# Project Machine Learning
# — Milestone 3 —

Cederic Aßmann, Friedrich Hagedorn, Jakub Sliwa

January 29, 2025

## 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 third milestone. The tasks within its scope include evaluation of our final model, as well as utilizing multiple eXplainable AI (XAI) methods in order to investigate model's interpretability[1].

The structure of the report is as follows: Section 1 provides an overview of the report's purpose and structure. Section 2 briefly describes different XAI methods, namely Gradient × Input, Shapley Values and Layer-wise Relevance Propagation (LRP), which were used to examine model's explainability. Section 3 focuses on the experiments conducted during this milestone, which include applying various previously described XAI methods, and discusses obtained results. This section also covers additional evaluation of the final method. Section 4 analyzes time-efficiency of our automated prediction system and gives possible recommendations, interprets the achieved results and summarizes the work done within the scope of the entire project. In the Appendix, all the False Negative cases from the test dataset evaluation are presented A, since these, due to their irreversible consequences, should be avoided at all cost.

## 2 XAI & Interpretability

Deep neural networks have had a significant impact on both science and industry. However, understanding the mechanisms behind their predictions is crucial, especially in the medical domain, where reliability and interpetability are of particular importance. One approach to evaluate a model's explainability is to assign an attribution value to each input feature, thereby quantifying its contribution to the model's output. By organizing the relevance scores of all input features to match the shape of the input, attribution maps can be generated. Visualizing them as heatmaps may provide valuable insights into the specific regions of an image that influenced the network's predictions. In this section we present XAI methods we utilized in order to further investigate the model's interpretability.

### 2.1 Gradient × Input (SmoothGrad)

Gradient × Input is a backpropagation-based method that computes attribution values in a single forward and backward pass Ancona et al. (2018). Attribution values are determined by calculating the partial derivatives of the network's output with respect to each input feature and multiplying them by the input itself, as shown in Eq. 1, where attribution value and network's output are denoted by $R_i(x)$ and $\hat{\theta}(x)$ respectively.

$$\mathrm{R_i}(x) = x_i \cdot \frac{\partial \hat{\theta}(x)}{\partial x_i} \tag{1}$$

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

To improve gradient stability, we applied SmoothGrad, a technique that can be combined with various sensitivity map algorithms. This method involves adding small amounts of noise to the input, executing Gradient × Input multiple times (ten iterations in our case), and averaging the resulting attribution maps. This approach not only enhances stability but also reduces noise in the attribution maps Smilkov et al. (2017).

## 2.2 Shapley Values

Shapley Values, originally derived from game theory, were proposed as a unified measure of feature importance Lundberg and Lee (2017). The authors of SHAP proposed to perceive any explanation of a model's predictions as an independent *explanation model*. They argued that SHAP is the only method that satisfies three key properties: local accuracy, missingness, and consistency, and other methods simply aim to approximate it. Shapley Values allocate to each feature the change in the expected prediction when conditioned on that feature. They explain the transition from a baseline value, which represents the prediction of the model without any available features, to the actual output of the model. Since the order in which features are added into the expectation matters, Shapley Values are computed by averaging feature importance across all possible orderings. The exact calculation of Shapley Values is computationally expensive ($O(2^N)$, where $N$ is the number of features); therefore in this work, we approximate Shapley Values using a Monte Carlo sampling approach. For a set of $N$ instances, random subsets $S \subseteq N \setminus \{i\}$ are sampled, and the marginal contribution of each instance $i$ is computed as the difference between the model's prediction with and without $i$ included in the subset:

$$\Delta_i(S) = f(S \cup \{i\}) - f(S).$$

The Shapley Value for instance $i$ is then approximated by averaging the marginal contributions over multiple random subsets:

$$\phi_i = \frac{1}{M} \sum_{j=1}^{M} \Delta_i(S_j),$$

where $M$ is the number of sampled subsets. This method reduces the computational complexity which would otherwise be unfeasible in the case of large histopathology slides ($N = 10.000$) for example, while providing an efficient and scalable means of approximating Shapley Values.

