ProtoMIL: Multiple Instance Learning with Prototypical Parts for Whole-Slide Image Classification – Supplementary Materials

Dawid Rymarczyk[1,2], Adam Pardyl[1]*, Jarosław Kraus[1]*,
Aneta Kaczyńska[1]*, Marek Skomorowski[1], and Bartosz Zieliński[1,2]

[1] Faculty of Mathematics and Computer Science, Jagiellonian University,
6 Łojasiewicza Street, 30-348 Kraków, Poland
[2] Ardigen SA, 76 Podole Street, 30-394 Kraków, Poland
{dawid.rymarczyk,adam.pardyl,jarek.kraus,
aneta.kaczynska}@student.uj.edu.pl,
{marek.skomorowski,bartosz.zielinski}@uj.edu.pl

In this Supplementary Materials, we present additional details on the ProtoMIL model and similarity scores visualizations with more instances and prototypes for all datasets considered in our experiments.

1 ProtoMIL

1.1 Prototypes projection.

Prototypes projection is an important step in the training procedure because it visualizes the prototypes using training patches. For this purpose, it replaces every learned prototype with the nearest training patch from the bag with the same label as the prototype class. The prototype $\mathbf{p^c}$ of class c (negative or positive) can be replaced using the following formula

$$\mathbf{p^c} \leftarrow \arg\min_{\mathbf{z}\in Z} \|\mathbf{z} - \mathbf{p^c}\|_2,$$

where $Z = \{\mathbf{z} \in Z_{\mathbf{x}} | \mathbf{x} \in X \wedge y = c\}$ and y is a label of bag X.

1.2 Pruning.

During the prototype projection, every prototype is replaced with the representation of the nearest training patch from the bag with the same label. Generally, the representations of the nearest training patches correspond to the same label. However, in some cases, the nearest patches of a prototype correspond to more than one class. It is especially problematic in highly unbalanced datasets, frequently occurring in MIL tasks. To remove such misleading prototypes, we extend the prototype pruning algorithm from [1] to work in the MIL scenario. More precisely, we find k-nearest training patches for each prototype p_i^c belonging

[*] denotes equal contribution

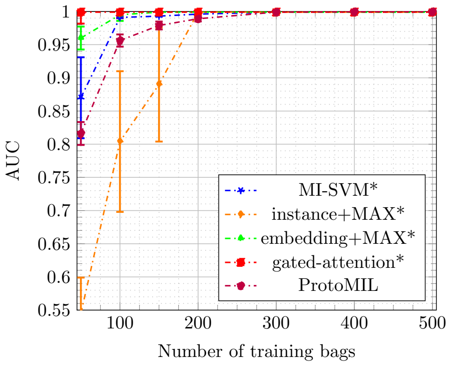


Fig. 1: Results for ProtoMIL and baseline MIL approaches on the MNIST Bags dataset depending on the number of training bags (x axis) using the AUC metric (y axis). One can observe that ProtoMIL achieves state-of-the-art results with a larger number of samples.

to class c. If out of those k patches less than r belong to bags labeled with class c, we assume that this prototype is not determinant and remove it. Moreover, in contrast to [1], we automatically select r to remove up to l% of prototypes (l and k are selected so that both classes still contain prototypes, and the drop in training accuracy is minimal). Finally, we fine-tuned attention and the final layers to compensate for the prototype removal.

2 Additional results

2.1 MNIST Bags

Experiment details. We experiment with the MNIST dataset, for which we generate the bags like proposed in [2]. Namely, a single bag contains grayscale images randomly sampled from the MNIST dataset. The bags' sizes are chosen using a normal distribution with a mean of 100 and a standard deviation of 20. A bag is considered positive if it contains at least one image labeled as "9". There are equal numbers of positive and negative bags. Notice that even though such dataset is class-balanced, it contains only 5% of images labeled as "9" (10% instances in the positive bags). We test ProtoMIL for different size of dataset (50, 100, 200, 300, 400, 500 bags). Every experiment is run with random 10-fold cross-validation and repeated five times with a different seed to obtain mean AUC as the evaluation metric. We train a model for 30, 20, and 10 epochs for warmup, fine-tuning, and end-to-end training, respectively. The number of prototypes per class is set to 10, with prototype size $64 \times 2 \times 2$ (determined experimentally).

Results. We compare our model to baseline MIL pooling methods from [2]. As shown, our ProtoMIL approach requires slightly more samples to achieve AUC scores competitive to the regular models (Figure 1). However, as presented in Figure 3a, it increases model interpretability by finding distinct parts of images and match them with intuitive positive and negative prototypes (see Figure 2).

