World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:18, No:3, 2024

Attention Multiple Instance Learning for Cancer Tissue Classification in Digital Histopathology Images

Afaf Alharbi, Qianni Zhang

Abstract—The identification of malignant tissue in histopathological slides holds significant importance in both clinical settings and pathology research. This paper presents a methodology aimed at automatically categorizing cancerous tissue through the utilization of a multiple instance learning framework. This framework is specifically developed to acquire knowledge of the Bernoulli distribution of the bag label probability by employing neural networks. Furthermore, we put forward a neural network-based permutation-invariant aggregation operator, equivalent to attention mechanisms, which is applied to the multi-instance learning network. Through empirical evaluation on an openly available colon cancer histopathology dataset, we provide evidence that our approach surpasses various conventional deep learning methods.

Keywords—Attention Multiple Instance Learning, Multiple Instance Learning, transfer learning, histopathological slides, cancer tissue classification.

I. Introduction

MEDICAL imaging plays a crucial role in the multifaceted approach to cancer care, offering invaluable insights for the detection, diagnosis, and management of various cancer types through advanced techniques. The field of histopathology stands at the forefront of cancer diagnostics, employing advanced staining techniques and molecular pathology to distinguish between benign and malignant lesions, to understand the tumor microenvironment, and to identify prognostic and predictive biomarkers critical for patient-specific treatment strategies. However, the current reliance on expert pathologists for visual assessment of tumor slides introduces significant challenges in terms of time consumption, financial resources, and limited availability of specialized professionals. Furthermore, visual evaluations are inherently susceptible to inconsistencies and imprecision arising from inter- and intra-observer variability, thereby compromising accurate diagnosis and subsequent treatment planning. The advent of digital pathology has revolutionized this landscape by enabling the application of computational methodologies on digital whole slide images (WSIs), thus automating the process, providing quantitative insights, and reducing subjective factors. Particularly, the latest advancements in Artificial Intelligence (AI) have demonstrated remarkable capabilities in discerning complex image patterns through autonomous acquisition of image interpretations. In some cases, well-designed AI models can even capture latent image information that may elude human perception, thereby enhancing the decision-making process. Consequently, these advancements have solidified AI methods as an invaluable resource for advancing medical image understanding and analysis.

The effectiveness of AI models heavily depends on the availability of a large volume of accurately labeled data for training purposes. However, when it comes to the analysis of WSIs, the process of labeling can present certain challenges due to the limited ability of labels assigned to WSIs or regions of interest (ROIs) to fully capture the complex tissue compositions contained within. Consequently, this can lead to misguided training, resulting in the development of unreliable models.

In the domain of medical imaging, the issue of weakly annotated data is a common concern, where a single label is assigned to an image to indicate its classification as either benign or malignant. To address this challenge, the adoption of the Multi-Instance Learning scheme (MIL) proves to be particularly suitable. Under a binary classification structure, a ”bag” is labeled as positive if it contains one or more positive instances, and negative if it only consists of negative instances. Extensive research and analysis have been dedicated to exploring various models and learning algorithms for MIL within this field [1], [2].

Multiple Instance Learning (MIL) proves to be an effective approach for analyzing histopathological images, primarily due to its ability to reason about subsets of data, often represented by patches, which serve as fundamental units in histopathology computations. In the context of weakly supervised multiple instance learning for histopathological tumor segmentation [3], MIL considers a collection of instances grouped together as a ”bag” with a single assigned class label. The primary objective is to develop a model capable of predicting the label for the entire bag in various ways. Additionally, there is significant interest in identifying the key instance(s) within a bag that contribute to its assigned label [3]. Within the medical field, the latter concern holds substantial importance due to its implications for clinical practice and legal considerations. To address the challenge of bag classification, several strategies have been proposed. These include analyzing similarities among bags [4], embedding instances into a lower-dimensional

A. Alharbi is with School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK (corresponding author, e-mail: a.alharbi@qmul.ac.uk).

