

# Key Patches Are All You Need: A Multiple Instance Learning Framework For Robust Medical Diagnosis

D. J. Araújo<sup>1</sup>, M. R. Verdelho<sup>1</sup>, A. Bissoto<sup>2</sup>, J. C. Nascimento<sup>1</sup>, C. Santiago<sup>1</sup>, C. Barata<sup>1,3</sup>

<sup>1</sup>Institute for Systems and Robotics, LARSyS, Instituto Superior Técnico, Portugal

<sup>2</sup>Institute of Computing, Recod.ai Lab, University of Campinas, Brazil <sup>3</sup>Lisbon ELLIS Unit

diogoparaujo@tecnico.ulisboa.pt

## Abstract

*Deep learning models have revolutionized the field of medical image analysis, due to their outstanding performances. However, they are sensitive to spurious correlations, often taking advantage of dataset bias to improve results for in-domain data, but jeopardizing their generalization capabilities. In this paper, we propose to limit the amount of information these models use to reach the final classification, by using a multiple instance learning (MIL) framework. MIL forces the model to use only a (small) subset of patches in the image, identifying discriminative regions. This mimics the clinical procedures, where medical decisions are based on localized findings. We evaluate our framework on two medical applications: skin cancer diagnosis using dermoscopy and breast cancer diagnosis using mammography. Our results show that using only a subset of the patches does not compromise diagnostic performance for in-domain data, compared to the baseline approaches. However, our approach is more robust to shifts in patient demographics, while also providing more detailed explanations about which regions contributed to the decision. Code is available at: <https://github.com/diogoja99/Medical-Multiple-Instance-Learning>.*

## 1. Introduction

Deep learning (DL) architectures revolutionized the field of medical image analysis, achieving performances that rival even those of more experienced clinicians. It is undeniable that DL models can extract relevant and sometimes new information from medical data. However, there is still a high degree of uncertainty associated with the information that is being used by these models and whether it maps to actual (novel) concepts, or if the models are identifying spurious correlations and taking advantage of dataset bias [2, 9]. Thus, in order to really leverage DL systems in healthcare, it is necessary to ensure that these models are simultaneously

explainable and able to achieve good performances outside the datasets they were trained on.

The evolution in the DL field has led to the proposal of different ways for extracting information from images. In this scope, convolutional neural networks (CNNs) are still the most common architectures in medical image analysis. However, in recent years, vision transformers (ViTs) have also gained popularity [1, 4]. CNNs and ViTs adopt different feature extraction paradigms: CNNs explore the local neighborhood, while ViTs are able to capture the image context by using self-attention blocks that leverage spatial information and distant relationships. ViTs also adopt an explicit patch-based strategy, as opposed to the traditional full-image analysis performed by CNNs. Nevertheless, both architectures end up learning patch-based representations that are then aggregated into a single representation vector (e.g., through the global average pooling operation in CNNs and the class token in ViTs).

Patch or region-based analysis resembles clinical practice for medical image inspection, where doctors search for localized findings and criteria to perform a diagnosis. However, contrary to DL models, clinicians do not need to process all regions in a medical image, as they are able to automatically identify the key regions that match a malignant diagnosis. An example is the 7-point checklist method used in skin image analysis [16]. This approach focuses solely on the presence or absence of certain dermoscopic features within the lesion, regardless of their spatial arrangement. Another example is breast cancer, where radiologists identify and classify findings in mammography.

The development of DL models capable of identifying regions of interest (ROIs) in medical images and using only those regions to perform a diagnosis is a promising line of research. On one hand, these methods are more aligned with clinical practice. Additionally, showing the ROIs grants some measure of explainability to the model. On the other hand, by forcing the model to use only a part of the image to perform a diagnosis, we can: i) improve its robustness to bias, as the information that the model can use is limited and

thus it must select the most discriminative one; and ii) identify spurious correlations learned by the model (*e.g.*, one or more ROIs matching artifacts instead of clinical findings).

The multiple instance learning (MIL) framework, commonly used in weakly-supervised problems, emerges as a natural direction to enforce CNNs and ViTs to look for ROIs. Under the MIL framework, an image is considered a ‘bag’, and each patch within the image is an ‘instance’. The classification of the entire image depends on the presence or absence of ‘key instances’, where we can limit their number to be small, forcing the model to make a decision with less information.

In this paper, we explore the benefits of incorporating a MIL framework on top of the feature extraction procedures of both CNNs and ViTs. Using as test beds two medical problems (skin and breast cancer) we show that MIL can be easily integrated into the pipelines of both CNNs and ViTs and that it can be used to select the most relevant patches for both approaches, reducing the amount of information used by the classifier. The models that use MIL achieve competitive performances against the standard CNNs and ViTs, showing that discriminative information is localized in a (small) subset of image regions. Moreover, by identifying these regions, we can provide the user with explanations for the model’s decision. Surprisingly, we also observed that, by using MIL, we obtain diagnostic systems that generalize better to new datasets, with different distributions and characteristics than those used for model training.