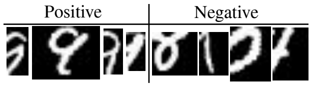


Fig. 2: Sample positive and negative prototypes of ProtoMIL trained on the MNIST Bags dataset. Notice that the positive prototypes correspond to parts of "9" while the negative prototypes contain parts of the other digits (like "8" or "4"). It is expected because a bag is considered positive if it contains at least one image of "9".

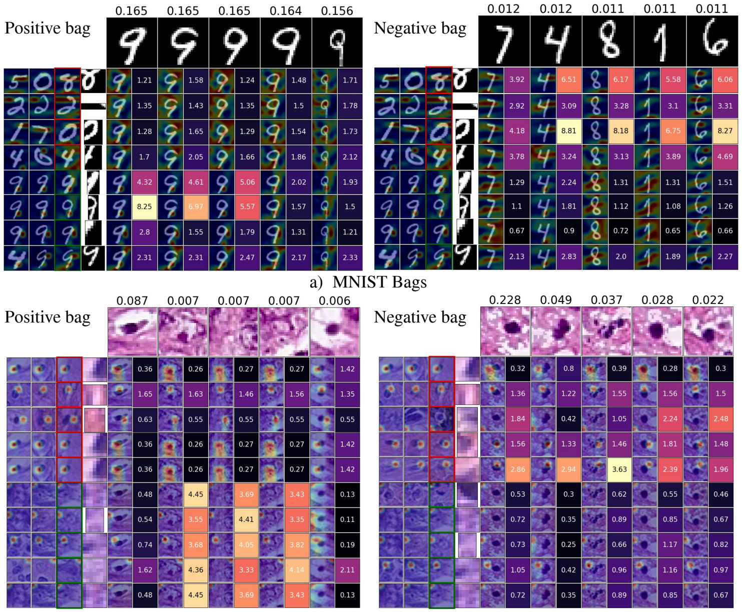
We experiment on two histological datasets as out toy task: Colon Cancer and Bisque breast cancer. The former contains 100 H&E images with $22,444$ manually annotated nuclei of four different types: epithelial, inflammatory, fibroblast, and miscellaneous. To create bags of instances, we extract 27×27 nucleus-centered patches from each image, and the goal is to detect if the bag contains one or more epithelial cells, as colon cancer originates from them. On the other hand, the Bisque dataset consists of 58 H&E breast histology images of size 896×768, out of which 32 are benign, and 26 are malignant (contain at least one cancer cell). Each image is divided into 32×32 patches, resulting in 672 patches per image. Patches with at least 75% of the white pixels are discarded, resulting in 58 bags of various sizes.

We apply extensive data augmentation for both datasets, including random rotations, horizontal and vertical flipping, random staining augmentation, staining normalization, and instance normalization. We use ResNet-18 convolutional parts with the first layer modified to 3×3 convolution with stride 1 to match the size of smaller instances. We set the number of prototypes per class to 10 with a size of $128 \times 2 \times 2$. Warmup, fine-tuning, and end-to-end training take 60, 20, and 20 epochs, respectively. 10-fold cross-validation with 1 validation fold and 1 test fold is repeated 5 times.

Results. Table 1 presents our results compared to both traditional and attention-based MIL models. On the Bisque dataset, our model significantly outperforms all baseline models. However, due to the small size of the Colon Cancer dataset, ProtoMIL overfits, resulting in poorer AUC than attention-based models. Nevertheless, in both cases, ProtoMIL provides finer explanations than all baseline models (see Figure 3b and Supplementary Materials).

2.2 Messidor dataset

Experiment details. The Messidor dataset contains 1200 retinal images: 654 with a positive label (diabetic retinopathy) and 546 with a negative one. To create bags



a) MNIST Bags

b) Colon Cancer

Fig. 3: Similarity scores between five crucial instances of a bag (columns) and eight or ten prototypical parts (rows) for a positive and negative bag (left and right side, respectively) from the MNIST Bags (a) and Colon Cancer datasets (b). Each prototypical part is represented by a part of image and three nearest training patches, and each instance is represented by the image and the value of attention weight a_i. Moreover, each cell contains a similarity score and a heatmap corresponding to prototype activation. One can observe that instances of a negative bag usually activate negative prototypes (four upper prototypes in red brackets), while the instances of positive bags mostly activate positive prototypes (four bottom prototypes in green brackets).