Dr. Q. Zhang is with School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK (e-mail: Qianni.zhang@qmul.ac.uk).

model that is subsequently sub-sampled and used as input for bag classification models [5], [6], or considering an aggregated response from instance-specific classifiers [7]–[9]. The last approach can provide interpretable data and outcomes, although its accuracy remains relatively low [10]. Moreover, the significance of employing MIL at the instance level remains subject to debate [4]. In the case of binary classification tasks, whole slide images (WSIs) for malignant cases may contain both malignant and benign patches, while WSIs for benign cases only contain benign patches. When there is an absence of annotations for each individual extracted patch, the MIL framework emerges as an ideal solution for histopathological image classification.

The primary objective of this study is to utilize the Multiple Instance Learning (MIL) technique for the automated classification of various regions of interest (ROIs) across multiple levels. To overcome the limitations of weak supervision in tissue classification, we incorporate the attention mechanism into the MIL framework. This integration allows us to effectively address the challenges associated with the inadequate labeling of training data and improve the accuracy and reliability of the tissue classification process. On top of that, available pre-trained transfer learning models are employed as the baseline comparisons to our proposed Attention-MIL method.

II. Methodology

In supervised machine learning, sufficient labelled training datasets are essential to obtain a reliable and robust model. However, if the training dataset available is not adequate, the performance of the model will considerably be decreased. As biomedical studies and clinical practice produce various data with no annotation daily, it is necessary to develop and implement better alternative approaches for studying or investigating such unannotated clinical and biomedical studies. Therefore, we have applied two weakly supervised methods, which represents our main methods as follows:

A. Multiple Instance Learning

The fundamental method that has been used for a a given problem of supervised learning as a follows a search for a model that can predict a value $y \in \{0, 1\}$ for a single given instance $X \in R$. However: In MIL method, a bag including several instances can be used $X = \{x1, ..., xK\}$, which displays no specific dependencies or patterns together. In the MIL, the assumption is that for different bags, K can be vary and a Y label is connected with the bag as a single binary label. For the instances inside a bag,there are separate labels are existed. Accessing these labels, however, is not available even in the training process. The assumptions of the MIL problem can be defined and form as following:

$$X = \begin{cases} 0, & \text{if } \sum a = 1 \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

Such assumptions suggest that it must be permutation-invariant for a MIL model. Furthermore, the two statements may be reformulated using the maximum operator in a a form:

$$Y = \max_k \{y_k\}$$

It would be challenging to learn a model that attempts to optimize a goal based on maximum over instance labels. One of the reasons behind these challenges is that gradient vanishing issues would be faced in the gradient-based learning. Another reason is this model could be utilised only in the case of using an instance-level classifier.

To make the learning task simpler, we introduce a MIL training model by optimizing the function of log likelihood which basically distributes the bag label with the parameter [0,1] according to the Bernoulli distribution.

B. Attention Mechanism

As we mentioned earlier Attention Mechanism has been proposed by [11] and based on this mechanism, we have applied the A-MIL model. This model proposes using a weighted average where neural network can be used to determine the weights. The weights must sum to 1 at all times in order to be constant for the size of a bag. The average satisfies Theorem 1 where the weights and embedding with the functions values form the f function. If:

$$H = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\} \tag{2}$$

is a bag containing instance embeddings, the attention-based MIL pooling operation is defined.

$$\mathbf{z} = \sum_{k=1}^{K} a_k \mathbf{h}_k \tag{3}$$

where

$$a_k = \frac{\exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_k^\top\right)\right\}}{\sum_{j=1}^{K} \exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_j^\top\right)\right\}} \tag{4}$$

The concepts of V and W are learned via training. This attention scores enables us to interpret the trained model by revealing the influence each instance has on the drawn conclusion and acting as a similarity measure for comparison between the instances.

