Project Machine Learning
— Milestone 2 —

Cederic Aßmann, Friedrich Hagedorn, Jakub Sliwa

December 30, 2024

1 Introduction

The main objective of our participation in the *Machine Learning Project* during the Winter Semester 2024/25 is to reconstruct the results achieved by Ilse et al. (2018) and published in the paper titled *Attention-based Deep Multiple Instance Learning*. The replication process has been divided into three milestones. Hence, this report aims to present our progress in achieving the final goal within the confines of the second milestone. The tasks within its scope include hyperparameter fine-tuning for MNIST-Bags and conducting experiments analogous to these from milestone 1, but on a real-world histopathology dataset - Camelyon16[1].

The structure of the report is as follows: Section 1 provides an overview of the report's purpose and structure. Section 2 describes the process of hyperparameter fine-tuning for MNIST-Bags, along with its impact on test results across various MIL approaches. Moreover the detailed quantitative results for the MNIST-Bags experiments, conducted with the chosen hyperparameters, are provided. Section 3 introduces the real-world histopathology dataset, Camelyon16, detailing its origins, the transformation of slides into patches, and its relevance to the MIL problem. Section 4 focuses on the experiments conducted during this milestone, beginning with a detailed explanation of our methodology, followed by a discussion of the results and findings. This section also covers model interpretability and presents approaches we used to enhance it. Section 5 addresses the challenges encountered and outlines the work remaining for the final milestone. In the Appendix, the Camelyon16 hyperparameter fine-tuning results as well as the validation performance metrics are presented A.

2 Hyperparameter tuning MNIST

To optimize model performance, we conducted a hyperparameter tuning using Weights and Biases sweeps with a Bayesian search strategyWandB (2024). Bayesian search was chosen due to its efficiency in exploring high-dimensional spaces compared to exhaustive grid search or random search. The mean bag size in this hyperparameter tuning was set to 50 with a variance of 10. We focused on the following hyperparameters:

Learning Rate: The learning rate $\eta \in [0.0001, 0.001]$ significantly influences how quickly a model converges. A low learning rate ensures stability, but may result in slower convergence, while a high learning rate risks overshooting optimal minima.

Weight Decay: Weight decay $\lambda \in [10^{-6}, 10^{-2}]$ helps regularize the model by penalizing large weights, mitigating overfitting. The range was selected to balance between under-regularization and over-regularization.

Attention dimension: The attention dimension $d_{\text{attn}} \in \{64, 128, 256\}$ determines the size of the attention space in the model, impacting its capacity to capture complex patterns. Higher dimensions can improve performance, but increase computational cost and risk overfitting.

[1]The full implementation is available at `https://git.tu-berlin.de/cederic/attdmil`.

Results of the hyperparameter sweep: A selection of training and validation logs is provided in this WandB report. The general training loss characteristic and validation AUC scores are similar to the figures in milestone 1. The sweep identified the following optimal hyperparameter configuration: **Learning Rate:** 0.00015, **Weight Decay:** 0.00963 and **Attention Dimension:** 128.

Performance improvement: Repeating the training process with these optimized hyperparameters led to general improvements in performance across different MIL approaches and training bag sizes. As in Ilse et al. (2018) and in milestone 1, the performance is measrued by the AUC score.

The Gated Attention and Attention model for $\mu = 10$, see Tab. 1, showed consistent improvement using the optimal hyperparameters across almost all number of training bags, while other approaches like Embedding+mean experienced slight decreases in performance for larger training bag sizes. Most MIL approaches for a mean bag size of $\mu = 50$, including Embedding+mean and Gated Attention, performed absolutely well for different training bag sizes, see Tab. 2. Improvements, compared to the milestone 1 experiments, were observed for most configurations, but certain configurations, e.g. Embedding+mean, displayed minor fluctuations in performance. Moreover, the Attention and Gated Attention models achieved near-perfect results for a mean bag size of $\mu = 100$, see Tab. 3. Performance improvements were generally stable, though Embedding+mean showed slight decreases for larger number of training bags.

These results indicate that the optimal learning rate and weight decay parameter successfully balanced convergence speed and regularization, while the attention dimension of 128 provided sufficient capacity for learning complex relationships within the data without overfitting.
Hyperparameter tuning using Bayesian search and Wandb sweeps proved effective for optimizing our MIL models, yielding a configuration that outperformed previous experiments from milestone 1.

Table 1: Test AUC, $\mu = 10$, for MNIST-Bags dataset with a different number of training bags. Optimized hyperparameters. Green: Improvement to M1, Red: No improvement to M1.

