed by the model utilizing the MoE architecture. This classification problem involves a hierarchy of superclasses and classes, as depicted in the table. The multimodal MoE architecture, equipped with an FCM-based gating network, demonstrates the ability to emulate pathologists' diagnoses and provide accurate predictions for each class.

Table 1
The list of abbreviations in this paper.

Abbreviation	Definition
CNN	convolutional neural network
FCM	flow cytometry
H&E	hematoxylin-and-eosin
MIL	multiple instance learning
MLP	multi-layer perceptron
MoE	mixture of experts
WSI	whole slide image
AITL	angioimmunoblastic T-cell lymphoma
ATLL	adult T-cell leukemia/lymphoma
CHL	classical Hodgkin lymphoma
DLBCL	diffuse large B-cell lymphoma
FL	follicular lymphoma
RL	reactive lymphoid hyperplasia

slide digitized by a WSI scanner, and these images can have sizes of up to about 100,000 × 100,000 pixels. Due to computational resource limitations, it is common to analyze small image patches extracted from WSIs as inputs for classification models such as convolutional neural networks (CNNs). Some studies in digital pathology concentrate on lymphoma classification, with methods typically applied to annotated cancerous images.[11–13] However, in most pathological image datasets, class labels are assigned to WSIs rather than individual image patches due to the high cost of pathologists' annotations of tumor regions. When predicting slide-level labels only from image patches without patch-level labels, MIL techniques are known to be effective.[24]

Multiple instance learning In MIL classification, the model is fed a *bag* that represents a set of instances, where class labels are not assigned to individual instances but rather to entire bags. In the simplest binary classification problem, a positive bag contains at least one positive instance, whereas a negative bag comprises only negative instances. When the classification problem extends to multi-class classification, each bag representing a class contains at least one class-specific instance. Note that the majority of MIL works in digital pathology concentrate on binary classification, distinguishing between cancer and normal cases. There have been relatively few papers published on multi-class classification in this context. The MIL classification model aggregates information from each instance to produce the final prediction for an input bag. In machine learning for digital pathology, the classification problem of WSIs can be formulated as MIL by regarding a single WSI (bag) as a collection of multiple image patches (instances). Particularly, attention-based MIL methods[14–16,24] have been successful in providing explainability in WSI classification, including lymphoma diagnosis.[14] These methods automatically compute important image patches for classification and can provide attention regions corresponding to tumor regions.

2.2. Multimodal analysis

In various medical fields, including pathological diagnosis, physicians make diagnoses based not only on images but also on additional auxiliary data such as clinical records. Hence, it would be effective to not only use images but also incorporate auxiliary information into the machine learning method. Multimodal analysis methods that combine images with clinical records have been studied in the medical field. For example, Yala et al.[18] improved risk prediction by combining mammography images and conventional risk factors.

Yap et al.[17] reported improved accuracy in classifying skin lesions and detecting melanoma by using dermoscopic images and clinical records. Li et al.[25] combined tabular clinical data with histological images in an MIL setting for lymph node metastasis prediction, where attributes including age, genes, and tumor location, were used as inputs along with multiscale histological images. Chen et al.[20] proposed a transformer-based multimodal model for survival prediction using images and genetic data. Their method enabled the visualization of co-attentions between images and genetic information. Another study[21] encoded the relationship between images and clinical records using a transformer mechanism, which provided exploratory and explanatory attention in classification and achieved high explainability. In pathology and other medical fields, the use of diagnostic information as auxiliary input, in addition to images, contributes to improving task accuracy.

2.3. Flow cytometry in lymphoma

FCM is a technique for measuring the optical and fluorescence characteristics of single cells.[37,38] In hematopathology, markers are attached to cells in collected specimens, and the presence or absence of abnormal cell populations for each marker is examined. In the pathological diagnosis of lymphoma, FCM data are commonly provided to pathologists along with H&E-stained tissue specimens from the laboratory. These data are presented in the form of "abnormal cell populations for CD20 and CD10" after post-processing by technologists.

Lymphoma, the focus of our study, consists of groups of subtypes. For example, DLBCL and FL belong to the B-cell lymphoma group, while AITL and ATLL belong to the T-cell lymphoma group. In this study, we call these groups of subtypes as *superclasses* while each subtype is referred to as a *class*. FCM data is valuable for distinguishing superclasses in lymphoma diagnosis, and pathologists often refer to FCM results when they examine H&E-stained tissue specimens. CD20 is a marker strongly associated with B-cell lymphoma, and the presence of abnormal cells for CD20 indicates a high possibility of B-cell lymphoma.

However, there are reliability issues with FCM. For example, some DLBCL cases do not show B-cell-related abnormal cell populations. The ability to detect abnormal cells varies depending on the class. In our experimental dataset, more than half of patients with AITL, a T-cell lymphoma, show no abnormal cell populations in T-cell-related markers. If FCM data from cases without abnormal cell populations are used for classification, they are likely to be misclassified as other superclasses, such as RL, which does not show abnormal cell populations in FCM. Therefore, it is challenging to use FCM data alone for classification or superclass classification. Previous studies using FCM for machine learning have focused only on the classification of B-cell lymphoma, which can be easily identified solely by FCM.[39,40] However, to address a wide range of superclasses and classes, it is essential to appropriately utilize FCM data.

2.4. Mixture of experts

When dealing with a classification problem involving superclasses and classes, as in our study, the mixture-of-experts (MoE) architecture, which incorporates hierarchies as a network structure, is effective.[23] MoE is a machine learning method where the final output is obtained by weighting the outputs of multiple networks, known as *experts*, using a gating network. Gating networks can employ soft gating,[23,41] which probabilistically weights the outputs of experts, or hard gating,[42] which selects the output of a single expert. MoE has been applied in digital pathology studies. For example, there is a segmentation method for histopathological images that introduces experts corresponding to different magnifications,[2] and a classification method for cytopathological images that introduces experts corresponding to different input modalities.[43]

In MoE architectures, the gating network usually receives the same input as the experts. By using FCM data as input for the gating network, we aim to construct a classification model that closely resembles the decision-making process of pathologists who utilizes FCM data for superclass classification[1]. For example, the first expert can be trained to classify DLBCL and FL effectively, the second expert can be trained to classify AITL and ATLL effectively, and the third expert can be trained to classify CHL and RL effectively. By incorporating FCM data, which is useful

[1] In an MoE trained from scratch, it is generally not guaranteed that each expert will be interpretable. However, when training data with labels for both superclass and (sub)class are available, we can pre-train each expert using only the data from each superclass. This allows us to create an MoE with experts specialized for each superclass.

for superclass classification, into the gating process, the model can focus on the output of the first expert if the FCM data indicates a B-cell-related abnormal cell population. However, due to the reliability issues with FCM data, cases without abnormal cell populations in FCM may still receive a higher weight for the output of the third expert, even if the correct superclass for the input case is B-cell lymphoma. To address the variability of FCM, we propose an MoE classification model that employs a robust gating method by combining images and FCM. Additionally, we aim to develop a highly explainable model by combining the MoE structure with the MIL mechanism, as MIL is effective for the classification of WSIs.

3. Proposed method

3.1. Problem setup

The dataset consists of N patients (cases) represented as $\mathcal{D} = \{(\mathbb{X}_n, \mathbb{T}_n, \mathbb{Y}_n)\}_{n=1}^{N}$, where each case comprises three components: a WSI denoted as $\mathbb{X}_n$, FCM table data represented by $\mathbb{T}_n$, and a corresponding label denoted as $\mathbb{Y}_n$. $\mathbb{T}_n$ is an L-dimensional binary vector that indicates the presence or absence of abnormal cell populations for each marker. $\mathbb{Y}_n$ is a K-dimensional one-hot vector that represents the subtype (class). Each element of FCM data $\mathbb{T}_n$ takes the value 1 if the presence of abnormal cell population for the corresponding marker is confirmed, or 0 otherwise. FCM data are available as inputs even for test cases, as they are provided to pathologists during the observation of H&E-stained tissue specimens for practical diagnosis. In the dataset with C superclasses and K classes, the superclasses are denoted as $c = 1, \ldots, C$ and the classes are denoted as $k = 1, \ldots, K$. Each of the K classes belong to exactly one of the C superclasses. In this study, the aim is to identify the class for the n-th patient when a WSI $\mathbb{X}_n$ and FCM data $\mathbb{T}_n$ are provided as inputs to the machine learning model. The output of the model is a vector $\hat{\mathbb{Y}}_n$ which aims to match the true one-hot encoded class $\mathbb{Y}_n$. Since the WSI $\mathbb{X}_n$ is a large digital image that cannot be directly fed into a classification model, the class for the n-th patient is determined by using M_n small image patches $\{\boldsymbol{x}_{n,m}\}_{m=1}^{M_n}$ extracted from $\mathbb{X}_n$. This problem setting is formulated as an MIL classification problem, where a model obtains the output $\hat{\mathbb{Y}}_n = f(\{\boldsymbol{x}_{n,m}\}_{m=1}^{M_n}, \mathbb{T}_n)$ from an input bag B_n consisting of multiple image patches $\{\boldsymbol{x}_{n,m}\}_{m=1}^{M_n}$ and FCM data $\mathbb{T}_n$.

3.2. Proposed model

Fig. 2 illustrates an overview of the proposed classification model, which consists of four components: (i) feature extractors, (ii) multimodal gating network, (iii) multiple sub-networks, and (iv) attention-based MIL. In the following discussions, the subscript n for the patient is omitted for notational simplicity.