C. Architecture of A-MIL

Our network architecture primarily consists of a Convolutional Neural Network (CNN) with attention layers. The architecture begins with two convolutional layers followed by max pooling. The output is then flattened and passed through two regression layers. Subsequently, an attention layer is introduced, followed by a sigmoid fully connected output layer.

To illustrate the pipeline structure of the A-MIL system, we refer to Fig. 1. Each patch within a bag undergoes feature extraction to obtain instance-level labels. These labels are extracted by each instance in the dense layer. The attention processing layer calculates the attention scores based on these instance labels. Moreover, attention weights are assigned to facilitate focused aggregation, enabling the capture

of bag-level characteristics. In the A-MIL framework, the weights of various instances within a bag can vary, providing flexibility in capturing the importance of different instances. The bag-level classifier leverages the attention aggregation algorithm to effectively capture the descriptive nature of the bag.

Fig. 1 provides a visual representation of the required feature extractor in A-MIL, while Table I presents the input and output dimensions for each layer in the A-MIL architecture.

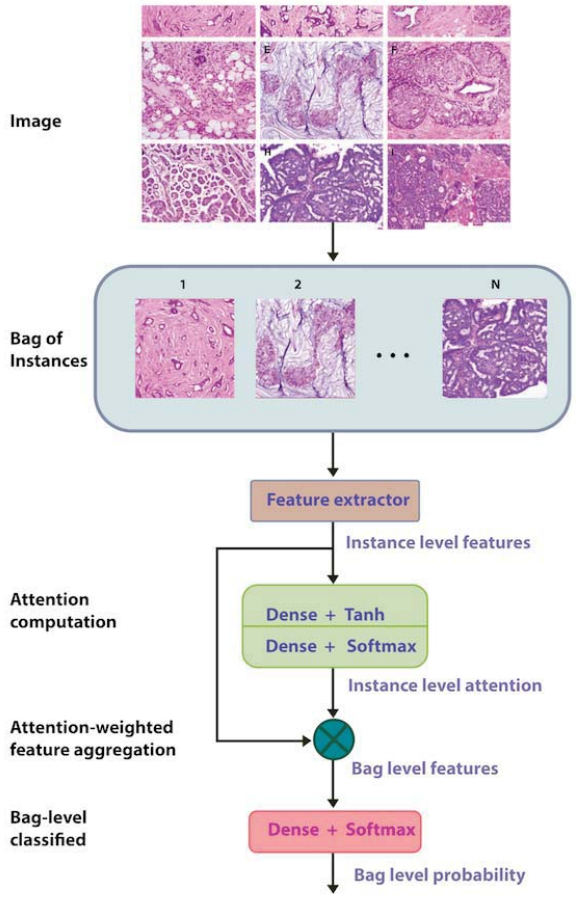

Fig. 1 A-MIL Architecture

TABLE I
AMIL MODEL ARCHITECTURE

Layers	Dimension	
Conv2D	input	27x27x3
	output	24x24x36
Max pooling2D	input	12x12x36
	output	10x10x48
Conv2D	input	10x10x48
	output	5x5x48
Max pooling2D	input	5x5x48
	output	1200
Flatten	input	512
	output	512
Dense	input	512
	output	512
Dropout	input	512
	output	512
Dense	input	512
	output	512
Drpout	input	512
	output	512
MIL:Attention	input	512
	output	1
Multiply	input	1
	output	512
FC1:Sigmoid	input	512
	output	1

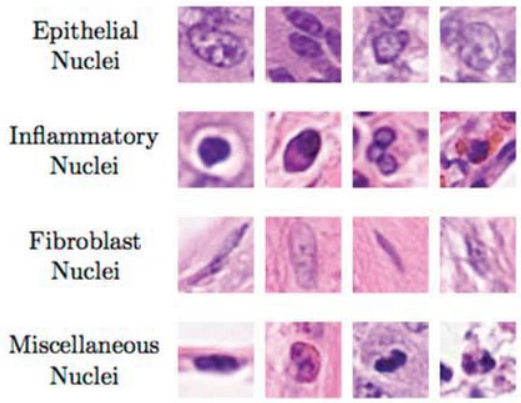

Fig. 2 Example sample from the COLON CANCER Dataset, displaying different forms of nuclei including epithelial, inflammatory, and miscellaneous