# train bags	50	100	150	200	300	400	500
Instance+max	0.530 ± 0.131	0.863 ± 0.11	0.959 ± 0.008	0.974 ± 0.004	0.982 ± 0.004	0.988 ± 0.003	0.991 ± 0.002
Instance+mean	0.672 ± 0.016	0.706 ± 0.023	0.721 ± 0.014	0.700 ± 0.018	0.710 ± 0.012	0.738 ± 0.009	0.738 ± 0.008
Emb.+max	0.722 ± 0.029	0.903 ± 0.016	0.961 ± 0.008	0.972 ± 0.005	0.982 ± 0.005	0.987 ± 0.002	0.99 ± 0.003
Emb.+mean	0.716 ± 0.019	0.831 ± 0.011	0.868 ± 0.015	0.913 ± 0.013	0.972 ± 0.003	0.983 ± 0.003	0.986 ± 0.001
Attention	0.884 ± 0.019	0.94 ± 0.013	0.948 ± 0.012	0.98 ± 0.005	0.988 ± 0.002	0.991 ± 0.001	0.993 ± 0.003
Gated Attention	0.879 ± 0.011	0.943 ± 0.015	0.961 ± 0.009	0.98 ± 0.004	0.985 ± 0.004	0.991 ± 0.001	0.993 ± 0.001

Table 2: Test AUC, $\mu = 50$, for MNIST bags dataset with a different number of training bags. Optimized hyperparameters. Green: Improvement to M1, Red: No improvement to M1.

# train bags	50	100	150	200	300	400	500
Instance+max	0.547 ± 0.166	0.849 ± 0.204	0.883 ± 0.218	0.884 ± 0.218	0.994 ± 0.001	0.996 ± 0.001	0.996 ± 0.001
Instance+mean	0.830 ± 0.022	0.804 ± 0.016	0.784 ± 0.037	0.772 ± 0.024	0.829 ± 0.027	0.832 ± 0.013	0.785 ± 0.014
Emb.+max	0.946 ± 0.021	0.982 ± 0.006	0.993 ± 0.002	0.994 ± 0.001	0.996 ± 0.001	0.997 ± 0.001	0.997 ± 0.001
Emb.+mean	0.882 ± 0.013	0.962 ± 0.002	0.981 ± 0.003	0.986 ± 0.001	0.965 ± 0.052	0.993 ± 0.001	0.995 ± 0.000
Attention	0.977 ± 0.005	0.992 ± 0.002	0.995 ± 0.002	0.996 ± 0.001	0.995 ± 0.001	0.997 ± 0.000	0.998 ± 0.000
Gated Attention	0.979 ± 0.004	0.994 ± 0.001	0.996 ± 0.001	0.996 ± 0.000	0.996 ± 0.001	0.997 ± 0.001	0.997 ± 0.001

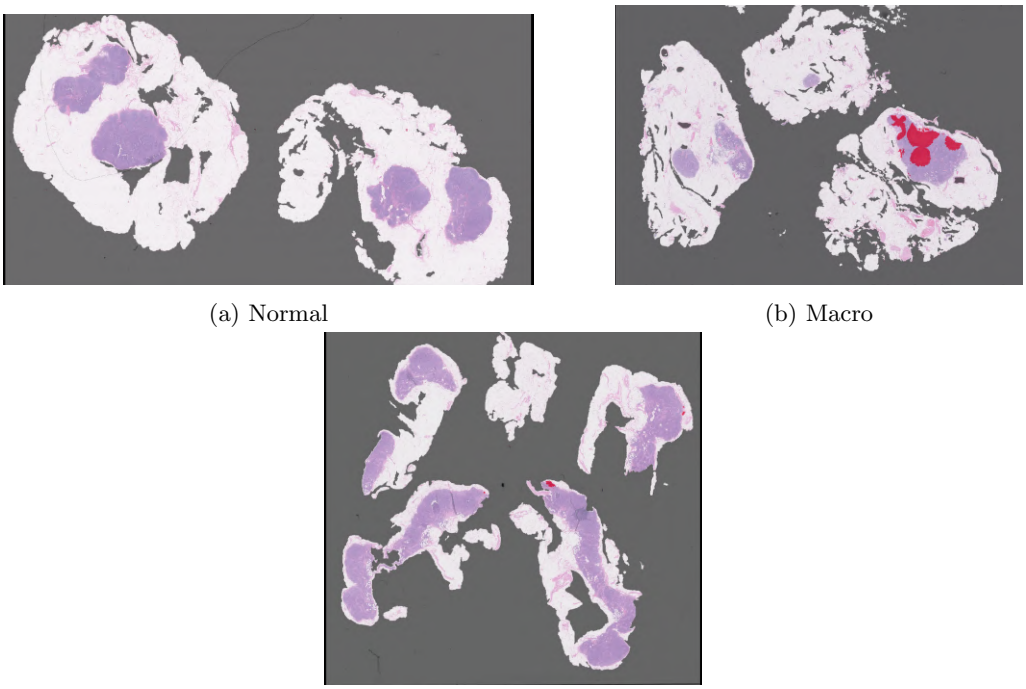
Table 3: Test AUC, $\mu = 100$, for MNIST bags dataset with a different number of training bags. Optimized hyperparameters. Green: Improvement to M1, Red: No improvement to M1.

