Triple-kernel Gated Attention-based Multiple Instance Learning with Contrastive Learning for Medical Image Analysis

Huafeng Hu[1], Ruijie Ye[2], Jeyan Thiyagalingam[3*], Frans Coenen[2] and Jionglong Su[4*]

[1]Department of Electrical and Electronic Engineering, University of Liverpool based at Xi'an Jiaotong-Liverpool University, Suzhou, 215123, Jiangsu, China.
[2*]Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, United Kingdom.
[3*]Scientific Computing Department, Science and Technologies Facilities Council, Harwell Campus, Oxford, OX11 0GD, United Kingdom.
[4*]School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, Jiangsu, China.

*Corresponding author(s). E-mail(s): t.jeyan@stfc.ac.uk; jionglong.su@xjtlu.edu.cn;
Contributing authors: huafeng.hu@xjtlu.edu.cn; sgrye2@liverpool.ac.uk; Coenen@liverpool.ac.uk;

Abstract

In machine learning, multiple instance learning is a method evolved from supervised learning algorithms, which defines a "bag" as a collection of multiple examples with a wide range of applications. In this paper, we propose a novel deep multiple instance learning model for medical image analysis, called triple-kernel gated attention-based multiple instance learning with contrastive learning. It can be used to overcome the limitations of the existing multiple instance learning approaches to medical image analysis. Our model consists of four steps. i) Extracting the representations by a simple convolutional neural network using contrastive learning for training. ii) Using three different kernel functions to produce the importance of each instance from the

entire image and forming an attention map. iii) Based on the attention map, aggregating the entire image together by attention-based MIL pooling. iv) Feeding the results to the classifier for prediction. The results on different datasets demonstrate that the proposed model outperforms state-of-the-art methods on binary and weakly supervised classification tasks. It can provide more efficient classification results for various disease models and additional explanatory information.

Keywords: Deep Learning, Multiple Instance Learning, Medical Image Analysis

1 Introduction

In machine learning, image classification typically assumes that all images are labeled with different classes. However, human pathological images may exhibit various disease characteristics in actual medical procedures, so we cannot simply assign a unique class to the whole image. This typical problem is called multiple instance learning (MIL), which was proposed by Dietterich et al. in 1997 [1]. It is a learning problem with a bag with multiple instances as the training unit. As most medical images have relatively high resolution and weakly labeled small datasets, the MIL method is a common method for medical image analysis [2]. Several research has been conducted in which the MIL method is applied to medical problems, such as drug activity prediction problem [1], dementia classification in brain MRI [3], and computer-aided detection (CAD) [4].

In recent years, with the rapid development of deep learning, the combination of MIL and neural network models has become a development trend [5]. Xu et al. first used a deep neural network as the feature extractor with the MIL algorithm as the classifier for medical image analysis [6]. Yousefi et al. proposed a framework to combine the CNN-based MIL with random forest to improve the performance for mass detection on breast data [7]. However, these researches are more of an attempt to combine CNN and MIL for medical image analysis that do not fully explain the underlying logic. Ilse et al. presented an attention-based strategy that improves the interpretability of MIL while also enhancing its flexibility [8]. Since then, the study of attention-based MIL has attracted much attention. Yao et al. proposed attention-based deep MIL for whole slide imaging classification [9]. In [10], an attention-based time-incremental CNN was proposed for achieving both spatial and temporal fusion of information from electrocardiogram for multi-class detection. Han et al. extended the attention-based deep MIL method to three-dimensional space for accurate screening of COVID-19 [11]. However, both methods require more data for their model training. In the case of some relatively rare disease, scarcity in the data present a challenge to the research. Rymarczyk et al. presented a kernel function on improving the performance of attention-based deep

MIL model on kinds of dataset [12]. However, the performance of their model is not stable with a reasonable explanation.

1.1 Motivations

Although some of the studies mentioned above have made significant progress in MIL methods, they all have shortcomings. The motivation of this paper is to overcome three existing limitations.

1. Diseased cells only occupy a part of the whole image for medical images. For example, breast cancer cells in the early stage usually cover less than five percent of the entire mammogram, which leads to a high imbalance in the proportion of examples in the positive bag, leading to misclassification of these positive bags by the model. In addition, the maximum pooling method is widely used in deep learning, and its characteristic of retaining only the largest value may lead to the lack of key information. In addition, due to the small data size of the medical image and under weak supervision, the model easily loses key features due to overfitting issues.
2. The current models commonly extract features from the given patch by CNN, such as ROI, because training traditional windows sliding feature extractors is very time-consuming and inefficient for high-resolution medical images. However, this simplified learning scheme may not obtain optimal features when classifying medical images.
3. The training process of the deep learning model is more like a black box, and the interpretation of the intermediate process is not outstanding. However, due to the particularity of medical images, doctors need more information to support subsequent diagnoses when using the model. Therefore, we need to explain the intermediate process further.

1.2 Contributions

This paper proposes a novel deep MIL model for medical image analysis called Triple-kernel Gated Attention-based MIL with contrastive learning (TGA-MIL). It is used to overcome the limitations of the existing MIL approach. The model consists of four steps. First, extracting the representations by a simple CNN model using contrastive learning for training. Second, using three different kernel functions to produce the importance of each instance from the entire image and form an attention map. Third, the attention map aggregates the entire image together by attention-based MIL pooling. Finally, feeding the results to the classifier for prediction. We use the TGA-MIL method on MNIST, two classical MIL datasets, and various medical image datasets, i.e., USBC breast cancer, colon cancer, and DDSM dataset, to test and show that it can be used for binary, multi-class, and weakly supervised classification tasks. This paper makes the following key contributions:

1. We propose a general framework called TGA-MIL for MIL problems, which combines three different kernels to generate an attention map. Compared to

state-of-the-art models, the results show that TGA-MIL outperforms other models in classification accuracy on different datasets. Moreover, we use contrastive learning for feature extraction in MIL. We successfully apply it to the MIL problem in the medical field;
2. We propose a novel concatenation of the representations from three kernels, i.e., Laplace (LA), Radial Basis Function (RBF), and Inverse Multiquadric (IM), to improve the representativeness of the features and optimize the weight of the attention map, as well as to improve the learning ability of the model for the properties of input data, which is finally manifested in the improvement of the classification results on five different datasets. We show that the concatenation of three different representations outperforms the traditional method of using three different representations as base learners for ensemble learning; and
3. We apply and optimize the gate attention-based MIL, and use the attention map in the model to interpret the training process for medical image analysis.

2 Related Work

2.1 Multiple Instance Learning

In machine learning, MIL is a method evolved from supervised learning algorithms, which defines a "bag" as a collection of multiple examples with a wide range of applications. [13]. Dietterich et al. completed one of the seminal studies in this subject [1]. Typically, MIL-based frameworks utilize either mean pooling or maximum pooling, with the latter being the more common. Both operators are non-trainable, which limits their capacity. Although MIL pooling operators with global adaptive parameters are widely used in many fields, their flexibility is limited [13].

Over the last 20 years, MIL has been effectively used in various areas, such as CAD [14], image classification [15], image segmentation [16], image annotation [17], object tracking [18], human action recognition [19], and interaction detection [20]. The challenge of diagnosing chronic obstructive pulmonary disease using breast CT also appears to have improved [21]. Jia et al. structured this goal as a MIL issue and created a weakly labeled histopathology image dataset to segment cancerous regions with weak supervision [22]. Most research focuses on the bag-level MIL scenario since building the instance-level classification method requires the true label of an instance and considers learning an optimal classification model for the target.

2.2 Deep MIL

Previous MIL research considered selected features to represent instances, hence additional feature extraction was unnecessary. However, new research into the use of fully-connected neural networks in MIL suggests that it may

still be advantageous [23]. Similarly, combining MIL with deep learning in computer vision enhances accuracy dramatically. Kraus et al. devised a method for classifying and segmenting microscope images using the Noisy-AND pooling function that combines deep CNNs with MIL [24]. Zhou et al. proposed using simply image-level annotation to diagnose diabetic retinopathy using a MIL approach with AlexNet [25]. However, in image classification, the reasonable use of attention-based methods to combine deep learning with MIL is more effective and illustrative [8].

