Is Attention Interpretation? A Quantitative Assessment On Sets

Anonymous Author(s)
Affiliation
Address
email

Abstract

The debate around the interpretability of attention mechanisms is centered on whether attention scores can be used as a proxy for the relative amounts of signal carried by sub-components of data. We propose to study the interpretability of attention in the context of set machine learning, where each data point is composed of an unordered collection of instances with a global label. For classical multiple-instance-learning problems and simple extensions, there is a well-defined "importance" ground truth that can be leveraged to cast interpretation as a binary classification problem, which we can quantitatively evaluate. By building synthetic datasets over several data modalities, we perform a systematic assessment of attention-based interpretations. We find that attention distributions are indeed often reflective of the relative importance of individual instances, but that silent failures happen where a model will have high classification performance but attention patterns that do not align with expectations. Based on these observations, we propose to use ensembling to minimize the risk of misleading attention-based explanations.

1 Introduction

Attention mechanisms have become a popular tool in multiple areas of machine learning, in particular in natural language processing (NLP) where their introduction significantly increased performance (Devlin et al., 2018). Attention-based models have also been successful in the context of computer vision (Dosovitskiy et al., 2020) and have in particular been attractive in digital histopathology applications (cancer diagnosis based on stained microscopy images) (Ilse et al., 2018; Redekop et al., 2021; Lu et al., 2021; Tourniaire et al., 2021), where a patch-based approach is particularly well-adapted to analyse the large whole-slide images (WSIs) with corrupting artefacts typically exploited in this field.

Besides the performance gain provided by attention mechanisms in many applications, one of their alluring aspects is the promise of interpretability: attention relies on a dynamically weighted average of representations of data subcomponents, and it feels natural that these weights should be informative of the relative importance of these subcomponents for the final prediction. This potential interpretability is particularly attractive for biomedical applications, both in a clinical setting and for research. Indeed, insights into automatic diagnostic tools is both a regulatory requirement (Selbst and Powles, 2017) and a necessary safeguard to understand and diagnose failure modes for critical decisions (Cluzeau et al., 2020). In a biomedical research context, attention-based interpretability could lead to new breakthroughs in understanding the mechanisms that underlie diseases and help find new targets for diagnosis and therapy.

While intuitively promising, there is still no clear understanding of the extent to which attention distributions provide meaningful information about the amount of signal carried by data subcom-

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

ponents. This has been the object of a debate within the context of NLP (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), which started at a conceptual level but was then moved forward by experimental assessments (Serrano and Smith, 2019; Vashishth et al., 2019). These studies found imperfect and task-dependent agreement between attention and other importance attribution metrics, but are limited by the constraints inherent to NLP: the difficulty of building robust ground truths and evaluation metrics for token importance (Madsen et al., 2022).

Given the recent interest in using attention in the context of biomedical applications, we propose to study the quality of attention-based explanations of instance importance in a simpler context, where we can conceive synthetic tasks with a well-defined ground truth, therefore allowing more control on the evaluation. Indeed, histopathological (and biomolecular) applications of attention can be characterised as multiple-instance learning (MIL) problems or simple extensions thereof. The goal of this work is to establish synthetic analogies for the MIL-like problems encountered in biomedical applications, with well-defined instance-level importance labels, and to quantitatively assess the quality of attention-based explanations, how frequently they are misleading, and potential solutions.

Our manuscript is organised as follows: we first introduce MIL as an abstract set classification problem, as well as some multi-population extensions. We show how these problems permit a quantitative assessment of instance importance attributions and why they map satisfyingly to some biomedical problems. We then describe the synthetic datasets we constructed as analogies and the attention-based models used to classify them, and conduct experiments to show to which extent attention-based explanations can be trusted. Finally, we argue for an ensemble-based solution to respond to the potential weaknesses of single-model explanations.

2 Importance attribution as a binary classification task

2.1 Multiple-instance learning and its extensions

2.1.1 Problem Formulation

Multiple-instance learning (MIL) is a classical weakly-supervised learning binary classification problem (Maron and Lozano-Pérez, 1997; Dietterich et al., 1997; Oquab et al., 2015) in which data points X_i are made of unordered collections of vectors $X_i = \{x_{i1}, \ldots, x_{iM_i}\}$. The individual vectors x_{im} are referred to as "instances", while the data points X_i are called bags of instances. Each instance x_{im} has a binary label $y_{im} \in \{1, 0\}$ (also referred to as positive and negative), which is not available at training time, but defines the label Y_i of the bag X_i as:

$$Y_i = \min\left(1, \sum_{m=1}^{M_i} y_{im}\right), \tag{1}$$

which simply means that Y_i is 1 if at least one of the y_{im} is 1, and is 0 otherwise.

This is a formalization of classification problems used in multiple biomedical applications, such as patient diagnosis from histopathology images. Images are typically processed as collections of patches, of which only a few might contain clinically relevant regions. Another interesting medical application of MIL is the classification of tumors using single-cell molecular profiles. In this case, samples are a mixture of healthy and cancerous cell profiles, but only patient-level labels are available.

2.1.2 Multi-Population MIL

Inspired by the biological applications of MIL, especially in the context of cancer, we propose to extend MIL to a multi-population setting with non-trivial interactions, which we can formalise as logical problems.

Multi-population AND

- There are three instance populations with three instance labels: $y_{im} \in \{0, 1, 2\}$.
- Bags have a binary label Y_i given as "the set of $\{y_{im}\}$ contains 1 AND contains 2". Namely, Y_i is 1 only if it contains population 1 and 2, but 0 if only one of the two is present. Population 0 is irrelevant.

This problem can model tumours where multiple cell communities can develop and support each other's growth by collaboration: the presence of both cellular communities accelerates disease progression and leads to worse prognosis (Tabassum and Polyak, 2015). In this case, population 0 would correspond to uninformative cells such as healthy cells in the tumour microenvironment while populations 1 and 2 would represent two cancerous populations that can collaborate.

Multi-population XOR

- There are three instance populations with three instance labels: $y_{im} \in \{0, 1, 2\}$.
- Bags have a binary label Y_i given as "the set of $\{y_{im}\}$ contains 1 XOR contains 2", *i.e.* Y_i is 1 only if it contains population 1 but not 2 or 2 but not 1. Population 0 is irrelevant.

This problem can model tumours where two cell communities can co-evolve but reduce their joint fitness such as by increasing drug response when both are present (Miller et al., 1991).

2.2 Quantifying key instance attribution

The simple setting of MIL lends itself to quantifying the interpretability of importance distributions over bags of instances such as those provided by attention. For standard MIL this is often called key-instance attribution (Liu et al., 2012), which amounts to identifying positive instances inside positive bags. When ground truth instance-level labels are known, this can be formulated as a supervised binary classification problem. In this work, we train models with weak, bag-level labels but want to evaluate the attention scores as a prediction score to identify positive label instances.

Of course, we cannot expect attention scores to be well calibrated and to allow their immediate interpretation as a probability score for being "important". We therefore need to be careful with some of the standard classification metrics based on discretising prediction scores, such as accuracy or F_1. What we require of our attention scores is that they discriminate well between positive and negative instances for some threshold, which can be verified by inspecting the area under the receiver operating characteristic curve (AUC-ROC) or the average precision score (AUC-PR). For the sake of clarity, we will refer to the AUC of importance attribution as iAUC, so as not to confuse metrics for the bag-level classification and those for evaluating attention-based explanations. As illustrated in supplementary figures, bag with unbalanced proportion of instance types necessitate the use of AUC-PR instead of AUC-ROC since the later can lead to inflated scores because of the accurate prediciton of the majority class.

The multi-population extensions of MIL, *i.e.* AND and XOR don't have canonically defined importances. We propose to extend the "key instance" label by assigning it to populations 1 and 2 for both problems defined in section 2.1.2, while classifying population 0 as unimportant, since its presence or absence does not impact the bag labels.