(i) Feature extractors Feature extractors $g_{\text{img}} : \boldsymbol{x} \to \boldsymbol{h}$ and $g_{\text{FCM}} : \mathbb{T} \to \boldsymbol{h}^{(\text{FCM})}$ map image patches $\{\boldsymbol{x}_m\}_{m=1}^{M}$ and FCM $\mathbb{T}$ in a bag to feature vectors $\{\boldsymbol{h}_m\}_{m=1}^{M}$ and $\boldsymbol{h}^{(\text{FCM})}$. The functions g_{img} and g_{FCM} are implemented by CNN and multi-layer perceptron (MLP) respectively. A feature extractor g_{img} which is pre-trained with images of all classes in advance can obtain better common features to be input to the sub-networks mentioned later.

(ii) Multimodal gating network The proposed method uses FCM data as an input to the gating network in the MoE architecture following the practice of pathologists who use FCM as a reference for superclass classification in diagnosis. However, performing gating for superclass classification using only FCM data might lead to misclassification due to the detection ability of abnormal cell populations. To address this, multimodal

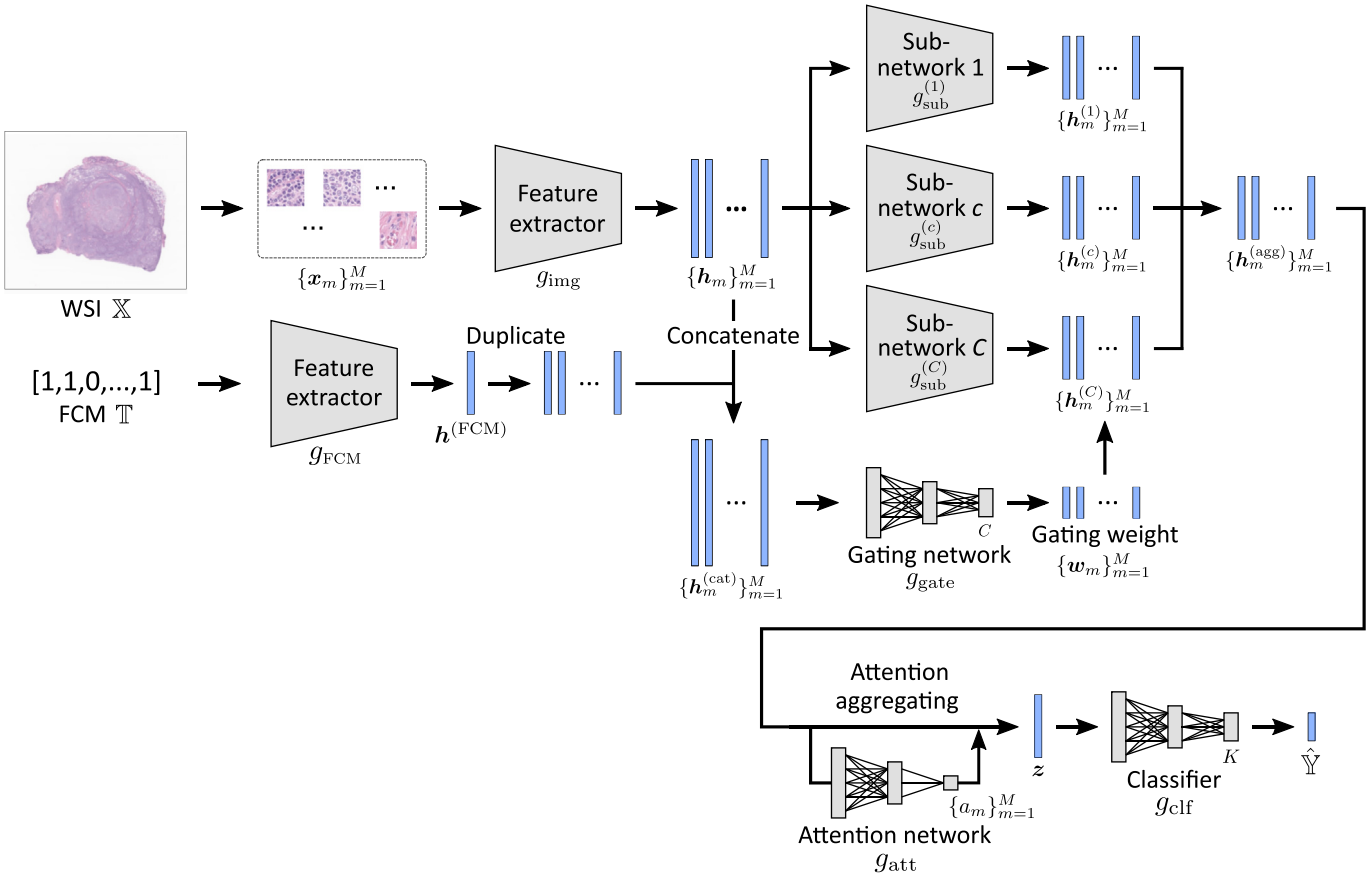

Fig. 2. The proposed classification model, which received a set of multiple image patches and FCM data, consists of four components: (i) feature extractors, (ii) multimodal gating network, (iii) multiple sub-networks, and (iv) attention-based MIL.

inputs combining image patches and FCM data enable the computation of gating weights that are robust to the variability of FCM for each patch. An FCM feature $\boldsymbol{h}^{(\mathrm{FCM})}$ is replicated for each image patch in a bag, and the gating weight for an image patch $\boldsymbol{x}_m$ is calculated by an MLP-structured gating network g_{gate} as follows:

$$\boldsymbol{w}_m = g_{\mathrm{gate}}\left(\boldsymbol{h}_m^{(\mathrm{cat})}\right),$$

where $\boldsymbol{h}_m^{(\mathrm{cat})} = \mathrm{cat}(\boldsymbol{h}_m, \boldsymbol{h}^{(\mathrm{FCM})})$ is a feature vector obtained by concatenating the image feature $\boldsymbol{h}_m$ and the FCM feature $\boldsymbol{h}^{(\mathrm{FCM})}$. Each element $w_m(c)$ of the C-dimensional vector $\boldsymbol{w}_m$ represents the weight assigned to the output of the c-th sub-network for the image patch $\boldsymbol{x}_m$ ($\Sigma_c w_m(c) = 1$). To control the balance between soft and hard gating, a temperature parameter T is introduced in the softmax activation function of the gating network.[44] A smaller temperature parameter results in sparser weights, emphasizing the sub-networks on which each patch should focus.

(iii) Multiple sub-networks Sub-networks $g_{\mathrm{sub}}^{(c)}$ receive a common image feature $\boldsymbol{h}_m$ and output features specific to class classification within different superclasses. Each sub-network corresponds to an expert in the standard MoE architecture and is pre-trained to provide encoded features for accurately predicting classes in the corresponding superclass. For instance, consider the three superclasses "B-cell," "T-cell," and "Others." Then, the sub-network $g_{\mathrm{sub}}^{(1)}$ provides image features for accurately classifying DLBCL and FL, which are B-cell lymphomas. Second, the sub-network $g_{\mathrm{sub}}^{(2)}$ provides image features for accurately classifying AITL and ATLL, which are T-cell lymphomas. Finally, the sub-network $g_{\mathrm{sub}}^{(3)}$ provides image features for accurately classifying CHL and RL, which are other lymphoma and non-lymphoma. Each image feature $\boldsymbol{h}_m$ is mapped as $\boldsymbol{h}_m^{(c)} = g_{\mathrm{sub}}^{(c)}(\boldsymbol{h}_m)$ by the c-th sub-network, which provides an appropriate representation for superclass c. Following the method described in[23] (see also Section 2.4 above), the C feature vectors obtained by the C sub-networks for each image patch are aggregated using the gating weight $\boldsymbol{w}_m$ as follows:

$$\boldsymbol{h}_m^{(\mathrm{agg})} = \sum_{c=1}^{C} w_m(c)\boldsymbol{h}_m^{(c)}.$$

The features of each image patch in a bag, aggregated as described above, are expected to be more discriminative due to the multimodal gating, which plays a role in superclass classification. By pre-training an MIL classification model for each superclass (e.g., a model to classify DLBCL and FL) using the common feature extractor and trainable sub-networks, it is possible to acquire sub-networks specialized for each superclass.

(iv) Attention-based MIL The image features $\boldsymbol{h}_m^{(\mathrm{agg})}$, which are weighted sums of outputs of the sub-networks based on the gating weights, are aggregated into a single bag feature using attention-based MIL.[24] An attention network g_{att}, with an NN architecture, computes an attention weight a_m that indicates the contribution of an image feature $\boldsymbol{h}_m^{(\mathrm{agg})}$ to the classification. The aggregated feature representing an input bag is calculated by using the weighted image features and attention weights as follows,

$$\boldsymbol{z} = \sum_{m=1}^{M} a_m \boldsymbol{h}_m^{(\mathrm{agg})}.$$

The classifier g_{clf} provides the prediction $\hat{\mathbb{Y}}$ with $\boldsymbol{z}$ as input. The entire classification model is optimized to minimize the cross-entropy loss function between the prediction $\hat{\mathbb{Y}}$ and the correct label $\mathbb{Y}$.

Training of the proposed model To successfully train the proposed model, it is essential to pre-train a feature extractor capable of representing all classes and sub-networks proficient in classifying within each superclass. Before initiating training for the proposed model, the feature extractor and sub-networks are initialized by pre-training two types of MIL classification models illustrated in Fig. 3. Both model structures are basic attention-based MIL classification models, as explained in the previous paragraph. Firstly, MIL classification for all classes is conducted, enabling the model to identify an input WSI into target six classes. By training this classification model, we obtain the common feature extractor g_{img} capable of computing image features representing all six classes. Following that, three two-class MIL classification models for three superclasses are trained. The previously pre-trained feature extractor, with fixed model parameters, is utilized in all three models. For training sub-network $g_{\mathrm{sub}}^{(c)}$, only WSIs belonging to superclass c are used. When the number of classes in the c-th superclass is denoted as K_c, the output of the model for superclass c is the K_c-dimensional vector $\hat{\mathbb{Y}}_c$, representing the predictions for classes within superclass c. In our problem setting, K_c is set to two for any c, and each sub-network is trained so that the computed features effectively discriminate between two classes within the corresponding superclass. The proposed model is then trained using these pre-trained feature extractor and sub-networks as the initial parameters.

