

Single-Stage Broad Multi-Instance Multi-Label Learning (BMIML) with Diverse Inter-Correlations and its Application to Medical Image Classification

Qi Lai, Jianhang Zhou, Yanfen Gan, Chi-Man Vong, *Senior Member, IEEE*,
and C.L. Philip Chen, *Fellow, IEEE*

Abstract—In many real-world applications, one object (e.g., image) can be represented or described by multiple instances (e.g., image patches) and simultaneously associated with multiple labels. Such applications can be formulated as *multi-instance multi-label learning* (MIML) problems and have been extensively studied during the past few years. Existing MIML methods are useful in many applications but most of which suffer from relatively low accuracy and training efficiency due to several issues: i) *the inter-label correlations* (i.e., the probabilistic correlations between the multiple labels corresponding to an object) are neglected; ii) *the inter-instance correlations* (i.e., the probabilistic correlations of different instances in predicting the object label) cannot be learned *directly* (or jointly) with other types of correlations due to the missing instance labels; iii) diverse inter-correlations (e.g., inter-label correlations, inter-instance correlations) can only be learned in multiple stages. To resolve these issues, a new single-stage framework called *broad multi-instance multi-label learning* (BMIML) is proposed. In BMIML, there are three innovative modules: i) an *auto-weighted label enhancement learning* (AWLEL) based on broad learning system (BLS) is designed, which simultaneously and efficiently captures the inter-label correlations while traditional BLS cannot; ii) A specific MIML neural network called *scalable multi-instance probabilistic regression* (SMIPR) is constructed to effectively estimate the inter-instance correlations using the object label only, which can provide additional probabilistic information for learning; iii) Finally, an *interactive decision optimization* (IDO) is designed to combine and optimize the results from AWLEL and SMIPR and form a single-stage framework. As a result, BMIML can achieve simultaneous learning of diverse inter-correlations between whole images, instances, and labels in single stage for higher classification accuracy and much faster training time. In this work, medical image classifications is employed as an illustration. Experiments show that BMIML is highly competitive to (or even better than) existing methods in accuracy and much faster than most MIML methods even for large medical image data sets (> 90K images).

Index Terms—Multi-instance learning, multi-label learning, simultaneous learning, medical image classification, single-stage framework.

I. INTRODUCTION

IN traditional supervised learning, one object is only represented by a single instance and associated with a single label. However, in many real-world applications, one object can be naturally described by a collection of instances (called

a *bag*) and has multiple class labels simultaneously. Such applications can be formulated as *multi-instance multi-label learning* (MIML) problem [1] and have been extensively applied in many fields such as image classification [2], [3], video annotation [4], [5], biomedicine [6] and protein function prediction [7]. Out of many MIML applications, medical image classification is one of the most popular research areas for its practical use. Nowadays, with higher pressure on public health and a shortage of professionals on different types of medical imaging [8], it is necessary to further investigate general, effective, and efficient automated methods for clinical use. In recent works [9]–[11], extensive applications have been proposed to explore a possible way of automated disease classification. Literature [12]–[14] show that the MIML-based methods have great potential in automated disease classification and clinical diagnosis in the medical field. These indicate the feasibility of a MIML-based automated approach for disease classification. Therefore, medical image classification is employed as an illustrative application in this work.

Medical image classification task is typically formulated as a multi-class or *multi-label learning* (MLL) problem. Strictly speaking, the medical image is usually multi-labeled, and for each image, the distribution of different labels is often imbalanced. As shown in Fig. 1, *Label 1* is the dominant position and is accurately predicted while *Label 3* is almost ignored since *Label 3* only occupies a small part of the images. For this reason, the easily recognized labels usually result in a dominant position, which always leads to relatively poor performance. MIML has been applied to deal with the above problem, which offers a way for understanding the correlations between the input images and the output labels. In MIML setting, an image can be divided into several segments or patches (i.e., *instances*) so that the multi-label classification tasks can be performed at the instance-level as shown in Fig. 1. Meanwhile, a collection of instances is called a *bag* which can represent an image (training sample) and the bag is assigned with a set of multiple class labels (i.e., *label set*).

Practically, clinicians consider diverse correlations in medical image classification as illustrated in Fig. 1, where the solid lines indicate the correlations between bags (*global view*), instances (*local view*), and labels while the dash lines indicate the inter-correlations (partially ignored in existing works but practically all are required) between *bags-bags*, *instances-instances*, *labels-labels*. In other words, all correlations are practically used to estimate the correlation between bags and

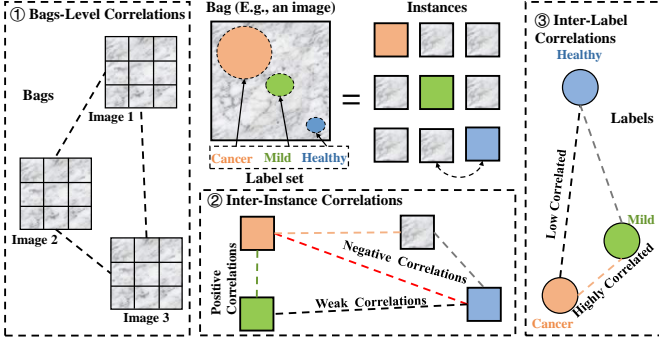


Fig. 1. The details of diverse correlations between bags (*global view*), instances (*local view*), and the multiple labels. Solid lines indicate the correlations between bags, instances, and labels. Dash lines indicate the inter-correlations (partially ignored in existing works but practically all are required) between bags-bags, instances-instances, labels-labels.

labels to achieve the best possible classification performance [15]. For example, to identify if an image is relevant to suspicion lesions (i.e., the correlation between bags and labels) or not, the following information should be considered simultaneously:

- 1) the bag-level correlations that reveal the difference of medical images of distinct diseases from the perspective of the whole image (*global view*);
- 2) the inter-instance correlations that reveal which parts of an image are significant to distinguish from different diseases (*local view*);
- 3) the inter-label correlations that quantitatively indicate the margin between two diseases with the class probabilities (or confidence) of the intraclass samples.

Therefore, for more effective medical image classification, the diverse correlations which were partially *neglected* in existing methods should be simultaneously considered [16].

However, how to make use of inter-instance correlations [17], [18] (i.e., the probabilistic correlations of different instances in predicting the bag labels) remains a challenging research topic because, in almost all available data sets, only image/bag-level (*global view*) labels are available while instance-level (*local view*) labels are missing due to the heavy burden in manual labeling for the clinicians. For this reason, traditional supervised learning methods are *unable* to learn the inter-instance correlations *directly*.

Although existing MIML methods can learn the inter-instance correlations *indirectly* through multiple independent learning procedures, this indirect multi-stage way will affect the model performance in accuracy and efficiency. Moreover, considering diverse correlations will bring time-consuming which is another challenge for existing MIML methods [19], especially in large data sets. Thus, it is necessary to design a unified single-stage interactive framework that can learn the information of whole images/bags (*global view*) and instances (*local view*) simultaneously and improve efficiency. However, existing MIML methods do not provide the way of simultaneous learning so that it becomes highly nontrivial to implement over MIML methods. To our best knowledge, there is no such simultaneous learning mechanism of diverse correlations for

MIML in existing works as summarized in Table I.

Recently, efficient discriminative learning called *Broad Learning System* (BLS) [20] was proposed. The main advantage of BLS is its efficient network training under random feature mapping with the ability to jointly learning of multiple sub-networks. In BLS, the original inputs are transferred as the mapping features and placed in the *feature layer* (a sub-network), and the structure is extended to the *enhancement layer* (another sub-network) in a broad sense. Both the feature and enhancement layers are then connected to the output layers. Thus, BLS offers the necessary mechanism of simultaneous/joint learning efficiently.

Although BLS can deal with MLL tasks (e.g., one sample corresponding to several labels), it does not consider the inter-label correlations [21], [22] which must be considered in MLL. Moreover, BLS requires that all inputs are independent of each other and simply sets the entire data matrix X as input [20]. However, medical image classification is always a MIML problem in which the instances of the input samples are highly relevant, so it is impossible to assume all inputs independently. Therefore, the application of BLS in medical image classification becomes nontrivial and challenging.

In this paper, a novel approach for medical image classification called *Broad Multi-Instance Multi-Label network* (BMIML) is proposed. Concretely, the BMIML is based on BLS which can jointly learn multiple sub-networks in a broad sense so that the diverse correlations between bags, instances, and labels can be simultaneously captured. However, standard BLS cannot capture the inter-label correlation which is necessary for handling MIML problems. Also, it cannot utilize the inter-instance correlations for training *directly*. Thus, in BMIML, an interactive framework is newly designed that includes three novel modules: i) *auto-weighted label enhancement learning* (AWLEL), ii) *scalable multi-instance probabilistic regression* (SMIPR), and iii) an *interactive decision optimization* (IDO). On the one hand, AWLEL as part of MLL in BMIML can model diverse correlations, including the inter-label correlation which can improve the accuracy of BMIML. On the other hand, SMIPR as part of multi-instance learning in BMIML is a way to model the inter-instance correlations using bag labels only. Finally, IDO works as a bridge to connect AWLEL and SMIPR to integrate their results into a network, forming an interactive single-stage framework that can deal with MIML problems efficiently and effectively.