The rest of our paper is organized as follows. In Section 2, we discuss the application of patch selection methods in medical image analysis and how MIL-based approaches can be used in this context. Our approach, which focuses on incorporating various MIL frameworks after the feature extraction pipeline of a CNN or ViT, is discussed in Section 3. The experimental setup and the results are described in Sections 4 and 5. Finally, our conclusions and findings are summarized in Section 6.

## 2. Related Work

Most state-of-the-art classification models for medical image analysis are either based on CNNs or ViTs [1, 4]. While these architectures are conceptually different, both can be viewed as extractors of patch-level features. These features are then aggregated into a single vector that represents the entire image. Pooling operators, in particular global average pooling, are usually used for CNNs, while ViTs integrate the information of all patches into the class token. In the end, the representation is fed to an MLP head that performs the binary or multi-class classification. This means that, when training a model, we are allowing it to explore all the available information to define the decision boundaries. The drawback is that the model can learn spurious correlations and use them to achieve higher performances

during training and in-domain validation [9, 30].

Attention blocks, in particular spatial ones, can be seen as a mechanism to reduce the amount of image information used by the models [10], as they act as patch selectors. Despite their popularity, spatial attention blocks are not sparse (apart from a few exceptions [20]), which means that all regions in the image end up contributing to the decision. Moreover, they consist of additional layers of parameters to be learned end-to-end, increasing the model complexity, and their placement in the architecture is not trivial.

A particular type of attention is self-attention used by ViTs [8]. Here, multi-head self-attention (MSA) blocks [28] are used to extract complex features by leveraging patch correlations. Self-attention also affects the class embedding that will be fed to the classifier. However, while the visualization of the output of the MSA blocks allows the identification of relevant patches, they are all used to build the class embedding. To overcome this issue and reduce the amount of information, Liang *et al.* [19] introduced the Expediting ViT (EViT), which progressively discards less relevant patches. The EViT model uses attention scores to determine the significance of each patch towards the model’s output. The  $k$  most relevant patches are categorized as “attentive”, while the remaining ones are deemed “inattentive” and are subsequently merged into a single embedding. EViT showcases a promising direction for information selection in ViTs. However, the ratio between attentive and inattentive patches must be empirically defined by the user.

MIL-based frameworks have been explored in the field of medical image analysis to process high-resolution images, such as those of histopathology [3, 13, 17, 24]. MIL is particularly well suited in this context, as often we only have access to image level labels (*e.g.*, tumor staging), but the relevant information is localized in a small portion of the image that we want to identify. To achieve this goal, the original image is partitioned into big patches that are then independently processed by a CNN and aggregated in the end, using a variety of strategies such as max or attention pooling [13] and transformers [24].

It is clear that MIL frameworks are explainable, as they highlight relevant patches in an image. Moreover, when certain operators are used (max or top-k pooling) they can be seen as a proxy to a spatial attention module that does not require the learning of additional model parameters. However, the application of MIL in the medical domain has two limitations: i) it is often applied to binary classification problems, while several medical problems are multi-class (*e.g.*, skin image analysis); and ii) by dividing the image into patches and processing each one independently, we may be losing relevant features. Regarding the latter issue, we propose to apply MIL only after the feature extraction processes of CNNs and ViTs. This allows us to select relevant patches and reduce the amount of information used by

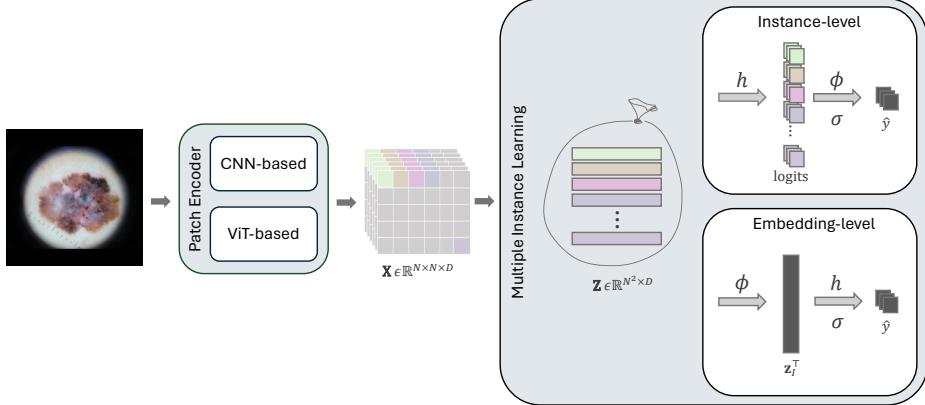


Figure 1. Overview of the proposed approach. An encoder block (CNN or ViT-based) extracts patch representations from the input image. Each patch will be an instance of a bag. Then, a MIL block determines the bag label using an *instance* or *embedding-level* approach.

the classifier, while still exploring the capabilities of these two architectures to extract information from images. For the multi-class problem, there have been some attempts to extend MIL to this setting [18, 22, 29], without definitive results. In this work, we propose a generalized MIL formulation for a multi-class problem and show that the binary problem is a particular case of this formulation.