3 Methods

3.1 Attention-Based Deep MIL

Permutation-invariant models are best-suited to handle MIL tasks as they introduce an inductive bias tailored to sets of instances where order does not matter. To this end, the Deep Sets architecture (Zaheer et al., 2018) was designed to produce an independent latent representation of each instance, which are then aggregated with a permutation invariant function such as the mean. The aggregated latent representation is further processed to produce a bag label, as shown in fig. 1.

Attention-based aggregation is another permutation-invariant operation that dynamically performs weighted averages using the attention scores. This was shown to improve performance and provide insights into the data through the assigned weights (Ilse et al., 2018). With attention-based aggregation, a data point $X = \{x_1, \ldots, x_M\}$ is mapped to a prediction y as follows:

$$z_i = \phi(x_i), \quad Z = \sum_{m=1}^{M} a_m z_m, \quad y = \rho(Z), \tag{2}$$

where ϕ and ρ are approximated by neural networks and a_m is the attention scores of instance x_m, defined as:

$$a_m = \frac{\exp\{\mathbf{w}^\top \tanh[\mathbf{V}\phi(x_m)^\top]\}}{\sum_{j=1}^{M} \exp\{\mathbf{w}^\top \tanh[\mathbf{V}\phi(x_j)^\top]\}}, \tag{3}$$

and, $\mathbf{V} \in \mathbb{R}^{L\times K}$ and $\mathbf{w} \in \mathbb{R}^{L\times 1}$ are trainable parameters. Notice that as $\sum_{j=1}^{M} a_m = 1$, Eq. 3 defines normalized discrete weights over the instances.

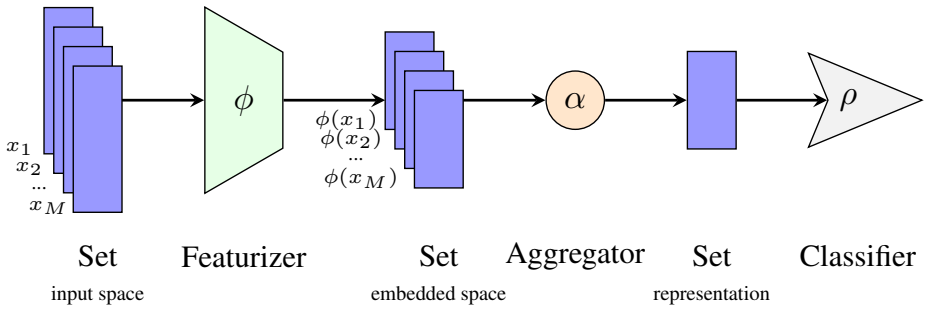


Figure 1: Deep-Sets-like permutation invariant networks map bags of instances x_i to bags of latent representations $\phi(x_i)$ which are then aggregated as a set representation. This set representation is processed by a classifier to obtain a prediction ρ. In all experiments in this paper, the aggregator α is the attention mechanism described in eq. (2).

3.2 Synthetic Datasets

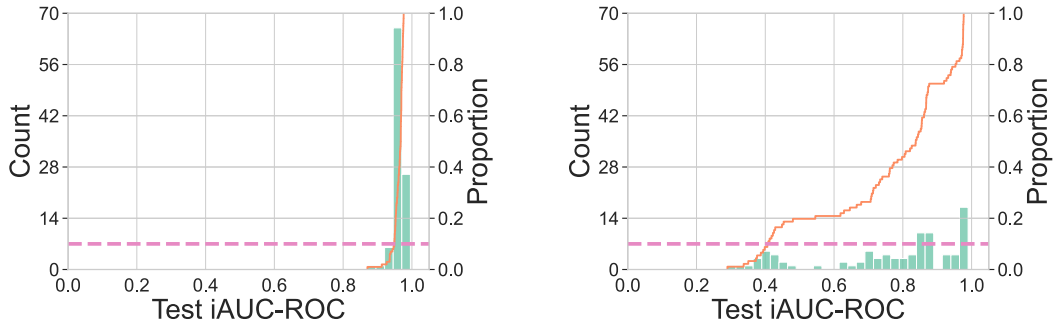
We generate synthetic datasets with well-defined ground truth instance labels using three data modalities. These instance labels were kept hidden from the model at all times and only used to evaluate the performance of the attention attribution.

The first type of datasets, referred to as *Gaussian MIL*, *Gaussian AND* or *Gaussian XOR*, was built by sampling instances from normal distributions, $\mathcal{N}(\mu, \sigma = 0.5)$ with $\mu \in \mathbb{R}^4$. Populations 0, 1 and 2 correspond to three choices of μ: $\mu_0 = (0, 0, 0, 0)^\top$, $\mu_1 = (1, 1, 1, 1)^\top$ and $\mu_2 = (-1, 1, 1, 1)^\top$.

The second type of datasets trades 4-dimensional vectors for 28×28 pixels images of MNIST handwritten digits and are referred to as *MNIST MIL*, *MNIST AND* or *MNIST XOR*. The bags defined by first specifying the list of digits allowed in each population and then randomly sampling images of the specified digits from the original MNIST dataset (LeCun and Cortes, 2005). Images of the digit "3" are given the instance label 1 in every problem while images of the digit "9" have instance label 2 in the XOR and AND cases. Any other digits are considered unimportant (label 0).

The last data modality mimics data produced by single-cell proteomics experiments. We used experimental single-cell mass-cytometry (CyTOF) measurements from breast cancer tumours (Wagner et al., 2019) to produce pseudo-samples by randomly selecting epithelial cells. Each cell is characterised by 27 protein abundance measurements from a panel of markers chosen for cell phenotyping. The work that collected and published these data (Wagner et al., 2019) grouped cells into 9 super-clusters of functionally and phenotypically distinct cells, including 7 clusters of luminal cells and two clusters of basal cells (B1 and B2). Basal cells are indicative of more dangerous tumours, in particular, super-cluster B2 was found to be strongly associated with triple-negative tumours (Elias, 2010). We therefore define the *CyTOF MIL*, *CyTOF AND* or *CyTOF XOR* with populations 0, 1, and 2 respectively corresponding to luminal cells, B2 cells and B1 cells.

In all settings, we generate bags of 250 instances, which are drawn from bi- or trinomial distributions of populations 0, 1 and 2. In the MIL setting, we use a binomial distribution with equi-probable outcomes while in the multi-population settings we use a trinomial distribution where population 0 has probability 0.4 and populations 1 and 2 have probability 0.3. While not described in this paper, we have confirmed that our results are quite robust to changes in these parameters except for extreme cases (extremely low fractions of some population or very small bags).



(a) (b)

Figure 2: **(a)** Configuration with stable iAUC-ROC. **(b)** Configuration with significant fraction of low iAUC-ROC. Both configurations were trained on the Gaussian MIL setting. The left Y-axis refers to the histogram (in green), while the right Y-axis refers to the cumulative frequency plot (orange line). The magenta line is a guide for the eye showing the 10% threshold used to define bad configurations in table 1.

3.3 Interpretability Evaluation

As said earlier, we train models with bag-level labels but we are mostly interested in the attention mechanism performance as instance importance discriminator. A fair assessment is only possible with the correct metric, which is why we compared the well known AUC-ROC with the AUC-PR. Both metrics are threshold-independent and result from a trade-off: for AUC-ROC, the True Positive Rate (TPR, or Sensitivity or Recall) and False Positive Rate (FPR, or 1-Specificity) are considered whereas for AUC-PR, the trade-off is between Precision and Recall. Precision is the fraction of relevant instances among all instances retrieved by the model and Recall is the fraction of relevant instances that actually were detected. As discussed in (Sofaer et al., 2019), the AUC-ROC is prone to overestimation in cases where the proportion of positive instances is much smaller than of negative ones because the TN count is then disproportionately large and pulls the FPR towards zero. In this work, we validate the results from Sofaer et al. and justify the use of AUC-ROC with our carefully generated synthetic datasets.