2.3 Attention-based MIL

The purpose of embedding attention processes into deep learning is to mimic human brain activity by concentrating on a few crucial regions. Attention is responsible for several breakthroughs in natural language processing, notably the Transformer architecture [26]. The attention-based deep learning framework is a widely used embedding attention scheme. Pappas et al. sought to employ a network instead of a linear regression model to compute the attention weights on instances [27]. Qi et al. sought to classify, and segment point sets using the attention-based MIL operator [28]. Ilse et al. proposed two kinds of attention-based MIL operators to enhance the performance of neural networks [8]. This proposal is shown to outperform the max and mean operators. Furthermore, Han et al. proposed to apply the attention technique to 3D data with automated instance generation. All these studies motivate us to further research attention-based MIL.

2.4 Contrastive learning

Contrastive learning [29] is a self-supervised learning approach whose basic idea is to make base models perform certain auxiliary tasks based on temporal correspondence [30], and cross-modal consistency [31]. It achieves great success and attention in the field of machine learning. Contrastive representation learning has played a significant role in natural language processing in the past two decades. For example, in 2008, a two-class classification task with contrastive representation learning [32], was successful in determining whether and how the middle word of a context window is related to its context. Moreover, the Bidirectional Encoder Representation from Transformer (BERT) [26] model utilizes contrastive learning to extract bidirectional word representations with the Transformer architecture's decoder and distinguishes itself in multiple downstream tasks with transfer learning. It demonstrates the unique capability of contrastive learning to learn highly effective representations of original images [29]. There are many ways to construct auxiliary tasks with data augmentation, e.g., rotation prediction [33] and automatic colorization [34]. These auxiliary tasks are built to train new weights of a base neural network to extract features efficiently. The CT scan images of COVID-19 tend to be limited because many CT scan datasets are not sharable due to privacy concerns [35]. Besides, labeling images manually is time-consuming and requires

a lot of experience, making it an uphill task. Because of this, a self-supervised learning model is necessary in such cases. The application of self-supervised learning can enable a base neural network to learn feature representations more efficiently than those without it, allowing the size of datasets to be significantly increased by image augmentations. As a result, it can save much time for researchers in annotating medical image datasets.

With the development of contrastive self-supervised learning, there are now many popular methods, e.g., Momentum Contrastive (MoCo) [36], and Simple Framework of Contrastive Learning (SimCLR) [37]. MoCo focuses on building a consistent dictionary to speed up the learning process of contrastive learning. The SimCLR has larger batch sizes and extensive data augmentation, further facilitating the contrastive learning process [37]. Therefore, to explore how contrastive learning can positively affect medical image analysis, we attempt to apply this strategy to our medical image classification task. Moreover, Chaitanya et al. proposed a novel contrastive learning framework by leveraging domain-specific and problem-specific cues for medical image analysis [38]. They improved the performance of contrastive learning in dense prediction issues. Wu et al. proposed a new contrastive learning framework with a shared model by federated learning for medical image analysis [39]. The results showed that feature exchanges could be used to improve the labeling efficiency of medical images. Wang et al. sought to alleviate the limited labeling issue on the medical image analysis, and they proposed an uncertainty weighted integration method incorporating contrastive learning to extract representations [40]. Moreover, adversarial networks are also an alternative method to handle this issue. For example, Wang et al. proposed a 3D auto-context-based locality adaptive multi-modality generative adversarial networks for high quality medical image analysis, and the results showed their method could boost the training data with limited labels [41]. Luo et al. proposed adaptive rectification adversarial networks on this field [42]. In our research, we choose SimCLR to learn representations without manual labels.

3 Methodology

We propose a self-supervised image classification method. The whole framework is given in Figure 1. In this section, to make this work clearer, we describe related background formulas and introduce our model.

3.1 Multiple Instance Learning

The training set in MIL comprises multi-instance bags with classification labels, with each bag containing some instances without classification labels. A *positive bag* is defined as having at least one positive instance in a multi-instance bag. A *negative bag* is defined as having no positive instance in a bag. Multiple instance learning aims to build a multi-instance classifier by learning multi-instance bags with classification labels and applying the classifier to predict unknown multi-instance bags. The data unit of the MIL data set is the

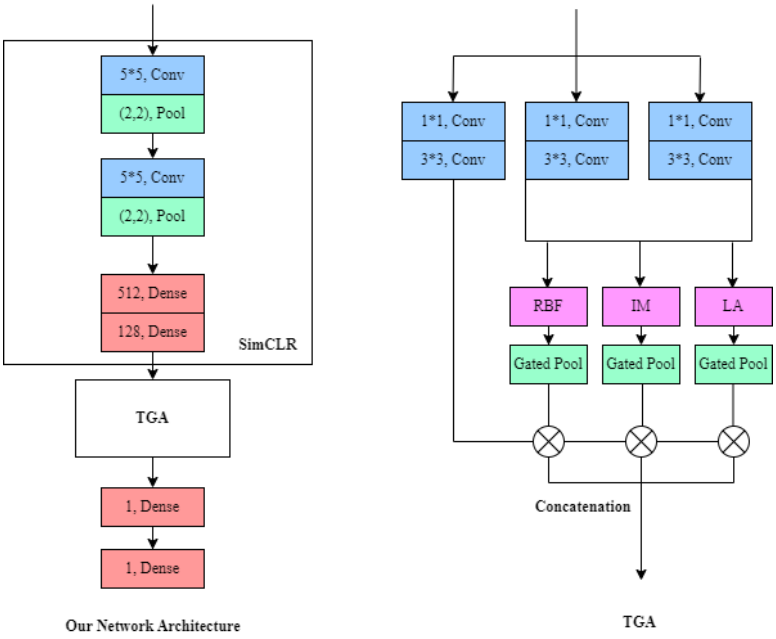


Fig. 1: The framework of our TGA-MIL.

bag. Taking the binary classification of MIL as an example, we assume each instance as $x \in \chi$ with a label $y \in \{0, 1\}$ which is unknown to the learner. Let $B = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a bag with label $c(B)$ given by

$$c(B) = \begin{cases} 0, & \text{iff } \sum y_i = 0 \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

This formula is only applicable in the case of using instance-level classifiers with a given label. However, each instance is a patch extracted from the original image in medical images. In actual situations, there is no given label for each instance. It is difficult to train a model that only learns to optimize the target based on the largest instance label in the real world. Since the labels of instances can be unknown in a weakly supervised task, there is a problem that the instance-level classifier may be undertrained. This leads to an increase in the number of misclassified cases.

The most common MIL approach is the embedding-based approach, which involves three steps in classifying a bag of instances [8]. First, obtaining a function f to extract the representations of instances. Second, designing a symmetric function to combine transformed instances. Finally, using a function g to modify combined instances. However, this approach is usually difficult to obtain key instances in improving the classification performance of the classifier. In this regard, an additional instance-level approach is introduced to provide an estimated score for obtaining key instances.

3.2 Self-training the CNN Feature Extractor Using Contrastive Learning

Since MIL is a weakly supervised problem, we use self-supervised contrastive learning to learn the feature extractor f. Specifically, we consider SimCLR from [37], a state-of-the-art self-supervised learning framework that learns robust representations without manual labels. SimCLR is a strategy whose auxiliary task mainly focuses on learning the efficient representations depending on the optimization of the reciprocal information between the extracted features from different random image augmentations of a single object. Our model considers image cropping, flipping, and Gaussian noise as image augmentation methods. The training process guarantees consistency between sub-images from the same image. Feature extractors obtain the representation of training samples for further classification tasks.