### 3. Proposed Approach

Our approach, illustrated in Fig. 1, relies on MIL strategies to obtain the image classification using only a subset of its patches. The input image is processed with an encoder block that extracts patch features. Then, a MIL block predicts the classification of the image (bag), based on those patch (instance) representations. The following sections describe each block in detail.

#### 3.1. Patch Encoder Block

The first component of our approach is a patch encoder block. This block is responsible for generating the representation vector of each of the  $N \times N$  patches in the image. Two types of encoders may be adopted: CNNs, where patch representations correspond to each pixel of the output feature map; or ViTs, where patch representations correspond to the final representation of each patch token.

##### 3.1.1 CNN Encoder

Most popular CNNs follow the same type of architecture, consisting of a sequence of convolutional blocks (with convolutional, pooling, and normalization layers), followed by a classification head. The convolutional blocks process the image using small kernels that extract low-level to high-level features from each region in the image. Their output is a  $N \times N$  feature map,  $\mathbf{X} \in \mathbb{R}^{N \times N \times D}$ , with  $N$  much smaller than the size of the original image, illustrated in Fig. 1.

Due to the convolution operations, each pixel in this feature map can be interpreted as the representation of a patch (receptive field) in the image. Typically, the feature map is then transformed using a global average pooling, resulting in a representation vector for the entire image that is the input to the classification head. The underlying premise of this step, however, is that the relevant information for the classification task is spread across the entire image.

To avoid the above premise, we discard the global average pooling and treat each pixel in the feature map as an instance for the MIL classifier. Concretely, we assume that the feature map  $\mathbf{X}$  contains the representation of all  $N \times N$  patches in the image.

##### 3.1.2 ViT Encoder

ViT-based architectures use the multi-head self-attention (MSA) mechanism [27] to extract complex features based on patch correlations in images while taking into account positional information. The input image is first transformed into a sequence of  $N^2$  patches. Then, these patches are processed by several linear projections and MSA layers.

Each MSA consists of running several self-attention mechanisms in parallel on a sequence comprising the patch representations and an additional class token. The resulting attention maps hold information regarding the pairwise similarities between patches. Effectively, each MSA layer modifies the patches and the class token representations by combining the information contained in the entire sequence through weighted averages.

In the standard ViT, the final image classification is obtained by applying an MLP head to the class token, which harnesses information from all patches. To avoid this, we discard the class token and use the final patch representations, denoted as  $\mathbf{X} \in \mathbb{R}^{N \times N \times D}$ , as input to the MIL block. This means that each patch representation captures

the global context of the image, unlike in the CNN encoder, where they only depend on their local neighborhood.

### 3.2. MIL Block

The MIL block aims to apply a MIL classifier to the patch representations obtained by the patch encoder block. The tensor  $\mathbf{X}$  is first flattened to a matrix  $\mathbf{Z} \in \mathbb{R}^{N^2 \times D}$ , which is our collection of instances, represented as a bag in Fig. 1. The  $j$ -th line vector in  $\mathbf{Z}$ ,  $\mathbf{z}_j \in \mathbb{R}^{1 \times D}$ , is the representation of the  $j$ -th patch in the input image. The proposed MIL classifier consists of three key operations:

- $h$ , a linear projection function that maps the input from an embedding of dimension  $D$  to the logits with dimension  $C$  (the number of classes), given by  $h(\mathbf{Z}) = \mathbf{W}\mathbf{Z}^\top + \mathbf{b}$ , where  $\mathbf{W} \in \mathbb{R}^{C \times D}$  and  $\mathbf{b} \in \mathbb{R}^C$  is a bias term;
- $\phi$ , a permutation-invariant pooling that aggregates the patch-level information into image-level by applying a top- $k$  average, where  $k = 1$  leads to max pooling and  $k = N^2$  leads to average pooling; and
- $\sigma$ , a non-linear activation function.

The order of these three functions,  $\{h, \phi, \sigma\}$ , determines the specific MIL approach used: *instance-level* or *embedding-level*, as detailed in the following sections.

This formulation is a generalization of the classical MIL approach for binary problems. However, it should be emphasized that the binary case represents a special case of our formulation, where we set  $C = 1$  and  $\sigma$  is the sigmoid function. On the other hand, in a multi-class problem,  $C > 2$  and  $\sigma$  is the softmax function.

#### 3.2.1 Instance-level Approach

The *instance-level* approach is characterized by performing class predictions on each patch. It can be implemented in two different modes, although in both cases the first step is to apply  $h$  to the patch representations. This projects the patch features into a new embedding space of dimension  $C$ , corresponding to the patch logits.

For the first *instance-level* mode (I-1), the second step is to apply  $\sigma$ , which converts the logits obtained in the previous step into class probabilities for each patch. Then, in the final step, the pooling function  $\phi$  is applied, resulting in the predicted class probabilities  $\hat{y} \in [0, 1]^C$ .

The second *instance-level* mode (I-2) reverses the order of these two steps. It applies the pooling  $\phi$ , followed by  $\sigma$ , to convert the pooled logits into probabilities.