Role of attentions The proposed classification model provides two types of attentions: gating weights and attention weights. The gating weights indicate which sub-network outputs should be focused on, whereas

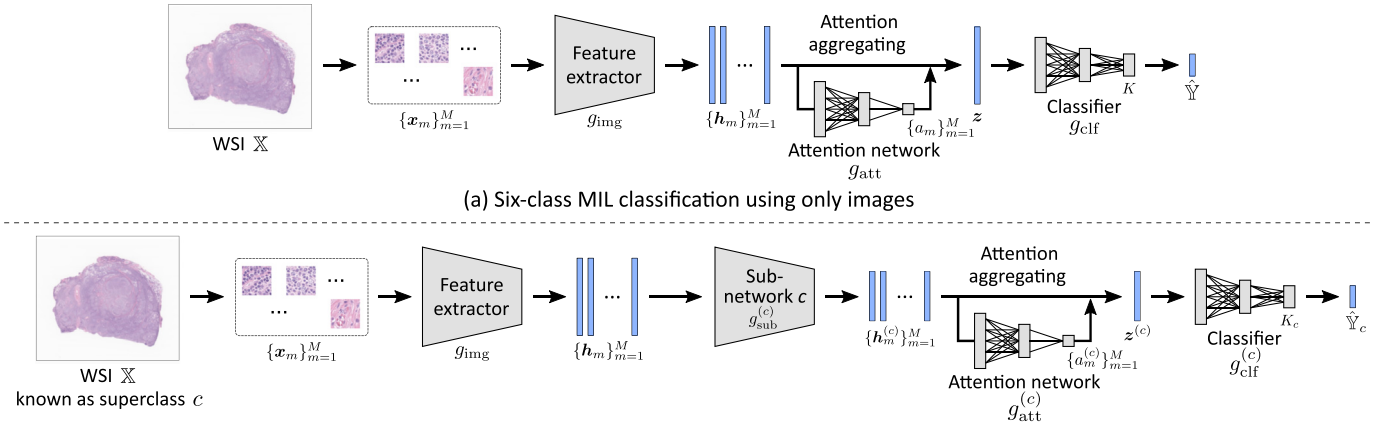

Fig. 3. Pre-training involves the common feature extractor and multiple sub-networks, accomplished through training two types of MIL classification models. (a) Initially, a six-class MIL classification is trained using all-class data to obtain a feature extractor with representations of all classes. (b) Subsequently, to pre-train the sub-networks, two-class MIL classification is performed for each superclass using data from the corresponding classes.

the attention weights indicate which image patches should be focused on. In other words, the gating weights indicate which superclass each image patch belongs to, considering the FCM data, and can quantitatively indicate how each image patch exhibits features specific to B-cell lymphoma, T-cell lymphoma, or Others. On the other hand, attention weights are expected to represent something like tumorness since they indicate which image patches possess class-specific characteristics. However, for RL cases (which are not lymphomas), attention weights show a degree of RL-specific features rather than tumors. For example, if the gating weight for the sub-network of B-cell lymphoma is large for an image patch, and its attention weight is large as well, it indicates a tumor region of B-cell lymphoma. In contrast, if the gating weight for the sub-network of T-cell lymphoma is large for an image patch, but its attention weight is small, it indicates a normal region of T-cell lymphoma. In this study, the highly explainable classification model is realized by visualizing these two types of weights to show the classification result.

4. Experiments

We conduct experiments to compare classification accuracy and visualize computed weights to validate the effectiveness of the proposed method.

4.1. Experimental setting

Dataset In the experiment, we utilized a private clinical dataset diagnosed at Kurume University. The dataset consists of H&E-stained tissue specimens and FCM provided by a single laboratory company. A class label was assigned to each case based on the definitive diagnosis according to the WHO classification, determined by an expert hematopathologist, taking into account additional immunohistochemical staining and genetic testing. The dataset comprises $N = 600$ clinical cases, including six classes: 100 DLBCL, 100 FL, 100 AITL, 100 ATLL, 100 CHL, and 100 RL cases. DLBCL and FL belong to superclass 1, specifically B-cell lymphoma, while AITL and ATLL are classified under superclass 2, representing T-cell lymphoma. CHL and RL fall into superclass 3, categorized as Others. In practical pathology, lymphoma comprises approximately 100 classes with varying case numbers for each class. However, in this study, we focused on only six classes, categorized into three superclasses. The number of cases for each class was uniformly set to 100, aiming to effectively train the proposed method and assess its effectiveness in this initial phase.

H&E-stained tissue specimens were digitized into WSIs using a WSI scanner (Aperio GT450; Leica Biosystems, Germany) at a magnification of 40x (0.26 μm/pixel). FCM data was obtained through a flow cytometer and utilized as inputs in the form of binary vectors, indicating the presence or absence of abnormal cell populations, which were processed by a technologist. In this study, we employed a total of 18 antibodies shown in Table 2, resulting in an 18-dimensional binary vector representation of the FCM data ($\mathbb{T}_n \in \{0,1\}^{18}$). The FCM data contains valuable information for superclass classification, as mentioned earlier. However, there are cases where abnormal cells are not identified by markers associated with the actual superclass. Table 3 displays the presence or absence of abnormal cell populations among the 600 cases used in the experiment. As observed in the table, a majority of the CHL and RL cases lack abnormal cell populations. Additionally, there are cases of B-cell and T-cell lymphoma where abnormal cell populations cannot be identified, leading to their classification as Others solely based on FCM data. While FCM serves as a powerful tool for assessing cellular function, it is crucial to recognize its limitations and the fact that it may not always provide reliable data.

Table 2
The types of markers used in FCM data.

Associated superclass	Marker
B-cell lymphoma	CD10, CD19, CD20, CD23, kappa, lambda
T-cell lymphoma	CD2, CD3, CD4, CD5, CD7, CD8
Others	CD11c, CD16, CD25, CD30, CD34, CD56

Table 3
Presence or absence of abnormal cell populations in the 600 cases of the experiment. Cases where the FCM data shows no abnormal cell populations are expected to be classified as Others within the superclass.

Superclass	Class	With abnormal cells	Without abnormal cells
B-cell lymphoma	DLBCL	87	13
	FL	96	4
T-cell lymphoma	AITL	40	60
	ATLL	84	16
Others	CHL	1	99
	RL	3	97

Baseline methods We compared the proposed method with the following baseline methods.The model structures for methods 3, 4, and 5 are shown in Fig. 4. As mentioned in Section 2, various methods utilizing both images and auxiliary data have been proposed. However, our approach makes the final prediction solely from images. To confirm the suitability of the image features obtained through the proposed MoE architecture for classification, we aim to keep the models of the comparative methods as simple as possible. Note that method 1 is an optimistic baseline that assumes a hypothetical situation where the correct superclass classification is given as an oracle.

(1) Two-class MIL classification model with known superclass (optimistic method)

In this method, three distinct two-class MIL classification models illustrated in Fig. 3(b), which are learned to obtain sub-networks for each superclass during the pre-training step of the proposed method in Section 3, are utilized. As mentioned in Section 3, the classification model for superclass c is trained using only WSIs belonging to superclass c. The feature extractor trained in method 2 is used with the fixed parameters as the common feature extractor in all three models. During the testing step, classification is performed by assuming a hypothetical situation where the superclass of an input case is given as an oracle. For instance, when the input test cases are DLBCL or FL, the two-class MIL classification model for B-cell is used for testing. This method represents the ideal situation when the superclass can be classified with 100% accuracy. The sub-networks specialized for each superclass are used as the initial parameters of the corresponding sub-networks in all MoE-architecture method, including the proposed method.

(2) Six-class MIL classification model using only images

This method is a simple six-class MIL classification model illustrated in Fig. 3(a) that does not use MoE architectures and consists of only a feature extractor, an attention network, and a classifier. The training of this model is equivalent to one of the pre-training steps outlined in the proposed method in Section 3. The feature extractor trained in this method is used as the fixed feature extractor in other methods, ensuring that the models acquire common image features across all classes.

(3) Hierarchical classification using images and FCM

Hierarchical classification is performed through the two-class MIL classification models used in method 1, following the superclass classification using FCM data. For instance, when the superclass classifier g_{sclf} predicts the superclass of an input case as 1, the two-class classification model for superclass 1 (B-cell lymphoma) is employed to classify the input WSI as either DLBCL or FL. Thus, if g_{sclf} misclassifies the superclass of an input case, the final predicted class will be incorrect. Method 1 represents the ideal case for this method, assuming a superclass classification accuracy of 100%.

(4) MoE classification with gating network using FCM

In this method, a gating network that uses only FCM data as input outputs the same gating weights to all image patches in the WSI. The feature extractor trained in method 2 and the sub-networks trained in method 1 are used as the initial parameters in the model, optionally re-training the sub-networks. The gating network g_{gate} outputs a single three-dimensional weight vector $\boldsymbol{w}$ for all image patches in the bag.

(5) MoE classification with gating network using images

This method involves inputting image patches to both sub-networks and the gating network without FCM data, where the gating network outputs different gating weights for each image patch. Similar to method 4, the feature extractor trained in method 2 and the sub-networks trained in method 1 are used as the initial parameters in the model, optionally re-training the sub-networks.