4 Results

The basis of our analysis is a hyperparameter search for each task and data modality. We perform a grid search through possible configurations for our models and train each configuration with five random initialisations. Models are then ranked and selected on the basis of their performance on a validation set, and evaluated on a separate test set. More details on the hyperparameter search are provided in appendix A.

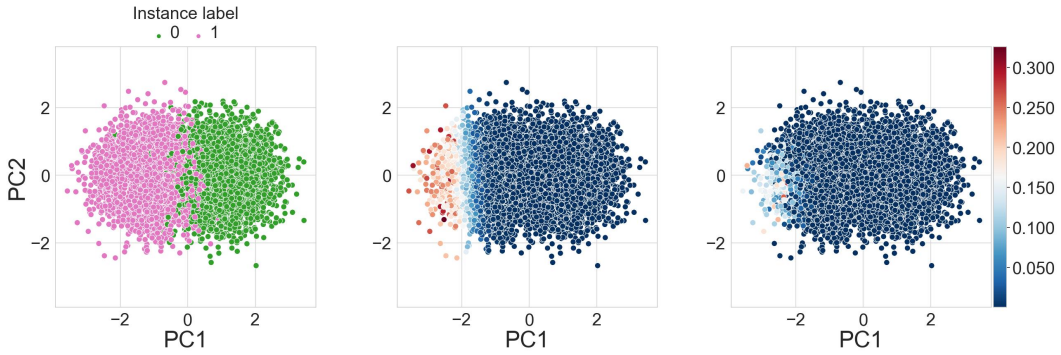
4.1 Models with high accuracy can have poorly behaved attention

To reproduce the process of selecting models in a setting where instance-level importances are unknown, we select five candidate model configurations from our hyperparameter search based on their validation accuracy and evaluate the interpretability of their attention distributions. We train 100 repetitions of each of those top five configurations with different random seeds and evaluate how well the attention scores separate negative from positive instances in bags with a positive label. As we show in fig. 2, some configurations have narrow distributions of iAUC-ROC centred around a reasonable value (0.95), meaning that all model realisations provide meaningful interpretations through their attention distributions while others have a non-negligible fraction of outliers with a very poor identification of important instances (iAUC-ROC around 0.4).

This pattern repeats over all problems and data modalities we evaluated. We summarise the results of our analysis in table 1, where we report the mean test iAUC-ROC across all configurations and the number of "bad" configurations, defined as those having 10% or higher fraction of realisations

Table 1: Evaluation of attention explanations performances. Multi-population problems tend to have more bad configurations than MIL, which can still have poor explanations. In general, AND problems also have an overall lower iAUC-ROC.

Data	Problem	Mean iAUC	# bad config.
Gaussian	MIL	0.84	2/5
	AND	0.70	5/5
	XOR	0.86	4/5
MNIST	MIL	0.91	0/5
	AND	0.69	5/5
	XOR	0.91	1/5
CyTOF	XOR	0.91	1/5
	AND	0.76	3/5
	XOR	0.84	2/5



(a) Instance Labels (b) Good Attention (c) Bad Attention

Figure 3: Low-dimensional projections of MIL data with showing attention scores for an example of a "good" and a "bad" model, as well as the instance labels shown for reference.

with an iAUC-ROC less than 0.65. Detailed results with all iAUC-ROC distributions are available in supplement S.1.

To further illustrate the difference in behavior between "good" and "bad" models, we show low-dimensional representations of one of our numerical datasets (Gaussian) in fig. 3, where the attention distributions are visible. "Good" models have an essentially constant attention over unimportant instances and show a sharp gradient on positive instances moving away from the class boundary, while "bad" models essentially have uniform attention over much of the dataset, with the exception of a small minority of the data for which the attention is higher but not as high as for "good" models.

4.2 Repetitions of the same model have little correlation between performance and interpretability

The stochasticity of training multiple neural networks with the same hyperparameters leads to the variability in the quality of the explanations provided by their attention maps. Of course, this stochasticity also leads to variability in the validation and test performance of these models. It is therefore natural to investigate whether, for a fixed configuration, there is a correlation between the classification performance at the bag level and the quality of the attention-based explanations. This analysis might provide a way to weed out problematic models at the validation stage.

For each problem and data modality, we use the top 5 configurations defined in section 4.1 to evaluate how well validation-time classification performance discriminates between models with low and high-quality explanations. As we show in fig. 4, high performance is not a good indicator of good explanations, and the correlation between accuracy and iAUC-ROC exists but is rather mild. A more

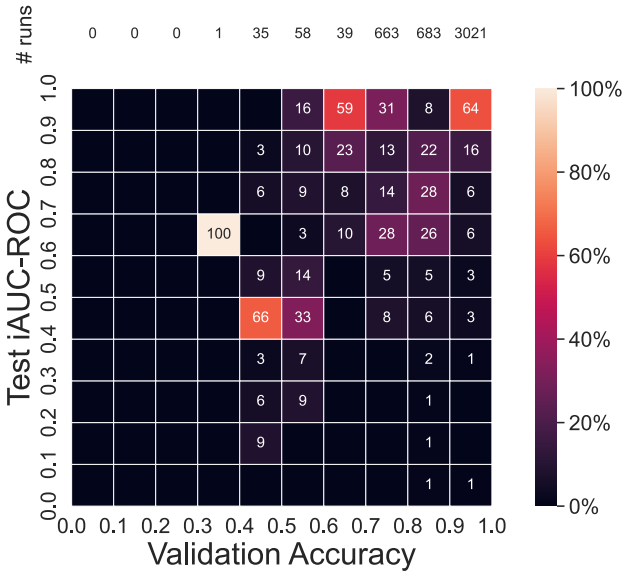


Figure 4: Relationship between validation accuracy and test iAUC-ROC for top configurations. Models are binned by validation accuracy and iAUC-ROC and each bin displays the fraction of total models *per column* (*i.e.* per accuracy bin). The total number of models in each column is reported at the top.

detailed picture separated by problem and data modality is available in supplement S.3. In the case of MIL problems on Gaussian data (supplementary figure S.19a), all models with the top configurations reach a validation accuracy of 100% while having varying iAUC-ROC values. On more complex problems, not all realisations reach perfect accuracies, and a limited amount of correlation can be observed. Indeed, as shown in supplementary figure S.19e, it is often the case that only the models with top validation accuracies reach the top values for the iAUC-ROC. Nevertheless, there is still significant variability among the models with top validation accuracies so that filtering out models with a poorer validation performance is not enough to avoid models with poor explanations.

We measure the Spearman correlation ρ between the validation-time accuracy of the 100 repetitions of each top configurations for all our classification tasks and the iAUC-ROC score and report them in table 2. For each problem, we further report the configurations with the highest and lowest spreads of iAUC-ROC values (ΔiAUC) between individual top-performing realisations. Namely, to compute ΔiAUC, we select the models in the highest decile of validation accuracy for each configuration and measure the spread between their maximum and minimum iAUC-ROC values. This provides a way of observing how specific configurations have a large variability of iAUC-ROC even when filtering for models with high classification performance.

4.3 Ensembling improves explanation robustness

While the risk of poor explanations is real, most trained models with good performance achieve satisfying interpretation-based explanation quality. We therefore propose to use ensembling to reduce the risk of encountering poorly-performing single models. Two strategies are possible:

- Single-configuration ensembling, where a fixed hyperparameter set is chosen based on validation performance and multiple realisations are trained with different random seeds.
- Multi-configuration ensembling, where we chose a number of high-performing models and ensemble realisations of each hyperparameter choice.

For both approaches, the ensembling is performed with the goal of obtaining *more robust attention-based explanations*. More concretely, for each bag of instances, each model produces an attention distribution over the instances and we compute the average attention scores across models.