Hence BMIML overcomes the weaknesses of the BLS and existing MIML framework by simultaneously learning the diverse (inter-)correlations. The illustrations of the diverse (inter-)correlations mentioned above are shown in Fig. 1 and Table I. The main contributions of BMIML are summarized below:

- 1) Through our proposed method BMIML, the diverse correlations between bags, instances, and multiple labels can be considered simultaneously for higher classification accuracy. This simultaneous consideration of diverse correlations cannot be done in existing MIML works as illustrated in Table I.
- 2) In BMIML, an interactive single-stage learning framework is newly designed which can simultaneously con-

TABLE I
DIVERSE INTER-CORRELATIONS EXPLOITED BY PROPOSED BMIML AND MIML METHODS

Approaches	Diverse Inter-correlations					
	Bag-Bag	Inter-instances	Inter-labels	Bag-Instance	Bag-Label	Instance-Label
MIMLNN [23]		✓		✓	✓	
MIMLSVM [1]	✓			✓	✓	
MIMLmiSVM [23]	✓			✓	✓	
MIMLkNN [24]	✓			✓	✓	✓
MIMLBoost [1]	✓			✓	✓	
MIMLfast [19]			✓	✓	✓	✓
DeepMIML [25]	✓	✓		✓	✓	✓
Proposed BMIML	✓	✓	✓	✓	✓	✓

sider the correlations of both global views (whole images/bags) and local view (image patches/instances) for image classification tasks. This single-stage framework can further improve classification accuracy, learning efficiency, and human burden, especially on large data sets. This is a non-trivial challenging task because local view labels are always missing in the training data set.

The organization of this article is as below. Section II provides a brief review of BLS and MIML. Section III details our proposed methods: BMIML, including AWLEL, SMIPR, and IDO. Section IV demonstrates the experimental results with analysis and discussion. At last, a conclusion is drawn in Section V.

II. PRELIMINARIES

A. Multi-instance Multi-label Learning (MIML)

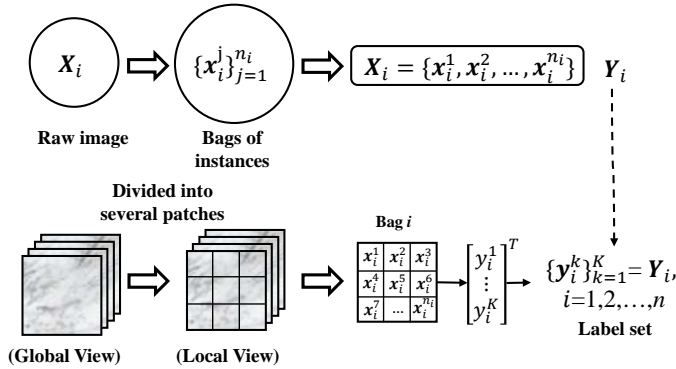


Fig. 2. A brief introduction of the MIML setting. X_i represents the i -th image in the dataset and then it was divided into several patches called instances x_i^j , where $j = 1, 2, \dots, n_i$, n_i is the number of the instances in the i -th image. Y_i is the label set associated with X_i . K indicates the number of categories.

Clinically, a medical image can be described by multiple semantic labels, as shown in Fig. 1. However, these labels are only closely related to their respective *regions/patches* (called *instances*) rather than the entire image [26], as illustrated in Fig. 1. For this reason, a more rational and natural strategy is to model medical image classification as a *multi-instance multi-label learning* (MIML) tasks [23]. As illustrated in Fig.2, given a training set $\{(X_i, Y_i)\}_{i=1}^n$ where $X_i = \{x_i^1, x_i^2, \dots, x_i^{n_i}\}$ ($i = 1, 2, \dots, n$) represents a bag of instances (image patches) x_i^j divided from the i -th original image X_i , Y_i is a K -dimensional label vector $[y_i^1, y_i^2, \dots, y_i^K]$

or X_i and $y_i^k \in \{-1, 1\}$, $k = 1, 2, \dots, K$ entry y_i^k indicates the membership corresponding to X_i with the k th class label. Unfortunately, as shown in Fig.2, the relation between Y_i and each instance x_i^j is not explicitly indicated in the training set, which is exactly our training target. Therefore, we introduce a probabilistic regression framework to construct the probabilistic correlations between instances x_i^j and bag label Y_i . Based on the training set, the MIML probabilistic regression aims to approximate a function that can predict the class probability (or confidence) of testing set as accurately as possible.

B. Broad Learning System (BLS)

BLS is simply introduced here and the readers can refer [20] for details. Given the training set $\{X, Y\} \in \mathbb{R}^{N \times (D+K)}$ where $X = [X_i] \in \mathbb{R}^{N \times D}$ is the input matrix where X_i denotes the i -th sample with the relevant output Y_i and $Y = [Y_i] \in \mathbb{R}^{N \times K}$ is the output matrix. D is the dimension of input vector X_i and K is the number of class labels. Then the input matrix X is mapped into a series of random features Z_{m_1} , $m_1 = 1$ to M_1 . Each feature mapping node Z_{m_1} can be represented as:

$$Z_{m_1} = \xi_{m_1}^z(Xw_{m_1}^z + \beta_{m_1}^z) \quad (1)$$

where m_1 is a user-specified parameter and $\xi_{m_1}^z$ is an activation function (e.g., sigmoid). $bmw_{m_1}^z$ and $\beta_{m_1}^z$ are the randomly generated weights and bias matrices with the proper dimensions for input X , respectively. Similarly, the enhancement nodes H_{m_2} , $m_2 = 1$ to M_2 are denoted by:

$$H_{m_2} = \xi_{m_2}^h(Z_{m_1}w_{m_2}^h + \beta_{m_2}^h) \quad (2)$$

where $\xi_{m_2}^h$ is a non-linear function (e.g., $\tanh(\cdot)$) which can be selected differently in building a model as well as $\xi_{m_1}^z$ and m_2 is a user-specified parameter. Here, the number of mapping nodes Z_{m_1} and enhancement nodes H_{m_2} can be same or different. It is set according to the actual situation and will not be described here. $w_{m_2}^h$ and $\beta_{m_2}^h$ are respectively random weights and bias matrices for the mapped features Z_{m_1} . Hence, the output nodes Y of BLS can be denoted by:

$$Y = [Z_1, Z_2, \dots, Z_{M_1}, H_1, H_2, \dots, H_{M_2}]W \quad (3)$$

where the weights W are connecting the layer of features nodes and the layer of enhancement nodes to the output nodes,

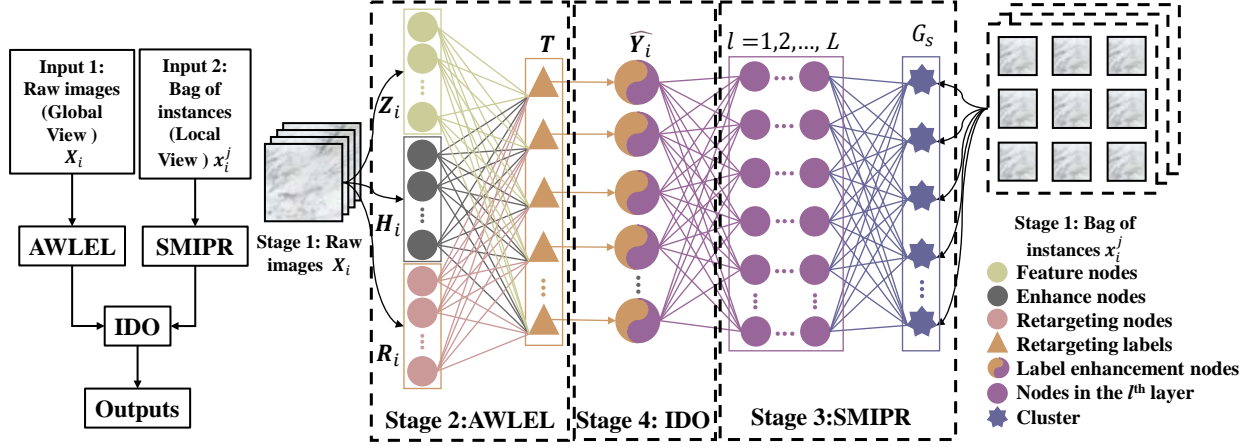


Fig. 3. Structure of the proposed BMIML.

and $W = A^+Y$, which can be easily computed using ridge regression approximation of pseudoinverse as follows:

$$\arg \min_W \|AW - Y\|_2^2 + \lambda \|W\|_2^2 \quad (4)$$

$$A^+ = \lim_{\lambda \rightarrow 0} (\lambda I + AA^T)^{-1} A^T \quad (5)$$

where $A = [Z_1, Z_2, \dots, Z_{M_1}, H_1, H_2, \dots, H_{M_2}]$. The value λ indicates the further constraints on the squared weights W . Consequently, we have

$$W = (\lambda I + AA^T)^{-1} A^T Y \quad (6)$$

Since BLS simply takes the entire data matrix X as input [20] i.e., all inputs are assumed independent of each other, and it cannot capture correlations between multiple labels. Therefore, BLS cannot directly employ MIML tasks. In our tasks, an improved BLS framework is designed by i) adding a retargeting layer enables BLS to capture the inter-label correlations; and ii) simultaneously modeling the diverse correlations between the bags, instances, and labels (see Table I).