# train bags	50	100	150	200	300	400	500
Instance+max	0.584 ± 0.205	0.890 ± 0.207	0.748 ± 0.238	0.999 ± 0.001	0.744 ± 0.255	1.000 ± 0.000	1.000 ± 0.000
Instance+mean	0.920 ± 0.022	0.905 ± 0.028	0.887 ± 0.026	0.864 ± 0.014	0.894 ± 0.016	0.850 ± 0.013	0.855 ± 0.021
Emb.+max	0.993 ± 0.003	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Emb.+mean	0.993 ± 0.002	0.994 ± 0.001	0.996 ± 0.000	0.997 ± 0.001	0.998 ± 0.000	0.999 ± 0.000	0.999 ± 0.000
Attention	0.999 ± 0.001	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.001	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Gated Attention	0.999 ± 0.001	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

3 Histopathology dataset overview

3.1 Camelyon16

The Camelyon16 dataset, introduced in 2015 as part of the Cancer Metastases in Lymph Nodes Challenge, aimed to assess the performance of machine learning algorithms in the automated detection of metastases within whole-slide images (WSIs) of sentinel lymph node sections. The dataset comprises slides obtained from 399 breast cancer patients who underwent surgery at two hospitals in the Netherlands[2] Ehteshami Bejnordi et al. (2017). The slides were acquired using two different scanners and were annotated under the supervision of expert pathologists. Each WSI is assigned a binary label, either '1' (indicating the presence of tumorous cells) or '0' (indicating their absence). In cases where tumorous cells are present, a distinction is made between micrometastases, characterized by tumor cell diameters ranging from 0.2 to 2 mm, and macrometastases, where the tumor cell diameter exceeds 2 mm. Examples of WSIs belonging to each class are depicted in Fig. 1.



(a) Normal

(b) Macro

(c) Micro

Figure 1: Examples of slides with different labels. Tumorous cells are in red.

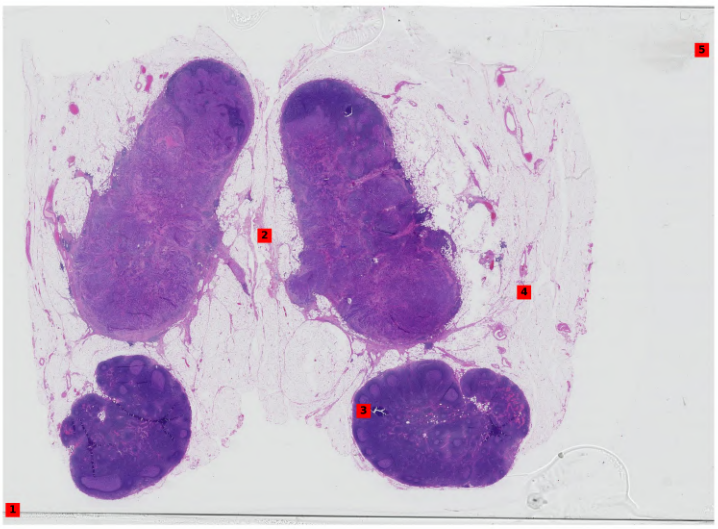
3.2 Train/Validation/Test split

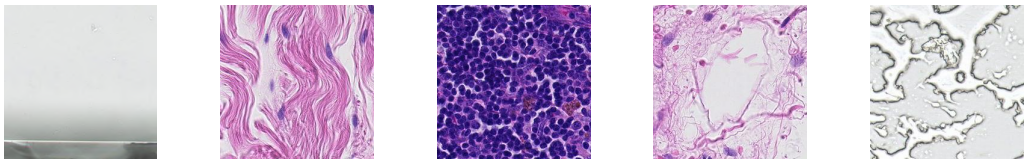
The train-test split of the Camelyon16 dataset was predefined by its authors, who stated that the split was performed randomly while ensuring that both subsets contain sufficient micro- and macro-metastatic images. The training set comprises 270 WSIs, including 159 normal slides, 53 with micrometastases, and 58 with macrometastases. It is further split into training and validation data by including 80% of each case aggregated together in the training set and the remaining 20% aggregated in the validation set, see metadata. This ensured that different cases were equally distributed across the datasets. The test set consists of 80 normal images and 49 metastatic images, of which 27 contain micrometastases and 22 contain macrometastases. A distinguishing feature of the Camelyon16 dataset is that all images were acquired from different patients, which is not always the case in similar datasets (e.g. TCGA). Yet, for the sake of the third milestone, we implemented a custom splitting function that ensures WSIs from a single patient are not distributed across multiple subsets.

[2]These hospitals were Radbound University Medical Center (RUMC) and University Medical Center Utrecht (UMCU).

3.3 Slide-to-Patch splitting

WSIs are high-resolution images with dimensions of approximately 150.000×150.000 pixels, requiring between 0.5 and 4 GB of storage space per slide. Due to memory limitations, processing and training of ML models on entire slides is infeasible Jafarinia et al. (2024). To address this, each slide is divided into smaller, uniformly sized patches through a systematic transformation process. First, a grid is defined over the slide, splitting it into patches of equal size. Then, for each patch, the tissue coverage is evaluated against a predefined threshold. Patches exceeding the threshold are retained, while those below it are discarded. The positions of the retained patches are stored in the metadata, enabling the reconstruction of the original slide. An example of such transformation is illustrated in Fig. 2, where positions of patches 1 - 5 are marked in the original slide.





(a) 1 (b) 2 (c) 3 (d) 4 (e) 5

Figure 2: Individual patches from a WSI with their positions marked in red.