Notice that in the special case of a binary problem, I-1 and I-2 lead to the same classification result, even though they estimate different probabilities for the two classes. Therefore, we only show results with I-1 for this setting.

#### 3.2.2 Embedding-level Approach

The *embedding-level* approach starts by aggregating patch representations with the pooling function,  $\phi$ . This leads to a new vector,  $\mathbf{z}_I \in \mathbb{R}^{1 \times D}$ , representing image-level features. Only then are these features transformed to logits with the linear projection  $h$ . Finally, the  $\sigma$  operator converts the logits to the predicted class probabilities,  $\hat{y} \in [0, 1]^C$ .

When  $\phi$  is the average pooling, this approach reverts back to the standard CNN strategy of applying a global average pooling before the classification head.

## 4. Experimental Setup

We evaluated the performance of the proposed approach in two medical image classification problems: skin cancer diagnosis in dermoscopy and breast cancer diagnosis in mammography. For each of these settings, we trained a set of baselines (standard CNNs and ViTs models) and our MIL approaches described in Section 3. In order to compare with a recent approach that also performs patch selection, we trained various EViT [19] with different keep rate values for the attentive patches. Our results are all evaluated in terms of class recall (R) and balanced accuracy (BA - the average of the recalls). Below, we describe the adopted datasets, as well as the training specifications.

### 4.1. Datasets

**Skin Cancer.** For dermoscopy image analysis, we address two main challenges: binary and multi-class classification. The ISIC 2019 dataset [5, 6, 26] is our primary dataset, which we partition into a training (80%) and validation (20%) sets. For the binary problem, the training phase consisted of using only the melanoma (MEL) and nevi (NV) classes from the ISIC 2019. We also evaluated the generalization capabilities of the proposed approach in several out-of-domain datasets: HIBA [15], PH<sup>2</sup> [21], and Derm7pt [14]. Each of the previous datasets contains images collected from patients of different demographic groups, allowing us to do a preliminary assessment of the fairness of the different models.

For the multi-class classification task, we employed the ISIC 2019 dataset [5, 6, 26] for training and validation purposes, while the HIBA dataset [15] served as our testing ground. These datasets encompass eight diagnostic categories: Actinic keratosis (AK), Basal cell carcinoma (BCC), Benign keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic nevus (NV), Squamous cell carcinoma (SCC), and Vascular lesion (VASC). Table 1 provides a detailed overview of the class distributions for the training, validation, and test datasets for both binary and multi-class scenarios.

**Breast Cancer.** For mammography image analysis, we evaluated our proposal on the binary task of distinguishing

Table 1. Summary of the overall distribution of the training, validation, and testing dermoscopic image datasets.

Classes	ISIC 2019		HIBA	PH <sup>2</sup>	Derm7pt
	Train	Val.	Test	Test	Test
AK	687	173	46	—	—
BCC	2,653	664	228	—	—
BKL	2,089	525	62	—	—
DF	191	48	39	—	—
MEL	3,611	904	194	40	252
NV	10,293	2,575	549	160	575
SCC	502	126	111	—	—
VASC	202	51	41	—	—
<b>Total</b>	20,228	5,066	1,270	200	827

breasts with findings from those with no findings. We employed the DDSM dataset [11, 12] for both training (90%) and validation (10%). Specifically, the training dataset comprised 2,428 cases identified with findings and 1,342 cases without findings. For validation, we evaluated 260 cases with findings against 137 cases without findings.

**Preprocessing.** All input images were resized to a uniform size of  $224 \times 224 \times 3$ . Mammography images were converted from grayscale to RGB by replicating the color channel. To preserve the original aspect ratio of both the dermoscopy and mammography images, we applied padding to ensure that all images had a square format.

## 4.2. Training Setup

**Encoder Block.** We explored a variety of CNN-based pre-trained backbones for the patch encoder block, encompassing ResNet-18 (RN-18), ResNet-50 (RN-50), VGG-16, DenseNet-169 (DN-169), and EfficientNetB3 (EN-B3). Additionally, we explored ViT models – DEiT-S and EViT-S with a keep rate ( $K_r$ ) of 0.7. Every model was pre-trained on the ImageNet1k dataset [7].

For ViT-based encoders, images were partitioned into  $14 \times 14$  patches. To match these resolutions in the CNN experiments, we collected the CNN feature maps with a spatial dimension of  $14 \times 14$ . Among the evaluated CNN backbones, EN-B3 emerged as the best model, leading to the creation of the MIL-EN-B3 model with 2.2M parameters. From the transformer variants, DEiT-S was chosen, forming the MIL-DEiT-S model with 22M parameters. Comparison between the various backbones can be seen in Supplementary Material.

For the *instance-level* approach, we conducted experiments with three MIL pooling operators: the *max* operator, the *average* operator, and the *top-k average* operator, with three different configurations:  $k \approx 12.5\%$ ,  $k = 25\%$ , and  $k = 50\%$ . Our experiments showed that  $k = 25\%$  was generally the best representation of the *top-k average* pool-

ing operator. In the case of the *embedding-level* approach, we used the following MIL pooling operators: the *column-wise global max* pooling operator, the *column-wise global average* pooling operator, and the *column-wise global top-k average* pooling operator, with  $k = 25\%$  (different  $k$  values were tested and can be seen in Supplementary Material).