III. PROPOSED BMIML

BLS is good at joint learning of different information and therefore suitable for learning diverse correlations simultaneously. Although BLS has demonstrated its strong classification ability in many fields [?], [?], [27], it does not work very well for semantically complex images (e.g., multi-label images) since it cannot consider the inter-label correlations and the property of weakly discriminative features in the image. In this section, aiming at improving the performance for medical image classification, an single-stage interactive framework is newly designed based on i) *auto-weighted label enhancement learning* (AWLEL) to process MLL in MIML, i.e., handling diverse correlations, and reformulating the original single-label space into an enhanced retargeted multi-label space by considering intra-class and inter-class scatters for better discrimination under weak features (as shown in Fig. 4); ii) a novel *scalable multi-instance probabilistic regression*

(SMIPR) to provide multi-instance probabilistic predictions by fully utilizing the inter-instance correlations (as shown in Fig. 5); and iii) using an *interactive decision optimization* (IDO) to combine the AWLEL and SMIPR, forming an end-to-end framework to deal with MIML tasks. The entire process of the proposed BMIML is summarized in Fig. 3, which has the following four computational stages.

A. Overview of BMIML Stages

Stage 1 (Preprocessing): The training data set includes original images X_i (global view) and instances x_i^j (local view, simply dividing from original images, detailed in Section IV-B), which are the inputs to the AWLEL and SMIPR, respectively.

Stage 2 (Auto-Weighted Label Enhancement Learning): AWLEL is designed based on the BLS, as shown in Fig. 3, stage 2. Different from the standard BLS, a new *retargeting layer* R_i is added in the BLS that aims to improve the issue of MLL tasks and can automatically generate the retargeted labels for instances from bag-level labels. In addition, it can guarantee to impose the constraint of large margin of classification boundary for the requirement of correct classification for each data point. The learning details for the proposed AWLEL module is described in Section III-B, and its optimization strategy is detailed in Section III-E.

Stage 3 (Scalable Multi-Instance Probabilistic Regression): SMIPR is a neural network specifically designed for MIML which performs probabilistic regression on each instance according to the retargeted labels T_i only generated by AWLEL. In other words, SMIPR can estimate the inter-instance correlations by using the images/bags labels *only*. Different from traditional neural network structure, the first layer of SMIPR is a clustering process to generate S disjoint groups of bags G_1, G_2, \dots, G_S , and calculate the corresponding medoids v_p of the clusters G_p , $p = 1$ to S . Since clustering helps uncover the underlying structure of the training data set, the medoid of each cluster may make full use of the instance information and encodes some distribution information of different bags. The detail is discussed in Section III-C.

Stage 4 (Interactive Decision Optimization): In almost all data sets, there are only bag-level (*global view*) labels while the instance-level (*local view*) labels are missing. Therefore, an *interactive decision optimization* (IDO) is designed as a bridge to connect the above two modules: AWLEL and SMIPR. In other words, IDO integrates the results of i) simultaneous learning of diverse correlations and ii) direct learning the instance class membership probability in a single network, which can achieve an end-to-end learning. The detailed is provided by Section III-D.

In summary, we aim to construct BMIML for the effective and efficient classification of medical images. For this purpose, an interactive end-to-end learning framework is designed, as shown in Figure 3. First, the AWLEL captures the inter-label correlation, which helps to enlarge the target gaps between the interclass samples. Then, SMIPR was employed to learn the inter-instance correlation according to the inter-label correlation so that it can better capture the local view information. Finally, the AWLEL and SMIPR are connected under IDO and therefore a single-stage multi-instance multi-label learning framework can be achieved. Also, for this reason, IDO cannot work independently for the multi-label image classification task.

B. Auto-Weighted Label Enhancement Learning (AWLEL)

In standard BLS, all inputs are assumed independent of each other and the entire data matrix \mathbf{X} is taken as input. Besides, the output matrix \mathbf{Y} in standard BLS is a strict zero-one matrix, i.e., only the label entry of each row is one, where $label \in \{1, 2, \dots, K\}$ is class label of sample \mathbf{X}_i , as shown in Fig. 4(a). Practically, a medical image is always associated with multiple labels and the distribution of labels is imbalanced. Strict zero-one indicator do not make sense and may be detrimental to classification. Moreover, a series of instances (divided from an image) in a bag are often dependent with each other (i.e., inter-instance correlations). Hence, there is also a probabilistic correlation between multiple labels (i.e., inter-label correlations) associated with a bag. To tackle this issue, another sub-network (called *retargeting nodes*) is added into the standard BLS which can enable BLS for multi-label tasks and capture the inter-label correlations. In our work, the retargeting nodes is defined as

$$\mathbf{R}_i = \xi_i^r (\mathbf{X}_i \mathbf{w}_{m_1}^z + \mathbf{Z}_{m_1} \mathbf{w}_{m_2}^h + \beta_i^r) \quad (7)$$

where ξ_i^r is a non-linear function (e.g., *Tribas*) and β_i^r is a regularization parameter controlling the degree of bias. The weights $\mathbf{w}_{m_1}^z$ and $\mathbf{w}_{m_2}^h$ can be generated from Eqs. (1) and (2), respectively. And then we define the *retargeted labels* of the i th training sample as below:

$$\mathbf{T}_i = (\mathbf{X}_i, \mathbf{Z}_{m_1}, \mathbf{H}_{m_2}, \mathbf{R}_i) \mathbf{w}_i^t \quad (8)$$

where the weight \mathbf{w}_i^t is jointly optimized by feature nodes, enhance nodes, and retargeting nodes. The simultaneously learning of random mapping and regression target of BLS is as follows:

$$\arg \min_{\mathbf{w}^t, \mathbf{T}} \|\mathbf{A} \mathbf{w}^t - \mathbf{T}\|_2^2 + \lambda \|\mathbf{w}^t\|_2^2 \quad (9)$$

where $\mathbf{T} \in \mathbb{R}^{N \times K}$ is the *retargeted labels* and consists of \mathbf{T}_i which can reflect the classification separability (see Fig.3 Stage 2) of each sample (global view) with respect to different class labels. To improve the interclass separability, Eq. (9) is reformulated as:

$$\arg \min_{\mathbf{w}_i^t, \mathbf{T}_i} \sum_{i=1}^N (\gamma_i \|\mathbf{A}_i \mathbf{w}_i^t - \mathbf{T}_i\|_2^2 + \lambda \|\mathbf{w}_i^t\|_2^2 + \theta \omega_i \|\mathbf{T}_i - \mathbf{Y}_i\|_2^2) \quad (10)$$

where the weighted penalty factors γ_i and ω_i control the effect of outliers and the balance between the loss components in the total loss, $\gamma_i = \left(\frac{1}{\|\mathbf{A}_i \mathbf{w}_i^t - \mathbf{T}_i\|_2} \right)$ and $\omega_i = \left(\frac{1}{\|\mathbf{T}_i - \mathbf{Y}_i\|_2} \right)$. \mathbf{A}_i is the i th row vector in the matrix \mathbf{A} , as illustrated in Section II-B. The value ϑ indicates further constraint on the squared difference of retargeted label and ground truth. Using the diagonal matrices $\mathbf{\Gamma} = [\gamma_1, \gamma_2, \dots, \gamma_N]^T$ and $\mathbf{\Omega} = [\omega_1, \omega_2, \dots, \omega_N]^T$ and combining with Eq. (10), we have:

$$\arg \min_{\mathbf{w}^t, \mathbf{T}} \left\| \sqrt{\mathbf{\Gamma}} \mathbf{A} \mathbf{w}^t - \mathbf{T} \right\|_2^2 + \lambda \|\mathbf{w}^t\|_2^2 + \vartheta \left\| \sqrt{\mathbf{\Omega}} (\mathbf{T} - \mathbf{Y}) \right\|_2^2 \quad (11)$$

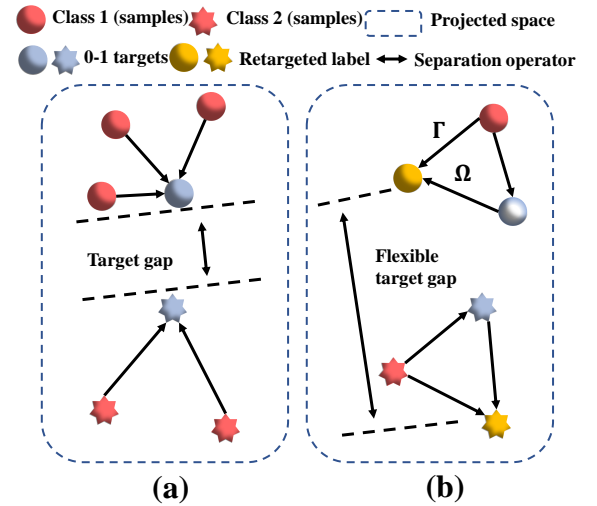


Fig. 4. The difference between the standard BLS and the AWLEL. In (a), all intraclass samples shrink to the fixed 0-1 targets in standard BLS projector space while in (b), AWLEL can auto-weight all intraclass samples to enlarge the gaps between interclass samples.