## 2.3 Layer-Wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) is an explanation framework for neural networks that operates by propagating the model's prediction backwards using local propagation rules. Each layer adheres to the conservation axiom, which ensures that the relevance scores assigned to the input variables sum to the network's output. In other words, the relevance received by a layer must be fully redistributed to the layer below, as expressed in Eq. 2, where $x_i$ and $y_j$ denote vectors of neurons representing the input and output of some layer, respectively. This equation reflects the LRP view on Gradient × Input Ali et al. (2022).

$$R(x_i) = \sum_j \frac{\partial y_j}{\partial x_i} \frac{x_i}{y_j} R(y_j) \tag{2}$$

In our implementation, we employed an optimization trick that omits explicitly stating the propagation rule by treating certain terms as constants (detached from the forward pass). Additionally, we utilized the $LRP - \epsilon$ variant, which introduces a small positive term in the denominator to reduce noise and produce more stable results Binder et al. (2016).

# 3 Final method

The final method for the Attention-Deep Multiple Instance Learning task was already proposed in milestone 2. For the details, regarding the hyperparameter selection and the final results in terms of a detailed evaluation including the accuracy, precision, recall, F1-score and AUC-score, see Sec. 4.3 in the milestone 2 report. In addition to these outcomes, we would like to highlight in this section further implications from the results and provide a clear and complemented overview about the chosen XAI methods.

## 3.1 Further Results

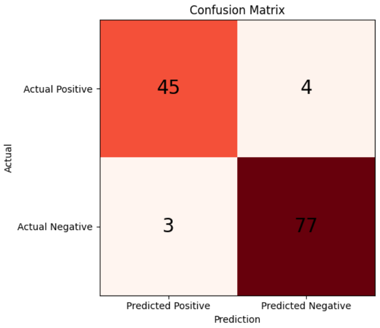


Figure 1: Confusion matrix, test dataset

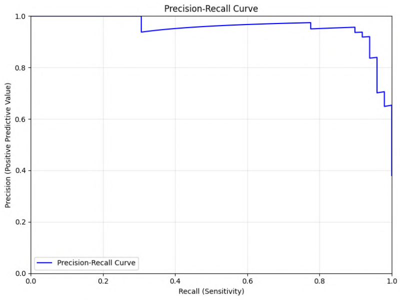


Figure 2: Precision-Recall curve, test dataset

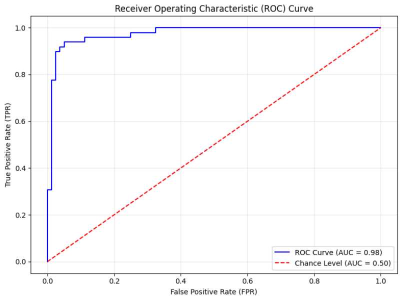


Figure 3: ROC curve, test dataset

In this section, we would like to address the critical part of the automated cancer prediction, focusing on the classification of the bags of instances into 'tumor' and 'non-tumor' categories. Accurate predictions in this context are essential, as False Negatives could lead to missed diagnoses with potentially severe consequences or delayed treatment, while False Positives may result in unnecessary follow-ups or interventions. The consequences are less severe than missing a cancer diagnosis, therefore the main focus should be on the number of false negative cases. To further evaluate the performance of our attention based MIL model with a focus on the aforementioned context, we employ a combination of metrics and visualizations that offer complementary insights into its efficacy and reliability. This further analysis is performed on the Camelyon16 test dataset, further details on how the dataset is splitted can be found in milestone 2 report. We analyze the following:

- Confusion Matrix: Highlighting True Positives, True Negatives, False Positives, and False Negatives so that the models performance can be determined for each test case.
- Precision-Recall Curve: This is particularly valuable in this imbalanced tumor prediction scenario where False Negatives carry significant implications.
- ROC Curve and AUC: In addition to the previously provided AUC score, we would like to also visualize the model's discriminative ability across varying classification thresholds.