**Baseline Models.** The baseline models for our experiments used the same backbones as the MIL models described above. Here, however, we use the full architecture, replacing only the classification layer with one specific to our medical problems.

**EViT Baselines.** We adopted the EViT-S configuration with 22.1M parameters as the standard for comparing our method. In all configurations, we placed the token reorganization block in three different layers: the 3<sup>rd</sup>, the 6<sup>th</sup>, and the 9<sup>th</sup> layers. The keep rate ( $K_r$ ) determines the number of attentive tokens retained by the token reorganization block. We explored different settings and settled on  $K_r = 0.6$  and  $K_r = 0.7$ . With these choices, the EViT model with  $K_r = 0.6$  preserves 43 patches, while the model with  $K_r = 0.7$  retains 68 patches out of 196 patches. The assessment of additional  $K_r$  values can be seen in the Supplementary Material.

**Training Configurations.** All models were trained using a class-weighted categorical Cross Entropy (CE) loss function since all datasets are highly unbalanced. Online augmentation strategies tailored to each task were used in order to enhance model robustness. Specifically, for dermoscopy image classification, we adopted the augmentation configuration outlined by Touvron *et al.* [25]. In contrast, the mammography image classification task incorporated random horizontal and vertical flips, along with random rotations. All tested models were implemented and trained using PyTorch on NVIDIA GeForce RTX 3090 and 4090.

## 5. Experimental Results

The experimental results for the binary problems can be seen in Tables 2 and 4, where the latter table corresponds to the generalization experiments. The multi-class results are shown in Table 3. In the following subsections, we discuss the experimental results as well as the visualizations obtained with our MIL models.

### 5.1. Binary MIL

Our experimental results for the binary classification of dermoscopy and mammography images are summarized in Table 2. These results show that there is a marginal difference between the MIL models and their baseline counterparts. For the ISIC 2019 set, the standard deviation for BA stands at a modest 1.60%, and for the DDSM dataset, an even smaller standard deviation of 0.49% is observed. In terms of backbones, the one based on DEiT-S achieves better performances in the case of skin cancer, suggesting that

Table 2. Results for the binary problem in dermoscopy and mammography.

Models	ISIC 2019			DDSM		
	BA	R-MEL	R-NV	BA	R-F	R-NoF
EN-B3	90.7	85.5	95.8	96.1	94.6	97.6
DEiT-S	91.7	86.7	<b>96.7</b>	95.3	<b>95.0</b>	95.6
EViT	Kr = 0.6	91.4	86.6	96.3	<b>96.2</b>	94.6
	Kr = 0.7	90.7	85.4	95.7	95.8	93.1
MIL-EN-B3	Max	88.5	86.7	90.2	95.4	91.5
	<b>I</b>	Topk	89.5	85.6	93.3	95.2
		Avg	89.1	86.7	91.6	94.9
	Max	86.0	85.7	86.4	95.8	91.5
	<b>E</b>	Topk	89.2	85.7	92.7	95.8
		Avg	89.1	84.4	93.8	95.8
MIL-DEiT-S	Max	91.7	87.1	96.3	94.7	91.5
	<b>I</b>	Topk	91.4	86.6	96.2	94.9
		Avg	<b>91.8</b>	<b>87.5</b>	96.1	95.1
	Max	91.0	87.4	94.5	95.8	92.3
	<b>E</b>	Topk	91.5	86.9	96.1	94.7
		Avg	91.4	87.4	95.4	95.6

in this context the patch correlation may contain discriminative information. In the case of breast cancer, it seems that the performances are fairly similar. When comparing the DEiT-S and MIL-DEiT-S results with those of EViT, we conclude that: i) discarding several patches does not significantly affect the performance of the models; and ii) our MIL framework is very competitive against more complex models for information selection. Finally, regarding MIL with instances against the embedding versions, we conclude that performing an analysis at the patch level seems to be better in most settings.

In summary, the binary results underscore the potential of integrating a MIL into CNN and ViT pipelines to select key patches for diagnosis. This process effectively reduces the information used by the classifiers without significant performance loss, suggesting that the most discriminative information is concentrated in a few regions of the images.

## 5.2. Multi-class MIL

In this section, we discuss the results of our proposed multi-class MIL framework, as detailed in Table 3. The table compares the performance of our MIL methods with that of baseline models and EViT on the challenging task of multi-class classification of dermoscopy images. Here we show the results for ISIC 2019 and HIBA [15], which was the only test set where all classes matched the ones used for training. In this section, we will only discuss the results for ISIC 2019, while the HIBA results will be discussed in the next section.

Table 3. Results for the multi-class problem in dermoscopy.