Method	Colon Cancer	
	Accuracy	AUC
instance+max*	84.2% ± 2.1%	0.914 ± 0.010
instance+mean*	77.2% ± 1.2%	0.866 ± 0.008
embedding+max*	82.4% ± 1.5%	0.918 ± 0.010
embedding+mean*	86.0% ± 1.4%	0.940 ± 0.010
AbMILP*	88.4% ± 1.4%	0.973 ± 0.007
SA-AbMILP**	**90.8% ± 1.3%**	**0.981 ± 0.007**
ProtoMIL (our)	81.3% ± 1.9%	0.932 ± 0.014

Table 1: Results for Colon Cancer dataset. ProtoMIL achieves slightly worse results for the Colon Cancer dataset, probably due to its small size. Notice that values for comparison indicated with "*" and "**" comes from [2] and [3], respectively.

Method	Accuracy	F-score
MI-SVM*	54.5%	0.70
mi-SVM*	54.5%	0.71
EMDD*	55.1%	0.69
Citation k-NN*	62.8%	0.69
MILBoost*	64.1%	0.66
mi-Graph*	72.5%	0.75
MIL-GNN-Att*	72.9%	0.75
MIL-GNN-DP*	74.2%	**0.77**
AbMILP**	74.5%	0.74
SA-AbMILP**	75.2%	0.76
LSA-AbMILP**	**76.3%**	**0.77**
ProtoMIL (our)	70.0%	0.75

Table 2: Results for the Messidor dataset show that in terms of F-score, our ProtoMIL method is comparable with methods based on attention (AbMILP) or graph convolutions (MIL-GNN-ATT). Notice that values for comparison marked with "*" and "**" are taken from [4] and [3], respectively.

of instances, we crop overlapping patches of size 224 × 224 from each of 700 × 700 images, and patches with more than 70% black pixels are dropped as in [4]. Additionally, we apply extensive data augmentation, including random rotations, horizontal and vertical flipping, Gaussian noise, and patch normalization. We use ResNet-18 convolutional layers learned from scratch with 10 prototypes per class and prototype size of $1 \times 1 \times 128$. Warmup, fine-tuning, and end-to-end training take 30, 20, and 10 epochs, respectively. We perform 10 fold cross-validation repeated two times as in [4].

Results. Results of ProtoMIL in the case of F-score are comparable with the ones achieved in [4] and [3] (see Table 2). However, the accuracy is significantly lower, most possibly due to the data class imbalance. Nevertheless, our model provides a fine-grained interpretation of its decision, as presented in Figure 4.

2.3 Additional pruning results

3 Additional visualizations

References

1. Chen, C., Li, O., Tao, C., Barnett, A.J., Su, J., Rudin, C.: This looks like that: deep learning for interpretable image recognition. arXiv preprint arXiv:1806.10574 (2018)

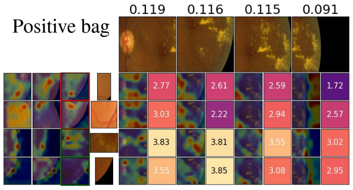


Fig. 4: Similarity scores between four crucial instances of a bag (columns) and four prototypical parts (rows) for a positive bag from the Messidor dataset. One can observe that ProtoMIL focuses on the disease factors, which are the brightest yellow spots on the image. Moreover, both positive and negative prototypes are activated since the retina with pathological changes still shows healthy features, such as veins. Please refer to Figure 3 for a detailed description of the visualization.

Dataset	Before pruning			After pruning		
	Proto. #	Accuracy	AUC	Proto. #	Accuracy	AUC
MNIST Bags 500	20 ± 0	99.2% ± 0.1%	0.999 ± 0.001	14.12 ± 0.28	99.2% ± 0.1%	0.999 ± 0.001
Messidor	20 ± 0	70.0% ± 0.9%	0.692 ± 0.012	16.70 ± 1.86	64.7% ± 1.3%	0.717 ± 0.013

Table 3: The influence of ProtoMIL pruning on the accuracy and AUC score. One can notice that even though the pruning removes around 30% of the prototypes, it usually does not noticeably decrease the AUC and accuracy of the model.

2. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
3. Rymarczyk, D., Borowa, A., Tabor, J., Zielinski, B.: Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1721–1730 (2021)
4. Tu, M., Huang, J., He, X., Zhou, B.: Multiple instance learning with graph neural networks. arXiv preprint arXiv:1906.04881 (2019)