3.3 Attention-based MIL Pooling

Ilse et al. presented kinds of MIL pooling inspired by the instance-level approach to modify the existing embedding-level approach [8]. Before introducing our innovation part, we briefly describe the two schemes proposed in Ilse's article to illustrate our scheme better.

3.3.1 Attention Pooling

Attention-based MIL is an embedding-based MIL approach. It starts by mapping instances from a given bag X into a low-dimensional space to obtain their embeddings $H = \{\mathbf{h}_1, \ldots, \mathbf{h_k}\}, \mathbf{h_i} \in \mathbb{R}^M$. It performs the following MIL pooling to obtain a representation of the whole bag:

$$\mathbf{h}_{\text{bag}} = \sum_{i=1}^{k} a_i \mathbf{h_i}, \tag{2}$$

where:

$$a_k = \frac{\exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_k^\top\right)\right\}}{\sum_{j=1}^{k} \exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_j^\top\right)\right\}}, \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$ and $\mathbf{V} \in \mathbb{R}^{L\times M}$ are parameters and the $\tanh(\cdot)$ is used to prevent the gradient from exploding. This module can be used to obtain the similarity between instances. Moreover, the sum of the attention weight a_i is 1, and a bigger weight means a more significant impact of the instance on the classification.

3.3.2 Gated Attention Pooling

In addition, since $\tanh(x)$ is approximately linear at $x \in [-1, 1]$, its ability to learn complex relationships is limited, leading to a decrease in the representativeness of the extracted features. Therefore, Ilse et al. proposed to additionally use the gating mechanism together with $\tanh(\cdot)$ non-linearity that yields [8]:

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

where $\mathbf{U} \in \mathbb{R}^{L \times M}$ are parameters, $\odot$ is an element-wise multiplication and $\text{sigm}(\cdot)$ is the sigmoid function. Compared with $\tanh(\cdot)$, gated attention introduces nonlinear characteristics to overcome the limitations of linear equations.

The basic idea of attention-based MIL consists of four steps. First, CNN is used to obtain representations from each bag. Second, the attention or gated attention mechanism is used to produce the attention weights by the representations. Third, attention-based MIL pooling is used to obtain a vector for each bag. Finally, fully-connected layers are used to classify the vector for the results.

3.4 Gated Attention-based MIL Using Three Kernels

Inspired by the successful use of kernel function in SVM, Tsai et al. successfully applied an RBF-based formulation for the attention mechanism in Transformer on translation field [43]. Moreover, Kim et al. proposed LA kernel instead of the dot product in the image processing field [44]. However, the instability of the results makes the overall performance inferior to the dot product. Although they are either not used in the image domain or the results are not satisfactory, their concept makes us think that a different kernel function can be used instead of the dot product in the gated attention-based pooling, i.e., $\odot$ in Equation 4. In our study, we use the previously described RBF and LA kernels, but also discuss the IM kernel that is widely used in SVM. Their formulas are as follows.

$$\text{LA}: \ k(v, u) = -\left\|v - u\right\|_1 \tag{5}$$

$$\text{RBF}: \ k(v, u) = \exp\left(-\frac{\left\|v - u\right\|_2^2}{2\sigma^2}\right) \tag{6}$$

$$\text{IM}: \ k(v, u) = \frac{1}{\sqrt{\left\|v - u\right\|^2 + c^2}} \tag{7}$$

where σ and c are trainable parameters. RBF can approximate any nonlinear function with arbitrary precision and has global approximation capability. The convergence speed is fast, and the learning generalization ability of the corresponding attention map is improved. However, since the performance of RBF

depends on the choice of the center of the data points, it leads to the instability of performance. LA kernel overcomes the limitations of the central dependency issue in RBF kernel. However, because it is a parameter-free kernel, we cannot fine-tune it during the training process. IM kernel is an improved version of RBF, which is used to neutralize the unstable nature of RBF. In summary, dot-product attention displays non-smooth predictions. We use triple kernels to help smooth out the interpolations and combine their strengths to improve the performance of our model.

We consider the instability of the kernel function in [44] and the problem of the limited amount of medical image data. Therefore, unlike Equation 2, we concatenate and transpose a_k generated by the three kernels. Afterward, we concatenate the three identical $\mathbf{h_k}$ and feed them to gated attention-based MIL pooling.

This method is similar to ensemble learning, so we compare it to the typical stacking method in subsequent comparative experiments, which combines data sets with multiple base learners and generates a new meta-model [45]. The specific process of stacking is used with 3 base learners, i.e., gated attention-based MIL with RBF, IM, and LA kernel, respectively.

4 Experiments

In our experiments, we evaluate the efficacy of our method using many different datasets as follows. Five classical MIL benchmark datasets, Musk1, Musk2, Fox, Tiger, Elephant [1]; an MNIST-based image dataset [46]; three medical datasets, USBC breast cancer [47], colon cancer [48], and DDSM [49]. We employ a standard assessment approach, 10-fold cross-validation, and five repeats in Musk1, Musk2, and the MNIST-based dataset to achieve a fair comparison. For consistency on the DDSM, we use the same experimental method from [50]. To compare the performance between different methods, we use metrics which includes the classification accuracy, precision, recall, F-score, and AUC. For computations, our models are implemented by Tensorflow and trained on the GTX1080Ti.

4.1 Musk1, Musk2, Fox, Tiger, and Elephant

4.1.1 Experimental Settings

In the first experiments, we will test our method against other deep MIL methods on five classical benchmark datasets, i.e., Musk1, Musk2, Fox, Tiger, and Elephant. Musk1 and Musk2 are used to identify whether a medication molecule will attach to a target protein. A positive molecule has at least one form that can bind well, whereas a negative molecule has no shapes that can bind well. In MIL contexts, this problem may be expressed fairly naturally: each molecule would be a bag, and the possible conformations would be instances in that bag [1]. Fox, Tiger, and Elephant contain features extracted from corresponding animal images. These datasets are made up of extracted

feature vectors from instances and do not need the learning of a feature extractor. Because the characteristics have already been established, the experiment involves directly feeding the feature to three kernel functions for predicting attention maps without contrastive learning.

4.1.2 Results

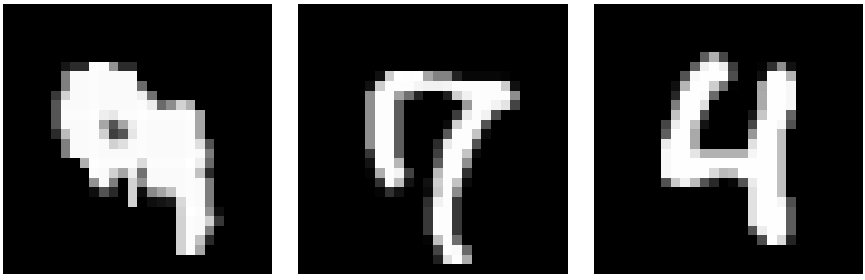
Experiments are repeated five times, each using 10-fold cross-validation to compare our TGA-MIL to other current designs on the MIL issue, as given in Table 1. The results show that our TGA-MIL surpasses the state-of-the-art models on four datasets except for Fox. Meanwhile, on the Fox dataset, our TGA-MIL also obtains the fourth-highest results. This shows that our method is more efficient.

Table 1: Results on classical MIL datasets. Experiments were repeated five times, with the average classification accuracy (±standard error) provided. The best results for each dataset are highlighted in bold.