Tab. 5 in milestone 2 report shows that Precision and Recall scores have a relatively low standard deviation of the results performed on 5 training runs. Therefore, we decided to show the confusion matrix 1 as well as the Precision-Recall curve 2 and the ROC curve 3 exemplary based on the evaluation on one training run. The confusion matrix depicts that there are in total 7 cases out of 129 falsely predicted. False Positives (3) and False Negatives (4) cases are balanced. The false negative cases (test_011, test_013, test_066, test_099), see App. 11, are the critical ones which a model should aim to completely avoid. Fairly, one needs to admit that these missed tumor cases are relatively difficult ones of micro tumors. The Precision-Recall curve shows in general the near perfect discriminative ability of the Attention MIL model, however, there is a trade-off between Precision and Recall. To be robust against False Negatives, the Recall score should be about 1.0. Fig. 2 suggests to lower the threshold (standard: positive case: $\hat{\theta}(X) > 0.5$, negative case: $\hat{\theta}(X) < 0.5$). This would lead to a higher Recall score but on the other hand the number of False Positives would increase. Regarding the tumor prediction task, we might accept this kind of error of the model as it has much lower impact on the patient and therefore lower the threshold. The ROC curve 3 shows nearly the same results and implications.

### 3.2 XAI

The milestone 2 report covered a detailed Sec. 4.4 about model interpretability using the attention weights and applying specific normalization techniques to enhance the meaningfulness of certain cases. This section provides further XAI results from Gradient × Input, Shapley Values and LRP. Moreover, we discovered an approach combining the attention scores with LRP, resulting in more meaningful explanations and guidance towards tumorous instances.

**LRP + Attention weights** The idea of combining the relevance scores together with the attention scores arose from the observation that for some cases one or the other method steers in the right direction but is not able to fully capture in a meaningful and wholesome way which instances are responsible for the tumor prediction. By default, the attention scores are normalized so that $a_k \in [0, 1]$. The Relevance scores as they should capture both excitatory and inhibitory instances are normalized by our default $\mathcal{R}(x_i) \in [-1, 1]$. We decided to apply the log normalization $log(\mathcal{R}(x_i)) \in [0, 1]$ onto the Relevance scores and renormalize the attention scores $a_k \in [-1, 1]$. Then, both scores are simply added together. This follows the idea to capture information out of both methods, enforcing the norm of values with equal sign and degrading the norm of values with unequal sign. The negative contributions arises in most cases from the attention scores as it has proven to be more reliable. In practice, interpretability benefits from this approach as depicted in Fig. 4 where it is compared to both methods performed separately and the ground truth annotation from a histopathologist.

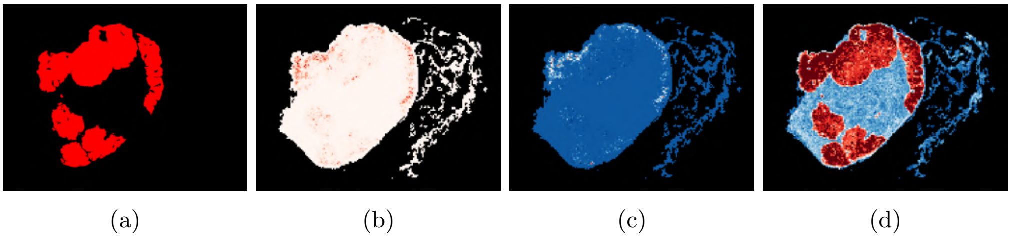
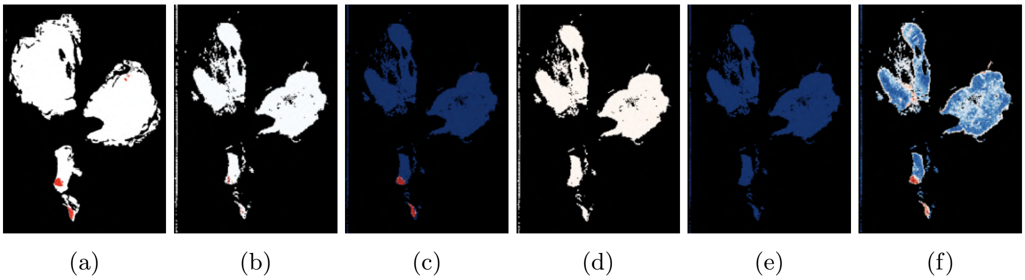