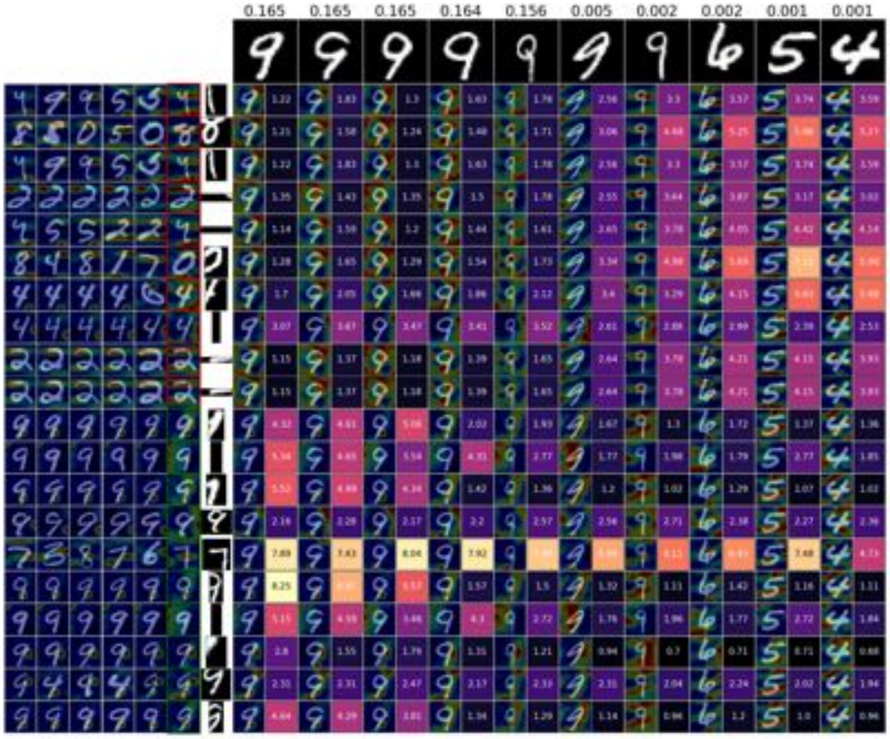


Fig. 5: Similarity scores for a positive bag from MNIST Bags.

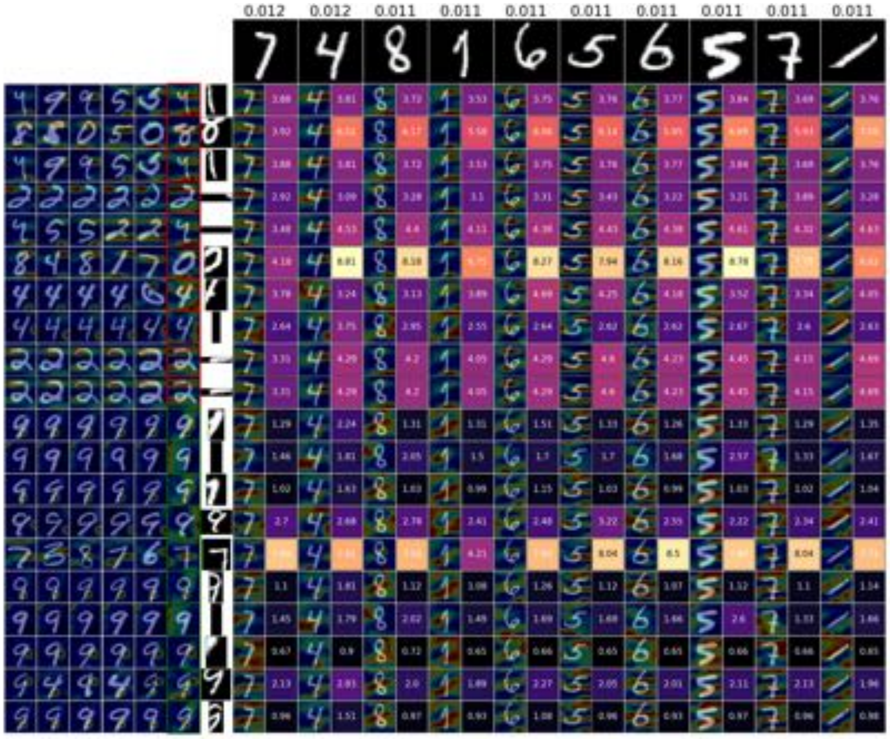


Fig. 6: Similarity scores for a negative bag from MNIST Bags.

