Patients and Slides are Equal: A Multi-level Multi-instance Learning Framework for Pathological Image Analysis

Fei Li[1,2], Mingyu Wang[1,2], Bin Huang[1,2], Xiaoyu Duan[4], Zhuya Zhang[1,2], Ziyin Ye[3(✉)], and Bingsheng Huang[1,2(✉)]

[1] Medical AI Lab, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China
huangb@szu.edu.cn

[2] Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen, China

[3] Department of Pathology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China
yeziyin@mail.sysu.edu.cn

[4] Department of Pathology, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen, China

Abstract. In current pathology image classification, methods mostly rely on patch-based multi-instance learning (MIL), which only considers the relationship between patches and slides. However, in clinical medicine, doctors use slide-level labels to summarize patient-level labels as a diagnostic result, indicating the involvement of three levels of patch, slide, and patient in actual pathology image analysis, which we refer to as the multi-level multi-instance learning (ML-MIL) problem. To address this issue, we propose a novel and general framework called Patients and Slides are Equal (P&SrE), inspired by the doctor's diagnostic process of repeatedly confirming labels at the patient and slide level. In this framework, we treat patients and slides as instances at the same level and use transformers and attention mechanisms to build connections between them. This allows for interaction between patient-level and slide-level information and the correction of their respective features to achieve better classification performance. We evaluate our method on two datasets using two state-of-the-art MIL methods as baselines. The results show that our method improves the performance of the baselines on both slide and patient levels. Our method provides a simple and effective solution to the common problem of ML-MIL in medical clinical scenarios and has broad potential applications.

Keywords: Multiple instance learning · Multi-level Labels · Pathology Images · Transformer

F. Li, M. Wang and B. Huang—Contribute equally to this work.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 63–71, 2023.
https://doi.org/10.1007/978-3-031-43904-9_7

1 Introduction

Pathological image analysis is a vital area of research within medical image analysis, focused on utilizing computer technology to aid doctors in diagnosing and treating diseases by analyzing pathological tissue slide images [5]. Advancements in pathological image analysis have been made in early cancer diagnosis, tumor localization, and grading, and treatment planning [3,10]. Multi-instance learning [2] is the primary analysis method used, which involves analyzing tasks based on slide labels and patches. Despite this, the clinical pathological analysis presents certain challenges and complexities, with the ultimate diagnosis relying on patients rather than slides.

Specifically, in clinical problems of pathological image analysis, doctors usually summarize patient-level labels based on slide labels as the diagnostic results [1,6]. For example, for the pathological discrimination diagnosis task of intestinal tuberculosis(ITB) and Crohn's desease(CD), the categories of postoperative slides are divided into three types (normal, CD, ITB), and doctors will summarize the binary results of patients (ITB or CD) based on slide-level labels [6]. Similar situations exist in other tasks, such as the classification of breast cancer metastases in lymph nodes, where slide categories may have different classifications, and the corresponding diagnosis of the same patient is whether the cancer has spread to the regional lymph nodes (N-stage) [1]. Therefore, as shown in Fig. 1, actual pathological image analysis involves the relationships of patches, slides, and patients, which is called a multi-level multi-instance learning (ML-MIL) problem. Among them, for patients and slides, patients are bags while slides are instances, and for slides and patches, slides are bags while patches are instances.

There are generally two methods to solve the ML-MIL problem. The first method is to directly average the prediction values of slides or take the maximum prediction value [9]. This method is relatively simple, but the information exchange between slides is not fully utilized, which may lead to errors in the summary result. The second method is to treat slide-patient as a new MIL problem according to the traditional MIL thinking, where slides are regarded as instances and patient labels as bags. Although this method seems reasonable, the number of patients is usually relatively small, and deep learning models usually require a large amount of data for training. Therefore, the insufficient number of samples at the slide-patient level may make it difficult for the model to learn enough information.