As shown in Fig.4 (a) and (b), we aim to overcome the limitation of BLS and promote effective separability. Therefore, we expect the samples are drawn from the same class and gather to the corresponding targets. This allows adaptive learning of intraclass targets while enlarging the gaps between interclass targets, resulting to more generalized properties. As formulated in Eq. (10), unlike standard BLS, the retargeted labels (\mathbf{T}_i) can be flexibly balanced between strict zero-one targets (\mathbf{Y}_i) and regression results (\mathbf{w}^t), leading to better classification results. In addition, normal samples can receive

higher weights to increase their contributions, while lower weights are assigned to suspicious outliers to reduce their negative effects [27]. Finally, similar to \mathbf{Y}_i , the retargeted label can reformulate as $\mathbf{T}_i = [t_i^1, t_i^2, \dots, t_i^K]$ where t_i^k ($k = 1, 2, \dots, K$) denotes a class label rather than the real-valued y_i^k . Then we can obtain the retargeted label \mathbf{T} as follows:

$$\mathbf{T} = \begin{bmatrix} t_1^1 & \dots & t_1^K \\ \vdots & \ddots & \vdots \\ t_i^1 & \dots & t_i^K \end{bmatrix} \quad (12)$$

C. Scalable Multi-instance Probabilistic Regression (SMIPR)

The MIML regression task is the natural extension of traditional (*single instance or single label*) regression to the MIML setting. MIML regression models the sample in the same way as MIML classification, with the important difference that each bag is relevant to several *real-valued* outcomes but not categorical classes. However, each instance in the bag makes a (possibly different) contribution to the bag label [28]. For this reason, it becomes necessary to make full use of the probabilistic correlations of different instances in the bag instead of a single score-maximizing instance in predicting the object label. According to the definition about the class-conditional probability density and the prior probability, we can formulate the probability of the joint distribution at the instance-level as below:

$$P(\mathbf{x}_i^j, \hat{y}_i^c) = P(\mathbf{x}_i^j) P(\hat{y}_i^c | \mathbf{x}_i^j) \quad (13)$$

where \mathbf{x}_i^j indicates the j th instance in the i th bag while \hat{y}_i^c indicates the c th class probability of the i th bag and $\hat{y}_i^c \in \hat{\mathbf{Y}}_i = [\hat{y}_i^1, \hat{y}_i^2, \dots, \hat{y}_i^K]$. Note that the index c stands for the most probable class, and K equals to the number of the correct classes (ground truth). According to the MIML property, the bag includes a series of instances corresponding to K possible classes and therefore we have

$$P(\hat{\mathbf{Y}}_i | \mathbf{X}_i) = \prod_{c=1}^K P(\hat{y}_i^c | \mathbf{x}_i^j) \quad (14)$$

where \mathbf{X}_i indicates the i th bag while $\hat{\mathbf{Y}}_i = [\hat{y}_i^1, \hat{y}_i^2, \dots, \hat{y}_i^K]^T$ indicates the K -dimensional output vector. Since Eq. (14) is computationally intractable, a specifically designed MIML probabilistic regression function g is designed to solve Eq. (14). The function g of an input bag \mathbf{X}_i on each of the output vector of the possible label \mathbf{Y}_i is illustrated in Fig. 5. Inspired by minimum squared error criteria [29], no matter whether there is any interdependence between the values of $g(\mathbf{X}_i, \mathbf{Y}_i)$ for different values of c , the squared error attains its absolute minimum if the probabilistic regression function $g(\mathbf{X}_i, \mathbf{Y}_i)$ is identical to the class probability $P(\hat{\mathbf{Y}}_i | \mathbf{X}_i)$:

$$g(\mathbf{X}_i, \mathbf{Y}_i) = P(\hat{\mathbf{Y}}_i | \mathbf{X}_i) \quad (15)$$

Fig. 5 shows the scalable multi-instance probabilistic regression (SMIPR) structure employed by BMIML. The regression problem is to determine the \mathbf{W}_{PR} from training

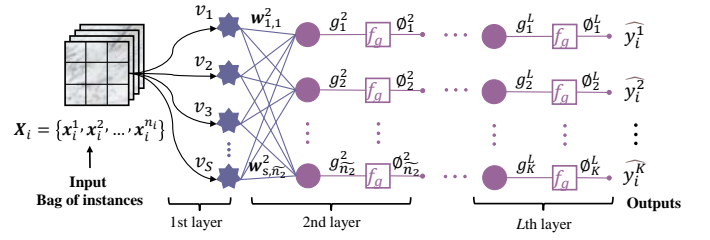


Fig. 5. Scalable multi-instance probabilistic regressions (SMIPR) structure.

set $\{(\mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^n$. We define $\mathbf{W}_{PR} = [\mathbf{w}_{p,q}^l]$, and $\mathbf{w}_{p,q}^l$ indicates the weight connecting the p th node in $(l-1)$ th layer and the q th node in l th layer. For the probabilistic regression structure, by regarding each bag as an individual object, the training set $\{\mathbf{X}_i\}_{i=1}^n$ is clustered in the first layer ($l=1$) into S disjoint groups of bags $\{G_p\}_{p=1}^S$ ($G_{s1} \cap_{s1 \neq s2} G_{s2} = \emptyset$) with $\bigcup_{p=1}^S G_p = \{\mathbf{X}_i\}_{i=1}^n$ by *k-means algorithm* [30]. After the clustering process, the training set is divided into S partitions and their medoids v_p are decided as:

$$v_p = \arg \min_{\mathbf{A} \in G_p} \sum_{\mathbf{B} \in G_p} \text{dist}(\mathbf{A}, \mathbf{B}) \quad (16)$$

where $\text{dist}(\mathbf{A}, \mathbf{B})$ denotes the *Hausdorff distance* [31] between two bags of instances $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{N1}\}$ and $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{N2}\}$, which can be defined as: $\text{dist}(\mathbf{A}, \mathbf{B}) = \max\{\max_{\mathbf{a} \in \mathbf{A}} \min_{\mathbf{b} \in \mathbf{B}} \|\mathbf{a} - \mathbf{b}\|, \max_{\mathbf{b} \in \mathbf{B}} \min_{\mathbf{a} \in \mathbf{A}} \|\mathbf{b} - \mathbf{a}\|\}$ where $\|\mathbf{a} - \mathbf{b}\|$ measures the distance between instances \mathbf{a} and \mathbf{b} . When the number of the layers is set to 2 (i.e., $l=2$), the numbers of input and output nodes are fixed so that $\mathbf{W}_{PR} = [\mathbf{w}_{p,q}^l]_{S \times K}$ where S is the maximum number of the clusters (input), and K is the maximum number of output classes. For the l th $2 < l < L$ layer, the maximum number of nodes is defined as \tilde{n}_l while the number of output nodes ($l=L$) is still set to K , which is the scalable part of the multi-instance probabilistic regression structure, as shown in Fig. 5. Then the weights $[\mathbf{w}_{p,q}^l]$ can be optimized by minimizing the following sum-of-squares error function:

$$E = \frac{1}{2} \sum_{i=1}^n \sum_{q=1}^K \{g_q^l(\mathbf{X}_i) - t_i^q\}^2 \quad (17)$$

where t_i^q is the desired output values in output layer ($l = L$ and $q = 1$ to K) of \mathbf{X}_i on the q th class with the elements $[t_i^q]_{n \times K = \mathbf{T}}$, and $g_q^l(\mathbf{X}_i)$ is defined as:

$$g_q^l(\mathbf{X}_i) = \begin{cases} \sum_{p=1}^{\tilde{n}_{l-1}} \mathbf{w}_{p,q}^l \phi_p^{l-1}(\mathbf{X}_i) & \text{if } l > 2 \\ \sum_{p=1}^S \mathbf{w}_{p,q}^l \phi_p^l(\mathbf{X}_i) & \text{if } l = 2 \end{cases} \quad (18)$$

where p and q are the numbers of nodes in the $(l-1)$ th and l th layer, respectively. Finally, $\phi_p^l(\mathbf{X}_i)$ can be calculated as below:

$$\phi_p^l(\mathbf{X}_i) = \begin{cases} \sum_{p=1}^{\tilde{n}_{l-1}} \mathbf{w}_{p,q}^l f_g(\phi_p^{l-1}(\mathbf{X}_i)) & \text{if } l > 2 \\ \text{dist}(\mathbf{X}_i, v_p) & \text{if } l = 2 \end{cases} \quad (19)$$

where $f_g(\cdot)$ is an activation function (e.g., *sigmoid*) and the weights $\mathbf{w}_{p,q}^l = \begin{cases} 0, & p = q \\ 1, & p \neq q \end{cases}$ if $l=2$, $\mathbf{w}_{p,q}^l$ ($l > 2$) can be updated using gradient descent:

$$\mathbf{w}_{p,q}^{l+1} = \mathbf{w}_{p,q}^l + \eta(f_g(\phi_p^l(\mathbf{X}_i))\Delta^l) \quad (20)$$

where η is a learning rate and Δ^l is denoted as the gradient of the l th layer obtained in back-propagation:

$$\Delta^l = \begin{cases} \Delta^{l+1}\mathbf{w}^{l+1}\mathbf{F}^l, & \text{if } 2 < l < L \\ (\mathbf{T} - g(\mathbf{X}))^T \mathbf{F}^l & \text{if } l = L \end{cases} \quad (21)$$

$$\text{where } \mathbf{F}^l = \begin{bmatrix} f'_g(\phi_{i,1}^l) & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ 0 & \cdots & f'_g(\phi_{i,p}^l) & \cdots & 0 \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & \cdots & f'_g(\phi_{n_l,1}^l) \end{bmatrix}.$$