Figure 4: **LRP + Attention weights: a)** Ground truth annotation of tumor region in red **b)** Raw Attention weights $a_k$ **c)** Raw Relevance scores $\mathcal{R}(x_i)$ **d)** Relevance scores + Attention scores, (normalized with LRP+Att. technique); (ID: *tumor_110* from Camelyon16)

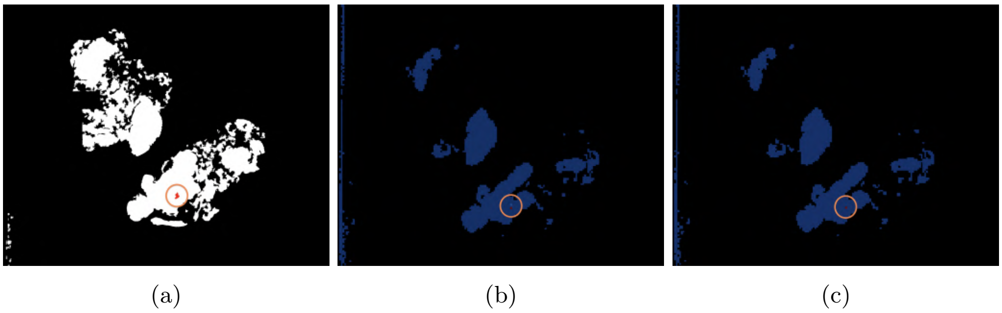
**Comparing XAI methods** The following section is about a comparison between the Gradient × Input (SmoothGrad), LRP, Shapley Value and Attention Weight explanations. All methods are compared with each other and the ground truth annotations serve as the reference. A general observation is that there is not one XAI method which is robustly applicable to every tumor case. The macro tumor regions are generally much better captured than the micro tumor regions, see also our findings regarding False Negatives in Sec. 3.1. The Shapley Value calculation does not scale well as the number of instances increases ($O(2^N)$) and unfortunately one bag of the Camelyon16 dataset consists of ∼10.000 instances. Therefore, Shapley Values are not the best choice as an interpretability method for the MIL problem. Nevertheless, the insights this method gives, are in some cases e.g. in Fig. 5 more meaningful and closer to the ground truth annotations than the other methods.



(a) (b) (c) (d) (e) (f)

Figure 5: **Shapley Values: a)** Ground truth annotation of tumor region in red **b)** SmoothGrad (log normalized) **c) Shapley Values d)** Raw Attention weights $a_k$ **e)** Raw Relevance scores $\mathcal{R}(x_i)$ **f)** Relevance scores + Attention weights (normalized with LRP+Att. technique); (ID: *test_002* from Camelyon16)

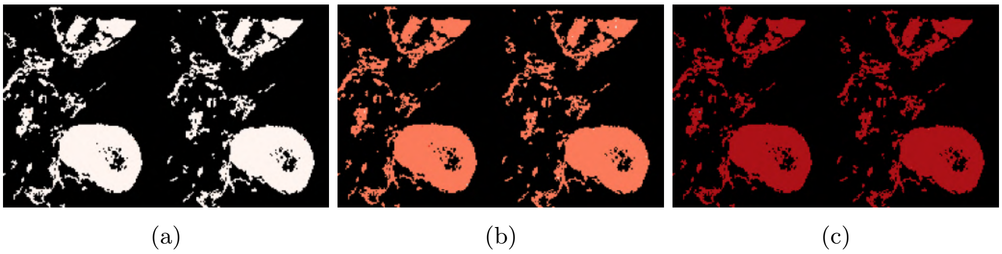
The SmoothGrad and the LRP explanations do not differ much as our Attention MIL model does not consist of many different layers where the benefits and refinements of LRP could be fully utilized compared to the solely gradient at the input layer, see Fig. 6.