Methods	Musk1	Musk2	Fox	Tiger	Elephant
mi-Net [51]	0.889 ± 0.039	0.858 ± 0.049	0.613 ± 0.035	0.824 ± 0.034	0.858 ± 0.037
MI-Net [51]	0.887 ± 0.041	0.859 ± 0.046	0.622 ± 0.038	0.830 ± 0.032	0.862 ± 0.034
MI-Net with DS [51]	0.894 ± 0.042	0.874 ± 0.043	0.630 ± 0.037	0.845 ± 0.039	0.872 ± 0.032
MI-Net with RC [51]	0.898 ± 0.043	0.873 ± 0.044	0.619 ± 0.047	0.836 ± 0.037	0.873 ± 0.044
Attention [8]	0.892 ± 0.040	0.858 ± 0.048	0.615 ± 0.043	0.839 ± 0.022	0.868 ± 0.022
Gated Attention [8]	0.900 ± 0.050	0.863 ± 0.042	0.603 ± 0.029	0.845 ± 0.018	0.857 ± 0.027
mi-Net Attention [52]	0.900 ± 0.063	0.870 ± 0.048	0.630 ± 0.026	0.845 ± 0.028	0.865 ± 0.024
ELDB [53]	0.902 ± 0.016	0.857 ± 0.039	$\mathbf{0.648 \pm 0.014}$	0.767 ± 0.013	0.843 ± 0.012
TGA-MIL (ours)	$\mathbf{0.910 \pm 0.033}$	$\mathbf{0.881 \pm 0.040}$	0.628 ± 0.020	$\mathbf{0.846 \pm 0.015}$	$\mathbf{0.875 \pm 0.020}$

4.2 MINST-based Dataset

4.2.1 Experimental Settings



(a) (b) (c)

Fig. 2: Sample images that are easily misclassified (a) “9”, (b) “7”, (c) “4”.

Representations in the classical MIL benchmark datasets have been pre-extracted, so there are limitations in the measurement of classification performance. To demonstrate the capacity of our approach in an experiment that is both classical and more challenging, we turn our attention to the MNIST dataset in the second experiment. To fairly compare the capabilities of our TGA-MIL method with the original attention-based MIL methods, we carry out the same processing as [8] on the MNIST dataset. As shown in Figure 2, the MNIST dataset is easy to misclassify the images of "9", "7", and "4". A bag is created by selecting a random number of 28×28 grayscale images from the MNIST dataset. We define positive bag to be one that contains at least one image "9". In the test set, we use a fixed number of 100 bags. For comparison, we follow the CNN architecture according to [8], called LeNet 5 without contrastive learning [54]. The optimal hyperparameters are shown in Table 2. We also apply data augmentation, e.g., random rotations, random cropping, and horizontal and vertical flipping. In the experiments, we design a random positive number with 10 as the mean and 1 as the variance for each bag. The integer closest to this random number is the number of instances in the bag. Besides, we use varying numbers of training bags, i.e., 50, 100, 150, 200, 250, 300. Using these settings, we test how varying the number of training bags and instances will affect MIL models. Since our training data is randomly selected, it is easy to produce a high degree of imbalance between positive and negative samples. Therefore, in this experiment, we only use AUC, which is less sensitive to the imbalance of positive and negative samples, to compare the classification performance between different models.

Table 2: The hyperparameters for MNIST-based dataset

Optimizer	β_1, β_2	Learning rate	Maximum of Epochs	Selection criteria
Adam	$0.9, 0.999$	0.0001	50	lowest loss

4.2.2 Results

Table 3: MNIST-based dataset with a different number of training bags. Experiments were repeated five times, with the average AUC (±standard error) provided. The best results for different numbers of training bags are highlighted in bold.

Number of Training bags	50	100	150	200	250	300
Max-pooling	0.531 ± 0.063	0.701 ± 0.092	0.940 ± 0.003	0.957 ± 0.001	0.970 ± 0.001	0.972 ± 0.001
Mean-pooling	0.611 ± 0.053	0.627 ± 0.083	0.925 ± 0.007	0.964 ± 0.004	0.969 ± 0.001	0.970 ± 0.001
Attention [8]	0.727 ± 0.043	0.901 ± 0.005	$\mathbf{0.955 \pm 0.006}$	0.970 ± 0.002	0.969 ± 0.001	0.976 ± 0.001
Gated Attention [8]	0.733 ± 0.041	$\mathbf{0.906 \pm 0.008}$	0.945 ± 0.001	0.974 ± 0.002	0.977 ± 0.001	0.975 ± 0.002
TGA-MIL (ours)	$\mathbf{0.753 \pm 0.034}$	0.900 ± 0.020	0.950 ± 0.001	$\mathbf{0.975 \pm 0.001}$	$\mathbf{0.980 \pm 0.002}$	$\mathbf{0.983 \pm 0.002}$

The results of AUC for MNIST-based dataset are presented in Figure 3 and Table 3. The findings of the experiment are given as follows,

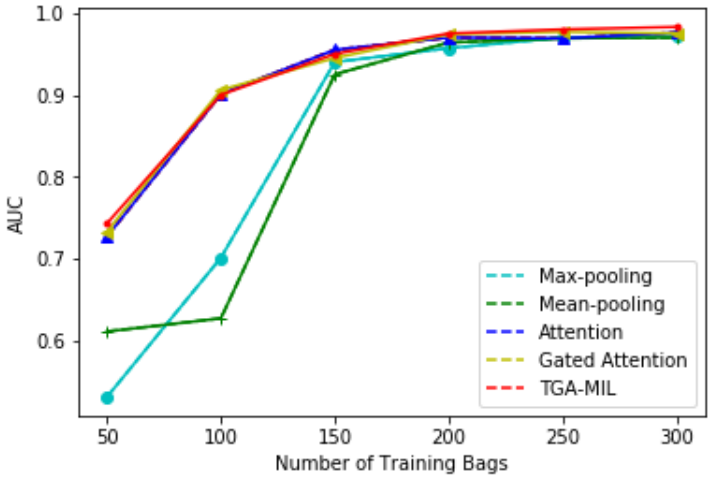


Fig. 3: Results for MNIST-based dataset with different number of training bags.

1. When the number of training sets is small (only 50 bags), the stability of all methods is relatively low (the variance is the largest). Our method increases the number and stability of representations through different kernels, which increases the AUC performance by at least 2% compared to other methods and significantly reduces the variance;
2. When the number of data set is moderate (100 and 150), our method does not obtain the best AUC results, but the gap with the best method is about 0.5%;
3. When the number of data sets is relatively large (200, 250, and 300 bags), the performance of all methods on the MNIST-based dataset tends to be stable, and the results are close. This is because the data set is relatively basic and not challenging. However, our method can further improve the maximum performance of the original method through three different kernels and obtain the highest AUC; and
4. Figure 4 gives the difference between our TGA-MIL and the attention weights generated by attention-based MIL and gated attention-based MIL. We can obtain that when "4", "7", "9" appears simultaneously in our method, the attention weights corresponding to "9" are enlarged, while the attention weights corresponding to "4" and "7" are relatively reduced. When there are only "7", its corresponding attention weights are still stable and will not be ignored by the model.