D. Interactive Decision Optimization (IDO)

To combine the classification result from AWLEL and probabilistic regression from SMIPR, an interactive module called IDO is designed, which forms an end-to-end learning framework to reduce user intervention (individual learning of bags and instances) and achieve better classification results. By combining Eqs. (8), (10) (14) and (17), the predicted label of a bag \mathbf{X}_i can be obtained as follows:

$$\hat{\mathbf{Y}}_i = \{f_{decision}^c(g_k^l(\mathbf{X}_i))\}_{c=1}^K \quad (22)$$

s.t. $g_k^l(\mathbf{X}_i) \in [\min(\mathbf{T}_1), \max(\mathbf{T}_1)]$

where $f_{decision}^c$ is the decision function for c th class ($c = 1, 2, \dots, K$), and it can be formulated as

$$f_{decision}^c(r) = \begin{cases} 1, & \rho(r) > \tau \\ 0, & \rho(r) \leq \tau \end{cases} \quad (23)$$

where τ is the user-defined decision threshold, and it is individually set for every c , and $\rho(\cdot)$ represents the softmax function. In our experiment, we set $\tau = 0.8$, that is, only when the probability of belonging to class c is larger than 0.8, it can be classified as class c .

E. Optimization Strategy

In this section, we give the optimal solution of Eq. (10) through the strategy of the ADMM algorithm [32]. For simplicity, Eq. (10) is reformulated with the *Lagrangian function* as

$$f_L(\mathbf{w}^t) = \left\| \sqrt{\Gamma}(\mathbf{A}\mathbf{w}_t - \mathbf{T}) \right\|_2^2 + \lambda \|\mathbf{w}_t\|_2^2 \quad (24)$$

$$f_L(\mathbf{T}) = \left\| \sqrt{\Gamma}(\mathbf{A}\mathbf{w}_t) - \mathbf{T} \right\|_2^2 + \vartheta \left\| \sqrt{\Omega}(\mathbf{T} - \mathbf{Y}) \right\|_2^2 \quad (25)$$

Fix \mathbf{T} Update \mathbf{w}^t : When \mathbf{T} are known, taking the derivative of Eq. (24) and setting it to 0. Then Eq. (24) can be written as the following optimization with respect to \mathbf{w}^t :

$$2\mathbf{A}^T\Gamma(\mathbf{A}\mathbf{w}_t - \mathbf{T}) + 2\lambda\mathbf{w}_t = 0$$

$$\Rightarrow \mathbf{w}_t = (\lambda\mathbf{I} + \mathbf{A}^T\Gamma\mathbf{A})^{-1}\mathbf{A}^T\Gamma\mathbf{T} \quad (26)$$

Fix \mathbf{w}^t Update \mathbf{T} : Since \mathbf{w}^t is fixed, similarly setting the derivative of Eq. (24) to 0, we arrive at

$$2\Gamma(\mathbf{A}\mathbf{w}_t - \mathbf{T}) + 2\vartheta\Omega(\mathbf{T} - \mathbf{Y}) = 0$$

$$\Rightarrow \mathbf{T} = (\Gamma + \vartheta\Omega)^{-1}(\Gamma\mathbf{A}\mathbf{w}_t + \vartheta\Omega\mathbf{Y}) \quad (27)$$

Based on the above results, we alternatively update \mathbf{T} and \mathbf{w}_t through the Eqs. (9) and (27) until convergence or the termination condition is satisfied.

Algorithm 1 BMIML

Input: The matrix representation of i^{th} samples in all n training samples: \mathbf{X}_i ; the set of training instances matrix (bags): $\{\mathbf{x}_i^j\}_{j=1}^{n_i}$; the label matrix for all n training samples and bags: $\mathbf{Y}_i, i = 1$ to n , decision threshold τ .

Output: Predicted Label $\hat{\mathbf{Y}}_i$.

Steps of label enhancement learning:

Calculate Z and H in the board learning system with the input X according to Eq.(1)-Eq.(3);

Calculate \mathbf{w}^t and \mathbf{T} by solving the problem of Eq. (11) using Eq.(26)-Eq.(27);

Steps of multi-instance probabilistic regression:

Do

For $t = 1$ to n

Generate distance matrix of X_i according to Eq.(16);

Clustering instances in X_i into S clusters:

$$\bigcup_{p=1}^S G_p = \{\mathbf{X}_i\}_{i=1}^n;$$

Update the weights of probabilistic regression W_{PR} according to Eq.(20);

END

until Convergence

Classification:

Predict the label according to Eq.(22):

$$\hat{\mathbf{Y}}_i \leftarrow f_{decision}^c(g_L^k(\mathbf{X}_i))$$

IV. EXPERIMENTAL

A. Datasets

TABLE II
PROPERTIES OF DATASETS

dataset	Instances	Bags	Labels	Image Type	Resolution
<i>NuCLS</i>	5,432	1,358	7	WSI	Various
<i>Breast</i>	2,416	151	22	WSI	1024*1024
<i>Pannuke</i>	31,616	7,904	5	WSI	256*256
<i>ODR</i>	90,000	10,000	8	fundus photos	576*576
<i>NIH</i>	896,960	112,120	14	X-ray	512*512

Our experiment was conducted over 5 real-world data sets from TCGA and Github for multi-label medical image classification about *whole-slide images* (WSIs), *X-ray*, and *computed tomography* (CT), etc. The **NuCLS data set** [33] is collected by TCGA, which contains 1358 WSIs for breast cancer with 7

possible labels. The **Breast Cancer Semantic Segmentation data set** (BCSS) [34] consists of 151 hematoxylin and eosin stained WSIs corresponding to 22 histologically-confirmed breast cancer cases. **Pannuke data set** [35] consists of 7904 WSIs across 19 different tissue types with 5 possible labels. The **ODR data set** [36] contains 10,000 *color retinal fundus* images annotated with 8 possible labels. *ODR* is collected by Shanggong Medical Technology Co., Ltd. from different hospitals/medical centers in China. In these institutions, fundus images are captured by various cameras in the market, such as Canon, Zeiss and Kowa, under various image resolutions. The largest data set **NIH Cheat X-ray data set** collected by the *NClinical Center* (clinicalcenter.nih.gov) and *National Library of Medicine* (www.nlm.nih.gov) contains 112,120 images with 14 possible labels, and each image is represented with a bag of 4 instances. The properties of these data sets are summarized in Table II. For each data set, 60% of the data are randomly selected for training, 10% for validation, and the remaining 30% for testing. In our experiment, the results are recorded after 10 epochs of model training where the instances in the bags were shuffled in each epoch.

B. Settings

To verify the advantage of BMIML on the task of multi-label medical image classification, seven state-of-the-art MIL approaches were compared: MIMLNN [23], MIMLSVM [1], MIMLmiSVM [23], MIMLkNN [24], MIMLBOOST [1],

MIMLfast [19], DeepMIML [25]. For fair comparison, the parameters of all the compared approaches are determined in the same way if no value is suggested in their literature. Instances are simply divided according to the size of the original image. In our experiment, we try to ensure that the size of each instance is about $64 * 64$. Thus, the number of instances in each bag is equal to the resolution of the original image divided by 64 (See Table II for details of the data sets). Of course, other methods can also be used to generate the instances. Four commonly used MIML metrics are employed for performance evaluation: *hamming loss* (HL), *one error* (OE), *ranking loss* (RL), and *average precision* (AP). All definitions of these metrics can be found in [22, 40]. For better performance evaluation, 10-fold cross validation is conducted on a machine with i7-9700k 3.60GHz CPU and 32 GB RAM memory.