(a) (b) (c)

Figure 6: **LRP and SmoothGrad: a)** Ground truth annotation of tumor region in red. **b) SmoothGrad c) Raw Relevance scores** $\mathcal{R}(x_i)$; (ID: *test_063* from Camelyon16)

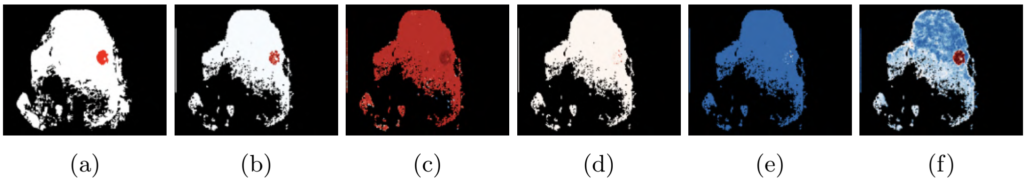
Fig. 7 depicts that for the negative cases, not a single instance consists of tumorous tissue, the raw attention weights are more accurate than the explanations of SmoothGrad and LRP. Where the attention weights are all 0, the Relevance scores seem to be more or less uniformly distributed but non-zero, which leads to these patterns of assigned importance across the whole bag of instances.



(a) (b) (c)

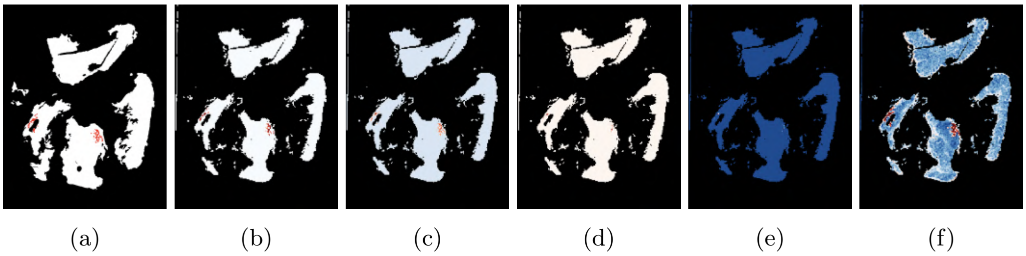
Figure 7: **Attention weights are best for negative cases: a) Raw Attention weights** $a_k$ **b)** SmoothGrad **c)** Raw Relevance scores $\mathcal{R}(x_i)$; (ID: *test_012* from Camelyon16)

The final observation from this comparison of these different XAI methods is that at some point, interpretability and discriminative power of the Attention MIL model is bounded by the dataset itself. More data, more variety in the data could strengthen the model which leads to more accurate information laying in the model which can be interpreted more reliable. Through the combination of LRP and the attention weights we were able to finetune the XAI methods and generate convincing explanations, see Fig. 8 and Fig. 9, but to some degree we just combined information originating from the same model, see not so convincing explanations in Fig. 10.