D. Evaluation Metrics

In our study, we have employed evaluation metrics to assess the performance of both the A-MIL model and the transfer learning models for COLON CANCER classification. Various statistical measures have been utilized to evaluate the classification models, including Accuracy (AC), Sensitivity (SN), Specificity (SP), F-Score, and Confusion Matrix. These metrics serve as valuable tools for analyzing and interpreting our results.

III. Experimental Setup

A. Datasets

In this section, we provide an overview of the dataset used in our project, specifically designed for identifying weakly-labeled histopathology images in the COLON CANCER dataset. The dataset comprises 100 histopathological images obtained from 9 patients, captured at a 20× optical resolution. Each image is cropped from non-overlapping sections of whole-slide images, resulting in a size of 500 × 500 pixels. The images contain various tissue types, including both normal and malignant tissues, with a majority of nuclei labeled for each cell. The dataset consists of 22,444 nuclei labeled with their respective tissue types, such as epithelial, inflammatory, fibroblast, and others. Fig. 2 illustrates a sample from the COLON CANCER Dataset, showcasing different forms of nuclei detected, namely epithelial, inflammatory, and miscellaneous.

1) A-MIL Model Setup: For the A-MIL model, we generated bags consisting of 27 × 27 patches. A bag

was assigned a positive label if it contained one or more nuclei. Augmentation techniques were applied using the transformation function proposed by Sirinukunwattana et al. [12]. Fig. 2 provides a glimpse of the COLON CANCER Dataset utilized in our study. To ensure reliable and accurate performance, we conducted our experiments five times using a 10-fold cross-validation approach, consisting of one test fold and one validation fold. Specifically, we trained our model for up to 100 epochs for each fold, utilizing an improved version of the models that have demonstrated high success on multiple datasets. Three main dimensions, namely 64, 128, and 256, were measured due to their impact on the performance of our approach. The initialization technique proposed by [13] was applied to set the weights of the entire layers, while biases were set to zero. We employed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and utilized the cross-entropy loss function. The learning rate was set to $(5 \times 10^{-5})/(5 \times 10^{-6})$, and the weight decay was set to $(5 \times 10^{-4})/(1 \times 10^{-4})$. The model with the lowest validation loss was selected as the best model, and the best model from each fold was evaluated on the test set. Evaluation metrics including AUC, Accuracy, Precision, Recall, and F1-score were employed to compare the performance of our model. The experiments were conducted using Python3 with TensorFlow, and the computations were performed on an Nvidia GeForce RTX 2080 GPU.

2) Training Process of A-MIL: To achieve reliable and accurate performance, our experiment has been repeated five times and 10-fold cross-validation approach has been applied which includes one test fold and one validation fold. Precisely, our model trained up to 100 epoch for each fold, we use an improved version of models that has demonstrated high success on multiple datasets [12], [14], [15].

3) Transfer Learning Models Setup: For the transfer learning models, the dataset has two classes labeled as Benign and Malignant. The dataset has been split into training 80% for training set, 20% for validation set and 10% for testing set. As our datasets contain images of two class, we have used this training set to train a classifier to classify each of the classes. Ultimately, we measure classifier consistency by test it on our test images sets. Also, the dataset has been augmented and shuffled.

DenseNet201 has been used as pretrained weights which is already trained in the ImageNet competition. The learning rate was set to be 0.0001. Additionally, batch normalization has been applied and Softmax also has been used as the activation function, Adam as the optimizer and binary-cross-entropy as the loss function. VGG and ResNet models have followed the same preprocessing and training process that have been applied on DenseNet model implementation. Evaluation matrices have been used to evaluate the results of proposed models implementation including Accuracy, Sensitivity (SN), Specificity (SP) and AUC.