C. Performance Comparison

Medium data sets The comparison results on three medium data sets are listed in Table III. BMIML achieves the best performance in most cases, MIMLNN and MIMLkNN work steadily on all the data sets but are not competitive when compared with BMIML. Although MIMLSVM achieves comparable results with our proposed methods in some cases, it is less effective on large data sets in Table IV. MIMLBoost and DeepMIML can handle only two smallest data sets (*NuCLS*

TABLE III
COMPARISON RESULTS (MEAN \pm STD.) ON THREE MEDIUM DATA SETS

Methods	MIMLNN	MIMLSVM	MIMLmiSVM	MIMLkNN	MIMLBoost	MIMLfast	DeepMIML	BMIML
<i>NuCLS</i>								
H.L. \downarrow	.125 \pm .004	.106 \pm .008	.494 \pm .017	.233 \pm .005	.116 \pm .025	.253 \pm .028	.202 \pm .030	.088\pm.030
O.E. \downarrow	.264 \pm .010	.132 \pm .027	.136 \pm .043	.284 \pm .022	.029\pm.001	.583 \pm .061	.525 \pm .008	<u>.037\pm.015</u>
R.L. \downarrow	.077 \pm .002	.041\pm.020	.368 \pm .017	.380 \pm .023	.099 \pm .005	.392 \pm .004	.325 \pm .019	<u>.043\pm.010</u>
A.P. \uparrow	.857 \pm .041	.941 \pm .006	.856 \pm .028	.757 \pm .007	.921 \pm .009	.722 \pm .011	.815 \pm .046	.968\pm.007
<i>Breast</i>								
H.L. \downarrow	.293 \pm .060	.297 \pm .011	.511 \pm .041	.297 \pm .033	.460 \pm .030	.318 \pm .021	.541 \pm .032	.290\pm.017
O.E. \downarrow	.219 \pm .013	.206 \pm .032	.183 \pm .003	.250 \pm .062	.013\pm.001	.500 \pm .016	.500 \pm .003	<u>.094\pm.001</u>
R.L. \downarrow	.204 \pm .007	.196 \pm .050	.438 \pm .028	.483 \pm .010	.943 \pm .041	.493 \pm .022	.502 \pm .046	.172\pm.004
A.P. \uparrow	.822 \pm .028	.832 \pm .064	.770 \pm .071	.599 \pm .025	.624 \pm .019	.591 \pm .016	.530 \pm .026	.854\pm.021
<i>Pannuke</i>								
H.L. \downarrow	.299 \pm .036	.285 \pm .041	.510 \pm .018	.299 \pm .005	N/A	.377 \pm .011	N/A	.276\pm.005
O.E. \downarrow	.250 \pm .012	.167\pm.024	.182 \pm .033	.200 \pm .022	N/A	.600 \pm .032	N/A	.212 \pm .038
R.L. \downarrow	.209 \pm .030	.189 \pm .006	.438 \pm .009	.509 \pm .036	N/A	.465 \pm .031	N/A	.151\pm.014
A.P. \uparrow	.806 \pm .042	.823 \pm .045	.770 \pm .013	.441 \pm .040	N/A	.439 \pm .060	N/A	.846\pm.003

$\uparrow(\downarrow)$ indicates that the larger (smaller) the value, the better the performance; Bold indicates the best performance of this metric; underline indicates the next best performance of this metric; N/A represents that no result was obtained in 72 hours.

TABLE IV
CLASSIFICATION AVERAGE PRECISION (AP) (MEAN \pm STD.) OF COMPARISON ALGORITHMS ON TWO LARGE DATA SETS WITH VARIOUS DATA SIZES

Dataset (Size)	MIMLNN	MIMLSVM	MIMLmiSVM	MIMLkNN	MIMLBoost	MIMLfast	DeepMIML	BMIML	
ODR	#2K	.670±.080	.649±.002	.700±.088	.214±.022	.580±.088	.465±.070	<u>.686±.002</u>	.727±.056
	#4K	.741±.047	.747±.010	N/A	.225±.060	.604±.036	.466±.048	N/A	.778±.028
	#6K	.756±.020	.775±.003	N/A	.243±.031	N/A	.483±.005	N/A	.835±.039
	#8K	.778±.031	.797±.014	N/A	.294±.090	N/A	.506±.043	N/A	.878±.047
	#10K	.794±.018	.846±.041	N/A	.342±.066	N/A	.512±.056	N/A	.917±.030
NIH	#30K	.391±.090	.508±.080	N/A	.271±.082	N/A	.344±.026	N/A	.536±.046
	#60K	.396±.002	.511±.052	N/A	.274±.091	N/A	.350±.075	N/A	.574±.002
	#90K	.396±.0081	.519±.019	N/A	.274±.026	N/A	.370±.036	N/A	.603±.041
	#120K	.397±.041	.527±.066	N/A	N/A	N/A	.376±.028	N/A	.661±.039

N/A means that no result was obtained in 72 hours.

Bold indicates the best performance of this metric; underline indicates the next best performance of this metric.

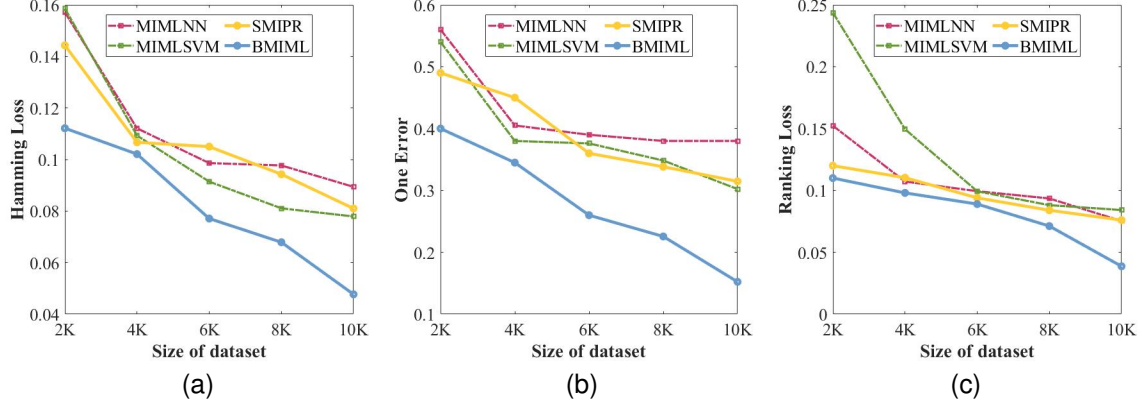


Fig. 6. Comparison results on *ODR* with varying data size; the values smaller, the performance better.

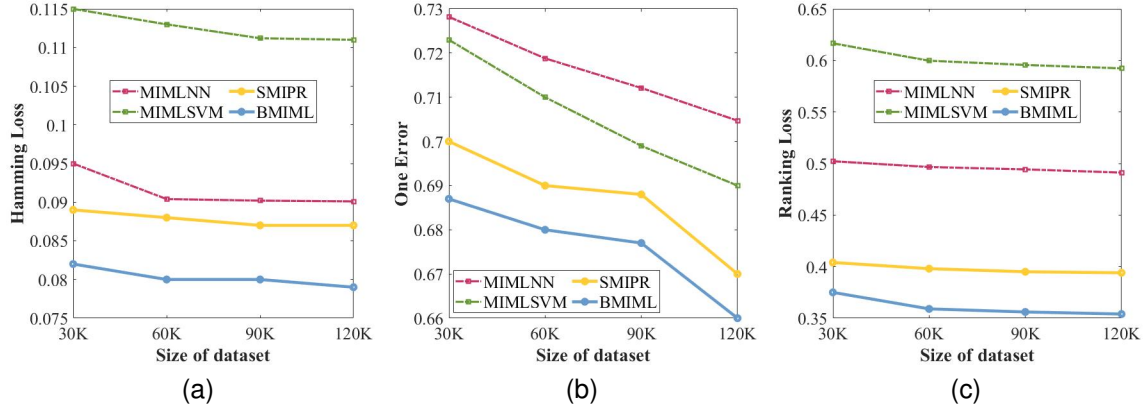


Fig. 7. Comparison results on *NIH* with varying data size; the values smaller, the performance better.

and *Breast*), and do not yield very good performance. MIML-fast works very poorly over all metrics on these three data sets. When the number of instances increases, its accuracy drops obviously.

Large data sets *ODR* and *NIH* contain 10,000 and 112,120 bags respectively, which are too large for most existing MIML approaches. Therefore, the comparison was conducted on their subsets with various data sizes. The number of bags in *ODR* ranges from 2,000 to 10,000, and the number of bags in *NIH* ranges from 30,000 to 120,000, and the average precision (AP) is shown in Table IV. For *NIH*, MIMLmiSVM and MIMIBoost cannot return any result after 72 hours even for the smallest data size (30K). Similarly, in *ODR*, MIMIBoost can only handle up to 4,000 bags while MIMLmiSVM up to 2,000 bags. In Table IV, the AP performance of MIMLkNN and MIMLfast are not comparable with other methods. For this reason, their performances on HL, OE and RL are not shown in Figs. 6 and 7. In Figs. 6 and 7, the trends of HL, OE, and RL drop along with increasing data sizes on the two large data sets *ODR* and *NIH*, respectively, while our proposed BMIML is obviously better than the others. Furthermore, BMIML is much more stable and effective than other methods on *NIH* data set for four evaluation the metrics (HL, OE, RL, AP).