3.4 Multiple Instance Learning

In the Camelyon16 dataset[3], annotations are available for each patch, which is usually not the case in other histopathology datasets where only slide-level labels are provided. This characteristic makes the application of MIL particularly suitable for such problems. Drawing a comparison to the MNIST-Bags dataset, a whole-slide image can be viewed as a bag of instances, while the individual patches correspond to single images of digits. Thus, a WSI can be conceptualized as a bag of individual patches extracted from it. The model's objective is then to predict whether an unseen slide contains tumorous cells or not. Furthermore, it is desirable to identify which patches contribute most to the final slide-level prediction, which can be inferred using the obtained attention weights.

[3]Preprocessing of histopathology data is also typically required, which involves tasks such as tissue detection and color normalization to account for variations caused by different scanning devices. Additionally, a feature extraction process is necessary. However, these preprocessing steps have already been performed by previous researchers for this dataset.

4 Histopathology method and evaluation

4.1 Overview

In milestone 2, we used the existing baseline MIL architectures from milestone 1 to apply them on the histopathology dataset. Since we already implemented the LeNet5 feature extractor on our own, this time we rely on an existing foundation model for histopathology data Lecun et al. (1998). Wang et al. (2022) proposed a transformer-based unsupervised feature extractor in their work. The model is called 'CTranspath' and processes the patches of a slide $p \in \mathbb{R}^{H \times W \times 3}$ as a whole (without another partition on patch level) through CNN layers followed by swin transformer blocks. The final result is a 1×768 latent meaningful feature vector. Fortunately, a feature matrix $F \in \mathbb{R}^{K \times 768}$ with feature vectors for every K patches of one bag (slide) of the Camelyon16 dataset is provided. Therefore, the training pipeline in this milestone consists only of the MIL aggregation parts and the prediction head (one fully-connected layer) to provide the final bag level prediction $\hat{\theta}(X)$. The main focus of our work lies in model interpretability.

4.2 Optimization and Metrics

Since the problem statement does not significantly differ from milestone 1, we mostly refer to sections 3.3 and 3.4 from the previous report. We optimize the neg-log-likelihood using the Adam optimization Kingma and Ba (2017) to fit the predicted probabilities to the true label distribution of bags. As for the baseline method, we evaluate the model's performance through several metrics, namely accuracy/error, precision, recall, f1-score and area under curve (AUC). During training, we calculate these metrics every 5 epochs on a curated validation set. After the training is finished, either based on reaching the maximum number of epochs (100) or because of the stopping criterion with a patience of 2, which is based on a combined validation loss + error, the aforementioned metrics are evaluated finally on the test set 3.2.

4.3 Experiments and Results

For the experiments in milestone 2, we are moving away from the approach of exactly replicating the results of Ilse et al. (2018), because we are using the Camelyon16 dataset, which differs from the breast cancer dataset Drelie Gelasca et al. (2008) and colon cancer dataset Sirinukunwattana et al. (2016). Nonetheless, our research questions are the following: (i) does the novel attention based MIL approach hold on real world histopathology data in terms of bag classification performance and outperform the previously used mean/max operators, (ii) how meaningful are the model's attention weights from the attention embedding approach and the model's instance scores from the instance mean/max approach. Detailed instructions on how to replicate our experiments can be found in the ReadMe.md of our repository. A selection of training logs is provided in this WandB report.

Details We set up our experiments by using all 399 available bags (slides), split into the training, validation and test set by the aforementioned strategy 3.2. Every MIL approach - in total six different architectures - used in milestone 1, is taken into consideration in these experiments as well. The Adam optimizer operates in our experiments on the default values $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to weight the exponential decay of the two moments. The weight decay parameter, which should improve the generalization of the model due to regularization of the weights, is chosen based on a hyperparameter tuning, as well as the learning rate. We decided to perform a discrete hyperparameter tuning for better comparability and due to computation efficiency and time reasons. The parameter for the feature dimension is fixed by the output of the CTranspath Wang et al. (2022) feature extractor's output and the parameter for the attention space dimension is set to $d_{att} = 256$, as tests for $d_{att} = \{128, 256, 512\}$ revealed only minor changes. A detailed description of our hyperparameter tuning is provided in Tab. 4. The whole experiment is set up as a grid search in Weights and Biases WandB (2024). As stated in the Histopathology MIL problem formulation 3.4, the key instances are the tumor patches in a bag.

Table 4: Different hyperparameter settings used in the Histopathology experiments.

Parameter	Values
mode	{instance, embedding}
pooling type	{mean, max, attention, gated attention}
# bags	399
attspace dim	256
learning rate	{1e-4, 5e-4, 1e-3, 5e-3}
weight decay	{1e-2, 1e-3, 1e-4}

Training The results of the hyperparameter tuning are provided in Sec. 4.3, but in this section we provide general observations of the neg-log likelihood training loss, which occur in every hyperparameter setting. Fig. 3 shows that the attention (red) and gated attention (yellow) approach as well as the max (purple) instance approach lead to fast convergence of the training loss. In turn, we want to highlight that the mean operator fails to converge for both the embedding (green) and instance (blue) approach for all different hyperparameter settings. We refrained from further inspection of this behavior because the working approaches are sufficient to inspect model's interpretability. The max embedding (pink) approach converges in general but more slowly, then the training is stopped based on the stopping criterion. In summary, the attention approaches and the max instance approach succeeded and thus will be the main focus in Sec. 4.4 on model interpretability and visualization.