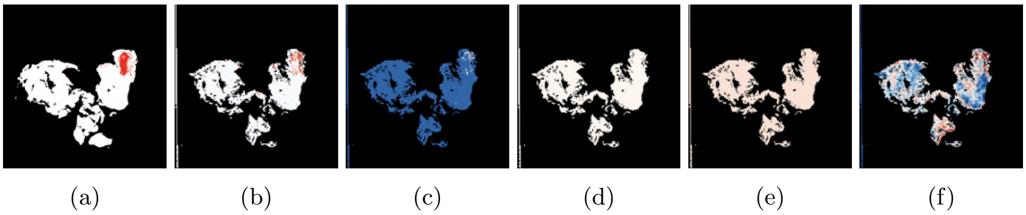
(a) (b) (c) (d) (e) (f)

Figure 8: **LRP + Attention weights (is all you need): a)** Ground truth annotation of tumor region in red **b)** SmoothGrad (log normalized) **c)** Shapley Values **d)** Raw Attention weights $a_k$ **e)** Raw Relevance scores $\mathcal{R}(x_i)$ **f) Relevance scores + Attention weights** (normalized with LRP+Att. technique); (ID: *test_027* from Camelyon16)



(a) (b) (c) (d) (e) (f)

Figure 9: **LRP + Attention weights (is really! all you need): a)** Ground truth annotation of tumor region in red **b)** SmoothGrad (log normalized) **c)** Shapley Values **d)** Raw Attention weights $a_k$ **e)** Raw Relevance scores $\mathcal{R}(x_i)$ **f) Relevance scores + Attention weights** (normalized with LRP+Att. technique); (ID: *test_061* from Camelyon16)



(a) (b) (c) (d) (e) (f)

Figure 10: **LRP + Attention weights (is definitely NOT all you need): a)** Ground truth annotation of tumor region in red **b)** SmoothGrad (log normalized) **c)** Shapley Values **d)** Raw Attention weights $a_k$ **e)** Raw Relevance scores $\mathcal{R}(x_i)$ **f)** Relevance scores + Attention weights (normalized with LRP+Att. technique); (ID: *test_051* from Camelyon16)

# 4 Discussion

## 4.1 Time-Efficiency Analysis and Human-Machine Collaboration

We analyze the efficiency gains of our automated prediction system compared to manual analysis, assuming human experts require 34 seconds Randell et al. (2014) per sample with perfect accuracy, while our method computes predictions instantaneously. For acceptable error rates of 5%, 10%, and 20%, the theoretical time savings are 95%, 90%, and 80% respectively, calculated as $T(e, x, N) = (1 - e)xN$ where $e$ is the acceptable error rate, $N$ is the total number of samples and $x$ is the time. For our test set of $N = 129$ samples, with human analysis time of $x = 34$s, we calculate the following time savings:

Table 1: Time Savings Analysis

| Error Rate | Human Time | Time Saved | Time Saved (%) |
|---|---|---|---|
| 0% (Human) | 73min 6s | 0 | 0% |
| 5% | 3min 39s | 69min 27s | 95% |
| 10% | 7min 19s | 65min 47s | 90% |
| 20% | 14min 37s | 58min 39s | 80% |

In practical terms, this reduces the total analysis time in a human machine collaboration from ∼73 minutes to between ∼4 minutes (5% error rate) and ∼15 minutes (20% error rate). However, practical implementation requires a more nuanced analysis, as incorrect predictions necessitate human intervention. Our system allows for confidence-based deferrals, where predictions below a certain threshold are automatically routed to human experts. Based on our confusion matrix, Fig. 1, and considering these possible outcomes in real world applications, we discern two critical scenarios:

1. Low-confidence predictions requiring immediate human review

2. False predictions above the confidence threshold requiring subsequent correction

The practical utility of the method depends heavily on the chosen confidence threshold, which trades off between automation rate and error correction overhead.

**Recommendations** The system is most effective when deployed as a preliminary screening tool with the following considerations:

- Set confidence thresholds based on institutional risk tolerance, as mentioned in Sec. 3.1
- Maintain systematic error documentation for model improvement
- Implement clear protocols for human review of low-confidence cases

This analysis suggests that while our method can significantly reduce workload, it should be viewed as a guidance tool rather than a replacement for expert analysis. As demonstrated in Fig. 8 and Fig. 9, the system can effectively highlight tumor regions in clear cases, potentially accelerating the histopathologist's review process. However, Figure 10 illustrates the method's limitations, where tumor regions are not as clearly identified. Success depends on establishing appropriate confidence thresholds and maintaining human oversight, particularly for cases where the model shows uncertainty or produces potentially misleading interpretations.