IV. Results Evaluation

A. A-MIL

This sections explaining the results and analysis of our experiment on our COLON CANCER dataset on both A-MIL model. Table II demonstrates that the evaluation scores of A-MIL implementation which can provide a deep explanation of A-MIL performance on our dataset. It can be seen that precision achieved the highest performance score with 0.900, followed by the accuracy scores at 0.888. F-score and AUC curve scores obtained the lowest values at 0.874 and 0.856 respectively. By checking the loss of training and validation sets, it is obvious that we were able to control the performance of the model during training.

TABLE II
Evaluation Matrices Scores Using A-MIL Model

Method	Accuracy	Precision	Recall	F1-Score	AUC
A-MIL	**0.90**	**0.91**	**0.88**	**0.88**	**0.817**

B. Transfer Learning Models

This section provides the results of Transfer Learning Models Implementation on colon cancer dataset.

1) DenseNet Model: After training our model, the results of the evaluation matrices that have been applied on our dataset can be clearly shown in Table III. The ROC curve achieved 0.857 score, this score approved the effectiveness of using the DenseNet Model.

TABLE III
Evaluation Matrices Scores Using DenseNet Model

Method	Accuracy	Precision	Recall	F1-Score	AUC
DenseNet	**0.86**	**0.87**	**0.87**	**0.87**	**0.85**

2) VGG Model: The VGG model has been applied using the preprocessing steps and training process that have been applied during DenseNet implementation. After training our model, the results of the evaluation matrices that have been applied to assess the performance of DenseNet model application on our dataset can be clearly shown in Table IV. The table illustrates that the model performed similarly on our evaluation matrices. It is noticeable that the accuracy of VGG model increased during the training process iteration increased. Receiver operating characteristic curve achieved 0.722 score.

TABLE IV
Evaluation Matrices Scores Using VGG Model

Method	Accuracy	Precision	Recall	F1-Score	AUC
VGG	**0.764**	**0.78**	**0.78**	**0.77**	**0.722**

3) ResNet Model: The results of the applied evaluation matrices that has been used to evaluate the performance of ResNet model application on colon cancer dataset can be clearly shown in Table V. Both accuracy and precision achieved 0.84, similarly recall and F1-score reached 0.83 and 0.82 respectively. Auc curve obtained 0.769 score, this score shows lower performance for ResNet in comparison to DenseNet.

TABLE V
EVALUATION MATRICES SCORES USING RESNET MODEL

Method	Accuracy	Precision	Recall	F1-Score	AUC
ResNet	**0.84**	**0.84**	**0.83**	**0.82**	**0.769**

4) Comparative Analysis between A-MIL Model and Transfer Learning Model: This section presents a comprehensive analysis comparing the performance of the A-MIL model with our benchmark Transfer Learning model. In summary, the A-MIL method demonstrates superior performance compared to the Transfer Learning (TL) models, particularly in terms of precision. The precision score for the A-MIL model is 0.91, which is significantly higher than the scores achieved by the TL models. Among the TL models, DenseNet achieves the highest precision score at 0.87, followed by ResNet with 0.84. The VGG pretrained model exhibits the lowest precision score, with a notably lower score of 0.78, Table VI.

TABLE VI
COMPARATIVE ANALYSIS BETWEEN A-MIL MODEL AND TRANSFER LEARNING MODE

Method	Accuracy	Precision	Recall	F1-Score	AUC
VGG	0.764	0.78	0.78	0.77	0.722
ResNet	0.84	0.84	0.83	0.82	0.769
DenseNet	0.86	0.87	0.87	0.87	0.857
A-MIL	**0.90**	**0.91**	**0.88**	**0.88**	**0.817**

V. CONCLUSION AND FUTURE WORK

This paper presents a comparative analysis between the Attention Multiple Instance Learning (A-MIL) model and various Transfer Learning models for the classification of colon cancer images. The performance of three pre-trained transfer learning models is evaluated and compared to the performance of the A-MIL model. The results of the study demonstrate that the application of A-MIL achieves superior performance in the classification task of colon cancer images, thereby validating the effectiveness of the A-MIL model compared to the application of transfer learning models. Future research endeavors will focus on further refining the A-MIL model to enhance its effectiveness and improve the accuracy of cancer classification tasks.