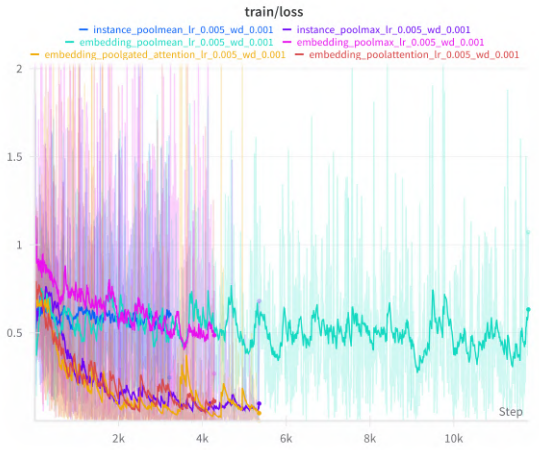


Figure 3: Training loss for learning rate = 0.005 and weight decay = 0.001.

Results The hyperparameter tuning revealed that there is a slight positive correlation (see App. Fig. 12) between the AUC score on the test set and the learning rate which suggests and verifies to use the largest learning rate lr = 0.005 of our set. Throughout all combinations of the grid search, the weight decay parameter has a small negative correlation with the AUC score (see App. Fig. 13). Taking into consideration this observation and the absolute results, the optimal weight decay parameter is wd = 0.001. This WandB report provides detailed information about all results of the hyperparameter tuning regarding the influence of parameters and their correlation with a chosen metric, e.g. test/AUC. With the chosen learning rate and weight decay, the qualitative results on the test set, listed in Tab. 5, justify hypothesis (i) in terms of the discriminative power the attention and gated attention approaches have. Nevertheless, the instance approach with max operator performs comparably well in most test metrics. In contrast, the other three approaches have significantly less discriminative power. The max embedding approach has a relatively high standard deviation for precision and recall, which suggests that with more training and fine-tuning its results could be more stable, making this approach more robust.

Table 5: Results on the Camelyon16 dataset. Experiments were run 5 times and an average (± a standard error of the mean) is reported.

Method	Accuracy	Precision	Recall	F1-Score	AUC
Instance+max	0.918 ± 0.004	**0.941 ± 0.010**	0.837 ± 0.013	0.885 ± 0.006	0.963 ± 0.008
Instance+mean	0.583 ± 0.051	0.425 ± 0.033	0.233 ± 0.204	0.250 ± 0.152	0.490 ± 0.020
Embedding+max	0.645 ± 0.062	0.779 ± 0.245	0.388 ± 0.274	0.398 ± 0.158	0.705 ± 0.009
Embedding+mean	0.626 ± 0.031	0.528 ± 0.048	0.331 ± 0.081	0.397 ± 0.046	0.532 ± 0.008
Attention	**0.935 ± 0.004**	0.933 ± 0.022	**0.894 ± 0.015**	**0.913 ± 0.004**	0.965 ± 0.011
Gated Attention	0.924 ± 0.019	0.930 ± 0.065	0.873 ± 0.035	0.898 ± 0.019	**0.967 ± 0.010**

4.4 Model interpretability

While the models demonstrate substantial discriminative abilities, the (ii) research question aims to discover the interpretability of the learned attention weights for the embedding approach Sec. 4.4 and the instance scores for the instance max approach Sec. 4.4.

Attention weights The idea behind attention weights is to represent the importance of each single patch p contributing to the decision of the prediction head. As $\sum_{k=1}^{K} a_k = 1$ applies, a heatmap visualization (0: white (low attention score), 1: darkred (high attention score)) is a useful method to provide insights. If attention weights provide meaningful information, MIL can be used to detect tumor regions, which is close to a rough segmentation. The first observation during training of the attention embedding approach is that during the first epochs, when the validation metrics perform poorly and the training loss has not yet converged, the attention weights are distributed almost uniformly in target regions. As the performance on the validation set increases and the training loss decreases, the meaningful attention weights seem to 'collapse'/'overfit' to a small number of highly contributing patches. This behavior is reasonable with respect to the formulation of the MIL problem, see Ilse et al. (2018), because the objective of MIL takes the deviation of prediction from ground truth just on bag level into account. Therefore, one expressive key instance, indicating whether there is a tumor in the bag/slide or not, is sufficient to impact a bag level classifier. Thus, we conclude that even though (i) is sufficiently fulfilled, (ii) is mostly not. This holds for both 'macro' and 'micro' cases (see Fig. 4 and Fig. 5 respectively). Additionally, some examples of high attention weights at the borders of the slide occur, even though the corresponding patches do not represent the actual cells (see Fig. 6). Further work should improve the extraction method of relevant patches from a slide in the preprocessing step, see Sec. 3.3, so that it is ensured that these 'border' patches are not included in a bag at all.