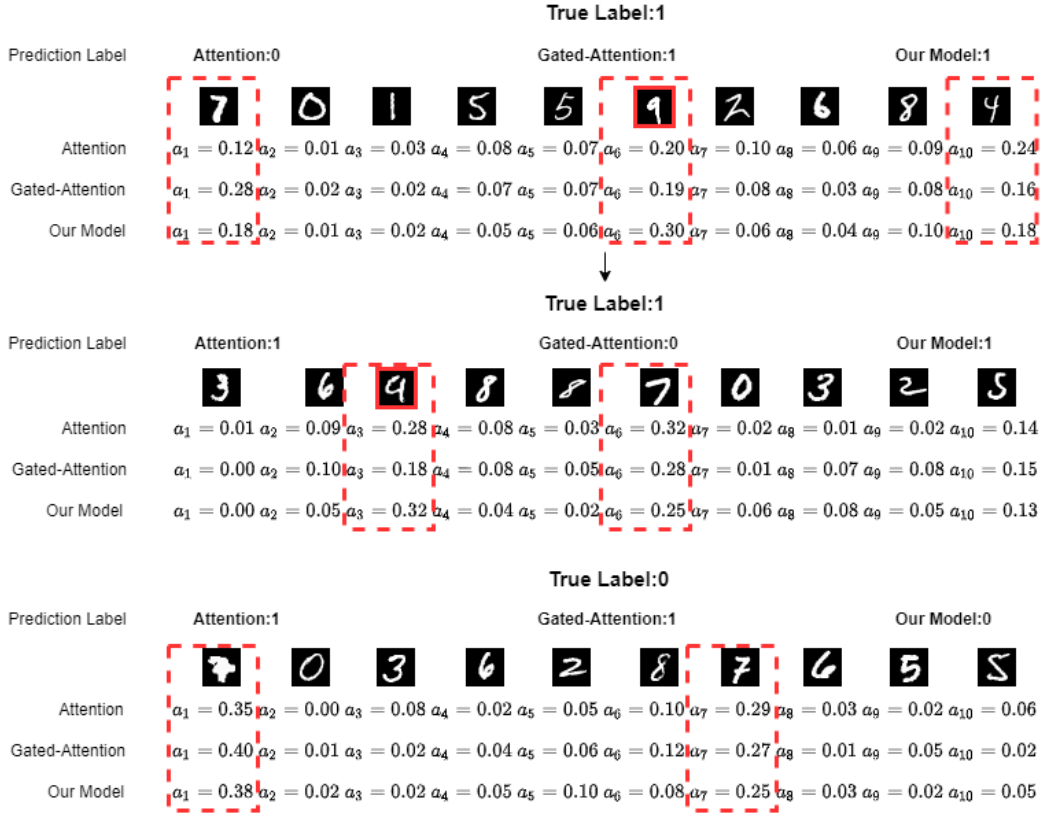


Fig. 4: Examples of different models with corresponding attention weights and prediction labels with 300 bags.

4.3 USBC Breast Cancer and Colon Cancer Datasets

4.3.1 Experimental Settings

The automatic identification of malignant areas in entire images stained with Hematoxylin and Eosin (H&E) is a popular research task. Current supervised methods utilize pixel-level annotations [55]. However, the preparation of large amounts of H&E data requires pathologists to spend much time, which is difficult to achieve in real life. Therefore, solutions using WSI will reduce the workload of pathologists. In this experiment, we test our method in classifying two weakly-labeled histopathology images of the breast cancer dataset from USBC [47] and the colon cancer dataset [48]. The description of each dataset is given as follows:

The USBC breast cancer dataset contains 58 H&E images with weakly labels, each measuring 896×768. If a photo contains breast cancer cells, it is classified as malignant; otherwise, it is classified as benign. Every image is divided into 32×32 patches and 672 patches per bag. We remove the patch which has 75% or more white pixels.
Colon cancer dataset contains 100 H&E images. The images are derived from various tissue appearances in both normal and cancerous areas. The majority of nuclei in each cell were indicated in each picture. There are four classes of nuclei in the dataset, including epithelial, inflammatory, fibroblast, and miscellaneous nuclei. A bag consists of patches with the resolution of 27×27. Furthermore, epithelial cell tagging is important from a therapeutic standpoint since epithelial cells are the source of colon cancer. Therefore, if a bag includes one or more epithelial nuclei, it is assigned a positive label.

We train the model weights on both datasets using the Adam optimizer with a constant learning rate of 0.0001. For MIL model training, a mini-batch size of 1 is used. SimCLR is used to train the feature extractor using patches derived from the training sets of the datasets. We utilize the Adam optimizer for SimCLR, with a min-batch size of 128 and an initial learning rate of 0.0001. ResNet is the CNN backbone used in MIL models and SimCLR. Specifically, for SimCLR, we use data augmentations, including random cropping, horizontal/vertical flipping, and random zoom. Warmup, fine-tuning, and end-to-end training take 60, 20, and 20 epochs, respectively. 10-fold cross-validation with one validation fold and one test fold is repeated five times. We have designed several experimental models with corresponding abbreviations for comparisons, as given in Table 4.

4.3.2 Results

We present results in Table 5 and Table 6 for USBC breast and colon cancer, respectively. The findings of two histological datasets are as follows,

Abbreviations	Experimental design
GA-RBF	Gated attention-based MIL with RBF kernel
GA-IM	Gated attention-based MIL with IM kernel
GA-LA	Gated attention-based MIL with LA kernel
S-AGR	Stacking with attention, gated attention and GA-RBF
S-AGI	Stacking with attention, gated attention and GA-IM
S-AGL	Stacking with attention, gated attention and GA-LA
S-RIL	Stacking with GA-RBF, GA-IM, and GA-LA

Table 4: The description of abbreviations with corresponding experimental design.

1. Our method obtain the highest value in comparing the five metrics of the two data sets, especially for the two most important indicators for medical images, accuracy, and recall. These two indicators fully show that our algorithm can still achieve higher performance than other algorithms on classical MIL datasets and data in the medical field;
2. We achieve at least 1.0% improvement in classification accuracy compared to the baseline method on the USBC breast cancer. In addition, compared to the other experimental group we designed, at least an improvement of 0.6% is achieved. In the comparative experiment, one kernel function is improved by about 1% relative to the baseline model. This is enough to demonstrate that the kernel function in our design is conducive to improving the selection effect of the attention map, and the participation of SimCLR and concatenation methods has better performance than the general stacking method; and
3. The localization performance indicates the capability of different models to delineate positive instances. Heat maps of different models from the USBC breast dataset are illustrated in Figure 5. It can be seen in the figure that compared to the two baseline methods, the heat map generated by our TGA-MIL increases the weights of the corresponding instances in the ground truth and significantly reduces the weights corresponding to the external non-key instances. It is sufficient to demonstrate that our model can enable the model to pay more attention to the key instances, learn more realistic and effective representations, and improve classification performance. This approach is very conducive to reducing the number of false negatives and can also be used to explain why our method achieves the highest recall.

Table 5: Results on USBC breast cancer dataset. Experiments were repeated five times, with the average (±standard error) provided. **Note:** The abbreviations in the table have been described in Table 4. The best results for each metric are highlighted in bold.

Methods	accuracy	Precision	Recall	F-score	AUC
Max-pooling	0.609 ± 0.018	0.594 ± 0.021	0.449 ± 0.097	0.516 ± 0.063	0.608 ± 0.028
Mean-pooling	0.738 ± 0.021	0.730 ± 0.021	0.661 ± 0.051	0.659 ± 0.027	0.806 ± 0.008
Attention [8]	0.738 ± 0.019	0.711 ± 0.020	0.728 ± 0.037	0.700 ± 0.030	0.785 ± 0.019
Gated Attention [8]	0.747 ± 0.016	0.719 ± 0.015	0.730 ± 0.022	0.718 ± 0.020	0.793 ± 0.023
mi-Net Attention [52]	0.750 ± 0.020	0.722 ± 0.020	0.725 ± 0.020	0.711 ± 0.022	0.790 ± 0.030
ELDB [53]	0.760 ± 0.018	0.720 ± 0.018	0.735 ± 0.029	0.721 ± 0.032	0.800 ± 0.028
GA-RBF	0.751 ± 0.014	0.716 ± 0.012	0.748 ± 0.021	0.725 ± 0.018	0.793 ± 0.020
GA-IM	0.749 ± 0.013	0.729 ± 0.013	0.743 ± 0.023	0.721 ± 0.019	0.779 ± 0.020
GA-LA	0.737 ± 0.018	0.731 ± 0.020	0.747 ± 0.020	0.712 ± 0.021	0.768 ± 0.025
S-AGR	0.757 ± 0.014	0.740 ± 0.014	0.760 ± 0.020	0.721 ± 0.018	0.801 ± 0.017
S-AGI	0.758 ± 0.013	0.742 ± 0.011	0.750 ± 0.015	0.732 ± 0.018	0.823 ± 0.020
S-AGL	0.756 ± 0.013	0.725 ± 0.017	0.758 ± 0.012	0.725 ± 0.017	0.813 ± 0.020
S-RIL	0.764 ± 0.011	$\mathbf{0.758 \pm 0.015}$	0.763 ± 0.010	0.737 ± 0.009	0.840 ± 0.009
TGA-MIL (ours)	$\mathbf{0.770 \pm 0.010}$	0.756 ± 0.011	$\mathbf{0.768 \pm 0.008}$	$\mathbf{0.742 \pm 0.018}$	$\mathbf{0.831 \pm 0.007}$

Table 6: Results on colon cancer dataset. Experiments were repeated five times, with the average (±standard error) provided. **Note:** The abbreviations in the table have been described in Table 4. The best results for each metric are highlighted in bold.