### 4.2 Interpretation of Results

The results from our attention-based MIL model show strong performance in tumor classification, with important insights coming from our XAI analysis. The false negative cases show a clear pattern - all of them contain micro tumor regions where only a small subset of the bag consists of tumorous instances. This matches to our observations in Sec. 3.1 and in milestone 2 report about how the model performs differently on macro versus micro tumor cases. The combination of LRP and attention weights worked particularly well for macro tumor cases, as shown in Figures 8 and 9. This combination helps solve problems we found with individual XAI methods: the attention weights by themselves sometimes missed parts of tumor regions, while raw LRP scores tended to highlight areas less precisely. By normalizing and combining these signals, we get better tumor region identification that matches more closely with expert annotations. When we looked at negative cases (Fig. 7), we found that attention weights give clearer interpretations than both SmoothGrad and LRP for non-tumorous tissue. In these cases, the attention weights correctly show almost zero importance across the bag, while the other methods produce less useful uniform distributions of relevance scores. While Shapley Values in some cases provided superior explanations (as discussed in Sec. 3.2), their computational limitations (1 bag of ∼10.000 instances took ∼3-4 hours computation time) made them impractical for regular use. We also found only small differences between SmoothGrad and LRP results (Fig. 6), which makes sense given our network's structure.

### 4.3 Outlook & Summary

**Outlook** Based on our findings, we suggest two key areas for future improvement. First, micro tumor detection remains a significant challenge, particularly in cases where very few instances in a bag are positive. This could be addressed through specialized attention mechanisms that are more sensitive to sparse positive instances. Second, while our XAI methods showed promise, there is substantial room for improvement. The most critical challenge for Shapley Values is their computational complexity - despite their superior explanatory power in several cases, calculating them for large bags is prohibitively expensive, even when approximation methods are used. Research into optimized computation techniques or future technological advancements in computing could make this valuable tool more practical. Additionally, our LRP implementation could be refined by exploring alternative propagation rules and incorporating more sophisticated network architectures that could better leverage LRP's layer-wise analysis capabilities. Moreover, it might be very interesting to compare the performance and explainability of the MIL attention approach across different datasets.

**Summary** All in all, the paper 'Attention-based Deep Multiple Instance Learning' Ilse et al. (2018) is fully replicated for both the MNIST baseline and the Histopathology dataset and further extended with a focus on XAI methods. The quantitative results correspond to the findings and results of Ilse et al. (2018). Solely, the mean and max pooling operator for the embedding approach did not work out as good as expected but as the focus for the histopathology dataset lies in the interpretability and evaluation of the attention approach, we have not put an extra amount of effort into improving the aforementioned methods. The attention weights, extracted from the pooling attention layer, served as a well-interpretable quantity across most cases of the Camelyon16 dataset. However, these, as well as other XAI methods like Shapley Values, Grad × Input and LRP, have in particular some downsides in meaningfulness applied onto the micro tumor region cases. These can be overcome in parts by renormalizing and aggregating certain methods with each other (e.g. LRP + Attention Weights). Other than the major finding that there is not one XAI method which can be reliably applied to each case out of the Camelyon16 dataset, there were no serious pitfalls we encountered during the 'Attention-based Deep Multiple Instance Learning' project.

# References

A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf. Xai for transformers: Better explanations through conservative propagation, 2022. URL https://arxiv.org/abs/2202.07304.

M. Ancona, E. Ceolini, C. Öztireli, and M. Gross. Towards better understanding of gradient-based attribution methods for deep neural networks, 2018. URL https://arxiv.org/abs/1711.06104.

A. Binder, G. Montavon, S. Bach, K.-R. Müller, and W. Samek. Layer-wise relevance propagation for neural networks with local renormalization layers, 2016. URL https://arxiv.org/abs/1604.00825.