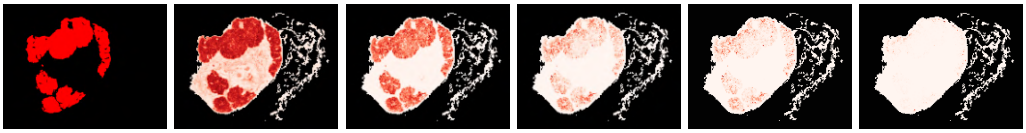


Figure 4: **Attention macro tumor region:** Left: Ground truth annotation of tumor region in red. Left+1: Attention weights a_k for every patch of one bag (ID: *tumor_110* from Camelyon16) after epochs $3, 4, 5, 6, 9$.

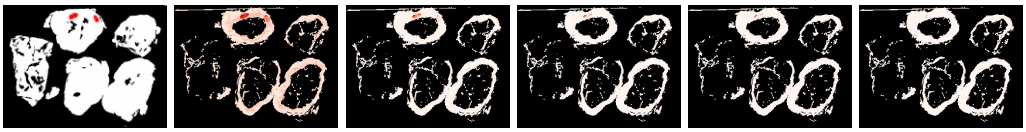


Figure 5: **Attention micro tumor region:** Left: Ground truth annotation of tumor region in red. Left+1: Attention weights a_k for every patch of one bag (ID: *tumor_083* from Camelyon16) after epochs $3, 4, 5, 6, 9$.

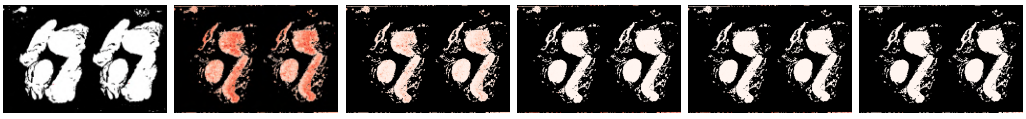


Figure 6: **Attention border highlights:** Left: Ground truth annotation of tumor region in red. Left+1: Attention weights a_k for every patch of one bag (ID: *tumor_086* from Camelyon16) after epochs $3, 4, 5, 6, 9$.

Developing more meaningful/interpretable MIL approaches and eXplainable AI methods is a possible direction to follow in milestone 3. For the current milestone, we introduce different visualization techniques without utilization of raw attention weights. Both techniques aim to manipulate the raw attention weights for visualization purposes only, so that research question (ii) is fulfilled as well.

Logarithmic normalization Using logarithmic normalization, large values are compressed and small values are expanded, which leads to greater uniformity of attention weights a_k. This normalization highlights small differences and provides smoother scaling than the raw attention weights of which the 'macro' tumor regions benefit. It is defined as $a_k^{\log} = \frac{\log(a_k+\epsilon)-\log(a_{\min}+\epsilon)}{\log(a_{\max}+\epsilon)-\log(a_{\min}+\epsilon)}$, where ϵ is a small constant to prevent numerical problems.

Percentile normalization Using percentile normalization, extreme outliers are removed from the attention weights vector a. This has the advantage that in the target regions of tumor patches the attention weights are more uniformly distributed. We discovered that this technique is in turn beneficial in 'micro' tumor regions. Due to the applied heuristic, extreme outliers are excluded, yet an emphasis on large attention weights for a small number of patches still persists. It is defined as $a_k^{\text{perc}} = \frac{\min\left(\max(a_k, a_{p_{\min}}), a_{p_{\max}}\right) - a_{p_{\min}}}{a_{p_{\max}} - a_{p_{\min}}}$, where $p_{min} = 1$ and $p_{max} = 99$ are the percentiles used.

Normalization evaluation Fig. 7 shows the benefits of the logarithmic normalization in comparison to the raw attention weights and the percentile normalization. The logarithmic normalization enables capturing even small tumor regions of the 'macro' tumor (depicted by the blue circle in the figure), whereas the raw attention weights provide no meaningful information about either location or size of the tumor. The percentile normalization highlights only the most confident areas of potential tumorous cells. In turn, the percentile normalization is well suited for the 'micro' tumor cases, which is shown in Fig. 8. The blue circles indicate the region of interest for tumor detection, whereas the orange circle highlights incorrectly marked patches. Therefore, percentile normalization is not a sufficiently accurate technique at all, but serves as an improvement in visualizing meaningful attention weights for micro tumor regions. In general, we observe a considerably higher significance of the attention weights in the macro tumor cases than in the micro tumor ones.