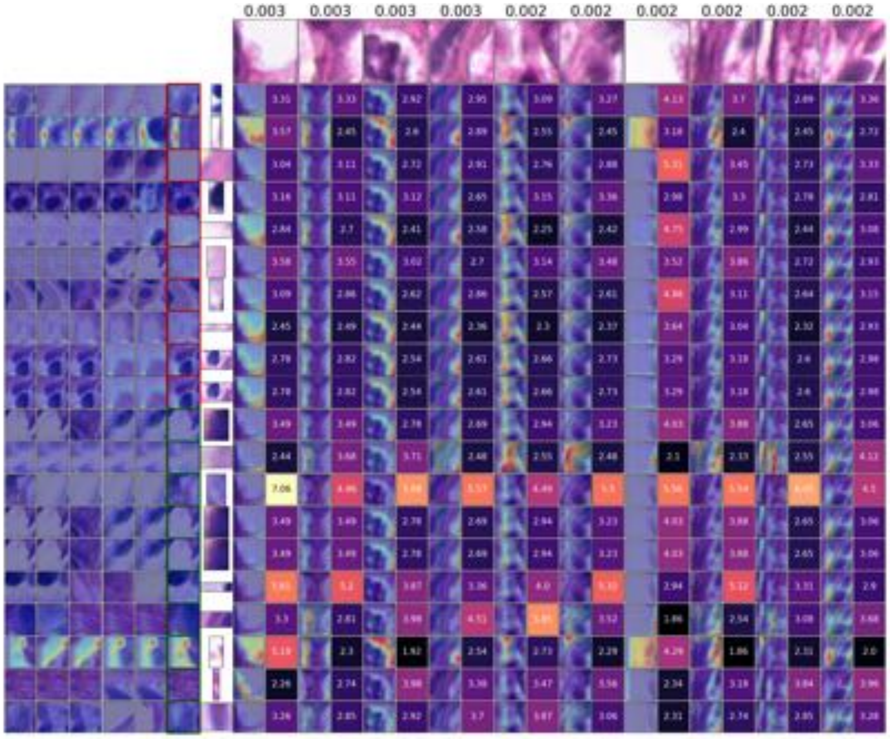


Fig. 7: Similarity scores for a positive bag from Bisque dataset.

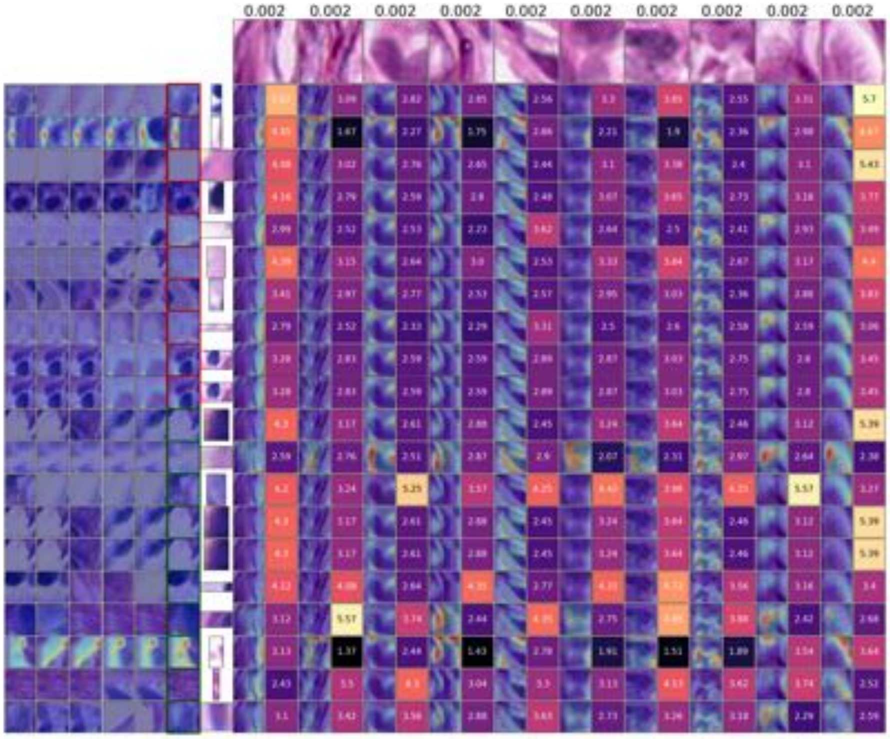


Fig. 8: Similarity scores for a negative bag from Bisque dataset.