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.

S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

R. Randell, R. A. Ruddle, R. G. Thomas, C. Mello-Thoms, and D. Treanor. Diagnosis of major cancer resection specimens with virtual slides: impact of a novel digital pathology workstation. *Human Pathology*, 45(10):2101–2106, 2014. ISSN 0046-8177. doi: https://doi.org/10.1016/j.humpath.2014.06.017. URL https://www.sciencedirect.com/science/article/pii/S004681771400272X.

D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise, 2017. URL https://arxiv.org/abs/1706.03825.

# A Appendix

## A.1 False Negative Cases

The identified False Negatives (test_011, test_013, test_066, test_099) are all from the same class. They are all consisting of micro tumor regions which are assumed to make the bag classification more difficult as the proportion of tumorous instances is really small. Fig. 11 depict the ground truth annotations of the micro tumor regions in red.

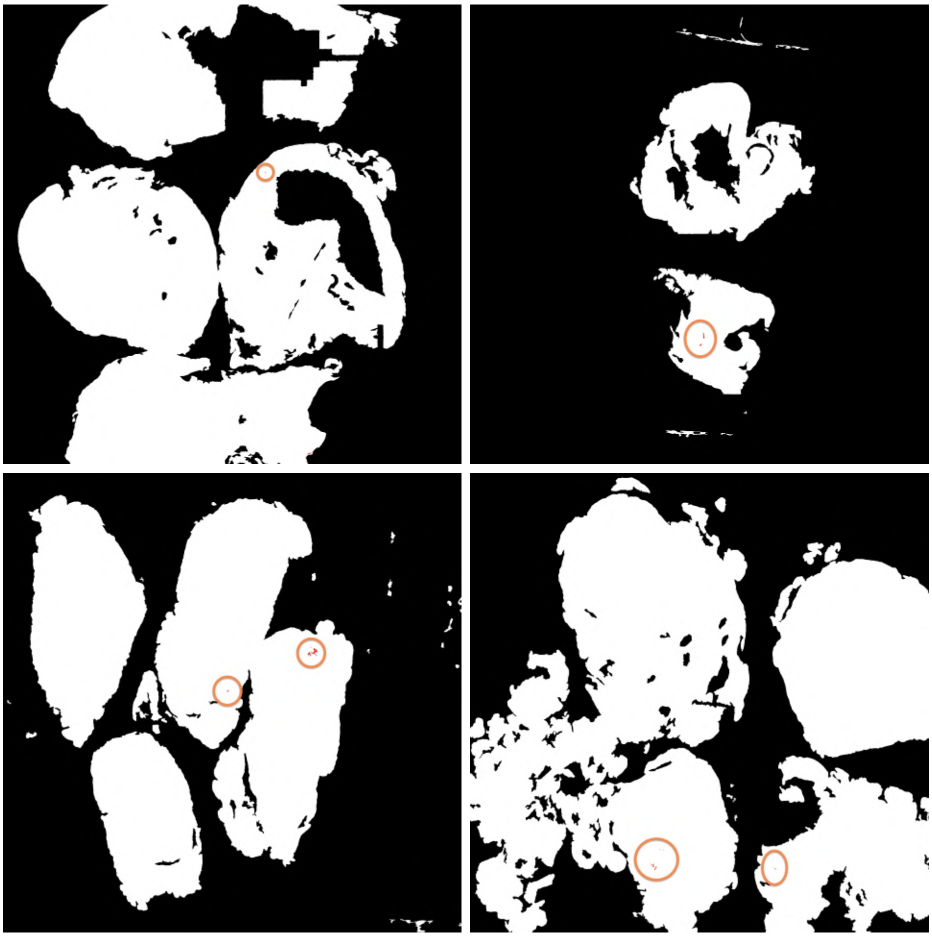


Figure 11: **False Negatives:** Ground truth annotations for the four identified false negative cases from the test set.