Methods	accuracy	Precision	Recall	F-score	AUC
Max-pooling	0.810 ± 0.013	0.870 ± 0.014	0.783 ± 0.019	0.821 ± 0.019	0.910 ± 0.009
Mean-pooling	0.832 ± 0.012	0.867 ± 0.011	0.754 ± 0.030	0.813 ± 0.015	0.902 ± 0.008
Attention [8]	0.900 ± 0.009	0.946 ± 0.013	0.851 ± 0.009	0.902 ± 0.010	0.959 ± 0.008
Gated Attention [8]	0.890 ± 0.010	0.950 ± 0.015	0.840 ± 0.029	$\mathbf{0.899 \pm 0.022}$	0.955 ± 0.009
mi-Net Attention [52]	0.900 ± 0.015	0.952 ± 0.011	0.850 ± 0.035	0.870 ± 0.025	0.951 ± 0.015
ELDB [53]	0.915 ± 0.012	0.951 ± 0.010	0.855 ± 0.027	0.878 ± 0.025	0.978 ± 0.010
GA-RBF	0.894 ± 0.012	0.914 ± 0.010	0.825 ± 0.026	0.871 ± 0.017	0.963 ± 0.007
GA-IM	0.902 ± 0.010	0.917 ± 0.008	0.807 ± 0.023	0.892 ± 0.014	0.969 ± 0.008
GA-LA	0.872 ± 0.009	0.920 ± 0.009	0.786 ± 0.030	0.792 ± 0.035	0.953 ± 0.021
S-AGR	0.906 ± 0.008	0.944 ± 0.008	0.832 ± 0.015	0.887 ± 0.012	0.972 ± 0.010
S-AGI	0.902 ± 0.007	0.924 ± 0.010	0.867 ± 0.014	0.877 ± 0.011	0.973 ± 0.008
S-AGL	0.886 ± 0.008	0.923 ± 0.007	0.794 ± 0.026	0.813 ± 0.012	0.965 ± 0.015
S-RIL	0.915 ± 0.009	0.938 ± 0.013	0.865 ± 0.010	0.876 ± 0.012	0.978 ± 0.007
TGA-MIL (ours)	$\mathbf{0.927 \pm 0.010}$	$\mathbf{0.955 \pm 0.015}$	$\mathbf{0.881 \pm 0.018}$	0.886 ± 0.018	$\mathbf{0.983 \pm 0.009}$

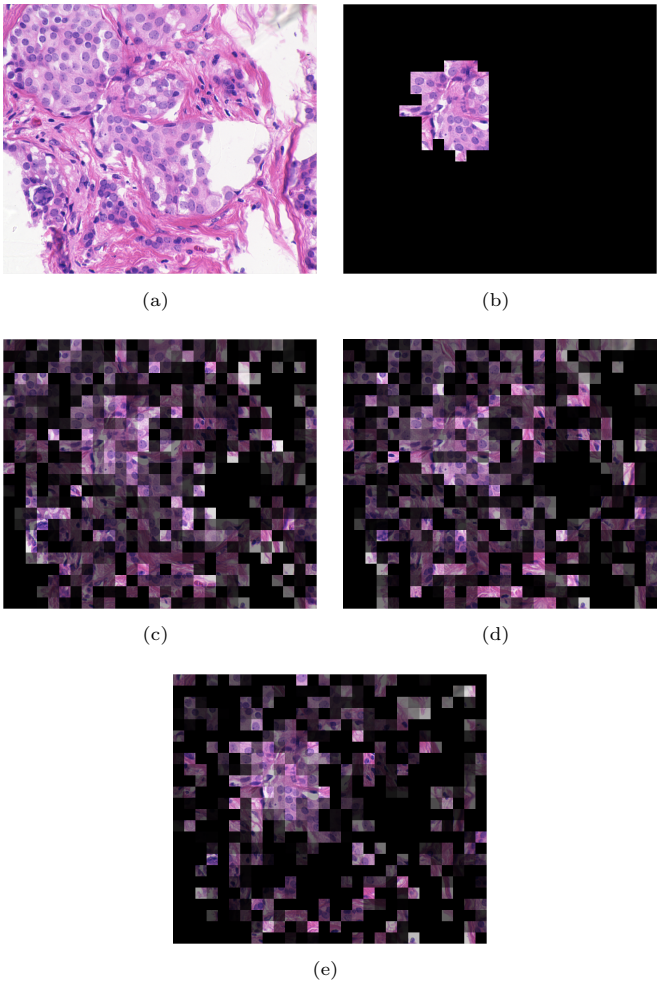


Fig. 5: An example of different methods generates a heat map comparison based on the attention map for USBC breast cancer dataset. Note that the attention weight is normalized to [0,1] and multiplied by each instance to produce the corresponding heat map. (a) Original image from USBC breast cancer dataset, (b) Ground truth instances from given labels, (c) Heat map from attention-based MIL, (d) Heat map from gate attention-based MIL, (e) Heat map from TGA-MIL.

4.3.3 Ablation Study

In our ablation study, we study the impact of using different numbers of kernels on the performance of these two datasets. As Table 7 demonstrates, the performance of three kernels outperforms others on three metrics, i.e., accuracy, F-score, and AUC. Meanwhile, all metrics on three kernels obtain the lowest

standard errors. Therefore, three kernels are the most stable model with the best performance. In Table 8, the model of three kernels performs best on all metrics. As the results show, three kernels would be the most suitable model.

Table 7: Ablation study on USBC breast cancer dataset. Experiments were repeated five times, with the average (±standard error) provided. **Note:** k stands for the number of kernels with their names in parentheses. The best results for each metric are highlighted in bold.

Methods	accuracy	Precision	Recall	F-score	AUC
$k = 1$ (RBF)	0.751 ± 0.014	0.716 ± 0.012	0.748 ± 0.021	0.725 ± 0.018	0.793 ± 0.020
$k = 1$ (IM)	0.749 ± 0.013	0.729 ± 0.013	0.743 ± 0.023	0.721 ± 0.019	0.779 ± 0.020
$k = 1$ (LA)	0.737 ± 0.018	0.731 ± 0.020	0.747 ± 0.020	0.712 ± 0.021	0.768 ± 0.025
$k = 2$ (RBF+IM)	0.762 ± 0.021	0.751 ± 0.020	0.755 ± 0.030	0.738 ± 0.027	0.829 ± 0.030
$k = 2$ (IM+LA)	0.758 ± 0.020	0.759 ± 0.019	0.745 ± 0.021	$\mathbf{0.743 \pm 0.022}$	0.819 ± 0.023
$k = 2$ (RBF+LA)	0.765 ± 0.024	$\mathbf{0.760 \pm 0.024}$	0.757 ± 0.030	0.741 ± 0.025	0.808 ± 0.010
k =3 (RBF+IM+LA)	$\mathbf{0.770 \pm 0.010}$	0.756 ± 0.011	$\mathbf{0.768 \pm 0.008}$	$\mathbf{0.742 \pm 0.018}$	$\mathbf{0.831 \pm 0.007}$

Table 8: Results on colon cancer dataset. Experiments were repeated five times, with the average (±standard error) provided.