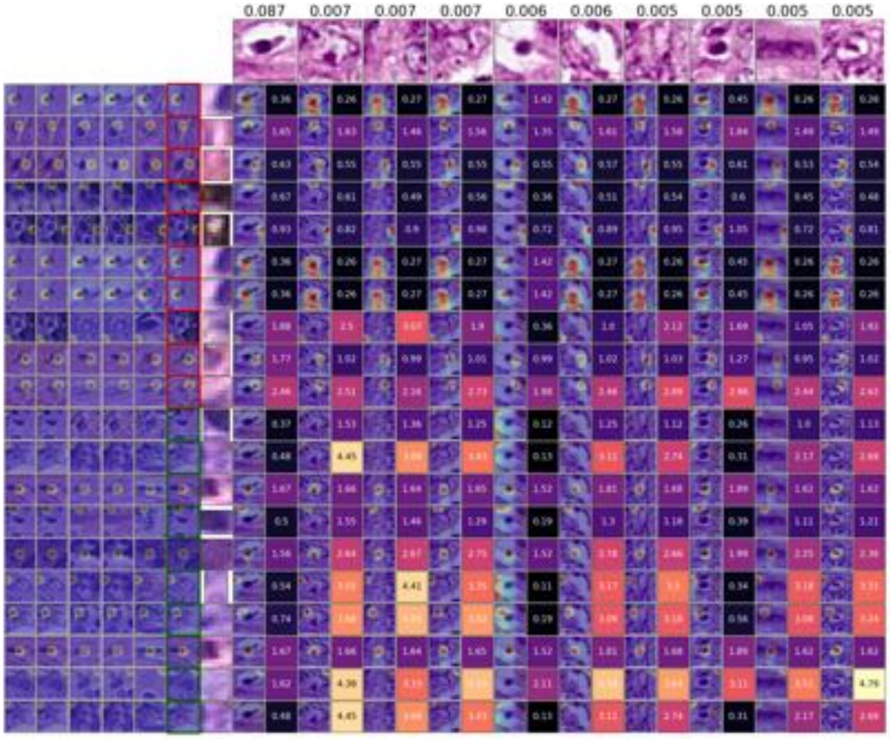


Fig. 9: ProtoMIL analysis matrix for a positive example from Colon Cancer dataset.

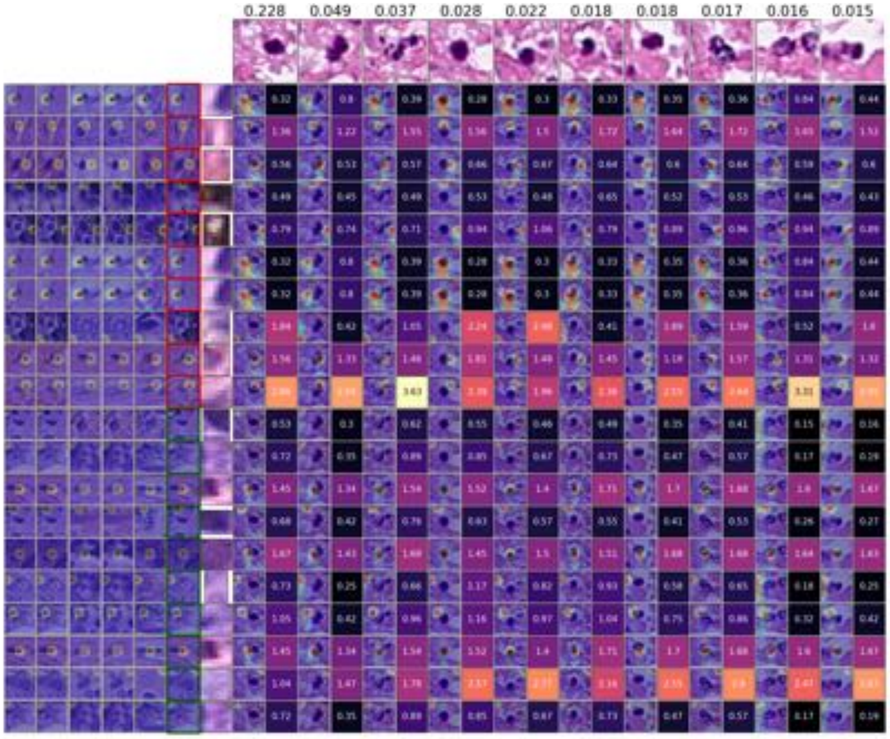


Fig. 10: ProtoMIL analysis matrix for a negative example from Colon Cancer dataset.