Implementation details We explain the details of the model architectures and the training procedure. ResNet50[45] is employed as a feature extractor g_{img}, and a 2048-dimensional vector after global average pooling is obtained as its output $\boldsymbol{h}_m$. Among all experiments, the six-class MIL classification model shown in Fig. 3(a) is initially trained. This involves learning the classification model for method 2 and acquiring the common feature extractor for each of the other methods. For the six-class classification model, the aforementioned feature extractor pre-trained with ImageNet is fine-tuned, and the learned feature extractor g_{img} is employed for all other methods. The feature extractor for FCM data, denoted as g_{FCM}, is a two-layer neural network that transforms the 18-dimensional FCM vector $\mathbb{T}$ into a 128-dimensional feature vector $\boldsymbol{h}^{(\text{FCM})}$ using 128 output units and the ReLU activation function. The increase in the dimensionality of the output vector compared to the input is intended to handle the numerical data from FCM itself. Although the input space is small in this case as it is based on the presence or absence of abnormal cell populations, we are implementing a generalized model. The structures of the other networks are as follows. The gating network g_{gate} is a three-layer MLP with a hidden layer consisting of 512 units and the ReLU activation function, which outputs a three-dimensional weight vector $\boldsymbol{w}_m$ through the softmax activation function. However, the gating network g_{gate} specific to method 4 is a three-layer MLP with a hidden layer of 128 units and the ReLU activation function, also producing a three-dimensional weight vector $\boldsymbol{w}_m$ through the softmax activation function. The sub-networks $g_{\text{sub}}^{(c)}$ are three-layer MLPs with 1024 hidden units and the ReLU activation function, yielding

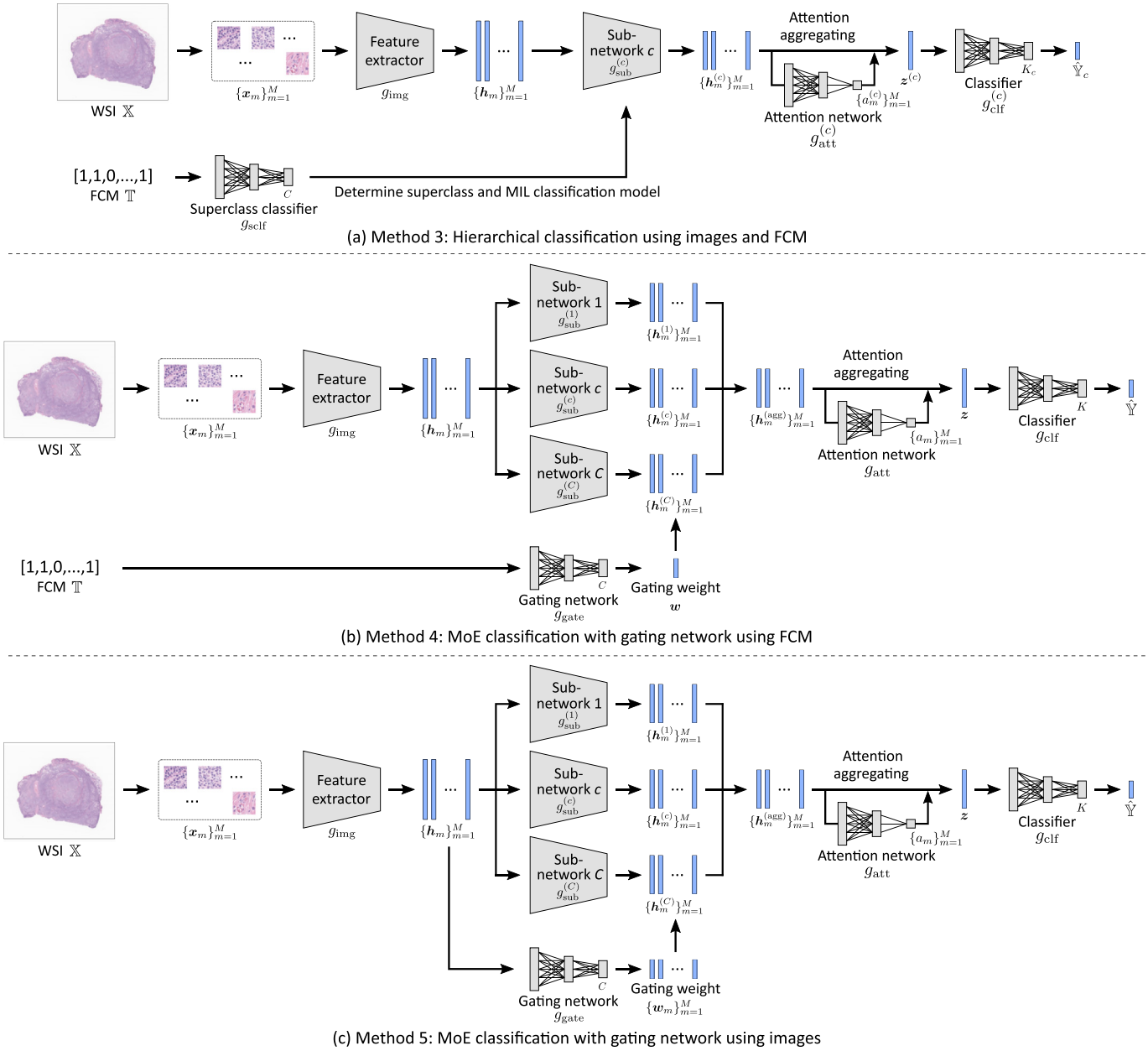

Fig. 4. Model structures of the compared methods: (a) Hierarchical classification using images and FCM, (b) MoE classification with gating network using FCM, and (c) MoE classification with gating network using images.

Table 4
Comparison of classification performance for each method. Each value represents the mean and standard error from five-fold cross-validation. The MoE-architecture models offer options to re-train sub-networks (Frozen/Trained) and the temperature parameter ($T = 0.5/T = 0.8$) of gating weights.

Method	Option	Accuracy	Precision	AUC
Method 1	Oracle	0.842 ± 0.014	0.846 ± 0.014	0.980 ± 0.002
Method 2		0.702 ± 0.013	0.713 ± 0.015	0.909 ± 0.003
Method 3		0.695 ± 0.013	0.726 ± 0.019	0.869 ± 0.010
Method 4	Frozen, $T = 0.5$	0.557 ± 0.012	0.623 ± 0.032	0.852 ± 0.014
	Trained, $T = 0.5$	0.582 ± 0.039	0.593 ± 0.049	0.863 ± 0.017
	Frozen, $T = 0.8$	0.553 ± 0.027	0.579 ± 0.027	0.845 ± 0.011
	Trained, $T = 0.8$	0.580 ± 0.035	0.606 ± 0.040	0.856 ± 0.017
Method 5	Frozen, $T = 0.5$	0.558 ± 0.013	0.571 ± 0.022	0.858 ± 0.008
	Trained, $T = 0.5$	0.660 ± 0.016	0.681 ± 0.017	0.893 ± 0.009
	Frozen, $T = 0.8$	0.555 ± 0.020	0.562 ± 0.026	0.859 ± 0.009
	Trained, $T = 0.8$	0.642 ± 0.010	0.660 ± 0.011	0.890 ± 0.007
Proposed	Frozen, $T = 0.5$	0.670 ± 0.030	0.679 ± 0.030	0.910 ± 0.012
	Trained, $T = 0.5$	0.722 ± 0.016	**0.738 ± 0.015**	0.917 ± 0.010
	Frozen, $T = 0.8$	0.710 ± 0.014	0.718 ± 0.016	0.911 ± 0.006
	Trained, $T = 0.8$	**0.723 ± 0.007**	0.737 ± 0.008	**0.924 ± 0.009**

the 512-dimensional vector $\boldsymbol{h}_m^{(c)}$. The attention network g_{att} is a three-layer MLP with 128 hidden units and the hyperbolic tangent activation function, outputting a scalar. The classifier g_{clf} is a three-layer MLP with 128 hidden units and the ReLU activation function that outputs a six-dimensional class prediction vector through the softmax activation function. For the six-class MIL classification (method 2), the numbers of hidden units in the attention network g_{att} and the classifier g_{clf} are set to 512, as the length of input feature vectors is 2048. The superclass classifier g_{sclf} is a three-layer MLP with 128 hidden units and the ReLU activation function that provides a three-dimensional superclass prediction vector through the softmax activation function. The initialization of network parameters is random, except for the ResNet50-based feature extractor g_{img} trained in the six-class classification. In MoE-based methods, the parameters of the sub-networks $g_{\text{sub}}^{(c)}$ are initialized with those of the sub-networks trained in the two-class MIL classification model within each superclass, as illustrated in Fig 3(b).

As a pre-processing step, all WSIs are divided into tiled image patches of 224×224 pixels at 40x magnification. The average chromaticity within image patches is thresholded using Otsu's binarization to eliminate glass regions from the entire slide, extracting only image patches of tissue areas. As a result of the aforementioned processing, each WSI contains 100 or more image patches. An input bag consists of 100 randomly selected 224×224-pixel image patches from the entire tissue at 40x magnification, along with an 18-dimensional FCM vector ($M_n = 100$ for any n). During the training step, up to 5000 image patches are randomly selected from each WSI at each epoch, resulting in the creation of a maximum of 50 bags for each patient. No image patch overlaps between bags, and each image patch is used only once in a single epoch of training. To enhance the model's robustness to the randomness of image patches within the bag, 5000 image patches are randomly selected from each WSI in every epoch. The determination of the number of image patches in a bag and the maximum number of bags for each WSI is based on previous literature regarding the classification of lymphoma WSIs.[14] On the other hand, since FCM data is provided for each case, the same FCM data is assigned to bags of the same case. During the test step, a maximum of 50 bags are created from a single case, and the class predictions of all bags for the case are averaged to make the case-level class prediction. The selection of patches and the creation of bags in testing are the same across all methods. The network parameters are optimized through 10-epoch training using SGD momentum. The learning rates are set to 0.01 (method 2) and 0.001 (all other methods), while the momentum is set to 0.9. The learning rate is scheduled to decrease by a factor of 0.1 every 5 epochs. For the superclass classifier g_{sclf}, it is trained to predict superclass of an input case using only FCM data. The optimization is performed with 50-epoch training using SGD momentum, with the learning rate and momentum set to 0.001 and 0.9, respectively. In the training of the superclass classifier g_{sclf}, the learning rate is scheduled to decrease by a factor of 0.1 every 10 epochs.