Data	Problem	Spearman ρ	High ΔiAUC	Low ΔiAUC
Gaussian	MIL	0.04 ± 0.07	0.56	0.04
	AND	0.53 ± 0.31	0.66	0.07
	XOR	0.62 ± 0.17	0.77	0.27
MNIST	MIL	-0.01 ± 0.01	0.47	0.03
	AND	0.12 ± 0.23	0.41	0.12
	XOR	0.21 ± 0.15	0.16	0.03
CyTOF	MIL	0.14 ± 0.24	0.47	0.13
	AND	0.17 ± 0.15	0.56	0.32
	XOR	0.23 ± 0.23	0.54	0.24

Table 2: Predictivity of classification performance for informative explanations. We report the Spearman correlation between the validation accuracy and the iAUC as well as the highest and lowest ΔiAUC found among the models.

Data	Problem	% bad configs.		
		N=1	N=20 (single)	N=20 (mult.)
Gaussian	MIL	18.0	0.0	0.0
	AND	40.0	7.3	0.0
	XOR	17.3	0.0	0.0
MNIST	MIL	0.02	0.0	0.0
	AND	21.0	0.0	0.0
	XOR	4.0	0.0	0.0
CyTOF	MIL	2.3	0.0	0.0
	AND	16.7	6.7	2.7
	XOR	6.7	0.0	0.0

Table 3: Impact of ensembling on the fraction of models with bad explanations. We compare three situations: no ensembling (N=1), and ensembling 20 models with either single configuration ensembling (N=20, single) or multi-configuration ensembling (N=20, mult.).

This yields a valid attention distribution for the ensemble in the sense that the averaged distribution also sums to 1.

As we show in table 3, ensembling does improve the fraction of models with bad explanations (as defined in section 4.1), and multi-configuration ensembling provides the best option for most cases. The results we report for single-configuration ensembling are the average of the results obtained for the top five configurations found for each problem through hyperparameter search. As we show in more details in supplement S.5, this average hides the fact that single-configuration ensembling fails badly for some configurations, while multi-configuration ensembling does not present this failure mode.

4.4 AUC-ROC and AUC-PR are equivalent for balanced datasets

The results presented in appendix S.6 endorse the claim from Sofaer et al. that the AUC-ROC can be misleading in situations where the proportion of positive instances is much smaller than of negative ones. Indeed, the AUC-PR drops significantly with the proportion of positive instances while the AUC-ROC remains indifferent to this change. They also show that, for balanced datasets such as the ones we carefully generated, the AUC-ROC and AUC-PR outputs are in a comparable range. Interestingly, the AUC-PR was always slightly larger than the AUC-ROC when 50% of the instances were positive and clearly smaller when the proportion drops.

4.5 Attention network hidden layer size does not influence explanation

The most influential part of the architecture on the attention weight interpretability was thought to be the hidden layer size of the attention network. To inspect this idea, that exact parameter was varied across ten different values. The Accuracy, the iAUC-ROC and the iAUC-PR were reported in appendix S.7 for each hidden layer size. Only subtle variations of the iAUC metrics were observed, and no value appears to have a clear advantage in terms of iAUC. Interestingly, for the Gaussian MIL modality, the size leading to the highest iAUC scores is one. The curve appears a little more staggered for MNIST MIL, but one can observe that increasing the attention network hidden size does not improve the iAUC nor the Accuracy.

5 Discussion

Our experiments confirm that, most of the time, attention mechanisms provide meaningful information about the relative importance of instances in set classification problems like MIL. Nevertheless, silent failure modes exist where individual models can have good performance at the main weakly supervised task but produce attention maps that are not aligned with the amount of signal carried by data sub-components. This finding is somewhat worrying: with a bit of bad luck, a researcher could train a good model with poor interpretability and generate new hypotheses based on nonsensical explanations, which could lead to resource waste if they are the basis for experimental studies. However, attention-based explanations should not altogether be discarded, but be considered with care. As our ensembling experiments show, sporadically appearing bad-behaving models can be mitigated, but not altogether avoided in a multi-model setup as silent failures seem to fall in the minority. In some settings, however, ensembling by averaging attention scores does not improve the failure rate. We suspect that this is due to poor agreement between the attention assignment of different models, leading to poor ensemble performance, which could be improved by switching to majority voting. If this is the case, we could avoid false positive labelling of important instances by requiring a clear consensus between different models, which we hope to explore in future work. In any case, some responsible downstream analysis and validation of patterns highlighted by attention mechanisms is warranted when trying to discover new features in data, keeping in mind that there is a small but non-zero probability that the patterns might be misleading.

6 Conclusion

We showed across a variety of set-classification tasks and data modalities that silent failure modes exist for attention-based key instance attributions, where attention does not correlate with instance importance. While ensembling multiple random initialisations of the same model and multiple model architecture mitigates the issue, there often remains a probability that explanations based on attention could be misleading, which can range from problematic for scientific discovery to dangerous when using explanations to verify predictions in application settings. This should not be a reason to abandon attention as a tool for identifying important sub-components of data for a given model, but shows that downstream verification of potential patterns is necessary. We have hinted at the fact that a more fine-grained approach to ensembling could help filter false positives and this is definitely an interesting avenue for further research. Other important directions which we plan to pursue is the identification of the features of tasks where silent failure is less common, as well as understanding which aspects of model architecture impact the quality of importance attribution.

References

Jean Marc Cluzeau, Xavier Henriquel, George Rebender, Guillaume Soudain, Luuk van Dijk, Alexey Gronskiy, David Haber, Corentin Perret-Gentil, and Ruben Polak. 2020. *Concepts of Design Assurance for Neural Networks (CoDANN) - AI Roadmap*. Technical Report. EASA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. `https://doi.org/10.48550/ARXIV.1810.04805`

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. 1997. Solving the Multiple Instance Problem with Axis-Parallel Rectangles. *Artificial Intelligence* 89, 1 (Jan. 1997), 31–71. https://doi.org/10.1016/S0004-3702(96)00034-3

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. 2020. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Anthony D. Elias. 2010. Triple-Negative Breast Cancer: A Short Review. *American Journal of Clinical Oncology* 33, 6 (Dec. 2010), 637–645. https://doi.org/10.1097/COC.0b013e3181b8afcf

Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-Based Deep Multiple Instance Learning. In *International Conference on Machine Learning*. PMLR, 2127–2136.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. https://doi.org/10.48550/ARXIV.1902.10186

Yann LeCun and Corinna Cortes. 2005. The Mnist Database of Handwritten Digits. http://yann.lecun.com/exdb/mnist/.

Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. 2012. Key Instance Detection in Multi-Instance Learning. In *Proceedings of the Asian Conference on Machine Learning*. PMLR, 253–268.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. 2021. Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images. *Nature Biomedical Engineering* 5, 6 (June 2021), 555–570. https://doi.org/10.1038/s41551-020-00682-w

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-Hoc Interpretability for Neural NLP: A Survey. https://doi.org/10.48550/arXiv.2108.04840 arXiv:2108.04840 [cs]

Oded Maron and Tomás Lozano-Pérez. 1997. A Framework for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems*, Vol. 10. MIT Press.

Bonnie E. Miller, Todd Machemer, Martin Lehotan, and Gloria H. Heppner. 1991. Tumor Subpopulation Interactions Affecting Melphalan Sensitivity in Palpable Mouse Mammary Tumors. *Cancer Research* 51, 16 (Aug. 1991), 4378–4387.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic. 2015. Is Object Localization for Free? – Weakly-supervised Learning with Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ekaterina Redekop, Karthik V. Sarma, Adam Kinnaird, Anthony Sisk, Steven S. Raman, Leonard S. Marks, William Speier, and Corey W. Arnold. 2021. Attention-Guided Prostate Lesion Localization and Grade Group Classification with Multiple Instance Learning. In *Medical Imaging with Deep Learning*.