Models	ISIC 2019		HIBA
	BA	BA	BA
EN-B3	82.2	32.6	
	83.6	37.6	
EViT	Kr = 0.6	83.6	36.2
	Kr = 0.7	<b>84.3</b>	36.1
MIL-EN-B3	Max	74.1	36.5
	<b>I-1</b>	Topk	78.4
		Avg	79.9
	Max	76.4	36.3
	<b>I-2</b>	Topk	76.2
		Avg	77.5
MIL-DEiT-S	Max	72.3	34.9
	<b>E</b>	Topk	78.9
		Avg	77.6
	Max	82.2	<b>39.4</b>
	<b>I-1</b>	Topk	81.7
		Avg	81.6
MIL-DEiT-S	Max	75.4	35.4
	<b>I-2</b>	Topk	79.0
		Avg	82.6
	Max	82.4	33.8
	<b>E</b>	Topk	82.2
		Avg	82.6

When comparing our multi-class MIL framework with the baseline models, we find that the latter performs better in this task. There is a more noticeable difference in performance when using a CNN as the backbone of our model. This disparity might stem from how the EN-B3 model and the MIL model handle feature extraction. Specifically, the EN-B3 model may use different types of features than the MIL model, which extracts feature maps from a previous layer in the network. This leads to models that have a significantly different number of parameters, which may also impact their ability to memorize in-domain features. Specifically, our MIL-EN-B3 model operates with only 2.2M parameters, as opposed to the more substantial 11M parameters of the EN-B3 model. This hypothesis is further supported by the similar performance between the *instance-level* and *embedding-level* MIL approaches. The *embedding-level* approach mainly serves as a bridge between the MIL framework and traditional CNN-based architectures. The performance comparison between the *embedding-level* MIL-EN-B3 and the EN-B3 baseline mirrors the gap observed with the *instance-level* MIL, suggesting that the disparities in the results could in fact be attributed to the different sizes of the model architectures.

Regarding the two *instance-level* approaches for multi-

Table 4. Generalization results for binary classification of dermoscopy images.

Models	HIBA			PH <sup>2</sup>			Derm7pt		
	BA	R-MEL	R-NV	BA	R-MEL	R-NV	BA	R-MEL	R-NV
EN-B3	81.5	68.0	94.9	88.8	82.5	95.0	76.2	57.9	94.4
DEiT-S	82.0	67.0	96.9	86.6	75.0	98.1	74.0	53.2	94.8
EViT	Kr = 0.6	81.5	68.0	94.9	88.8	80.0	97.5	73.4	52.8
	Kr = 0.7	82.7	71.1	94.4	86.6	75.0	98.1	74.9	56.7
MIL-EN-B3	I	Max	<b>85.1</b>	<b>76.3</b>	93.8	84.4	75.0	93.8	<b>78.7</b>
	I	Topk	80.4	63.9	96.9	<b>89.7</b>	<b>82.5</b>	96.9	77.0
	I	Avg	82.7	72.7	92.7	85.0	77.5	92.5	76.5
	E	Max	85.0	74.7	95.0	85.3	77.5	93.1	76.1
	E	Topk	83.9	75.3	92.5	88.1	82.5	93.8	76.7
	E	Avg	82.4	68.0	96.7	83.4	70.0	96.9	75.3
MIL-DEiT-S	I	Max	82.1	67.5	96.7	87.2	77.5	96.9	74.4
	I	Topk	81.0	64.9	<b>97.1</b>	84.1	72.5	95.6	71.8
	I	Avg	83.2	69.6	96.7	87.8	80.0	95.6	74.8
	E	Max	81.3	68.6	94.2	85.3	75.0	95.6	74.7
	E	Topk	83.1	71.1	95.1	89.7	80.0	<b>99.4</b>	75.1
	E	Avg	82.2	69.6	94.7	85.0	75.0	95.0	76.0

class classification (I-1 and I-2), it seems that I-1 performs better across the two backbones. This leads us to recommend the I-1 formulation in future multi-class MIL applications. Once more, the *instance-level* MIL seems to consistently outperform the embedding approach, reinforcing the importance of performing a patch-based analysis rather than aggregating all or a subset of the image information.

The multi-class classification task is inherently more challenging than its binary counterpart. Nevertheless, the results from the EViT model still prove that information selection is desirable and leads to improved performance compared to traditional models that classify over the entire image. Moreover, our MIL models with DEiT-S backbone still hold their own against both the baselines and EViT. These findings challenge the assumption that larger, more complex models are always better. In essence, our results argue for a more targeted, efficient approach to medical image analysis.

### 5.3. Generalization Across Diverse Demographics

We evaluated the robustness of our MIL-based models across varied dermoscopy image datasets, each representing distinct patient demographics not seen in our training or validation sets. Table 4 displays the binary results for skin tests from the HIBA, PH<sup>2</sup>, and Derm7pt datasets, while Table 3 shows the results for HIBA.