Table 5
Confusion matrices of (a) the proposed method and (b) method 3.

(a) The confusion matrix of the proposed method (Trained, $T = 0.8$).

		Predict					
		DLBCL	FL	AITL	ATLL	CHL	RL
Correct	DLBCL	74	13	1	6	4	2
	FL	14	80	0	1	0	5
	AITL	0	1	67	10	12	10
	ATLL	2	0	15	74	4	5
	CHL	2	1	14	1	64	18
	RL	0	1	9	4	11	75

(b) The confusion matrix of method 3.

		Predict					
		DLBCL	FL	AITL	ATLL	CHL	RL
Correct	DLBCL	75	12	0	0	4	9
	FL	16	80	0	0	0	4
	AITL	1	1	30	7	41	20
	ATLL	1	0	12	71	2	14
	CHL	1	0	0	0	83	16
	RL	0	1	0	2	19	78

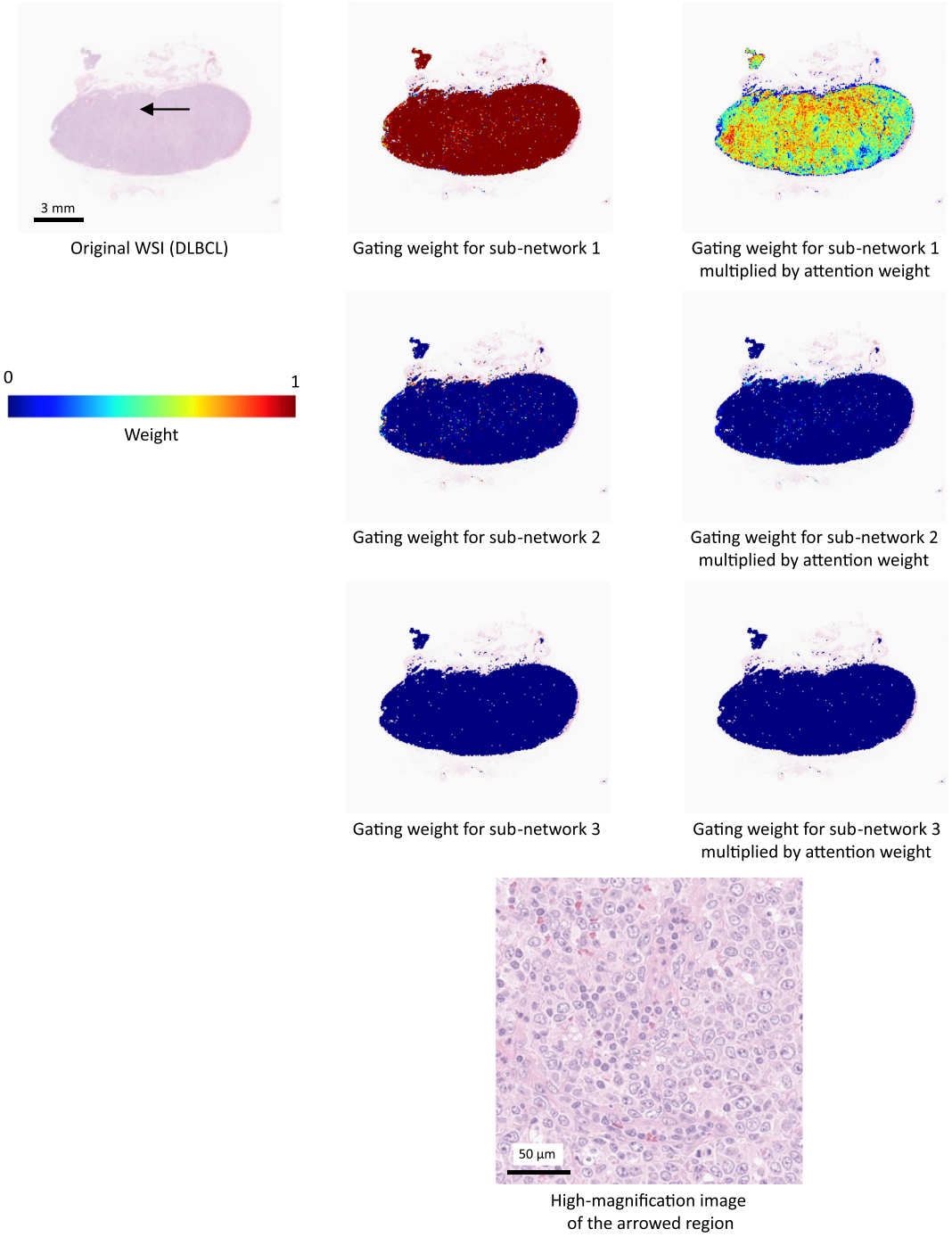

Fig. 5. An example of the visualization result for gating weights for each sub-network (middle column) and those multiplied by attention weights (right column). The high-value regions in the visualized images in the right column are interpreted as the class-specific regions that contribute to the final class prediction.

Our dataset is private and not publicly available. However, we published python source code using PyTorch library on GitHub https://github.com/mmmoe-ML/MMMoE.

4.2. Quantitative evaluation

The classification performance of each method was evaluated using five-fold cross-validation. In each fold, the training, validation, and test data were divided into 60%, 20%, and 20%, respectively, at the WSI level. The experiments were conducted five times, with changes in the data splitting to ensure that all WSIs were used for testing at least once. The same data splitting was applied to all compared methods, stratified across the six classes. Validation data was utilized to select the best model during training, and the model showing the smallest validation loss was chosen for testing after 10-epoch training. The results are presented in Table 4, where each value represents the mean and standard error from five-fold cross-validation. Precisions and AUCs were calculated for each class, and then macro-averaged. In method 3, the AUC was calculated after determining the classification model used through superclass classification by FCM-based MLP. For the models with MoE architecture, the parameters of the feature extractor were fixed, and the parameters of sub-networks were optionally re-trained (indicated as Frozen/Trained, respectively) to ensure that the roles of each sub-network were not significantly altered. To sparsify the gating weights and highlight the salient

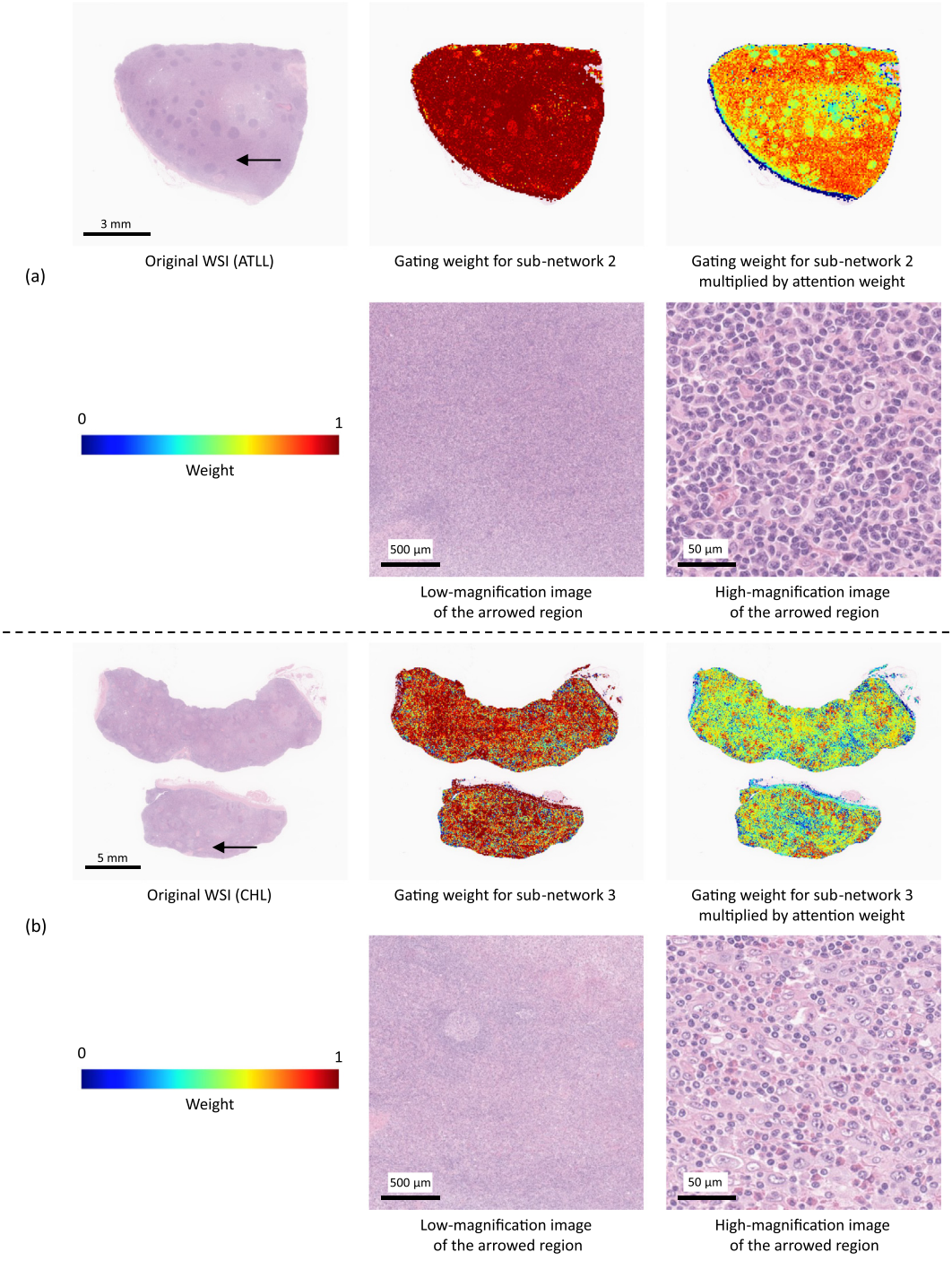

Fig. 6. The visualization result of gating and attention weights: (a) the visualization results of an ATLL case where abnormal cells are found in FCM data, and (b) the visualization results of a CHL case where there are no abnormal cells in FCM data but appropriate weights can be output by the gating network.

sub-networks, the temperature parameter T of the softmax function at the output of the gating network was employed. When the temperature parameter is set to smaller value, the weight values after the softmax function become more sparse. We selected two values, $T = 0.5$ and 0.8, to examine their effect on classification performance. However, it is important to note that in practical applications, the optimal temperature parameter should be determined using validation data.