Andrew D Selbst and Julia Powles. 2017. Meaningful Information and the Right to Explanation. *International Data Privacy Law* 7, 4 (Nov. 2017), 233–242. https://doi.org/10.1093/idpl/ipx022

Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2931–2951. https://doi.org/10.18653/v1/P19-1282

Helen R. Sofaer, Jennifer A. Hoeting, and Catherine S. Jarnevich. 2019. The Area under the Precision-Recall Curve as a Performance Metric for Rare Binary Events. *Methods in Ecology and Evolution* 10, 4 (2019), 565–577. https://doi.org/10.1111/2041-210X.13140

Doris P. Tabassum and Kornelia Polyak. 2015. Tumorigenesis: It Takes a Village. *Nature Reviews Cancer* 15, 8 (Aug. 2015), 473–483. https://doi.org/10.1038/nrc3971

Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, and Herve Delingette. 2021. Attention-Based Multiple Instance Learning with Mixed Supervision on the Camelyon16 Dataset. In *Proceedings of the MICCAI Workshop on Computational Pathology*. PMLR, 216–226.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention Interpretability Across NLP Tasks. *arXiv:1909.11218 [cs]* (Sept. 2019). arXiv:1909.11218 [cs]

Johanna Wagner, Maria Anna Rapsomaniki, Stéphane Chevrier, Tobias Anzeneder, Claus Langwieder, August Dykgers, Martin Rees, Annette Ramaswamy, Simone Muenst, Savas Deniz Soysal, Andrea Jacobs, Jonas Windhager, Karina Silina, Maries van den Broek, Konstantin Johannes Dedes, Maria Rodríguez Martínez, Walter Paul Weber, and Bernd Bodenmiller. 2019. A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell* 177, 5 (May 2019), 1330–1345.e18. `https://doi.org/10.1016/j.cell.2019.03.005`

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. `https://doi.org/10.48550/ARXIV.1908.04626`

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. 2018. Deep Sets. *arXiv:1703.06114 [cs, stat]* (April 2018). arXiv:1703.06114 [cs, stat]

A Hyperparameter Searches

Parameter	Values
Batch size	100
Epoch	500
Learning rate	0.005
Weight decay	0.0001
Loss function	Cross-entropy
Optim. algorithm	Adam
Hidden layer size	2, 4, 6, 8, 10
Attention size	1, 2, 4, 6, 8, 10
Featurizer depth	1, 2, 3
Classifier depth	1, 2, 3

Table 4: Parameter grid for Gaussian data.

Parameter	Values
Batch size	100
Epoch	500
Learning rate	0.005
Weight decay	0.0001
Loss function	Cross-entropy
Optim. algorithm	Adam
Hidden layer size	8, 16, 32, 64
Attention size	1, 2, 4, 6, 8, 10
Featurizer depth	1, 2
Classifier depth	1, 2

Table 5: Parameter grid for MNIST data.

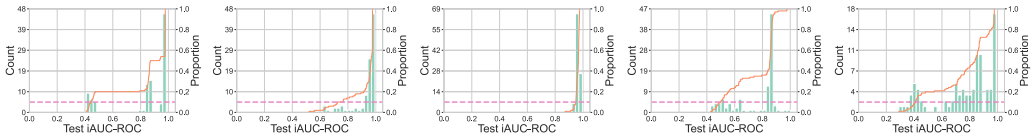
Parameter	Values
Batch size	100
Epoch	500
Learning rate	0.005
Weight decay	0.0001
Loss function	Cross-entropy
Optim. algorithm	Adam
Hidden layer size	2, 4, 6, 8, 10
Attention size	1, 2, 4, 6, 8, 10
Featurizer depth	1, 2, 3
Classifier depth	1, 2, 3

Table 6: Parameter grid for CyTOF data.

Supplementary material for ”Is Attention Interpretation? A Quantitative Assessment On Sets”

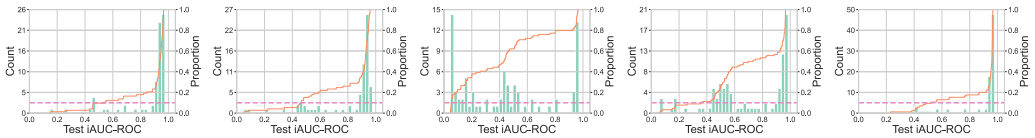
S.1 iAUC-ROC Distributions for Top Models

Models marked with an asterisk have a significant proportion of bad runs, i.e. 10% or more of them achieved an iAUC below 0.65.



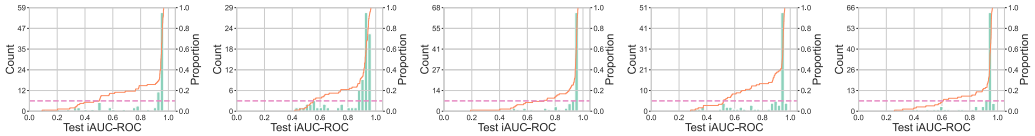
(a) Model 1 * (b) Model 2 (c) Model 3 (d) Model 4 * (e) Model 5 *

Figure S.1: Gaussian MIL



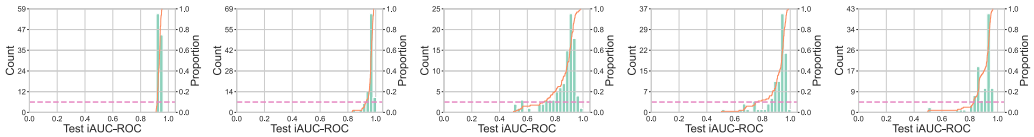
(a) Model 1 * (b) Model 2 * (c) Model 3 * (d) Model 4 * (e) Model 5 *

Figure S.2: Gaussian AND



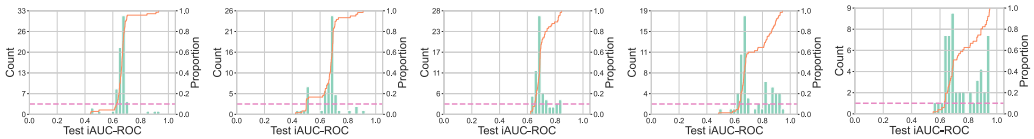
(a) Model 1 * (b) Model 2 * (c) Model 3 (d) Model 4 * (e) Model 5 *

Figure S.3: Gaussian XOR



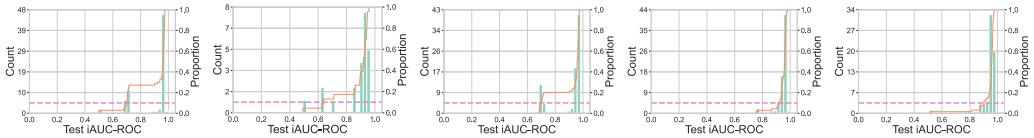
(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4 (e) Model 5

Figure S.4: MNIST MIL



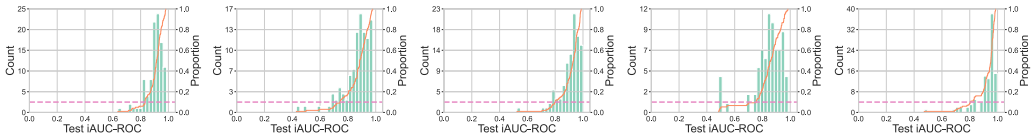
(a) Model 1 * (b) Model 2 * (c) Model 3 * (d) Model 4 * (e) Model 5 *

Figure S.5: MNIST AND



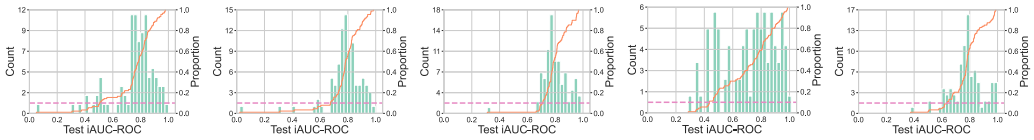
(a) Model 1 (b) Model 2 * (c) Model 3 (d) Model 4 (e) Model 5