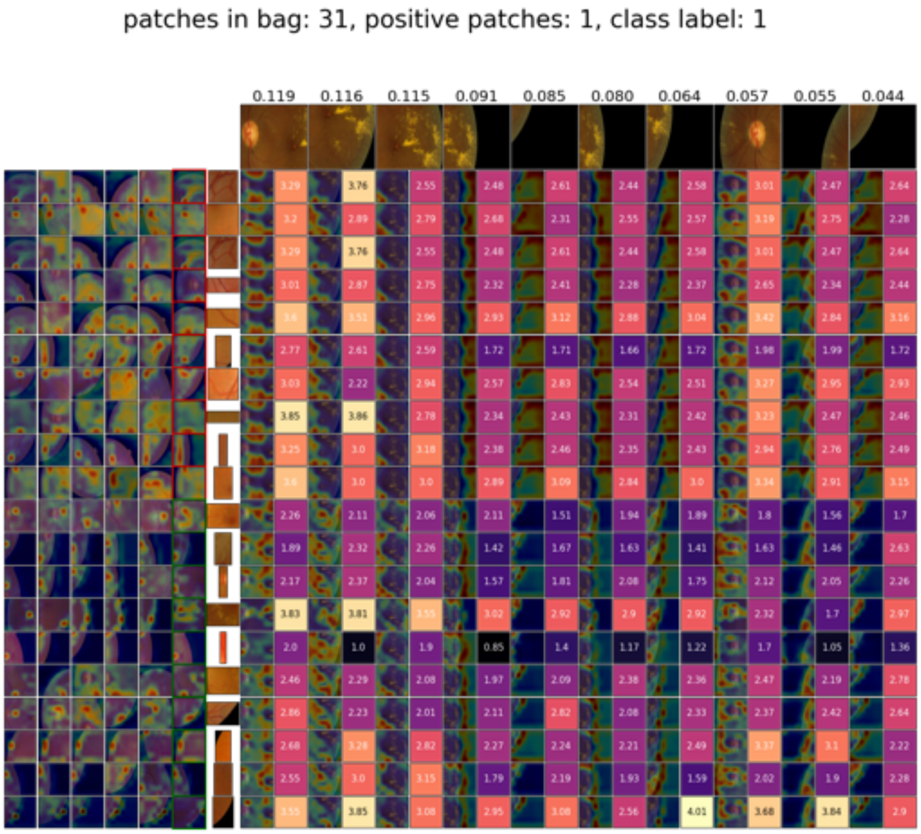


Fig. 11: Similarity scores for a positive bag from Messidor dataset.

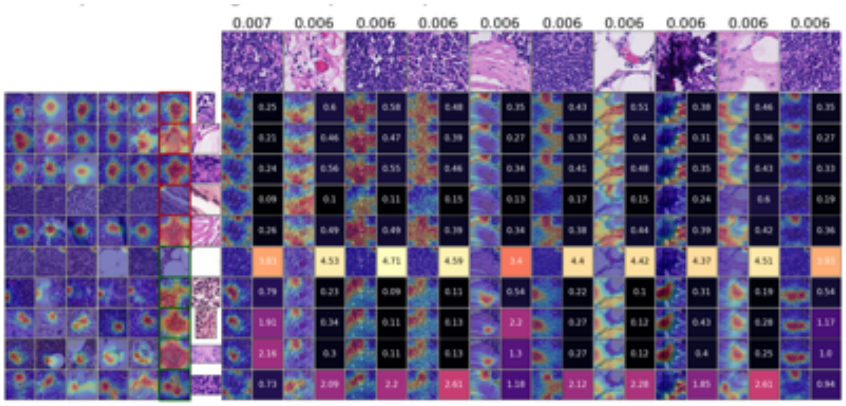


Fig. 12: Similarity scores for a positive bag from Camelyon16 dataset.

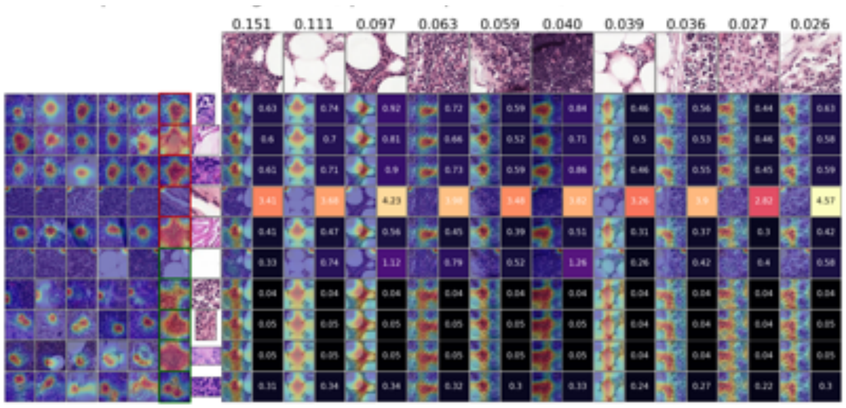


Fig. 13: Similarity scores for a negative bag from Camelyon16 dataset.