Methods	accuracy	Precision	Recall	F-score	AUC
k =1 (RBF)	0.894 ± 0.012	0.914 ± 0.010	0.825 ± 0.026	0.871 ± 0.017	0.963 ± 0.007
k =1 (IM)	0.902 ± 0.010	0.917 ± 0.008	0.807 ± 0.023	0.892 ± 0.014	0.969 ± 0.008
k =1 (LA)	0.872 ± 0.009	0.920 ± 0.009	0.786 ± 0.030	0.792 ± 0.035	0.953 ± 0.021
k =2 (RBF+IM)	0.889 ± 0.020	0.907 ± 0.021	0.842 ± 0.030	0.860 ± 0.023	0.955 ± 0.012
k =2 (IM+LA)	0.920 ± 0.015	0.945 ± 0.015	0.865 ± 0.021	0.882 ± 0.015	0.976 ± 0.014
k =2 (RBF+LA)	0.918 ± 0.008	0.937 ± 0.004	0.879 ± 0.023	0.872 ± 0.030	0.957 ± 0.017
k =3 (RBF+IM+LA)	$\mathbf{0.927 \pm 0.010}$	$\mathbf{0.955 \pm 0.015}$	$\mathbf{0.881 \pm 0.018}$	$\mathbf{0.886 \pm 0.018}$	$\mathbf{0.983 \pm 0.009}$

4.4 DDSM

4.4.1 Experimental Settings

In this experiment, we use a public dataset called DDSM [49]. This public dataset consists of 2620 digitized film-screen screening mammograms with pixel-level ground truth annotation for tumors [49]. Each mammogram includes two standard projections, the CC view and the mediolateral oblique MLO view, along with localization information. Specialists supplied the localization information stored in DDSM. We use the mammogram images from Lumisys scanner, which has the highest resolution in DDSM as our whole dataset. The subset of DDSM has 666 images in the benign class and 657 images in the malignant class [50]. In the experiment, without cross-validation, we randomly split the whole dataset into a training set, a validation set, and a test set according to proportions of 80%, 10%, and 10%, respectively. For this experiment, each image from DDSM is cropped into 224×224 instances without overlapping to form a bag. The hyperparameters of base model are shown in Table 9. The SimCLR is also used for our TGA-MIL with the initial parameters for feature extraction by pre-trained on ImageNet.

Table 9: The hyperparameters for DDSM dataset

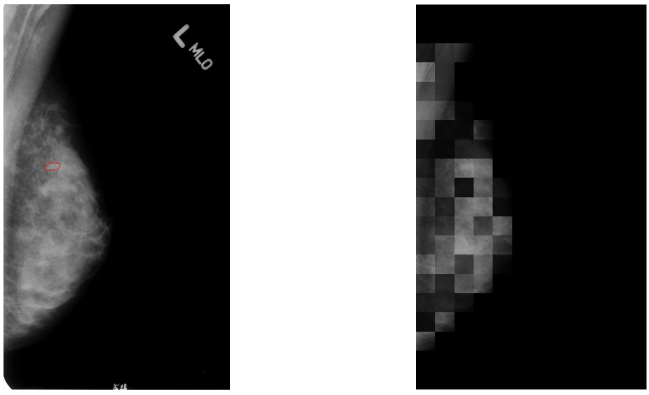
Optimizer	β_1, β_2	Learning rate	Maximum of Epochs	Batch size
Adam	$0.9, 0.999$	0.0001	50	1 (bag)

4.4.2 Results

The sensitivity of each method is given in Table 10. It is not difficult to see that the previous algorithm has been outdated. Compared to the previously proposed model, the original two attention-based MIL algorithms or our newly proposed TGA-MIL algorithm have made considerable progress. Even if the previous algorithm label is instance-based, and we only have a bag-based label, our new algorithm still increases sensitivity by 1.1%. Moreover, unlike previous algorithms, TGA-MIL can provide more attention to the key instances for the model, thereby reducing the time consumption while improving the performance of the algorithm in the sliding windows method. In Figure 6, we can see that the external boundary can be ignored without manually removing the black instance, and the areas that may have cancerous cells are automatically highlighted.

Table 10: The overall detection performance (malignant vs. benign) of our method and other state-of-the-art methods. The best result is highlighted in bold.

Algorithms	Sensitivity
K-means and SVM [56]	83%
Cascaded Deep Learning and Random Forests [57]	77.2%
ANN [58]	75.9%
Feed Forward Neural Network [59]	74.6%
Extreme Learning Machine [60]	81.8%
Faster-RCNN [61]	71.2%
CNN-based Framework [50]	85.2%
Attention [8]	86.2%
Gated Attention [8]	86.4%
mi-Net Attention [52]	86.7%
ELDB [53]	85.8%
TGA-MIL (ours)	**87.8**%



(a) (b)

Fig. 6: An example of DDSM dataset with corresponding heat map by our TGA-MIL. Note that the attention weight is normalized to [0,1] and multiplied by each instance for producing the correspond heat map. (a) Original image from DDSM dataset, the ground truth is surrounded by the red circle, (b) Heat map from TGA-MIL.

5 Conclusions and Future Work

This paper presents a novel MIL approach for medical image analysis, called triple-kernel gated attention-based multiple instance learning with contrastive learning (TGA-MIL). In contrast to gated attention-based MIL approach, it use SimCLR for initial CNN parameters instead of pre-trained from ImageNet

and concatenate three different kernels, LA, RBF, and IM, for extracting representations. The experiments on nine datasets (Musk1, Musk2, Fox, Tiger, Elephant, MNIST-based dataset, USBC breast cancer dataset, colon cancer dataset, DDSM dataset) confirm that our method is on par or outperforms the current state-of-the-art methodology based on various metrics. In contrast, our method uses the attention map to focus on more representative parts, thus solving the problem of insufficient labels. This overcomes the limitation that the whole image cannot be used as input data. Also, the performance using the whole image is close to that of using only the ROI, which illustrates the practicality of our method. Finally, unlike previous algorithms like black boxes, TGA-MIL can provide more attention to the key instances for the model, thereby reducing the time consumption while improving the performance of the algorithm in the sliding windows method.

Future research can be carried out in two aspects. First, we applied the method of contrastive learning to perform self-supervised learning to overcome the adverse effects of unlabeled instances. However, we directly use the SimCLR method in this part. In the future, we will design contrastive learning that is more in line with medical images to replace SimCLR and improve the practicality of the model in the medical field. Second, we use the heat map generated according to the attention weight to explain which parts of the model will be more concentrated when used to understand the progress of the model. However, for medical images, there may be further developed, such as how the representation generated by the feature extractor affects the subsequent formation so that the doctor can better understand the internal use mechanism of the model when using it.

Acknowledgments. Our previous article [62] has been selected as one of the best papers in PIC-2021, following which we are invited to submit this work Applied Intelligence. We would like to thank PIC-2021 for giving us this opportunity to be recommended. This work was supported in part by the Key Program Special Fund at Xi'an Jiaotong-Liverpool University (KSF-A-22).

Declarations

Conflict of interests. The authors declare that they have no conflict of interest.

Data Availability Statement. The data will be made available on reasonable request.