Figure S.6: MNIST XOR



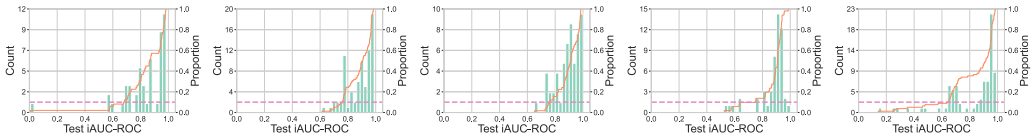
(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4 (e) Model 5

Figure S.7: CyTOF MIL



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

Figure S.8: CyTOF AND

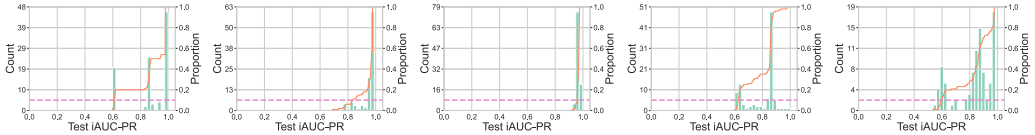


(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4 * (e) Model 5 *

Figure S.9: CyTOF XOR

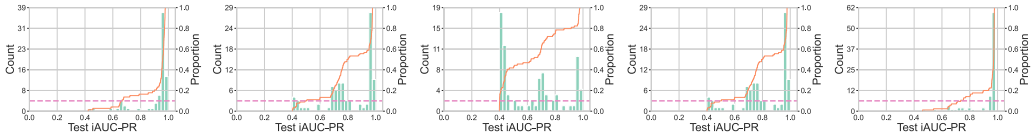
S.2 iAUC-PR Distributions for Top Models

Models marked with an asterisk have a significant proportion of bad runs, i.e. 10% or more of them achieved an iAUC-PR below 0.65.



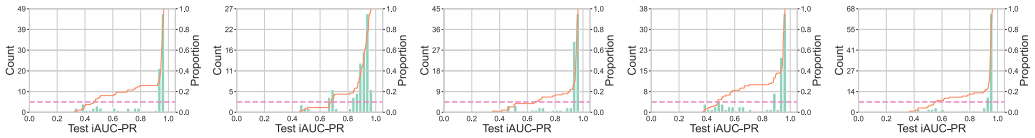
(a) Model 1 * (b) Model 2 (c) Model 3 (d) Model 4 * (e) Model 5 *

Figure S.10: Gaussian MIL



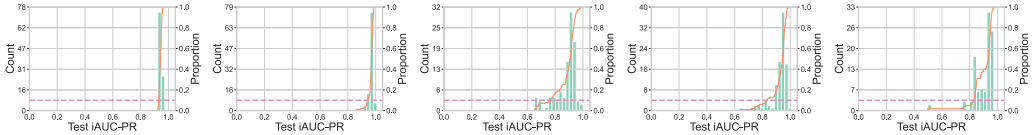
(a) Model 1 * (b) Model 2 * (c) Model 3 * (d) Model 4 * (e) Model 5

Figure S.11: Gaussian AND



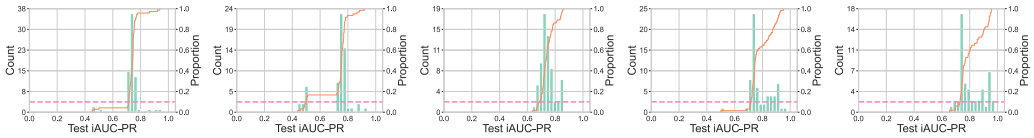
(a) Model 1 * (b) Model 2 * (c) Model 3 * (d) Model 4 * (e) Model 5 *

Figure S.12: Gaussian XOR



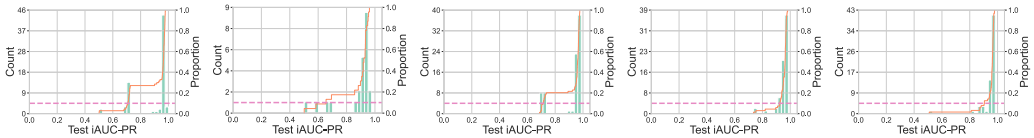
(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4 (e) Model 5

Figure S.13: MNIST MIL



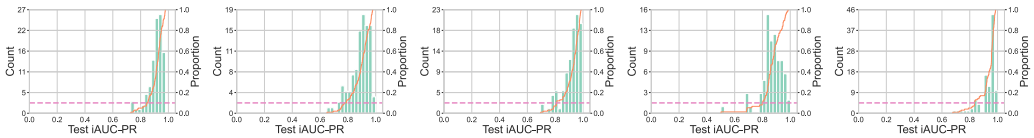
(a) Model 1 (b) Model 2 * (c) Model 3 (d) Model 4 (e) Model 5

Figure S.14: MNIST AND



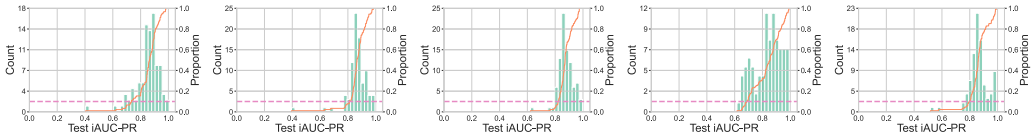
(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4 (e) Model 5

Figure S.15: MNIST XOR



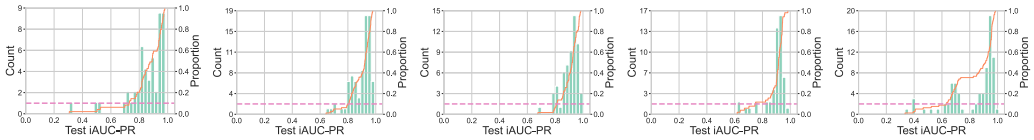
(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4 (e) Model 5