To address the multi-level multi-instance learning (ML-MIL) problem in medical field, we propose a novel framework called Patients and Slides are Equal (P&SrE). Inspired by the iterative labeling process in medical diagnosis, this framework treats patients and slides as instances at the same level and uses transformers and attention mechanisms to build connections between them. This simple yet effective method allows for interaction between patient-level and slide-level information to correct their respective features and improve classification performance. Our framework consists of two steps: first, at the patch-slide level, a common MIL framework is used to train a MIL neural network and obtain

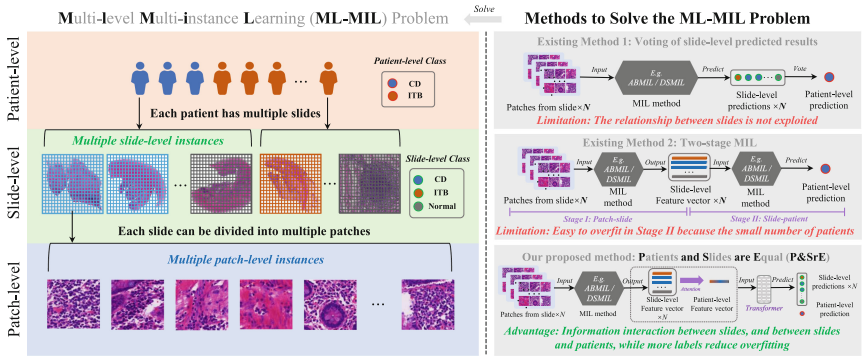

Fig. 1. Description and solutions for the ML-MIL problem.

slide-level feature vectors; then, at the slide-patient level, we use self-attention mechanisms to combine the slides of the same patient into patient-level feature vectors, and treat these patient-level feature vectors together with all slide-level feature vectors of the same patient as instances at the same level, which are inputted into transformers for feature interaction and prediction of patient- and slide-level labels. Our method can effectively solve the problem of difficult training due to the scarcity of samples at the highest level in ML-MIL, and can be integrated into two state-of-the-art methods to further improve performance. We conducted rigorous experiments on two datasets and demonstrated the effectiveness of our method. Our contributions include:

1) Proposing a novel general framework to address the unique "patch-slide-patient" ML-MIL problem in the medical field. Before this, no other framework had directly tackled this specific problem, making our proposal a ground-breaking step in the application of ML-MIL in healthcare;
2) Proposing a simple yet highly effective method that leverages self-attention mechanisms and transformer models to enhance the interaction between slide and patient information. This innovative approach not only improves the classification performance at the patient level but also at the slide level, showcasing its effectiveness and versatility;
3) Conducting extensive experiments on two separate datasets. Our method was seamlessly integrated with two prior state-of-the-art methods, demonstrating its compatibility and adaptability. The experiments resulted in improved performance, indicating that our method enhances the efficacy of these existing approaches.

2 Method

2.1 Overview

Our proposed method P&SrE is illustrated in Fig. 2. Specifically, the framework consists of two parts. The first part is the slide-patch level MIL based on a state-

of-the-art MIL method. The second part is the patient-slide level MIL, which generates patient-level features using attention mechanism and interacts the features with transformer. To enhance readability, we first provide the following symbolization for ML-MIL: For a patient X_i, it has a patient-level classification label Y_i. For patient X_i, there may exist N_i slides S_i=$\{s_j|\text{j=1 to } N_i\}$, where the classification label for each slide s_j is denoted as z_j. For each slide s_j, it may be divided into M_j patches P_j=$\{p_k|\text{k=1 to } M_j\}$. Here, i,j, and k are indices for patient, slide, and patch levels, respectively.

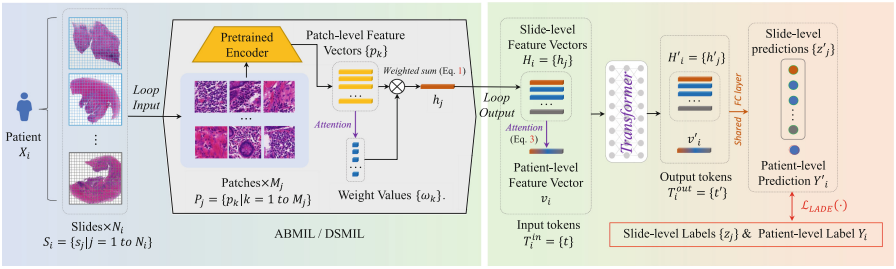

Fig. 2. Overview of the proposed framework P&SrE. This framework consists of two-level MIL parts: Slide-patch level MIL and patient-slide level MIL

2.2 Slide-Patch Level MIL

Our proposed framework has strong scalability as it can be based on any attention-based MIL method. Therefore, we directly use the state-of-the-art (SOTA) MIL methods, ABMIL [8] and DSMIL [9] for the slide-patch stage. These two methods differ in their attention computing approach for each patch.