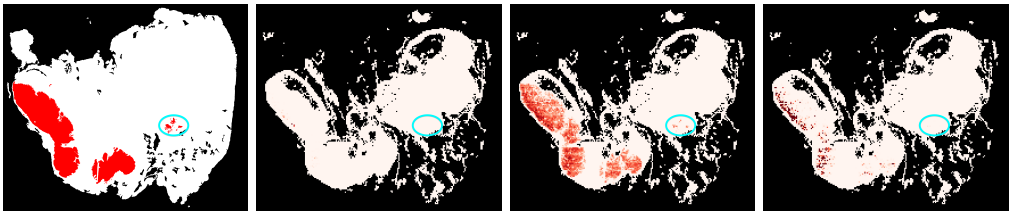


Figure 7: **Normalization to highlight macro tumor regions:** From left to right: ground truth annotation, raw attention weights, logarithmic normalization, percentile normalization. (ID: *test_016* from Camelyon16, Zoom: no zoom applied)

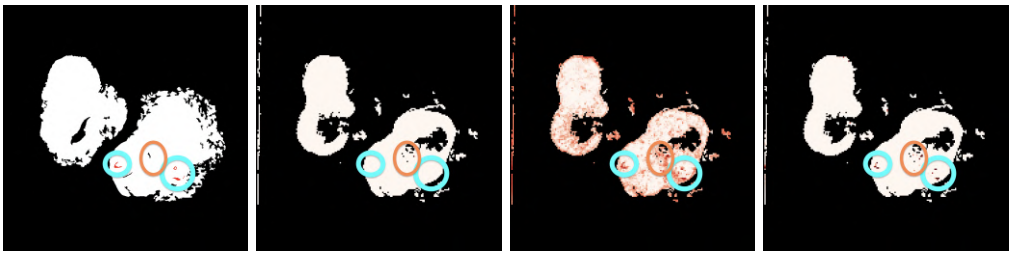


Figure 8: **Normalization to highlight micro tumor regions:** From left to right: ground truth annotation, raw attention weights, logarithmic normalization, percentile normalization. (ID: *test_038* from Camelyon16, Zoom: zoom applied to highlight region of interest)

Instance scores Apart from attention heatmaps, the instance approaches provide for every single patch of a bag the output of the fully connected layer, which serves as a score for the corresponding instance. These scores are then aggregated and a final bag-level classification is performed. Therefore, these scores should contain meaningful information about tumor region patches and, after they have been postprocessed, can serve as an instance score heatmap, providing broad tumor segmentation areas. Fig. 9 and Fig. 10 show the significance of instance scores for a macro and a micro tumor region, respectively. In comparison to the attention heatmaps, the instance scores serve as a more reliable method of model interpretation, since no further scaling or normalization is needed to disregard single collapsed or overfitted attention weights.

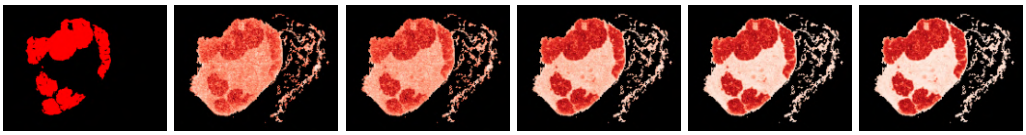


Figure 9: **Instance macro tumor region:** Left: Ground truth annotation of tumor region in red. Left+1: Instance scores for every patch of one bag (ID: *tumor_110* from Camelyon16) after epochs $2, 3, 4, 5, 6$.

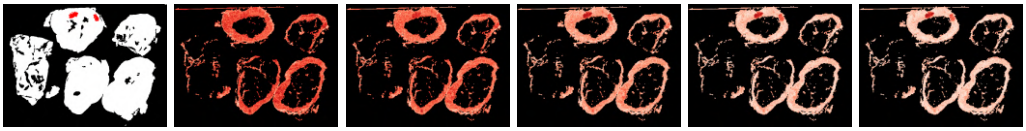


Figure 10: **Instance micro tumor region:** Left: Ground truth annotation of tumor region in red. Left+1: Instance scores for every patch of one bag (ID: *tumor_083* from Camelyon16) after epochs $2, 3, 4, 5, 6$.

5 Discussion

5.1 Challenges & Possibilities

The training process for most approaches requires approximately 20 minutes for 100 epochs. However, not all approaches were trained for the full 100 epochs, as the stopping criterion was often met earlier. Retraining or extending the training of MIL models with new data is feasible but highly dependent on the characteristics of the incoming data. Successful integration depends on normalizing and aligning the new data with the distribution of the Camelyon16 dataset. Although retraining with out-of-distribution data might be possible, there is no guarantee that the models will generalize effectively to such data. Generalizability is primarily assessed using a curated test set, which, while rigorously validated, originates from the preprocessed Camelyon16 dataset. The confidence scores generated by instance-based approaches are processed through the MIL pooling layer, meaning that the raw instance probabilities are only intermediate outputs. These scores, as discussed, improve model interpretability. In contrast, embedding-based approaches apply a nonlinear activation function as the final operation, allowing the predicted probabilities to be interpreted as confidence measures. The performance metrics computed on the test set evaluate the quality of these confidence scores. A comparison with the validation metrics, see Fig. 14, reveals consistent results among the successful approaches, indicating reasonable reliability. However, the mean-based approaches, which perform poorly on the test set, demonstrate significantly better results on the validation set.
To conclude, in milestone 2 the MIL approach is successfully applied to real world histopathology data. Especially the attention-based and max instance-based approaches achieve excellent test performance results regarding the discriminative ability of the bag level classifier. The mean approach performs significantly worse. Both the attention and max approaches provide meaningful information from different layers to create an interpretable heatmap visualization for tumor region detection. Regarding the significance of the attention weights, Javed et al. (2022) introduced Additive MIL, which can be added to any MIL approach by some function recomposition. This approach offers determining exact patch contribution towards a prediction, benefits for multi-class problems and distinction between excitatory and inhibitory contributions. In the latest research by Hense et al. (2024) xMIL-LRP is introduced, which offers the same distinction between both types of contributions and the same conservation. Moreover, it also includes context sensitivity to enable capturing dependencies between features in a whole bag.