The results demonstrate that the proposed method achieved the highest accuracy among the compared methods, except for the optimistic method. While both the proposed method and method 2 use a single feature vector to perform final classification by aggregating image features, the proposed method obtains more discriminative features by aggregating features with higher expressive power. In methods 3 and 4, which directly use FCM data for superclass classification, the FCM data have a dominant influence on the classification, and misclassification of cases with no abnormal cell population strongly deteriorates the results. The accuracy of FCM-based superclass classification using an MLP in method 3 was 0.831, leading to lower overall classification performance for method 3. While the proposed multimodal gating is effective, it is not flawless and may occasionally assign inappropriate weights based on FCM data. For instance, there are cases of T-cell lymphoma without abnormal cell populations in FCM data. To address misclassifications by the multimodal gating network, it is considered that re-training the sub-networks could enhance the overall classification performance.

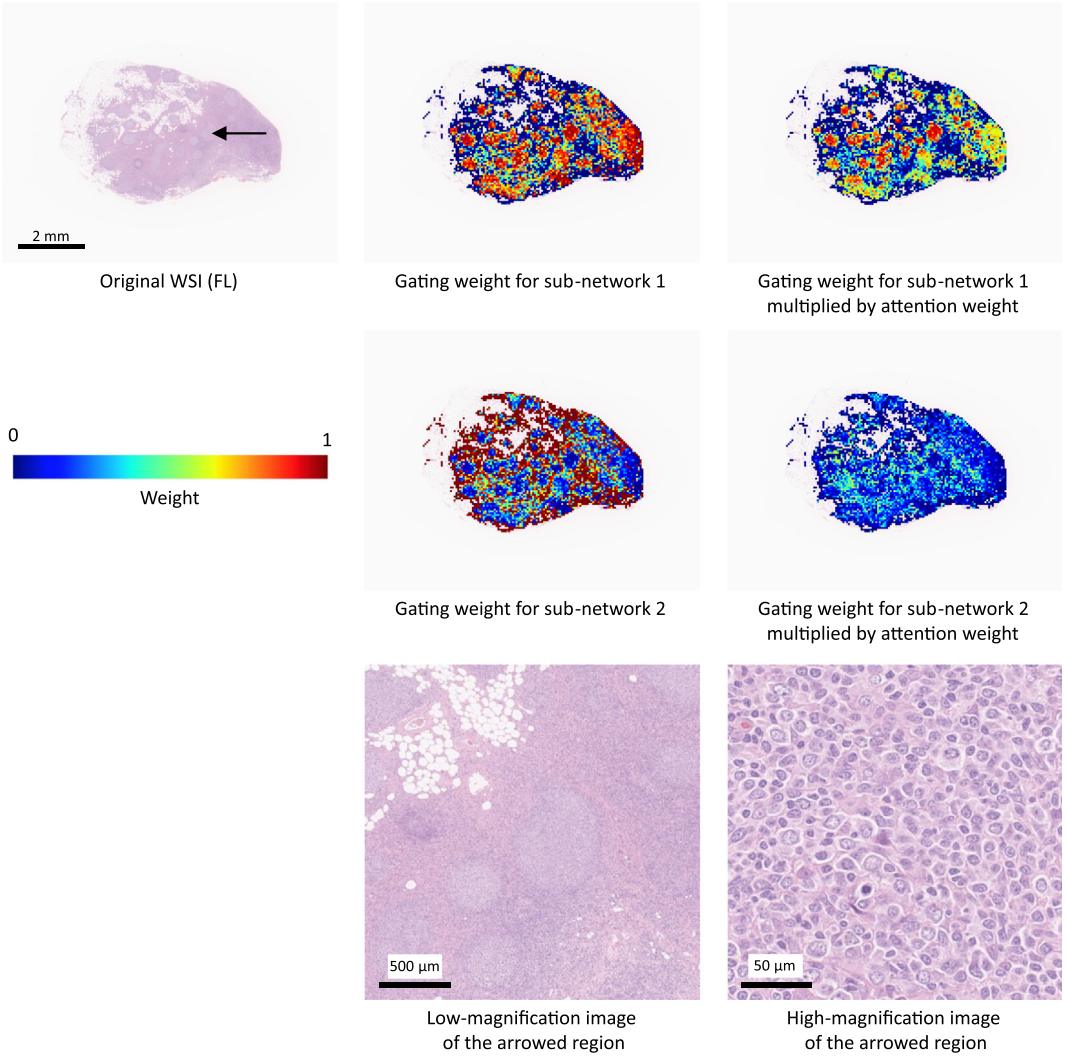

Fig. 7. The visualization results of gating and attention weights for an FL case.

Additionally, Table 5 presents the confusion matrices of the proposed method and method 3. The multimodal gating in the proposed method improves classification accuracy and provides robust results for AITL cases, even when abnormal cell populations are not detected. However, the attempt to incorporate both image and FCM data, even in cases without abnormal cell populations in FCM, has introduced a new issue where CHL is misclassified as a T-cell case. Introducing a gating network that incorporates additional medical record information, such as patient details, might enhance superclass classification and overall accuracy in classification. Although the improvement in accuracy achieved by the proposed method is relatively modest, the task of multi-class classification of lymphoma solely from H&E-stained WSIs is inherently challenging. A significant contribution lies in the visualization of multiple weights, as detailed in the next section, a capability unique to the proposed method.

4.3. Visualization of weights

We conducted visualization of gating and attention weights to provide an explanation for the decision-making process. In this experiment, we calculated gating and attention weights for all image patches in the entire WSI. Gating weights were generated as continuous values between 0 and 1 using the softmax function, allowing us to visualize them as a heatmap ranging from blue to red. Blue represents a weight of 0, while red represents a weight of 1. During the classification experiment, the attention network outputs were normalized to a range between 0 and 1 using the softmax function within the bag, ensuring that the sum of attention weights in the bag became 1. However, in the visualization experiment, the attention network outputs for all image patches in the tissue slide were normalized to a range of 0 to 1 based on the minimum and maximum values. Fig. 5 illustrates an example of the visualization of gating weights for each sub-network, as well as their multiplication by attention weights. The middle column displays the visualization of gating weights, where red regions indicate image patches that the model identified as superclass-specific regions associated with the sub-network using information from images and FCM data. The right column demonstrates the visualization of the multiplication of gating weights and attention weights, where the remaining red regions are interpreted as class-specific regions (i.e., tumor regions) that contribute to the classification within the superclass. The high-magnification image displays the region with higher attention, allowing us to identify the large cells commonly observed in DLBCL.

The cases discussed below were carefully chosen as representative examples that are easily interpretable. To analyze the visualization results of correctly predicted cases, an expert pathologist (one of the authors, who is an institution member with over 15 years of experience diagnosing more than 10000 cases of lymphoma) provided valuable insights about the relationship between tumor regions and visualized weights.

Fig. 6 shows (a) the visualization results of an ATLL case where abnormal cells are found in FCM data, and (b) the visualization results of a CHL case where there are no abnormal cells in FCM data but appropriate weights can be output by the gating network. The FCM data of this ATLL