D. Module Analysis

To evaluate the performance of the two proposed modules (AWLEL and SMIPR) ablation studies are conducted. The number of layers l in BMIML and SMIPR is both set to 3. For **medium datastes**, as shown in Table V, BMIML achieves the best performance and the proposed SMIPR alone performs the next best in most cases which validate our idea to consider both global view and local view rather than local view only. The performance of AWELE alone in various metrics is not competitive to SMIPR and BMIML since the ability of BLS feature learning is relatively weak. As illustrated in Table VI, for **large data sets** AWELE does not work well while SMIPR is relatively better but still not comparable to BMIML. With the increasing data set sizes, the advantage of BMIML becomes more and more obvious. Combined with Tables VI and VII, it can be observed that for large data sets, the combination of AWELE and SMIPR not only improve accuracy but also training efficiency.

E. Efficiency Comparison

The training time of each approach on the three data sets is shown in Table VII and their trends (based on \log_{10}) are drawn in Fig 8 for easier comparison. Obviously, MIMLfast

TABLE V
CLASSIFICATION PERFORMANCE (MEAN \pm STD.) OF AWLEL, SMIPR, AND BMIML ON THREE MEDIUM DATA SETS

Datasets	AWLEL	SMIPR	IDO	H.L. \downarrow	O.E. \downarrow	R.L. \downarrow	A.P. \uparrow
<i>NuCLS</i>	✓			.463 \pm .021	.250 \pm .011	.169 \pm .042	.736 \pm .080
		✓		.105 \pm .005	.056 \pm .020	.051 \pm .003	.908 \pm .012
	✓	✓	✓	.088\pm.030	.037\pm.015	.043\pm.010	.968\pm.007
<i>Breast</i>	✓			.600 \pm .020	.500 \pm .013	.543 \pm .050	.548 \pm .016
		✓		.291 \pm .040	.193 \pm .003	.190 \pm .006	.833 \pm .019
	✓	✓	✓	.290\pm.017	.094\pm.001	.172\pm.004	.854\pm.021
<i>Pannuke</i>	✓			.617 \pm .070	.400 \pm .023	.594 \pm .061	.432 \pm .084
		✓		.290 \pm .031	.238 \pm .022	.187 \pm .030	.825 \pm .051
	✓	✓	✓	.276\pm.005	.212\pm.038	.151\pm.014	.846\pm.003

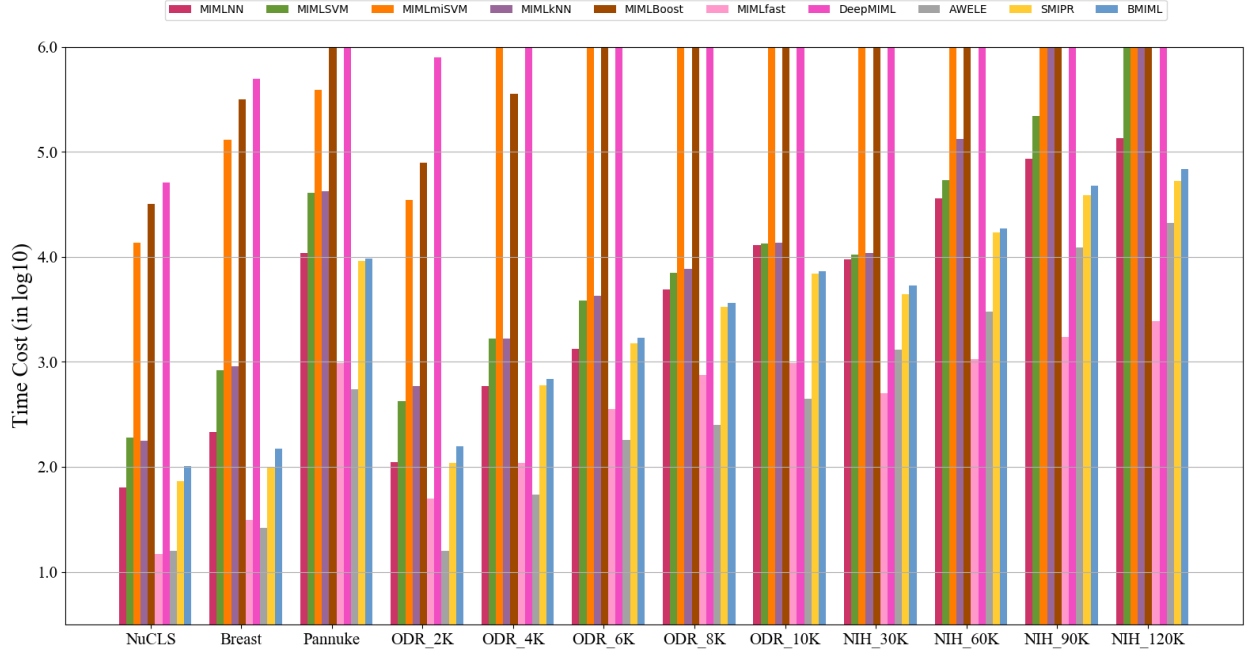


Fig. 8. Training time comparison (in seconds).

TABLE VI
CLASSIFICATION AVERAGE PRECISION (AP) (MEAN \pm STD.) OF AWLEL, SMIPR AND BMIML ON TWO LARGE DATASETS WITH VARIOUS DATA SIZES

datasets (size)	AWLEL	SMIPR	BMIML
<i>ODR</i>	#2K .462 \pm .064	.714 \pm .008	.727\pm.056
	#4K .469 \pm .025	.755 \pm .016	.778\pm.028
	#6K .502 \pm .016	.785 \pm .054	.835\pm.039
	#8K .508 \pm .076	.837 \pm .023	.878\pm.047
	#10K .519 \pm .033	.864 \pm .033	.917\pm.030
<i>NIH</i>	#30K .310 \pm .025	.535 \pm .021	.536\pm.046
	#60K .318 \pm .030	.554 \pm .045	.574\pm.002
	#90K .320 \pm .011	.580 \pm .002	.603\pm.041
	#120K .331 \pm .061	.612 \pm .033	.661\pm.039

is the most efficient one. However, as illustrated in Tables III and IV, MIMLfast does not work well in the four MIML metrics (HL, OE, RL, AP) because MIMLfast only employs a simple linear classifier and lacks preprocessing the raw images. Although such a framework greatly improves efficiency, it does not work well on the raw images. MIMLBoost is most time-consuming, followed by MIMLmiSVM and MIMLkNN. As shown in Table VII, the advantage of our proposed BMIML is obvious. On *ODR*, MIMLBoost can obtain results in 72

hours for the two smallest subsets only, while MIMLmiSVM can handle only 2000 samples. In contrast, BMIML takes only 19 hours even for the largest size (120K). On *NIH*, MIMLBoost and MIMLmiSVM fail to obtain any result in 72 hours even with the smallest size, while MIMLkNN and MIMLSVM cannot work when the data size reaches 90k, but BMIML can still work well and efficiently. On the largest data (*NIH_120K*), the advantage of BMIML is even more obvious. Except for MIMLfast, none of the existing methods can deal with large data sets faster than BMIML. In Table VII, both MIMLfast and AWELE can achieve high efficiency, but when the data size reaches 30k, the efficiency of AWELE decreases significantly. As observed in Tables IV, VI and VII, MIMLfast is sensitive to the number of instances, and AWLEL is sensitive to the number of bags. In other words, the time cost is not only related to the number of bags but also to the number of instances in each bag.

V. CONCLUSION

In this paper, an accurate and efficient BMIML framework was successfully developed, which is suitable for multi-label image classification in medical scenarios. The proposed frame-

TABLE VII
TRAINING TIME COMPARISON (IN SECONDS)