Our MIL models consistently outperformed their baseline counterparts on unseen data. For example, the *instance-level* MIL-EN-B3 model using max pooling outperformed the EN-B3 baseline by a BA of 3.6% on the HIBA dataset. In addition, the multi-class results on the HIBA test set (see

Table 3) show that even if there is a performance drop when the MIL models are compared with the baselines, they still generalize better to unseen data. Finally, contrary to what is stated in the literature, the *embedding-level* approach did not consistently outperform the *instance-level* models. In fact, the *instance-level* models often outperformed their *embedding-level* counterparts. These results suggest that the key patches identified by the *instance-level* MIL models may have significant medical relevance, contributing to improved generalization to unseen data.

In summary, the generalization results show that our MIL models deal better with unseen data, being potentially more fair across different demographics, despite using less information. This establishes MIL as a promising method to improve fairness in medical image analysis.

### 5.4. Explainability of MIL Models

Our results suggest that the key regions identified by the *instance-level* MIL models are correlated with meaningful information within the image, thereby increasing the model’s robustness to dataset bias compared to baseline counterparts. Nevertheless, it is critical to assess whether these identified regions truly capture clinical findings or are simply artifacts. To clarify this, we compared heatmaps generated by the EN-B3 baseline to those generated by our MIL models, focusing on the binary classification of melanoma versus nevus in the PH<sup>2</sup> dataset.

Figure 3 shows the Grad-Cam visualizations for the baseline EN-B3 model, highlighting the areas that influence its predictions for melanoma (MEL) and nevus (NV) clas-

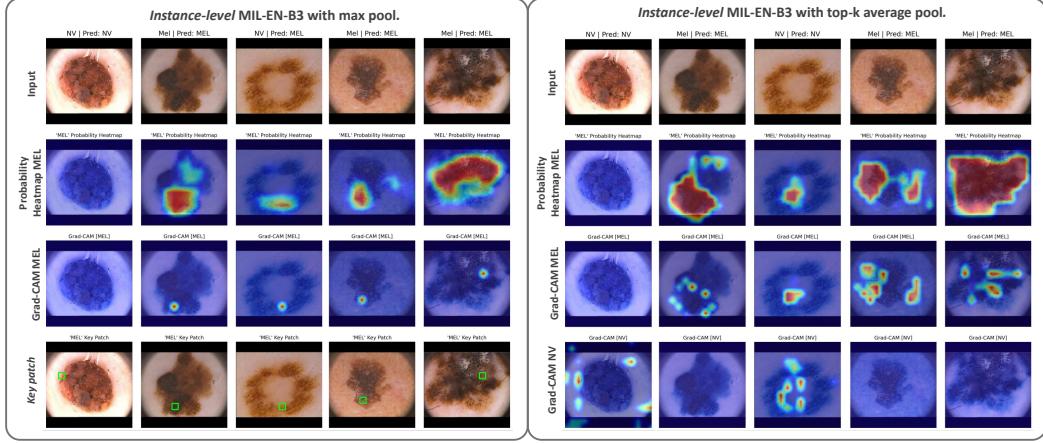


Figure 2. Visualization of the key patches identified by two different MIL approaches. On the left, we have the *instance-level* MIL-EN-B3 using *max* polling, and on the right, we have the *instance-level* MIL-EN-B3 using the *top-k average* pooling operator, with  $k \approx 12.5\%$ . The images used for visualization are taken from the PH<sup>2</sup> test set and refer to the binary classification task of melanoma vs. nevus.

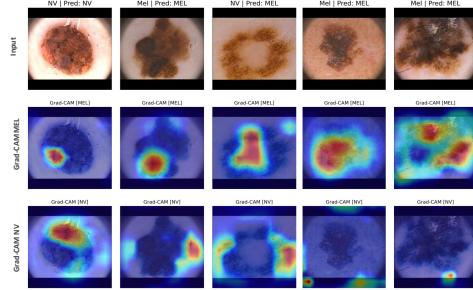


Figure 3. Grad-Cam [23] heatmap visualizations generated by the EN-B3 baseline model for images from the PH<sup>2</sup> test set.

sifications, while Figure 2 illustrates the key regions identified by the *instance-level* MIL-EN-B3 model for the same lesions. In the MIL setting, we compare two pooling strategies: *max* pooling and *top-k average* pooling, which averages the values of the 25 most relevant patches.

Notably, our *instance-level* MIL approaches consistently highlight *key patches* that, similar to ROIs in clinical diagnosis, lie within or at the edges of lesions for melanoma cases, or bordering healthy tissue for nevus cases. This preference for clinically relevant areas confirms the ability of our models to extract medically relevant features, validating their performance on the validation set and their ability to generalize to unseen data.

When comparing the heatmaps, it is clear that the EN-B3 model produces coarse heatmaps, whereas the MIL’s *instance-level* approaches produce a more detailed delineation of relevant regions, providing finer explanations. This clarity and specificity reinforces MIL’s position as a more clinically translatable tool that can potentially provide

clearer explanations of the decision-making process.