References

[1] Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997)

[2] Carbonneau, M.-A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition **77**, 329–353 (2018)

[3] Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J.V., Rueckert, D., Initiative, A.D.N., *et al.*: Multiple instance learning for classification of dementia in brain mri. Medical image analysis **18**(5), 808–818 (2014)

[4] Chen, Y., Bi, J., Wang, J.Z.: Miles: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12), 1931–1947 (2006)

[5] Dimitriou, N., Arandjelović, O., Caie, P.D.: Deep learning for whole slide image analysis: an overview. Frontiers in medicine **6**, 264 (2019)

[6] Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Eric, I., Chang, C.: Deep learning of feature representation with multiple instance learning for medical image analysis. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1626–1630 (2014). IEEE

[7] Yousefi, M., Krzyżak, A., Suen, C.Y.: Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. Computers in biology and medicine **96**, 283–293 (2018)

[8] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136 (2018). PMLR

[9] Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis **65**, 101789 (2020)

[10] Yao, Q., Wang, R., Fan, X., Liu, J., Li, Y.: Multi-class arrhythmia detection from 12-lead varied-length ecg using attention-based time-incremental convolutional neural network. Information Fusion **53**, 174–182 (2020)

[11] Han, Z., Wei, B., Hong, Y., Li, T., Cong, J., Zhu, X., Wei, H., Zhang, W.: Accurate screening of covid-19 using attention-based deep 3d multiple instance learning. IEEE transactions on medical imaging **39**(8), 2584–2594 (2020)

[12] Rymarczyk, D., Borowa, A., Tabor, J., Zieliński, B.: Kernel self-attention in deep multiple instance learning. arXiv preprint arXiv:2005.12991 (2020)

[13] Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. Advances in neural information processing systems, 570–576 (1998)

[14] Fung, G., Dundar, M., Krishnapuram, B., Rao, R.B.: Multiple instance learning for computer aided diagnosis. Advances in neural information processing systems **19**, 425 (2007)

[15] Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: ICML, vol. 98, pp. 341–349 (1998). Citeseer

[16] Wu, J., Zhao, Y., Zhu, J.-Y., Luo, S., Tu, Z.: Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 256–263 (2014)

[17] Yang, C., Dong, M., Hua, J.: Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp. 2057–2063 (2006). IEEE

[18] Babenko, B., Yang, M.-H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE transactions on pattern analysis and machine intelligence **33**(8), 1619–1632 (2010)

[19] Yi, Y., Lin, M.: Human action recognition with graph-based multiple-instance learning. Pattern Recognition **53**, 148–162 (2016)

[20] Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 28–35 (2012). IEEE

[21] Cheplygina, V., Sørensen, L., Tax, D.M., Pedersen, J.H., Loog, M., de Bruijne, M.: Classification of copd with multiple instance learning. In: 2014 22nd International Conference on Pattern Recognition, pp. 1508–1513 (2014). IEEE

[22] Jia, Z., Huang, X., Eric, I., Chang, C., Xu, Y.: Constrained deep weak supervision for histopathology image segmentation. IEEE transactions on medical imaging **36**(11), 2376–2388 (2017)

[23] Wu, J., Yu, Y., Huang, C., Yu, K.: Deep multiple instance learning for image classification and auto-annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3460–3469 (2015)

[24] Kraus, O.Z., Ba, J.L., Frey, B.J.: Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics **32**(12), 52–59 (2016)

[25] Zhou, L., Zhao, Y., Yang, J., Yu, Q., Xu, X.: Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images. IET Image Processing **12**(4), 563–571 (2018)

[26] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[27] Pappas, N., Popescu-Belis, A.: Multilingual hierarchical attention networks for document classification. arXiv preprint arXiv:1707.00896 (2017)

[28] Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660 (2017)

[29] Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: A framework and review. IEEE Access (2020)

[30] Li, X., Liu, S., De Mello, S., Wang, X., Kautz, J., Yang, M.-H.: Joint-task self-supervised learning for temporal correspondence. arXiv preprint arXiv:1909.11895 (2019)

[31] Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.-F., Wang, W.Y., Zhang, L.: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6629–6638 (2019)

[32] Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, pp. 160–167 (2008)

[33] Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)

[34] Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European Conference on Computer Vision, pp. 649–666 (2016). Springer

[35] Bai, H.X., Hsieh, B., Xiong, Z., Halsey, K., Choi, J.W., Tran, T.M.L., Pan, I., Shi, L.-B., Wang, D.-C., Mei, J., *et al.*: Performance of radiologists

in differentiating covid-19 from non-covid-19 viral pneumonia at chest ct. Radiology **296**(2), 46–54 (2020)

[36] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)

[37] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607 (2020). PMLR

[38] Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in Neural Information Processing Systems **33**, 12546–12558 (2020)

[39] Wu, Y., Zeng, D., Wang, Z., Shi, Y., Hu, J.: Federated contrastive learning for volumetric medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 367–377 (2021). Springer

[40] Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y.: Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. Medical Image Analysis **79**, 102447 (2022)

[41] Wang, Y., Zhou, L., Yu, B., Wang, L., Zu, C., Lalush, D.S., Lin, W., Wu, X., Zhou, J., Shen, D.: 3d auto-context-based locality adaptive multi-modality gans for pet synthesis. IEEE transactions on medical imaging **38**(6), 1328–1339 (2018)

[42] Luo, Y., Zhou, L., Zhan, B., Fei, Y., Zhou, J., Wang, Y., Shen, D.: Adaptive rectification based adversarial network with spectrum constraint for high-quality pet image synthesis. Medical Image Analysis **77**, 102335 (2022)

[43] Tsai, Y.-H.H., Bai, S., Yamada, M., Morency, L.-P., Salakhutdinov, R.: Transformer dissection: A unified understanding of transformer's attention via the lens of kernel. arXiv preprint arXiv:1908.11775 (2019)

[44] Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., Teh, Y.W.: Attentive neural processes. arXiv preprint arXiv:1901.05761 (2019)

[45] Wolpert, D.H.: Stacked generalization. Neural networks **5**(2), 241–259 (1992)

[46] Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine **29**(6), 141–142 (2012)

[47] Gelasca, E.D., Byun, J., Obara, B., Manjunath, B.: Evaluation and benchmark for biological image segmentation. In: 2008 15th IEEE International Conference on Image Processing, pp. 1816–1819 (2008). IEEE

[48] Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.-W., Snead, D.R., Cree, I.A., Rajpoot, N.M.: Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE transactions on medical imaging **35**(5), 1196–1206 (2016)

[49] Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, W.P.: The digital database for screening mammography. In: Proceedings of the 5th International Workshop on Digital Mammography, pp. 212–218 (2000). Medical Physics Publishing

[50] Hu, H., Coenen, F., Ma, F., Thiyagalingam, J., Su, J.: Location-aware convolutional neural networks based breast tumor detection (2018)

[51] Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition **74**, 15–24 (2018)

[52] Yi, J., Zhou, B.: Attention awareness multiple instance neural network. arXiv preprint arXiv:2205.13750 (2022)

[53] Yang, M., Zhang, Y.-X., Wang, X., Min, F.: Multi-instance ensemble learning with discriminative bags. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2021)

[54] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., *et al.*: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

[55] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)

[56] Martins, O., Braz Junior, G., Corrêa Silva, A., Cardoso de Paiva, A., Gattass, M., *et al.*: Detection of masses in digital mammograms using k-means and support vector machine. ELCVIA: electronic letters on computer vision and image analysis **8**(2), 039–50 (2009)

[57] Dhungel, N., Carneiro, G., Bradley, A.P.: Automated mass detection in mammograms using cascaded deep learning and random forests. In: 2015

International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8 (2015). IEEE

[58] Bellotti, R., De Carlo, F., Tangaro, S., Gargano, G., Maggipinto, G., Castellano, M., Massafra, R., Cascio, D., Fauci, F., Magro, R., *et al.*: A completely automated cad system for mass detection in a large mammographic database. Medical physics **33**(8), 3066–3075 (2006)

[59] Delogu, P., Fantacci, M.E., Kasae, P., Retico, A.: Characterization of mammographic masses using a gradient-based segmentation algorithm and a neural classifier. Computers in Biology and Medicine **37**(10), 1479–1491 (2007)

[60] Wang, Z., Yu, G., Kang, Y., Zhao, Y., Qu, Q.: Breast tumor detection in digital mammography based on extreme learning machine. Neurocomputing **128**, 175–184 (2014)

[61] Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. Scientific reports **8**(1), 4165–4171 (2018)

[62] Zhang, S., Zou, B., Xu, B., Su, J., Hu, H.: An efficient deep learning framework of covid-19 ct scans using contrastive learning and ensemble strategy. In: 2021 IEEE International Conference on Progress in Informatics and Computing (PIC), pp. 388–396 (2021). IEEE