5.2 Future Work

For milestone 3, we could investigate out-of-distribution learning and applications on larger datasets, such as the cancer genome atlas TCGA (2006) dataset. Another possibility would be to focus on improving the preprocessing step, since the current patch extraction method considers some border artifacts to be valid patches, which may disturb the attention mechanism, as illustrated in Fig. 6. The most promising direction to follow, would be to further investigate model interpretability, utilizing the aforementioned Additive MIL and/or xMIL-LRP techniques, which would help us obtain more meaningful insights into the model.

References

E. Drelie Gelasca, J. Byun, B. Obara, and B. Manjunath. Evaluation and benchmark for biological image segmentation. In *2008 15th IEEE International Conference on Image Processing*, pages 1816–1819, San Diego, CA, USA, 2008. IEEE. ISBN 978-1-4244-1765-0. doi: 10.1109/ICIP.2008.4712130. URL http://ieeexplore.ieee.org/document/4712130/.

B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 12 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.14585. URL https://doi.org/10.1001/jama.2017.14585.

J. Hense, M. J. Idaji, O. Eberle, T. Schnake, J. Dippel, L. Ciernik, O. Buchstab, A. Mock, F. Klauschen, and K.-R. Müller. xMIL: Insightful Explanations for Multiple Instance Learning in Histopathology, Nov. 2024. URL http://arxiv.org/abs/2406.04280. arXiv:2406.04280.

M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

H. Jafarinia, A. Alipanah, D. Hamdi, S. Razavi, N. Mirzaie, and M. H. Rohban. Snuffy: Efficient whole slide image classifier, 08 2024. URL https://arxiv.org/pdf/2408.08258v2.

S. A. Javed, D. Juyal, H. Padigela, A. Taylor-Weiner, L. Yu, and A. Prakash. Additive MIL: Intrinsically Interpretable Multiple Instance Learning for Pathology, Oct. 2022. URL http://arxiv.org/abs/2206.01794. arXiv:2206.01794 [cs].

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization, Jan. 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998. ISSN 1558-2256. doi: 10.1109/5.726791. URL https://ieeexplore.ieee.org/document/726791/?arnumber=726791. Conference Name: Proceedings of the IEEE.

K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot. 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, May 2016. ISSN 1558-254X. doi: 10.1109/TMI.2016.2525803. URL https://ieeexplore.ieee.org/document/7399414/?arnumber=7399414. Conference Name: IEEE Transactions on Medical Imaging.

TCGA. The Cancer Genome Atlas Program (TCGA) — cancer.gov. https://www.cancer.gov/ccg/research/genome-sequencing/tcga, 2006. [Accessed 14-12-2024].

WandB. Weights & Biases: The AI Developer Platform — wandb.ai. https://wandb.ai/site/, 2024. [Accessed 13-11-2024].

X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, Oct. 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102559. URL https://www.sciencedirect.com/science/article/pii/S1361841522002043.

A Appendix

A.1 Camelyon16 hyperparameter tuning

The results of the WandB hyperparameter tuning provide a correlation and importance analysis, Fig. 12, of the learning rate, Fig. 11, and weight decay, Fig. 13, regarding the AUC score on the test set.

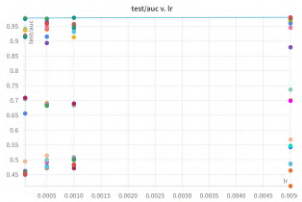


Figure 11: Results of AUC test score (y-axis) for different MIL approaches displayed against different learning rates (x-axis).

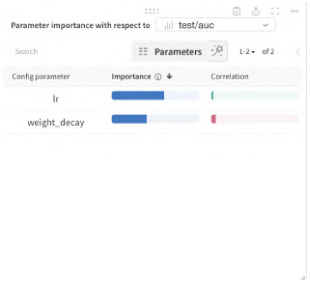


Figure 12: Correlation and importance of hyperparameters.

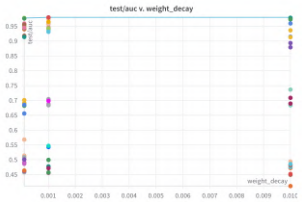


Figure 13: Results of AUC test score (y-axis) for different MIL approaches displayed against different weight decays (x-axis).

A.2 Camelyon16 validation performance metrics

These are the validation metrics during training after every 5th epoch. The validation metrics are similar to the test metric performances but especially the mean approaches perform better on the validation set than on the test set which suggests not a high reliability of the method for these approaches.

Figure 14: **Camelyon16 validation metrics:** For all six different approaches the accuracy, precision, recall, f1-score, AUC score and the validation loss is displayed.