For ABMIL, the attention of each patch is computed by an MLP. Specifically, for M_j patches p_k, an encoder is applied to obtain the patch feature matrix F_i, where,$F_i \in R^{M_j \times 1024}$. Then, F_i is passed through an fc layer followed by a Tanh activation and another fc layer followed by a sigmoid activation to obtain two feature matrices, F'_i and F''_i, both $\in R^{M_j \times 128}$. These matrices are element-wise multiplied and then passed through an fc layer to obtain the weight of each patch, ω_k.

For DSMIL, the attention of each patch is based on the cosine distance between instances and key instances. First, an fc layer is applied to the patch feature matrix F_i to obtain the importance score θ_k for each patch. The patch with the highest score is selected as the key instance. Then, the feature matrix F_i is mapped to a matrix $Q_i \in R^{M_j \times 128}$ and the cosine similarity between all instances and the key instance is computed as the weight of each patch, ω_k.

Although ABMIL and DSMIL compute attention differently, both methods compute the attention-weighted sum of patch instances features as the bag representation of the slide. Therefore, the slide feature output by both methods can be generalized as:

$$h_j = \sum_{k=1}^{M_j} \omega_k * p_k / \sum_{k=1}^{M_j} \omega_k \tag{1}$$

Finally, we obtain the feature vector set H_i=$\{h_j|$j=1 to $N_i\}$ for all slides $\{s_j\}$ of patientX_i through patch-slide MIL.

2.3 Patient-Slide Level MIL

After performing patch-slide level MIL, we move on to patient-slide level MIL. In general MIL algorithms, the patient is regarded as the bag and the slide as the instance. However, considering the diagnostic process in clinical practice, we propose to treat both patients and slides as instances at the same level. Specifically, our P&SrE framework for patient-slide level consists of two parts: patient-level feature generation based on self-attention and patient-slide feature interaction based on Transformer [11].

Patient-Level Feature Generation Based on Self-attention. Doctors usually select certain key slides for careful observation and information aggregation during diagnosis, similar to the self-attention mechanism. Therefore, we directly use a fully connected (FC) layer to integrate the feature-level features into patient-level features v_i through attention mechanism, serving as patient instances. Specifically, given the feature vector collection $\{h_j\}$ from multiple slides in the previous step, we input it to the FC layer and apply the sigmoid activation function to output the weight α_j for each h_j. Then, we perform a weighted average of the vectors based on this weight to obtain the patient feature v_i:

$$\alpha_j = FC(\{h_j|j = 1 \ to \ N_i\}) \tag{2}$$

$$v_j = \sum_{j=1}^{N_i} \alpha_j * h_j / \sum_{j=1}^{N_i} \alpha_j \tag{3}$$

Patient-Slide Feature Interaction Based on Transformer. This process is where our framework shines. After doctors summarize the patient-level results, they typically review the slides to double-check the diagnosis results. This patient-slide feature interaction (PSFI) naturally lends itself to the construction of a Transformer, and information exchange and integration between slides and patient level are bidirectional. Thus, self-attention is more ideal for this purpose than other kinds of attention (such as cross-attention or doctors' attention). By using the self-attention-based transformer structure, each input token is treated equally (i.e., viewed as the same instance level), and tokens can interact extensively with each other, enabling mutual correction between patients and slides and even between slides. Specifically, we merge the slide feature set $\{h_j\}$ and the patient feature v_i into the input tokens $T_i^{in} = \{h_1, h_2, ..., h_{N_i}, v_i\} = \{t\}$,

and then input them into a multi-layer transformer through self-attention and feed-forward neural network layers to obtain the interaction information between slides and output tokens T_i^{out}:

$$\beta_{k,l} = softmax(W^Q {t_k}^T (W^K t_l)/\sqrt{d}) \tag{4}$$

$$\bar{t}_k = \sum_{l=1}^{N_i+1} \beta_{k,l} W^V t_l \tag{5}$$

$$t'_k = RELU(\bar{t}_k W^R + b_1) W^O + b_2 \tag{6}$$

where d is the dimension of the token, and t_k and t_l come from T_i^{in}. $\beta_{k,l}$ is multi-head attention matrix, and W^Q, W^K, and W^V are weight matrices of query, key, and value, respectively. W^R and W^O are transformation matrices. b_1 and b_2 are bias vectors. This update procedure is repeated for L layers, where the t'_k are fed to the successive transformer layer. Finally, we obtain the output tokens T_i^{out}=$\{h'_1, h'_2, ..., h'_{N_i}, v'_i\}$. Then, all output tokens are input into a shared FC layer, and the patient's predicted logits Y'_i and the predicted classification logits $\{z'_j | j = 1\ to\ N_i\}$ for each slide are output.