REFERENCES

[1] Thomas G. Dietterich a, Richard H. Lathrop band Lozano-Pérez, Tomás, "Solving the multiple instance problem with axis-parallel rectangles",Artificial intelligence,Elsevier. vol.89,pp. 31–71,1997.
[2] Maron, Oded and Lozano-Pérez, Tomás,"A framework for multiple-instance learning", Advances in neural information processing systems, Citeseer,pp.570–576,1998.
[3] Cosatto, Eric and Laquerre, Pierre-Francois and Malon, Christopher and Graf, Hans-Peter and Saito, Akira and Kiyuna, Tomoharu and Marugame, Atsushi and Kamijo, Ken'ichi," Automated gastric cancer diagnosis on h&e-stained sections; ltraining a classifier on a large scale with multiple instance machine learning", Medical Imaging 2013: Digital Pathology, International Society for Optics and Photonics, vol. 8676, pp.867–605, 2013.
[4] liu2012key,Liu, Guoqing and Wu, Jianxin and Zhou, Zhi-Hua,"Key instance detection in multi-instance learning" ,Asian Conference on Machine Learning,PMLR, pp.253–268,2012.
[5] Cheplygina, Veronika and Sørensen, Lauge and Tax, David MJ and de Bruijne, Marleen and Loog, Marco, "Label stability in multiple instance learning",International Conference on Medical Image Computing and Computer-Assisted Intervention, pp.539–546,Springer, 2015.
[6] Andrews, Stuart and Tsochantaridis, Ioannis and Hofmann, Thomas, "Support vector machines for multiple-instance learning",Advances in neural information processing systems,vol. 15, MIT; 1998,2003.
[7] Chen, Yixin and Bi, Jinbo and Wang, James Ze,"Multiple-instance learning via embedded instance selection",IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28,pp. 1931–1947, IEEE,2006.
[8] Ramon, Jan and De Raedt, Luc,"Multi instance neural networks",Proceedings of the ICML-2000 workshop on attribute-value and relational learning, pp. 53–60,2000.
[9] Raykar, Vikas C and Krishnapuram, Balaji and Bi, Jinbo and Dundar, Murat and Rao, R Bharat, "Bayesian multiple instance learning: automatic feature selection and inductive transfer",Proceedings of the 25th international conference on Machine learning, pp. 808–815,2008.
[10] Zhang, Cha and Platt, John and Viola, Paul,"Multiple instance boosting for object detection", Advances in neural information processing systems, vol. 18, pp. 1417–1424, Citeseer, 2005.
[11] Kandemir, Melih and Hamprecht, Fred A,"Computer-aided diagnosis from weak supervision: A benchmarking study", Computerized medical imaging and graphics, vol. 42, Elsevier, 2015.
[12] Sirinukunwattana, Korsuk and Raza, Shan E Ahmed and Tsang, Yee-Wah and Snead, David RJ and Cree, Ian A and Rajpoot, Nasir M,"Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images",IEEE transactions on medical imaging,vol. 35, pp. 1196–1206, IEEE, 2016.
[13] Glorot, Xavier and Bengio, Yoshua,"Understanding the difficulty of training deep feedforward neural networks",Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.
[14] Wang, Xinggang and Yan, Yongluan and Tang, Peng and Bai, Xiang and Liu, Wenyu,"Revisiting Multiple Instance Neural Networks", arXiv preprint arXiv:1610.02501, 2016.
[15] LeCun, Yann and Bottou, Léon and Bengio, Yoshua and Haffner, Patrick,"Gradient-based learning applied to document recognition",Proceedings of the IEEE,vol. 86, pp. 2278–2324, IEEE, 1998.