Figure S.16: CyTOF MIL



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

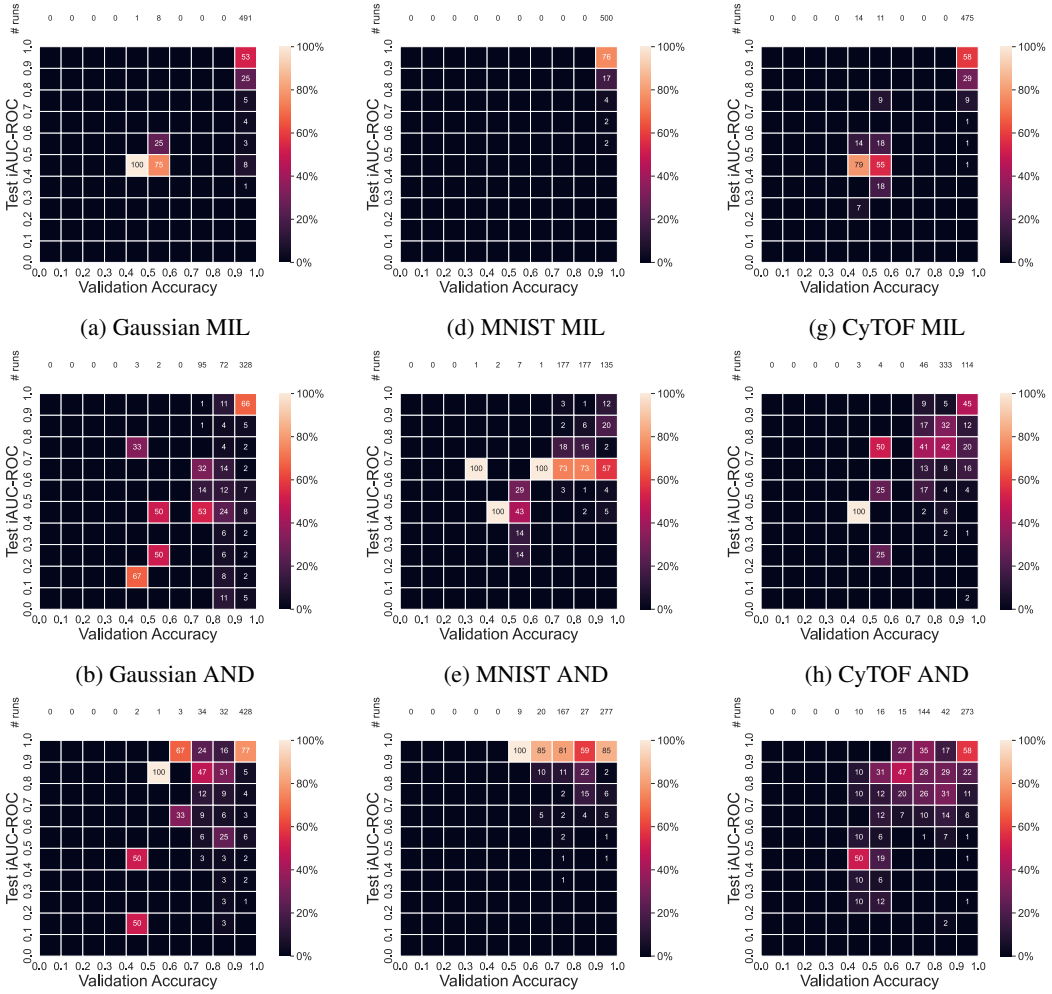
Figure S.17: CyTOF AND



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

Figure S.18: CyTOF XOR

S.3 Correlations between iAUC-ROC and Accuracy



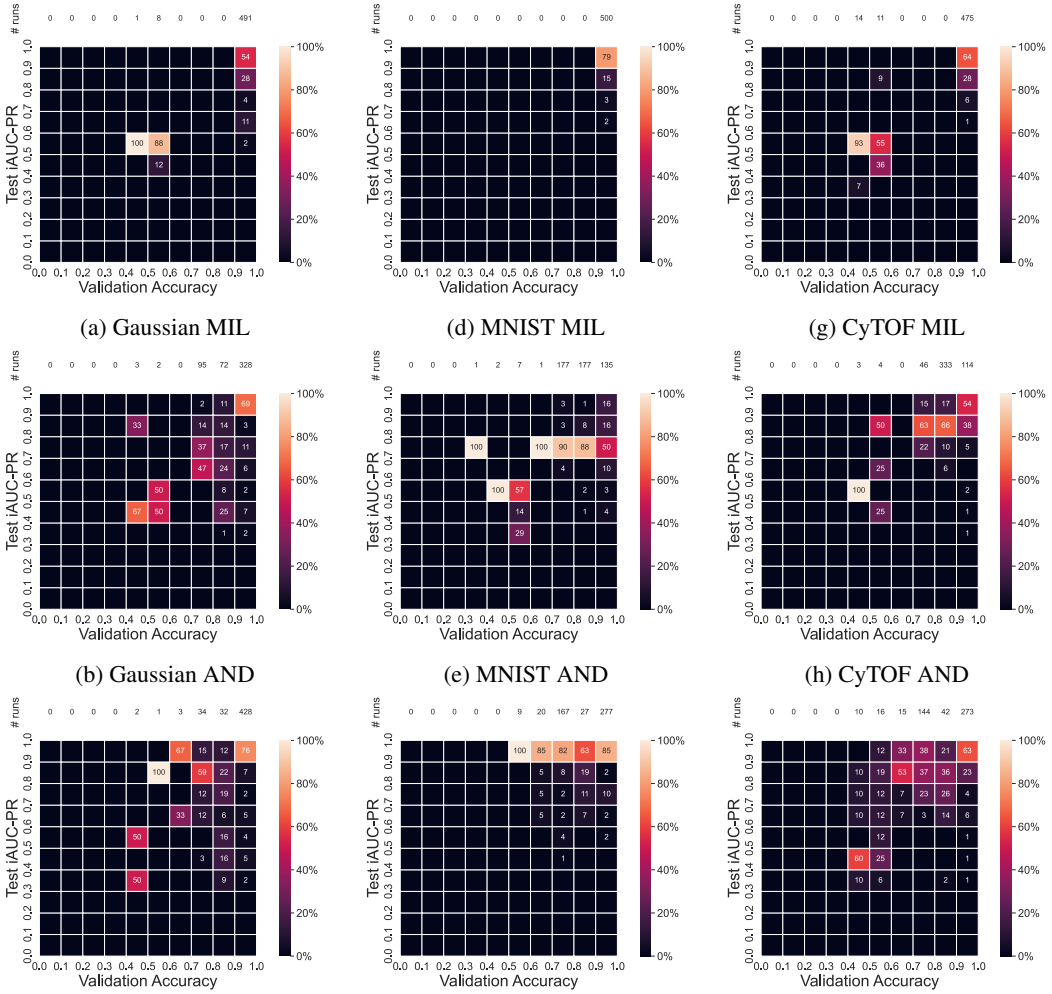
(a) Gaussian MIL (d) MNIST MIL (g) CyTOF MIL

(b) Gaussian AND (e) MNIST AND (h) CyTOF AND

(c) Gaussian XOR (f) MNIST XOR (i) CyTOF XOR

Figure S.19: Relationship between validation accuracy and test iAUC-ROC for top configurations, separated by problem and data modality. Models are binned by validation accuracy and iAUC-ROC and each bin displays the fraction of total models *per column* (*i.e.* per accuracy bin). The total number of models in each column is reported at the top.

S.4 Correlations between iAUC-PR and Accuracy



(a) Gaussian MIL (d) MNIST MIL (g) CyTOF MIL

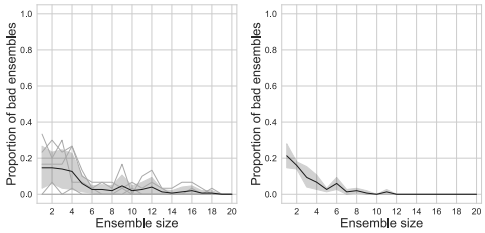
(b) Gaussian AND (e) MNIST AND (h) CyTOF AND

(c) Gaussian XOR (f) MNIST XOR (i) CyTOF XOR

Figure S.20: Relationship between validation accuracy and test iAUC-PR for top configurations, separated by problem and data modality. Models are binned by validation accuracy and iAUC-PR and each bin displays the fraction of total models *per column* (*i.e.* per accuracy bin). The total number of models in each column is reported at the top.

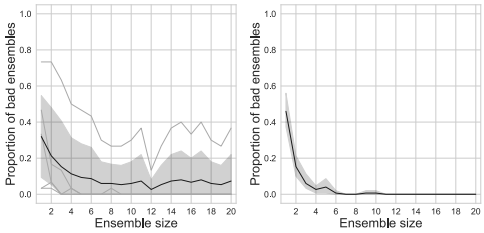
S.5 Ensembling

Proportion of bad ensembles for single- and multi-configuration ensembles. Bad ensembles are characterised by an iAUC of 0.65 or below. For each ensemble size, 30 different ensembles were produced. In the single-configuration plots, the light grey lines show the results for the individual configurations while the black line shows their average. In the multi-configuration case, the process was repeated five times. The 95% confidence interval is indicated by the grey area.