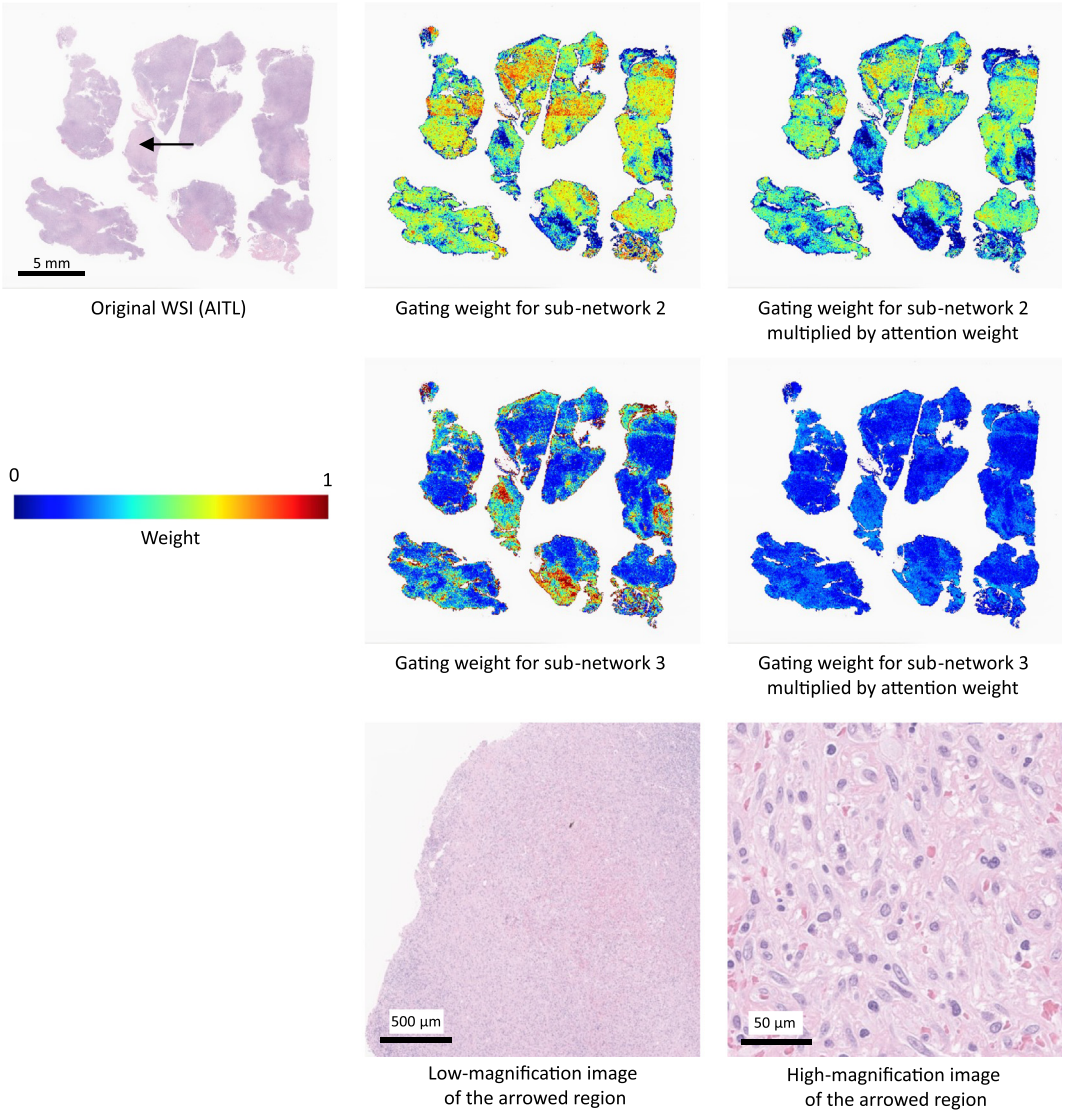

Fig. 8. The visualization results of gating and attention weights for an AITL case where abnormal cells are found in FCM data.

case show significant T-cell lymphoma-specific information, and the outputs of sub-network 2 were emphasized at almost the entire tissue region. On the other hand, the weights for sub-network 2 after multiplying the attention weights show lower values in the normal follicular regions, and it can be confirmed that the model predicted the correct class by focusing on the interfollicular T-cell regions. Although abnormal cell populations are not usually identified in CHL cases, the absence of abnormal cells in the dataset used in this study is a reason for superclass classification as CHL or RL. As a result, the outputs of sub-network 3 are focused by combining images and FCM data, and attention weights are higher in the regions where the density of cell nuclei in the tissue specimen is low (i.e., the color is lighter). The Hodgkin cells characteristic of Hodgkin lymphoma often occur in these areas of low cell density, and the visualization results are consistent with the pathologist's perception of the diagnosis. We can see that Hodgkin cells actually exist in the high-magnification image.

Figs. 7 and 8 show the visualization results for an FL case and an AITL case respectively, both of which exhibit abnormal cells in FCM data. In the FL case, the outputs of sub-network 1 are focused in the follicular region, while the outputs of sub-network 2 are focused between follicular regions. The structure in which B-cells form a follicular structure and T-cells are placed around it is typical of FL cases, and we can see that the model predicted the final class as FL using the tumor features in follicular regions. In the AITL case, the model focused on the outputs of sub-network 3 in the light-colored regions of the tissue specimen, which shows that those regions were suspected as the specific regions associated with CHL. However, unlike the CHL case in Fig. 6 (b), because there is no actual Hodgkin cells and they were not considered tumor area, the case was classified as AITL based on information from the region whose outputs of sub-network 2 were focused. The light-colored regions of the tissue specimen in the AITL case, which are shown in the magnified version, are due to the proliferation of histiocytes and differ from those in CHL.

Fig. 9 shows the visualization results of AITL cases where abnormal cells were not found in both FCM data. In both cases, the FCM data show all zeros, which cannot be discriminated from CHL and RL. In other words, if superclass classification is based solely on FCM data, the outputs of sub-network 3 are expected to be emphasized in the entire tissue. However, by inputting images at the same time, the proposed method can correctly classify these cases as AITL by focusing on the outputs of sub-network 2.

From the above, we confirmed that each network of the proposed classification model is appropriately trained and can present a basis for decision making that is consistent with the pathologist's diagnostic process.

5. Conclusion

In this paper, we proposed a multimodal classification method for lymphoma pathology diagnosis, incorporating both images and auxiliary

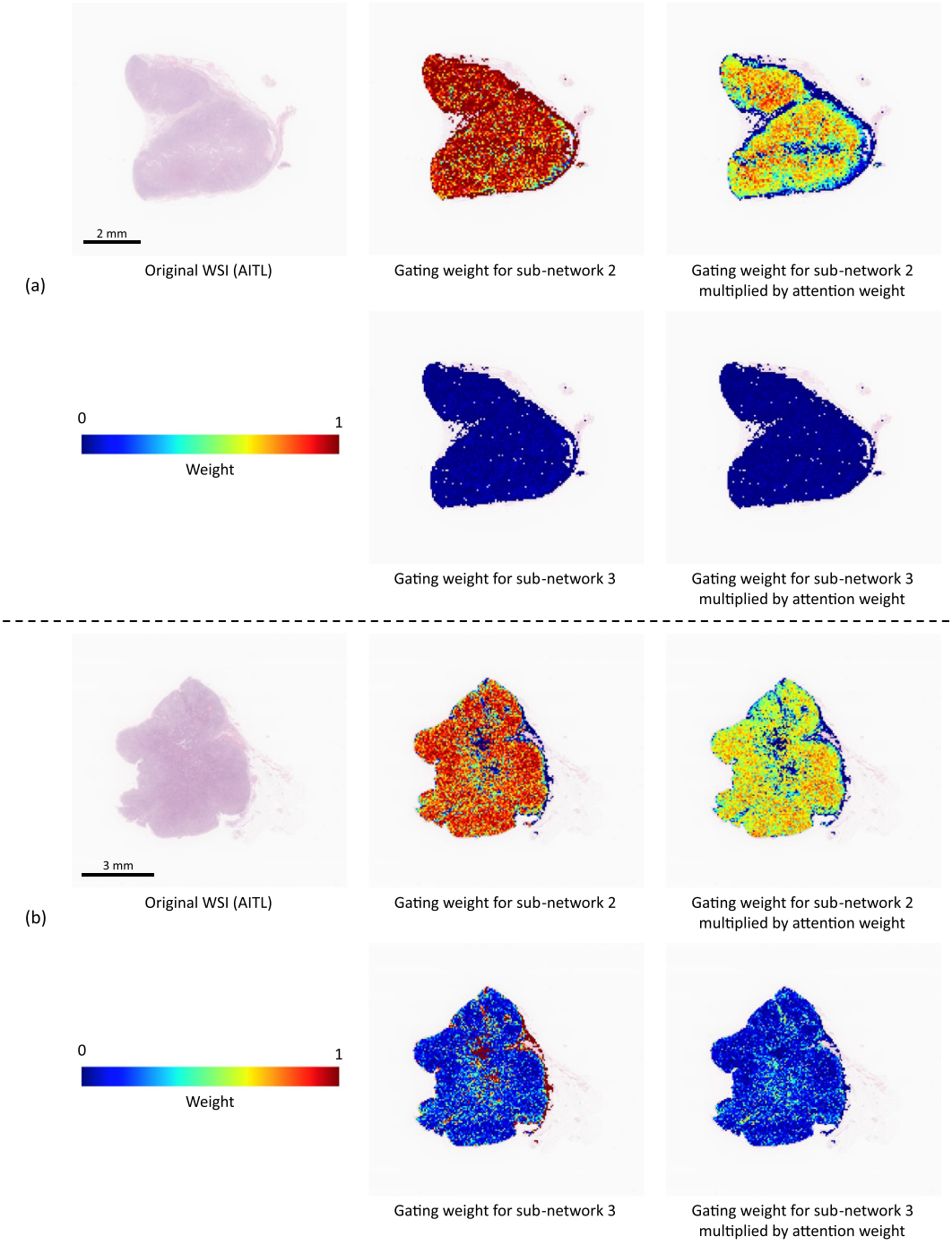

Fig. 9. The visualization results of gating and attention weights: both AITL cases have no abnormal cell populations in their FCM data.

information. By integrating MoE into the MIL framework, our proposed method constructs an effective classification model for problem settings characterized by hierarchical class structures, encompassing superclasses and classes. Notably, the gating network in the MoE structure allows the input of both FCM (useful for superclass classification) and image features simultaneously, enabling robust weight outputs from sub-networks to address the limitations of FCM power. To assess the effectiveness of our method, we conducted experiments involving a six-class classification task with 600 lymphoma cases. The proposed method achieved a classification accuracy of 72.3%, surpassing the 69.5% obtained through the straightforward combination of FCM and images, as well as the 70.2% achieved by the method using only images. Comparative methods suitable for borrowing from existing literature were unavailable, as our model needed to clearly define the role of inputs, mirroring the actual diagnostic process. This involved using FCM exclusively for superclass classification, with the final classification based on image data. Furthermore, the combination of multiple weights in the MoE and MIL allows for the visualization of specific cellular and tumor regions, resulting in a highly explanatory model that conventional methods cannot achieve. Validation from an expert hematopathologist further confirmed the visualization results in multiple classes.

The proposed method can offer accurate and highly explanatory diagnostic assistance at pathology facilities where WSI and FCM are available. However, there are several challenges to address for clinical application. In this study, as an initial step, the number of target classes was set to six, whereas the total number of lymphoma subtypes is approximately 100.