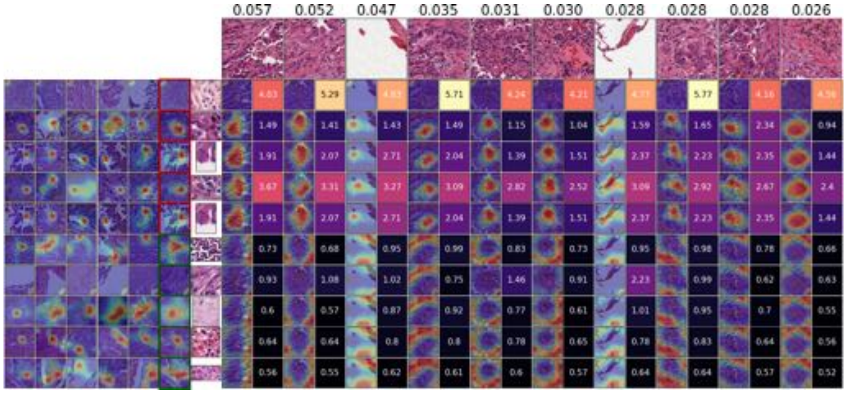


Fig. 14: Similarity scores for a LUAD bag from TCGA-NSCLC dataset.

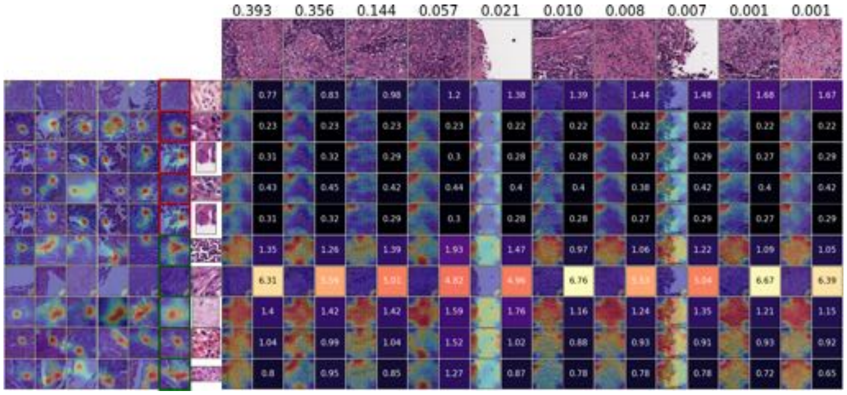


Fig. 15: Similarity scores for a LUSC bag from TCGA-NSCLC dataset.

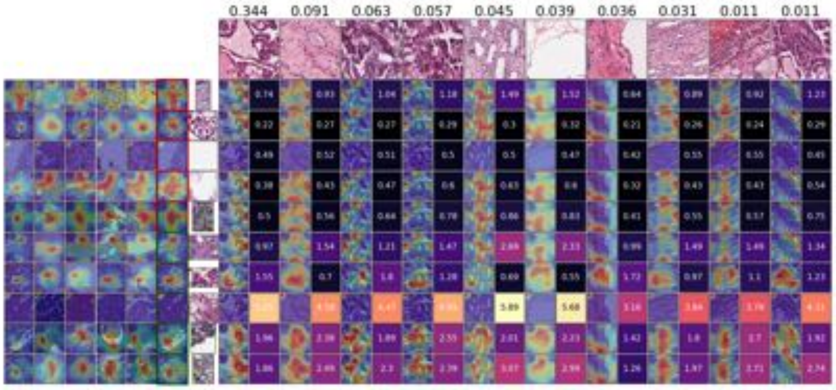


Fig. 16: Similarity scores for a positive bag from TCGA RCC dataset.

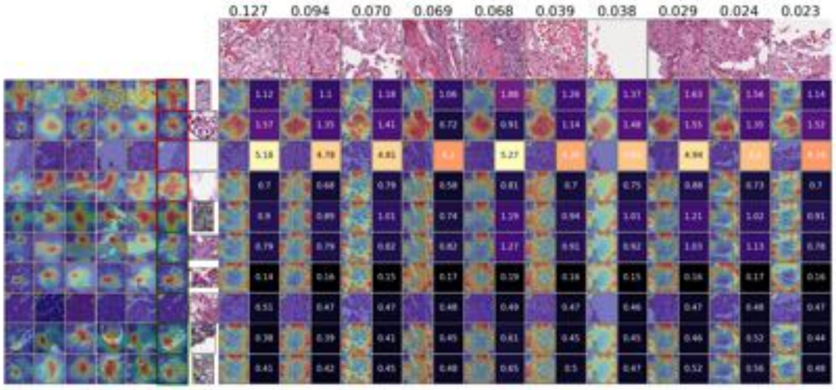


Fig. 17: Similarity scores for a negative bag from TCGA RCC dataset.