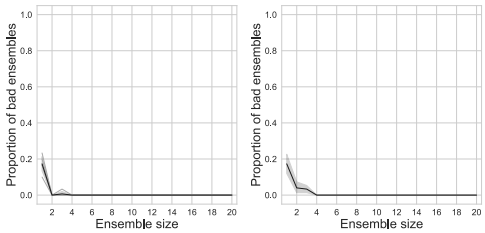
(a) Single-Configuration (b) Multi-Configuration

Figure S.21: Gaussian MIL



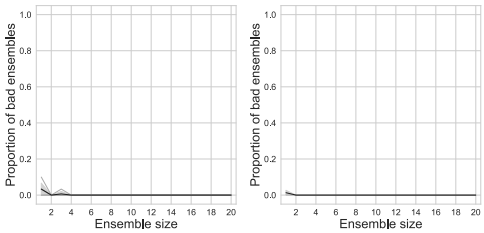
(a) Single-Configuration (b) Multi-Configuration

Figure S.22: Gaussian AND



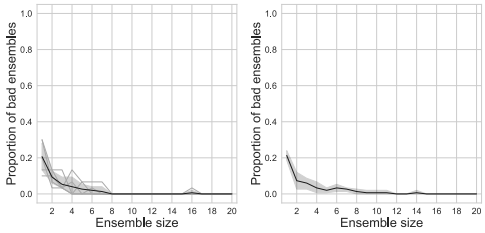
(a) Single-Configuration (b) Multi-Configuration

Figure S.23: Gaussian XOR



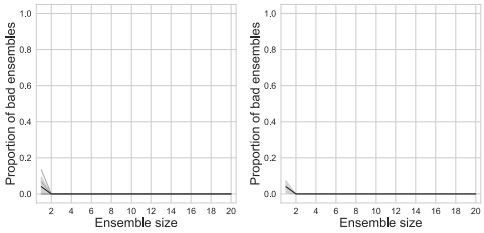
(a) Single-Configuration (b) Multi-Configuration

Figure S.24: MNIST MIL



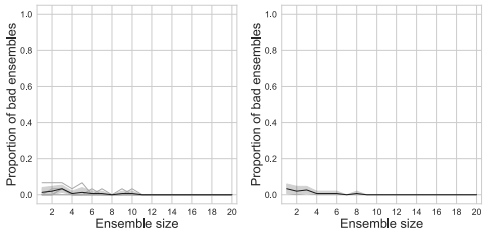
(a) Single-Configuration (b) Multi-Configuration

Figure S.25: MNIST AND



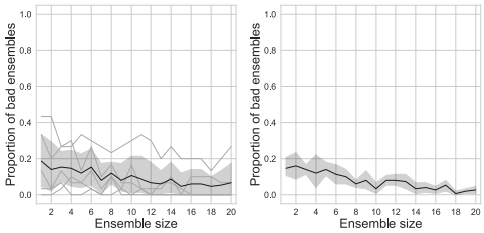
(a) Single-Configuration (b) Multi-Configuration

Figure S.26: MNIST XOR



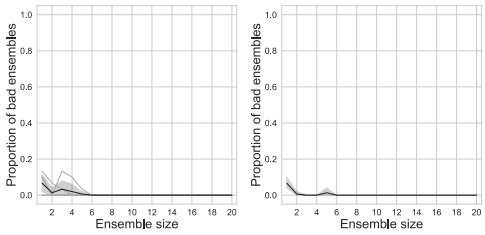
(a) Single-Configuration (b) Multi-Configuration

Figure S.27: CyTOF MIL



(a) Single-Configuration (b) Multi-Configuration

Figure S.28: CyTOF AND



(a) Single-Configuration (b) Multi-Configuration

Figure S.29: CyTOF XOR

S.6 Impact of instance proportion on metrics

Comparison of Accuracy, iAUC-ROC and iAUC-PR metrics for unbalanced bags in terms of negative to positive proportions. The 95% confidence interval is indicated by the light area.

Parameter	Values
Batch size	100
Epoch	1000
Learning rate	0.005
Weight decay	0.0001
Loss function	Cross-entropy
Optim. algorithm	Adam
Featurizer layers	4, 8, 8
Attention layers	8, 8, 1
Classifier layers	8, 2

Table S.1: Parameters used to investigate the impact of instance proportion with Gaussian data.

Parameter	Values
Batch size	100
Epoch	1000
Learning rate	0.02
Weight decay	0.0001
Loss function	Cross-entropy
Optim. algorithm	Adam
Featurizer layers	1, 8, 8
Attention layers	8, 8, 1
Classifier layers	8, 2

Table S.2: Parameters used to investigate the impact of instance proportion with MNIST data.

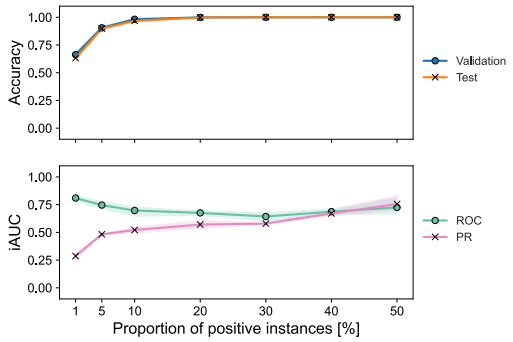


Figure S.30: Gaussian MIL

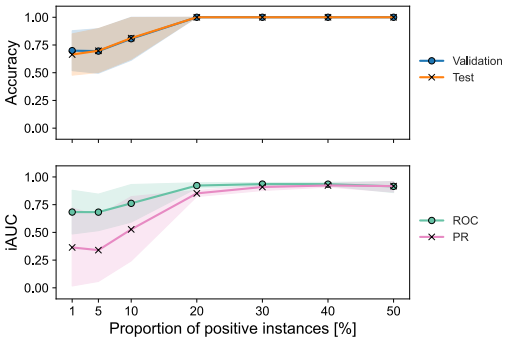


Figure S.31: MNIST MIL

S.7 Impact of attention network architecture on metrics

Comparison of Accuracy, iAUC-ROC and iAUC-PR metrics different sizes of attention network hidden layer. The 95% confidence interval is indicated by the light area.

Parameter	Values
Batch size	100
Epoch	1000
Learning rate	0.005
Weight decay	0.0001
Loss function	Cross-entropy
Optim. algorithm	Adam
Featurizer layers	4, 8, 8
Attention layers	8, *, 1
Classifier layers	8, 2

Table S.3: Parameters used to investigate the impact of the attention network hidden layer size with Gaussian data. The star indicates the element which was varied.

Parameter	Values
Batch size	100
Epoch	1000
Learning rate	0.02
Weight decay	0.0001
Loss function	Cross-entropy
Optim. algorithm	Adam
Featurizer layers	1, 8, 8
Attention layers	8, *, 1
Classifier layers	8, 2

Table S.4: Parameters used to investigate the impact of the attention network hidden layer size with MNIST data. The star indicates the element which was varied.

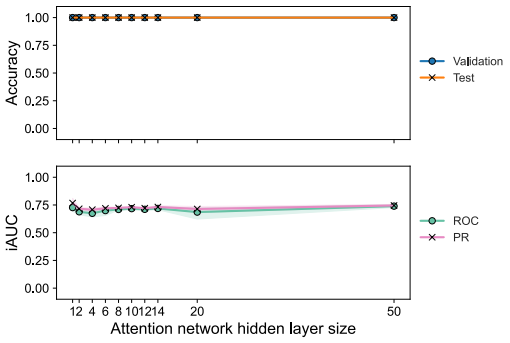


Figure S.32: Gaussian MIL

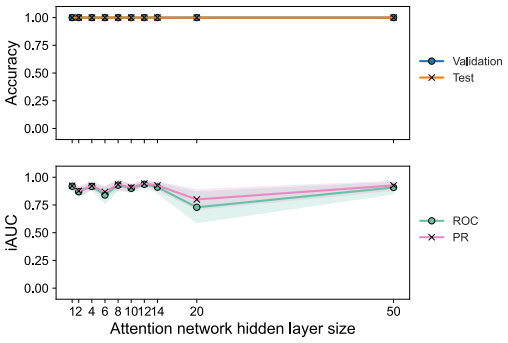


Figure S.33: MNIST MIL