For diagnostic support in clinical practice, it is imperative to increase the number of classes covered by the classification, aligning with actual diagnoses (e.g. mantle cell lymphoma and MALT lymphoma). As the number of classes increases, a class imbalance issue may arise in clinical settings, where some classes may have fewer cases than others. We can also apply the method for addressing class imbalance problems[46,47] to challenges in digital pathology.

In this paper, external validation could not be conducted as there was no available dataset with curated WSI and FCM data. It is well-known that the variation in tissue processing protocols can adversely impact the results of machine learning algorithms. Currently, several methods exist to address such issues.[48,49] Notably, domain adversarial learning has demonstrated effectiveness in the classification of lymphoma for various tissue slides obtained from multiple institutions.[14]

Our proposed method is not confined to lymphoma diagnosis but can be applied to similar problem settings with hierarchical class structures. Additionally, as auxiliary data is not limited to FCM, the proposed method can be employed whenever there is valuable information for superclass classification in patient data, such as medical records.

Declaration of interests

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.

Acknowledgements

This work was partially supported by MEXT KAKENHI (20H00601), JST CREST (JPMJCR21D3, JPMJCR22N2) including AIP challenge program, JST Moonshot R&D (JPMJMS2033-05), JST AIP Acceleration Research (JPMJCR21U2), NEDO (JPNP18002, JPNP20006) and RIKEN Center for Advanced Intelligence Project.

References

1. Gabriele Campanella. Hanna Matthew G, Geneslaw Luke, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images Nature medicine 2019;25:1301–1309.
2. Tokunaga Hiroki, Teramoto Yuki, Yoshizawa Akihiko, Bise Ryoma. Adaptive Weighting Multi-Field-of-View CNN for Semantic Segmentation in Pathology in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 12597–12606. 2019.
3. Tanizaki Kosuke, Hashimoto Noriaki, Inatsu Yu, Hontani Hidekata, Takeuchi Ichiro. Computing valid p-values for image segmentation by selective inference in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 9553–9562. 2020.
4. Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks Medical Image Analysis 2020;65, 101789.
5. Chen Richard J, Chen Chengkuan, Li Yicong, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 16144–16155. 2022.
6. Roger Schaer, Sebastian Otálora, Oscar Toro, Manfredo Atzori, Henning Müller. Deep learning-based retrieval system for gigapixel histopathology cases and the open access literature. Journal of Pathology Informatics 2019:10.
7. Hegde Narayan, Hipp Jason D, Liu Yun, et al. Similar image search for histopathology: SMILY *NPJ digital medicine*. 2019;2:1–9.
8. Shivam Kalra, Tizhoosh HR. Choi Charles, et al. Yottixel–an image search engine for large archives of histopathology whole slide images Medical Image Analysis 2020;101757.
9. Chen Richard J, Lu Ming Y, Williamson Drew FK, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning Cancer Cell 2022;40:865–878.
10. Hölscher David L, Bouteldja Nassim, Joodaki Mehdi, et al. Next-Generation Morphometry for pathomics-data mining in histopathology *Nature Communications*. 2023;14:470.
11. Li Dongguang, Bledsoe Jacob R, Zeng Yu, et al. A deep learning diagnostic platform for diffuse large B-cell lymphoma with high accuracy across multiple hospitals *Nature communications*. 2020;11:6004.
12. Damir Vrabac, Akshay Smit, Rebecca Rojansky, et al. DLBCL-Morph: Morphological features computed using deep learning for an annotated digital DLBCL image set Scientific Data 2021;8:135.
13. Hiroaki Miyoshi, Kensaku Sato, Yoshinori Kabeya, et al. Deep learning shows the capability of high-level computer-aided diagnosis in malignant lymphoma Laboratory Investigation 2020:1-11.
14. Hashimoto Noriaki, Fukushima Daisuke, Koga Ryoichi, et al. 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*: 3852–3861. 2020.
15. Li Bin, Li Yin, Eliceiri Kevin W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*: 14318–14328. 2021.
16. Zhang Hongrun, Meng Yanda, Zhao Yitian, et al. 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* : 18802–18812. 2022.
17. Jordan Yap, William Yolland, Philipp Tschandl. Multimodal skin lesion classification using deep learning Experimental dermatology 2018;27:1261–1267.
18. Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, Regina Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction Radiology 2019;292:60–66.
19. Yang Jialiang Ju, Jie Guo Lei, et al. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning Computational and structural biotechnology journal 2022;20:333–342.
20. Chen Richard J, Lu Ming Y, Weng Wei-Hung, et al. Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images in *Proceedings of the IEEE/CVF International Conference on Computer Vision*: 4015–4025. 2021.
21. Yusuke Takagi, Noriaki Hashimoto, Hiroki Masuda, et al. Transformer-based personalized attention mechanism for medical images with clinical records. Journal of Pathology Informatics 2023, 100185.
22. Ashish Vaswani, Noam Shazeer. Parmar Niki, et al. Attention is all you need Advances in neural information processing systems 2017;30.
23. Jacobs Robert A, Jordan Michael I, Nowlan Steven J, Hinton Geoffrey E. Adaptive mixtures of local experts Neural computation 1991;3:79–87.
24. Ilse Maximilian, Tomczak Jakub, Welling Max. Attention-based deep multiple instance learning in *International conference on machine learning*: 2127–2136PMLR 2018.
25. Li Hang, Yang Fan, Xing Xiaohan, 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: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*: 529–539Springer 2021.
26. Lerousseau Marvin, Vakalopoulou Maria, Classe Marion, et al. Weakly supervised multiple instance learning histopathological tumor segmentation in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*: 470–479Springer 2020.
27. Wetstein Suzanne C, Jong Vincent MT, Stathonikos Nikolas, et al. Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images *Scientific reports*. 2022;12:15102.
28. Seyed Mousavi Hojjat, Vishal Monga, Ganesh Rao, Rao Arvind UK. Automated discrimination of lower and higher grade gliomas based on histopathological image analysis. Journal of pathology informatics 2015:6.
29. Hou Le, Samaras Dimitris, Kurc Tahsin M, Gao Yi, Davis James E, Saltz Joel H. Patch-based convolutional neural network for whole slide tissue image classification in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 2424–2433. 2016.
30. Xu Yan, Jia Zhipeng, Ai Yuqing, et al. Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*: 947–951IEEE 2015.
31. Gao Yi, Liu William, Arjun Shipra, et al. Multi-scale learning based segmentation of glands in digital colonrectal pathology images in *Medical Imaging 2016: Digital Pathology*;9791:97910MInternational Society for Optics and Photonics 2016.
32. Cireşan Dan C, Giusti Alessandro, Gambardella Luca M, Schmidhuber Jürgen. Mitosis detection in breast cancer histology images with deep neural networks in *International Conference on Medical Image Computing and Computer-assisted Intervention*:411–418Springer 2013.
33. Cruz-Roa Angel, Basavanhally Ajay, González Fabio, et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks in *Medical Imaging 2014: Digital Pathology*;9041:904103International Society for Optics and Photonics 2014.
34. Bejnordi Babak Ehteshami. Veta Mitko, Van Diest Paul Johannes, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer Jama 2017;318:2199–2210.
35. Zhu Xinliang, Yao Jiawen, Zhu Feiyun, Huang Junzhou. Wsisa: Making survival prediction from whole slide histopathological images in *Proceedings of the IEEE conference on computer vision and pattern recognition*: 7234–7242. 2017.
36. Noriaki Hashimoto, Yusuke Takagi, Hiroki Masuda, et al. Case-based similar image retrieval for weakly annotated large histopathological images of malignant lymphoma using deep metric learning Medical Image Analysis 2023;85, 102752.
37. Michael Brown, Carl Wittwer. Flow cytometry: principles and clinical applications in hematology Clinical chemistry 2000;46:1221–1229.
38. Aysun Adan, Günel Alizada, Yağmur Kiraz, Yusuf Baran, Ayten Nalbant. Flow cytometry: basic principles and applications Critical reviews in biotechnology 2017;37:163–176.
39. Valentina Gaidano, Valerio Tenace, Nathalie Santoro, et al. A clinically applicable approach to the classification of B-cell non-Hodgkin lymphomas with flow cytometry and machine learning Cancers 2020;12:1684.
40. Sebastian Böttcher, Robby Engelmann, Georgiana Grigore, et al. Expert-independent classification of mature B-cell neoplasms using standardized flow cytometry: a multicentric study Blood advances 2022;6:976–992.
41. Shazeer Noam, Mirhoseini Azalia, Maziarz Krzysztof, et al. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer in *International Conference on Learning Representations* 2017.

42. Gross Sam, Ranzato Marc'Aurelio, Szlam Arthur. Hard mixtures of experts for large scale weakly supervised vision in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 6865–6873. 2017.
43. Sahasrabudhe Mihir, Sujobert Pierre, Zacharaki Evangelia I, et al. Deep multi-instance learning using multi-modal data for diagnosis of lymphocytosis *IEEE Journal of Biomedical and Health Informatics*. 2020;25:2125–2136.
44. Hinton Geoffrey, Vinyals Oriol, Dean Jeff. Distilling the knowledge in a neural network *arXiv preprint arXiv:1503.02531*. 2015.
45. He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition in *Proceedings of the IEEE conference on computer vision and pattern recognition*: 770–778 2016.
46. Lin Tsung-Yi, Goyal Priya, Girshick Ross, He Kaiming, Dollár Piotr. Focal loss for dense object detection in *Proceedings of the IEEE international conference on computer vision*: 2980–2988. 2017.
47. Cui Yin, Jia Menglin, Lin Tsung-Yi, Song Yang, Belongie Serge. Class-balanced loss based on effective number of samples in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*: 9268–9277. 2019.
48. Ganin Yaroslav, Ustinova Evgeniya, Ajakan Hana, et al. Domain-adversarial training of neural networks *The journal of machine learning research*. 2016;17:2096–2030.
49. David Tellez, Geert Litjens, Péter Bándi, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology Medical image analysis 2019;58, 101544.