## 6. Conclusions

This work demonstrates the potential of integrating MIL into the pipeline of CNNs and ViTs to select relevant patches and use less information in the classification stage. Our findings reveal that despite a significant reduction in the amount of information, MIL models can achieve results comparable to more complex networks. By focusing on the most discriminative image patches, similar to clinical practice, MIL models show a strong ability to generalize across different datasets and demographic groups. This suggests a promising direction towards creating more explainable, efficient, and fair medical image analysis systems. Moreover, the assessment of selected regions underscores MIL’s alignment with clinical relevance, providing a more interpretable decision-making process that mirrors the diagnostic approach of medical experts. Future work will focus on validating the identified key regions against specific medical concepts, as well as exploring these regions to improve model performance and further enhance clinical applicability and fairness.

## Acknowledgements

This work was supported by LARSyS funding (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020) and projects 2023.02043.BDANA, 10.54499/2022.07849.CEECIND/CP1713/CT0001, MIA-BREAST [10.54499/2022.04485.PTDC], PT SmartRetail [PRR - C645440011-00000062], Center for Responsible AI [PRR - C64500882-00000055].

## References

[1] Reza Azad, Amirhossein Kazerouni, Moein Heidari, Ehsan Khodapanah Aghdam, Amirali Molaei, Yiwei Jia, Abin Jose, Rijo Roy, and Dorit Merhof. Advances in medical image analysis with vision transformers: a comprehensive review. *Medical Image Analysis*, page 103000, 2023. 1, 2

[2] Alceu Bissoto, Catarina Barata, Eduardo Valle, and Sandra Avila. Even small correlation and diversity shifts pose dataset-bias issues. *Pattern Recognition Letters*, 2024. 1

[3] Kaitao Chen, Shiliang Sun, and Jing Zhao. Camil: Causal multiple instance learning for whole slide image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1120–1128, 2024. 2

[4] Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C Thai, Kathleen Moore, Robert S Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, 79:102444, 2022. 1, 2

[5] Noel C. F. Codella, David A. Gutman, Emre Celebi, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Nabin Mishra, Aadi Kalloo, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan C. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv: Computer Vision and Pattern Recognition*, 2016. 4

[6] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Verónica Vilaplana, Ofer Reiter, Ofer Reiter, Allan C. Halpern, Susana Puig, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild. *arXiv: Image and Video Processing*, 2019. 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2020. 2

[9] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1, 2

[10] Tiago Gonçalves, Isabel Rio-Torto, Luís F Teixeira, and Jaime S Cardoso. A survey on attention mechanisms for medical applications: are we moving toward better algorithms? *IEEE Access*, 10:98909–98935, 2022. 2

[11] Michael D. Heath, Michael D. Heath, Kevin W. Bowyer, Kevin W. Bowyer, Daniel B. Kopans, Daniel B. Kopans, W. Philip Kegelmeyer, W. Philip Kegelmeyer, Richard H. Moore, Richard H. Moore, Kai-Chun Chang, K.I. Chang, S. Munishkumaran, and S. Munishkumaran. Current status of the digital database for screening mammography. *Digital Mammography / IWDM*, 1998. 5

[12] Michael D. Heath, Michael D. Heath, Kevin W. Bowyer, Kevin W. Bowyer, D. B. Kopans, Daniel B. Kopans, Roscoe M. Moore, and Richard H. Moore. The digital database for screening mammography. *null*, 2007. 5

[13] Maximilian Ilse, Jakub M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. *International Conference on Machine Learning*, 2018. 2

[14] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019. 4

[15] María Agustina Ricci Lara, María Victoria Rodríguez Kowalcuk, Maite Lisa Eliceche, María Guillermmina Ferreresso, Daniel Luna, Susana Carrera Benítez, and Luis Daniel Mazzuoccolo. A dataset of skin lesion images collected in argentina for the evaluation of ai tools in this population. *Scientific Data*, 2023. 4, 6

[16] Giuseppe Di Leo, Consolatina Liguori, Antonio Pietrosanto, Gabriella Fabbrocini, and M. Sclavenzi. Elm image processing for melanocytic skin lesion diagnosis based on 7-point checklist: a preliminary discussion. 2004. 1

[17] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 2

[18] Xi-Lin Li. A multiclass multiple instance learning method with exact likelihood. *arXiv: Machine Learning*, 2018. 3

[19] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and P. Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv.org*, 2022. 2, 4

[20] Pedro Henrique Martins, Vlad Niculae, Zita Marinho, and André FT Martins. Sparse and structured visual attention. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 379–383. IEEE, 2021. 2

[21] Teresa Mendonça, Pedro M. Ferreira, Jorge S. Marques, André R. S. Marçal, and Jorge Rozeira. Ph 2 - a dermoscopic image database for research and benchmarking. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2013. 4

[22] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. *International Conference on Learning Representations*, 2014. 3

[23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 8

[24] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image

classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 2

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021. 5

[26] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 2018. 4

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 3

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *null*, 2017. 2

[29] Xinyu Xu and Baoxin Li. Evaluating multi-class multiple-instance learning for image categorization. In *Computer Vision – ACCV 2007*, pages 155–165, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. 3

[30] Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, and Aaron Courville. Can subnetwork structure be the key to out-of-distribution generalization? In *International Conference on Machine Learning*, pages 12356–12367. PMLR, 2021. 2