Training Progress and Loss Function. During training, we sampled one patient at a time and pre-extracted their batch-level features for all slides, in order to save GPU memory. Due to the issue of class imbalance in both slide level and patient level, we use the LADE [7] loss function.

3 Experiments and Results

3.1 Dataset and Evaluation

CD-ITB Dataset. CD-ITB is a private dataset consisting of 853 slides from 163 patients, with binary patient-level labels of CD or ITB in a ratio of 103:60 and tri-class slide-level labels of CD, ITB, and normal slides in a ratio of 436:121:296, respectively. On average, there were 5 slides per patient. The slides were scanned at a magnification of 40× (0.25 μm/px), and annotations were curated by experienced pathologists. We adopted a patient-level stratification approach for 5-fold cross-validation, with 20% of the training set randomly assigned as the validation set for each fold. The dataset comprises an average of 2.3k instances per bag, with the largest bag containing over 16k instances.

Camelyon17 Dataset. Camelyon17 [1] is a publicly dataset, and its training set comprises 500 slides from 100 breast cancer patients with lymph node metastases. The slides are classified into four distinct categories, namely negative, ITC, micro, and macro, in proportions of 318:36:59:87, respectively. There were 5 slides per patient on average. The patients are divided into two groups

based on their pN stage, namely lymph node positive and lymph node negative, in proportions of 24:76, respectively. The data folding method is the same as the CD-ITB dataset. The average number of instances per bag is approximately 6.1k, and the largest bag contains over 23k instances.

Metrics. We report class-wise weighted accuracy (Acc), precision(Pre), Recall, and F1-score (F1). To avoid randomness, we run all experiments five times and report the averaged metrics.

3.2 Implementation Details

We utilized ResNet50, which was pre-trained on ImageNet1K, to extract features from patches. Each patch was of size 512×512 pixels. For both ABMIL and DSMIL networks, we kept the original parameters for the number of channels at each layer. Following the reference [4], we employed a transformer with 8 heads and 8 layers in the patient-slide feature interactions. All networks are implemented using PyTorch and trained on a NVIDIA RTX TITAN GPU with 24 GB memory. We employed two Adam optimizers with a maximum learning rate of 1e−4 and a cosine annealing update strategy that gradually decreased the learning rate to 1e−12 over 300 epochs.

3.3 Comparisons and Results

We compared our strategy with two state-of-the-art MIL methods to evaluate its performance. To investigate the impact of self-attention and transformers on slide-level and case-level results, we conducted ablation experiments: "ABMIL + P&SrE (with/without PSFI)" and "DSMIL + P&SrE (with/without PSFI)", respectively. For slide-level classification, we used mean pooling and max pooling to pool feature vectors of patches into a representative vector for the slide, which was then fed into a fully connected layer for classification. At the patient level, we used two approaches for prediction: MaxS, where the feature of the instance that achieves the maximum positive probability from the slide-level MIL model is selected to patient-level model, and MaxMinS, where the mean value of features of the maximum and minimum positive probability from the slide-level MIL model is selected to patient-level model.

The results of 5-fold CV at the slide and patient levels are reported in Table 1 and Table 2, respectively. Our P&SrE framework improves both ABMIL and DSMIL methods at both levels. ABMIL with P&SrE improves the F1 score from 0.565 to 0.579 for the CD-ITB dataset and from 0.529 to 0.571 for the Camelyon17 dataset at the slide-level, and improves the F1 score from 0.522 to 0.599 for the CD-ITB dataset and from 0.842 to 0.861 for the Camelyon17 dataset at the patient-level. Therefore, the ablation experiments demonstrate the effectiveness of P&SrE in enhancing the classification performance at both the slide and patient levels.