Datasets	MIMLNN	MIMLSVM	MIMLmiSVM	MIMLkNN	MIMLBoost	MIMLfast	DeepMIML	AWLEL	SMIPR	BMIML
NuCLS	63.7	189.6	13672.8	178.4	32165.2	14.7	50980.1	15.9	73.4	102.1
Breast	213.3	832.5	130212.2	899.38	314913.4	31.4	499114.2	26.2	99.3	149.94
Pannuke	10918.4	40600.3	390637.3	42383.2	N/A	972.5	N/A	550.3	9174.7	9691.3
ODR_2K	110.5	424.5	34689.6	584.57	78909.2	49.7	788860.1	15.8	108.6	157.3
ODR_4K	592.2	1677.0	N/A	1669.3	356740.7	109.4	N/A	54.3	600.4	685.4
ODR_6K	1326.3	3848.5	N/A	4248.5	N/A	357.4	N/A	180.7	1510.7	1690.7
ODR_8K	4875.1	7056.7	N/A	7656.9	N/A	745.9	N/A	252.6	3321.5	3651.5
ODR_10K	12832.4	13446.8	N/A	13680.2	N/A	972.5	N/A	444.6	6872.3	7349.1
NIH_30K	9454.7	10577.6	N/A	10839.8	N/A	500.9	N/A	1303.6	4406.9	5352.3
NIH_60K	35718.9	54054.6	N/A	131671.5	N/A	1062.5	N/A	3012.7	16994.8	18657.2
NIH_90K	85520.1	217528.8	N/A	N/A	N/A	1720.9	N/A	12303.6	38416.6	47335.3
NIH_120K	135587.2	N/A	N/A	N/A	N/A	2420.1	N/A	21077.3	53180.8	68459.3

work consists of three novel modules i) auto-weighted label enhancement learning (AWELE), ii) scalable multi-instance probabilistic regression (SMIPR), and iii) interactive decision optimization (IDO). AWELE fully takes into account the inter-correlations of the bags, instances, and labels from the training sample, leading to more effective classification. Compared to the existing indirect methods, SMIPR utilizes the inter-instance correlations directly which can reduce the information loss incurred during the conversion process so that it is more effective and efficient than existing indirect methods. IDO works as a bridge to interactively combine and optimize the results from AWELE and SMIPR. Therefore, an interactive end-to-end single network for MIML becomes possible, which has never been done in the literature. Extensive experiments were conducted on several real-world medical image databases. The results demonstrate that the proposed BMIML is: i) highly effective (improved by up to 2% – 40.7% on AP) under the four-evaluation metrics (HL, OR, RL, AP) than other state-of-the-art MIML algorithms; ii) significantly more efficient (about 16.56% – 99.99% faster) than most existing algorithms while dealing with large data sets (except for MIMLfast, which is with very poor accuracy). In the future, we will try to employ other kinds of images rather than medical images only.

REFERENCES

- [1] Z.-H. Zhou and M.-L. Zhang, “Multi-instance multi-label learning with application to scene classification,” in *Advances in neural information processing systems*, 2006, pp. 1609–1616.
- [2] M. Jie and Z. Hong, “Image classification algorithm based on lts-hd multi instance multi label rbf,” in *2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2017, pp. 190–194.
- [3] L. Song, J. Liu, B. Qian, M. Sun, K. Yang, M. Sun, and S. Abbas, “A deep multi-modal cnn for multi-instance multi-label image classification,” *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 6025–6038, 2018.
- [4] X.-Y. Zhang, H. Shi, C. Li, and P. Li, “Multi-instance multi-label action recognition and localization based on spatio-temporal pre-trimming for untrimmed videos,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 886–12 893.
- [5] S. Biswas and J. Gall, “Multiple instance triplet loss for weakly supervised multi-label action localisation of interacting persons,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2159–2167.
- [6] T. Li, Y. Yang, and H.-B. Shen, “Hmiml: Hierarchical multi-instance multi-label learning of drosophila embryogenesis images using convolutional neural networks,” in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 907–912.
- [7] J.-S. Wu, S.-J. Huang, and Z.-H. Zhou, “Genome-wide protein function prediction through multi-instance multi-label learning,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 5, pp. 891–902, 2014.
- [8] Q. Chang, H. Qu, Y. Zhang, M. Sabuncu, C. Chen, T. Zhang, and D. N. Metaxas, “Synthetic learning: Learn from distributed asynchronous discriminator gan without sharing medical image data,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 856–13 866.
- [9] T. Zhao, K. Cao, J. Yao, I. Nogues, L. Lu, L. Huang, J. Xiao, Z. Yin, and L. Zhang, “3d graph anatomy geometry-integrated network for pancreatic mass segmentation, diagnosis, and quantitative patient management,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 743–13 752.
- [10] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, “High-resolution encoder-decoder networks for low-contrast medical image segmentation,” *IEEE Transactions on Image Processing*, vol. 29, pp. 461–475, 2019.
- [11] Y. Li, Y. Iwamoto, L. Lin, R. Xu, R. Tong, and Y.-W. Chen, “Volumenet: A lightweight parallel network for super-resolution of mr and ct volumetric data,” *IEEE Transactions on Image Processing*, vol. 30, pp. 4840–4854, 2021.
- [12] K. Xu, Z. Zhao, J. Gu, Z. Zeng, C. W. Ying, L. K. Choon, T. C. Hua, and P. K. Chow, “Multi-instance multi-label learning for gene mutation prediction in hepatocellular carcinoma,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 6095–6098.
- [13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 318–14 328.
- [14] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng, “Learning calibrated medical image segmentation via multi-rater agreement modeling,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 341–12 351.
- [15] L. Wang, Y. Liu, H. Di, C. Qin, G. Sun, and Y. Fu, “Semi-supervised dual relation learning for multi-label classification,” *IEEE Transactions on Image Processing*, vol. 30, pp. 9125–9135, 2021.
- [16] Y. Xing, G. Yu, C. Domeniconi, J. Wang, Z. Zhang, and M. Guo, “Multi-view multi-instance multi-label learning based on collaborative matrix factorization,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5508–5515.
- [17] Y. Li, S. Wang, Q. Tian, and X. Ding, “A boosting approach to exploit instance correlations for multi-instance classification,” *IEEE transactions on neural networks and learning systems*, vol. 27, no. 12, pp. 2740–2747, 2015.
- [18] Z. Chi, Z. Wang, and W. Du, “Explicit metric-based multiconcept multi-instance learning with triplet and superbag,” *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [19] S.-J. Huang, W. Gao, and Z.-H. Zhou, “Fast multi-instance multi-label learning,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2614–2627, 2018.
- [20] C. P. Chen and Z. Liu, “Broad learning system: An effective and efficient incremental learning system without the need for deep architecture,” *IEEE transactions on neural networks and learning systems*, vol. 29, no. 1, pp. 10–24, 2017.
- [21] H. D. Nguyen, X.-S. Vu, and D.-T. Le, “Modular graph transformer networks for multi-label image classification,” in *Proceedings of the*

- AAAI Conference on Artificial Intelligence, vol. 35, no. 10, 2021, pp. 9092–9100.
- [22] J. Ma and Y. Liu, “Latent topic-aware multi-label classification,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 558–573.
 - [23] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, “Multi-instance multi-label learning,” *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.
 - [24] M.-L. Zhang, “A k-nearest neighbor based multi-instance multi-label learning algorithm,” in *2010 22nd IEEE international conference on tools with artificial intelligence*, vol. 2. IEEE, 2010, pp. 207–212.
 - [25] J. Feng and Z.-H. Zhou, “Deep miml network,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
 - [26] J. Li, G. Zhao, Y. Tao, P. Zhai, H. Chen, H. He, and T. Cai, “Multi-task contrastive learning for automatic ct and x-ray diagnosis of covid-19,” *Pattern Recognition*, vol. 114, p. 107848, 2021.
 - [27] F. Chu, T. Liang, C. P. Chen, X. Wang, and X. Ma, “Weighted broad learning system and its application in nonlinear industrial process modeling,” *IEEE transactions on neural networks and learning systems*, vol. 31, no. 8, pp. 3017–3031, 2019.
 - [28] e. a. Herrera, F., *Multiple Instance Learning. Foundations and Algorithms. Multiple Instance Learning: Foundations and Algorithms*. SpringerISBN: 978-3-319-47758-9, 2016.
 - [29] H. Ney, “On the probabilistic interpretation of neural network classifiers and discriminative training criteria,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 2, pp. 107–119, 1995.
 - [30] C. Yuan and H. Yang, “Research on k-value selection method of k-means clustering algorithm,” *J*, vol. 2, no. 2, pp. 226–235, 2019.
 - [31] D. Karimi and S. E. Salcudean, “Reducing the hausdorff distance in medical image segmentation with convolutional neural networks,” *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019.
 - [32] B. Wahlberg, S. Boyd, M. Annergren, and Y. Wang, “An admm algorithm for a class of total variation regularized estimation problems,” *IFAC Proceedings Volumes*, vol. 45, no. 16, pp. 83–88, 2012.
 - [33] M. Amgad, L. A. Atteya, H. Hussein, K. H. Mohammed, E. Hafiz, M. A. Elsebaie, A. M. Alhusseiny, M. A. AlMoslemany, A. M. Elmatboly, P. A. Pappalardo *et al.*, “Nucls: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation,” *arXiv preprint arXiv:2102.09099*, 2021.
 - [34] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. Salem, A. F. Ismail, A. M. Saad *et al.*, “Structured crowdsourcing enables convolutional segmentation of histology images,” *Bioinformatics*, vol. 35, no. 18, pp. 3461–3467, 2019.
 - [35] J. Gamper, N. A. Koohbanani, K. Benes, S. Graham, M. Jahanifar, S. A. Khurram, A. Azam, K. Hewitt, and N. Rajpoot, “Pannuke dataset extension, insights and baselines,” *arXiv preprint arXiv:2003.10778*, 2020.
 - [36] N. Li, T. Li, C. Hu, K. Wang, and H. Kang, “A benchmark of ocular disease intelligent recognition: one shot for multi-disease detection,” in *International Symposium on Benchmarking, Measuring and Optimization*. Springer, 2020, pp. 177–193.