Table 1. Slide-level 5-fold CV results (%)

Method	CD-ITB dataset				Camelyon17 dataset			
	Pre	Recall	Acc	F1	Pre	Recall	Acc	F1
Mean pooling (on patch-level feature vectors)	45.0	40.2	40.2	41.6	47.4	31.6	31.6	35.5
Max pooling (on patch-level feature vectors)	39.6	33.1	33.1	34.9	46.9	25.4	25.4	29.9
ABMIL [8]	57.6	57.2	57.2	55.4	55.3	63.7	63.7	50.9
(ours) ABMIL + P&SrE (w/o PSFI)	57.0	57.3	57.3	56.5	58.6	65.2	65.2	52.9
(ours) ABMIL + P&SrE	**59.0**	**59.7**	**59.7**	**57.9**	**61.0**	**66.9**	**66.9**	**57.1**
DSMIL [9]	57.0	57.3	57.3	56.9	55.4	63.8	63.8	50.5
(ours) DSMIL + P&SrE (w/o PSFI)	56.7	57.4	57.4	56.6	55.5	64.6	64.6	51.7
(ours) DSMIL + P&SrE	**58.5**	**59.1**	**59.1**	**57.3**	**55.8**	**66.4**	**66.4**	**52.3**

Table 2. Patient-level 5-fold CV results (%)

Method	CD-ITB dataset				Camelyon17 dataset			
	Pre	Recall	Acc	F1	Pre	Recall	Acc	F1
ABMIL (MaxS)	36.9	63.3	46.6	46.6	78.0	94.4	75.0	85.0
ABMIL+ (MaxMinS)	38.2	65.0	48.5	48.2	78.3	94.7	76.0	85.7
ABMIL(baseline+Mean) (on probabilities of slides)	38.2	43.3	53.4	40.6	**98.3**	75.0	**80.0**	85.1
(ours) ABMIL + P&SrE (w/o PSFI)	48.3	60.0	59.5	52.2	81.5	87.1	75.2	84.2
(ours) ABMIL + P&SrE	**56.8**	**71.0**	**66.3**	**59.9**	78.0	**96.1**	76.4	**86.1**
DSMIL + (MaxS)	50.8	56.7	63.8	53.5	78.7	97.4	78	87.1
DSMIL + (MaxMinS)	50.0	60.0	63.2	54.6	79.3	85.5	72	82.3
DSMIL(baseline)+Mean (on probabilities of slides)	43.8	57.7	53.3	48.1	**100**	73.7	**80.0**	84.9
(ours) DSMIL + P&SrE (w/o PSFI)	49.7	**61.7**	62.8	54.9	78.0	**97.9**	77.4	86.8
(ours) DSMIL + P&SrE	**60.0**	56.7	**69.7**	**57.7**	80.2	96.8	79.4	**87.7**

4 Limitations

Our study has some limitations that should be addressed. For instance, we did not explore the possibility of treating patches as an equivalent level to slides and patients. The primary reason is that the vast number of patches required for analysis is significantly larger than that of slides and patients, which presents a computational challenge for training. As a result, we have not yet explored this avenue. In the future, we plan to leverage clustering and active learning methods to reduce the number of patches and enable the interaction of all three levels with the Transformer, which would further enhance the accuracy and efficiency of our proposed method.

5 Conclusion

This study proposes a highly scalable and versatile framework to address M-MIL problems. We first classify the process from patch to slide to the patient in medical pathology diagnosis as a multi-level MIL problem. Based on existing state-of-the-art MIL methods, we then extend the framework to P&SrE, which

conducts feature extraction and interaction at the slide-patient level. By introducing a transformer, the framework enables iterative interaction and correction of information between patients and slides, resulting in better performance at both the patient level and slide level compared to existing state-of-the-art algorithms on two validation datasets.

Acknowledgements. This study was supported by Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515010571), National Natural Science Foundation of China (No. 82271958, 81971684, 81801761), Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions (No. 2022SHIBS0003), and Guangdong Provincial Clinical Research Center for Digestive Diseases (No. 2020B1111170004).

References

1. Bandi, P., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE Trans. Med. Imaging **38**(2), 550–560 (2018)
2. Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: a survey of problem characteristics and applications. Pattern Recognit. **77**, 329–353 (2018)
3. Cui, M., Zhang, D.Y.: Artificial intelligence and computational pathology. Lab. Invest. **101**(4), 412–422 (2021)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Fuchs, T.J., Buhmann, J.M.: Computational pathology: challenges and promises for tissue analysis. Comput. Med. Imaging Graph. **35**(7–8), 515–530 (2011)
6. Gecse, K.B., Vermeire, S.: Differential diagnosis of inflammatory bowel disease: imitations and complications. Lancet Gastroenterol. Hepatol. **3**(9), 644–653 (2018)
7. Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., Chang, B.: Disentangling label distribution for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6626–6636 (2021)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
9. Li, B., Li, Y., Eliceiri, K.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, pp. 14318–14328 (2021)
10. Serag, A., et al.: Translational AI and deep learning in diagnostic pathology. Front. Med. **6**, 185 (2019)
11. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)